UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Robust and Efficient Methods in Semi-supervised Inference and Causal Inference

Permalink

https://escholarship.org/uc/item/2841f4kd

Author

Zhang, Yuqian

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Robust and Efficient Methods in Semi-supervised Inference and Causal Inference

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Yuqian Zhang

Committee in charge:

Professor Jelena Bradic, Chair
Professor Sanjoy Dasgupta
Professor Ronghui Xu
Professor Danna Zhang
Professor Wenxin Zhou

2022

Copyright
Yuqian Zhang, 2022
All rights reserved.

The dissertation of Yuqian Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to express my greatest appreciation to my advisor Professor Jelena Bradic. Her passion and curiosity for leading-edge research are extremely impressive and inspiring. I am very fortunate and pleasant to be guided by her in my Ph.D. study. She never hesitates to offer help to me, and her advice is always helpful and constructive. It is very joyful to work with her, and she is very considerate and supportive, especially during the tough Covid-19 times. Under her guidance, I have not only learned state-of-the-art statistical research but also acknowledged how to survive and succeed in my academic career. I really appreciate such a valuable and memorable experience with her.

Next, I hope to thank Professor Abhishek Chakrabortty. I am delighted to work with him, and I have learned a lot from his generous suggestions. I am also thankful to Professor Lily Xu, Anru Zhang, and Wenxin Zhou for their kindly help and support on my job applications. I would also express my appreciation to Professor Sanjoy Dasgupta and Danna Zhang for their support and serving on my thesis committee.

It is never easy to live in a foreign country. Thankfully, I have been surrounded by many cute friends. I want to thank my roommates, Jiaqi Chen, Xindong Tang, Jiang Wang, Zi Yang, who made my Ph.D. life full of happiness. I would like to thank my other colleagues and friends in UCSD, in particular, Mingjie Chen, Wentao Deng, Toni Gui, Bingni Guo, Weijie Ji, Mengying Lan, Woonam Lim, Tuo Lin, Zhiling Liu, Chunyi Lv, Jiajie Shi, Yening Shu, Zian Wang, Yujie Xu, Xuyu Zhang, Muhan Zhao, etc. for their accompany. Also, I wish to thank my friends outside USCD, including Chenyang Duan, Ziheng Liao, Jibiao Shen, Tianchen Song, Xueqian Wu, Ying Zhou etc. I have enjoyed my time with

them.

Lastly, I want to give my deepest gratitude to my parents, Yunyan Xianyu and Zhixiong Zhang, for their endless support and love. I am incredibly proud of my mother, Yunyan Xianyu, for her bravery and dedication in facing Covid-19 in the front line.

In this dissertation, some materials have been published, or been submitted for publication.

Chaper 1, in full, is a reprint of the material as it appears in Biometrika. Zhang, Yuqian; Bradic, Jelena. High-dimensional semi-supervised learning: in search of optimal inference of the mean, Biometrika, asab042, 2021. The dissertation author was the primary investigator and author of this paper.

Chaper 2, in full, has been submitted for publication of the material. Zhang, Yuqian; Chakrabortty, Abhishek; Bradic, Jelena. Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. The dissertation author was one of the primary investigators and authors of this material.

Chaper 3, in full, has been submitted for publication of the material. Bradic, Jelena; Ji, Weijie; Zhang, Yuqian. High-dimensional inference for dynamic treatment effects. The dissertation author was one of the primary investigators and authors of this material.

Chaper 4, in full, is currently being prepared for submission for publication of the material. Zhang, Yuqian; Bradic, Jelena; Ji, Weijie. Dynamic treatment effects: high-dimensional inference under model misspecification. The dissertation author was the primary investigator and author of this material.

VITA

2016	B. S. in Mathematics and Applied Mathematics, Wuhan University
2022	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

Y. Zhang, J. Bradic and W. Ji, "Dynamic treatment effects: high-dimensional inference under model misspecification", *Preprint*, 2021. `arXiv:2111.06818`.

J. Bradic, W. Ji, Y. Zhang, "High-dimensional inference for dynamic treatment effects", *Preprint*, 2021. `arXiv:2110.04924`.

Y. Zhang, A. Chakrabortty, J. Bradic, "Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap", *Preprint*, 2021. `arXiv:2104.06667`.

Y. Zhang and J. Bradic, "High-dimensional semi-supervised learning: in search of optimal inference of the mean", *Biometrika*, 2021. asab042.

ABSTRACT OF THE DISSERTATION

Robust and Efficient Methods in Semi-supervised Inference and Causal Inference

by

Yuqian Zhang

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2022

Professor Jelena Bradic, Chair

We consider the mean response estimation and inference in semi-supervised settings in the first two chapters. Such settings consist of a relatively small labeled dataset and an extensive unlabeled dataset. Chapter 1 considers the classical semi-supervised setup that the outcome is missing completely at random (MCAR). Our goal is to improve the efficiency of the supervised sample mean estimator using the additional unlabeled data. We proposed a semi-supervised mean estimator based on flexible working models, including high-dimensional and non-parametric models. In Chapter 2, we further consider the situation

that a selection bias may appear. Our goal is to remove the bias originating from the dependence between the missing and outcome. We propose a semi-supervised doubly robust mean estimator with valid inference results when some product rate condition holds. Our work fills in the gap between the semi-supervised literature and the missing data literature. We allow selection bias – this extends the semi-supervised literature. We also allow extremely unbalanced labeled/unlabeled groups and violate the usual positivity condition, which is always assumed throughout the missing data literature.

The last two chapters consider the estimation and inference of the dynamic treatment effect (DTE) when the treatment variable is longitudinal and the covariates are possibly high dimensional. Chapter 3 proposes a doubly robust DTE estimator based on (imputed) Lasso-type nuisance estimators. We established root-n inference when all the nuisance models are correctly specified and some sparsity conditions hold. Chapter 4 further provides root-n inference for the DTE even when model misspecification occurs. This is achieved based on special "moment targeting" nuisance estimators. We provide valid inference as long as one of the nuisance models is correctly specified at each time spot – such a result is better than all the existing literature, even containing the low-dimensional works.

Chapter 1

High-dimensional semi-supervised learning: in search of optimal inference of the mean

1.1 Introduction

We consider a semi-supervised setting with n independent and identically distributed pairs $(X_i, Y_i)_{i=1}^n \sim P_{(X,Y)}$ of observations, with covariates $X_i \in \mathbb{R}^{p-1}$ and the outcome $Y_i \in \mathbb{R}$. We presuppose the existence of an additional set of m observations, $(X_i)_{i=n+1}^{n+m}$. With $\tau = \lim_{m,n\to\infty} n/(m+n) \in [0,1]$ denoting the ratio of the fully observed data and data with the missing outcomes, we are particularly focused on the case of $\tau = 0$, i.e., $m \gg n$. The semi-supervised learning setting can be viewed as a particular missing data setting, where the outcome is missing completely-at-random. Although the missing data literature, in general, addresses a more general setting of the outcomes missing-at-random [SRR99],

semi-supervised learning has a particular caveat that the missing data's size is enormous, $m \gg n$. With $m \gg n$, typical missing-at-random approaches [BR05] no longer apply. The positivity/overlap condition, see, e.g., [RLSR12], is no longer satisfied; with $\tau = 0$, the probability of observing the outcome converges to zero, therefore implying that the semi-supervised setting is not a simple subset of the missing-at-random setting. Instead, we treat the missingness size, an impediment from the missing-at-random perspective, as a semi-supervised strength. In the case of infinite missingness of the response, we are left with infinite additional information regarding the covariates' distribution, P_X. Mimicking the known P_X setting, we remove the bias in estimating the outcome model and show that semi-supervised-double-robust inference is achievable.

Our main contribution is in constructing new semi-supervised estimates of $\theta = E(Y)$ and in providing root-n inferential guarantees while allowing for misspecification of the distribution of $Y \mid X$. An impediment to providing optimal inferences about θ lies in the inability to estimate $E(Y \mid X)$ with root-n guarantees. Sparse regularizers, random forests, nonparametric (smoothing) estimators, or neural networks do not admit root-n consistency. While there is vast literature on semi-supervised learning, comparatively little is known about making inferences about θ; see [Zhu05]. Recent results of [WL08, EACR$^+$16, MC18] consider the class of low-dimensional graph-oriented semi-supervised algorithms. Semi-supervised learning in the context of classification has had a long tradition; see [CSZ09, GB05]. A small but growing literature has considered the development of semi-supervised inferential procedures. The recent work of [ZBC19] is a special case of our construction. Authors utilize the least-squares approach in linear models whenever $p = o(n^{1/2})$. Our results are based on $n^{-1} \log(p) = o(1)$ together with many possible estimators, e.g., random forests

and neural networks. [CC18] develop the semi-supervised regression method with improved efficiency when the linear model is misspecified. [GC18] consider semi-supervised prediction, while [CG20] propose semi-supervised explained variance estimates. We, therefore, view our contribution as complementary to this growing literature.

We believe that our new estimating tools will be useful beyond the specific class of environments studied here. We illustrate this point by applying our findings to heterogeneous treatment effects. Existing approaches of [CCD$^+$18, KSBY19]; and [CCD$^+$17] build learners that can conform to many machine learning methods [WA18, AIW18]. However, they do not consider the semi-supervised setting with the outcome and the treatment missing. We discover that the asymptotic variance size is reduced regardless of whether additional information on the treatment is available. Moreover, treatment assignment can potentially depend on all covariates with no explicit sparsity requirement. The method also shares the low-dimensional asymptotic efficiency of [CAC18].

1.2 Efficient estimation of the mean

1.2.1 From de-biasing to double-robustness

Let $\beta^* \in \mathbb{R}^p$, the population slope, be an l_2 projection defined as

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^p} E\left(Y - \beta_1 - X^{\mathbb{T}}\beta_{-1}\right)^2.$$

Here, β_{-j} denotes β with the j-th coordinate removed. For $\varepsilon = Y - \beta_1^* - X^{\mathbb{T}}\beta_{-1}^*$ and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon)$ with $E(\varepsilon \mid X) \neq 0$ we do not necessarily assume that the regression model is linear. With μ and C, denoting the mean and the covariance of X_i, respectively, we

use $V_i = X_i - \mu$, and $Z_i = C^{-1/2}(X_i - \mu)$. With $\tilde{X}_i = (1, X_i^{\mathbb{T}})^{\mathbb{T}}$ and $\tilde{V}_i = (1, V_i^{\mathbb{T}})^{\mathbb{T}}$, let $\tilde{\mu} = (1, \mu^{\mathbb{T}})^{\mathbb{T}}$ and $\tilde{C} = \mathrm{cov}(\tilde{X})$ denote the mean and covariance of $\tilde{X} = (1, X^{\mathbb{T}})^{\mathbb{T}}$. The mean of the response, $\theta = E(Y)$, can be seen as a linear contrast of β^*:

$$\theta = \tilde{\mu}^{\mathbb{T}} \beta^*.$$

When $p \gg n$, a good candidate estimate of β^*, is a regularized estimator, $\hat{\beta}$, e.g., Lasso [Tib97] or square-root Lasso [BCW11]. However, such estimators suffer from slower than root-n consistency: when the outcome model is linear, $\|\hat{\beta} - \beta^*\|_2^2 = o_P\{s \log(p)/n\}$ with $s = |\{j : \beta_j^* \neq 0\}|$. Hence, a plug-in estimate will not achieve root-n inference regarding θ, even if the outcome model is correct, unless s is a constant. Existing literature provides easy solutions with many possible ways to remove the bias of regularization. Each of these could potentially achieve root-n inference of θ but would, however, require strong assumptions on the models: the outcome must be well specified as well as sparse enough. For example, let $\hat{\beta}_{\mathrm{db}} = \hat{\beta} + n^{-1} \sum_{i=1}^n \hat{\Theta} \tilde{X}_i (Y_i - \tilde{X}_i^{\mathbb{T}} \hat{\beta})$, denote the de-biased Lasso [VdGBRD14]. Here, $\hat{\Theta}$, is a candidate estimate of $\tilde{\Sigma}^{-1}$, $\tilde{\Sigma} = E\tilde{X}\tilde{X}^{\mathbb{T}} \in \mathbb{R}^{p \times p}$. Root-$n$ inference of θ would then require outcome sparsity $s = o\{n^{1/2}/\log(p)\}$ as well as $|\{k \neq j : (\tilde{\Sigma}^{-1})_{j,k} \neq 0\}| = o\{n/\log(p)\}$ [VdGBRD14].

However, $\hat{\beta}_{\mathrm{db}}$ does not directly use the additional covariate information available in the semi-supervised setting. Let us consider a particular case where P_X, and with it, $\tilde{\Sigma}^{-1}$ and $\tilde{\mu}$ are known. In this case, we could use an improved de-biased semi-supervised estimator $\tilde{\beta} = \hat{\beta} + n^{-1} \sum_{i=1}^n \tilde{\Sigma}^{-1} \tilde{X}_i (Y_i - \tilde{X}_i^{\mathbb{T}} \hat{\beta})$, which then leads to

$$\tilde{\mu}^{\mathbb{T}} \tilde{\beta} = \tilde{\mu}^{\mathbb{T}} \hat{\beta} + n^{-1} \sum_{i=1}^n e_1^{\mathbb{T}} \tilde{X}_i^{\mathbb{T}} (Y_i - \tilde{X}_i \hat{\beta}) = \tilde{\mu}^{\mathbb{T}} \hat{\beta} + n^{-1} \sum_{i=1}^n (Y_i - \tilde{X}_i^{\mathbb{T}} \hat{\beta}),$$

where $e_1 = (1, 0, 0, \ldots, 0)^{\mathbb{T}}$. Interestingly, by algebraic manipulation, it is not difficult to see that the right-hand side above becomes $\bar{Y} + (\tilde{\mu} - \bar{X})^{\mathbb{T}}\hat{\beta}$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ and $\bar{X} = n^{-1}\sum_{i=1}^{n} \tilde{X}_i$, therefore matching with the low-dimensional estimator of [ZBC19]. There seems to be an intricate connection between the above estimator and the double-robust, missing-at-random estimators of [BR05]. However, there is an important difference. If $T_j = 1$ for $j = 1, \ldots, n$ and zero otherwise, i.e., T is the indicator of the observed data. Missing-at-random treats T as a random variable whereas semi-supervised learning treats T as fixed, non-random. Semi-supervised learning can be viewed as missing-at-random conditional on $(T_i)_{i=1}^{m+n}$ being fixed. Then, the missing-at-random average treatment effect of the treated matches the above estimator

$$\tilde{\mu}^{\mathbb{T}}\hat{\beta} + (n+m)^{-1}\sum_{i=1}^{n+m} T_i(Y_i - X_i\hat{\beta})/P(T_i = 1 \mid X_i),$$

where $P(T_i = 1 \mid X_i) = P(T_i = 1) = n/(n+m)$. However, missing-at-random double-robust estimates require $P(T_i = 1 \mid X_i) > 0$ whereas in the semi-supervised setting we have $P(T_i = 1 \mid X_i) \to 0$ with $m \gg n$.

In the semi-supervised setting, we aim to show that the above estimator's sample equivalent will suffice for root-n inference on θ. Let

$$\tilde{\theta} = \hat{\mu}^{\mathbb{T}}\hat{\beta} + n^{-1}\sum_{i=1}^{n}(Y_i - \tilde{X}_i^{\mathbb{T}}\hat{\beta}), \qquad \hat{\mu} = (n+m)^{-1}\sum_{i=1}^{n}\tilde{X}_i.$$

Our estimator will use cross-fitting, which plays a crucial role in establishing the double-robust property of the proposed estimator, i.e., in controlling the term t_2 in the decomposition

$$\tilde{\theta} - \theta = t_1 + t_2 + t_3,$$

where $t_1 = \theta - n^{-1}\sum_{i=1}^{n} Y_i$, $\quad t_2 = (n^{-1}\sum_{i=1}^{n}\tilde{X}_i - \hat{\mu})^{\mathbb{T}}(\hat{\beta} - \beta^*)$, $\quad t_3 = (n^{-1}\sum_{i=1}^{n}\tilde{X}_i - \hat{\mu})^{\mathbb{T}}\beta^*$.

The cross-fitting technique helps in removing the bias arising from t_2. With the use of cross-fitting, $\hat{\beta}$'s and X_i's influences in t_2 are separated and tight control of t_2 is achieved under minimal conditions. Without cross-fitting, $|\tilde{\theta}-\theta| \leq \|n^{-1}\sum_{i=1}^{n}\tilde{X}_i-\hat{\mu}\|_\infty\|\hat{\beta}-\beta^*\|_1$ where the right-hand side is $O_P(n^{-1/2})$ as long as $s \leq n^{1/2}/\log(p)$. Instead, with the use of cross-fitting, we can guarantee root-n consistency as long as $s \leq n/\log(p)$. Cross-fitting can be traced back to the natural ideas of cross-validation. Historical background is provided by [Sto74] and [Gei75] for example. More recently, [RWG19] show that sample splitting increases the accuracy and robustness of inference. [CCD$^+$17] use cross-fitting to define double-robust missing-at-random estimates.

We start by splitting the labeled observations into K sets, I_k, each of size N, and split the unlabeled observations into sets I'_k. Let $J_k = I_k \cup I'_k$ with $|J_k| = M$. Let $\hat{\beta}^{(-k)}$ denote an estimate of β^* computed on all but the kth labeled observations, $\hat{\beta}^{(-k)} = \hat{\beta}(\{(\tilde{X}_i, Y_i) : i \in \{1, 2, \ldots, n\} \setminus I_k\}) \in \mathbb{R}^p$. Then, we propose

$$\hat{\theta}^{(k)} = \hat{\mu}^{(k)^\mathbb{T}}\hat{\beta}^{(-k)} + N^{-1}\sum_{i\in I_k}\left(Y_i - \tilde{X}_i^\mathbb{T}\hat{\beta}^{(-k)}\right), \qquad \hat{\mu}^{(k)} = M^{-1}\sum_{i\in J_k}\tilde{X}_i. \tag{1.1}$$

Finally, we propose the following semi-supervised estimator, which aggregates the above estimates:

$$\hat{\theta} = K^{-1}\sum_{k=1}^{K}\hat{\theta}^{(k)}.$$

We will show that this estimator becomes an unbiased estimator of θ, even in finite samples.

1.2.2 From the mean to the coefficient of determination

A crucial statistical problem is the estimation of the Proportion of Variance Explained (PVE),

$$\text{PVE} = \text{var}(\tilde{X}^{\mathbb{T}}\beta^*)/\sigma_Y^2.$$

Estimation of PVE with $p \gg n$ is difficult due to the numerous overfitting issues. In this section, we propose a semi-supervised coefficient of determination, R^2, an estimator of PVE. The estimation of the explained variance, $b^2 = \text{var}(\tilde{X}^{\mathbb{T}}\beta^*)$, [CG20] can be done with the cross-fitted residuals

$$\hat{b}^{2^{(k)}} = \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)} + 2N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\hat{\varepsilon}_i, \quad \hat{\varepsilon}_i = Y_i - \hat{\theta} - \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i, \tag{1.2}$$

and $\hat{b} = K^{-1}\sum_{k=1}^{K}\hat{b}^{2^{(k)}}$, where the estimates of $\tilde{V}_i$ are $\hat{V}_i = \tilde{X}_i - \hat{\mu}^{(k)}$ and their covariance $\hat{C}^{(k)} = M^{-1}\sum_{i\in J_k}\hat{V}_i\hat{V}_i^{\mathbb{T}}$. The motivation behind this careful construction is governed by bias-propagation in the high-dimensional setting; as we will show, the residuals as defined above are, however, root-n consistent. This, in turn, provides a more stable estimate and enables theoretically weak conditions. To see that the naive estimate $Y_i - \hat{\beta}^{\mathbb{T}}\tilde{X}_i$ may not guarantee root-n consistency, we only need to observe that in such a case, $Y_i - \hat{\beta}^{\mathbb{T}}\tilde{X}_i = \varepsilon_i + (\theta - \hat{\beta}^{\mathbb{T}}\tilde{\mu}) - (\hat{\beta} - \beta^*)^{\mathbb{T}}(\tilde{X}_i - \tilde{\mu})$, while the term $\theta - \hat{\beta}^{\mathbb{T}}\tilde{\mu}$ is not necessarily root-n consistent whenever $p \gg n$. Our cross-fitted construction can be seen as a bias-corrected estimate of the residuals. We propose a new estimator of the variance of the response, $\sigma_Y^2 = \text{var}(Y)$,

$$\hat{\sigma}_Y^{2^{(k)}} = N^{-1}\sum_{i\in I_k}(Y_i - \hat{\theta})^2 + N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\left(\hat{C}^{(k)} - \hat{V}_i\hat{V}_i^{\mathbb{T}}\right)\hat{\beta}^{(-k)}, \tag{1.3}$$

and with it $\hat{\sigma}_Y^2 = K^{-1}\sum_{k=1}^{K}\hat{\sigma}_Y^{2^{(k)}}$. Our results also hold for the truncated version $\hat{\sigma}_{Y,\text{trunc}}^2 = \max(\hat{\sigma}_Y^2, 0)$. A classical estimate, the simple sample variance, $S_Y^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$,

does not utilize any additional knowledge of the covariates. Alternatively, one may consider $n^{-1}\sum_{i=1}^{n}(Y_i-\hat{\theta})^2$. However, both of these estimates can be improved. Our theoretical results demonstrate a persistent variance magnification,

$$n^{-1}\sum_{i=1}^{n}(Y_i-\hat{\theta})^2=\sigma_Y^2+n^{-1}\sum_{i=1}^{n}\{\beta^{*\mathbb{T}}(\tilde{V}_i\tilde{V}_i^{\mathbb{T}}-\tilde{C})\beta^*\}+T+O_P(n^{-1}),$$

where $E(T)=0$ and $T=n^{-1}\sum_{i=1}^{n}(\,2\beta^{*\mathbb{T}}\tilde{V}_i\varepsilon_i+\varepsilon_i^2-\sigma_\varepsilon^2)$. Hence, our estimator adds a correction term so that the contribution of the middle term disappears. Therefore, R^2 can be obtained by plugging in the estimators of b^2, (1.2), and the variance of the response σ_Y^2,(1.3),

$$R^2=K^{-1}\sum_{k=1}^{K}\hat{b}^{2^{(k)}}/\hat{\sigma}_Y^{2^{(k)}}. \tag{1.4}$$

1.2.3 Root-n consistency

We establish the root-n consistency of the proposed semi-supervised estimators. Constants in what follows, possibly changing from line to line, are independent of the sample size.

