Title
Sparsity Double Robust Inference of Average Treatment Effects

Authors
Bradic, Jelena
Wager, Stefan
Zhu, Yinchu

Publication Date
2019-05-02

Peer reviewed

Sparsity Double Robust Inference of Average Treatment Effects

Jelena Bradic*, Stefan Wager†, and Yinchu Zhu‡

*University of California, San Diego
†Stanford University
‡University of Oregon

May 3, 2019

Abstract

Many popular methods for building confidence intervals on causal effects under high-dimensional confounding require strong "ultra-sparsity" assumptions that may be difficult to validate in practice. To alleviate this difficulty, we here study a new method for average treatment effect estimation that yields asymptotically exact confidence intervals assuming that either the conditional response surface or the conditional probability of treatment allows for an ultra-sparse representation (but not necessarily both). This guarantee allows us to provide valid inference for average treatment effect in high dimensions under considerably more generality than available baselines. In addition, we showcase that our results are semi-parametrically efficient.

1 Introduction

Average treatment effect estimation is a core problem in causal inference, and has been the topic of a considerable amount of recent literature [Imbens and Rubin, 2015]. In this paper, we focus on the task average treatment effect estimation with high-dimensional confounders: We have access to n *i.i.d.* samples $(X_i, Y_i, W_i) \in \mathcal{X} \times \mathbb{R} \times \{0, 1\}$, where X_i denotes high-dimensional pre-treatment features ($\mathcal{X} \subset \mathbb{R}^p$ with $p \gg n$), W_i is the treatment assignment, and Y_i is our outcome of interest. Causal effects are defined via potential outcomes $\{Y_i(0), Y_i(1)\}$, such that we observe $Y_i = Y_i(W_i)$ and the average treatment effect is defined as $\tau = \mathbb{E}Y_i(1) - Y_i(0)$ [Neyman, 1923, Rubin, 1974]. Finally, we assume that there are no unmeasured confounders, i.e., the treatment assignment W_i may not be randomized, but can be treated as such once we control for X_i, i.e., $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i$ [Rosenbaum and Rubin, 1983]. Throughout, we also assume overlap, such that $\eta \leq \mathbb{P}\left(W_i \,\big|\, X_i = x\right) \leq 1 - \eta$ for all x and some $\eta > 0$.

In the low-dimensional case, one of the most prominent approaches to average treatment effect estimation is via augmented inverse-propensity weighting [Robins et al., 1994],

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}\left(X_i\right) - \hat{\mu}_{(0)}\left(X_i\right) + \frac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i)} \left(Y_i - \hat{\mu}_{(W_i)}(X_i)\right) \right), \tag{1}$$

where $e(x) = \mathbb{P}\left(W_i \,\middle|\, X_i = x\right)$ is the propensity score, $\mu_{(w)}(x) = \mathbb{E}\left(Y_i(w) \,\middle|\, X_i = x\right)$ are conditional response surfaces, and the quantities above with hats are estimates thereof. A celebrated property of this estimator is that it is double robust, meaning that it is consistent whenever either $\hat{e}(x)$ or the $\hat{\mu}_{(w)}(x)$ are consistent [Scharfstein et al., 1999]. Moreover, $\hat{\tau}$ is $\sqrt{n}$-consistent and semiparametrically efficient whenever the following risk bounds hold [Farrell, 2015]

$$\mathbb{E}\big(\hat{\mu}_{(W)}(X) - \mu_{(W)}(X)\big)^2 \mathbb{E}(\hat{e}(X) - e(X))^2 = o\left(\frac{1}{n}\right). \tag{2}$$

This statement is not sensitive to the structure of the estimators $\hat{e}(x)$ or the $\hat{\mu}_{(w)}(x)$ provided we use an appropriate type of sample splitting [Chernozhukov et al., 2018a, Zheng and van der Laan, 2011], and thus allows for considerable methodological flexibility. For example, Farrell et al. [2018] establish conditions under which (2) holds when $\hat{e}(x)$ or the $\hat{\mu}_{(w)}(x)$ are fit using neural networks. These results on augmented inverse-propensity weighting can also be applied when X_i is high dimensional; however, in this case, the required risk bound can be difficult to satisfy. In particular, except in extreme cases, the condition (2) effectively requires both $\mu_{(w)}(x)$ and $e(x)$ to admit very sparse representations.

In this paper, we study a doubly robust construction that is specifically designed for the high-dimensional case, and can be used for valid inference of τ under substantially weaker sparsity assumptions than standard augmented inverse-propensity weighting. We focus on the case where $\mu_{(w)}(x)$ and $e(x)$ have a high dimensional linear-logistic specification (we omit intercepts for conciseness of presentation),

$$\mu_{(w)}(x) = x'\beta_{(w)}, \quad e(x) = 1/\left(1 + \exp(-x'\theta)\right), \quad \beta_{(w)}, \theta \in \mathbb{R}^p, \tag{3}$$

and consider an estimator that is $\sqrt{n}$-consistent for τ under the condition that either θ or the $\beta_{(w)}$ (but not necessarily both) satisfy the type of sparsity condition that is usually required for high-dimensional inference [Javanmard and Montanari, 2014, van de Geer et al., 2014, Zhang and Zhang, 2014]. We refer to this property as sparsity double robustness.

The issue of sparsity doubly robustness has been an open question since the recent development of high-dimensional inference. This literature requires sparsity level $o(\sqrt{n}/\log p)$ for inference, a condition stronger than $o(n/\log p)$ needed for consistent estimation. Such a gap has only been addressed very recently in Javanmard and Montanari [2018], who found that the sparsity level of only one parameter needs to satisfy $o(\sqrt{n}/\log p)$, not both. However, their work only addresses the linear models and heavily relies on the Gaussianity assumption of the design. In this paper, we show that such sparsity doubly robustness result holds true for nonlinear models without Gaussian designs.

Our method starts with a functional form that closely resembles (1). However, we choose our estimators of $\mu_{(w)}(x)$ and $e(x)$ in ways that carefully exploit the geometry of sparseness in (3) and are thus able to improve on its performance. A closely related estimator has been independently studied by Tan [2019+], who considered potentially misspecified models but did not provide results on sparsity doubly robustness. Our main construction is as follows, modulo some algorithmic tweaks (including a type of sample splitting):

$$\hat{\theta}_{(w)} = \text{argmin}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}\{W_i \neq w\} X_i' \theta + \mathbb{1}\{W_i = w\} \exp\left(-X_i'\theta\right)\right) + \lambda_\theta \left\|\theta\right\|_1 \right\} \tag{4}$$

$$\hat{\beta}_{(w)} = \text{argmin}_{\beta} \left\{ \frac{1}{n} \sum_{W_i = w} \exp(-X_i'\hat{\theta}_{(w)}) \left(Y_i - X_i'\beta\right)^2 + \lambda_\beta \left\|\beta\right\|_1 \right\} \tag{5}$$

$$\begin{aligned} \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \Big[& \Big(X_i'\hat{\beta}_{(1)} + W_i[1 + \exp(-X_i'\hat{\theta}_{(1)})](Y_i - X_i'\hat{\beta}_{(1)}) \Big) \\ & - \Big(X_i'\hat{\beta}_{(0)} + (1 - W_i)[1 + \exp(-X_i'\hat{\theta}_{(0)})](Y_i - X_i'\hat{\beta}_{(0)}) \Big) \Big]. \end{aligned} \tag{6}$$

As discussed in Section 2, we can study this estimator from two different perspectives. If $\beta_{(w)}$ is very sparse, then the solution to (5) converges at a fast rate, while the solution to the propensity model (4) effectively debiases $\hat{\beta}_{(w)}$ even if $\hat{\theta}_{(w)}$ is not particularly accurate. Meanwhile, if $\theta_{(w)}$ is very sparse, then the converse holds. Our proof exploits this idea to establish sparsity double robustness.

The idea of fitting a propensity model that can also leverage the shape of the conditional response surface has generated considerable interest in recent years. The key observation here is that, in addition to being a consistent estimator when θ is very sparse, (4) also "balances" the inverse-propensity weighted features among the treated and control samples in finite samples [Chan et al., 2015, Hainmueller, 2012, Imai and Ratkovic, 2014, Tan, 2017, Zhao, 2019]

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx \frac{1}{n} \sum_{W_i = w} \frac{X_i}{1 + \exp(-X_i'\hat{\theta}_{(w)})}. \tag{7}$$

The advantage of balancing is that, if the linear model for Y is well specified, then balancing as in (7) is sufficient for eliminating confounding, even when $\hat{\theta}_{(w)}$ itself may be inconsistent or misspecified [Athey et al., 2018, Hirshberg and Wager, 2018, Kallus, 2018, Zhao and Percival, 2017, Zubizarreta, 2015]. Note that, here, we estimate separate models for $\mathbb{P}\left(W_i = 0 \,\middle|\, X_i = x\right)$ and $\mathbb{P}\left(W_i = 1 \,\middle|\, X_i = x\right)$, parametrized by $\theta_{(0)}$ and $\theta_{(1)}$ respectively. This parametrization is based on (3) and reads

$$\mathbb{P}\left(W_i = w \,\middle|\, X_i = x\right) = 1/(1 + \exp(-x'\theta_{(w)})) \qquad \text{for} \qquad w \in \{0, 1\}.$$

Notice that by (3), we have that $\theta_{(1)} = \theta$ and $\theta_{(0)} = -\theta$. Asymptotically, we expect both parameter vectors to be consistent, $-\hat{\theta}_{(0)}, \hat{\theta}_{(1)} \approx \theta$, but finite-sample differences between $\hat{\theta}_{(0)}$ and $\hat{\theta}_{(1)}$ play a key role in enabling the balance [Imai and Ratkovic, 2014].

Our main finding is that an estimator constructed via the above "balancing" principle achieves sparsity double robustness, meaning that it attains $\sqrt{n}$-consistency given strong enough sparsity assumptions $o(\sqrt{n}/\log p)$ on either θ or the $\beta_{(w)}$, but not necessarily both. As discussed further below, this property is considerably stronger than the standard double robustness property (2) in the high-dimensional setup (3).

1.1 Related Work

Double robust and/or semiparametrically efficient estimation has a long tradition in the literature on causal inference [Chernozhukov et al., 2018a, Farrell, 2015, Hahn, 1998, Hirano et al., 2003, Newey and Robins, 2018, Robins and Rotnitzky, 1995, Robins et al., 1994, Scharfstein et al., 1999, Tan, 2010, van der Laan and Rubin, 2006]. More recently, it has been shown that with high dimensional confounders, we can improve the behavior of double-robust-type estimators by having them directly exploit the geometry of sparsity.

As one of the first result in this direction, Athey et al. [2018] showed, given sufficient sparsity on the outcome function in (3), $\|\beta_{(w)}\|_0 \ll \sqrt{n}/\log(p)$, we can achieve $\sqrt{n}$-consistency without any assumptions on the propensity score beyond overlap by simply using weights that balance moments as follows (the $\hat{\beta}_{(w)}$ are estimated via the lasso):

$$\begin{aligned} \hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n} X_i' \left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right) + \hat{\gamma}_i(W_i)(2W_i - 1)\left(Y_i - X_i'\hat{\beta}_{(W_i)}\right), \\ \hat{\gamma}(w) &= \operatorname{argmin}_{\gamma}\left\{\frac{1}{n^2}\sum_{W_i = w}\gamma_i^2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\left(1 - \gamma_i \mathbb{1}\left\{W_i = w\right\}\right) X_i\right\|_\infty^2\right\}. \end{aligned} \tag{8}$$

Conceptually, this approach is related to several papers that stress the important of covariate balance for accurate estimation of treatment effects [Chan et al., 2015, Imai and Ratkovic, 2014, Kallus, 2018, Zhao, 2019, Zubizarreta, 2015]. Hirshberg and Wager [2018] establish conditions under which this estimator is efficient.

The main downside of the approximate residual balancing estimator (8) is that it always requires sparsity of the outcome model, and cannot use a well specified and sparse propensity model to compensate for a complex outcome model. Our sparsity double robustness result, which only requires strong sparsity of either θ or the $\beta_{(w)}$ in (3) directly addresses this limitation; and, as shown in our experiments, yields substantial gains in accuracy when θ is in fact sparse.

Our result is most closely related to a recent proposal by Chernozhukov et al. [2018b], who studied any linear functional whose Riesz representer admits an (approximate) linear representation. In another paper, Chernozhukov et al. [2018a] considers theoretical results for estimators based on learning conditional mean function and the propensity score. In both papers, the key condition is that the product of ℓ_2-loss for learning the two nuisance parameters is $o(n^{-1/2})$, a condition referred to as rate double robustness; see Definition 2 in Smucler et al. [2019]. Sufficient conditions for rate double robustness have been provided in these works in terms of sparsity levels. For example, Remark 5.2 of Chernozhukov et al. [2018a] shows that rate double robustness is guaranteed when the product of two sparsity levels is $o(n)$, while Remark 7 of Chernozhukov et al. [2018b] points out that under the assumption of bounded ℓ_1-norm of both parameters, rate double robustness holds whenever one of the sparsity levels is $o(\sqrt{n}/\log p)$.

The sparsity doubly robustness in this paper contributes to the literature by providing a different perspective. We show that efficient estimation is also possible in certain cases in which rate double robustness might not hold. One such example is when the logistic parameter has bounded ℓ_1-norm and has sparsity level $o(\sqrt{n}/\log p)$ and the conditional parameter has sparsity level $o(n^{3/4}/\log p)$ with potentially large ℓ_1-norm. In this example, we can still derive $1/\sqrt{n}$-consistency although we are not aware of any results that can guarantee rate double robustness.

In addition, our work is also different from Chernozhukov et al. [2018b] in terms of specification. In the context of average treatment effect estimation, the formulation in Chernozhukov et al. [2018b] means that we need there to exist (potentially sparse) vectors $\xi_{(0)}$ and $\xi_{(1)}$ whose ℓ_1-norms are bounded (see Definition 3 or 4 therein) as well as such that $|1/(1-e(x)) - x'\xi_{(0)}| \approx 0$ and $|1/e(x) - x'\xi_{(1)}| \approx 0$ uniformly across x. This may be a reasonable assumption if x was in fact constructed as a basis expansion of some simpler measured features; however, it appears to be difficult to justify more generally. One contribution of this paper relative to Chernozhukov et al. [2018b] is that we achieve sparsity double robustness using the natural linear-logistic specification (3).

We also note two recent papers that consider estimators that resemble ours. Ning et al. [2018] consider an estimator that, in the spirit of Belloni et al. [2014], first fit a penalized covariate-balancing propensity model, and then re-fit without penalty those coefficients that correspond to features that are relevant to outcome modeling. Meanwhile, Tan [2019+] augments a penalized covariate-balancing propensity model in an outcome regression; it turns out that his covariate-balancing mechanism designed to address the issue of misspecification is also helpful for relaxing sparsity requirements. Neither paper, however, achieves sparsity double robustness as discussed here; rather, they require both the outcome parameter vector β and the propensity parameter vector θ to be ultra-sparse—or, if there is misspecification they require the population minimizers of both the outcome and propensity loss functions to be ultra-sparse. Under the framework of Smucler et al. [2019], Rotnitzky et al. [2019], Ning et al. [2018], Tan [2019+] are classified as examples of model double robustness, which means that one of the models (either conditional mean or propensity score) is misspecified. Rate double robustness requires that the product of the ℓ_2-norms of the estimation errors in two models is of the order $o(n^{-1/2})$.

2 Sparsity Double Robust Estimation

Whenever a parameter is identified through a moment condition, like (1), a direct loss minimization that does not take into account this moment condition may not guarantee desirable properties. Controlling inferential features of high-dimensional estimates is extremely difficult; most, if not all, require strict sparsity conditions. We aim to control optimality at estimation by directly embedding the leading term of the bias into a constraint of newly designed estimators.

The main idea behind our construction is that we use estimators $\hat{\beta}_{(0)}$, etc., of $\beta_{(0)}$, etc., that have two complementary properties. When the underlying parameter $\beta_{(0)}$ is ultra-sparse, then $\hat{\beta}_{(0)}$ converges to $\beta_{(0)}$ in ℓ_1-norm. Furthermore, even when $\beta_{(0)}$ is not ultra-sparse, $\hat{\beta}_{(0)}$ still has a useful covariate-balancing property implied by its Karush-Kuhn-Tucker (KKT) conditions that can be put to good use (and similar guarantees hold for $\hat{\beta}_{(1)}$, $\hat{\theta}_{(0)}$ and $\hat{\theta}_{(1)}$). We then provide two separate consistency and asymptotic normality proofs for our estimator: One that assumes that $\beta_{(w)}$ is ultra-sparse and relies on KKT conditions for the $\hat{\theta}_{(w)}$ estimator to debias a very accurate $\hat{\beta}_{(w)}$ estimator, and a second that assumes that θ is ultra-sparse and relies KKT conditions for the $\hat{\beta}_{(w)}$ estimator to debias a very accurate $\hat{\theta}_{(w)}$ estimator. Of course, only one of these arguments needs to hold for us to achieve asymptotic normality, and thus our estimator is sparsity double robust. This argument was inspired by the one used by Chernozhukov et al. [2018b]; however, as discussed in the related works section, Chernozhukov et al. [2018b] make the somewhat unusual assumption that $1/e(x)$ can be approximated by a sparse linear model (rather than the assumption we make here, i.e., a sparse logistic model for $e(x)$).

2.1 Doubly Robust Balancing via Moment Targeting

In this section, we briefly sketch the argument behind our main formal result, and use it to motivate the form of our estimator. One of the main ingredients is moment targeting: We design estimators such that they satisfy certain moment conditions that help reduce the bias at estimation. The construction for estimators for $\beta_{(w)}$ and $\theta_{(w)}$ is based on the structure of the bias in the final estimator for $\mathbb{E}\mu_{(w)}(X_i)$. We emphasize that the argument here is only heuristic; formal arguments are given in the appendix.

Given these preliminaries, observe that the treatment effect estimator under consideration can be written in a familiar form

$$\hat{\tau} = \hat{\mu}_{(1)} - \hat{\mu}_{(0)},$$

where $\hat{\mu}_{(w)}$ is an estimate of $\mu_{(w)} = \mathbb{E}Y_i(w)$. We use

$$\hat{\mu}_{(w)} = \frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\beta}_{(1)} + \hat{\gamma}_i(w)\mathbb{1}\{W_i = w\}(Y_i - X_i'\hat{\beta}_{(w)}), \qquad \hat{\gamma}_i(w) = 1 + \exp(-X_i'\hat{\theta}_{(w)}),$$

and construct $\hat{\mu}_{(0)}$ analogously. Our goal is to choose $\hat{\theta}_{(w)}$ (and hence $\hat{\gamma}_i(w)$) such as to control the errors $\hat{\mu}_{(w)} - \mu_{(w)}$ under flexible sparsity conditions.