Assumption 1.1. *Let the covariance matrix C be such that $\lambda_{\min}(C)>0$ and $\lambda_{\max}(C)\leq c_1$ and $\sup_{\|a\|_2=1}E|a^{\mathbb{T}}Z|^{2+c}<c_1$ as well as $E|Y|^{2+c}<c_1$, for positive constants $c,c_1>0$.*

Assumption 1.2. *The responses are such that $E|Y|^{4+c}<c_1$ whereas the covariance matrix C satisfies, $\lambda_{\min}(C)>0$ and $\lambda_{\max}(C)\leq c_1$ and $\sup_{\|a\|_2=1}E|a^{\mathbb{T}}Z|^{4+c}<c_1$ for positive constants $c,c_1>0$.*

Assumption 1.3. *$\hat{\beta}$ is an estimator for β^* that satisfies $\|\hat{\beta}-\beta^*\|_2=O_P(1)$, as $n,p\to\infty$.*

Condition 1.1 or 1.2, used one at a time, provide a well-defined linear approximation model β^*. A bounded variance of Y simplifies exposition; all of the results still hold even if this condition is removed. However, the results would be less interpretable. Condition 1.3 allows for a wide variety of estimates of β^*: Lasso, Dantzig, Square-root Lasso, Elastic-net [ZH05] or Slope [BVDBS$^+$15] are plausible. Similarly, different structural forms of β^* are permissible; a considerably weaker form of sparsity, l_r sparsity with $r \in (0, 1)$, would be effective as long as $\|\beta^*\|_r^r = o[\{n/\log(p)\}^{1-r/2}]$ [YZ10], for example. As per Conditions 1.1 and 1.2, bounded $2+c$ and $4+c$ moments allow heavy-tailed distributions for the covariates as well as the noise; see, e.g., the Huber estimate of [SZF20].

Theorem 1.1. *Let Conditions 1.1 and 1.3 hold. Then, as* $m, n, p \to \infty$, $\hat{\theta} - \theta = O_P(n^{-1/2})$. *Moreover, if Condition 1.2 hold as well,* $\hat{\sigma}_Y^2 - \sigma_Y^2 = O_P(n^{-1/2})$.

Regarding $\hat{\theta}$, Condition 1.1 can be relaxed to bounded $1 + c$ moments. Importantly, we do not rely on a strong signal-to-noise ratio to achieve root-n consistency. If $s = p$, one can show that the Lasso estimate equals zero with high-probability, in which case the proposed estimate will be the same as the naive $\bar{Y}$. Hence, there is no loss in efficiency, and it seems that the semi-supervised mean estimate is advantageous in almost all cases. We discuss some aspects of the variance in the Section 1.4.

1.2.4 Asymptotic normality

In this section, we proceed to prove that semi-supervised estimates are asymptotically normal and that they improve the efficiency of estimation by borrowing strength from the additional dataset.

Assumption 1.4. *$\hat{\beta}$ is an estimator of β^* that satisfies $\|\hat{\beta} - \beta^*\|_2 = o_P(1)$, as $n, p \to \infty$.*

Theorem 1.2. *Let Conditions 1.1 and 1.4 hold. Then, as $m, n, p \to \infty$,*

$$n^{1/2}(\hat{\theta} - \theta) \to N\left(0, \sigma_\varepsilon^2 + \tau b^2\right), \tag{1.5}$$

in distribution, provided that $\sigma_\varepsilon^2 + \tau b^2 > c$ for some constant $c > 0$.

Compared with requirements for inference in high-dimensional linear models, Conditions 1.1 and 1.4 are milder. Where we require only moderately sparse regimes $s = o(n/\log p)$, high-dimensional and even doubly-robust methods require more strict settings; see, e.g., [SRR19, Tan20b, Tan20a, BWZ19]. In particular, we do not require any sparsity structure on Σ^{-1}, a condition that has been typically assumed throughout the literature, if the variance is unknown. Lastly, we do not require homogeneity of the errors, ε.

Regarding efficiency, observe that

$$\mathrm{var}(n^{1/2}\bar{Y}) = \sigma_Y^2 = \sigma_\varepsilon^2 + b^2 \geq \sigma_\varepsilon^2 + \tau b^2.$$

where $\sigma_\varepsilon^2 + \tau b^2$ is the asymptotic variance of $\hat{\theta}$ as in (1.5). Hence, the semi-supervised estimator $\hat{\theta}$ is asymptotically at least as accurate as $\bar{Y}$ and is often more accurate. Namely, the additional unlabeled data reduce the asymptotic variance by $(1 - \tau)b^2$. The more unlabeled data we observe, the more accurate the proposed estimator $\hat{\theta}$ becomes. When $\tau = 0$, the asymptotic variance is equivalent to the case of known P_X.

Throughout the chapter, we mainly focus on the case of the signal-to-noise ratio, $\mathrm{snr} = b^2/\sigma_\varepsilon^2$, being bounded away from 0 and ∞. However, observe that the two extremes are not particularly informative. Namely, the case of $\mathrm{snr} = 0$ illustrates that no

estimator can improve the naive $\bar{Y}$. Conversely, the case of snr $= \infty$ and $\tau = 0$, illustrates that semi-supervised estimator can potentially lead to a better than $n^{1/2}$ convergence rate. Set $\rho_j = \mathrm{Corr}(Z_j, Y)$ for each $j \in \{1, 2, \ldots, p-1\}$. Then, $b^2 = {\beta^*_{-1}}^{\mathbb{T}} C \beta^*_{-1} = \{C^{-1}E(VY)\}^{\mathbb{T}} C C^{-1} E(VY) = \sigma_Y^2 \sum_{j=1}^{p-1} \rho_j^2$. If $\tau < 1$ and $\sigma_Y^2 \sum_{j=1}^{p-1} \rho_j^2 > c$ for some $c > 0$, i.e., when at least one of the covariates has positive marginal correlation with the response, $\hat{\theta}$ is asymptotically more accurate than $\bar{Y}$.

Our estimator is also optimal in the following sense. The asymptotic variance in Theorem 1.2 is the same as that of [ZBC19], proved under a low-dimensional setting; see their Theorem 2.4. Moreover, it also achieves the oracle lower bound presented in their Proposition 3.1. The following result presents theoretically valid root-n confidence intervals of θ, while only requiring consistency of $\hat{\beta}$ at an arbitrarily slow rate.

Theorem 1.3. *Let Conditions 1.1 and 1.4 hold. With $\hat{\varepsilon}_i$ defined in* (1.2)*, we define $\hat{\sigma}_\varepsilon^2 = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$. Then, whenever $m, n, p \to \infty$, $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_P(1)$, $\hat{b}^2 = b^2 + o_P(1)$, and a valid confidence intervals about θ, at significance level α, is defined as*

$$\mathrm{CI}(\theta) = \left(\hat{\theta} - z_{1-\alpha/2} \{\hat{\sigma}_\varepsilon^2/n + \hat{b}^2/(m+n)\}^{1/2},\ \hat{\theta} + z_{1-\alpha/2} \{\hat{\sigma}_\varepsilon^2/n + \hat{b}^2/(m+n)\}^{1/2} \right), \quad (1.6)$$

with $z_{1-\alpha/2}$ being $(1-\alpha/2)$-quantile of a standard normal distribution.

A few comments are in order. If we are willing to assume Condition 1.2, we show that $\hat{b}^2 - b^2 = O_P(\|\hat{\beta} - \beta^*\|_2^2 + n^{-1/2})$. In contrast, a naive plug-in estimate of b^2, ${\hat{\beta}^{(-k)}}^{\mathbb{T}} \hat{C} \hat{\beta}^{(-k)}$ would only guarantee $O_P(\|\hat{\beta} - \beta^*\|_2)$. Therefore, our result on $\hat{b}^2$ can be seen as complementary to [CG20]. We provide the same convergence rate, whenever $b^2 > c$, $c > 0$, however, with weaker assumptions: we allow heavy-tailed X and ε and misspecified linear model. An

asymptotically normal result holds once $\|\hat{\beta}-\beta^*\|_2 = o_P(n^{-1/4})$; the details of the asymptotic theory regarding $\hat{b}$ are contained in Theorem 1.6 under a more general setting.

Next, we discuss the high-dimensional R^2 semi-supervised estimate. We begin by highlighting the asymptotic results on the variance estimate, followed by a simple corollary regarding the asymptotics of R^2.

Theorem 1.4. *Let Conditions 1.2 and 1.4 hold. Then, as* $m, n, p \to \infty$,

$$n^{1/2}(\hat{\sigma}_Y^2 - \sigma_Y^2) \to N\left\{0, \mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau \mathrm{var}(\beta^{*\mathbb{T}}\tilde{V})^2\right\}, \tag{1.7}$$

in distribution, provided that $\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\beta^{*\mathbb{T}}\tilde{V})^2 > c$ *for some constant* $c > 0$. *Moreover, for* $\hat{\sigma}_\nu^2$ *and* $\hat{\sigma}_\xi^2$ *defined in* (1.15) *and* (1.17), *respectively, we have*

$$\hat{\sigma}_\nu^2 + n(m+n)^{-1}\hat{\sigma}_\xi^2 = \mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\beta^{*\mathbb{T}}\tilde{V})^2 + o_P(1). \tag{1.8}$$

A sufficient condition regarding Theorem 1.4 includes $\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon) > 0$ – whenever $\sigma_\varepsilon^2 > c_1$, and $\mathrm{corr}(\varepsilon^2, \beta^{*\mathbb{T}}\tilde{V}\varepsilon) > -1 + c_2$, for some $c_1, c_2 > 0$, the asymptotic variance in (1.7) is positive. Now we are ready to state the asymptotic normality of R^2 as a simple corollary of a more general result; see Theorem 1.6.

Corollary 1.1. *Let Conditions 1 and 4 hold. Then, for* R^2 *defined in* (1.4), *we have* $R^2 = PVE + o_P(1)$, *whenever* $m, n, p \to \infty$. *Moreover, if Condition 2 holds with* $\|\hat{\beta} - \beta^*\|_2 = o_P(n^{-1/4})$, *then, as* $m, n, p \to \infty$,

$$n^{1/2}V^{-1/2}(R^2)(R^2 - \mathrm{PVE}) \to N(0, 1)$$

in distribution, provided $V(R^2) > 0$, *where*

$$V(R^2) = \mathrm{var}[\sigma_Y^{-4}b^2\varepsilon^2 + \sigma_Y^{-4}\sigma_\varepsilon^2\{2\varepsilon\beta^{*\mathbb{T}}\tilde{V} + \tau(\beta^{*\mathbb{T}}\tilde{V})^2\}] + \tau\sigma_Y^{-8}\sigma_\varepsilon^4\mathrm{var}\{(\beta^{*\mathbb{T}}\tilde{V})^2\}.$$

1.3 Beyond linear outcome models

Recall that our estimation towards the mean depends on the linear projection of $g^0(x) = E(Y \mid X = x)$. A question arises naturally: can we use general machine learning algorithms to estimate $g^0(x)$ and design non-linear projection for optimal estimation of θ? Are we able to construct confidence intervals, and will the asymptotic variances of the estimators be improved? We provide positive answers to both questions.

A natural extension of $\hat{\theta}$ can be defined as

$$\hat{\theta}_{\text{gen}} = K^{-1} \sum_{k=1}^{K} \hat{\theta}_{\text{gen}}^{(k)}, \text{ where } \hat{\theta}_{\text{gen}}^{(k)} = M^{-1} \sum_{i \in J_k} \hat{g}^{(-k)}(X_i) + N^{-1} \sum_{i \in I_k} \left\{ Y_i - \hat{g}^{(-k)}(X_i) \right\}, \quad (1.9)$$

and $\hat{g}^{(-k)}$ is the estimate of g^0 computed on all but the k-th labeled observations. We suppose the existence of some $g^* = g_d^* : \mathbb{R}^p \to \mathbb{R}$, such that $\mu_{2,X}\{\hat{g}^{(-k)}(x) - g^*(X)\} = o_P(1)$ as $n \to \infty$, and possibly $p, q \to \infty$ and where $\mu_r(f) = E\{f - E(f)\}^r$ is the r-th central moment, and $\mu_{r,X}(f) = E_X\{f - E_X(f)\}^r$ with E_X denoting the conditional expectation on the marginal distribution P_X. Here, d denotes the degree of freedom of the working model. Note that $g^*(x) = g^0(x)$ is unnecessary. Here, $g^* = g_d^*$ can be chosen as the projection of the underlying curve $g^0(x)$ to a functional class $\mathcal{G}_d$, i.e.,

$$g^* = \arg\min_{g \in \mathcal{G}_d} E\{g^*(X) - g^0(X)\}^2. \quad (1.10)$$

With a small abuse in notation, let $\varepsilon = Y - g^*(X)$ denote the unexplained error of the model. To better interpret our results, we assume that $E(\varepsilon) = 0$ and $E\{\varepsilon g^*(X)\} = 0$, which is satisfied once $b + ag \in \mathcal{G}_d$ for all $a, b \in \mathbb{R}$ and $g \in \mathcal{G}_d$. We demonstrate in Theorem 1.5 that, $\hat{\theta}_{\text{gen}}$ of (1.9), is asymptotically normal with asymptotic variance

$$V_{\text{gen}}(\theta) = \sigma_{\varepsilon,\text{gen}}^2 + \tau b_{\text{gen}}^2,$$

where $b^2_{\text{gen}} = \text{var}\{g^*(X)\}$ denotes the explained variance of the model g, and $\sigma^2_{\varepsilon,\text{gen}} = E\{Y - g^*(X)\}^2 = \text{var}(Y) - b^2_{\text{gen}}$ denotes the unexplained variance. When g^* is defined as in (1.10), b^2_{gen} and $\sigma^2_{\varepsilon,\text{gen}}$ are the largest explained variance and smallest unexplained variance among the functional class $\mathcal{G}_d$, respectively. The unexplained variance can be estimated using a cross-fitting scheme

$$\hat{\sigma}^2_{\varepsilon,\text{gen}} = n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \left\{ Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i) \right\}^2, \tag{1.11}$$

with $\hat{h}^{(-k)}(X_i) = \hat{g}^{(-k)}(X_i) - M^{-1} \sum_{i \in J_k} \hat{g}^{(-k)}(X_i)$. As for the explained variance, (1.2) can be generalized through a bias-corrected cross-fitting estimator

$$\hat{b}^2_{\text{gen}} = (m+n)^{-1} \sum_{k=1}^{K} \sum_{i \in J_k} \{\hat{h}^{(-k)}(X_i)\}^2 + 2n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \hat{h}^{(-k)}(X_i)\{Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i)\}. \tag{1.12}$$

Now, $\hat{V}_{\text{gen}}(\theta) = \hat{\sigma}^2_{\varepsilon,\text{gen}} + n\hat{b}^2_{\text{gen}}/(m+n)$ and a α-level confidence interval can be constructed as

$$\text{CI}_{\text{gen}}(\theta) = \left(\hat{\theta}_{\text{gen}} - z_{1-\alpha/2} \left\{ \hat{V}_{\text{gen}}(\theta)/n \right\}^{1/2}, \hat{\theta}_{\text{gen}} + z_{1-\alpha/2} \left\{ \hat{V}_{\text{gen}}(\theta)/n \right\}^{1/2} \right). \tag{1.13}$$

The asymptotic normality of non-linear R^2 is established in Theorem 1.6.

Theorem 1.5. *Suppose that $E|Y|^{2+c} < C$ and $E|g^*(X)|^{2+c} < C$ for some $C < \infty$. Then, as long as $\mu_{2,X}\{\hat{g}^{(-k)}(x) - g^*(X)\} = o_P(1)$ for each k, as $n, p \to \infty$ (or $n, p, d \to \infty$), $\hat{\theta}_{\text{gen}}$ satisfies*

$$n^{1/2} V^{-1/2}_{\text{gen}}(\theta)(\hat{\theta}_{\text{gen}} - \theta) \to N(0,1), \qquad \hat{V}_{\text{gen}}(\theta) = V_{\text{gen}}(\theta) + o_P(1),$$

provided that $V_{\text{gen}}(\theta) > 0$.

The asymptotic variance above depends on the explained variance b^2_{gen}: the larger the explained variance is, the more efficient estimation of θ is. In particular, a worst case of

$b^2_{\text{gen}} = 0$ corresponds to the sample mean estimator. When $g^*(x) = g^0(x)$, the asymptotic variance is optimal; it matches the oracle lower bound of Proposition 3.1 in [ZBC19] and one can see a clear efficiency gain through $b^2_{\text{gen}}(g^*) \leq b^2_{\text{gen}}(g^0)$.

1.4 Further discussion on the variance

1.4.1 Asymptotic inference for the variance

When we are interested in estimating and perhaps constructing confidence intervals regarding the variance of Y, we require the same set of simple assumptions used in obtaining inferential statements regarding the mean of Y. Even when $\hat{\beta}$ is a biased estimate whose bias is bounded asymptotically (but is not diminishing) we are able to guarantee $n^{1/2}$ consistency of the estimate, (1.3). For consistent $\hat{\beta}$, even without specified rate assumptions, we can guarantee more, in distribution,

$$n^{1/2}(\hat{\sigma}^2_Y - \sigma^2_Y) \to N\left\{0, \text{var}\left(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon\right) + \tau\text{var}\left(\beta^{*\mathbb{T}}\tilde{V}\right)^2\right\}.$$

The result above remains correct even when there is a large dependence of ε_i on $\tilde{V}_i$. The result simplifies a lot if both the covariates X_i and the errors ε_i have Gaussian distribution; in that case, the asymptotic variance becomes $2\sigma^4_\varepsilon + 4\sigma^2_\varepsilon b^2 + 2\tau b^4$. Moreover, under the same set of assumptions, we can consistently estimate the asymptotic variance of $\hat{\sigma}^2_Y$. To do so, we estimate the two components of the asymptotic variance separately. Let's focus on estimating $\text{var}(\beta^{*\mathbb{T}}\tilde{V})^2$ first. To that end, we construct consistent estimates of $(\beta^{*\mathbb{T}}\tilde{V}_i)^2 - E(\beta^{*\mathbb{T}}\tilde{V})^2$, $\xi_i^{(k)}$, as follows

$$\xi_i^{(k)} = \hat{\beta}^{(-k)^{\mathbb{T}}}\left(\hat{V}_i\hat{V}_i^{\mathbb{T}} - \hat{C}^{(k)}\right)\hat{\beta}^{(-k)}. \tag{1.14}$$

Then we set

$$\hat{\sigma}_{\xi}^2 = N^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} {\xi_i^{(k)}}^2 \approx \mathrm{var}(\beta^{*\mathbb{T}} \tilde{V})^2. \tag{1.15}$$

Next, we estimate $\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}} \tilde{V} \varepsilon)$. To that end we define

$$\eta_i^{(k)} = \hat{\varepsilon}_i^2 + 2\hat{\beta}^{(-k)^{\mathbb{T}}} \hat{V}_i \hat{\varepsilon}_i + \ \hat{\beta}^{(-k)^{\mathbb{T}}} \hat{C}^{(k)} \hat{\beta}^{(-k)},$$

and observe that $\hat{\sigma}_Y^2$ is an average of $\eta_i^{(k)}$. Then, we create a cross-fitted residuals of the following form

$$\nu_i^{(k)} = \eta_i^{(k)} - \hat{\sigma}_Y^2 \tag{1.16}$$

and show

$$\hat{\sigma}_{\nu}^2 = N^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} {\nu_i^{(k)}}^2 \approx \mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}} \tilde{V} \varepsilon). \tag{1.17}$$

In Theorem 1.4, we showcase that as long as any consistent estimate of $\hat{\beta}$ is used, the confidence interval

$$\mathrm{CI}(\sigma_Y^2) = \left(\hat{\sigma}_Y^2 - z_{1-\alpha/2} \{ \hat{\sigma}_{\nu}^2/n + \hat{\sigma}_{\xi}^2/(m+n) \}^{1/2},\ \hat{\sigma}_Y^2 + z_{1-\alpha/2} \{ \hat{\sigma}_{\nu}^2/n + \hat{\sigma}_{\xi}^2/(m+n) \}^{1/2} \right) \tag{1.18}$$

will be asymptotically correct.

1.4.2 Variance estimation discussion

Based on Theorem 1.4, when the data follows Gaussian distribution, it is not difficult to see that when $\tau \leq 1$, i.e., $m \geq n$

$$\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}} \tilde{V} \varepsilon) + \tau \mathrm{var}(\beta^{*\mathbb{T}} \tilde{V})^2 \leq \mathrm{var}(Y - \theta)^2 = \mathrm{var}(n^{1/2} S_Y^2) + o(1).$$

Namely, the constructed confidence interval for σ_Y^2 as presented in (1.18) is asymptotically more accurate in the sense of having smaller width asymptotically, than the interval that is

solely based on $\{Y_i\}_{i=1}^n$

$$\left(S_Y^2 \Big/ \left\{1 - z_{\alpha/2}(\hat{\gamma} - 1)^{1/2}/n^{1/2}\right\}, S_Y^2 \Big/ \left\{1 + z_{\alpha/2}(\hat{\gamma} - 1)^{1/2}/n^{1/2}\right\} \right),$$

or its robust alternatives (see for example [HBH05]) where $\hat{\gamma}$ is any consistent estimator for the kurtosis. One of the choices for $\hat{\gamma}$ can be

$$\hat{\gamma} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_i \frac{(Y_i - \bar{Y})^4}{S_Y^4} - \frac{3(n-1)^2}{(n-2)(n-3)} + 3.$$

This, in turn, implies that the proposed semi-supervised estimator, $\hat{\sigma}_Y^2$ is asymptotically more accurate than the sample variance, S_Y^2.

In general, the efficiency of the proposed semi-supervised estimator $\hat{\sigma}_Y^2$ would depend on the particular model of non-linearity, i.e., on the particular deviations from the linear model. We illustrate the discussion with two specific examples. To that end, we introduce a proportionality coefficient r as the proportion of the decrease achieved by the semi-supervised estimator compared to S_Y^2. We define such coefficient with

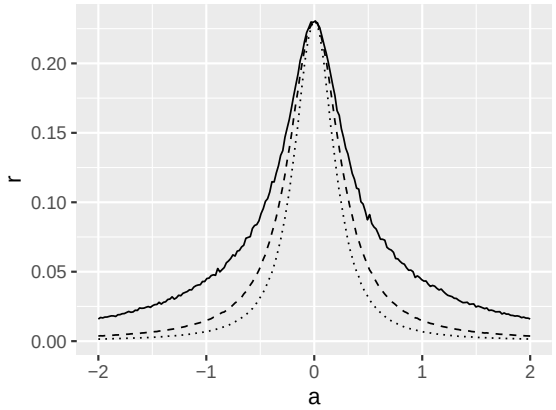
$$r = \frac{\text{var}(Y - \theta)^2 - \text{var}(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon)}{\text{var}(Y - \theta)^2}. \tag{1.19}$$

Here, we have assumed that $m \gg n$ and the effect of τ is negligible.

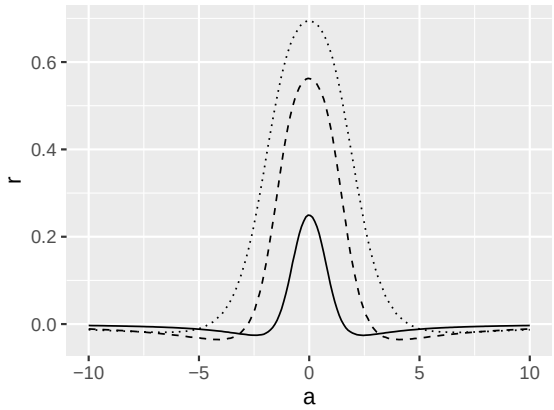
The first example discusses a heteroscedastic linear model where the variance of the error depends quadratically on the covariates. The second discusses larger deviations from normality, where the response model is highly non-linear. In particular, we consider

$$Y_i = \sum_{j=1}^p X_{ij} + \left(a \sum_{j=1}^p X_{ij}^2 + \sum_{j=1}^p X_{ij} \right) \eta_i \qquad \text{(Example 1)}$$

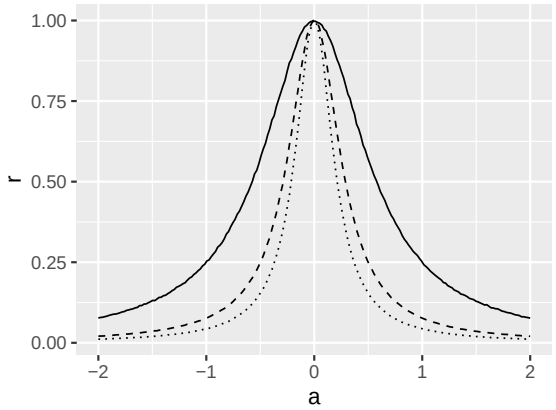
$$Y_i = a \left| \log\Big(0.8 \left| \sum_{j=1}^p X_{ij} \right| + 0.01\Big) \right| + \sum_{j=1}^p X_{ij} + \eta_i \qquad \text{(Example 2)}$$



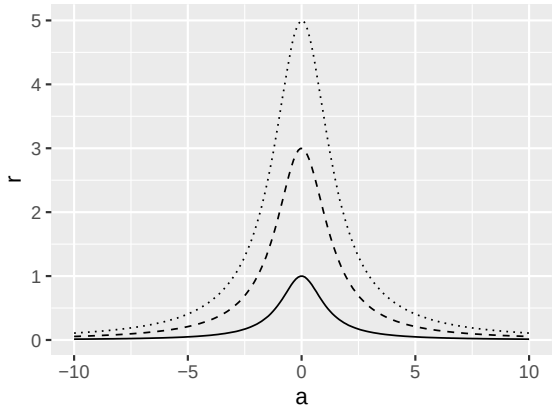
(a) The proportion of decrease versus the size of the heteroscedasticity of the linear model



(b) The proportion of decrease versus the size of the non-linearity of Example 2



(c) Signal to noise ratio versus the size of the heteroscedasticity of the linear model



(d) Signal to noise ratio versus the size of the non-linearity of Example 2

Figure 1.1: Proportion of the decrease in asymptotic variance achieved by the proposed semi-supervised estimator $\hat{\sigma}^2_Y$ as a function of the coefficient a representing the heteroscedasticity in (a) and (c) and non-linearity in (b) and (d) subfigures. Different colors correspond to different dimensionality settings, in (a) and (c), we have $p = 1$ (solid), $p = 10$ (dashed) and $p = 20$ (dotted); in (b) and (d), we have $p = 1$ (solid), $p = 3$ (dashed) and $p = 5$ (dotted).

where a measures the size of the deviation from the linear model. In the above $X_i, \eta_i \sim N(0, 1)$. When $r > 0$ we see that the proposed estimator is more efficient than S^2_Y.

From Example 1, we observe that efficiency persists over a broad range of heteroscedastic model specifications. We also observe that the larger the magnitude of a is – the more significant the effect of heteroscedasticity is – the smaller the signal is in the lin-

ear model. This, in turn, results in smaller r values. Example 2 is showing a more complex scenario; smaller magnitudes of a indicate not too great deviations from the linear model and result in greater efficiency in $\hat{\sigma}_Y^2$. For larger a, the linear approximation is too far from the data generating process. Results are presented in Figure 1.1, where we also showcase the Signal to Noise ratio corresponding to each setting.

1.4.3 Inference of variance by a general machine learning model

In addition to the confidence interval of mean $E(Y)$, one may also be interested in inference towards the variance $\sigma_Y^2 = \text{var}(Y)$, the explained variance b_{gen}^2 and the unexplained variance $\sigma_{\varepsilon,\text{gen}}^2$.

The general semi-supervised estimators towards $\sigma_{\varepsilon,\text{gen}}^2$ and b_{gen}^2 are proposed in (1.11) and (1.12) when we construct asymptotic confidence intervals of the mean. As for the variance of Y, recall that in our setting, $\sigma_Y^2 = \sigma_{\varepsilon,\text{gen}}^2 + b_{\text{gen}}^2$. Hence, the variance can be estimated by the sum of estimated explained variance and unexplained variance

$$\hat{\sigma}_{Y,\text{gen}}^2 = \hat{\sigma}_{\varepsilon,\text{gen}}^2 + \hat{b}_{\text{gen}}^2. \tag{1.20}$$

The estimation of PVE can also be handled by the usage of a general machine learning model. An extension of R^2 can be defined as

$$R_{\text{gen}}^2 = K^{-1} \sum_{k=1}^{K} \hat{b}_{\text{gen}}^{2^{(k)}} / \hat{\sigma}_{Y,\text{gen}}^{2^{(k)}}.$$

Because of the limited length of the paper, here we only propose the asymptotic normality results of the generalized estimators.

Theorem 1.6. *Suppose that we have n independent and identically distributed samples $(Y_i, X_i) \sim P$ whose marginal distributions are (P_Y, P_X). In addition, suppose that we observe*

a supplementary set of m independent and identically distributed samples X_i that are drawn from the same distribution P_X. Moreover, suppose that $E|Y|^{4+c} < C$, $E|g^*(X)|^{4+c} < C$ and the estimation error satisfies $\mu_{4,X}\{\hat{g}^{(-k)}(X) - g^*(X)\} = o_P(1)$, then

$$\frac{n^{1/2}(\hat{\sigma}^2_{Y,\text{gen}} - \sigma^2_Y)}{\{V(\sigma^2_Y)\}^{1/2}} \to N(0,1),$$

in distribution, where

$$V(\sigma^2_Y) = \text{var}\left\{\varepsilon^2 + 2\varepsilon(g^*(X) - \theta) + \frac{n}{m+n}(g^*(X) - \theta)^2\right\} + \frac{mn}{(m+n)^2}\text{var}\left\{(g^*(X) - \theta)^2\right\}.$$

The asymptotic variance $V(\sigma^2_Y)$ can be estimated by

$$\hat{V}(\sigma^2_Y) = n^{-1}\sum_{k=1}^{K}\sum_{i\in I_k}\left(\nu_i - N^{-1}\sum_{i\in I_k}\nu_i\right)^2 + \frac{m}{(m+n)^2}\sum_{k=1}^{K}\sum_{i\in I_k}\left(\xi_i - N^{-1}\sum_{i\in I_k}\xi_i\right)^2,$$

with $\hat{V}(\sigma^2_Y)/V(\sigma^2_Y) = 1 + o_P(1)$, where

$$\nu_i = \left(Y_i - \hat{\theta}\right)^2 - \frac{m}{m+n}\left\{\hat{g}^{(-k)}(X_i) - M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i)\right\}^2,$$
$$\xi_i = \left\{\hat{g}^{(-k)}(X_i) - M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i)\right\}^2.$$

Moreover, if in addition that $\mu_{2,X}\{\hat{g}^{(-k)}(X) - g^*(X)\} = o_P(n^{-1/2})$ and either a) $g(x)$ is linear on x, or b) $\mu_{2,X}\{\hat{g}^{(-k)}(X) - g^*(X)\}\mu_{2,X}\{g^*(X) - g^0(X)\} = o_P(n^{-1})$, then

$$\frac{n^{1/2}(\hat{\sigma}^2_{\varepsilon,\text{gen}} - \sigma^2_{\varepsilon,\text{gen}})}{\{V(\sigma^2_{\varepsilon,\text{gen}})\}^{1/2}} \to N(0,1), \qquad \frac{n^{1/2}(\hat{b}^2_{\text{gen}} - b^2_{\text{gen}})}{\{V(b^2_{\text{gen}})\}^{1/2}} \to N(0,1), \tag{1.21}$$
$$n^{1/2}V^{-1/2}(R^2_{\text{gen}})(R^2_{\text{gen}} - \text{PVE}) \to N(0,1),$$

in distribution, provided that $V(\sigma^2_{\varepsilon,\text{gen}}) > c$, $V(b^2_{\text{gen}}) > c$ and $V(R^2_{\text{gen}}) > c$, where $V(\sigma^2_{\varepsilon,\text{gen}}) =$

$\mathrm{var}(\varepsilon^2)$ *and*

$$V(b_{\mathrm{gen}}^2) = \mathrm{var}\left[2\varepsilon\{g^*(X)-\theta\} + \frac{n}{m+n}\{g^*(X)-\theta\}^2\right] + \frac{mn}{(m+n)^2}\mathrm{var}\left[\{g^*(X)-\theta\}^2\right],$$

$$\begin{aligned} V(R_{\mathrm{gen}}^2) = \mathrm{var}[&\sigma_Y^{-4} b_{\mathrm{gen}}^2 \varepsilon^2 + \sigma_Y^{-4}\sigma_{\varepsilon,\mathrm{gen}}^2\{2\varepsilon(g^*(X)-\theta) + n(m+n)^{-1}(g^*(X)-\theta)^2\}] \\ &+ n(m+n)^{-1}\sigma_Y^{-8}\sigma_{\varepsilon,\mathrm{gen}}^4 \mathrm{var}\{(g^*(X)-\theta)^2\}. \end{aligned}$$

Based on Theorem 1.6, an asymptotic confidence intervals for σ_Y^2, at significant level α, is proposed as

$$\mathrm{CI}_{\mathrm{gen}}(\sigma_Y^2) = \left(\hat{\sigma}_{Y,\mathrm{gen}}^2 - z_{1-\alpha/2}\left\{\hat{V}(\sigma_Y^2)/n\right\}^{1/2},\ \hat{\sigma}_{Y,\mathrm{gen}}^2 + z_{1-\alpha/2}\left\{\hat{V}(\sigma_Y^2)/n\right\}^{1/2}\right). \tag{1.22}$$

For a general non-linear model, as in Theorem 1.6, we can see that the asymptotic normality results of estimating explained variance and unexplained variance (1.21) require rates on $\hat{g} - g^*$ and $g^* - g^0$. Here we provide some insights on the two rates. As the degree of freedom d grows, the estimation error $\mathrm{err}_1 = \mu_{2,X}\{\hat{g}^{(-k)}(X) - g^*(X)\}$ grows and the misspecification error $\mathrm{err}_2 = \mu_{2,X}\{g^*(X) - g^0(X)\}$ decreases. To obtain (1.21), we need $\mathrm{err}_1 = o_P(n^{-1/2})$ and $\mathrm{err}_1\mathrm{err}_2 = o_P(n^{-1})$, which is weaker than $\mathrm{err}_3 = \mu_{2,X}\{\hat{g}^{(-k)}(X) - g^0(X)\} = o_P(n^{-1/2})$.

Here we take the ReLu network as an example and showcase the conditions when the asymptotic normalities (1.21) hold. Following the settings as in [FLM21], let W and L being the number of parameters and the number of layers, respectively. Assume the conditions in Theorem 2 of [FLM21], $g^0(X) \in \mathcal{W}^{r,\infty}((-1,1)^p) = \{g : \max_{\alpha,|\alpha|\le r} \mathrm{ess\,sup}_{x\in(-1.1)^d} |D^\alpha g(x)| \le 1\}$. For $W \propto n^a$ with any $a > 0$ and $L \propto \log n$, omitting the logarithm terms, we have err_1, err_2 are of the order n^{-a} and $n^{-2ar/p}$ respectively. Hence, the asymptotic normalities (1.21) hold when $a \in (0, 1/2)$ and $p < 2r$. In other words, the degree of freedom d, or W in the

neural network example, is flexible, and we are able to obtain the asymptotic normalities for a wide range of d.

1.5 Data missing-at-random

Now, we turn to missing-at-random setting, where whether we observe Y_i depends on X_i. Suppose that we have $m+n$ independent and identically distributed samples $(T_i, Y_i, X_i) \sim P$, whose marginal distributions are (P_T, P_Y, P_X). Here, $T_i \in \{0,1\}$ denotes the labeling: Y_i is observable if and only if $T_i = 1$. Assume the missing-at-random condition: $Y_i \perp T_i \mid X_i$. Let $Y_i^o = T_i Y_i$. Let $n = \sum_{i=1}^{m+n} T_i$ be the amount of labeled samples. Here, n is a random variable. Estimating the mean of Y is equivalent to the estimation of the average treatment effect of the treated. Let $\hat{g}^{(-k)}(x)$ and $\hat{s}^{(-k)}(x)$ be estimates of $g^0(x) = E(Y \mid X = x)$ and $s^0(x) = E(T \mid X = x)$ computed on all but the observations in k-th fold, respectively. Then,

$$\hat{\theta}_{\text{MAR}} = (m+n)^{-1} \sum_{k=1}^{K} \sum_{i \in J_k} \left[\hat{g}^{(-k)}(X_i) + \frac{T_i\{Y_i^o - \hat{g}^{(-k)}(X_i)\}}{\hat{s}^{(-k)}(X_i)} \right] \tag{1.23}$$

is an estimate of the mean $\theta_{\text{MAR}} = E(Y)$ under the missing-at-random setting. Here, the mean estimator (1.23) is a special case of the double/debiased machine learning estimator of [CCD$^+$18], where they require a positive overlap assumption $P\{s^0(X) > c\} = 1$ for some constant $c > 0$. In our semi-supervised setting, we do allow that $\tau = \lim_{m,n\to\infty} n/(m+n) = 0$, i.e. $s^0(X) = E(T) \to 0$. Hence, it is natural to ask if we can relax the positive overlap to a more general condition on $s^0(x)$, rather than forcing $s^0(x)$ being a constant as in semi-supervised learning? In Theorem 1.7 bellow, we showcase that only $P\{s^0(X) > c_1 E(T)\} = 1$ is needed, and that $E(T) \to 0$ is allowed.

Theorem 1.7. *Suppose that* $E|Y|^{2+c} < c_1$, $E(\varepsilon^2 \mid X) < c_1$ *and* $E(\varepsilon^2) > c_2$ *for some* $c, c_1, c_2 > 0$. *Suppose the expected number of labeled samples grows to infinity, i.e.* $(m+n)E(T) \to \infty$. *Besides, for a given covariate, the ratio of the probability of observing the corresponding response to the overall labeling probability is bounded away from zero, i.e.* $P\{s^0(X) > c_1 E(T)\} = 1$, *for some* $c_1 > 0$. *Suppose* $K < \infty$ *and the estimators of the outcome and the propensity score model,* $\hat{g}$ *and* $\hat{s}$, *have estimation errors satisfying* $E_X\{\hat{g}^{(-k)}(X) - g^0(X)\}^2 = o_P(1)$, $E_X\{1 - s^0(X)/\hat{s}^{(-k)}(X)\}^2 = o_P(1)$, *and*

$$E_X\{\hat{g}^{(-k)}(X) - g^0(X)\}^2 \cdot E_X\{1 - s^0(X)/\hat{s}^{(-k)}(X)\}^2 = o_P[(m+n)^{-1}\{E(T)\}^{-1}]$$

as $m+n, p \to \infty$. *Then, the estimator* $\hat{\theta}_{\text{MAR}}$, (1.23), *is asymptotically normally distributed*

$$(m+n)^{1/2}\left[E\{g^0(X)\}^2 + E\left\{T\varepsilon/s^0(X)\right\}^2\right]^{-1/2}(\hat{\theta}_{\text{MAR}} - \theta) \to N(0,1). \tag{1.24}$$

Moreover, if $E_X[\{\hat{g}(X) - g^0(X)\}^2\{1 - s^0(X)/\hat{s}^{(-k)}(X)\}^2] = o_P(1)$, *then,*

$$\hat{V}_{\text{MAR}} = (m+n)^{-1}\sum_{k=1}^{K}\sum_{i\in J_k}\left[\hat{g}^{(-k)}(X_i) + \frac{T_i\{Y_i^o - \hat{g}^{(-k)}(X_i)\}}{\hat{s}^{(-k)}(X_i)} - \hat{\theta}_{\text{MAR}}\right]^2$$

is consistent, in that as $m+n, p \to \infty$, $\hat{V}_{\text{MAR}}(\theta) = [E\{g^0(X)\}^2 + E\left\{T\varepsilon/s^0(X)\right\}^2]\{1 + o_P(1)\}$.

Hence, an asymptotic $(1-\alpha)$-level confidence interval for the mean θ_{MAR} could be defined as:

$$\left(\hat{\theta}_{\text{MAR}} - z_{1-\alpha/2}\hat{V}_{\text{MAR}}(\theta)(m+n)^{-1/2},\ \hat{\theta}_{\text{MAR}} + z_{1-\alpha/2}\hat{V}_{\text{MAR}}(\theta)(m+n)^{-1/2}\right).$$

Observe that under the assumptions of Theorem 1.7,

$$(m+n)^{-1}[E\{g^0(X)\}^2 + E\{T\varepsilon/s^0(X)\}^2] = O[\{(m+n)E(T)\}^{-1}],$$

which is of the same order as n^{-1}, since $n \sim \text{Binomial}\{m+n, E(T)\}$. That is, the mean estimate $\hat{\theta}_{\text{MAR}}$ is $n^{1/2}$-consistent. Hence, the accuracy depends on the number of labeled

samples rather than the total number of samples. The consistency rate under the missing-at-random setting coincides with [CC18], see Section 2 of their Supplementary Material. Unlike assuming $s^0(X)$ to be known, we consider the consistency rates of $E_X\{1-s^0(X)/\hat{s}^{(-k)}(X)\}^2$. Whenever, $E(T) \to 0$, the rate of $1 - s^0(X)/\hat{s}^{(-k)}(X)$ depends on $(m+n)E(T)$, rather than $m+n$ alone. Hence the estimation error of $\hat{s}$ cannot be simply ignored for a large m. An illustrative example is that of the case of T is independent of X with the empirical mean $\bar{T} = (m+n)\sum_{i=1}^{m+n} T_i$. One can easily check that, for $T_i \sim$ Bernoulli$\{E(T)\}$, we have $1 - E(T)/\bar{T} = O_P[\{(m+n)E(T)\}^{-1/2}]$.

1.6 Heterogeneous treatment effects

Suppose that in addition to previous settings, we have access to a treatment indicator $D_i \in \{0, 1\}$, $i = 1, \cdots, m+n$. Following the potential outcomes framework, [SNDS90,Rub74, Hol88] we then hypothesize the presence of potential outcomes $Y_i(0)$ and $Y_i(1)$ corresponding to, respectively, the response the i-th subject would have experienced with and without the treatment. We then observe that the average treatment effect (ATE)

$$\delta = E\{E(Y \mid X, D = 1) - E(Y \mid X, D = 0)\} = \tau_1 - \tau_0. \tag{1.25}$$

Similarly as in Section 1.2, we hypothesize the existence of the l_2 slopes $\beta_w^* = \min_{\beta \in \mathbb{R}^p} E\{(Y - \tilde{X}^{\mathbb{T}}\beta)^2 \mid D = w\}$, defined at the population level for $w \in \{0, 1\}$. A standard way of constructing the average treatment effects estimates is to posit a model on the treatment assignment and then adjust for possible confounding. Treatments are assigned to subjects according to an underlying scheme that depends on the subjects' features. Their

dependence can be captured by

$$D_i = e(X_i) + \zeta_i, \tag{1.26}$$

where $e(X_i)$ is an unknown propensity score function [**?**]. In the following, we assume two primitive conditions: a widely regarded overlap condition regarding the treatment missingness and an identifiability condition.

Assumption 1.5. *Let* $P\{c \leq e(X) \leq 1-c\} = 1$ *and* $P\{c \leq \hat{e}(X) \leq 1-c\} = 1$ *with some constant* $c \in (0,1)$. *For* $\varepsilon_i = Y_i(D_i) - \{D_i\beta_1^* + (1-D_i)\beta_0^*\}^{\mathbb{T}}\tilde{X}_i$, *let* $E(\zeta \mid X) = 0$, *as well as* $P\{E(\varepsilon^2 \mid X) < C\} = 1$ *with some constant* $C > 0$.

Let $\hat{\beta}_1$, $\hat{\beta}_0$, $\hat{e}$ denote estimators for β_1^*, β_0^*, e, respectively, satisfying $E_{P_X}\{(\hat{\beta}_w^{(-k)} - \beta_w^*)^{\mathbb{T}}\tilde{X}\}^2 = O_P(a_{n,p}^2)$, $E_{P_X}\{\hat{e}^{(-k)}(X) - e(X)\}^2 = O_P(b_{m+n,p}^2)$, and $E[\{E(Y \mid X) - \beta_w^{*\mathbb{T}}\tilde{X}\}^2 \mid D = w] = O_P(c_p^2)$. Here, $a_{n,p}$, $b_{m+n,p}$ and c_p are non-negative sequences of numbers with c_p describing how close the linear model is to the true underlying curve. The semi-supervised estimator (1.1) needs to be adjusted for the confounding effects. To that end, we introduce

$$\hat{\tau}_\omega^{(k)} = \hat{\mu}^{(k)^{\mathbb{T}}}\hat{\beta}_\omega^{(-k)} + N^{-1}\sum_{i\in I_k} w_i^{(-k)}(\omega)\left(Y_i - \tilde{X}_i^{\mathbb{T}}\hat{\beta}_\omega^{(-k)}\right), \qquad \hat{\mu}^{(k)} = M^{-1}\sum_{i\in J_k}\tilde{X}_i.$$

In the above, the weights, $w_i^{(-k)}(\omega)$, correspond to the ratio of the observed treatment proportion; then, the framework from Section 1.2.1 will still lead to root-n consistent estimates. We denote these weights as

$$w_i^{(-k)}(\omega) = \omega D_i/\hat{e}^{(-k)}(X_i) + (1-\omega)(1-D_i)/\{1-\hat{e}^{(-k)}(X_i)\}.$$

Then, the estimate of the average treatment effect can be defined as a difference of $\hat{\delta}^{(k)} = \hat{\tau}_1^{(k)} - \hat{\tau}_0^{(k)}$ and $\hat{\delta} = K^{-1}\sum_{k=1}^K \hat{\delta}^{(k)}$.