To motivate our choice of $\hat{\theta}_{(1)}$, let us first consider the case where $\beta_{(1)}$ is very sparse, i.e., $\|\beta_{(1)}\|_0 \ll \sqrt{n}/\log p$. Notice that

$$\begin{aligned}\hat{\mu}_{(1)} - \mu_{(1)} = n^{-1}\sum_{i=1}^{n}(X_i'\beta_{(1)} - \mu_{(1)}) + n^{-1}\sum_{i=1}^{n} W_i\varepsilon_{i,(1)}\hat{\gamma}_i(1) \\ + n^{-1}\sum_{i=1}^{n}\left[1 - W_i\hat{\gamma}_i(1)\right] X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right),\end{aligned}$$

where $\varepsilon_{i,(w)} = Y_i(w) - X_i'\beta_{(w)}$. The first two terms on the right hand side are asymptotically normal with mean zero under weak consistency conditions on $\hat{\theta}_{(1)}$ that only require a moderate amount of sparsity on θ. Meanwhile, the last term can be bounded using Holder's inequality,

$$\begin{aligned}&\left|n^{-1}\sum_{i=1}^{n}\left[1 - W_i\hat{\gamma}_i(1)\right] X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right)\right| \\ &= \left|n^{-1}\sum_{i=1}^{n}\left[1 - W_i(1 + \exp(-X_i'\hat{\theta}_{(1)}))\right] X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right)\right| \\ &\leq \left\|n^{-1}\sum_{i=1}^{n}\left[1 - W_i(1 + \exp(-X_i'\hat{\theta}_{(1)}))\right] X_i\right\|_\infty \left\|\hat{\beta}_{(1)} - \beta_{(1)}\right\|_1.\end{aligned}$$

Under sparsity assumption $\|\beta_{(1)}\|_0 = o(\sqrt{n}/\log p)$, we can typically obtain $\|\hat{\beta}_{(1)} - \beta_{(1)}\|_1 = o_P(1/\sqrt{\log p})$ via sparse methods [e.g., Negahban et al., 2012]. Meanwhile, the KKT conditions for the estimator in (4) with $w = 1$ automatically yields [Tan, 2017]

$$\left\|n^{-1}\sum_{i=1}^{n}\left[1 - W_i(1 + \exp(-X_i'\hat{\theta}_{(1)}))\right] X_i\right\|_\infty = O_P(\sqrt{n^{-1}\log p}),$$

thus bounding the bias to the order of $o_P(n^{-1/2})$. This is the first example of moment targeting. The KKT condition of the estimator provides a convenient moment condition for the purpose of bias reduction.

The above argument closely mirrors the argument used by Athey et al. [2018] to obtain $\sqrt{n}$-consistent estimates of τ when $\|\beta_{(w)}\|_0 \ll \sqrt{n}/\log p$. The main difference with our approach is that Athey et al. [2018] do not fit a model for θ, but instead directly optimize the weights $\hat{\gamma}$ via quadratic programming as in Javanmard and Montanari [2014] and Zubizarreta [2015]. That in turn, leads to somewhat loss of flexibility whenever the outcome model is not sparse.

Here, the fact that we also model θ enables us to alternatively exploit sparsity in θ and correspondingly relax assumptions on $\beta_{(1)}$. To do so, note that

$$\begin{aligned}\hat{\mu}_{(1)} - \mu_{(1)} = n^{-1}\sum_{i=1}^{n}(X_i'\beta_{(1)} - \mu_{(1)}) + n^{-1}\sum_{i=1}^{n} W_i\varepsilon_{i,(1)}\hat{\gamma}_i(1) \\ + n^{-1}\sum_{i=1}^{n}\left[1 - W_i(1+\exp(-X_i'\theta))\right] X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right) \\ + n^{-1}\sum_{i=1}^{n} W_i\left[\exp(-X_i'\theta) - \exp(-X_i'\hat{\theta}_{(1)})\right] X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right).\end{aligned}$$

Again, the sum of the first three terms above is asymptotically Gaussian on the $\sqrt{n}$-scale under only weak assumptions on $\hat{\beta}_{(1)}$. To handle the last term, we can use Taylor expansion to argue that (we will make this rigorous in the proof of our main result)

$$\begin{aligned}&\left| n^{-1}\sum_{i=1}^{n} W_i \exp(-X_i'\hat{\theta}_{(1)})X_i'(\hat{\theta}_{(1)} - \theta)X_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right)\right| \\ &\qquad \lesssim \left\| n^{-1}\sum_{i=1}^{n} W_i\exp(-X_i'\hat{\theta}_{(1)})X_iX_i'\left(\hat{\beta}_{(1)} - \beta_{(1)}\right)\right\|_\infty \left\|\hat{\theta}_{(1)} - \theta\right\|_1.\end{aligned}$$

Now, given sufficient sparsity on θ, i.e., $\|\theta\|_0 \ll \sqrt{n}/\log p$ we can verify that $\|\hat{\theta}_{(1)} - \theta\|_1 = o_P(1/\sqrt{\log p})$. Meanwhile, the the KKT condition for the estimator $\hat{\beta}_{(1)}$ in (5) automatically yields that the first component above is $O_P(\sqrt{n^{-1}\log p})$. Thus, we also expect $\hat{\mu}_{(0)}$ to be accurate when θ is very sparse, even if β is not. This is another example of moment targeting in that the KKT condition for $\hat{\beta}_{(1)}$ again provides a convenient bound for bounding the bias.

2.2 Sample splitting for optimality

The above discussion provides some helpful conceptual guidance on how to pick good estimators of the unknown $\beta_{(w)}$ and $\theta_{(w)}$. To achieve optimality in most general terms, we invoke a special scheme of sample-splitting similar to cross-fitting. Under the usual cross-fitting scheme, the influence function is evaluated on observations that are not used to estimate the nuisance parameters (in our case $\beta_{(w)}, \theta_{(w)}$). Cross-fitting has been used to reduce bias terms in many semiparametric and high-dimensional models [see, e.g., Chernozhukov et al., 2018a, Newey and Robins, 2018, Schick, 1986, Zheng and van der Laan, 2011]. Here, our approach requires us to only cross-fit $\hat{\beta}_{(w)}$, but not $\hat{\theta}_{(w)}$.

The entire sample is divided into two parts $\mathcal{J}$ and $\mathcal{J}^c$. For $F \in \{\mathcal{J}, \mathcal{J}^c,\}$, estimators trained using the sample F are denoted with $\hat{\beta}_{(w),F}$ and $\hat{\theta}_{(w),F}$, respectively. For expositional simplicity, we assume that $|\mathcal{J}| = |\mathcal{J}^c| = n/2$. Then, for $(w, F) \in \{0,1\} \times \{\mathcal{J}, \mathcal{J}^c\}$, we define the estimator of the mean

$$\hat{\mu}_{(w),F} = \frac{1}{|F|} \sum_{i \in F} \left(X_i' \hat{\beta}_{(w),F^c} + \hat{\gamma}_i(w,F) \mathbb{1}\{W_i = w\}(Y_i - X_i' \hat{\beta}_{(w),F^c}) \right) \tag{9}$$

where the weight function is defined in-sample

$$\hat{\gamma}_i(w,F) = 1 + \exp(-X_i' \hat{\theta}_{(w),F}),$$

$\hat{\theta}_{(w),F}$ is defined in Algorithm 1, and $\hat{\beta}_{(w),F}$ is given by

$$\hat{\beta}_{(w),F} = \arg\min_{\beta} \left\{ \frac{1}{|F|} \sum_{i \in F} \mathbb{1}\{W_i = w\} \exp(-X_i' \hat{\theta}_{(w),F}) \left(Y_i - X_i' \beta\right)^2 + \lambda_\beta \|\beta\|_1 \right\}. \tag{10}$$

Algorithm 1 presents details of the propensity estimation. The loss functions in (10) and (11) were recently utilized in Tan [2019+] but the proposed average treatment effects estimator therein does not achieve sparsity double robustness.

Algorithm 1 Optimistic penalized covariate-balancing propensity estimation

---

Require: - a training sample $F \in \{\mathcal{J}, \mathcal{J}^c\}$, a treatment status indicator $w \in \{0,1\}$ a tuning parameter $\lambda_\theta \asymp \sqrt{\log(p)/n}$ and a pre-defined constant κ

Compute

$$\check{\theta}_{(w),F} \leftarrow \arg\min_{\theta} \left\{ \frac{1}{|F|} \sum_{i \in F} \left[\mathbb{1}\{W_i \neq w\} X_i' \theta + \mathbb{1}\{W_i = w\} \exp\left(-X_i' \theta\right) \right] + \lambda_\theta \|\theta\|_1 \right\} \tag{11}$$

if $\|\check{\theta}_{(w),F}\|_1 > \kappa$, **then**

$$\hat{\theta}_{(w),F} \leftarrow \arg\min_{\theta} \left\{ \|\theta\|_1, \text{ s.t. } \left\| \frac{1}{|F|} \sum_{i \in F} \left[1 - \mathbb{1}\{W_i = w\} \big(1 + \exp(-X_i' \theta)\big) \right] X_i \right\|_\infty \leq \lambda_\theta \right\}$$

else

$\hat{\theta}_{(w),F} \leftarrow \check{\theta}_{(w),F}$

end if

return $\hat{\theta}_{(w),F}$

The method presented here splits the sample into two subsamples. Although one can easily follow the same principle and split the sample into multiple subsamples, we do not pursue this option here for notational simplicity. We now define

$$\hat{\mu}_{(1)} = (\hat{\mu}_{(1),\mathcal{J}} + \hat{\mu}_{(1),\mathcal{J}^c})/2.$$

Similarly, we can define $\hat{\mu}_{(0)}$. Then, the average treatment effect estimator is defined as

$$\hat{\tau} = \hat{\mu}_{(1)} - \hat{\mu}_{(0)}. \tag{12}$$

3 Formal Results

We now turn to a formal characterization of the average treatment effect estimator in (12), with the aim of providing asymptotic Gaussianity whenever one, but not both, of θ, β is estimated consistently. We begin by listing some theoretical assumptions necessary for the development of the theoretical guarantees.

First, we assume that the covariate space and the parameter space are both subsets of Euclidean space; specifically, we assume that $X \in [a,b]^p$ and $\theta \in \mathcal{B}_1(r) \subset \mathbb{R}^p$ for some bounded $r > 0$, where $\mathcal{B}_1(r)$ is an ℓ_1 ball with radius r.

The results discussed below hold whenever, the tuning parameters (in Algorithm 1) are chosen to be proportional to $\sqrt{\log(p)/n}$. Moreover, we also assume that κ is chosen to be larger than $\|\theta_{(w)}\|_1$. Our procedure is not particularly sensitive to the choice of κ; in practice it suffices to choose a large enough number.

Assumption 1 (Eigenvalue)**.** *The minimum and maximum eigenvalues of $\mathbb{E}[X_i X_i']$ are contained in a bounded interval that does not contain zero.*

Our next assumption controls the regularity properties of the errors within both models (3). Let $\varepsilon_{i,(w)} = Y_i - X_i'\beta_{(w)}$ and $v_{i,(w)} = \mathbb{1}\{W_i = w\} - e_{(w)}(X_i)$. Note that in the context of models (3) the unconfoundedness assumption implies $\varepsilon_{i,(w)} \perp v_{i,(w)} | X_i$ and from now on we will work with this slightly weaker assumption.

Assumption 2 (Model)**.** *X_i has a bounded sub-Gaussian norm. Moreover, for $w \in \{0,1\}$, $\varepsilon_{i,(w)}$ is sub-Gaussian.*

Now, observe that Assumption 2 is very weak and in particular it is not implying consistent estimation in the outcome model. The boundedness of $\|X_i\|_\infty$ and $\|\theta\|_1$ guarantees the overlap condition.

Finally, in the context of the average treatment effects, in order to provide confidence intervals an estimate of the asymptotic variance of $\hat{\tau}$ is needed. We show an asymptotic variance of $\hat{\tau}$ takes the form of

$$\begin{aligned}
&\mathbb{E}\left[X_i'\left(\beta_{(1)} - \beta_{(0)}\right) - \tau\right]^2 + \mathbb{E}\left[W_i \varepsilon_{i,(1)} \gamma_i(1)\right]^2 + \mathbb{E}\left[(1 - W_i)\varepsilon_{i,(0)}\gamma_i(0)\right]^2 \\
&:= \Omega + V_{(1)} + V_{(0)}.
\end{aligned}$$

Observe that Ω is the variability induced primarily from the variability of the design X. The other two terms can be viewed as properly normalized unexplained variance of the models (3).

To define variance estimates, we define estimates of Ω, $V_{(1)}$ and $V_{(0)}$ separately. We set

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} \left(X_i'(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}) - \hat{\tau} \right)^2, \tag{13}$$

as well as

$$\begin{aligned}
&\hat{V}_{(w)} = n^{-1} \sum_{i=1}^{n} (2W_i - 1)^2 \hat{\varepsilon}_{i,(w)}^2 \hat{\gamma}_i^2(w) \mathbf{1}\{W_i = w\}, \\
&\hat{\varepsilon}_{i,(w)} = Y_i - X_i'\hat{\beta}_{(w)}, \qquad \hat{\gamma}_i(w) = 1 + e^{-(2W_i - 1)X_i'\hat{\theta}_{(w)}}.
\end{aligned} \tag{14}$$

In the above display, $\hat{\beta}_{(w)}$ and $\hat{\theta}_{(w)}$ could be the ones computed on one sample, $\mathcal{J}$ or could be different ones; for example to borrow strength across samples we consider

$$\hat{\beta}_{(w)} = (\hat{\beta}_{(w),\mathcal{J}} + \hat{\beta}_{(1),\mathcal{J}^c})/2, \qquad \hat{\theta}_{(w)} = (\hat{\theta}_{(w),\mathcal{J}} + \hat{\theta}_{(w),\mathcal{J}^c})/2.$$

Now, we define the variance estimate as

$$\hat{V} = \hat{\Omega} + \hat{V}_{(0)} + \hat{V}_{(1)} \tag{15}$$

for $\hat{\Omega}$ and $\hat{V}_{(w)}$ defined in (13) and (14), respectively. We show that the construction above is appropriate for such circumstances and leads to asymptotically optimal confidence sets.

Theorem 1. *Let Assumptions 1 and 2 hold. Then, as long as one of the following two conditions holds,*

(i) (Ultra-sparse outcome model) $\|\beta_{(w)}\|_0 = o(\sqrt{n}/\log p)$ *and* $\|\theta\|_0 = o(n/\log p)$,

(ii) (Ultra-sparse propensity model) $\|\theta\|_0 = o(\sqrt{n}/\log p)$ *and* $\|\beta_{(w)}\|_0 = O(n^{3/4}/\log p)$,

we have the following representation of $\hat{\tau}$ as defined in (12)

$$\sqrt{n}(\hat{\tau} - \tau) = n^{-1/2} \sum_{i=1}^{n} \psi(Y_i, W_i, X_i, \tau, e_{(1)}(X_i)) + o_P(1),$$

where

$$\begin{aligned} &\psi(Y_i, W_i, X_i, \tau, e_{(1)}(X_i)) = \\ &\qquad X_i' \left(\beta_{(1)} - \beta_{(0)}\right) + W_i \left(\frac{Y_i - X_i'\beta_{(1)}}{e_{(1)}(X_i)}\right) - (1 - W_i) \left(\frac{Y_i - X_i'\beta_{(0)}}{1 - e_{(1)}(X_i)}\right) - \tau. \end{aligned} \tag{16}$$

In particular, under these assumptions we have

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, V_*), \qquad V_* = \mathbb{E}\left[\psi^2(Y_i, W_i, X_i, \tau, e_{(1)}(X_i))\right].$$

Moreover, under the same assumptions, for $\hat{V}$ defined in (15) we have

$$\hat{V} = V_* + o_P(1),$$

and in turn $\sqrt{n}(\hat{\tau} - \tau)/\sqrt{\hat{V}} \xrightarrow{d} \mathcal{N}(0, 1)$.

By well known results [e.g., Hahn, 1998, Newey, 1994, Robins and Rotnitzky, 1995], ψ in (16) is the efficient influence function and V_* is the semiparametric efficiency lower bound. Therefore, our estimator $\hat{\tau}$ in (12) is a semiparametrically efficient estimator. As discussed before, the sparsity requirement in Theorem 1 is considerably weaker that needed by existing estimators in the high-dimensional linear-logistic model (3), including the methods discussed in Athey et al. [2018], Belloni et al. [2014], Farrell [2015], Ning et al. [2018] and Tan [2019+], in that we only need either the outcome model or the propensity model to be ultra sparse (but not both).

4 Numerical Experiments

In this section we present numerical work where we contrast the behavior of the introduced method with the existing approaches. We consider the following design setting, $X_i \sim N(0, \Sigma)$ with $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho = 0.6$. We set the sample size and the number of covariates to be $n = 500$ and $p = 600$, respectively. The following structure for the parameters is used. We consider the propensity model where

$$\theta = a_\theta(1, 0, 1, 0, 1, 0, ..., 0, 1, 0, 0..., 0)'$$

with $\|\theta\|_0 = s_\theta$. We vary the value of s_θ and set a_θ such that $\sqrt{\theta' \Sigma_X \theta} = 1$.

Similarly, for the outcome model we consider

$$\beta_{(1)} = a_\beta(1, 0, 1, 0, 1, 0, ..., 0, 1, 0, 0..., 0)'$$

and set $\beta_{(0)} = -\beta_{(1)}$. In other words, non-zero entries appear only on indices with odd numbers. For a_β, we consider two cases. In the first, we have homoskedastic Erros: The error term is generated from a centered $\chi^2(1)$ distribution (since it is light tailed and asymmetric). We set $\sqrt{\beta'_{(1)} \Sigma_X \beta_{(1)}} = \sqrt{2}$. We use $\sqrt{2}$ to get an R^2 of 50% (because the error has variance 2). The second has heteroskedastic errors: $\varepsilon_{i,(1)}$ is generated according to

$$(4 \times \mathbf{1}\{e(X_i) \le 0.5\} + \mathbf{1}\{e(X_i) > 0.5\})\, \xi_i$$

with ξ_i being a centered $\chi^2(1)$ variable independent of X_i. Observe that $\varepsilon_{i,(0)}$ is still a centered $\chi^2(1)$ variable. We also consider other values of a_β such that the R^2 in the homoskedastic case is 10%. We report the mean squared error (MSE) and coverage probability of 95% confidence interval (CP). We compare our methods with two popular alternatives:

- **AIPW.** Augmented inverse propensity weighting [Robins et al., 1994] is a popular method for estimating the average treatment effect. We implement this using the hdm package in R.
- **ARB.** Approximate residual balancing was proposed by Athey et al. [2018]. This method can handle cases in which the propensity score is hard to estimate.

Table 1: Mean squared error (MSE) and coverage probability (CP) across two models both with R-squared= 0.5. Comparison includes augmented inverse propensity weighting (AIPW), approximate residual balancing (ARB), and sparsity double robust estimation (SDR) in (12). Parameters s_θ and s_β denote sparsity of the propensity and outcome models, respectively.