An asymptotic $(1-\alpha)$-level confidence interval for the ATE could then be defined as

$$\left(\hat{\delta} - z_{1-\alpha/2}\hat{V}_\delta^{1/2} n^{-1/2},\ \hat{\delta} + z_{1-\alpha/2}\hat{V}_\delta^{1/2} n^{-1/2}\right). \tag{1.27}$$

The estimator of $V_\delta = V_1 + \tau V_2$, (1.30), is defined as $\hat{V}_\delta = K^{-1}\sum_{k=1}^{K}\{\hat{V}_1^{(k)} + n(m+n)^{-1}\hat{V}_2^{(k)}\}$. Observe that $V_1 = \mathrm{var}\{r(Y - \beta_1^{*\mathbb{T}}\tilde{X}) - \rho(Y - \beta_0^{*\mathbb{T}}\tilde{X})\}$. Then, a natural plug-in estimator can be defined as $\hat{V}_1^{(k)} = N^{-1}\sum_{i\in I_k}\nu_{\delta,i}^2$, where

$$\nu_{\delta,i} = r_i^{(-k)}(Y_i - \hat{\beta}_1^{(-k)^{\mathbb{T}}}\tilde{X}_i) - \rho_i^{(-k)}(Y_i - \hat{\beta}_0^{(-k)^{\mathbb{T}}}\tilde{X}_i) - \{\hat{\delta} - (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathbb{T}}\hat{\mu}^{(k)}\},$$

recall $\hat{\mu}^{(k)}$ is defined as (1.1). The second component, $V_2 = E\{(\beta_1^* - \beta_0^*)^{\mathbb{T}}(\tilde{X} - \tilde{\mu})\}^2$, is estimated as $\hat{V}_2^{(k)} = N^{-1}\sum_{i\in I_k}\xi_{\delta,i}^2$ for $\xi_{\delta,i} = (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathbb{T}}(\tilde{X}_i - \hat{\mu})$. The next theorem is the main result of this section.

Theorem 1.8. *Let Conditions 1.1 and 1.5 hold. Then, under the setting of this section*

$$\hat{\delta} - \delta = O_P(n^{-1/2} + a_{n,p}b_{m+n,p} + b_{m+n,p}c_p),$$

whenever $a_{n,p} = O(1)$. *Therefore, whenever* $a_{n,p}b_{m+n,p} = o(1)$ *and* $b_{m+n,p}c_p = o(1)$, $\hat{\delta}$ *is consistent. If however,* $a_{n,p}b_{m+n,p} = O(n^{-1/2})$ *and* $b_{m+n,p}c_p = O(n^{-1/2})$, $\hat{\delta}$ *is a* $n^{1/2}$*-consistent estimate of* δ. *Additionally, the asymptotic normality follows*

$$n^{1/2}(\hat{\delta} - \delta) \to N(0, V_\delta) \tag{1.28}$$

in distribution, whenever,

$$a_{n,p} = o(1), \quad b_{m+n,p} = o(1), \quad a_{n,p}b_{m+n,p} = o(n^{-1/2}), \quad b_{m+n,p}c_p = o(n^{-1/2}), \tag{1.29}$$

with an asymptotic variance

$$V_\delta = \mathrm{var}\left(\varepsilon\zeta/[e(X)\{1-e(X)\}]\right) + \tau(\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{C}(\beta_1^* - \beta_0^*), \tag{1.30}$$

provided that $V_\delta > c$ *for some* $c > 0$, *and* $\tau = \lim_{m,n\to\infty} n/(m+n)$. *Moreover,* $\hat{V}_\delta = V_\delta + o_P(1)$.

Suppose the sparsity of the outcome and the treatment model are s_Y and s_D, respectively. For illustration purposes suppose that both models are parametric and linear. Then, $c_p = 0$, the rates $a_{n,p}$ and $b_{m+n,p}$ for a Lasso estimate, become $a_{n,p} = O_P[\{s_Y \log(p)/n\}^{1/2}]$, $b_{m+n,p} = O_P[\{s_D \log(p)/(m+n)\}^{1/2}]$. Therefore, $s_Y = o\{n/\log(p)\}$, $s_D = o\{(m+n)/\log(p)\}$ and $s_Y s_D = o[(m+n)/\{\log(p)\}^2]$ are required to achieve asymptotic normality. Then, when m is large enough, in that $s_D n \log p/m \to 0$, we require $s_Y = o\{n/\log(p)\}$, which is extremely mild, i.e., consistency in estimation of the propensity model at any arbitrary rate. If both D and Y were unavailable in the unlabeled data, the estimation error on the propensity score would depend on n rather than $m+n$ with the same sparsity assumptions as in [CCD$^+$17, SRR19] and others. At the same time, we achieve a more efficient estimator, regardless of whether D is available in the unlabeled data or not, i.e., reducing the size of the asymptotic variance. When the outcome model is misspecified, even if $c_p = O(1)$ that the linear model does not reach the underlying curve as p grows, we can still obtain the asymptotic normality (1.28) provided m is large enough so that $b_{m+n,p} = o(n^{-1/2})$. Supervised settings have more stringent conditions; see, e.g., [SRR19, Tan20b]. If one is only interested in obtaining a root-n consistency, the outcome model can be completely misspecified, including completely dense high-dimensional models. They can be estimated using machine learning methods, such as random forests, Bayesian classification, regression tree, and deep neural networks; one just needs to replace the linear projection $\tilde{X}_i^{\mathbb{T}} \hat{\beta}^{(-k)}$ by any $\hat{g}^{(-k)}(w, X_i)$,

$$\begin{aligned}\hat{\delta}_{\text{gen}} =& (m+n)^{-1} \sum_{k=1}^{K} \sum_{i \in J_k} \left\{ \hat{g}^{(-k)}(1, X_i) - \hat{g}^{(-k)}(0, X_i) \right\} \\ &+ n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \left[\frac{D_i \{Y_i - \hat{g}^{(-k)}(1, X_i)\}}{\hat{e}^{(-k)}(X_i)} - \frac{(1-D_i)\{Y_i - \hat{g}^{(-k)}(0, X_i)\}}{1 - \hat{e}^{(-k)}(X_i)} \right], \end{aligned} \tag{1.31}$$

where $\hat{g}^{(-k)}(w, X_i)$ is an estimate of $E(Y \mid X, D = w)$ trained on $(D_i, Y_i, X_i)_{i \in \{1,2,\ldots,n\} \setminus I_k}$, for

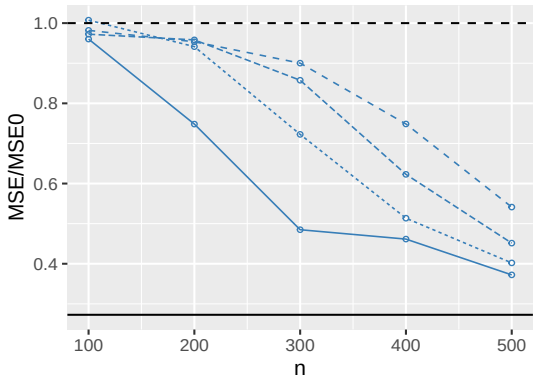
$w \in \{0, 1\}$. Moreover, an asymptotic confidence interval can be extended from (1.27), by replacing the linear outcome model with a general non-linear model.

1.7 Finite-sample experiments

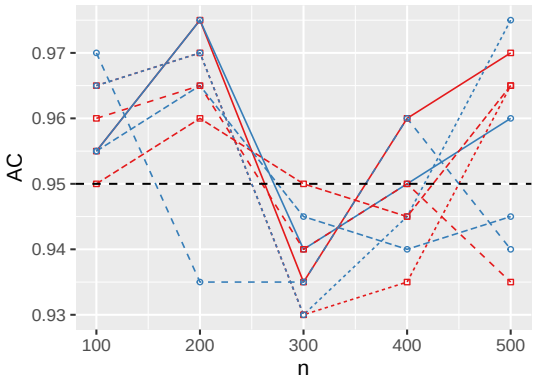
1.7.1 Numerical experiments

In this section, we illustrate the finite-sample properties of $\hat{\theta}$. We consider semi-supervised estimators based on ordinary least squares (SSL-OLS), 10-fold cross-validated lasso (SSL-Lasso), additive model (SSL-Additive), XGBoost (SSL-XGBoost), multilayer perceptron (SSL-MLP), and random forest (SSL-RF) for which vanilla, pre-set tuning parameters are used. We compare with the sample mean $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and with the semi-supervised least squares estimator (SSLS) proposed in [ZBC19], whenever $p < n$. We consider confidence intervals (1.13), where the significance level is $\alpha = 0.05$ throughout. Each set of results is based on 200 repetitions with $K = 5$. The black solid line in all the plots denotes the optimal ratio $\{\sigma_Y^2 - m b_{\text{gen}}^2(g^0)/(m+n)\}/\sigma_Y^2$. We will see that, as long as the sample size n is large enough, our proposed semi-supervised estimators $\hat{\theta}$ is better than the sample mean $\bar{Y}$ in the sense of mean squared error.

Model 5.1. Let $X_i \sim^{\text{iid}} N_{p-1}(0, I_{p-1})$, with $p = 500$, $m = 10n$, $Y_i = s^{-1/2} \sum_{j=1}^{s} X_{ij} + \delta_i$, $s \in \{30, 50, 70, 90\}$, $\delta_i \sim^{\text{iid}} N(0, 0.25)$. Results are presented in Figure 3.1, where we observe that our SSL-Lasso estimator is more efficient than the sample mean, Figure 1.2a, regardless of the level of sparsity. Figure 1.2b illustrates robustness in terms of the average coverage probability of the SSL-Lasso estimate.



(a) The ratio of mean squared errors.

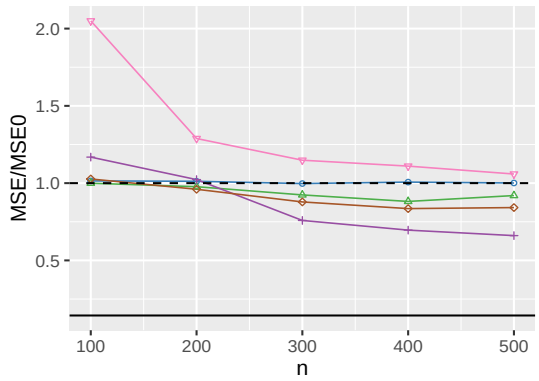


(b) The average coverage of $\bar{Y}$ and $\hat{\theta}$.

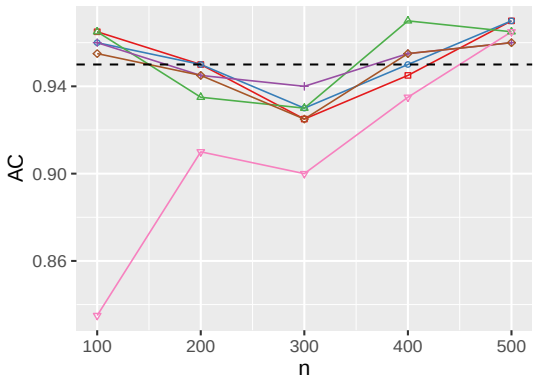
Figure 1.2: Model 5.1: Comparison of SSL-Lasso and the sample mean. The mean squared error of the sample mean is denoted as MSE0. The plot includes sample mean (red squares) and SSL-Lasso (blue circles) estimates. The sparsity level of the linear coefficients, s, is denoted with long dashed, dashed, dotted and solid lines for $s = 90$, $s = 70$, $s = 50$ and $s = 30$, respectively.

Model 5.2. Let X_i and δ_i be as in Model 5.1. and consider a non-linear model $Y_i = 3\cos(X_{i1} + X_{i2} + X_{i3}) + \delta_i$, with $p = 51$, $m = 10n$. We compare our SSL estimator with a variety of baseline procedures and the semi-supervised least squares estimator $\hat{\theta}_{\text{SSLS}}$ of [ZBC19]. Figure 1.3a illustrates that SSLS is less efficient than the sample-mean estimator, that our SSL-Lasso is equivalent to the sample-mean, and that all other SSL-methods are more efficient with SSL-XGBoost outperforming the rest. Figure 1.3b demonstrates extremely poor finite-sample coverage of SSLS and nominal coverage of our proposal.

Model 5.3. Let $X_i \sim^{\text{iid}} N_{p-1}(0, C)$, be equi-correlated with $C_{ij} = \{1 - 1/(2p)\}1_{\{i=j\}} + 1/(2p)1_{\{i\neq j\}}$, with $p = 1001$, $m = 10n$. We consider a non-linear additive outcome model $Y_i = \sum_{j=1}^{p-1} 0.7^{j-1} \sin(X_{ij}) + \delta_i$, where $\delta_i \sim^{\text{iid}} N(0, 0.25)$. Figure 1.4a demonstrates significant gain in reduction of MSE of the proposed method with the SSL-Lasso in the lead. Figure 1.4b presents strong finite sample coverage.

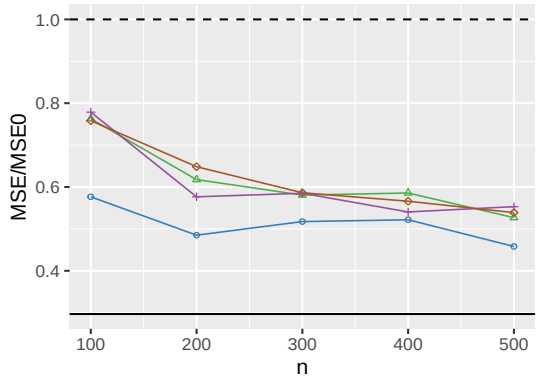


(a) The ratio of mean squared errors.

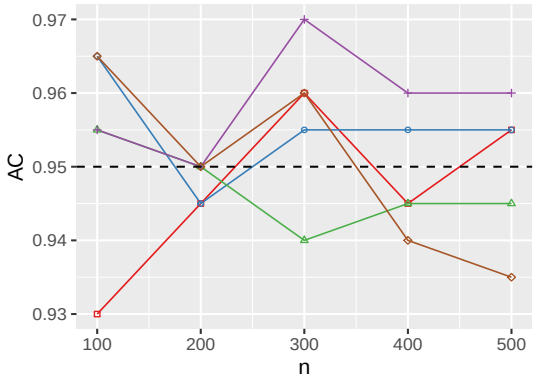


(b) The average coverage.

Figure 1.3: Model 5.2: Comparison of SSL-method estimators and the SSLS [ZBC19]. The plot includes sample mean (red squares), SSL-Lasso (blue circles), SSL-Additive (green up triangles), SSL-XGBoost (purple pluses), SSL-RF (brown diamonds), and SSLS (pink down triangles) estimates.

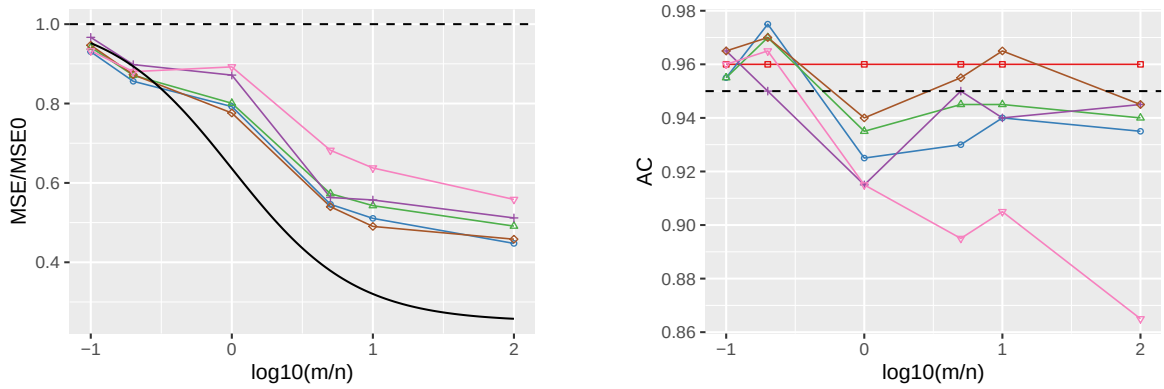


(a) The ratio of mean squared errors.



(b) The average coverage.

Figure 1.4: Model 5.3: Comparison of SSL-method and the sample mean. The plot includes sample mean (red squares), SSL-Lasso (blue circles), SSL-Additive (green up triangles), SSL-XGBoost (purple pluses), and SSL-RF (brown diamonds) estimates.



(a) The ratio of mean squared errors.

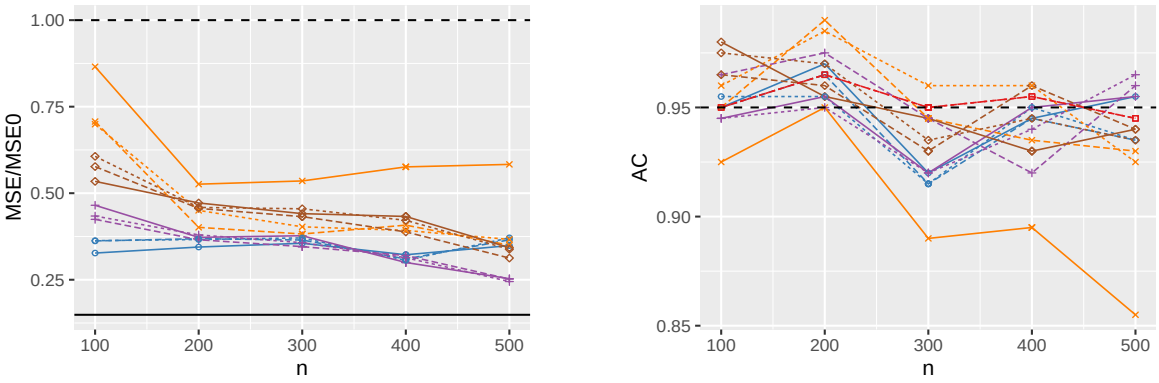
(b) The average coverage.

Figure 1.5: Model 5.4: Impact of the size of additional data. The plot includes sample mean (red squares), SSL-Lasso (blue circles), SSL-Additive (green up triangles), SSL-XGBoost (purple pluses), SSL-RF (brown diamonds), and SSLS (pink down triangles) estimates.

Model 5.4. Here we observe behavior with varying m. Let X_i and δ_i be as in Model 5.1. and consider the non-linear outcome of Model 5.3.. Set $p = 201$, $n = 500$ and let m vary from $0.1n$ to $10n$. We compare with $\bar{Y}$ and SSLS of [ZBC19]. We see substantial gains in efficiency. SSL-RF dominates the other estimators, both in terms of MSE, Figure 1.5a, and coverage Figure 1.5b. SSLS is losing coverage with a larger m. When m is small, the ordinary least squares estimate's impact is not significant, and SSLS is similar to the sample mean $\bar{Y}$. As m grows, the instability of least-squares and the unfitness of SSLS is exposed.

Model 5.5. Is sample-splitting needed? Let $X_i \sim^{\text{iid}} \text{Lognormal}_{p-1}(0, C)$, with C as in Model 5.3. with $p = 101$ and $m = 10n$. Let $Y_i = \sum_{j=1}^{3}\{\log(X_{ij}+1)^2 + 0.1\} + \delta_i$, where $\delta_i \sim^{\text{iid}} N(0, 0.25)$. We varied K from 1 to 5 and then to 20. We observe that some methods, like SSL-MLP, benefit significantly from sample splitting: without it, they under-cover, Figure 1.6b, and have the largest MSE, Figure 1.6a.

Model 5.6. In finite samples, the randomness from the K-partition creates an additional variance. We repeat the random K-partition for S times, and for each time, we



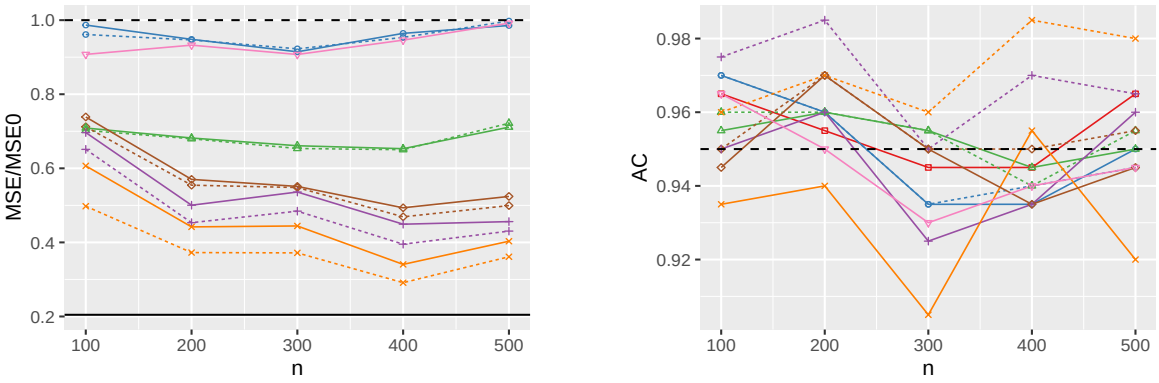
(a) The ratio of mean squared errors.

(b) The average coverage.

Figure 1.6: Model 5.5: Is sample-splitting needed? The plot includes sample mean (red squares), SSL-Lasso (blue circles), SSL-XGBoost (purple pluses), SSL-MLP (orange crosses), and SSL-RF (brown diamonds) estimates. The number of folds, K, is denoted with solid, dashed, and long dashed lines for $K = 1$(without cross-fitting), $K = 5$, and $K = 20$, respectively.

obtain an estimate $\hat{\theta}^s$ and the corresponding estimated asymptotic variance $\hat{V}(\hat{\theta}^s)$. Here we compare $\hat{\theta}^1$ with the average $\tilde{\theta} = S^{-1}\sum_{s=1}^{S}\hat{\theta}^s$. An asymptotic confidence interval based on $\tilde{\theta}$ can be constructed using an estimated variance $\tilde{V}(\tilde{\theta}) = S^{-1}\sum_{s=1^S}\{\hat{V}(\hat{\theta}^s) + (\hat{\theta}^s - \tilde{\theta})^2\}$. The outcome model is non-linear with one interaction term $Y_i = X_{i1}X_{i2} + 0.5(X_{i3} + 0.5)^2 + \delta_i$ and X_i and δ_i are as in Model 5.1. with $p = 4$, $m = 10n$. Figures 1.7a and 1.7b illustrate that partitions do not matter much for the least-squares procedure: SSL-Lasso, SSL-Additive, and SSL-RF do not vary much. However, highly non-linear methods, such as SSL-MLP and SSL-XGBoost, benefit significantly from repeating the partitioning process.

Model 5.7. (ATE) Let $X_{ij} \sim^{\text{iid}} \text{Uniform}(-1, 1)$, $p = 11$, $D_i \sim \text{Bernoulli}[1/\{1 + \exp(5^{1/2}\sum_{j=1}^{5} X_{ij}/2)\}]$. Linear Setting: the outcome model is $Y_i = D_i(1 + \beta_1^{\mathbb{T}}X_i) + (1 - D_i)\beta_0^{\mathbb{T}}X_i + \delta_i$, where $\delta_i \sim^{\text{iid}} N(0, 0.2^2)$ and $\beta_0 = -(0.5^{1/2}, 0.5, 0.5^{3/2}, 0.5^2, 0.5^2, 0, 0, 0, 0, 0)$, $\beta_1 = -\beta_0$. Non-linear Setting: the outcome model is $Y_i = D_i\{X_{i1}X_{i2} + 0.5(X_{i3} + 0.5)^2\} + (1-D_i)\{X_{i1}X_{i2} - 0.5(X_{i3}+0.5)^2\} + \delta_i$. For the Linear Setting our proposed estimator and the



(a) The ratio of mean squared errors.

(b) The average coverage.

Figure 1.7: Model 5.6: Does partitioning matter? The plot includes sample mean (red squares), SSL-OLS (blue circles), SSL-Additive (green up triangles), SSL-XGBoost (purple pluses), SSL-MLP (orange crosses), SSL-RF (brown diamonds), and SSL-SSLS (pink down triangles) estimates. The number of cross-fitting repetitions, S, is denoted with solid and dashed lines for $S = 1$ and $S = 5$, respectively.

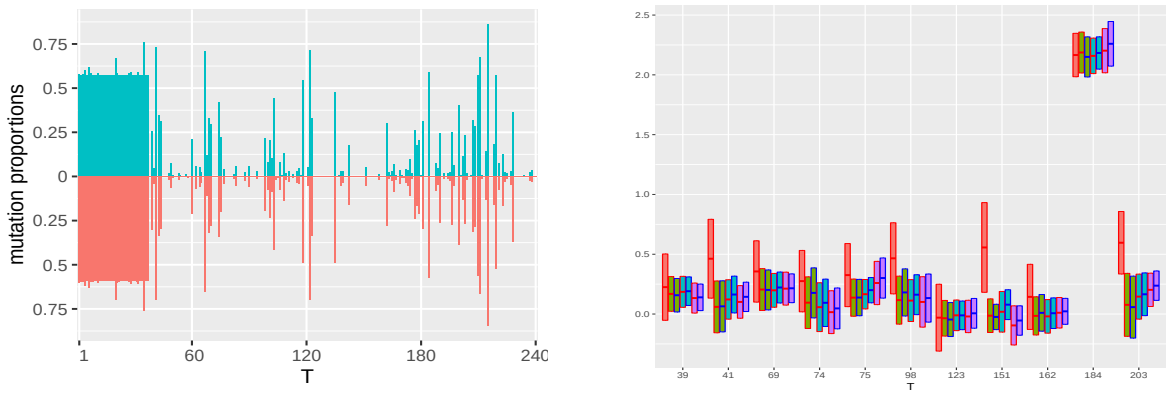
estimator of [CCD$^+$17], estimate the propensity and the outcome model by cross-validated generalized and linear ridge regression. For the Non-Linear Setting, the outcome models are estimated by ridge regression, additive model, and multilayer perceptron. Parameters α and β of [CAC18] are estimated by cross-validated adaptive lasso, where the initial weights are estimated by linear regression or generalized linear regression; the parameter γ is estimated by cross-validated lasso; the kernel is chosen to be 6-th order Gaussian, and the bandwidth is estimated by the plug-in method. Table 1.1 contains all the results. We found that the biases of our SSL and the supervised estimator of [CCD$^+$17] are not sensitive to the choice of the tuning parameters, while the bias of [CAC18] is. Under the linear outcome models, the two SSL estimators have smaller mean squared errors than the supervised estimator; under non-linear outcome models, our semi-supervised mlp+ridge estimator outperforms the others.

Table 1.1: Experiments for the ATE including Bias: average of the estimation biases, Emp SE: empirical standard error, ASE: average of estimated standard errors, RMSE: root-mean-square error, and AC: average coverage of the 95% confidence intervals

Estimator	Bias	Emp SE	ASE	RMSE	AC
$n = 100, m = 200$			Linear Outcome		
Zhang & Bradic (ridge+ridge)	0.0010	0.0881	0.0812	0.0879	0.935
[CCD$^+$17] (ridge+ridge)	0.0097	0.1295	0.1238	0.1295	0.930
[CAC18]	-0.0147	0.0885	0.0801	0.0895	0.925
$n = 500, m = 1000$			Linear Outcome		
Zhang & Bradic (ridge+ridge)	-0.0025	0.0333	0.0351	0.0333	0.945
[CCD$^+$17] (ridge+ridge)	-0.0052	0.0588	0.0546	0.0588	0.965
[CAC18]	-0.0093	0.0329	0.0352	0.0341	0.940
$n = 200, m = 400$			Non-Linear Outcome		
Zhang & Bradic (ridge+ridge)	0.0031	0.0660	0.0672	0.0659	0.965
[CCD$^+$17] (ridge+ridge)	0.0051	0.0714	0.0737	0.0714	0.955
Zhang & Bradic (additive+ridge)	0.0027	0.0622	0.0638	0.0621	0.960
[CCD$^+$17] (additive+ridge)	0.0054	0.0705	0.0731	0.0706	0.960
Zhang & Bradic (mlp+ridge)	-0.0027	0.0518	0.0497	0.0518	0.935
[CCD$^+$17] (mlp+ridge)	0.0015	0.0570	0.0596	0.0569	0.960
[CAC18]	-0.0209	0.0637	0.0655	0.0669	0.970
$n = 500, m = 1000$			Non-Linear Outcome		
Zhang & Bradic (ridge+ridge)	-0.0005	0.0384	0.0413	0.0383	0.970
[CCD$^+$17] (ridge+ridge)	-0.0014	0.0433	0.0457	0.0432	0.955
Zhang & Bradic (additive+ridge)	-0.0001	0.0385	0.0395	0.0383	0.975
[CCD$^+$17] (additive+ridge)	-0.0006	0.0436	0.0455	0.0435	0.960
Zhang & Bradic (mlp+ridge)	-0.0025	0.0256	0.0275	0.0255	0.975
[CCD$^+$17] (mlp+ridge)	-0.0017	0.0361	0.0354	0.0361	0.940
[CAC18]	-0.0143	0.0377	0.0408	0.0402	0.945

1.7.2 Real data

We consider the dataset of [BSB+06], available at the Stanford University HIV Drug Resistance Database [RGK+03] `https://hivdb.stanford.edu`. It is known that mutations are common in HIV, and some of the mutations may affect HIV drug resistance. We provide estimation and inference for the average treatment effect of a specific mutation on the reverse transcriptase (RT) to the drug resistance. The outcome is lamivudine (3TC), a nucleoside reverse transcriptase inhibitor (NRTI), drug resistance. The treatment, D, denotes the existence of a mutation on the T-th position of the HIV's RT. Explanatory variables X_j, where $j \in \{1, 2, \ldots, 240\} \setminus \{T\}$, denotes existence of a mutation on the j-th position. We consider the subtype B sequence. Redundant viruses obtained from the same individuals were excluded. We obtained $n = 423$ pairs of supervised data $(D_{i,T}, Y_i, \{X_{i,j}\}_{j \neq T})_{i=1}^{n}$ and $m = 2458$ pairs of additional unlabeled covariates $(D_{i,T}, \{X_{i,j}\}_{j \neq T})_{i=n+1}^{m+n}$. Fix $T \in \{1, 2, \ldots, 240\}$. Before we perform our semi-supervised methods, we first check whether there is a significant difference between the distribution of X in the two groups; see the back-to-back bar chart of the labeled and unlabeled group's mutation proportions on different RT positions in Figure 1.8a. The p-value based on Pearson Statistic was obtained using a permutation distribution [AK05] and resulted in a value of 0.178. We do not have any significant evidence that the covariates' distributions differ between the supervised and unlabeled groups. Estimators of the propensity score and the outcome model are: (logistic) lasso + lasso, XGBoost + XGBoost, and random forest + random forest. In order to improve the stability of the estimator, we trim each $\hat{e}^{(-k)}(X_i)$ to $(0.01, 0.99)$. We compare with the sample estimator $(\sum_{i=1}^{n} D_i)^{-1} \sum_{i=1}^{n} D_i Y_i - \{\sum_{i=1}^{n}(1 - D_i)\}^{-1} \sum_{i=1}^{n}(1 - D_i) Y_i$, suitable only for homogeneous



(a) (b)

Figure 1.8: Real data. Left: A back-to-back bar chart comparing the labeled and unlabeled group's mutation proportions on RT positions between 1-240. The blue color on the top denotes the unlabeled group, and the red color on the bottom denotes the labeled group. Right: Confidence intervals of the average treatment effect. We compare the sample mean of the labeled samples (red border and red fill), supervised [CCD$^+$17] estimators (red border), and our SSL-method estimators (blue border). Estimators of the propensity score and the outcome model are: logistic + Lasso (green fill), XGBoost + XGBoost (aqua fill), RF + RF (purple fill).

effects. Figure 1.8b shows the confidence intervals for δ on several positions based on different estimators. We can see that there is a large average treatment effect on position 184, a small average treatment effect on positions 39, 69, and potentially a small average treatment effect on positions 41, 75, and 203. The sample estimator is most different from the rest on positions 41, 98, 151, and 203. The sample estimator is biased when the distribution of X on treated and control is different. It implies that the mutations on positions 41, 98, 151, and 203 are significantly dependent on the other positions' mutations. Moreover, our confidence intervals are shorter than those of [CCD$^+$17]. It coincides with the fact that additional unlabeled data provide improved asymptotic efficiency.

1.8 Proofs of main results

Notation A constant $c > 0$, independent of n, p, m may change value from one line to the other. For any vector $a \in \mathbb{R}^p$ and $r > 0$, $\|a\|_r = (\sum_{j=1}^p a_j^r)^{1/r}$, $\|a\|_0 = \#\{j \leq p : a_j \neq 0\}$. For any matrix $A \in \mathbb{R}^{p\times p}$, we denote $\|A\|_2 = \sup_{z\neq 0} \|Az\|_2/\|z\|_2$. We define $\mu_r(f) = E\{f - E(f)\}^r$ being the r-th central moment, and $\mu_{r,X}(f) = E_X\{f - E_X(f)\}^r$. Recall that $E_X(f) = \int f dP_X$ is the conditional expectation on the marginal distribution P_X. With a slight abuse of notation, for any function g,

$$E_{I_k^c}(g) = E\{g \mid (Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}\}$$

and $(Y, X) \sim P_{Y,X}$ independent of $(Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}$ in the proof of Theorems 1.1, 1.2, 1.3, 1.4, 1.5, 1.6;

$$E_{I_k^c}(g) = E\{g \mid (D_i, Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}\}$$

and $(D, Y, X) \sim P_{D,Y,X}$ independent of $(D_i, Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}$ in the proof of Theorem 1.8.

$$E_{J_k^c}(g) = E\{g \mid (T_i, Y_i, X_i)_{i\in\{1,2,\ldots,m+n\}\setminus J_k}\}$$

and $(T, Y, X) \sim P_{T,Y,X}$ independent of $(T_i, Y_i, X_i)_{i\in\{1,2,\ldots,m+n\}\setminus J_k}$ in the proof of Theorem 1.7.

1.8.1 Auxiliary Lemmas

We begin by presenting three simple results that will be useful throughout the document.

Lemma 1.1 (Lemma B.1 in [CCD+17]). *Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables. If for any $c > 0$, $P(|X_n| > c|Y_n) = o_P(1)$. Then, $X_n = o_P(1)$. In particular, this*

occurs if $E(|X_n|^q \mid Y_n) = o_P(1)$ *for any* $q \geq 1$. *Typical examples we used in our proofs are a)* $E(X_n^2 \mid Y_n) = o_P(1)$, *b)* $X_n = \sum_{i=1}^n Z_{n,i}/n$, *where* $(Z_{n,i})$ *is a row-wise independent and identically distributed triangular array, conditional on* Y_n, *with* $E(|Z_{n,1}| \mid Y_n) = o_P(1)$.

Lemma 1.2. *Let* $\{X_n\}$ *and* $\{Y_n\}$ *be sequences of random variables. If* $E(X_n^2 \mid Y_n) = O_P(1)$, *then* $X_n = O_P(1)$. *Consequently, if* $(Z_{n,i})$ *is a row-wise independent and identically distributed triangular array conditional on* Y_n, *with* $\mathrm{var}(Z_{n,1} \mid Y_n) = O_P(1)$, *or a stronger condition that* $E(Z_{n,1}^2 \mid Y_n) = O_P(1)$. *Then,* $\sum_{i=1}^n Z_{n,i}/n = E(Z_{n,1}) + O_P(n^{-1/2})$.

Proof of Lemma 1.2. For any $c > 0$, since $E(X_n^2 \mid Y_n) = O_P(1)$, there exists $C > 0$ such that, for all $n \geq 1$,

$$P\{E(X_n^2 \mid Y_n) > C\} < c_1/2.$$

Hence,

$$\begin{aligned}
&P\{|X_n| > (2C/c)^{1/2}\} = E[1_{\{|X_n|>(2C/c)^{1/2}\}}] \\
&= E\left[1_{\{E(X_n^2|Y_n)\leq C\}} E(1_{\{|X_n|>(2C/c)^{1/2}\}} \mid Y_n)\right] + E\left(1_{\{E(X_n^2|Y_n)>C\}} E[1_{\{|X_n|>(2C/c)^{1/2}\}} \mid Y_n]\right) \\
&< E\left[1_{\{E(X_n^2|Y_n)\leq C\}} E\{cX_n^2/(2C) \mid Y_n\}\right] + E[1_{\{E(X_n^2|Y_n)>C\}}] \\
&= cE\left[1_{\{E(X_n^2|Y_n)\leq C\}} E(X_n^2 \mid Y_n)\right]/(2C) + P\{E(X_n^2 \mid Y_n) > C\} \\
&\leq c/2 + c/2 = c.
\end{aligned}$$

That is, $X_n = O_P(1)$. It follows that $\sum_{i=1}^n Z_{n,i}/n = E(Z_{n,1}) + O_P(n^{-1/2})$ since

$$E\left[n\left\{n^{-1}\sum_{i=1}^n Z_{n,i} - E(Z_{n,1})\right\}^2 \mid Y_n\right] = \mathrm{var}(Z_{n,1} \mid Y_n) = O_P(1).$$

■

Lemma 1.3. *Let $(Z_{n,i})$ be a row-wise independent and identically distributed triangular array with $E(Z_{n,1}) = 0$ and $E|Z_{n,1}|^q < c_1$ for $q > 1$ and $C < \infty$. Let $X_n = \sum_{i=1}^n Z_{n,i}/n$. Then, $X_n = o_P(1)$.*

Proof of Lemma 1.3. Let $Y_{n,i} = Z_{n,i}1_{\{|Z_{n,i}|<n\}}$. For any $c > 0$,

$$P(|X_n| \geq c) \leq P\{\cup_{i=1}^n (Z_{n,i} \neq Y_{n,i})\} + P\left(\left|\sum_{i=1}^n Y_{n,i}\right| \geq nc\right)$$

Let $r \in (1, q \wedge 2)$, then $E|Z_{n,1}|^r \leq (E|Z_{n,1}|^q)^{r/q} < c_1^{r/q}$. By Markov's Inequality,

$$P\{\cup_{i=1}^n (Z_{n,i} \neq Y_{n,i})\} \leq nP(|Z_{n,1}| \geq n) \leq nE|Z_{n,1}|^q/n^q = n^{1-q}E|Z_{n,1}|^q = o(1)$$

and

$$\begin{aligned} P\left(\left|\sum_{i=1}^n Y_{n,i}\right| \geq nc\right) &\leq E\left|\sum_{i=1}^n Y_{n,i}\right|^2 /(nc)^2 = nE\left[Z_{n,1}^2 1_{\{|Z_{n,i}|<n\}}/(nc)^2\right] \\ &= E\left[|Z_{n,1}|^r |Z_{n,1}|^{2-r} 1_{\{|Z_{n,i}|<n\}}/(nc^2)\right] \leq n^{1-r}E|Z_{n,1}|^r/c^2 = o(1). \end{aligned}$$

Hence, $P(|X_n| \geq c) = o_P(1)$. That is, $X_n = o_P(1)$. ∎

1.8.2 Proofs of the main theorems

Proof of Theorem 1.1. This proof provides $n^{1/2}$ consistencies of $\hat{\theta}$ and $\hat{\sigma}_Y^2$.

Part 1. We first assume Condition 1.1 and 1.3 and show that $\hat{\theta} - \theta = O_P(n^{-1/2})$. By the definition of β^*, as in Lemma 1 of [ZBC19],

$$E(\varepsilon) = 0,\ E(\tilde{X}\varepsilon) = 0,\ \theta = \beta^{*\mathbb{T}}\tilde{\mu},\ \sigma_Y^2 = b^2 + \sigma_\varepsilon^2.$$

By the definition of $\hat{\theta}^{(k)}$ as in (1.1),

$$\hat{\theta}^{(k)} - \theta = N^{-1}\sum_{i\in I_k}(Y_i - \theta) - N^{-1}\sum_{i\in I_k} \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i + M^{-1}\sum_{i\in J_k} \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i. \tag{1.32}$$

Now we will show that each of the terms on the RHS is of the order $O_P(n^{-1/2})$. From Conditions 1.1, 1.3 and recall that $\tilde{V} = \tilde{X} - \tilde{\mu}$ is independent of $(Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}$, while $\hat{\beta}^{(-k)}$ is a function of $(Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}$,

$$E(Y-\theta)^2 \le (E|Y-\theta|^{2+c})^{2/(2+c)} < c_1,$$

$$E(\beta^{*\mathbb{T}}\tilde{V})^2 = E(Y-\theta)^2 - \sigma_\varepsilon^2 \le E(Y-\theta)^2 < c_1,$$

$$E_{I_k^c}\left\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\right\}^2 = (\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}(\hat{\beta}^{(-k)}-\beta^*) \le \|\hat{\beta}^{(-k)}-\beta^*\|_2^2\|\tilde{C}\|_2 = O_P(1),$$

and by triangle inequality, $E_{I_k^c}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V})^2 = O_P(1)$. Then, by Lemma 1.2,

$$N^{-1}\sum_{i\in I_k}(Y_i-\theta) = O_P(N^{-1/2}), \tag{1.33}$$

$$N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i = O_P(N^{-1/2}), \tag{1.34}$$

$$M^{-1}\sum_{i\in J_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i = O_P(M^{-1/2}). \tag{1.35}$$

Therefore, $\hat{\theta}^{(k)} - \theta = O_P(N^{-1/2}) + O_P(N^{-1/2}) + O_P(M^{-1/2}) = O_P(N^{-1/2})$, since $M \ge N$. When $K < \infty$,

$$\hat{\theta} = K^{-1}\sum_{k=1}^{K}\hat{\theta}^{(k)} = \theta + O_P(n^{-1/2}). \tag{1.36}$$

Part 2. Now we assume Condition 1.2 and 1.3 and show that $\hat{\sigma}_Y^2 - \sigma_Y^2 = O_P(n^{-1/2})$. Recall the definition of $\hat{\sigma}_Y^{2^{(k)}}$ in Section 1.2.2,

$$\hat{\sigma}_Y^{2^{(k)}} = N^{-1}\sum_{i\in I_k}(Y_i-\hat{\theta})^2 + M^{-1}\sum_{i\in J_k}\left(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\right)^2 - N^{-1}\sum_{i\in I_k}\left(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\right)^2,$$

We first approximate the terms on the RHS by replacing $\hat{\theta}$ and $\hat{V}_i$ by θ and $\tilde{V}_i$, respectively. Recall (1.33) and (1.36),

$$N^{-1}\sum_{i\in I_k}(Y_i-\hat{\theta})^2 = N^{-1}\sum_{i\in I_k}(Y_i-\theta)^2 + (\hat{\theta}-\theta)^2 - 2(\hat{\theta}-\theta)N^{-1}\sum_{i\in I_k}(Y_i-\theta)$$
$$= N^{-1}\sum_{i\in I_k}(Y_i-\theta)^2 + O_P(N^{-1})$$

Besides, by definition, $\hat{V}_i = \tilde{V}_i - (\hat{\mu}^{(k)} - \tilde{\mu})$, where $\hat{\mu}^{(k)} - \tilde{\mu} = M^{-1}\sum_{i\in J_k} \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i$. Recall (1.34) and (1.35),

$$M^{-1}\sum_{i\in J_k}\Big(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\Big)^2 = M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 - \{\hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})\}^2$$
$$= M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(M^{-1}), \tag{1.37}$$

and

$$N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i)^2 = N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + \{\hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})\}^2$$
$$- 2\,\hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i$$
$$= N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(M^{-1}+N^{-1/2}M^{-1/2}). \tag{1.38}$$

Hence,

$$\hat{\sigma}_Y^{2^{(k)}} = N^{-1}\sum_{i\in I_k}(Y_i-\theta)^2 - N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(N^{-1}). \tag{1.39}$$

Now we will show that each of the terms on the RHS of (1.39) is of the order $O_P(N^{-1/2})$. By Lemma 1.2, it suffices to show

$$E(Y-\theta)^4 = O(1), \tag{1.40}$$

$$E_{I_k^c}(\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V})^4 = O_P(1). \tag{1.41}$$

Here, (1.40) follows by the assumption that $E|Y|^{4+c} < c_1$. Besides, recall that $\tilde{V} = \tilde{X} - \tilde{\mu} = (0, V^{\mathbb{T}})^{\mathbb{T}}$, where $V = C^{1/2}Z$. By Condition 1.2, we have bounded 4-th moments

$$E(\beta^{*\mathbb{T}}\tilde{V})^4 = E(\ {\beta^*_{-1}}^{\mathbb{T}} C^{1/2}Z)^4 = b^4 E(\ {\beta^*_{-1}}^{\mathbb{T}} C^{1/2}Z/b)^4 \leq b^4 \sup_{\|a\|_2=1} E(a^{\mathbb{T}}Z)^4 = O(1),$$

$$E_{I_k^c}\left\{(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}\right\}^4 \leq \|\hat{\beta}^{(-k)} - \beta^*\|^4_{\tilde{C}} \sup_{\|a\|_2=1} E(a^{\mathbb{T}}Z)^4 = O_P(1),$$

and hence (1.41) follows. Now, we obtain

$$\hat{\sigma}_Y^{2^{(k)}} - \sigma_Y^2 = O_P(N^{-1/2}) \tag{1.42}$$

and the proof is finalized by noticing that for finite K, the rate above is inherited for the averaged estimator

$$\hat{\sigma}_Y^2 = \sigma_Y^2 + O_P(n^{-1/2}). \tag{1.43}$$

■

Proof of Theorem 1.2. This proof provides an asymptotic normal result for $n^{1/2}(\hat{\theta} - \theta)$ by relying on Conditions 1.1 and 1.4. Recall from (1.32),

$$\begin{aligned}\hat{\theta}^{(k)} - \theta &= N^{-1}\sum_{i\in I_k}(Y_i - \theta) - N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i + M^{-1}\sum_{i\in J_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i \\ &= N^{-1}\sum_{i\in I_k}\varepsilon_i + M^{-1}\sum_{i\in J_k}\beta^{*\mathbb{T}}\tilde{V}_i - N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}_i + M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}_i\end{aligned}$$

Since

$$E_{I_k^c}\left\{(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}\right\}^2 = \|\hat{\beta}^{(-k)} - \beta^*\|^2_{\tilde{C}} \leq \|\hat{\beta}^{(-k)} - \beta^*\|_2^2\|\tilde{C}\|_2 = o_P(1),$$

and by Lemma 1.1,

$$N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}_i = o_P(N^{-1/2}), \qquad M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}}\tilde{V}_i = o_P(M^{-1/2}).$$

Therefore,

$$\hat{\theta}^{(k)} - \theta = N^{-1} \sum_{i \in I_k} \varepsilon_i + M^{-1} \sum_{i \in J_k} \beta^{*\mathbb{T}} \tilde{V}_i + o_P(N^{-1/2}).$$

When $K < \infty$,

$$n^{1/2}(\hat{\theta} - \theta) = n^{-1/2} \sum_{i=1}^{n} \varepsilon_i + n^{1/2} M^{-1} \sum_{i \in J_k} \beta^{*\mathbb{T}} \tilde{V}_i + o_P(1).$$

By Condition 1.1, $E|\varepsilon|^{2+c} < c_1,\ E|\ \beta^{*\mathbb{T}} \tilde{V}|^{2+c} < c_1$. With a slight abuse of notation, assume that $\sigma_\varepsilon^2 = \lim_{n\to\infty} E(\varepsilon^2), \quad \tau b^2 = \lim_{n\to\infty} nE(\ \beta^{*\mathbb{T}} \tilde{V})^2/(m+n)$ both exists. We continue the analysis by analyzing three separate cases.

a) When $\sigma_\varepsilon^2 > 0$ and $\tau b^2 > 0$,

$$\frac{E|\varepsilon|^{2+c}}{\{E(\varepsilon^2)\}^{1+c/2}} < c_1, \quad \frac{E|(n/(m+n))^{1/2}\ \beta^{*\mathbb{T}} \tilde{V}|^{2+c}}{\{nE(\ \beta^{*\mathbb{T}} \tilde{V})^2/(m+n)\}^{1+c/2}} < c_1,$$

i.e. the Lyapunov condition holds. By Lindeberg-Feller Central Limit Theorem,

$$n^{-1/2} \sum_{i=1}^{n} \varepsilon_i \to N\left(0, \sigma_\varepsilon^2\right), \quad n^{1/2} M^{-1} \sum_{i \in J_k} \beta^{*\mathbb{T}} \tilde{V}_i \to N\left(0, \tau b^2\right).$$

in distribution. By Slutsky's Theorem and multivariate delta method,

$$n^{1/2}(\hat{\theta} - \theta) = n^{-1/2} \sum_{i=1}^{n} \varepsilon_i + n^{1/2} M^{-1} \sum_{i \in J_k} \beta^{*\mathbb{T}} \tilde{V}_i + o_P(1) \to N\left(0, \sigma_\varepsilon^2 + \tau b^2\right). \tag{1.44}$$

b) When $\sigma_\varepsilon^2 = 0$, recall the assumption that $\sigma_\varepsilon^2 + \tau b^2 > 0$, we have $\tau b^2 > 0$. In this case, by Lemma 1.1 and Lindeberg-Feller Central Limit Theorem,

$$n^{-1/2} \sum_{i=1}^{n} \varepsilon_i = o_P(1), \quad n^{1/2} M^{-1} \sum_{i \in J_k} \beta^{*\mathbb{T}} \tilde{V}_i \to N\left(0, \tau b^2\right).$$

By Slutsky's Theorem, (1.44) holds.

c) When $\tau b^2 = 0$, similarly as in b), (1.44) holds. ∎

Proof of Theorem 1.3. This proof provides consistency results for $\hat{\sigma}_\varepsilon^2$ and $\hat{b}^2$ by assuming Conditions 1.1 and 1.4.