Homoscedastic errors

	$s_\theta = 2,\ s_\beta = 2$		$s_\theta = 2,\ s_\beta = 30$	
	MSE	CP	MSE	CP
AIPW	0.041	0.954	0.063	0.950
ARB	0.039	0.928	0.043	0.916
SDR	0.042	0.952	0.037	0.960
	$s_\theta = 30,\ s_\beta = 2$		$s_\theta = 30,\ s_\beta = 30$	
	MSE	CP	MSE	CP
AIPW	0.058	0.932	0.097	0.954
ARB	0.039	0.882	0.043	0.924
SDR	0.035	0.972	0.038	0.968

Heteroscedastic errors

	$s_\theta = 2,\ s_\beta = 2$		$s_\theta = 2,\ s_\beta = 30$	
	MSE	CP	MSE	CP
AIPW	0.223	0.920	0.191	0.958
ARB	0.145	0.922	0.129	0.950
SDR	0.132	0.972	0.081	0.990
	$s_\theta = 30,\ s_\beta = 2$		$s_\theta = 30,\ s_\beta = 30$	
	MSE	CP	MSE	CP
AIPW	0.245	0.934	0.258	0.962
ARB	0.113	0.948	0.102	0.946
SDR	0.080	0.992	0.074	0.994

The results are reported in Tables 1 and 2. Table 1 indicated that in the baseline case with extremely sparse $\beta_{(w)}$ and θ, AIPW, approximate residual balancing method and SDR perform very similarly. Note that this is as expected; all of the methods should be achieving the same asymptotic variance. However, when either the propensity score model or the conditional mean function or both are not extremely sparse, the SDR method delivers smaller MSE. This confirms our theoretical results, which state that our method is guaranteed to provide efficient estimation even if there is lack of extreme sparsity.

Perhaps a more direct way of formalizing this intuition is through analyzing the rate of the remainder similar to the discussion in [Newey and Robins, 2018]; one way is to express the rate of the remainder for the asymptotic expansion in Theorem 1 in terms of $\|\beta_{(w)}\|_0$ and $\|\theta\|_0$. One can use our technical arguments to show that, compared to AIPW, the remainder of the SDR estimator has the same order or smaller order of magnitude. In Table

2, we also report the results by setting a_β such that in the homoscedasticity case we have an R-squared of 10%. The pattern is quite similar; we observe comparable performance when we have extreme sparsity in both models and the proposed method has lower MSE in the absence of such sparsity in either model.

Table 2: Mean squared error (MSE) and coverage probability (CP) across two models both with R-squared$=0.1$. Comparison includes augmented inverse propensity weighting (AIPW), approximate residual balancing (ARB), and sparsity double robust estimation (SDR) in (12). Parameters s_θ and s_β denote sparsity of the propensity and outcome models, respectively.

Homoscedastic errors

	$s_\theta=2$, $s_\beta=2$		$s_\theta=2$, $s_\beta=30$	
	MSE	CP	MSE	CP
AIPW	0.043	0.954	0.055	0.952
ARB	0.038	0.915	0.039	0.926
SDR	0.042	0.948	0.035	0.965
	$s_\theta=30$, $s_\beta=2$		$s_\theta=30$, $s_\beta=30$	
	MSE	CP	MSE	CP
AIPW	0.048	0.947	0.086	0.962
ARB	0.039	0.900	0.042	0.920
SDR	0.035	0.977	0.039	0.974

Heteroscedastic errors

	$s_\theta=2$, $s_\beta=2$		$s_\theta=2$, $s_\beta=30$	
	MSE	CP	MSE	CP
AIPW	0.206	0.936	0.211	0.932
ARB	0.116	0.950	0.132	0.946
SDR	0.073	0.980	0.079	0.990
	$s_\theta=30$, $s_\beta=2$		$s_\theta=30$, $s_\beta=30$	
	MSE	CP	MSE	CP
AIPW	0.207	0.934	0.204	0.936
ARB	0.105	0.942	0.122	0.928
SDR	0.057	0.992	0.060	0.996

Supplementary Materials

Supplementary materials collect details of the main results and proofs.

Notations: In the rest of the paper, we shall use the following notations. We will frequently use the function

$$q(z) := 1 + \exp(-z)$$

as well as $\dot{q}(z) = dq(z)/dz = -\exp(-z)$. We will use the notations H_A and H_B to denote $\mathcal{J}$ and $\mathcal{J}^c$, respectively; doing so allows us to easily see the symmetry between $\mathcal{J}$ and $\mathcal{J}^c$. Moreover, $b_n = n/2$, $b_n^{-1}\sum_{i\in H_A}$ and $b_n^{-1}\sum_{i\in H_B}$ will be denoted by $\mathbb{E}_{n,H_A}$ and $\mathbb{E}_{n,H_B}$. Here, we assume that n is an even number so b_n is an integer.

We define $\mathbb{S}^{p-1} = \{v \in \mathbb{R}^p : \ \|v\|_2 = 1\}$. For any $k_0 > 0$ and any $J \subseteq \{1, ..., p\}$, we define the cone set $\mathcal{C}(J, k_0) = \{x \in \mathbb{R}^p : \ \|x_{J^c}\|_1 \le k_0\|x_J\|_1\}$. We use $'$ to denote the transpose.

A Proof of Theorem 1

We notice that the KKT condition for $\hat{\beta}_{(w),F}$ defined in (10) reads

$$\left\|\mathbb{E}_{n,H_F} W_i \dot{q}(X_i'\hat{\theta}_{(1),F})(X_i'\hat{\beta}_{(1),F} - Y_i)X_i\right\|_\infty \le \lambda_\beta/4 \quad \text{for } F \in \{\mathcal{J}, \mathcal{J}^c\} = \{A, B\}, \tag{17}$$

Moreover, we notice that the solution for Algorithm 1 always satisfies

$$\left\|\mathbb{E}_{n,H_F}\left[1 - W_i q(X_i'\hat{\theta}_{(1),F})\right] X_i\right\|_\infty \le \lambda_\theta \quad \text{for } F \in \{\mathcal{J}, \mathcal{J}^c\} = \{A, B\} \tag{18}$$

Moreover, under Assumptions 1 and 2, we can define constants $M_1, ..., M_5 > 0$ such that the following conditions hold (which we will establish as a reasonable later on):

1. $\mathbb{P}(\|X\|_\infty \le M_1) = 1$ and $\|X_i\|_{\psi_2} \le M_1$ for some constant $M_1 > 0$.
2. $\mathbb{E}X_iX_i'W_i$, $\mathbb{E}X_iX_i'(1 - W_i)$, $\mathbb{E}X_iX_i'\exp(-X_i'\theta_{(1)})W_i$ have all the eigenvalues in a fixed interval $[M_2, M_3]$, where $M_2, M_3 > 0$ are constants.
3. For $j \in \{0, 1\}$, $\mathbb{E}(\varepsilon_{i,(j)} \mid X_i) = 0$ and there exists a constant $M_4 > 0$ such that $\mathbb{E}(\exp(t\varepsilon_{i,(j)}) \mid X_i) \le \exp(M_4t^2)$ $\forall t \in \mathbb{R}$.
4. $\|\theta_{(1)}\|_1 \le M_5$ for some constant $M_5 > 0$ and $\kappa_0 \ge M_5$ is suitably chosen.

A.1 Main results

Main body of the proof consists of three big components. Theorem 2 showcases the asymptotic normality result whenever the outcome model is ultra-sparse. Theorem 3 showcases the result of ultra-sparse propensity model. Theorem 1 is then completed with the help of Lemma 4 and Theorem 5 establishing consistency of estimation of the asymptotic variance.

Theorem 2 (Ultra-sparse outcome model)**.** *Let Assumptions 1 and 2 hold. Suppose that* $\|\beta_{(1)}\|_0 = o(\sqrt{n}/\log p)$ *and* $\|\theta_{(1)}\|_0 = o(n/\log p)$*. Then*

$$\sqrt{n}(\hat{\mu}_{(1)} - \mu_{(1)}) = n^{-1/2}\sum_{i=1}^{n}\left[W_i\varepsilon_{i,(1)}q(X_i'\theta_{(1)}) + \left(X_i'\beta_{(1)} - \mu_{(1)}\right)\right] + o_P(1).$$

Proof. See Section A.3. □

Theorem 3 (Ultra-sparse propensity model)**.** *Let Assumptions 1 and 2 hold. Suppose that* $\|\theta_{(1)}\|_0 = o(\sqrt{n}/\log p)$ *and* $\|\beta_{(1)}\|_0 = O(n^{3/4}/\log p)$*. Then*

$$\sqrt{n}(\hat{\mu}_{(1)} - \mu_{(1)}) = n^{-1/2}\sum_{i=1}^{n}\left[W_i\varepsilon_{i,(1)}q(X_i'\theta_{(1)}) + \left(X_i'\beta_{(1)} - \mu_{(1)}\right)\right] + o_P(1).$$

Proof. See Section A.4. □

Lemma 4. *Consider* V_* *defined in Theorem 1. Under the conditions of Theorem 1, we have*

$$V_* = \mathbb{E}W_i\varepsilon_{i,(1)}^2 q^2(X_i'\theta_{(1)}) + \mathbb{E}(1-W_i)\varepsilon_{i,(0)}^2 q^2(X_i'\theta_{(0)}) + \mathbb{E}(X_i'(\beta_{(1)} - \beta_{(0)}) - \tau)^2.$$

Proof. See Section A.5. □

Theorem 5. *Under the conditions of Theorem 1, we have* $\hat{V} = V_* + o_P(1)$*.*

Proof. See Section A.6. □

A.2 Auxiliary results

Proofs of the main results require a sequence of statements discussing properties of the newly proposed estimators.

A.2.1 Estimators for the propensity score's $\theta_{(1)}$

First we discuss the eigenvalue properties of various design matrices.

Lemma 6. *Let Assumptions 1 and 2 hold. Assume that* $s = o(n/\log p)$*. Then with probability approaching one, we have that*

1. $\mathbb{E}_{n,H_A}W_i(X_i'v)^2 \geq c_1^2\|v_J\|_2^2$ *for all* $v \in \bigcup_{|J|\leq s}\mathcal{C}(J,3)$ *and* $v \in \bigcup_{|J|\leq s}\mathcal{C}(J,1)$
2. $\mathbb{E}_{n,H_A}(X_i'v)^2/\|v\|_2^2 \leq c_2^2$ *for all* $v \in \bigcup_{|J|\leq s}\mathcal{C}(J,3)$.
3. $\|\mathbb{E}_{n,H_A}X_i(W_iq(X_i'\theta_{(1)}) - 1)\|_\infty \leq 0.5\lambda_\theta$ *for suitably chosen* $\lambda_\theta \asymp \sqrt{n^{-1}\log p}$
4. $\|\mathbb{E}_{n,H_A}X_iW_i\exp(X_i'\hat{\theta}_{(1),A})\varepsilon_{i,(1)}\|_\infty \leq \lambda_\beta/4$ *for suitably chosen* $\lambda_\beta \asymp \sqrt{n^{-1}\log p}$*,*

where $c_1, c_2, c_3 > 0$ *are constants depending only on* $M_1, ..., M_5$*. Analogous results hold if we replace* $\mathbb{E}_{n,H_A}$ *with* $\mathbb{E}_{n,H_B}$*.*

Proof of Lemma 6.
Proof of the first two claims. We invoke Theorem 16 of Rudelson and Zhou [2013]. Let $\Sigma_1 = \mathbb{E}W_iX_iX_i'$, $\tilde{X}_i = W_iX_i\Sigma_1^{-1/2}$ and $k_0 \in \{1,3\}$. Notice that $\tilde{X}_i$ is isotropic by definition. By the sub-Gaussian property of X_i and the assumption that eigenvalues of Σ_1 are bounded away from zero, it follows that $\tilde{X}_i$ also has bounded sub-Gaussian norm. For a fixed $\delta \in (0,1)$, we define $d(3k_0, A)$ as in Theorem 16 of Rudelson and Zhou [2013], where $A = \Sigma_1^{1/2}$. Clearly, $d(3k_0, A) \asymp s$ and $m \lesssim s$. Observe that Equation (39) therein holds because $n \gg s\log p$.

Therefore, by their Theorem 16, with probability approaching one,

$$1-\delta \leq \frac{\|\tilde{X}v\|_2/\sqrt{n}}{\|\Sigma_1^{1/2}v\|_2} \leq 1+\delta, \qquad \forall v \in \bigcup_{|J|\leq s} \mathcal{C}(J,k_0),$$

where $\tilde{X} = (\tilde{X}_1, ..., \tilde{X}_n)' \in \mathbb{R}^{n\times p}$. Since $\|\Sigma_1^{1/2}v\|_2/\|v\|_2$ is bounded away from zero and infinity, it follows that there exist constants $C_1, C_2 > 0$ such that

$$C_1 \leq \frac{\|\tilde{X}v\|_2/\sqrt{n}}{\|v\|_2} \leq C_2, \qquad \forall v \in \bigcup_{|J|\leq s} \mathcal{C}(J,k_0).$$

Since $\|v\|_2 \geq \|v_J\|_2$, we have that with probability approaching one,

$$C_1 \leq \frac{\|\tilde{X}v\|_2/\sqrt{n}}{\|v_J\|_2} \qquad \text{and} \qquad \frac{\|\tilde{X}v\|_2/\sqrt{n}}{\|v\|_2} \leq C_2 \qquad \forall v \in \bigcup_{|J|\leq s} \mathcal{C}(J,k_0).$$

This proves the first claim. The second claim follows by replacing $\tilde{X}$ with X.

Proof of the third claim. Notice that $W_i q(X_i'\theta_{(1)}) - 1 = v_{i,(1)} q(X_i'\theta_{(1)})$. Thus,

$$\mathbb{E}_{n,H_A} X_i (W_i q(X_i'\theta_{(1)}) - 1) = \mathbb{E}_{n,H_A} X_i q(X_i'\theta_{(1)}) v_{i,(1)}.$$

Notice that conditional on $\{X_i\}_{i\in H_A}$, $\{v_{i,(1)}\}_{i\in H_A}$ is independent across i with mean zero. Moreover, $v_{i,(1)}$ is also sub-Gaussian since it is bounded by 2. (To see this, simply notice that $v_{i,(1)} = W_i - e_{(1)}(X_i)$ and both W_i and $e_{(1)}(X_i)$ are bounded by 1.) By Hoeffding's inequality (e.g., Proposition 5.10 in Vershynin [2012a]), it follows that for $j \in \{1, ..., p\}$ and for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i\in H_A} X_{i,j} q(X_i'\theta_{(1)}) v_{i,(1)}\right| > t \mid \{X_i\}_{i\in H_A}\right) \leq \exp\left(1 - \frac{C_3 t^2}{\sum_{i\in H_A} X_{i,j}^2 [q(X_i'\theta_{(1)})]^2}\right),$$

where $C_3 > 0$ is a universal constant.

Since $\|X\|_\infty \leq M_1$, $\|\theta_{(1)}\|_1 \leq M_5$ and $q(X_i'\theta_{(1)}) = 1 + \exp(-X_i'\theta_{(1)})$, we have that

$$\sum_{i\in H_A} X_{i,j}^2 [q(X_i'\theta_{(1)})]^2 \leq b_n C_4,$$

where $C_4 = M_1^2(1+\exp(M_1 M_5))^2$. Hence, by the union bound, it follows that

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\sum_{i\in H_A} X_{i,j} q(X_i'\theta_{(1)}) v_{i,(1)}\right| > t\right) \leq p \exp\left(1 - \frac{C_3 t^2}{b_n C_4}\right).$$

Hence, by taking $\lambda_\theta = 2\sqrt{2 b_n C_3^{-1} C_4 \log p}$, we have

$$\mathbb{P}(\|\mathbb{E}_{n,H_A} X_i (W_i q(X_i'\theta_{(1)}) - 1)\|_\infty > 0.5\lambda_\theta) = \exp(1)/p \to 0.$$

This proves the third claim.

Proof of the fourth claim. The argument is essentially the same as the proof of the third claim. We outline the strategy. Notice that $\hat{\theta}_{(1),A}$ is computed using $\{(X_i, W_i)\}_{i\in H_A}$, which depends only on $\{(X_i, v_{i,(1)})\}_{i\in H_A}$. Since $\varepsilon_{i,(1)}$ and $v_{i,(1)}$ are independent conditional on X_i, it follows that condition on $\{(X_i, W_i)\}_{i\in H_A}$, $\{\varepsilon_{i,(1)}\}_{i\in H_A}$ is independent across i with mean zero. Therefore, exactly the same argument as above with $v_{i,(1)}$ replaced by $\varepsilon_{i,(1)}$ would yield the fourth claim. The proof is complete. □

Lemma 7. *For any $x \in \mathbb{R}$, $\exp(-x) - 1 + x \geq 0.4x^2 - 0.1x^3$.*

Proof of Lemma 7. Let $h(x) = \exp(-x) - 1 + x - 0.4x^2 + 0.1x^3$. Then $\ddot{h}(x) = d^2h(x)/dx^2 = \exp(-x) + 0.6x - 0.8$. We first show that $\ddot{h}(\cdot)$ is convex and then derive the minimum of $h(\cdot)$.

By taking the second derivative of $\ddot{h}(\cdot)$, we can see that $\ddot{h}(\cdot)$ is convex. To find the minimum of $\ddot{h}(\cdot)$, we consider the first order condition: $-\exp(-x) + 0.6 = 0$, i.e.,

$$\arg\min_{x\in\mathbb{R}} h(x) = \log(5/3).$$

This means that $\min_{x\in\mathbb{R}} \ddot{h}(x) = \ddot{h}(\log(5/3)) = 0.6 + 0.6 \times \log(5/3) - 0.8 > 0$. Therefore, $\ddot{h}(\cdot)$ is non-negative, which means that $h(\cdot)$ is convex.

Now we take the first order condition for $\min_{x\in\mathbb{R}} h(x)$, leading to

$$-\exp(-x) + 1 - 0.8x + 0.3x^2 = 0.$$

Clearly, $x = 0$ is a solution. Since $h(\cdot)$ is convex, this is the only solution. Therefore, $\min_{x\in\mathbb{R}} h(x) = h(0) = 0$. The proof is complete. □

A.2.2 Lasso-type estimator: $\check{\theta}$

For the next result, we introduce a simplified notation to help with the exposition.

Define

$$L_n(\theta) = \mathbb{E}_{n,H_A}[(1 - W_i)X_i'\theta + W_i \exp(-X_i'\theta)]$$

and $\dot{L}_n(\theta) = \mathbb{E}_{n,H_A}(1 - W_i q(X_i'\theta))X_i$. Define

$$\check{\theta}_{(1),A} = \arg\min_{\theta} \{L_n(\theta) + \lambda\|\theta\|_1\}.$$

Let $J \subset \{1, ..., p\}$ satisfy $\text{supp}(\theta_{(1)}) \subset J$.