Part 1. We first show that $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_P(1)$. Recall the definition of $\hat{\sigma}_\varepsilon^{2^{(k)}}$,

$$\hat{\sigma}_\varepsilon^{2^{(k)}} = N^{-1} \sum_{i \in I_k} (Y_i - \hat{\theta} - \ \hat{\beta}^{(-k)^{\mathbb{T}}} \hat{V}_i)^2.$$

Now we first approximate the RHS by replacing $\hat{\theta}$ and $\hat{V}_i$ by θ and $\tilde{V}_i$, respectively. Recall from (1.33) – (1.36),

$$\begin{aligned}
\hat{\sigma}_\varepsilon^{2^{(k)}} &= N^{-1} \sum_{i \in I_k} (Y_i - \theta - \ \hat{\beta}^{(-k)^{\mathbb{T}}} \tilde{V}_i)^2 + \{\hat{\theta} - \theta - \ \hat{\beta}^{(-k)^{\mathbb{T}}} (\hat{\mu}^{(k)} - \tilde{\mu})\}^2 \\
&\quad - 2\{\hat{\theta} - \theta - \ \hat{\beta}^{(-k)^{\mathbb{T}}} (\hat{\mu}^{(k)} - \tilde{\mu})\} N^{-1} \sum_{i \in I_k} (Y_i - \theta - \ \hat{\beta}^{(-k)^{\mathbb{T}}} \tilde{V}_i) \\
&= N^{-1} \sum_{i \in I_k} \varepsilon_i^2 + N^{-1} \sum_{i \in I_k} \{(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}_i\}^2 \\
&\quad - 2N^{-1} \sum_{i \in I_k} \varepsilon_i (\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}_i + O_P(N^{-1}). \qquad (1.45)
\end{aligned}$$

Remember the definition that $E_{I_k^c}(g) = E\{g \mid (Y_i, X_i)_{i \in \{1,2,\ldots,m+n\} \setminus I_k}\}$. By Condition 1.1, $E|\varepsilon|^{2+c} < c_1$, $E_{I_k^c}\{(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}\}^2 = o_P(1)$ and hence $E_{I_k^c}|\varepsilon(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}| = o_P(1)$. By Lemma 1.1,

$$N^{-1} \sum_{i \in I_k} \{(\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}_i\}^2 = o_P(1), \qquad N^{-1} \sum_{i \in I_k} \varepsilon_i (\hat{\beta}^{(-k)} - \beta^*)^{\mathbb{T}} \tilde{V}_i = o_P(1).$$

By Lemma 1.3,

$$N^{-1} \sum_{i \in I_k} \varepsilon_i^2 = \sigma_\varepsilon^2 + o_P(1).$$

Hence, $\hat{\sigma}_\varepsilon^{2^{(k)}} = \sigma_\varepsilon^2 + o_P(1)$. When $K < \infty$, $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_P(1)$.

Part 2. Now we show that $\hat{b}^2 = b^2 + o_P(1)$. By the definition of $\hat{b}^{2^{(k)}}$,

$$\hat{b}^{2^{(k)}} = M^{-1} \sum_{i \in J_k} (\hat{\beta}^{(-k)^{\mathbb{T}}} \hat{V}_i)^2 + 2N^{-1} \sum_{i \in I_k} \hat{\beta}^{(-k)^{\mathbb{T}}} \hat{V}_i (Y_i - \hat{\theta}) - 2N^{-1} \sum_{i \in I_k} (\hat{\beta}^{(-k)^{\mathbb{T}}} \hat{V}_i)^2.$$

We first approximate the terms of RHS by replacing $\hat{\theta}$ and $\hat{V}_i$ by θ and $\tilde{V}_i$, respectively. Recall from (1.37) and (1.38),

$$\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)} = M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(M^{-1}), \tag{1.46}$$

$$N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i)^2 = N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(N^{-1}).$$

Recall from (1.33) – (1.36),

$$\begin{aligned}
&N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i(Y_i-\hat{\theta})\\
&= N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta) + \ \hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})(\hat{\theta}-\theta)\\
&\quad - \hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})N^{-1}\sum_{i\in I_k}(Y_i-\theta) - (\hat{\theta}-\theta)N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i\\
&= N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta) + O_P(N^{-1}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{b}^{2^{(k)}} = M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + 2N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta)\\
- 2N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(N^{-1}).
\end{aligned}$$

The first term on the RHS can be expressed as

$$\begin{aligned}
M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 = M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 + 2M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\ \beta^{*\mathbb{T}}\tilde{V}_i\\
+ M^{-1}\sum_{i\in J_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2.
\end{aligned}$$

By Condition 1.1 and 1.4, $E|\ \beta^{*\mathbb{T}}\tilde{V}|^{2+c} < c_1$, $\quad E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\}^2 = o_P(1)$, which implies that $E_{I_k^c}|(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\ \beta^{*\mathbb{T}}\tilde{V}| = o_P(1)$. By Lemma 1.3, $M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 = b^2 + o_P(1)$.

By Lemma 1.1,

$$M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\ \beta^{*\mathbb{T}}\tilde{V}_i=o_P(1),\quad M^{-1}\sum_{i\in J_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2=o_P(1).$$

Hence,

$$M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2=b^2+o_P(1). \tag{1.47}$$

Similarly, $N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2=b^2+o_P(1)$. Recall from (1.33) – (1.36),

$$\begin{aligned}
&N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i(Y_i-\hat{\theta})\\
&=N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta)+\ \hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})(\hat{\theta}-\theta)\\
&\quad-\hat{\beta}^{(-k)^{\mathbb{T}}}(\hat{\mu}^{(k)}-\tilde{\mu})N^{-1}\sum_{i\in I_k}(Y_i-\theta)-(\hat{\theta}-\theta)N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i\\
&=N^{-1}\sum_{i\in I_k}\ \beta^{*\mathbb{T}}\tilde{V}_i(Y_i-\theta)+N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i(Y_i-\theta)+O_P(N^{-1}).
\end{aligned}\tag{1.48}$$

By Condition 1.1 and 1.4, $E|\ \beta^{*\mathbb{T}}\tilde{V}(Y-\theta)|^{2+c}<c_1$ and $E_{I_k^c}|(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}(Y-\theta)|=o_P(1)$. By Lemma 1.3, $N^{-1}\sum_{i\in I_k}\ \beta^{*\mathbb{T}}\tilde{V}_i(Y_i-\theta)=b^2+o_P(1)$, and by Lemma 1.1, $N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i(Y_i-\theta)=o_P(1)$. Hence, $N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i(Y_i-\hat{\theta})=b^2+o_P(1)$. Combining all the previous results,

$$\begin{aligned}
\hat{b}^{2^{(k)}}&=M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2+2N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta)\\
&\quad-2N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2+O_P(N^{-1}) \qquad (1.49)\\
&=b^2+o_P(1)+2\{b^2+o_P(1)\}-2\{b^2+o_P(1)\}=b^2+o_P(1). \qquad (1.50)
\end{aligned}$$

When $K<\infty$, $\hat{b}^2=b^2+o_P(1)$.

Part 3. Now assume Conditions 1.2 and 1.4, we provide consistency rate results for

$\hat{\sigma}_\varepsilon^2$ and $\hat{b}^2$. We first consider $\hat{\sigma}_\varepsilon^2$, recall (1.45),

$$\hat{\sigma}_\varepsilon^{2^{(k)}} = N^{-1}\sum_{i\in I_k}\varepsilon_i^2 + N^{-1}\sum_{i\in I_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2 - 2N^{-1}\sum_{i\in I_k}\varepsilon_i(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i + O_P(N^{-1}).$$

By Conditions 1.2 and 1.4, $E(\varepsilon^4) < c_1$,

$$E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\}^4 \le \|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^4 \sup_{\|a\|_2=1} E(a^{\mathbb{T}}Z)^4 = O(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^4)$$

and hence $E_{I_k^c}\{\varepsilon(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\}^2 = O(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2)$. By Lemma 1.2,

$$N^{-1}\sum_{i\in I_k}\varepsilon_i^2 = \sigma_\varepsilon^2 + O_P(N^{-1/2}),$$

$$N^{-1}\sum_{i\in I_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2 = \|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 N^{-1/2}),$$

$$N^{-1}\sum_{i\in I_k}\varepsilon_i(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i = O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}} N^{-1/2}).$$

Hence,

$$\begin{aligned}\hat{\sigma}_\varepsilon^{2^{(k)}} &= N^{-1}\sum_{i\in I_k}\varepsilon_i^2 + N^{-1}\sum_{i\in I_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2 \\ &\quad - 2N^{-1}\sum_{i\in I_k}\varepsilon_i(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i + O_P(N^{-1}) \\ &= \sigma_\varepsilon^2 + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 + N^{-1/2}).\end{aligned}$$

When $K<\infty$, $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 + N^{-1/2}) = \sigma_\varepsilon^2 + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_2^2 + n^{-1/2})$.

By the same strategy, now we show the consistency result for $\hat{b}^2$. Recall (1.49),

$$\begin{aligned}\hat{b}^{2^{(k)}} &= M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + 2N^{-1}\sum_{i\in I_k}\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i(Y_i-\theta) \\ &\quad - 2N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 + O_P(N^{-1}),\end{aligned}$$

where

$$M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 = M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 + 2M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\ \beta^{*\mathbb{T}}\tilde{V}_i$$
$$+ M^{-1}\sum_{i\in J_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2.$$

By Conditions 1.1 and 1.2, $E(\ \beta^{*\mathbb{T}}\tilde{V})^4 < c_1$, and recall that $E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\}^4 = O(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^4)$, hence

$$E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\ \beta^{*\mathbb{T}}\tilde{V}\}^2 = O(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2)$$

By Lemma 1.2,

$$M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 = b^2 + O_P(M^{-1}),$$
$$M^{-1}\sum_{i\in J_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^2 = \|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2 M^{-1/2}), \tag{1.51}$$
$$M^{-1}\sum_{i\in J_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\ \beta^{*\mathbb{T}}\tilde{V}_i = (\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^* + O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}M^{-1/2}). \tag{1.52}$$

Hence, $M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 = b^2+\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+O_P(M^{-1})$. Similarly, $N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2 = b^2+\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+O_P(N^{-1})$. Besides, recall (1.48). Then, simple algebra concludes

$$N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i(Y_i-\hat{\theta})$$
$$= N^{-1}\sum_{i\in I_k}\ \beta^{*\mathbb{T}}\tilde{V}_i(Y_i-\theta) + N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i(Y_i-\theta)+O_P(N^{-1}).$$

By Conditions 1.2 and 1.4,

$$E\{\ \beta^{*\mathbb{T}}\tilde{V}(Y-\theta)\}^{2+c} = O(1), \qquad E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}(Y-\theta)\}^2 = O(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}^2).$$

By Lemma 1.2,

$$N^{-1}\sum_{i\in I_k}\ \beta^{*\mathbb{T}}\tilde{V}_i(Y_i-\theta)=b^2+O_P(N^{-1/2}),$$

$$N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i(Y_i-\theta)=(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+O_P(\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}N^{-1/2}).$$

Hence, $N^{-1}\sum_{i\in I_k}\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i(Y_i-\hat{\theta})=b^2+(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+O_P(N^{-1/2})$. Combining all previous results,

$$\begin{aligned}\hat{b}^{2^{(k)}}&=b^2+\|\hat{\beta}^{(-k)}-\beta^*\|^2_{\tilde{C}}+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+2\{b^2+(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*\}\\&\quad-2\{b^2+\|\hat{\beta}^{(-k)}-\beta^*\|^2_{\tilde{C}}+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*\}+O_P(N^{-1/2})\\&=b^2+O_P(\|\hat{\beta}^{(-k)}-\beta^*\|^2_{\tilde{C}}+N^{-1/2}).\end{aligned}$$

When $K<\infty$, $\hat{b}^2=b^2+O_P(\|\hat{\beta}^{(-k)}-\beta^*\|^2_{\tilde{C}}+N^{-1/2})=b^2+O_P(\|\hat{\beta}^{(-k)}-\beta^*\|^2_2+n^{-1/2})$. ∎

Proof of Theorem 1.4. This proof provides an asymptotic normal result for $n^{1/2}(\hat{\sigma}^2_Y-\sigma^2_Y)$ and a consistent estimate for the asymptotic variance.

Part 1. We first show $n^{1/2}(\hat{\sigma}^2_Y-\sigma^2_Y)\to N\left(0,\mathrm{var}(\varepsilon^2+2\ \beta^{*\mathbb{T}}\tilde{V}\varepsilon)+\tau\mathrm{var}(\ \beta^{*\mathbb{T}}\tilde{V})^2\right)$. Recall from (1.39),

$$\hat{\sigma}^{2^{(k)}}_Y=N^{-1}\sum_{i\in I_k}(Y_i-\theta)^2-N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2+M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2+O_P(N^{-1}).$$

By (1.51) and (1.52),

$$M^{-1}\sum_{i\in J_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2=M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2+\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+o_P(M^{-1/2}).$$

Similarly,

$$N^{-1}\sum_{i\in I_k}(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\tilde{V}_i)^2=N^{-1}\sum_{i\in I_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2+\|\hat{\beta}^{(-k)}-\beta^*\|_{\tilde{C}}+2(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{C}\beta^*+o_P(N^{-1/2}).$$

Hence,

$$\hat{\sigma}_Y^{2^{(k)}} = M^{-1}\sum_{i\in J_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 + N^{-1}\sum_{i\in I_k}(Y_i-\theta)^2 - N^{-1}\sum_{i\in I_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 + o_P(N^{-1/2}).$$

When $K < \infty$, by the independency between $(Y_i, X_i)_{i=1}^n$ and $\{X_i\}_{i=n+1}^{m+n}$, and similarly as (1.44),

$$\begin{aligned} n^{1/2}(\hat{\sigma}_Y^2 - \sigma_Y^2) &= n^{-1/2}\sum_{i\in I_k}(\varepsilon_i^2 + 2\varepsilon_i\ \beta^{*\mathbb{T}}\tilde{V}_i - \sigma_\varepsilon^2) \qquad (1.53)\\ &\quad + n^{1/2}(m+n)^{-1}\sum_{i=1}^{m+n}\left\{(\ \beta^{*\mathbb{T}}\tilde{V}_i)^2 - b^2\right\} + o_P(1) \qquad (1.54)\\ &\to N\left\{0, \mathrm{var}(\varepsilon^2 + 2\ \beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\ \beta^{*\mathbb{T}}\tilde{V})^2\right\}, \end{aligned}$$

in distribution, provided that $\mathrm{var}(\varepsilon^2 + 2\ \beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\ \beta^{*\mathbb{T}}\tilde{V})^2 > 0$.

Part 2. Now we prove the consistency of $\hat{\sigma}_\nu^2 + n\hat{\sigma}_\xi^2/(m+n)$. It suffices to show

$$\hat{\sigma}_\nu^2 = E(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon - \sigma_\varepsilon^2)^2 + o_P(1)$$

and $\hat{\sigma}_\xi^2 = E\{\ \beta^{*\mathbb{T}}(\tilde{V}\tilde{V}^{\mathbb{T}} - \tilde{C})\beta^*\}^2 + o_P(1)$. Recall (1.14), $\xi_i^{(k)} = \ \hat{\beta}^{(-k)^{\mathbb{T}}}\left(\hat{V}_i\hat{V}_i^{\mathbb{T}} - \hat{C}^{(k)}\right)\hat{\beta}^{(-k)}$. Now define $\xi_i = \ \beta^{*\mathbb{T}}(\tilde{V}_i\tilde{V}_i^{\mathbb{T}} - \tilde{C})\beta^*$. Observe that by algebraic manipulation followed by a Cauchy - Schwarz inequality

$$\begin{aligned} \left|N^{-1}\sum_{i\in I_k}\xi_i^{(k)^2} - N^{-1}\sum_{i\in I_k}\xi_i^2\right| &= \left|N^{-1}\sum_{i\in I_k}(\xi_i^{(k)} - \xi_i)^2 + 2N^{-1}\sum_{i\in I_k}\xi_i(\xi_i^{(k)} - \xi_i)\right|\\ &\le \left|N^{-1}\sum_{i\in I_k}(\xi_i^{(k)} - \xi_i)^2\right| + 2\left\{N^{-1}\sum_{i\in I_k}\xi_i^2 N^{-1}\sum_{i\in I_k}(\xi_i^{(k)} - \xi_i)^2\right\}^{1/2}. \qquad (1.55) \end{aligned}$$

By Condition 1.2, $E|\ \beta^{*\mathbb{T}}(\tilde{V}\tilde{V}^{\mathbb{T}} - \tilde{C})\beta^*|^{2+c} < c_1$. By Lemma 1.3,

$$N^{-1}\sum_{i\in I_k}\xi_i^2 = E\{\ \beta^{*\mathbb{T}}(\tilde{V}\tilde{V}^{\mathbb{T}} - \tilde{C})\beta^*\}^2 + o_P(1).$$

Now, we need to prove $N^{-1}\sum_{i\in I_k}(\xi_i^{(k)}-\xi_i)^2 = o_P(1)$. Observe that

$$\xi_i^{(k)}-\xi_i = \left\{(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i)^2 - (\ \beta^{*\mathbb{T}}\tilde{V}_i)^2\right\} - \left[\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)} - b^2\right].$$

It suffices to show

$$N^{-1}\sum_{i\in I_k}\left\{(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V})^2 - (\ \beta^{*\mathbb{T}}\tilde{V})^2\right\}^2 = o_P(1), \qquad \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)} - b^2 = o_P(1).$$

By (1.46) and (1.47), we can see that $\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)} - b^2 = o_P(1)$.