Proposition 8. *We assume that $\|\dot{L}_n(\theta_{(1)})\|_\infty \leq c\lambda$, $\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta)(X_i\delta)^2 \geq \kappa_1\|\delta_J\|_2^2$, $\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta)|X_i'\delta|^3 \leq \kappa_2\|\delta_J\|_2^3$ and $\lambda^2 s \leq \kappa_1^4\kappa_2^{-2}/20$ for any $\delta \in \{v : \|v_J\|_1 \leq \|v_{J^c}\|_1(1 + c)/(1 - c)\}$, where $s = |J|$. Let $\Delta = \check{\theta}_{(1),A} - \theta_{(1)}$.*

Then,

$$\|\Delta_{J^c}\|_1 \leq (1 - c)^{-1}(1 + c)\|\Delta_J\|_1,$$

$$\|\Delta_J\|_2 \leq 10\kappa_1^{-1}\lambda\sqrt{s}, \qquad \text{and} \qquad \|\Delta\|_1 \leq 20(1 - c)^{-1}\kappa_1^{-1}\lambda s.$$

Proof of Proposition 8. Our proof uses the minoration argument from Belloni et al. [2011, 2016]. We notice that

$$L_n(\theta_{(1)} + \Delta) + \lambda\|\theta_{(1)} + \Delta\|_1 \leq L_n(\theta_{(1)}) + \lambda\|\theta_{(1)}\|_1.$$

Since $\|\theta_{(1)} + \Delta\|_1 = \|\theta_{(1)} + \Delta_J\|_1 + \|\Delta_{J^c}\|_1$, we have that

$$L_n(\theta_{(1)} + \Delta) - L_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \le \lambda\|\theta_{(1)}\|_1 - \lambda\|\theta_{(1)} + \Delta_J\|_1 \le \lambda\|\Delta_J\|_1. \tag{19}$$

By the convexity of $L_n(\cdot)$, we have that $L_n(\theta_{(1)} + \Delta) - L_n(\theta_{(1)}) - \Delta'\dot{L}_n(\theta_{(1)}) \ge 0$. Moreover,

$$|\Delta\dot{L}_n(\theta_{(1)})| \le \|\Delta\|_1\|\dot{L}_n(\theta_{(1)})\|_\infty \le c\lambda\|\Delta\|_1 \le c\lambda\|\Delta_J\|_1 + c\lambda\|\Delta_{J^c}\|_1.$$

The above two displays imply that

$$\begin{aligned}\lambda\|\Delta_J\|_1 \ge L_n(\theta_{(1)} + \Delta) - L_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 &\ge \Delta'\dot{L}_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \\ &\ge -(c\lambda\|\Delta_J\|_1 + c\lambda\|\Delta_{J^c}\|_1) + \lambda\|\Delta_{J^c}\|_1.\end{aligned}$$

Rearranging the terms, we obtain

$$\|\Delta_{J^c}\|_1 \le \frac{1+c}{1-c}\|\Delta_J\|_1. \tag{20}$$

We denote $A = \{v : \ \|v_{J^c}\|_1 \le (1-c)^{-1}(1+c)\|v_J\|_1\}$. We define the following quantity

$$r_A = \sup\left\{r > 0 : \inf_{\|\delta\|\le r, \ \delta\in A} \frac{L_n(\theta_{(1)} + \delta) - L_n(\theta_{(1)}) - \delta'\dot{L}_n(\theta_{(1)})}{\|\delta\|^2} \ge c_1\right\},$$

where $c_1 = 0.1\kappa_1$.

Step 1: show $r_A \ge 3\kappa_1\kappa_2^{-1}$.

Define $f : \mathbb{R} \mapsto \mathbb{R}$ by $f(x) = \exp(-x) - 1 + x$. Then we can rewrite

$$L_n(\theta_{(1)} + \delta) - L_n(\theta_{(1)}) - \delta'\dot{L}_n(\theta_{(1)}) = \mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)}) f(X'\delta).$$

By Lemma 7, we have that

$$\begin{aligned}&L_n(\theta_{(1)} + \delta) - L_n(\theta_{(1)}) - \delta'\dot{L}_n(\theta_{(1)}) \\ &\qquad \ge 0.4\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})(X_i\delta)^2 - 0.1\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})|X_i'\delta|^3 \\ &\qquad \ge 0.4\kappa_1\|\delta_J\|_2 - 0.1\kappa_2\|\delta_J\|_2^3.\end{aligned}$$

Thus, for any $\delta \in A$,

$$\frac{L_n(\theta_{(1)} + \delta) - L_n(\theta_{(1)}) - \delta'\dot{L}_n(\theta_{(1)})}{\|\delta_J\|_2^2} \ge 0.4\kappa_1 - 0.1\kappa_2\|\delta_J\|_2.$$

Therefore,

$$\begin{aligned}&\inf_{\delta\in A, \ \|\delta_J\|_2\le 3\kappa_1\kappa_2^{-1}} \frac{L_n(\theta_{(1)} + \delta) - L_n(\theta_{(1)}) - \delta'\dot{L}_n(\theta_{(1)})}{\|\delta_J\|_2^2} \\ &\qquad \ge \inf_{\delta\in A, \ \|\delta_J\|_2\le 3\kappa_1\kappa_2^{-1}} (0.4\kappa_1 - 0.1\kappa_2\|\delta_J\|_2) \ge 0.1\kappa_1 = c_1.\end{aligned}$$

Hence, we have that

$$r_A \ge 3\kappa_1\kappa_2^{-1}. \tag{21}$$

Step 2: show that $\|\Delta_J\|_2 \leq r_A$.
We proceed by contradiction. Assume that $\|\Delta_J\|_2 > r_A$. For $v \in \mathbb{R}^p$, define

$$Q(v) = L_n(\theta_{(1)} + v) - L_n(\theta_{(1)}) - v'\dot{L}_n(\theta_{(1)}).$$

By the convexity of $L_n(\cdot)$, $Q(\cdot)$ is also convex. Let $t = r_A/\|\Delta_J\|_2$. Since $\|\Delta_J\|_2 > r_A$, $t \in (0,1)$. Thus,

$$tQ(\Delta) + (1-t)Q(0) \geq Q(t\Delta + (1-t)\cdot 0).$$

Since $Q(0) = 0$, the above display implies that

$$Q(\Delta) \geq \frac{1}{t}Q(t\Delta) = r_A^{-1}\|\Delta_J\|_2 Q(t\Delta) \overset{\text{(i)}}{\geq} c_1 r_A \|\Delta_J\|_2. \tag{22}$$

where (i) follows by the definition of r_A and fact that $\|(t\Delta)_J\|_2 = r_A$ and $t\Delta \in A$. By (19), we have that

$$\begin{aligned}
\lambda\|\Delta_J\|_1 &\geq Q(\Delta) + \Delta'\dot{L}_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \\
&\overset{\text{(i)}}{\geq} c_1 r_A \|\Delta_J\|_2 + \Delta'\dot{L}_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \\
&\overset{\text{(ii)}}{\geq} c_1 r_A \|\Delta_J\|_2 - c\lambda\|\Delta\|_1 + \lambda\|\Delta_{J^c}\|_1,
\end{aligned}$$

where (i) follows by (22) and (ii) follows by $|\Delta\dot{L}_n(\theta_{(1)})| \leq \|\Delta\|_1\|\dot{L}_n(\theta_{(1)})\|_\infty \leq c\lambda\|\Delta\|_1$. Rearranging the terms, we obtain

$$c_1 r_A \|\Delta_J\|_2 + (1-c)\lambda\|\Delta_{J^c}\|_1 \leq \lambda\|\Delta_J\|_1.$$

Therefore,

$$r_A \leq \frac{\lambda\|\Delta_J\|_1}{c_1\|\Delta\|_2} \overset{\text{(i)}}{\leq} \frac{\lambda\sqrt{s}\|\Delta_J\|_2}{c_1\|\Delta\|_2} \leq c_1^{-1}\lambda\sqrt{s},$$

where (i) follows by Hölder's inequality.

By (21), $c_1 = 0.1\kappa_1$ and $r_A \geq 3\kappa_1\kappa_2^{-1}$. Hence, we have $3\kappa_1\kappa_2^{-1} \leq 10\kappa_1^{-1}\lambda\sqrt{s}$, which means $\lambda^2 s \geq 0.09\kappa_1^4\kappa_2^{-2}$. This contradicts the assumption of $\lambda^2 s \leq \kappa_1^4\kappa_2^{-2}/20$. Hence,

$$\|\Delta_J\|_2 \leq r_A. \tag{23}$$

Step 3: derive the desired result.
By (19), we have that

$$\begin{aligned}
\lambda\|\Delta_J\|_1 &\geq L_n(\theta_{(1)} + \Delta) - L_n(\theta_{(1)}) - \Delta'\dot{L}_n(\theta_{(1)}) + \Delta'\dot{L}_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \\
&\overset{\text{(i)}}{\geq} c_1\|\Delta_J\|_2^2 + \Delta'\dot{L}_n(\theta_{(1)}) + \lambda\|\Delta_{J^c}\|_1 \\
&\overset{\text{(ii)}}{\geq} c_1\|\Delta_J\|_2^2 - c\lambda\|\Delta\|_1 + \lambda\|\Delta_{J^c}\|_1,
\end{aligned}$$

where (i) follows by $\|\Delta_J\|_2 \leq r_A$ (due to (23)) and (ii) follows by $|\Delta\dot{L}_n(\theta_{(1)})| \leq \|\Delta\|_1\|\dot{L}_n(\theta_{(1)})\|_\infty \leq c\lambda\|\Delta\|_1$. Rearranging the terms, we obtain $c_1\|\Delta_J\|_2^2 + (1-c)\lambda\|\Delta_{J^c}\|_1 \leq \lambda\|\Delta_J\|_1$. Hence,

$$\|\Delta_J\|_2 \leq \frac{\lambda\|\Delta_J\|_1}{c_1\|\Delta_J\|_2} \overset{\text{(i)}}{\leq} \frac{\lambda\|\Delta_J\|_2\sqrt{s}}{c_1\|\Delta_J\|_2} = c_1^{-1}\lambda\sqrt{s} \overset{\text{(ii)}}{=} 10\kappa_1^{-1}\lambda\sqrt{s},$$

where (i) follows by Hölder's inequality and (ii) follows by $c_1 = 0.1\kappa_1$. Thus,

$$\|\Delta\|_1 = \|\Delta_J\|_1 + \|\Delta_{J^c}\|_1 \overset{\text{(i)}}{\leq} \left(1 + \frac{1+c}{1-c}\right)\|\Delta_J\|_1 \overset{\text{(ii)}}{\leq} \frac{2}{1-c}\|\Delta_J\|_2\sqrt{s} = \frac{20}{1-c}\kappa_1^{-1}\lambda s,$$

where (i) follows by (20) and (ii) follows by Hölder's inequality. □

With the help of Proposition 8 we are now able to establish estimation quality properties of the introduced estimator of $\theta_{(1)}$.

Lemma 9. *Let Assumptions 1 and 2 hold. Assume that* $\|\theta_{(1)}\|_0 = o(n/\log p)$. *Then*

$$\mathbb{P}\left(\check{\theta}_{(1),A} - \theta_{(1)} \in \mathcal{C}(\text{supp}(\theta_{(1)}), 3)\right) \to 1,$$

$$\|\check{\theta}_{(1),A} - \theta_{(1)}\|_1 = O_P\left(s_\theta\sqrt{n^{-1}\log p}\right)$$

and

$$\|\check{\theta}_{(1),A} - \theta_{(1)}\|_2 = O_P\left(\sqrt{s_\theta n^{-1}\log p}\right).$$

Proof of Lemma 9. We define the event

$$\mathcal{M} = \left\{\|\tilde{\theta}_{(1),A}\|_1 \leq M_5\right\} \bigcap \left\{\|\mathbb{E}_{n,H_A} X_i(1 - W_i q(X_i'\theta_{(1)}))\|_\infty \leq 0.5\lambda_\theta\right\}$$
$$\bigcap \left\{\min_{|J|\leq 2s_\theta} \inf_{\|v_{J^c}\|_1 \leq \|v_J\|_1} \frac{\mathbb{E}_{n,H_A} W_i (X_i'v)^2}{\|v_J\|_2^2} \geq c_1^2\right\}$$
$$\bigcap \left\{\max_{|J|\leq 2s_\theta} \max_{\|v_{J^c}\|_1 \leq \|v_J\|_1} \frac{\mathbb{E}_{n,H_A} (X_i'v)^2}{\|v\|_2^2} \leq c_2^2\right\}$$

where $c_1, c_2 > 0$ are constants from Lemma 6. By Lemma 6, $\mathbb{P}(\mathcal{M}) \to 1$.

Let $B = \text{supp}(\theta_{(1)})$. We apply Proposition 8 with $J = B$, $c = 0.5$ and $\lambda = \lambda_\theta$, obtaining that on the event $\mathcal{M}$, $\|\Delta_{B^c}\|_1 \leq 3\|\Delta_B\|_1$ and

$$\|\Delta\|_1 \leq 40\kappa_1^{-1}\lambda_\theta s_\theta. \tag{24}$$

Let N_0 denote the s_θ indices in B^c corresponding to the largest s_θ entries (in absolute value) of Δ. Let $N = B \bigcup N_0$. We now apply Lemma 6.9 of Bühlmann and Van De Geer [2011] to the vector Δ_{B^c}. Once we exclude the largest s_θ entries (in magnitude) in Δ_{B^c}, we obtain Δ_{N^c}. Hence, Lemma 6.9 of Bühlmann and Van De Geer [2011] implies that

$$\|\Delta_{N^c}\|_2 \leq s_\theta^{-1/2}\|\Delta_{B^c}\|_1 \leq s_\theta^{-1/2}\|\Delta\|_1 \overset{\text{(i)}}{\leq} 40\kappa_1^{-1}\lambda_\theta\sqrt{s_\theta},$$

where (i) follows by (24).

Since $|N| \leq 2s_\theta$ (due to the definition of N), $\|\Delta_{N^c}\|_1 \leq \|\Delta_N\|_1$ on the event $\mathcal{M}$. We now apply Proposition 8 with $J = N$, $c = 0.5$ and $\lambda = \lambda_\theta$, obtaining that on the event $\mathcal{M}$,

$$\|\Delta_N\|_2 \leq 10\kappa_1^{-1}\lambda\sqrt{|N|} \overset{\text{(i)}}{\leq} 10\kappa_1^{-1}\lambda_\theta\sqrt{2s_\theta},$$

where (i) follows by $|N| \leq 2s_\theta$. Hence, the above two displays imply

$$\|\Delta\|_2^2 = \|\Delta_N\|_2^2 + \|\Delta_{N^c}\|_2^2 \leq 400(1-c)^{-2}\kappa_1^{-2}\lambda^2 s_\theta + 200\kappa_1^{-2}\lambda^2 s_\theta.$$

Hence, $\|\Delta\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p})$. □

A.2.3 Dantzig-type estimator: $\hat{\theta}$

Lemma 10. *For any $z > 0$, there exists a constant $C_z > 0$ depending only on z such that for any $x \in [-z, z]$, $(1 - \exp(-x))x \geq C_z x^2$.*

Proof of Lemma 10. Let $f(x) = [(1 - \exp(-x))x]/x^2$. Then

$$df(x)/dx = x^{-1}\exp(-x)(1 + x - \exp(x)).$$

By the elementary inequality of $\exp(x) \geq 1 + x$ for any $x \in \mathbb{R}$, we have that $df(x)/dx \leq 0$ for any $x \in \mathbb{R}$. Thus, $f(x)$ is non-increasing on $\mathbb{R}$. Hence, $\inf_{x\in[-z,z]} f(x) = f(z)$. Let $C_z = f(z)$. Then $f(x) \geq C_z$, which implies $(1 - \exp(-x))x \geq C_z x^2$. The proof is complete. ☐

Proposition 11. *Suppose that $\|\mathbb{E}_{n,H_A} X_i(1 - W_i q(X_i'\theta_{(1)}))\|_\infty \leq \lambda$ and*

$$\inf_{\delta:\ \|\delta_{J^c}\|_1 \leq \|\delta_J\|_1} \mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})(X_i\delta)^2/\|\delta_J\|_2^2 \geq \kappa_1,$$

where $J \subset \{1, ..., p\}$ satisfies $\operatorname{supp}(\theta_{(1)}) \subseteq J$. Let $\Delta = \tilde{\theta}_{(1),A} - \theta_{(1)}$.

Then,

$$\|\Delta_{J^c}\|_1 \leq \|\Delta_J\|_1, \qquad \|\Delta\|_1 \leq 8D_1^{-1}\kappa_1^{-1}\lambda s, \qquad \textit{and}, \qquad \|\Delta_J\|_2 \leq 4D_1^{-1}\kappa_1^{-1}\lambda\sqrt{s},$$

where $s = |J|$ and $D_1 > 0$ is a constant depending only on M_1 and M_5.

Proof of Proposition 11. Since $\|\tilde{\theta}_{(1),A}\|_1 \leq \|\theta_{(1)}\|_1$ and $\|\tilde{\theta}_{(1),A}\|_1 = \|\theta_{(1)} + \Delta_J\|_1 + \|\Delta_{J^c}\|_1$ (due to $\operatorname{supp}(\theta_{(1)}) \subseteq J$), we have that $\|\Delta_{J^c}\|_1 \leq \|\Delta_J\|_1$. Hence, $\|\Delta\|_1 = \|\Delta_J\|_1 + \|\Delta_{J^c}\|_1 \leq 2\|\Delta_J\|_1$.

By construction, we have $\|\mathbb{E}_{n,H_A} X_i(1 - W_i q(X_i'(\theta_{(1)} + \Delta)))\|_\infty \leq \lambda$. Therefore,

$$\|\mathbb{E}_{n,H_A} X_i W_i(q(X_i'(\theta_{(1)} + \Delta)) - q(X_i'\theta_{(1)}))\|_\infty \leq 2\lambda.$$

This means

$$\left\|\mathbb{E}_{n,H_A} X_i W_i \exp(-X_i'\theta_{(1)})[1 - \exp(-X_i'\Delta)]\right\|_\infty \leq 2\lambda.$$

Let $\phi(x) = (1 - \exp(-x))x$. Therefore,

$$\begin{aligned}&\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})\phi(X_i'\Delta)\\ &\quad \leq \|\Delta\|_1 \left\|\mathbb{E}_{n,H_A} X_i W_i \exp(-X_i'\theta_{(1)})[1 - \exp(-X_i'\Delta)]\right\|_\infty \leq 2\lambda\|\Delta\|_1.\end{aligned}$$

Notice that by construction $\|\tilde{\theta}_{(1),A}\|_1 \leq \|\theta_{(1)}\|_1$. Since $\|\theta_{(1)}\|_1 \leq M_5$ and $\|X\|_\infty \leq M_1$ by assumption, $\|X\Delta\|_\infty \leq 2M_1M_4$ is also bounded. By Lemma 10, there exists a constant $D_1 > 0$ depending only on M_1M_5 such that $\phi(X_i'\Delta) \geq D_1(X_i'\Delta)^2$. It follows that

$$2\lambda\|\Delta\|_1 \geq D_1\mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})(X_i'\Delta)^2 \overset{\text{(i)}}{\geq} D_1\kappa_1\|\Delta_J\|_2^2,$$

where (i) follows by $\|\Delta_{J^c}\|_1 \leq \|\Delta_J\|_1$ and the assumption of $\inf_{\delta:\ \|\delta_{J^c}\|_1 \leq \|\delta_J\|_1} \mathbb{E}_{n,H_A} W_i \exp(-X_i'\theta_{(1)})(X_i\delta)^2/\|\delta_J\|_2^2 \geq \kappa_1$.

Since $\|\Delta\|_1 \leq 2\|\Delta_J\|_1 \leq 2\sqrt{s}\|\Delta_J\|_2$, we have that

$$4\lambda\sqrt{s}\|\Delta\|_2 \geq D_1\kappa_1\|\Delta_J\|_2^2,$$

which implies $\|\Delta_J\|_2 \leq 4D_1^{-1}\kappa_1^{-1}\lambda\sqrt{s}$ and thus $\|\Delta\|_1 \leq 2\sqrt{s}\|\Delta_J\|_2 \leq 8D_1^{-1}\kappa_1^{-1}\lambda s$. The proof is complete. ☐

With the help of Proposition 11 we are now able to complete the proof regarding the dantzig-type estimator as defined in Algorithm 1.

Lemma 12. *Let Assumptions 1 and 2 hold. Assume that* $\|\theta_{(1)}\|_0 = o(n/\log p)$.

Then, $\|\tilde{\theta}_{(1),A}\|_1 \le M_5$ *and* $\tilde{\theta}_{(1),A} - \theta_{(1)} \in \mathcal{C}(\mathrm{supp}(\theta_{(1)}), 1)$ *with probability approaching one.*

Moreover,

$$\|\tilde{\theta}_{(1),A} - \theta_{(1)}\|_1 = O_P(s_\theta \sqrt{n^{-1}\log p}),$$
$$\|\tilde{\theta}_{(1),A} - \theta_{(1)}\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p})$$

and

$$\sum_{i \in H_A} W_i \left(q(X_i'\tilde{\theta}_{(1),A}) - q(X_i'\theta_{(1)}) \right)^2 = o_P(n),$$

where $B = \mathrm{supp}(\theta_{(1)})$.

Proof of Lemma 12. We now combine Lemma 6 and Proposition 11 to obtain the desired result. Define the event

$$\mathcal{M} = \left\{ \|\tilde{\theta}_{(1),A}\|_1 \le M_5 \right\} \bigcap \left\{ \|\mathbb{E}_{n,H_A} X_i(1 - W_i q(X_i'\theta_{(1)}))\|_\infty \le \lambda_\theta \right\}$$
$$\bigcap \left\{ \min_{|J| \le 2s_\theta} \inf_{\|v_{J^c}\|_1 \le \|v_J\|_1} \frac{\mathbb{E}_{n,H_A} W_i (X_i'v)^2}{\|v_J\|_2^2} \ge c_1^2 \right\}$$
$$\bigcap \left\{ \max_{|J| \le 2s_\theta} \max_{\|v_{J^c}\|_1 \le \|v_J\|_1} \frac{\mathbb{E}_{n,H_A} (X_i'v)^2}{\|v\|_2^2} \le c_2^2 \right\}$$

where $c_1, c_2 > 0$ are constants from Lemma 6. By Lemma 6, $\mathbb{P}(\mathcal{M}) \to 1$.

Let $\Delta = \tilde{\theta}_{(1),A} - \theta_{(1)}$ and $B = \mathrm{supp}(\theta_{(1)})$. Notice that on the event $\mathcal{M}$, $\|\tilde{\theta}_{(1),A}\|_1 \le \|\theta_{(1)}\|_1 \le M_5$ and $\Delta \in \mathcal{C}(B, 1)$ (due to Proposition 11). Since $\mathbb{P}(\mathcal{M}) \to 1$, we have proved the first two claims. Now we prove the other claims in three steps.

Step 1: show $\|\tilde{\theta}_{(1),A} - \theta_{(1)}\|_1 = O_P(s_\theta \sqrt{n^{-1}\log p})$.

Since $\|X\theta_{(1)}\|_\infty \le \|X\|_\infty \|\theta_{(1)}\|_1 \le M_1 M_4$, we have that on the event $\mathcal{M}$,

$$\inf_{\|v_{J^c}\|_1 \le \|v_J\|_1} \frac{\mathbb{E}_{n,H_A} W_i (X_i'v)^2 \exp(-X_i'\theta_{(1)})}{\|v_J\|_2^2} \ge c_1^2 \exp(-M_1 M_4).$$

We apply Proposition 11 with J and obtain that on the event $\mathcal{M}$,

$$\|\Delta\|_1 \le 8 D_1^{-1} c_1^{-2} \exp(M_1 M_4) \lambda_\theta s_\theta, \tag{25}$$

where $D_1 > 0$ is a constant depending only on M_1 and M_4. Since $\lambda_\theta \asymp \sqrt{n^{-1}\log p}$ and $\mathbb{P}(\mathcal{M}) \to 1$, we obtain $\|\hat{\theta}_{(1),A} - \theta_{(1)}\|_1 = O_P(s_\theta \sqrt{n^{-1}\log p})$.

Step 2: show that $\|\tilde{\theta}_{(1),A} - \theta_{(1)}\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p})$.

Let N_0 denote the s_θ indices in B^c corresponding to the largest s_θ entries (in absolute value) of Δ. Let $N = B \bigcup N_0$. We now apply Lemma 6.9 of Bühlmann and Van De Geer [2011] to the vector Δ_{B^c}. Once we exclude the largest s_θ entries (in magnitude) in Δ_{B^c}, we obtain Δ_{N^c}. Hence, Lemma 6.9 of Bühlmann and Van De Geer [2011] implies that

$$\|\Delta_{N^c}\|_2 \le s_\theta^{-1/2} \|\Delta_{B^c}\|_1 \le s_\theta^{-1/2} \|\Delta\|_1 \le 8 D_1^{-1} c_1^{-2} \exp(M_1 M_4) \lambda_\theta \sqrt{s_\theta},$$

where (i) follows by (25).

We now apply Proposition 11 with $J = N$ and obtain that on the event $\mathcal{M}$,

$$\|\Delta_N\|_2 \leq 4D_1^{-1}c_1^{-2}\exp(M_1M_4)\lambda_\theta\sqrt{2s_\theta}.$$

Hence, the above two displays imply

$$\|\Delta\|_2^2 = \|\Delta_N\|_2^2 + \|\Delta_{N^c}\|_2^2 \leq 6\left(4D_1^{-1}c_1^{-2}\exp(M_1M_4)\right)^2\lambda_{\theta,}^2 s_\theta.$$

Hence, $\|\Delta\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p})$. The proof is complete. ☐

A.2.4 Combining Lasso-type and Dantzig-type estimators

Lemma 13. *Let Assumptions 1 and 2 hold. Assume that $\|\theta_{(1)}\|_0 = o(n/\log p)$.*

Then, $\|\hat\theta_{(1),A}\|_\infty \leq M_5 \vee \kappa_0$ and $\hat\theta_{(1),A} - \theta_{(1)} \in \mathcal{C}(\text{supp}(\theta_{(1)}, 3)$ with probability approaching one.

Moreover,

$$\|\hat\theta_{(1),A} - \theta_{(1)}\|_1 = O_P(s_\theta\sqrt{n^{-1}\log p}),$$
$$\|\hat\theta_{(1),A} - \theta_{(1)}\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p}),$$
$$\sum_{i\in H_A} W_i\left(q(X_i'\hat\theta_{(1),A}) - q(X_i'\theta_{(1)})\right)^2 = o_P(n)$$

and

$$\mathbb{E}_{n,H_A}(X_i'(\hat\theta_{(1),A} - \theta_{(1)}))^2 = O_P(s_\theta n^{-1}\log p).$$

Analogous results hold if we replace $\mathbb{E}_{n,H_A}$ and $\hat\theta_{(1),A}$ with $\mathbb{E}_{n,H_B}$ and $\hat\theta_{(1),B}$.

Proof of Lemma 13. By construction in Algorithm 1, $\|\hat\theta_{(1),A}\|_1 \leq \max\{\|\tilde\theta_{(1),A}\|_1, \kappa_0\}$. By Lemma 12, $\mathbb{P}(\|\tilde\theta_{(1),A}\|_1 \leq M_5) \to 1$. Thus, $\mathbb{P}(\|\hat\theta_{(1),A}\|_1 \leq M_5 \vee \kappa_0) \to 1$. By Hölder's inequality, $\|X\hat\theta_{(1),A}\|_\infty \leq \|X\|_\infty\|\hat\theta_{(1),A}\|_1 \leq M_1(M_5 \vee \kappa_0)$ with probability approaching one. Since the bounds for $\|\check\theta_{(1),A} - \theta_{(1)}\|_1$ and $\|\tilde\theta_{(1),A} - \theta_{(1)}\|_1$ are both $O_P(s_\theta\sqrt{n^{-1}\log p})$, we have $\|\hat\theta_{(1),A} - \theta_{(1)}\|_1 = O_P(s_\theta\sqrt{n^{-1}\log p})$. Similarly, $\|\hat\theta_{(1),A} - \theta_{(1)}\|_2 = O_P(\sqrt{s_\theta n^{-1}\log p})$.

Now we show the last bound. Let $\Delta = \hat\theta_{(1),A} - \theta_{(1)}$. By Taylor's theorem, there exists $\tau_i \in [0,1]$ such that

$$\left|q(X_i'\hat\theta_{(1),A}) - q(X_i'\theta_{(1)})\right| = \exp(-X_i'\theta_{(1)})\left|\exp(-X_i'\Delta) - 1\right| = \exp(-X_i'\theta_{(1)})\exp(-X_i'\Delta\tau_i)\left|X_i'\Delta\right|.$$

Notice that

$$\|X\Delta\|_\infty \leq \|X\|_\infty\|\Delta\|_1 \leq M_1(\|\theta_{(1)}\|_1 + \|\hat\theta_{(1),A}\|_1) \leq M_1(M_5 + \|\hat\theta_{(1),A}\|_1).$$

We have shown that with probability approaching one, $\|\hat\theta_{(1),A}\|_1 \leq M_5 \vee \kappa_0$. Thus, $\mathbb{P}(\mathcal{A}) \to 1$, where the event is defined as $\mathcal{A} = \{\|X\Delta\|_\infty \leq M_1(M_5 + M_5 \vee \kappa_0)\}$. Moreover, we have $\|X\theta_{(1)}\|_\infty \leq \|X\|_\infty\|\theta_{(1)}\|_1 \leq M_1M_5$. It follows that with probability one, for any $i \in H_A$,

$$\left|q(X_i'\tilde\theta_{(1),A}) - q(X_i'\theta_{(1)})\right| \leq \exp\left(M_1M_5 + M_1(M_5 + M_5 \vee \kappa_0)\right)|X_i'\Delta|.$$

Hence,

$$\sum_{i\in H_A} W_i \left(q(X_i'\tilde{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right)^2 \leq \exp\left(2M_1M_5 + 2M_1(M_5 + M_5 \vee \kappa_0)\right) \sum_{i\in H_A} (X_i'\Delta)^2.$$

By Lemmas 9 and 12, $\check{\theta}_{(1),A} - \theta_{(1)} \in \mathcal{C}(\text{supp}(\theta_{(1)}), 3)$ and $\tilde{\theta}_{(1),A} - \theta_{(1)} \in \mathcal{C}(\text{supp}(\theta_{(1)}, 1)$ with probability approaching one. Since $\Delta \in \{\check{\theta}_{(1),A} - \theta_{(1)},\ \tilde{\theta}_{(1),A} - \theta_{(1)}\}$ and $\mathcal{C}(\text{supp}(\theta_{(1)}, 1) \subset \mathcal{C}(\text{supp}(\theta_{(1)}, 3)$, we have that $\mathbb{P}(\Delta \in \mathcal{C}(\text{supp}(\theta_{(1)}), 3)) \to 1$.

It follows that on the event $\mathcal{M}$,

$$\mathbb{E}_{n,H_A}(X_i'\Delta)^2 \leq c_2^2\|\Delta\|_2^2 = O_P(s_\theta n^{-1}\log p).$$

By the above two displays, together with $s_\theta = o(n/\log p)$, we have

$$\sum_{i\in H_A} W_i \left(q(X_i'\tilde{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right)^2 = o_P(n).$$

□

A.2.5 Estimators for the outcome model's $\beta_{(1)}$

Lemma 14. *Let Assumptions 1 and 2 hold. Also assume that $\|\beta_{(1)}\|_0 = o(n/\log p)$. Then $\beta_{(1)}$ satisfies (17) and $\hat{\beta}_{(1),A} - \beta_{(1)} \in \mathcal{C}\left(\text{supp}(\beta_{(1)}), 3\right)$ with probability approaching one.*

Moreover,

$$\|\hat{\beta}_{(1),A} - \beta_{(1)}\|_1 = O_P(\|\beta_{(1)}\|_0\sqrt{n^{-1}\log p})$$

and

$$\mathbb{E}_{n,H_B}\left[X_i'(\hat{\beta}_{(1),A} - \beta_{(1)})\right]^2 = o_P(1).$$

Analogous results hold if we replace $\mathbb{E}_{n,H_A}$ and $\hat{\beta}_{(1),A}$ with $\mathbb{E}_{n,H_B}$ and $\hat{\beta}_{(1),B}$.

Proof of Lemma 14. We use the standard argument for Lasso. Let $s_\beta = \|\beta_{(1)}\|_0$ and $Q = \text{supp}(\beta_{(1)})$. Define the event

$$\mathcal{M} = \left\{\|\mathbb{E}_{n,H_A}X_iW_i\exp(X_i'\hat{\theta}_{(1),A})\varepsilon_{i,(1)}\|_\infty \leq \lambda_\beta/4\right\} \bigcap \left\{\min_{|J|\leq 2s_\beta}\min_{v\in\mathcal{C}(J,3)} \frac{\mathbb{E}_{n,H_A}W_i(X_i'v)^2}{\|v_J\|_2^2} \geq c_1^2\right\} \bigcap \left\{\max_{|J|\leq 2s_\beta}\max_{v\in\mathcal{C}(J,3)} \frac{\mathbb{E}_{n,H_B}(X_i'v)^2}{\|v\|_2^2} \leq c_2^2\right\},$$

where $c_1, c_2 > 0$ are constants from Lemma 6. By Lemma 6, $\mathbb{P}(\mathcal{M}) \to 1$.

Define $\Delta = \hat{\beta}_{(1)} - \beta_{(1)}$ and $\tilde{W}_i = W_i\exp(-X_i'\hat{\theta}_{(1),A})$. Since (17) is the KKT condition for the optimization program that defineds $\hat{\beta}_{(1)}$, $\hat{\beta}_{(1)}$ satisfies (17). Now we show the other claims in three steps.

Step 1: $\mathbb{P}(\Delta \in \mathcal{C}(Q, 3)) \to 1$.

Since $W_iY_i = W_iY_i(1)$, we have that $\tilde{W}_i(Y_i - X_i'\beta)^2 = \tilde{W}_i(Y_i(1) - X_i'\beta)^2$. Recall that $Y_i(1) - X_i'\beta_{(1)} = \varepsilon_{i,(1)}$. Thus, by construction, we have

$$\mathbb{E}_{n,H_A}\tilde{W}_i(\varepsilon_{i,(1)} - X_i'\Delta)^2 + \lambda_\beta\|\beta_{(1)} + \Delta\|_1 \leq \mathbb{E}_{n,H_A}\tilde{W}_i\varepsilon_{i,(1)}^2 + \lambda_\beta\|\beta_{(1)}\|_1.$$

Rearranging terms, we obtain

$$\begin{aligned}\mathbb{E}_{n,H_A}\tilde{W}_i(X_i'\Delta)^2 &\leq 2\mathbb{E}_{n,H_A}\varepsilon_{i,(1)}\tilde{W}_iX_i'\Delta + \lambda_\beta\left(\|\beta_{(1)}\|_1 - \|\beta_{(1)}+\Delta\|_1\right)\\ &= 2\mathbb{E}_{n,H_A}\varepsilon_{i,(1)}\tilde{W}_iX_i'\Delta + \lambda_\beta\left(\|\beta_{(1)}\|_1 - \|\beta_{(1)}+\Delta_Q\|_1 - \|\Delta_{Q^c}\|_1\right)\\ &\leq 2\mathbb{E}_{n,H_A}\varepsilon_{i,(1)}\tilde{W}_iX_i'\Delta + \lambda_\beta\left(\|\Delta_Q\|_1 - \|\Delta_{Q^c}\|_1\right)\\ &\leq 2\|\mathbb{E}_{n,H_A}\varepsilon_{i,(1)}\tilde{W}_iX_i\|_\infty\|\Delta\|_1 + \lambda_\beta\left(\|\Delta_Q\|_1 - \|\Delta_{Q^c}\|_1\right).\end{aligned}$$

Therefore, on the event $\mathcal{M}$, we have that

$$\mathbb{E}_{n,H_A}\tilde{W}_i(X_i'\Delta)^2 \leq \lambda_\beta\|\Delta\|_1/2 + \lambda_\beta\left(\|\Delta_Q\|_1 - \|\Delta_{Q^c}\|_1\right) = \frac{3}{2}\lambda_\beta\|\Delta_Q\|_1 - \frac{1}{2}\lambda_\beta\|\Delta_{Q^c}\|_1. \quad (26)$$

Hence, on the event $\mathcal{M}$, $\frac{3}{2}\lambda_\beta\|\Delta_Q\|_1 \geq \frac{1}{2}\lambda_\beta\|\Delta_{Q^c}\|_1$, which means $\Delta \in \mathcal{C}(Q,3)$. Since $\mathbb{P}(\mathcal{M}) \to 1$, we have proved $\mathbb{P}(\Delta \in \mathcal{C}(Q,3)) \to 1$.

Step 2: show $\|\Delta\|_1 = O_P(s_\beta\sqrt{n^{-1}\log p})$.

By definition of $\hat{\theta}_{(1),A}$, $\|\hat{\theta}_{(1),A}\|_1 \leq \|\theta_{(1)}\|_1 \leq M_5$. By assumption, $\|X\|_\infty \leq M_1$. Thus, $\min_{i\in H_A}\exp(-X_i'\hat{\theta}_{(1),A}) \geq \exp(-M_1M_5)$. Since on the event $\mathcal{M}$, $\Delta \in \mathcal{C}(Q,3)$, it follows that on this event,

$$\mathbb{E}_{n,H_A}\tilde{W}_i(X_i'\Delta)^2 \geq c_1^2\exp(-M_1M_5)\|\Delta_J\|_2^2.$$

Hence, (26) implies that

$$c_1^2\exp(-M_1M_5)\|\Delta_J\|_2^2 \leq \frac{3}{2}\lambda_\beta\|\Delta_Q\|_1 - \frac{1}{2}\lambda_\beta\|\Delta_{Q^c}\|_1 \leq \frac{3}{2}\lambda_\beta\|\Delta_Q\|_1 \leq \frac{3}{2}\lambda_\beta\|\Delta_Q\|_2\sqrt{s_\beta},$$

which means

$$\|\Delta_J\|_2 \leq \frac{3}{2}c_1^{-2}\exp(M_1M_5)\lambda_\beta\sqrt{s_\beta}. \quad (27)$$

Therefore, on the event $\mathcal{M}$,

$$\|\Delta\|_1 = \|\Delta_Q\|_1 + \|\Delta_{Q^c}\|_1 \leq 4\|\Delta_Q\|_1 \leq 4\sqrt{s_\beta}\|\Delta_Q\|_2 \leq 6c_1^{-2}\exp(M_1M_5)\lambda_\beta s_\beta.$$

Since $\mathbb{P}(\mathcal{M}) \to 1$ and $\lambda_\beta \asymp \sqrt{n^{-1}\log p}$, we have $\|\Delta\|_1 = O_P(s_\beta\sqrt{n^{-1}\log p})$.

Step 3: show that $\mathbb{E}_{n,H_B}\left[X_i'(\hat{\beta}_{(1),A} - \beta_{(1)})\right]^2 = o_P(1)$.