New, we consider $a_i = \beta^{*\mathbb{T}}\tilde{V}_i$ and $\Delta_i = \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \beta^{*\mathbb{T}}\tilde{V}_i$. Then

$$\begin{aligned} N^{-1}\sum_{i\in I_k}\left[(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i)^2 - (\ \beta^{*\mathbb{T}}\tilde{V}_i)^2\right]^2 &= N^{-1}\sum_{i\in I_k}\Delta_i^2(2a_i+\Delta_i)^2 \\ &= N^{-1}\sum_{i\in I_k}(\Delta_i^4 + 4a_i\Delta_i^3 + 4a_i^2\Delta_i^2). \end{aligned}$$

By Condition 1.2 and 1.4, $N^{-1}\sum_{i\in I_k} a_i^4 = E(\ \beta^{*\mathbb{T}}\tilde{V})^4 + o_P(1)$ with $E(\ \beta^{*\mathbb{T}}\tilde{V})^4 < c_1$ and by the fact that $\{(a+b+c)/3\}^4 \le (a^4+b^4+c^4)/3$,

$$\begin{aligned} &N^{-1}\sum_{i\in I_k}\Delta_i^4 \qquad (1.56)\\ &= N^{-1}\sum_{i\in I_k}\left\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i - \beta^{*\mathbb{T}}(\hat{\mu}^{(k)}-\tilde{\mu}) - (\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}(\hat{\mu}^{(k)}-\tilde{\mu})\right\}^4 \\ &\le 27N^{-1}\sum_{i\in I_k}\left[\left\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\right\}^4 + 27\left\{\ \beta^{*\mathbb{T}}(\hat{\mu}^{(k)}-\tilde{\mu})\right\}^4 + 27\left\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}(\hat{\mu}^{(k)}-\tilde{\mu})\right\}^4\right] \\ &= o_P(1), \qquad (1.57) \end{aligned}$$

here $N^{-1}\sum_{i\in I_k}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}_i\}^4 = o(1)$ results from $E_{I_k^c}\{(\hat{\beta}^{(-k)}-\beta^*)^{\mathbb{T}}\tilde{V}\}^4 = o(1)$ and Lemma 1.1. Hence, by Holder's Inequality,

$$N^{-1}\sum_{i\in I_k}\left\{(\ \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i)^2 - (\ \beta^{*\mathbb{T}}\tilde{V}_i)^2\right\}^2 = o_P(1).$$

Therefore, $N^{-1}\sum_{i\in I_k}(\xi_i^{(k)}-\xi_i)^2=o_P(1)$ and

$$N^{-1}\sum_{i\in I_k}\xi_i^{(k)^2}=N^{-1}\sum_{i\in I_k}\xi_i^2+o_P(1)=E\{\ \beta^{*\mathbb{T}}(\tilde{V}\tilde{V}^{\mathbb{T}}-\tilde{C})\beta^*\}^2+o_P(1).$$

When $K<\infty$,

$$\hat{\sigma}_\xi^2=K^{-1}\sum_{k=1}^{K}N^{-1}\sum_{i\in I_k}\xi_i^{(k)^2}=E\{\ \beta^{*\mathbb{T}}(\tilde{V}\tilde{V}^{\mathbb{T}}-\tilde{C})\beta^*\}^2+o_P(1).$$

To show $\hat{\sigma}_\nu^2=E(\varepsilon^2+2\beta^{*\mathbb{T}}\tilde{V}\varepsilon-\sigma_\varepsilon^2)^2+o_P(1)$, recall (1.16), $\nu_i^{(k)}=\hat{\varepsilon}_i^2+2\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\hat{\varepsilon}_i+\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)}-\hat{\sigma}_Y^2$, where $\hat{\varepsilon}_i=Y_i-\hat{\theta}-\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i$. Define $\nu_i=\varepsilon_i^2+2\beta^{*\mathbb{T}}\tilde{V}_i\varepsilon_i-\sigma_\varepsilon^2$. Similarly as in (1.55),

$$\begin{aligned}&\left|N^{-1}\sum_{i\in I_k}\nu_i^{(k)^2}-N^{-1}\sum_{i\in I_k}\nu_i^2\right|\\ \le&\left|N^{-1}\sum_{i\in I_k}(\nu_i^{(k)}-\nu_i)^2\right|+2\left\{N^{-1}\sum_{i\in I_k}\nu_i^2N^{-1}\sum_{i\in I_k}(\nu_i^{(k)}-\nu_i)^2\right\}^{1/2}.\end{aligned}$$

By Condition 1.2, $E|\varepsilon^2+2\beta^{*\mathbb{T}}\tilde{V}\varepsilon-\sigma_\varepsilon^2|^{2+c}<c_1$, and by Lemma 1.3,

$$N^{-1}\sum_{i\in I_k}\nu_i^2=E\left(\varepsilon^2+2\beta^{*\mathbb{T}}\tilde{V}\varepsilon-\sigma_\varepsilon^2\right)^2+o_P(1).$$

Now it remains to prove $N^{-1}\sum_{i\in I_k}(\nu_i^{(k)}-\nu_i)^2=o_P(1)$. It suffices to show

$$N^{-1}\sum_{i\in I_k}(\hat{\varepsilon}_i^2-\varepsilon_i^2)^2=o_P(1),\tag{1.58}$$

$$N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\hat{\varepsilon}_i-\beta^{*\mathbb{T}}\tilde{V}_i\varepsilon_i)^2=o_P(1),\tag{1.59}$$

$$\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)}-\hat{\sigma}_Y^2+\sigma_\varepsilon^2=o_P(1).\tag{1.60}$$

Recall that from (1.43), (1.46) and (1.47), we have $\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{C}^{(k)}\hat{\beta}^{(-k)}-b^2=o_P(1)$, $\hat{\sigma}_Y^2=\sigma_Y^2+o_P(1)$ and hence (1.60) holds. As for (1.58),

$$N^{-1}\sum_{i\in I_k}(\hat{\varepsilon}_i^2-\varepsilon_i^2)^2=N^{-1}\sum_{i\in I_k}(\hat{\varepsilon}_i-\varepsilon_i)^2\{(\hat{\varepsilon}_i-\varepsilon_i)+2\varepsilon_i\}^2.\tag{1.61}$$

By Condition 1.2, $E|\varepsilon|^{4+c} < c_1$. By Lemma 1.3,

$$N^{-1}\sum_{i\in I_k}\varepsilon_i^4 = E(\varepsilon^4) + o_P(1) \tag{1.62}$$

with $E(\varepsilon^4) < c_1$. Besides,

$$\begin{aligned} N^{-1}\sum_{i\in I_k}(\hat{\varepsilon}_i - \varepsilon_i)^4 &= N^{-1}\sum_{i\in I_k}(\hat{\theta} - \theta + \hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \ \beta^{*\mathbb{T}}\tilde{V}_i)^4 \\ &\leq 8(\hat{\theta}-\theta)^4 + 8N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \ \beta^{*\mathbb{T}}\tilde{V}_i)^4 = o_P(1), \end{aligned} \tag{1.63}$$

where the last equality results from (1.36) and (1.57). By (1.61), (1.62), (1.63) and Holder's Inequality, (1.58) holds. Now, for (1.59),

$$\begin{aligned} &N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i\hat{\varepsilon}_i - \beta^{*\mathbb{T}}\tilde{V}_i\varepsilon_i)^2 \\ &= N^{-1}\sum_{i\in I_k}\left\{(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \ \beta^{*\mathbb{T}}\tilde{V}_i)\varepsilon_i + \ \beta^{*\mathbb{T}}\tilde{V}_i(\hat{\varepsilon}_i - \varepsilon_i) + (\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \ \beta^{*\mathbb{T}}\tilde{V}_i)(\hat{\varepsilon}_i - \varepsilon_i)\right\}^2. \end{aligned}$$

Since

$$N^{-1}\sum_{i\in I_k}\varepsilon_i^4 = E\varepsilon^4 + o_P(1), \quad N^{-1}\sum_{i\in I_k}(\ \beta^{*\mathbb{T}}\tilde{V}_i)^4 = E(\ \beta^{*\mathbb{T}}\tilde{V})^4 + o_P(1),$$

$$N^{-1}\sum_{i\in I_k}(\hat{\varepsilon}_i - \varepsilon_i)^4 = o_P(1), \quad N^{-1}\sum_{i\in I_k}(\hat{\beta}^{(-k)^{\mathbb{T}}}\hat{V}_i - \ \beta^{*\mathbb{T}}\tilde{V}_i)^4 = o_P(1),$$

by Holder's Inequality, (1.59) holds. Now combining (1.58), (1.59) and (1.60), we have $N^{-1}\sum_{i\in I_k}(\nu_i^{(k)} - \nu_i)^2 = o_P(1)$ and hence

$$N^{-1}\sum_{i\in I_k}\nu_i^{(k)^2} = N^{-1}\sum_{i\in I_k}\nu_i^2 + o_P(1) = E\left(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon - \sigma_\varepsilon^2\right)^2 + o_P(1).$$

When $K < \infty$,

$$\hat{\sigma}_\nu^2 = K^{-1}\sum_{k=1}^{K}N^{-1}\sum_{i\in I_k}\nu_i^{(k)^2} = E\left(\varepsilon^2 + 2\beta^{*\mathbb{T}}\tilde{V}\varepsilon - \sigma_\varepsilon^2\right)^2 + o_P(1).$$

Therefore, $\hat{\sigma}_\nu^2 + n\hat{\sigma}_\xi^2/(m+n) = \mathrm{var}(\varepsilon^2 + 2\ \beta^{*\mathbb{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\ \beta^{*\mathbb{T}}\tilde{V})^2 + o_P(1)$. ∎

Proof of Corollary 1.1. When Conditions 1.1 and 1.4 hold, we have consistency results (1.42) and (1.50). By Slutsky's Theorem, for each $k \leq K$,

$$\hat{b}^{2^{(k)}}/\hat{\sigma}_Y^{2^{(k)}} = b^2/\sigma_Y^2 + o_p(1) = PVE + o_p(1)$$

and hence $R^2 = PVE + o_p(1)$.

The asymptotic normality result holds as a consequence of Theorem 1.6. ∎

Proof of Theorem 1.5. This proof provides an asymptotic normal result for $n^{1/2}(\hat{\theta}_{\text{gen}} - \theta)$ and the consistency of the asymptotic variance estimator. We first show the asymptotic normality. Let

$$\hat{\theta}_{\text{gen}}^{(k)} = M^{-1} \sum_{i \in J_k} \hat{g}^{(-k)}(X_i) + N^{-1} \sum_{i \in I_k} \left\{ Y_i - \hat{g}^{(-k)}(X_i) \right\},$$

where $M = (m+n)/K$ and $N = n/K$, then $\hat{\theta}_{\text{gen}} = K^{-1} \sum_{k=1}^{K} \hat{\theta}_{\text{gen}}^{(k)}$. Observe that

$$\begin{aligned} \hat{\theta}_{\text{gen}}^{(k)} = M^{-1} \sum_{i \in J_k} g^*(X_i) + N^{-1} \sum_{i \in I_k} \varepsilon_i + M^{-1} \sum_{i \in J_k} \left\{ \hat{g}^{(-k)}(X_i) - g^*(X_i) \right\} \\ - N^{-1} \sum_{i \in I_k} \left\{ \hat{g}^{(-k)}(X_i) - g^*(X_i) \right\}. \end{aligned}$$

By Lemma 1.1,

$$M^{-1} \sum_{i \in J_k} \left\{ \hat{g}^{(-k)}(X_i) - g^*(X_i) \right\} = o_P(M^{-1/2}),$$

$$N^{-1} \sum_{i \in I_k} \left\{ \hat{g}^{(-k)}(X_i) - g^*(X_i) \right\} = o_P(N^{-1/2}).$$

Hence, when $K < \infty$,

$$\hat{\theta}_{\text{gen}} = (m+n)^{-1} \sum_{i=1}^{m+n} g^*(X_i) + n^{-1} \sum_{i=1}^{n} \varepsilon_i + o_P(n^{-1/2}).$$

By Lindeberg-Feller Central Limit Theorem, as $m, n, p \to \infty$,

$$\frac{n^{1/2}\left\{(m+n)^{-1}\sum_{i=1}^{m+n} g^*(X_i) + n^{-1}\sum_{i=1}^{n}\varepsilon_i\right\}}{\mathrm{var}(\varepsilon) + \tau \mathrm{var}\{g^*(X)\}} \to N(0,1)$$

and hence

$$\frac{n^{1/2}(\hat{\theta}_{\mathrm{gen}} - \theta)}{\sigma^2_{\varepsilon,\mathrm{gen}} + \frac{n}{m+n} b^2_{\mathrm{gen}}} \to N(0,1).$$

Now, we showcase that $\hat{b}^2_{\mathrm{gen}}$ and $\hat{\sigma}^2_{\varepsilon,\mathrm{gen}}$ are consistent estimators of b^2_{gen} and $\sigma^2_{\varepsilon,\mathrm{gen}}$, respectively.

For $k \in K$ and $i \in I_k$, let

$$\nu_{\varepsilon,\mathrm{gen},i} = Y_i - \hat{\theta}_{\mathrm{gen}} - \hat{g}^{(-k)}(X_i) + M^{-1}\sum_{i \in J_k} \hat{g}^{(-k)}(X_i).$$

Then,

$$\begin{aligned}
&\left| N^{-1}\sum_{i\in I_k} \nu^2_{\varepsilon,\mathrm{gen},i} - N^{-1}\sum_{i\in I_k}\varepsilon_i^2 \right| \\
&\leq N^{-1}\sum_{i\in I_k}(\nu_{\varepsilon,\mathrm{gen},i} - \varepsilon_i)^2 + 2\left\{ N^{-1}\sum_{i\in I_k}\varepsilon_i^2 N^{-1}\sum_{i\in I_k}(\nu_{\varepsilon,\mathrm{gen},i} - \varepsilon_i)^2 \right\}^{1/2},
\end{aligned}$$

where $N^{-1}\sum_{i\in I_k}\varepsilon_i^2 = \sigma^2_{\varepsilon,\mathrm{gen}} + o_P(1)$. Besides,

$$\nu_{\varepsilon,\mathrm{gen},i} - \varepsilon_i = -(\hat{\theta}_{\mathrm{gen}} - \theta) - \left\{ \hat{g}^{(-k)}(X_i) - M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i) - g^*(X_i) + \theta \right\},$$

where $\hat{\theta}_{\mathrm{gen}} - \theta = o_P(1)$ and by Lemma 1.1,

$$\sum_{i\in I_k}\left\{ \hat{g}^{(-k)}(X_i) - M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i) - g^*(X_i) + \theta \right\} = o_P(1).$$

Hence,

$$N^{-1}\sum_{i\in I_k}\nu^2_{\varepsilon,\mathrm{gen},i} = N^{-1}\sum_{i\in I_k}\varepsilon_i^2 + o_P(1) = \sigma^2_{\varepsilon,\mathrm{gen}} + o_P(1)$$

and therefore, $\sigma^2_{\varepsilon,\mathrm{gen}} = \sigma^2_{\varepsilon,\mathrm{gen}} + o_P(1)$. Similarly, $b^2_{\mathrm{gen}} = b^2_{\mathrm{gen}} + o_P(1)$. ∎

Proof of Theorem 1.6. This proof provides asymptotic normalities for the variance, explained variance and unexplained variance estimators. We first work on the explained variance and the unexplained variance. With a slight abuse of notation, define

$$\begin{aligned}
\hat{b}^{2^{(k)}} &= M^{-1}\sum_{i\in J_k}\{\hat{h}^{(-k)}(X_i)\}^2 + 2N^{-1}\sum_{i\in I_k}\hat{h}^{(-k)}(X_i)\{Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i)\},\\
\tilde{b}^{2^{(k)}} &= M^{-1}\sum_{i\in J_k}\{\hat{g}^{(-k)}(X_i) - \mu^{(-k)}\}^2\\
&\qquad + 2N^{-1}\sum_{i\in I_k}\{\hat{g}^{(-k)}(X_i) - \mu^{(-k)}\}\{Y_i - \theta - \hat{g}^{(-k)}(X_i) + \mu^{(-k)}\},\\
\check{b}^{2^{(k)}} &= M^{-1}\sum_{i\in J_k}\{\tilde{g}^{*}(X_i)\}^2 + 2N^{-1}\sum_{i\in I_k}\varepsilon_i\tilde{g}^{*}(X_i),\\
\hat{\sigma}_{\varepsilon}^{2^{(k)}} &= N^{-1}\sum_{i\in I_k}\{Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i)\}^2,\\
\tilde{\sigma}_{\varepsilon}^{2^{(k)}} &= N^{-1}\sum_{i\in I_k}\{Y_i - \theta - \hat{g}^{(-k)}(X_i) + \mu^{(-k)}\}^2,\\
\check{\sigma}_{\varepsilon}^{2^{(k)}} &= N^{-1}\sum_{i\in I_k}\varepsilon_i^2.
\end{aligned}$$

where $\hat{h}^{(-k)}(X_i) = \hat{g}^{(-k)}(X_i) - M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i)$, $\mu^{(-k)} = E_{I_k^c}\{\hat{g}^{(-k)}(X)\}$ and $\tilde{g}^{*}(X_i) = g^{*}(X_i) - \theta$. The proof consists of 3 steps:

Step 1: $\hat{b}^{2^{(k)}} = \tilde{b}^{2^{(k)}} + o_P(N^{-1})$, $\hat{\sigma}_{\varepsilon}^{2^{(k)}} = \tilde{\sigma}_{\varepsilon}^{2^{(k)}} + o_P(N^{-1})$.

Step 2: $\tilde{b}^{2^{(k)}} = \check{b}^{2^{(k)}} + o_P(N^{-\frac{1}{2}})$, $\tilde{\sigma}_{\varepsilon}^{2^{(k)}} = \check{\sigma}_{\varepsilon}^{2^{(k)}} + o_P(N^{-\frac{1}{2}})$.

Step 3: $n^{1/2}\{V(\sigma^2_{\varepsilon,\text{gen}})\}^{-1/2}(\hat{\sigma}^2_{\varepsilon,\text{gen}} - \sigma^2_{\varepsilon,\text{gen}}) \to N(0,1)$, $n^{1/2}\{V(b^2_{\text{gen}})\}^{-1/2}(\hat{b}^2_{\text{gen}} - b^2_{\text{gen}}) \to N(0,1)$.

Step 1. Let $\Delta_1 = M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i) - \mu^{(-k)}$, $\Delta_2 = \hat{\theta}_{\text{gen}} - \theta$ and $\delta_i = \hat{g}^{(-k)}(X_i) - \mu^{(-k)} - \tilde{g}^{*}(X_i)$. Then, $\Delta_2 = O_P(n^{-1/2})$, and

$$\Delta_1 = M^{-1}\sum_{i\in J_k}\delta_i + M^{-1}\sum_{i\in J_k}\tilde{g}^{*}(X_i).$$

By Lemma 1.1 and 1.2,

$$M^{-1}\sum_{i\in J_k}\delta_i = o_P(M^{-1/2}), \qquad M^{-1}\sum_{i\in J_k}\tilde{g}^*(X_i) = O_P(M^{-1/2})$$

and hence $\Delta_1 = O_P(M^{-1/2})$. Observe that

$$\hat{b}^{2^{(k)}} = \tilde{b}^{2^{(k)}} - \Delta_1^2 + \Delta_1(\Delta_2 - \Delta_1) - \Delta_1 N^{-1}\sum_{i\in I_k}\left\{Y_i - \theta - \hat{g}^{(-k)}(X_i) + \mu^{(-k)}\right\}$$

$$+ (\Delta_1 - \Delta_2)N^{-1}\sum_{i\in I_k}\left\{\hat{g}^{(-k)}(X_i) - \mu^{(-k)}\right\},$$

where by Lemma 1.1 and Lemma 1.3,

$$N^{-1}\sum_{i\in I_k}\left\{Y_i - \theta - \hat{g}^{(-k)}(X_i) + \mu^{(-k)}\right\} = o_P(1), \quad N^{-1}\sum_{i\in I_k}\left\{\hat{g}^{(-k)}(X_i) - \mu^{(-k)}\right\} = o_P(1).$$

Therefore,

$$\hat{b}^{2^{(k)}} = \tilde{b}^{2^{(k)}} + o_P(N^{-1/2}). \tag{1.64}$$

Besides,

$$\hat{\sigma}_\varepsilon^{2^{(k)}} = \tilde{\sigma}_\varepsilon^{2^{(k)}} - 2(\Delta_1 - \Delta_2)N^{-1}\sum_{i\in I_k}(\varepsilon_i - \delta_i) + (\Delta_1 - \Delta_2)^2 = \tilde{\sigma}_\varepsilon^{2^{(k)}} + o_P(N^{-1/2}).$$

Step 2. Observe that

$$\tilde{b}^{2^{(k)}} = \check{b}^{2^{(k)}} + M^{-1}\sum_{i\in J_k}\delta_i\{\delta_i + 2\tilde{g}^*(X_i)\} + 2N^{-1}\sum_{i\in I_k}\delta_i\{\varepsilon_i - \tilde{g}^*(X_i) - \delta_i\},$$

$$\tilde{\sigma}_\varepsilon^{2^{(k)}} = \check{\sigma}_\varepsilon^{2^{(k)}} - 2N^{-1}\sum_{i\in I_k}\varepsilon_i\delta_i + N^{-1}\sum_{i\in I_k}\delta_i^2.$$

By Lemma 1.1,

$$M^{-1}\sum_{i\in J_k}\delta_i\{\delta_i+2\tilde{g}^*(X_i)\}=E_{I_k^c}(\delta^2)+2E_{I_k^c}\{\delta\tilde{g}^*(X)\}+o_P(M^{-\frac{1}{2}}),$$

$$N^{-1}\sum_{i\in I_k}\delta_i\{\varepsilon_i-\tilde{g}^*(X_i)-\delta_i\}=E_{I_k^c}(\delta\varepsilon)-E_{I_k^c}\{\delta\tilde{g}^*(X)\}-E_{I_k^c}(\delta^2)+o_P(N^{-\frac{1}{2}}),$$

$$N^{-1}\sum_{i\in I_k}\varepsilon_i\delta_i=E_{I_k^c}(\delta\varepsilon)+o_P(N^{-\frac{1}{2}}),$$

$$N^{-1}\sum_{i\in I_k}\delta_i^2=E_{I_k^c}(\delta^2)+o_P(N^{-\frac{1}{2}}).$$

Hence,

$$\tilde{b}^{2^{(k)}}=\check{b}^{2^{(k)}}+2E_{I_k^c}(\delta\varepsilon)-E_{I_k^c}(\delta^2)+o_P(N^{-\frac{1}{2}}), \tag{1.65}$$

$$\tilde{\sigma}_\varepsilon^{2^{(k)}}=\check{\sigma}_\varepsilon^{2^{(k)}}-2E_{I_k^c}(\delta\varepsilon)+E_{I_k^c}(\delta^2)+o_P(N^{-\frac{1}{2}}). \tag{1.66}$$

By assuming $E_{I_k^c}(\delta\varepsilon)=o_P(N^{-\frac{1}{2}})$ and $E_{I_k^c}(\delta^2)=o_P(N^{-\frac{1}{2}})$, we have

$$\tilde{b}^{2^{(k)}}=\check{b}^{2^{(k)}}+o_P(N^{-\frac{1}{2}}),\qquad \tilde{\sigma}_\varepsilon^{2^{(k)}}=\check{\sigma}_\varepsilon^{2^{(k)}}+o_P(N^{-\frac{1}{2}}). \tag{1.67}$$

Step 3. Observe that

$$K^{-1}\sum_{k=1}^{K}\check{b}^{2^{(k)}}=2n^{-\frac{1}{2}}\sum_{i=1}^{n}\varepsilon_i\tilde{g}^*(X_i)+(m+n)^{-1}\sum_{i=1}^{n}\{\tilde{g}^*(X_i)\}^2+(m+n)^{-1}\sum_{i=n+1}^{m+n}\{\tilde{g}^*(X_i)\}^2,$$
$$K^{-1}\sum_{k=1}^{K}\check{\sigma}_\varepsilon^{2^{(k)}}=\sum_{i=1}^{n}\varepsilon_i^2.$$

By Lindeberg-Feller Central Limit Theorem, as $m,n,p\to\infty$,

$$\frac{n^{1/2}(K^{-1}\sum_{k=1}^{K}\check{b}^{2^{(k)}}-b_{\text{gen}}^2)}{\{V(b_{\text{gen}}^2)\}^{1/2}}\to N(0,1),\quad \frac{n^{1/2}(K^{-1}\sum_{k=1}^{K}\check{\sigma}_\varepsilon^2-\sigma_{\varepsilon,\text{gen}}^2)}{\{V(\sigma_{\varepsilon,\text{gen}}^2)\}^{1/2}}\to N(0,1).$$