Recall that $\Delta \in \mathcal{C}(Q,3)$ on the event $\mathcal{M}$. By the definition of this event, it follows that on this event,

$$\mathbb{E}_{n,H_B}(X_i'\Delta)^2 \leq c_2^2\|\Delta\|_2^2 \overset{\text{(i)}}{\leq} \frac{9}{4}c_2^2c_1^{-4}\exp(2M_1M_5)\lambda_\beta^2 s_\beta,$$

where (i) follows by (27). By the assumption of $\|\beta_{(1)}\|_0 = o(n/\log p)$, it follows that $\mathbb{E}_{n,H_B}(X_i'\Delta)^2 = o_P(1)$. The proof is complete. □

Lemma 15. *Let $\{X_i\}_{i=1}^m$ be i.i.d sub-Gaussian random vectors. Let $\{\tilde{X}_i\}_{i=1}^m$ be an independent copy of $\{X_i\}_{i=1}^m$. Define $\mathcal{D}(J) = \{h \in \mathbb{R}^p : \text{support}(h) \subseteq J,\ \|h\|_2 = 1\}$, where $J \subseteq \{1,...,p\}$. Then*

$$\sup_{h_1,h_2\in\mathcal{D}(J)}\left|h_1'\left[\frac{1}{m}\sum_{i=1}^m(X_iX_i' - \tilde{X}_i\tilde{X}_i')\right]h_2\right| = O_P\left(\sqrt{|J|/m}\right).$$

Proof. Let $s = |J|$. Let $X_{i,J} = Z_i$ and $\tilde{X}_{i,J} = \tilde{Z}_i$. Denote $\Sigma = \mathbb{E}(Z_i Z_i') = \mathbb{E}(\tilde{Z}_i \tilde{Z}_i')$. By the sub-Gaussian assumption, it follows by Proposition 2.1 of Vershynin [2012b] that

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (Z_i Z_i' - \Sigma) \right\|_{\text{spectral}} = O_P\left(\sqrt{|J|/m}\right),$$

where $\|\cdot\|_{\text{spectral}}$ denotes the spectral norm, i.e., the maximal singular value. Similarly, we can show

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (\tilde{Z}_i \tilde{Z}_i' - \Sigma) \right\|_{\text{spectral}} = O_P\left(\sqrt{|J|/m}\right).$$

Thus,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (Z_i Z_i' - \tilde{Z}_i \tilde{Z}_i') \right\|_{\text{spectral}} = O_P\left(\sqrt{|J|/m}\right).$$

The desired result follows by notice that

$$\left| h_1' \left[\frac{1}{m} \sum_{i=1}^{m} (X_i X_i' - \tilde{X}_i \tilde{X}_i') \right] h_2 \right| \leq \|h_1\|_2 \|h_2\|_2 \left\| \frac{1}{m} \sum_{i=1}^{m} (Z_i Z_i' - \tilde{Z}_i \tilde{Z}_i') \right\|_{\text{spectral}}.$$

□

Lemma 16. *Let $\{X_i\}_{i=1}^m$ be i.i.d sub-Gaussian random vectors . Let $\{\tilde{X}_i\}_{i=1}^m$ be an independent copy of $\{X_i\}_{i=1}^m$. Let $J \subset \{1, ..., p\}$. Then*

$$\sup_{h_1, h_2 \in \mathcal{C}(J,3) \bigcap \mathbb{S}^{p-1}} \left| h_1' \left[\frac{1}{m} \sum_{i=1}^{m} (X_i X_i' - \tilde{X}_i \tilde{X}_i') \right] h_2 \right| = O_P\left(\sqrt{|J|/m}\right).$$

Proof of Lemma 16. Let $s = |J|$. Fix $\delta \in (0, 1)$. By Lemma 14 of Rudelson and Zhou [2013], we have

$$\bigcup_{|J| \leq s} \mathcal{C}(J, 3) \bigcap \mathbb{S}^{p-1} \subset (1-\delta)^{-1} \text{conv} \left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1} \right)$$

for some $d \asymp s$, where V_J is the space spanned by $\{e_j\}_{j \in J}$ and e_j is the j-th column of the $p \times p$ identity matrix.

Let $\Omega = \frac{1}{m} \sum_{i=1}^{m} (X_i X_i' - \tilde{X}_i \tilde{X}_i')$. We first fix $h_2 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}$. Notice that

$$\sup_{h_1 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2| \leq \frac{1}{1-\delta} \sup_{h_1 \in \text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)} |h_1' \Omega h_2| \tag{28}$$

$$\overset{\text{(i)}}{=} \frac{1}{1-\delta} \sup_{h_1 \in \bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2|, \tag{29}$$

where (i) follows by the fact that the maximum takes place at extreme points of the set $\text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)$ due to the convexity of the mapping $h_1 \mapsto |h_1' \Omega h_2|$.

Define the function $h_2 \mapsto f(h_2)$ by

$$f(h_2) = \sup_{h_1 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2| .$$

Notice that $f(\cdot)$ is convex since it is the supreme of convex functions. We observe that

$$\begin{aligned}
\sup_{h_1, h_2 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2| &= \sup_{h_2 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}} f(h_2) \\
&\overset{\text{(i)}}{\leq} \frac{1}{1-\delta} \sup_{h_2 \in \text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)} f(h_2) \\
&\overset{\text{(ii)}}{=} \frac{1}{1-\delta} \sup_{h_2 \in \bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}} f(h_2),
\end{aligned}$$

where (i) follows by (29) and (ii) follows by the fact that the maximum takes place at extreme points of the set conv $\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)$ due to the convexity of $f(\cdot)$. Therefore, we have

$$\sup_{h_1, h_2 \in \mathcal{C}(J) \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2| \leq \frac{1}{1-\delta} \sup_{h_1, h_2 \in \bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}} |h_1' \Omega h_2| = \frac{1}{1-\delta} \sup_{h_1, h_2 \in \mathcal{D}(J)} |h_1' \Omega h_2| .$$

Thus, the desired result follows by Lemma 15. □

A.3 Proof of Theorem 2

Proof of Theorem 2. We observe the following decomposition

$$2b_n(\hat{\mu}_{(1)} - \mu_{(1)}) = Q_A + Q_B,$$

where

$$Q_A = \sum_{i \in H_A} \left[W_i Y_i q(X_i' \hat{\theta}_{(1),A}) + \left(1 - W_i q(X_i' \hat{\theta}_{(1),A})\right) X_i' \hat{\beta}_{(1),B} - \mu_{(1)} \right]$$

and

$$Q_B = \sum_{i \in H_B} \left[W_i Y_i q(X_i' \hat{\theta}_{(1),B}) + \left(1 - W_i q(X_i' \hat{\theta}_{(1),B})\right) X_i' \hat{\beta}_{(1),A} - \mu_{(1)} \right].$$

We now characterize Q_A; completely analogous arguments hold for Q_B. Define

$$D_{1,n} = \sum_{i \in H_A} \left[1 - W_i q(X_i' \hat{\theta}_{(1),A}) \right] X_i' \left(\hat{\beta}_{(1),B} - \beta_{(1)} \right)$$

and

$$D_{2,n} = \sum_{i \in H_A} W_i \varepsilon_{i,(1)} \left(q(X_i' \hat{\theta}_{(1),A}) - q(X_i' \theta_{(1)}) \right).$$

Notice that

$$\begin{aligned}
Q_A &\overset{\text{(i)}}{=} \sum_{i \in H_A} \left[W_i \varepsilon_{i,(1)} q(X_i' \hat{\theta}_{(1),A}) + \left[1 - W_i q(X_i' \hat{\theta}_{(1),A}) \right] X_i' \left(\hat{\beta}_{(1),B} - \beta_{(1)} \right) + X_i' \beta_{(1)} - \mu_{(1)} \right] \\
&= \sum_{i \in H_A} \left[W_i \varepsilon_{i,(1)} q(X_i' \theta_{(1)}) + X_i' \beta_{(1)} - \mu_{(1)} \right] + D_{1,n} + D_{2,n}, \qquad (30)
\end{aligned}$$

where (i) follows by $W_iY_i = W_iY_i(1) = W_i(X_i'\beta_{(1)} + \varepsilon_{i,(1)})$.

Now we bound $D_{1,n}$ and $D_{2,n}$. Using Hölder's inequality, we have

$$|D_{1,n}| \leq \left\| \sum_{i \in H_A} \left[1 - W_i q(X_i'\hat{\theta}_{(1),A})\right] X_i \right\|_\infty \|\hat{\beta}_{(1),B} - \beta_{(1)}\|_1 \tag{31}$$

$$\overset{\text{(i)}}{\leq} b_n \lambda_\theta \|\hat{\beta}_{(1),B} - \beta_{(1)}\|_1 \tag{32}$$

$$\overset{\text{(ii)}}{=} b_n O\left(\sqrt{n^{-1}\log p}\right) O_P\left(\|\beta_{(1)}\|_0 \sqrt{b_n^{-1}\log p}\right) \overset{\text{(iii)}}{=} o_P(\sqrt{n}), \tag{33}$$

where (i) follows by (18), (ii) follows by $b_n \asymp n$, $\lambda_\theta = O(\sqrt{n^{-1}\log p})$ and Lemma 14 and (iii) follows by $\|\beta_{(1)}\|_0 = o(\sqrt{n}/\log p)$.

Notice that $\{\varepsilon_{i,(1)}\}_{i\in H_A}$ is independent across i conditional on $\{(X_i, W_i)\}_{i\in H_A}$ and that $\mathbb{E}(\varepsilon_{i,(1)} \mid \{(X_i, W_i)\}_{i\in H_A}) = 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[D_{2,n}^2 \mid \{(X_i, W_i)\}_{i\in H_A}\right] \\
&\quad= \sum_{i\in H_A} \mathbb{E}(\varepsilon_{i,(1)}^2 \mid \{(X_i, W_i)\}_{i\in H_A}) W_i \left(q(X_i'\hat{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right)^2 \\
&\quad= \sum_{i\in H_A} \mathbb{E}(\varepsilon_{i,(1)}^2 \mid X_i) \left(q(X_i'\hat{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right)^2 W_i \\
&\quad\leq \left(\max_{1\leq i\leq n} \mathbb{E}(\varepsilon_{i,(1)}^2 \mid X_i)\right) \left[\sum_{i\in H_A} W_i \left(q(X_i'\hat{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right)^2\right] \overset{\text{(i)}}{=} o_P(n),
\end{aligned}$$

where (i) follows by $\max_{1\leq i\leq n} \mathbb{E}(\varepsilon_{i,(1)}^2 \mid X_i) = O_P(1)$ (due to the assumption of sub-Gaussian $\varepsilon_{i,(1)}$) and Lemma 13. Hence,

$$D_{2,n} = o_P(\sqrt{n}). \tag{34}$$

In light of the decomposition in (30), it follows from (33) and (34) that

$$Q_A = \sum_{i\in H_A} \left[W_i \varepsilon_{i,(1)} q(X_i'\theta_{(1)}) + X_i'\beta_{(1)} - \mu_{(1)}\right] + o_P(\sqrt{n}).$$

Similarly, we can show that

$$Q_B = \sum_{i\in H_B} \left[W_i \varepsilon_{i,(1)} q(X_i'\theta_{(1)}) + X_i'\beta_{(1)} - \mu_{(1)}\right] + o_P(\sqrt{n}).$$

The desired result follows. □

A.4 Proof of Theorem 3

A.4.1 Preliminary results for proving Theorem 3

Lemma 17. *Let Assumption 1 hold. Suppose that $s \ll \sqrt{n}/\log p$. Then there exists a constant $c_3 > 0$ depending only on M_1 such that*

$$\mathbb{P}\left(\max_{v\in\bigcup_{|J|\leq s}\mathcal{C}(J,1)} \frac{\left[\mathbb{E}_{n,H_A}(X_i'v)^4\right]^{1/4}}{\|v\|_2} \leq c_3\right) \to 1.$$

Proof of Lemma 17. The proof proceeds in two steps. We first derive a large deviation bound for $\mathbb{E}_{n,H_A}(X_i'v)^4 - \mathbb{E}(X_i'v)^4$ and then use a covering argument and reduction principle to prove the desired result.

Step 1: bound $\mathbb{E}_{n,H_A}(X_i'v)^4 - \mathbb{E}(X_i'v)^4$ for any $v \in \mathbb{S}^{p-1}$.

Fix $v \in \mathbb{S}^{p-1}$. Let $Z_i = (X_i'v)^4 - \mathbb{E}(X_i'v)^4$. Notice that $X_i'v$ has a bounded sub-Gaussian norm by Assumption 1. It follows that $|Z_i|^{1/4}$ is sub-Gaussian. Hence, there exists a constant $C_1 > 0$ depending only on M_1 such that

$$\mathbb{P}(|Z_i|^{1/4} > z) \leq \exp(1 - C_1 z^2)$$

for any $z > 0$. Thus, $\mathbb{P}(|Z_i| > z) \leq \exp(1 - C_1 z^{1/2})$ for $z > 0$. Now we apply Theorem 1 of Merlevède et al. [2011]. Notice that since we have i.i.d data, we can take γ_1 in Equation (2.6) therein to be ∞ and thus we can take $\gamma = 1/2$ in their notation. It follows by their Theorem 1 and Remark 3 that there exist constants $C_2, ..., C_6 > 0$ depending only on M_1 such that for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{b_n}\left((X_i'v)^4 - \mathbb{E}(X_i'v)^4\right)\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^{b_n} Z_i\right| \geq t\right)$$
$$\leq b_n \exp\left(-C_2 t^{1/2}\right) + \exp\left(-\frac{t^2}{C_3 + C_4 b_n}\right) + \exp\left[-C_5 b_n^{-1} t^2 \exp\left(C_6 t^{1/4}/\sqrt{\log t}\right)\right]. \quad (35)$$

Moreover, the bounded sub-Gaussian norm of $X_i'v$ implies that there exists a constant $C_7 > 0$ depending on M_1 such that

$$\mathbb{E}(X_i'v)^4 \leq C_7. \quad (36)$$

Step 2: prove the desired result.

Fix $\delta \in (0, 1)$. Let $X_A = (X_1, ..., X_{b_n})' \in \mathbb{R}^{b_n \times p}$. By Lemma 14 of Rudelson and Zhou [2013], we have

$$\bigcup_{|J| \leq s} \mathcal{C}(J, 1) \bigcap \mathbb{S}^{p-1} \subset (1 - \delta)^{-1} \text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)$$

for $d = C_8 s$, where $C_8 > 0$ is a constant depending only on δ, V_J is the space spanned by $\{e_j\}_{j \in J}$ and e_j is the j-th column of the $p \times p$ identity matrix. Notice that

$$\sup_{v \in \bigcup_{|J| \leq s} \mathcal{C}(J,1) \bigcap \mathbb{S}^{p-1}} \|X_A v\|_4 \leq \frac{1}{1-\delta} \sup_{v \in \text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)} \|X_A v\|_4 \quad (37)$$
$$\overset{\text{(i)}}{=} \frac{1}{1-\delta} \sup_{v \in \bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}} \|X_A v\|_4, \quad (38)$$

where (i) follows by the fact that the maximum takes place at extreme points of the set $\text{conv}\left(\bigcup_{|J| \leq d} V_J \bigcap \mathbb{S}^{p-1}\right)$ due to the convexity of the mapping $v \mapsto \|X_A v\|_4$.

Now we use the standard covering argument. For any $J \subset \{1, ..., p\}$ with $|J| = s$, we can find a δ-net

$$\mathcal{T}_J = \{v_{(1)}^J, ..., v_{(N)}^J\}$$

for $V_J \bigcap \mathbb{S}^{p-1}$. By Lemma 20 of Rudelson and Zhou [2013], this can be done with $N \leq (3/\delta)^s$. Thus, we can use $\mathcal{T} = \bigcup_{|J|=d} \mathcal{T}_J$ as a δ-net for $\bigcup_{|J|\leq d} V_J \bigcap \mathbb{S}^{p-1}$. Notice that

$$|\mathcal{T}| \leq (3/\delta)^s \binom{p}{d} \leq \left(\frac{3ep}{d\delta}\right)^d < (3e\delta^{-1}p)^d \tag{39}$$

Let

$$S = \sup_{v \in \bigcup_{|J|\leq d} V_J \bigcap \mathbb{S}^{p-1}} \|X_A v\|_4.$$

For any $v_0 \in \bigcup_{|J|\leq d} V_J \bigcap \mathbb{S}^{p-1}$, we can find J with $|J| \leq s$ and $v_1 \in \mathcal{T}$ such that $\|v_1 - v_0\|_2 \leq \delta$ and $v_1, v_0 \in V_J \bigcap \mathbb{S}^{p-1}$. Therefore, $(v_1 - v_0)/\|v_1 - v_0\|_2 \in V_J \bigcap \mathbb{S}^{p-1}$. Now we observe that

$$\begin{aligned}
\|X_A v_0\|_4 &\leq \|X_A v_1\|_4 + \|X_A(v_1 - v_0)\|_4 \\
&\leq \max_{v\in\mathcal{T}} \|X_A v\|_4 + \|v_1 - v_0\|_2 \cdot \left\| X_A \frac{v_1 - v_0}{\|v_1 - v_0\|_2} \right\|_4 \\
&\leq \max_{v\in\mathcal{T}} \|X_A v\|_4 + \delta \left\| X_A \frac{v_1 - v_0}{\|v_1 - v_0\|_2} \right\|_4 \overset{\text{(i)}}{\leq} \max_{v\in\mathcal{T}} \|X_A v\|_4 + \delta S,
\end{aligned}$$

where (i) follows by $(v_1 - v_0)/\|v_1 - v_0\|_2 \in V_J \bigcap \mathbb{S}^{p-1}$. Since the above holds for any $v_0 \in \bigcup_{|J|\leq d} V_J \bigcap \mathbb{S}^{p-1}$, we have that

$$S = \sup_{v_0 \in \bigcup_{|J|\leq d} V_J \bigcap \mathbb{S}^{p-1}} \|X_A v_0\|_4 \leq \max_{v\in\mathcal{T}} \|X_A v\|_4 + \delta S,$$

which means that

$$S \leq \frac{1}{1-\delta} \max_{v\in\mathcal{T}} \|X_A v\|_4. \tag{40}$$

By (38) and (40), we have

$$\sup_{v \in \bigcup_{|J|\leq s} \mathcal{C}(J,1) \bigcap \mathbb{S}^{p-1}} \|X_A v\|_4 \leq (1-\delta)^{-2} \max_{v\in\mathcal{T}} \|X_A v\|_4. \tag{41}$$

By (35), it follows that for any $v \in \mathcal{T}$ and for any $x > 0$

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{i=1}^{b_n} \left((X_i' v)^4 - \mathbb{E}(X_i' v)^4 \right) \right| \geq b_n x \right) \\
&\leq b_n \exp\left(-C_2 \sqrt{b_n x}\right) + \exp\left(-\frac{b_n^2 x^2}{C_3 + C_4 b_n}\right) + \exp\left[-C_5 b_n x^2 \exp\left(C_6 \frac{(b_n x)^{1/4}}{\sqrt{\log(b_n x)}}\right)\right].
\end{aligned}$$

By the union bound and (39), it follows that for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{v\in\mathcal{T}}\left|\mathbb{E}_{n,H_A}(X_i'v)^4 - \mathbb{E}(X_i'v)^4\right| \geq x\right) \\
&= \mathbb{P}\left(\max_{v\in\mathcal{T}}\left|\sum_{i=1}^{b_n}\left((X_i'v)^4 - \mathbb{E}(X_i'v)^4\right)\right| \geq b_n x\right) \\
&\quad \leq \exp\left(\log b_n + d\log(3e\delta^{-1}) + d\log p - C_2\sqrt{b_n x}\right) \\
&\qquad + \exp\left(d\log(3e\delta^{-1}) + d\log p - \frac{b_n^2 x^2}{C_3 + C_4 b_n}\right) \\
&\qquad\quad + \exp\left[d\log(3e\delta^{-1}) + d\log p - C_5 b_n x^2 \exp\left(C_6\frac{(b_n x)^{1/4}}{\sqrt{\log(b_n x)}}\right)\right].
\end{aligned}$$

Since $d \asymp s$ and $b_n \asymp n$, the assumption of $s \ll \sqrt{n}/\log p$ implies that each of the three terms on the right-hand size of the above display tends to zero for any choice of $x > 0$. Hence, $\max_{v\in\mathcal{T}}\left|\mathbb{E}_{n,H_A}(X_i'v)^4 - E(X_i'v)^4\right| = o_P(1)$. By (36), we have

$$\mathbb{P}\left(\max_{v\in\mathcal{T}}\left|\mathbb{E}_{n,H_A}(X_i'v)^4\right| \leq 2C_7\right) \to 1.$$

Therefore, it follows by (41) that

$$\mathbb{P}\left(\sup_{v\in\bigcup_{|J|\leq s}\mathcal{C}(J,1)\bigcap\mathbb{S}^{p-1}}\|X_A v\|_4 \leq 2C_7/(1-\delta)^2\right) \to 1.$$

The proof is complete. □

Lemma 18. *Under the assumptions of Theorem 3, we have*

$$\sum_{i\in H_A} X_i'\delta_{\theta,A}W_i\dot{q}(X_i'\hat{\theta}_{(1),A})X_i'\left(\hat{\beta}_{(1),B} - \hat{\beta}_{(1)}\right) = o_P(\sqrt{n}).$$

Proof of Lemma 18. Let $\delta_{\theta,A} = \hat{\theta}_{(1),A} - \theta_{(1)}$ and $\delta_{\beta,B} = \hat{\beta}_{(1),B} - \hat{\beta}_{(1)}$. Let $\tilde{X}_i = X_{b_n+i}$, $\tilde{W}_i = W_{b_n+i}$ and $\tilde{Y}_i = Y_{b_n+i}$. Then we need to show

$$\sum_{i=1}^{b_n} W_i\dot{q}(X_i'\hat{\theta}_{(1),A})\delta_{\theta,A}'X_iX_i'\delta_{\beta,B} = o_P(\sqrt{n}).$$

We decompose

$$
\begin{aligned}
&\sum_{i=1}^{b_n} W_i \dot{q}(X_i'\hat{\theta}_{(1),A})\delta_{\theta,A}' X_i X_i'\delta_{\beta,B} \\
&\quad = \sum_{i=1}^{b_n} \tilde{X}_i'\delta_{\theta,A}\tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\tilde{X}_i'\delta_{\beta,B} \\
&\qquad + \sum_{i=1}^{b_n} \left[\dot{q}(\tilde{X}_i'\theta_{(1)}) - \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\right] \tilde{W}_i \delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B} \\
&\qquad + \sum_{i=1}^{b_n} \left[\dot{q}(X_i'\theta_{(1)}) W_i \delta_{\theta,A}' X_i X_i'\delta_{\beta,B} - \dot{q}(\tilde{X}_i'\theta_{(1)})\tilde{W}_i\delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B}\right] \\
&\qquad + \sum_{i=1}^{b_n} \left[\dot{q}(X_i'\hat{\theta}_{(1),A}) - \dot{q}(X_i'\theta_{(1)})\right] W_i\delta_{\theta,A}' X_i X_i'\delta_{\beta,B}. \qquad (42)
\end{aligned}
$$

We bound these four terms in four steps.
Step 1: show that $\sum_{i=1}^{b_n} \tilde{X}_i'\delta_{\theta,A}\tilde{W}_i\dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\tilde{X}_i'\delta_{\beta,B} = o_P(\sqrt{n})$.
By the KKT condition (17), we have

$$
\left\| \frac{1}{b_n}\sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})(\tilde{X}_i'\hat{\beta}_{(1),B} - \tilde{Y}_i)\tilde{X}_i \right\|_\infty \le \lambda_\beta/4.
$$

Notice that

$$
\frac{1}{b_n}\sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})(\tilde{X}_i'\beta_{(1)} - \tilde{Y}_i)\tilde{X}_i = -\mathbb{E}_{n,H_B} W_i X_i \dot{q}(X_i'\hat{\theta}_{(1),B})\varepsilon_{i,(1)}.
$$

By Lemma 6 (the version with $\mathbb{E}_{n,H_B}$), we have

$$
\left\| \frac{1}{b_n}\sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})(\tilde{X}_i'\beta_{(1)} - \tilde{Y}_i)\tilde{X}_i \right\|_\infty \le \lambda_\beta/4
$$

Hence,

$$
\left\| \sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\tilde{X}_i\tilde{X}_i'\delta_{\beta,B} \right\|_\infty = O_P(b_n\lambda_\beta)
$$

It follows by Lemma 13 that

$$
\begin{aligned}
&\left| \sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B} \right| \\
&\qquad \le \|\delta_{\theta,A}\|_1 \left\| \sum_{i=1}^{b_n} \tilde{W}_i \dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B})\tilde{X}_i\tilde{X}_i'\delta_{\beta,B} \right\|_\infty \\
&\qquad = O_P(b_n\lambda_\theta) O_P\left(\|\theta_{(1)}\|_0 \sqrt{b_n^{-1}\log p} \right) = o_P(\sqrt{n}).
\end{aligned}
$$

Step 2: show that $\sum_{i=1}^{b_n}\left[\dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B}) - \dot{q}(\tilde{X}_i'\theta_{(1)})\right]\tilde{W}_i\delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B} = o_P(\sqrt{n})$.

Lemma 3 implies $\mathbb{P}(\|\hat{\theta}_{(1),B}\|_\infty \leq M_5) \to 1$. Since

$$\|\tilde{X}\hat{\theta}_{(1),A}\|_\infty \leq \|\tilde{X}\|_\infty\|\hat{\theta}_{(1),A}\|_1 \leq M_1\|\hat{\theta}_{(1),A}\|_1,$$

we have $\mathbb{P}(\|\tilde{X}\hat{\theta}_{(1),B}\|_\infty \leq M_1M_5) \to 1$. By assumption, $\|\tilde{X}\theta_{(1)}\|_\infty \leq \|\tilde{X}\|_\infty\|\theta_{(1)}\|_1 \leq M_1M_5$. By Taylor's theorem, we have that

$$\max_{1\leq i\leq b_n}\frac{|\dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B}) - \dot{q}(\tilde{X}_i'\theta_{(1)})|}{|\tilde{X}_i'\delta_{\theta,B}|} \leq \max_{|t|\leq\max\{\|\tilde{X}\theta_{(1)}\|_\infty, \|\tilde{X}\hat{\theta}_{(1),B}\|_\infty\}}|\ddot{q}(t)| = O_P(1).$$

Therefore, we have

$$\begin{aligned}
&\left|\sum_{i=1}^{b_n}\left[\dot{q}(\tilde{X}_i'\hat{\theta}_{(1),B}) - \dot{q}(\tilde{X}_i'\theta_{(1)})\right]\tilde{W}_i\delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B}\right| \\
&\quad\leq O_P(1)\sum_{i=1}^{b_n}|\tilde{X}_i'\delta_{\theta,B}|\cdot|\tilde{X}_i'\delta_{\theta,A}|\cdot|\tilde{X}_i'\delta_{\beta,B}| \\
&\quad\leq O_P(1)\sqrt{\left(\sum_{i=1}^{b_n}|\tilde{X}_i'\delta_{\theta,B}|^2\cdot|\tilde{X}_i'\delta_{\theta,A}|^2\right)\left(\sum_{i=1}^{b_n}(\tilde{X}_i'\delta_{\beta,B})^2\right)} \\
&\quad\leq O_P(1)\sqrt{\sqrt{\left(\sum_{i=1}^{b_n}|\tilde{X}_i'\delta_{\theta,B}|^4\right)\left(\sum_{i=1}^{b_n}|\tilde{X}_i'\delta_{\theta,A}|^4\right)}\left(\sum_{i=1}^{b_n}(\tilde{X}_i'\delta_{\beta,B})^2\right)} \\
&\quad\overset{\text{(i)}}{\leq} O_P(1)\sqrt{\sqrt{O_P\left(b_n\|\delta_{\theta,B}\|_2^4\right)O_P\left(b_n\|\delta_{\theta,A}\|_2^4\right)}\left(\sum_{i=1}^{b_n}(\tilde{X}_i'\delta_{\beta,B})^2\right)} \\
&\quad\overset{\text{(ii)}}{\leq} O_P(1)\sqrt{\sqrt{O_P\left(b_n[s_\theta b_n^{-1}\log p]^2\right)O_P\left(b_n[s_\theta b_n^{-1}\log p]^2\right)}(s_\beta\log p)} \overset{\text{(iii)}}{=} o_P(\sqrt{n})
\end{aligned}$$

where (i) follows by Lemma 17, together with $\|\theta_{(1)}\|_0 \ll \sqrt{n}/\log p$ and the fact that $\mathbb{P}(\delta_{\theta,A}, \delta_{\theta,B} \in \mathcal{C}(\text{supp}(\theta_{(1)}), 1)) \to 1$ (Lemma 13), (ii) follows by Lemmas 13 and 14 and (iii) follows by $b_n \asymp n$, $s_\theta \ll \sqrt{n}/\log p$ and $s_\beta \ll n/\log p$.

Step 3: show that $\sum_{i=1}^{b_n}\left[\dot{q}(X_i'\theta_{(1)})W_i\delta_{\theta,A}'X_iX_i'\delta_{\beta,B} - \dot{q}(\tilde{X}_i'\theta_{(1)})\tilde{W}_i\delta_{\theta,A}'\tilde{X}_i\tilde{X}_i'\delta_{\beta,B}\right] = o_P(\sqrt{n})$.

By Lemmas 13 and 14, $\delta_{\theta,A} \in \mathcal{C}(\text{supp}(\theta_{(1)}), 3)$ and $\delta_{\beta,B} \in \mathcal{C}(\text{supp}(\beta_{(1)}), 3)$ with probability tending to one. Let $S = \text{supp}(\theta_{(1)})\bigcup\text{supp}(\beta_{(1)})$. Notice that $\mathcal{C}(\text{supp}(\theta_{(1)}), 1) \subset \mathcal{C}(S, 1) \subset \mathcal{C}(S, 3)$ and $\mathcal{C}(\text{supp}(\beta_{(1)}), 3) \subset \mathcal{C}(S, 3)$. It follows that $\mathbb{P}(\delta_{\theta,A}, \delta_{\beta,B} \in \mathcal{C}(S, 3)) \to 1$.

Hence, by $|S| \leq s_\theta + s_\beta$ and Lemma 16, we obtain

$$\begin{aligned}
&\left| \sum_{i=1}^{b_n} \left[\dot{q}(X_i'\theta_{(1)}) W_i \delta_{\theta,A}' X_i X_i' \delta_{\beta,B} - \dot{q}(\tilde{X}_i'\theta_{(1)}) \tilde{W}_i \delta_{\theta,A}' \tilde{X}_i \tilde{X}_i' \delta_{\beta,B} \right] \right| \\
&= \left| \delta_{\theta,A}' \left(\sum_{i=1}^{b_n} \left[\dot{q}(X_i'\theta_{(1)}) W_i X_i X_i' - \dot{q}(\tilde{X}_i'\theta_{(1)}) \tilde{W}_i \tilde{X}_i \tilde{X}_i' \right] \right) \delta_{\beta,B} \right| \\
&\quad \leq O_P\left(O_P\left(\sqrt{b_n (s_\beta + s_\theta)} \right) \right) \|\delta_{\theta,A}\|_2 \|\delta_{\beta,B}\|_2 \\
&\quad \overset{\text{(i)}}{=} o_P\left(\sqrt{n (s_\beta + s_\theta) s_\beta s_\theta n^{-2}} \log p \right) \overset{\text{(ii)}}{=} o_P(\sqrt{n}),
\end{aligned}$$

where (i) follows by Lemmas 13 and 14, together with $b_n \asymp n$ and (iii) follows by the conditions $s_\theta \ll \sqrt{n}/\log p$ and $s_\beta \lesssim n^{3/4}/\log p$.

Step 4: show that $\sum_{i=1}^{b_n} \left[\dot{q}(X_i'\hat{\theta}_{(1),A}) - \dot{q}(X_i'\theta_{(1)}) \right] W_i \delta_{\theta,A}' X_i X_i' \delta_{\beta,B} = o_P(\sqrt{n})$.

Similar to Step 2, we can show that $\max_{1 \leq i \leq b_n} |\dot{q}(X_i'\hat{\theta}_{(1),A}) - \dot{q}(X_i'\theta_{(1)})| / |X_i'\delta_{\theta,A}| \leq O_P(1)$. Hence,

$$\begin{aligned}
&\left| \sum_{i=1}^{b_n} \left[\dot{q}(X_i'\hat{\theta}_{(1),A}) - \dot{q}(X_i'\theta_{(1)}) \right] W_i \delta_{\theta,A}' X_i X_i' \delta_{\beta,B} \right| \\
&\leq O_P(1) \sum_{i=1}^{b_n} (X_i'\delta_{\theta,A})^2 \cdot |X_i'\delta_{\beta,B}| \\
&\leq O_P(1) \sqrt{ \left(\sum_{i=1}^{b_n} (X_i'\delta_{\theta,A})^4 \right) \left(\sum_{i=1}^{b_n} (X_i'\delta_{\beta,B})^2 \right) } \\
&\overset{\text{(i)}}{\leq} O_P(1) \sqrt{ O_P\left(b_n \|\delta_{\theta,A}\|_2^4 \right) O_P\left(b_n \|\delta_{\beta,B}\|_2^2 \right) } \\
&\overset{\text{(ii)}}{\leq} O_P(1) \sqrt{ O_P\left([s_\theta b_n^{-1} \log p]^2 \right) O_P\left(s_\beta \log p \right) } \overset{\text{(iii)}}{=} o_P(\sqrt{n}),
\end{aligned}$$

where (i) follows by Lemma 17, together with $\|\theta_{(1)}\|_0 \ll \sqrt{n}/\log p$ and the fact that $\mathbb{P}(\delta_{\theta,A} \in \mathcal{C}(\text{supp}(\theta_{(1)}), 1)) \to 1$ (Lemma 13), (ii) follows by Lemmas 13 and 14 and (iii) follows by $b_n \asymp n$, $s_\theta \ll \sqrt{n}/\log p$ and $s_\beta \ll n/\log p$. The desired result follows by the above four steps together with (42). ☐

Lemma 19. *Under the assumptions of Theorem 3, we have*

$$\sum_{i \in H_A} W_i \left[q(X_i'\theta_{(1)}) - q(X_i'\hat{\theta}_{(1),A}) \right] X_i' \left(\hat{\beta}_{(1),B} - \beta_{(1)} \right) = o_P(\sqrt{n}).$$

Proof of Lemma 19. Let $\delta_{\theta,A} = \hat{\theta}_{(1),A} - \theta_{(1)}$ and $\delta_{\beta,B} = \hat{\beta}_{(1),B} - \hat{\beta}_{(1)}$. Denote

$$\ddot{q}(a) = d^2 q(a)/da^2 = \exp(-a).$$

Define $Q_i = q(X_i'\theta_{(1)}) - q(X_i'\hat{\theta}_{(1),A}) + \dot{q}(X_i'\hat{\theta}_{(1),A})X_i'\delta_{\theta,A}$. Then

$$\sum_{i\in H_A} W_i \left[q(X_i'\theta_{(1)}) - q(X_i'\hat{\theta}_{(1),A})\right] X_i' \left(\hat{\beta}_{(1),B} - \beta_{(1)}\right)$$
$$= \sum_{i\in H_A} W_i Q_i X_i'\delta_{\beta,B} - \sum_{i\in H_A} X_i'\delta_{\theta,A} W_i \dot{q}(X_i'\hat{\theta}_{(1),A})X_i'\delta_{\beta,B}. \quad (43)$$

By Lemma 18,

$$\sum_{i\in H_A} X_i'\delta_{\theta,A} W_i \dot{q}(X_i'\hat{\theta}_{(1),A})X_i'\delta_{\beta,B} = o_P(\sqrt{n}). \quad (44)$$

By assumption, $\|X\theta_{(1)}\|_1 \leq \|X\|_\infty \|\theta_{(1)}\|_1 \leq M_1 M_5$. By Lemma 13, $\mathbb{P}(\|\hat{\theta}_{(1),A}\|_1 \leq M_5 \vee \kappa_0) \to 1$. Since $\|X\hat{\theta}_{(1),A}\|_\infty \leq \|X\|_\infty \|\hat{\theta}_{(1),A}\|_1 \leq M_1 \|\hat{\theta}_{(1),A}\|_1$, we have $\mathbb{P}(\|X\hat{\theta}_{(1),A}\|_\infty \leq M_1(M_5 \vee \kappa_0)) \to 1$. Therefore, by Taylor's theorem,

$$\max_{1\leq i\leq n} \frac{|Q_i|}{(X_i'\delta_{\theta,A})^2} \leq \frac{1}{2} \sup_{|t|\leq \max\{\|X\theta_{(1)}\|_\infty, \|X\hat{\theta}_{(1),A}\|_\infty\}} \ddot{q}(t) = O_P(1).$$

Therefore,

$$\left|\sum_{i\in H_A} W_i Q_i X_i'\delta_{\beta,B}\right| \leq \sqrt{\left(\sum_{i\in H_A} Q_i^2\right)\left(\sum_{i\in H_A} (X_i'\delta_{\beta,B})^2\right)}$$
$$\leq \sqrt{O_P(1)\left(\sum_{i\in H_A} (X_i'\delta_{\theta,A})^4\right)\left(\sum_{i\in H_A} (X_i'\delta_{\beta,B})^2\right)}$$
$$\overset{\text{(i)}}{\leq} \sqrt{O_P(1) b_n \left(s_\theta b_n^{-1} \log p\right)^2 (s_\beta \log p)} \overset{\text{(ii)}}{=} o_P(\sqrt{n}), \quad (45)$$

where (i) follows by Lemmas 13 and 14 and (ii) follows by $s_\theta \ll \sqrt{n}/\log p$, $s_\beta \ll n/\log p$ and $b_n \asymp n$.

In light of (43), the desired result follows by (44) and (45). □

A.4.2 Proof of Theorem 3

Proof of Theorem 3. Similar to the proof of Theorem 2, we observe the following decomposition

$$2b_n(\hat{\mu}_{(1)} - \mu_{(1)}) = Q_A + Q_B, \quad (46)$$

where

$$Q_A = \sum_{i\in H_A} \left[W_i Y_i q(X_i'\hat{\theta}_{(1),A}) + \left(1 - W_i q(X_i'\hat{\theta}_{(1),A})\right) X_i'\hat{\beta}_{(1),B} - \mu_{(1)}\right]$$

and

$$Q_B = \sum_{i\in H_B} \left[W_i Y_i q(X_i'\hat{\theta}_{(1),B}) + \left(1 - W_i q(X_i'\hat{\theta}_{(1),B})\right) X_i'\hat{\beta}_{(1),A} - \mu_{(1)}\right].$$

We further decompose Q_A; an analogous argument applies for Q_B. Since $W_i Y_i = W_i Y_i(1) = W_i(X_i'\beta_{(1)} + \varepsilon_{i,(1)})$, we have

$$Q_A = \sum_{i\in H_A} \left[\left(1 - W_i q(X_i'\theta_{(1)})\right) X_i'\beta_{(1)} + Y_i W_i q(X_i'\theta_{(1)}) - \mu_{(1)}\right] + D_{n,1} + D_{n,2} + D_{n,3},$$

where

$$\begin{cases} D_{1,n} = \sum_{i\in H_A}\left[1 - W_i q(X_i'\theta_{(1)})\right] X_i'\left(\hat{\beta}_{(1),B} - \beta_{(1)}\right) \\ D_{2,n} = \sum_{i\in H_A} W_i\varepsilon_{i,(1)}\left(q(X_i'\hat{\theta}_{(1),A}) - q(X_i'\theta_{(1)})\right). \\ D_{3,n} = \sum_{i\in H_A} W_i\left[q(X_i'\theta_{(1)}) - q(X_i'\hat{\theta}_{(1),A})\right] X_i'\left(\hat{\beta}_{(1),B} - \beta_{(1)}\right) \end{cases}$$

By the same argument as (34) in the proof of Theorem 2, we can show

$$D_{2,n} = o_P(\sqrt{n}).$$

Notice that $\hat{\beta}_{(1),B}$ is computed using observations in H_B and is thus independent of $\{(W_i, X_i)\}_{i\in H_A}$. Also notice that $1 - W_i q(X_i'\theta_{(1)}) = -v_{i,(1)} q(X_i'\theta_{(1)})$ and that conditional on $\{X_i\}_{i\in H_A}$, $\{v_{i,(1)}\}_{i\in H_A}$ has mean zero and is independent across i. Thus, we have

$$\begin{aligned} \mathbb{E}\left(D_{1,n}^2 \mid \{X_i\}_{i\in H_A}, \hat{\beta}_{(1),B}\right) &= \sum_{i\in H_A} \mathbb{E}\left(v_{i,(1)}^2 \mid \{X_i\}_{i\in H_A}\right)\left(q(X_i'\theta_{(1)})\right)^2\left[X_i'\left(\hat{\beta}_{(1),B} - \beta_{(1)}\right)\right]^2 \\ &\overset{\text{(i)}}{\leq} 4\sum_{i\in H_A}\left(q(X_i'\theta_{(1)})\right)^2\left[X_i'\left(\hat{\beta}_{(1),B} - \beta_{(1)}\right)\right]^2 \\ &\leq 4\left[\max_{i\in H_A}\left(q(X_i'\theta_{(1)})\right)^2\right]\sum_{i\in H_A}\left[X_i'\left(\hat{\beta}_{(1),B} - \beta_{(1)}\right)\right]^2 \overset{\text{(ii)}}{=} o_P(n), \end{aligned}$$

where (i) follows by $|v_{i,(1)}| = |W_i - e_{(1)}(X_i)| \leq |W_i| + |e_{(1)}(X_i)| \leq 2$ and (ii) follows by $\max_{1\leq i\leq n} q^2(X_i'\theta_{(1)}) = O(1)$ (due to the assumption of $\|X\theta_{(1)}\|_\infty = O(1)$) and Lemma 14. Hence,

$$D_{1,n} = o_P(\sqrt{n}).$$

Lemma 19 implies

$$D_{3,n} = o_P(\sqrt{n}).$$

Thus, we have proved

$$Q_A = \sum_{i\in H_A}\left[\left(1 - W_i q(X_i'\theta_{(1)})\right) X_i'\beta_{(1)} + Y_i W_i q(X_i'\theta_{(1)}) - \mu_{(1)}\right] + o_P(\sqrt{n}).$$

By an analogous argument, we can show that

$$Q_B = \sum_{i\in H_B}\left[\left(1 - W_i q(X_i'\theta_{(1)})\right) X_i'\beta_{(1)} + Y_i W_i q(X_i'\theta_{(1)}) - \mu_{(1)}\right] + o_P(\sqrt{n}).$$

Since $2b_n/n \to 1$, the desired result follows by (46) and

$$\left(1 - W_i q(X_i'\theta_{(1)})\right) X_i'\beta_{(1)} + Y_i W_i q(X_i'\theta_{(1)}) - \mu_{(1)} = \varepsilon_{i,(1)} W_i q(X_i'\theta_{(1)}) + X_i'\beta_{(1)} - \mu_{(1)}.$$

□

A.5 Proof of Lemma 4

Proof of Lemma 4. Rearranging terms, we have

$$
\begin{aligned}
V_* &= \mathbb{E}\left[\left\{W_i \varepsilon_{i,(1)} q(X_i'\theta_{(1)}) + \left(X_i'\beta_{(1)} - \mu_{(1)}\right)\right\} - \left\{(1-W_i)\varepsilon_{i,(0)} q(X_i'\theta_{(0)}) + \left(X_i'\beta_{(0)} - \mu_{(0)}\right)\right\}\right]^2 \\
&= \mathbb{E}(\psi_{i,1} + \psi_{i,2})^2,
\end{aligned}
$$

where $\psi_{i,1} = W_i \varepsilon_{i,(1)} q(X_i'\theta_{(1)}) - (1-W_i)\varepsilon_{i,(0)} q(X_i'\theta_{(0)})$ and $\psi_{i,2} = X_i'(\beta_{(1)} - \beta_{(0)}) - \tau$. Notice that $\mathbb{E}(\psi_{i,1} \mid X_i, W_i) = 0$. Therefore,

$$
V_* = \mathbb{E}\psi_{i,1}^2 + \mathbb{E}\psi_{i,2}^2.
$$

Since $W_i(1-W_i) = 0$, $\psi_{i,1}^2 = W_i \varepsilon_{i,(1)}^2 (q(X_i'\theta_{(1)}))^2 + (1-W_i)\varepsilon_{i,(0)}^2 (q(X_i'\theta_{(0)}))^2$. The proof is complete. □

A.6 Proof of Theorem 5

Proof of Theorem 5. By Lemma 4, it suffices to show the following claims:

1. $\hat{V}_1 = \tilde{V}_1 + o_P(1)$, where $\tilde{V}_1 = n^{-1}\sum_{i=1}^n W_i \varepsilon_{i,(1)}^2 (q(X_i'\theta_{(1)}))^2$.
2. $\hat{V}_2 = \tilde{V}_2 + o_P(1)$, where $\tilde{V}_2 = n^{-1}\sum_{i=1}^n (1-W_i) \varepsilon_{i,(0)}^2 (q(X_i'\theta_{(0)}))^2$.
3. $\hat{V}_3 = \tilde{V}_3 + o_P(1)$, where $\tilde{V}_3 = n^{-1}\sum_{i=1}^n (X_i'(\beta_{(1)} - \beta_{(0)}) - \tau)^2$.

This is because the law of large numbers would imply $\tilde{V}_1 + \tilde{V}_2 + \tilde{V}_3 = V_* + o_P(1)$. We show the above three claims in three steps.

Step 1: show $\hat{V}_1 = \tilde{V}_1 + o_P(1)$.

Notice that $W_i \varepsilon_{i,(1)}^2 = W_i (Y_i - X_i'\beta_{(1)})^2$. Let $\delta_\beta = \hat{\beta}_{(1)} - \beta_{(1)}$ and $\delta_\theta = \hat{\theta}_{(1)} - \theta_{(1)}$. Then we have

$$
\begin{aligned}
\hat{V}_1 - \tilde{V}_1 = &\underbrace{n^{-1}\sum_{i=1}^n W_i \left[(Y_i - X_i'\hat{\beta}_{(1)})^2 - (Y_i - X_i'\beta_{(1)})^2\right] (q(X_i'\hat{\theta}_{(1)}))^2}_{T_1} \\
&+ \underbrace{n^{-1}\sum_{i=1}^n W_i \varepsilon_{i,(1)}^2 \left[(q(X_i'\hat{\theta}_{(1)}))^2 - (q(X_i'\theta_{(1)}))^2\right]}_{T_2}. \qquad (47)
\end{aligned}
$$

We now bound these two terms. Notice that we have $\|\delta_{\beta,1}\|_1 = O_P(\|\beta_{(1)}\|_0 \sqrt{n^{-1}\log p})$, $\|\delta_{\theta,1}\|_1 = O_P(\|\theta_{(1)}\|_0 \sqrt{n^{-1}\log p})$ and $\|\hat{\theta}_{(1)}\|_1 = O_P(1)$. This is because these bounds hold for $\hat{\theta}_{(1),A}$ and $\hat{\theta}_{(1),B}$ (Lemma 13) as well as for $\hat{\beta}_{(1),A}$ and $\hat{\beta}_{(1),B}$ (Lemma 14). We observe that

$$
\begin{aligned}
|T_1| &\le \left(\max_{1\le i\le n}(q(X_i'\hat{\theta}_{(1)}))^2\right) n^{-1}\sum_{i=1}^{n}\left|(Y_i - X_i'\hat{\beta}_{(1)})^2 - (Y_i - X_i'\beta_{(1)})^2\right| \\
&\le \left(\max_{1\le i\le n}(q(X_i'\hat{\theta}_{(1)}))^2\right)\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + \left|2n^{-1}\sum_{i=1}^{n}\varepsilon_{i,(1)}X_i'\delta_{\beta,1}\right|\right) \\
&\le \left(\max_{1\le i\le n}(q(X_i'\hat{\theta}_{(1)}))^2\right)\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + 2\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_{i,(1)}X_i\right\|_\infty \|\delta_{\beta,1}\|_1\right) \\
&\overset{\text{(i)}}{=} O_P(1)\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + 2\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_{i,(1)}X_i\right\|_\infty \|\delta_{\beta,1}\|_1\right),
\end{aligned}
$$

where (i) follows by $\|X\hat{\theta}_{(1)}\|_\infty \le \|X\|_\infty\|\hat{\theta}_{(1)}\|_1 = O_P(1)$. Following an argument similar to the proof of the third claim in Lemma 6, we can easily show that $\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_{i,(1)}X_i\right\|_\infty = O_P(\sqrt{n^{-1}\log p})$; essentially, the argument is Hoeffding inequality and the union bound since elements of $X_i\varepsilon_{i,(1)}$ have bounded sub-Gaussian norms. Therefore, the above display implies

$$
\begin{aligned}
|T_1| &= O_P(1)\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + 2O_P(\sqrt{n^{-1}\log p})\|\delta_{\beta,1}\|_1\right) \\
&= O_P(1)\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + 2O_P(\sqrt{n^{-1}\log p}) \times O_P(\|\beta_{(1)}\|_0\sqrt{n^{-1}\log p})\right) \\
&\overset{\text{(i)}}{=} O_P\left(n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2\right) + o_P(1), \qquad (48)
\end{aligned}
$$

where (i) holds by $\|\beta_{(1)}\|_0 \ll n/\log p$. Let $\delta_{\beta,1,A} = \hat{\beta}_{(1),A} - \beta_{(1)}$ and $\delta_{\beta,1,B} = \hat{\beta}_{(1),B} - \hat{\beta}_{(1)}$. Notice that

$$\delta_{\beta,1} = (\delta_{\beta,1,A} + \delta_{\beta,1,B})/2.$$

By Lemma 14, $\mathbb{P}(\delta_{\beta,1,A}, \delta_{\beta,1,B} \in \mathcal{C}(\mathrm{supp}(\beta_{(1)}), 3)) \to 1$. By essentially the same argument as in the proof of the second claim of Lemma 6, we can show that

$$\max_{v\in\mathcal{C}(\mathrm{supp}(\beta_{(1)}),3))}\frac{n^{-1}\sum_{i=1}^{n}(X_i'v)^2}{\|v\|_2^2} = O_P(1).$$

By Lemma 14 and $\|\beta_{(1)}\|_0 \ll n/\log p$, we have $\|\delta_{\beta,1,A}\|_2 = o_P(1)$ and $\|\delta_{\beta,1,B}\|_2 = o_P(1)$. It follows that $n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1,A})^2 = O_P(\|\delta_{\beta,1,A}\|_2^2) = o_P(1)$ and $n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1,B})^2 = O_P(\|\delta_{\beta,1,B}\|_2^2) = o_P(1)$. Hence, the elementary bound yields

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 &= 0.25n^{-1}\sum_{i=1}^{n}\left(X_i'\delta_{\beta,1,A} + X_i'\delta_{\beta,1,B}\right)^2 \\
&\le 0.5n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1,A})^2 + 0.5n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1,B})^2 = o_P(1).
\end{aligned}
$$

In light of (48), we have

$$T_1 = o_P(1). \qquad (49)$$

Now we bound T_2. Let $f(x) = (q(x))^2$. Then

$$\dot{f}(x) = df(x)/dx = -2\exp(-2x) - 2\exp(-x).$$

By Taylor's theorem, there exists $r_i \in [0,1]$ such that

$$(q(X_i'\hat{\theta}_{(1)}))^2 - (q(X_i'\theta_{(1)}))^2 = (X_i'\delta_{\theta,1})\dot{f}\left(r_i X_i'\hat{\theta}_{(1)} + (1-r_i)X_i'\theta_{(1)}\right).$$

Since $\|\hat{\theta}_{(1)}\|_1 = O_P(1)$, $\|\theta_{(1)}\|_1 = O(1)$ and $\|X\|_\infty = O(1)$, we have that

$$\max_{1\le i\le n} |r_i X_i'\hat{\theta}_{(1)} + (1-r_i)X_i'\theta_{(1)}| = O_P(1).$$

Therefore,

$$\max_{1\le i\le n} \left| \frac{(q(X_i'\hat{\theta}_{(1)}))^2 - (q(X_i'\theta_{(1)}))^2}{X_i'\delta_{\theta,1}} \right| = O_P(1).$$

Let $\delta_{\theta,1,A} = \hat{\theta}_{(1),A} - \theta_{(1)}$ and $\delta_{\theta,1,B} = \hat{\theta}_{(1),B} - \theta_{(1)}$. Thus,

$$\begin{aligned}
|T_2| &\le n^{-1}\sum_{i=1}^n W_i\varepsilon_{i,(1)}^2 \left|(q(X_i'\hat{\theta}_{(1)}))^2 - (q(X_i'\theta_{(1)}))^2\right| \\
&\le O_P(1) n^{-1}\sum_{i=1}^n W_i\varepsilon_{i,(1)}^2 |X_i'\delta_{\theta,1}| \\
&\le O_P(1)\sqrt{\left(n^{-1}\sum_{i=1}^n \varepsilon_{i,(1)}^4\right) \times \left(n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1})^2\right)} \\
&\overset{\text{(i)}}{=} O_P(1)\sqrt{O_P(1) \times \left(n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1})^2\right)} \\
&= O_P(1)\sqrt{O_P(1) \times \left(0.25 n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1,A} + X_i'\delta_{\theta,1,B})^2\right)} \\
&\le O_P(1)\sqrt{O_P(1) \times \left(0.5 n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1,A})^2 + 0.5 n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1,B})^2\right)},
\end{aligned}$$

where (i) follows by the law of large numbers and the fact that $\varepsilon_{i,(1)}$ is sub-Gaussian. Now by essentially the same argument as above, we can show that

$$n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1,A})^2 = O_P(\|\delta_{\theta,1,A}\|_2^2)$$

and similarly $n^{-1}\sum_{i=1}^n (X_i'\delta_{\theta,1,B})^2 = O_P(\|\delta_{\theta,1,B}\|_2^2)$. By Lemma 13 and the condition of $\|\theta_{(1)}\|_0 \ll n/\log p$, we have that

$$\|\delta_{\theta,1,A}\|_2 = o_P(1)$$

and $\|\delta_{\theta,1,B}\|_2 = o_P(1)$. Therefore, the above display implies $T_2 = o_P(1)$. Thus, in light of (47) and (49), we have proved $\hat{V}_1 - \tilde{V}_1 = o_P(1)$.

Step 2: show $\hat{V}_2 = \tilde{V}_2 + o_P(1)$.
The argument is completely analogous to Step 1 and is thus omitted.
Step 3: show $\hat{V}_3 = \tilde{V}_3 + o_P(1)$.
Let $\xi_i = X_i'(\beta_{(1)} - \beta_{(0)}) - \tau$, $\hat{\xi}_i = X_i'(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}) - \hat{\tau}$ and $u_i = \hat{\xi}_i - \xi_i$. Notice that

$$\begin{aligned}
\left|\hat{V}_3 - \tilde{V}_3\right| &= \left|n^{-1}\sum_{i=1}^{n}\left((\xi_i + u_i)^2 - \xi_i^2\right)\right| \\
&\leq \left|n^{-1}\sum_{i=1}^{n} u_i^2\right| + 2\left|n^{-1}\sum_{i=1}^{n}\xi_i u_i\right| \\
&\leq \left|n^{-1}\sum_{i=1}^{n} u_i^2\right| + 2\sqrt{\left(n^{-1}\sum_{i=1}^{n}\xi_i^2\right) \times \left(n^{-1}\sum_{i=1}^{n} u_i^2\right)} \\
&= \left|n^{-1}\sum_{i=1}^{n} u_i^2\right| + \sqrt{O_P(1) \times \left(n^{-1}\sum_{i=1}^{n} u_i^2\right)}. \qquad (50)
\end{aligned}$$

Thus, it only remains to show $n^{-1}\sum_{i=1}^{n} u_i^2 = o_P(1)$. Let $\delta_{\beta,j} = \hat{\beta}_{(j)} - \beta_{(j)}$ for $j \in \{0, 1\}$. By elementary inequality of $(a + b + c)^2 \leq 4a^2 + 4b^2 + 2c^2$, we have

$$u_i^2 = \left(X_i'\delta_{\beta,1} - X_i'\delta_{\beta,0} - (\hat{\tau} - \tau)\right)^2 \leq 4(X_i'\delta_{\beta,1})^2 + 4(X_i'\delta_{\beta,0})^2 + 2(\hat{\tau} - \tau)^2.$$

Thus,

$$n^{-1}\sum_{i=1}^{n} u_i^2 \leq 4n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,1})^2 + 4n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,0})^2 + 2(\hat{\tau} - \tau)^2.$$

By the arguments in Step 1 and Step 2, we have already proved $n^{-1}\sum_{i=1}^{n}(X_i'\delta_{\beta,j})^2 = o_P(1)$ for $j \in \{0, 1\}$. Since $\sqrt{n}(\hat{\tau} - \tau) \to^d N(0, V_*)$, $\hat{\tau} - \tau = o_P(1)$. Therefore, $n^{-1}\sum_{i=1}^{n} u_i^2 = o_P(1)$. It follows by (50) that $\hat{V}_3 = \tilde{V}_3 + o_P(1)$. The proof is therefore complete. □

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.

Alexandre Belloni, Victor Chernozhukov, et al. L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34 (4):606–619, 2016.

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

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)*, 2015.

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, 2018a.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*, 2018b.

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

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

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

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2018.

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.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The 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.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8895–8906, 2018.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Yang Ning, Sida Peng, and Kosuke Imai. Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv preprint arXiv:1812.08683*, 2018.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(1):122–129, 1995.

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.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Andrea Rotnitzky, Ezequiel Smucler, and James M. Robins. Characterization of parameters with a mixed bias property. *arXiv:1904.03725*, 2019.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

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

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

Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *arXiv preprint arXiv:1710.08074*, 2017.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, Forthcoming, 2019+.

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 J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):1–40, 2006.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012a. ISBN 9780511794308. Cambridge Books Online.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012b.

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.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965–993, 2019.

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

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

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.