Hence,

$$n^{1/2}\{V(b_{\text{gen}}^2)\}^{-1/2}(\hat{b}_{\text{gen}}^2-b_{\text{gen}}^2)\to N(0,1),\quad n^{1/2}\{V(\sigma_{\varepsilon,\text{gen}}^2)\}^{-1/2}(\hat{\sigma}_{\varepsilon,\text{gen}}^2-\sigma_{\varepsilon,\text{gen}}^2)\to N(0,1).$$

Now, we show the asymptotic normal result for the variance estimator. Recall from (1.65) and (1.66),

$$\begin{aligned}
\hat{\sigma}^2_{Y,\text{gen}} &= \hat{b}^2_{\text{gen}} + \hat{\sigma}^2_{\varepsilon,\text{gen}} \\
&= K^{-1}\sum_{k=1}^{K}\left\{\check{b}^{2^{(k)}} + 2E_{I_k^c}(\delta\varepsilon) - E_{I_k^c}(\delta^2) + \check{\sigma}_{\varepsilon}^{2^{(k)}} - 2E_{I_k^c}(\delta\varepsilon) + E_{I_k^c}(\delta^2) + o_P(N^{-\frac{1}{2}})\right\} \\
&= K^{-1}\sum_{k=1}^{K}\left\{\check{b}^{2^{(k)}} + \check{\sigma}_{\varepsilon}^{2^{(k)}} + o_P(N^{-\frac{1}{2}})\right\},
\end{aligned}$$

where the bias terms $2E_{I_k^c}(\delta\varepsilon)$ and $E_{I_k^c}(\delta^2)$ canceled out. By Lindeberg-Feller Central Limit Theorem and Slutsky's Theorem, as $m, n, p \to \infty$,

$$\frac{n^{1/2}(\hat{\sigma}^2_{Y,\text{gen}} - \sigma^2_Y)}{\{V(\sigma^2_Y)\}^{1/2}} \to N(0,1).$$

Lastly, for the PVE estimation, by Step 1 and 2, we showcase that $\hat{b}^{2^{(k)}} = \check{b}^{2^{(k)}} + o_P(N^{-1/2})$ and $\hat{\sigma}_Y^{2^{(k)}} = \check{\sigma}_Y^{2^{(k)}} + o_P(N^{-1/2})$, where $\check{\sigma}_Y^{2^{(k)}} = \check{\sigma}_{\varepsilon}^{2^{(k)}} + \check{b}^{2^{(k)}}$. Besides, we also have $\hat{\sigma}_Y^{2^{(k)}} = \sigma^2_Y + o_P(1)$ by Lemma 1.1. Hence, for each $k \le K$,

$$\begin{aligned}
n^{1/2}\left(\frac{\hat{b}^{2^{(k)}}}{\hat{\sigma}_Y^{2^{(k)}}} - \frac{b^2_{\text{gen}}}{\sigma^2_Y}\right) &= n^{1/2}\left\{\frac{\sigma^2_Y(\hat{b}^{2^{(k)}} - b^2_{\text{gen}}) - b^2_{\text{gen}}(\hat{\sigma}_Y^{2^{(k)}} - \sigma^2_Y)}{\sigma^2_Y\hat{\sigma}_Y^{2^{(k)}}}\right\} \\
&= n^{1/2}\left[\frac{\sigma^2_Y\{\check{b}^{2^{(k)}} - b^2_{\text{gen}} + o_P(N^{-1/2})\} - b^2_{\text{gen}}\{\check{\sigma}_Y^{2^{(k)}} - \sigma^2_Y + o_P(N^{-1/2})\}}{\sigma^2_Y\{\sigma^2_Y + o_P(1)\}}\right] \\
&= n^{1/2}\sigma_Y^{-2}(\check{b}^{2^{(k)}} - b^2_{\text{gen}}) - n^{1/2}\sigma_Y^{-4}b^2_{\text{gen}}(\check{\sigma}_Y^{2^{(k)}} - \sigma^2_Y) + o_P(1).
\end{aligned}$$

It follows that

$$
\begin{aligned}
&n^{1/2}(R^2_{\text{gen}} - PVE) \\
&\quad = n^{1/2}\sigma_Y^{-2}K^{-1}\sum_{k=1}^{K}(\check{b}^{2^{(k)}} - b^2_{\text{gen}}) - n^{1/2}\sigma_Y^{-4}b^2_{\text{gen}}K^{-1}\sum_{k=1}^{K}(\check{\sigma}_Y^{2^{(k)}} - \sigma_Y^2) + o_P(1) \\
&\quad = n^{1/2}\sum_{i=1}^{n}\left(\sigma_Y^{-4}\sigma^2_{\varepsilon,\text{gen}}[\{\tilde{g}^*(X_i)\}^2 + \varepsilon_i\tilde{g}^*(X_i) - b^2_{\text{gen}}] - \sigma_Y^{-4}b^2_{\text{gen}}(\varepsilon_i^2 - \sigma_\varepsilon^2)\right) \\
&\qquad + n^{1/2}\sum_{i=n+1}^{m+n}\sigma_Y^{-4}\sigma^2_{\varepsilon,\text{gen}}[\{\tilde{g}^*(X_i)\}^2 - b^2_{\text{gen}}].
\end{aligned}
$$

By Lindeberg-Feller Central Limit Theorem,

$$n^{1/2}V^{-1/2}(R^2_{\text{gen}})(R^2_{\text{gen}} - \text{PVE}) \to N(0,1).$$

■

Proof of Theorem 1.7. This proof provides an asymptotic normal result for $n^{1/2}(\hat{\theta}_{\text{MAR}} - \theta_{\text{MAR}})$. Assume the following rates

$$E_{J_k^c}\{\hat{g}^{(-k)}(X) - g^0(X)\}^2 = O_P(a_{m+n,p}), \qquad E_{J_k^c}\{1 - s^*(X)/\hat{s}^{(-k)}(X)\}^2 = O_P(b_{m+n,p}).$$

By definition, the proposed estimator $\hat{\theta}_{\text{MAR}}$ can be rewritten as

$$\hat{\theta}_{\text{MAR}} = K^{-1}\sum_{k=1}^{K}\hat{\theta}^{(k)}_{\text{MAR}},$$

with

$$\hat{\theta}^{(k)}_{\text{MAR}} = M^{-1}\sum_{i\in J_k}\left[g^{(-k)}(X_i) + \frac{T_i\{Y_i^o - g^{(-k)}(X_i)\}}{\hat{s}^{(-k)}(X_i)}\right],$$

where $M = |J_k| = (m+n)/K$. Recall that for each i,

$$Y_i = g^0(X_i) + \varepsilon_i, \qquad T_i = s^*(X_i) + r_i.$$

Hence,

$$\begin{aligned}
&\hat{\theta}^{(k)}_{\mathrm{MAR}} - \theta \\
&\quad = M^{-1} \sum_{i \in J_k} \Bigg[g^0(X_i) + \frac{T_i \varepsilon_i}{s^*(X_i)} + T_i \varepsilon_i \left\{ \frac{1}{\hat{s}^{(-k)}(X_i)} - \frac{1}{s^*(X_i)} \right\} \\
&\qquad - \frac{r_i \left\{ \hat{g}^{(-k)}(X_i) - g^0(X_i) \right\}}{s^*(X_i)} - T_i \left\{ \hat{g}^{(-k)}(X_i) - g^0(X_i) \right\} \left\{ \frac{1}{\hat{s}^{(-k)}(X_i)} - \frac{1}{s^*(X_i)} \right\} \Bigg].
\end{aligned}$$

Since Y_i and T_i are independent conditional on X_i, the expectations of the terms on RHS are

$$\begin{aligned}
&E\left\{ g^0(X) + \frac{T\varepsilon}{s^*(X)} \right\} = \theta, \qquad E_{J_k^c}\left[T\varepsilon \left\{ \frac{1}{\hat{s}^{(-k)}(X)} - \frac{1}{s^*(X)} \right\} \right] = 0, \\
&E_{J_k^c}\left[\frac{r \left\{ \hat{g}^{(-k)}(X) - g^0(X) \right\}}{s^*(X)} \right] = 0.
\end{aligned}$$

Now, since $E_{J_k^c}(T \mid X) = s^*(X)$ and by the tower property of the conditional expectations, we have

$$\begin{aligned}
&E_{J_k^c}\left| T \left\{ \hat{g}^{(-k)}(X) - g^0(X) \right\} \left\{ \frac{1}{\hat{s}^{(-k)}(X)} - \frac{1}{s^*(X)} \right\} \right| \\
&= E_{J_k^c}\left\{ \left| \hat{g}^{(-k)}(X) - g^0(X) \right| \cdot \left| \frac{1}{\hat{s}^{(-k)}(X)} - \frac{1}{s^*(X)} \right| E_{J_k^c}(T \mid X) \right\} \\
&= E_{J_k^c}\left| \left\{ \hat{g}^{(-k)}(X) - g^0(X) \right\} \frac{s^*(X) - \hat{s}^{(-k)}(X)}{\hat{s}^{(-k)}(X)} \right|.
\end{aligned}$$

Now, by simple Holder's inequality and following the assumptions, the above is of the order of $O_P(a_{m+n,p} b_{m+n,p})$.

As for the the second moments, with similar reasoning, we have

$$\begin{aligned}
&E_{J_k^c}\left[T\varepsilon \left\{ \frac{1}{\hat{s}^{(-k)}(X)} - \frac{1}{s^*(X)} \right\} \right]^2 = E_{J_k^c}\left[\left\{ \frac{\hat{s}^{(-k)}(X) - s^*(X)}{\hat{s}^{(-k)}(X) s^*(X)} \right\}^2 E_{J_k^c}(T\varepsilon^2 \mid X) \right] \\
&= E_{J_k^c}\left[\left\{ \frac{\hat{s}^{(-k)}(X) - s^*(X)}{\hat{s}^{(-k)}(X) s^*(X)} \right\}^2 E_{J_k^c}(T \mid X) E_{J_k^c}(\varepsilon^2 \mid X) \right] \\
&= E_{J_k^c}\left[\frac{\left\{ \hat{s}^{(-k)}(X) - s^*(X) \right\}^2}{\left\{ \hat{s}^{(-k)}(X) \right\}^2 s^*(X)} E_{J_k^c}(\varepsilon^2 \mid X) \right] = O_P(b^2_{m+n,p} / E(T)),
\end{aligned}$$

as well as

$$E_{J_k^c}\left[\frac{r_i\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}}{s^*(X)}\right]^2 = E_{J_k^c}\left[\frac{\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}^2}{\{s^*(X)\}^2}E_{J_k^c}(r^2 \mid X)\right]$$
$$= E_{J_k^c}\left[\frac{\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}^2\{1-s^*(X)\}}{s^*(X)}\right] = O_P(a^2_{m+n,p}/E(T)).$$

By Lemma 1.2, we have

$$M^{-1}\sum_{i\in J_k} T_i\varepsilon_i\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\} = O_P(M^{-1/2}b_{m+n,p}\{E(T)\}^{-1/2})$$
$$= o_P(M^{-1/2}\{E(T)\}^{-1/2}),$$
$$M^{-1}\sum_{i\in J_k}\frac{r_i\left\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\right\}}{s^*(X_i)} = O_P(M^{-1/2}a_{m+n,p}\{E(T)\}^{-1/2})$$
$$= o_P(M^{-1/2}\{E(T)\}^{-1/2}).$$

By Lemma 1.1, and that $a_{m+n,p}b_{m+n,p} = o_P((m+n)^{-1/2}\{E(T)\}^{-1/2})$,

$$M^{-1}\sum_{i\in J_k} T_i\left\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\right\}\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\} = o_P(M^{-1/2}\{E(T)\}^{-1/2}).$$

Therefore,

$$\hat{\theta}^{(k)}_{\mathrm{MAR}} = \sum_{i\in J_k}\left\{g^0(X_i)+\frac{T_i\varepsilon_i}{s^*(X_i)}\right\} + o_P\left(M^{-1/2}\{E(T)\}^{-1/2}\right).$$

For $K<\infty$, we have

$$\hat{\theta}_{\mathrm{MAR}} = \sum_{i=1}^{m+n}\left\{g^0(X_i)+\frac{T_i\varepsilon_i}{s^*(X_i)}\right\} + o_P\left((m+n)^{-1/2}\{E(T)\}^{-1/2}\right).$$

Let $V_T = E\{T\varepsilon/s^*(X)\}^2$, and recall that $E(\varepsilon^2 \mid X) < c_1$,

$$V_T = E\left[\frac{1}{\{s^*(X)\}^2}E(T\varepsilon^2 \mid X)\right] = E\left[\frac{1}{\{s^*(X)\}^2}E(T \mid X)E(\varepsilon^2 \mid X)\right]$$
$$= E\left\{\frac{1}{s^*(X)}E(\varepsilon^2 \mid X)\right\} = E\left\{\frac{\varepsilon^2}{s^*(X)}\right\} \geq \frac{\{E(\varepsilon^2)\}^2}{E\{s^*(X)\varepsilon^2\}} > \frac{\{E(\varepsilon^2)\}^2}{C_1E(T)}.$$

which implies that

$$1_{\{|\frac{T\varepsilon}{s^*(X)}|>\delta(m+n)V_T^{1/2}\}} \le 1_{\{\frac{\varepsilon^2}{c_1^2\{E(T)\}^2}>\frac{\delta^2(m+n)^2\{E(\varepsilon^2)\}^2}{C_1E(T)}\}} = 1_{\{C_1\varepsilon^2>c_1^2\delta^2(m+n)^2\{E(\varepsilon^2)\}^2E(T)\}}$$
$$\le \frac{C_1\varepsilon^2}{c_1^2\delta^2(m+n)^2\left\{E(\varepsilon^2)\right\}^2E(T)}.$$

Hence, for any $\delta > 0$,

$$\begin{aligned}
&V_T^{-1}E\left[\left\{\frac{T\varepsilon}{s^*(X)}\right\}^2 1_{\{|\frac{T\varepsilon}{s^*(X)}|>\delta(m+n)V^{1/2}\}}\right] \\
&\le \frac{C_1E(T)}{\{E(\varepsilon^2)\}^2}E\left[\frac{T\varepsilon^2}{\{s^*(X)\}^2}\cdot\left|\frac{C_1\varepsilon^2}{c_1^2\delta^2(m+n)^2\left\{E(\varepsilon^2)\right\}^2E(T)}\right|^{c/2}\right] \\
&= \frac{C_1^{1+c/2}\{E(T)\}^{1-c/2}}{c_1^c\delta^c(m+n)^c\{E(\varepsilon^2)\}^{2+c}}E\left[\frac{|\varepsilon|^{2+c}}{\{s^*(X)\}^2}E(T\mid X)\right] \\
&= \frac{C_1^{1+c/2}\{E(T)\}^{1-c/2}}{c_1^c\delta^c(m+n)^c\{E(\varepsilon^2)\}^{2+c}}E\left\{\frac{|\varepsilon|^{2+c}}{s^*(X)}\right\} \\
&= \frac{C_1^{1+c/2}E|\varepsilon|^{2+c}}{c_1^{1+c}\delta^c(m+n)^c\{E(\varepsilon^2)\}^{2+c}\{E(T)\}^{c/2}} \to 0,
\end{aligned}$$

since $(m+n)^2E(T) \to \infty$. Therefore, by the Lindeberg Central Limit Theorem, as $m+n, p \to \infty$,

$$\{(m+n)V_T\}^{-1/2}\sum_{i=1}^{m+n}\frac{T_i\varepsilon_i}{s^*(X_i)} \to N(0,1)$$

in distribution. Besides, when $E\{g^0(X)\}^2 > C > 0$, we have

$$\frac{E|g^0(X)|^{2+c}}{[E\{g^0(X)\}^2]^{1+c/2}} < \infty,$$

by the Lindeberg-Feller Central Limit Theorem, as $m+n, p \to \infty$,

$$\left[(m+n)E\{g^0(X)\}^2\right]^{1/2}\sum_{i=1}^{m+n}\{g^0(X_i)-\theta\} \to N(0,1)$$

in distribution. Observe that

$$\operatorname{cov}\left\{g^0(X), \frac{T\varepsilon}{s^*(X)}\right\} = 0.$$

Then, by the delta method, as $m+n, p \to \infty$, we obtain

$$\left[\frac{m+n}{V_T + E\{g^0(X)\}^2}\right]^{1/2} \sum_{i=1}^{m+n} \left\{ g^0(X_i) + \frac{T_i \varepsilon_i}{s^*(X_i)} - \theta \right\} \to N(0,1), \tag{1.68}$$

in distribution. When $E\{g^0(X)\}^2 \to 0$, by Lemma 1.1, $\sum_{i=1}^{m+n}\{g^0(X_i) - \theta\} = o_P((m+n)^{-1/2})$. Since $[V_T + E\{g^0(X)\}^2]/V_T \to 1$, by the Slutsky's Theorem, (1.68) still holds. Now, recall that

$$\hat{\theta}_{\mathrm{MAR}} = \sum_{i=1}^{m+n} \left\{ g^0(X_i) + \frac{T_i \varepsilon_i}{s^*(X_i)} \right\} + R$$

where

$$R = o_P\left((m+n)^{-1/2}\{E(T)\}^{-1/2}\right).$$

Hence,

$$\frac{m+n}{V_T + E\{g^0(X)\}^2} R^2 = o_P\left(\frac{1}{E(T)[V_T + E\{g^0(X)\}^2]}\right) = o_P(1).$$

Therefore, as $m+n, p \to \infty$, the estimator $\hat{\theta}_{\mathrm{MAR}}$ is asymptotically normal

$$\left[\frac{m+n}{V_T + E\{g^0(X)\}^2}\right]^{1/2} (\hat{\theta}_{\mathrm{MAR}} - \theta) \to N(0,1).$$

Here,

$$(m+n)^{-1}[V_T + E\{g^0(X)\}^2] \le (m+n)^{-1}\left[\frac{\{E(\varepsilon^2)\}^2}{C_1 E(T)} + E\{g^0(X)\}^2\right] = O_P((m+n)^{-1}/E(T)).$$

Now we showcase that $\hat{V}_{\mathrm{MAR}}(\theta)$ is a consistent estimator of $V_{\mathrm{MAR}}(\theta) = V_T + E\{g^0(X)^2\}$. Let

$$\nu_{\theta,i} = g^{(-k)}(X_i) + \frac{T_i\{Y_i - g^{(-k)}(X_i)\}}{s^{(-k)}(X_i)} - \hat{\theta}_{\mathrm{MAR}}, \qquad \nu^*_{\theta,i} = g^0(X_i) + \frac{T_i\{Y_i - g^0(X_i)\}}{s^*(X_i)} - \theta.$$

Then, similarly as in (1.55),

$$\left| M^{-1}\sum_{i\in J_k} \nu^2_{\theta,i} - M^{-1}\sum_{i\in J_k} {\nu^*_{\theta,i}}^2 \right| \le M^{-1}\sum_{i\in J_k} \left(\nu_{\theta,i} - \nu^*_{\theta,i}\right)^2 + 2\left\{ M^{-1}\sum_{i\in J_k} {\nu^*_{\theta,i}}^2 M^{-1}\sum_{i\in J_k}\left(\nu_{\theta,i} - \nu^*_{\theta,i}\right)^2 \right\}^{1/2}.$$

We first consider the term $M^{-1}\sum_{i\in J_k}{\nu^*_{\theta,i}}^2$. Let $W_{n,i}={\nu^*_{\theta,i}}^2/V_{\mathrm{MAR}}(\theta)$. Then,

$$\begin{aligned}
nP(|W_{n,1}|>n) &\le E\left[|W_{n,1}|1_{\{|W_{n,1}|>n\}}\right] \le M^{-c/2}E|W_{n,1}|^{1+c/2} \\
&\le M^{-c/2}\frac{E|g^0(X)-\theta+T\varepsilon/s^*(X)|^{2+c}}{\{V_{\mathrm{MAR}}(\theta)\}^{1+c/2}} \\
&\le M^{-c/2}\frac{\left\{(E|g^0(X)-\theta|^{2+c})^{1/(2+c)}+(E|T\varepsilon/s^*(X)|^{2+c})^{1/(2+c)}\right\}^{2+c}}{\{V_{\mathrm{MAR}}(\theta)\}^{1+c/2}} \\
&\le M^{-c/2}\frac{E\left[|\varepsilon|^{2+c}/\{s^*(X)\}^{1+c}\right]}{\{V_{\mathrm{MAR}}(\theta)\}^{1+c/2}}+O(M^{-c/2}) \\
&\le \frac{C_1^{1+c/2}E|\varepsilon|^{2+c}}{c_1^{1+c}M^{c/2}\{E(\varepsilon^2)\}^{2+c}\{E(T)\}^{c/2}}+O(M^{-c/2})=o(1),
\end{aligned}$$

since $ME(T)\to\infty$. Besides, for any $0<c_1<2$, similarly,

$$M^{-1}E[W_{n,1}^2 1_{\{|W_{n,1}|\le M\}}] \le M^{-1}E(W_{n,1}^2|M/W_{n,1}|^{1-c/2}) = M^{-c/2}E|W_{n,1}|^{1+c/2}=o(1).$$

By general weak law of large numbers,

$$\frac{M^{-1}\sum_{i\in J_k}{\nu^*_{\theta,i}}^2}{V_{\mathrm{MAR}}(\theta)}=1+o_P(1).$$

Now, consider the term $M^{-1}\sum_{i\in J_k}(\nu_{\theta,i}-\nu^*_{\theta,i})^2$. Observe that

$$\begin{aligned}
\nu_{\theta,i}-\nu^*_{\theta,i} &= T_i\varepsilon_i\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\}-\frac{r_i\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\}}{s^*(X_i)} \\
&\quad - T_i\left\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\right\}\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\}-(\hat{\theta}_{\mathrm{MAR}}-\theta).
\end{aligned}$$

Recall that

$$\begin{aligned}
E_{J_k^c}\left[T\varepsilon\left\{\frac{1}{\hat{s}^{(-k)}(X)}-\frac{1}{s^*(X)}\right\}\right]^2 &= o_P\{1/E(T)\}, \\
E_{J_k^c}\left[\frac{r_i\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}}{s^*(X)}\right]^2 &= o_P\{1/E(T)\}.
\end{aligned}$$

Besides,

$$E_{J_k^c}\left[T\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}\left\{\frac{1}{\hat{s}^{(-k)}(X)}-\frac{1}{s^*(X)}\right\}\right]^2$$
$$=E_{J_k^c}\left[\left\{\hat{g}^{(-k)}(X)-g^0(X)\right\}^2\frac{\left\{\hat{s}^{(-k)}(X)-s^*(X)\right\}^2}{\left\{\hat{s}^{(-k)}(X)\right\}^2 s^*(X)}\right]=o_P\{1/E(T)\}.$$

By Lemma 1.1,

$$M^{-1}\sum_{i\in J_k}\left[T_i\varepsilon_i\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\}\right]^2=o_P\{1/E(T)\},$$
$$M^{-1}\sum_{i\in J_k}\left[\frac{r_i\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\}}{s^*(X_i)}\right]^2=o_P\{1/E(T)\},$$
$$M^{-1}\sum_{i\in J_k}\left[T_i\left\{\hat{g}^{(-k)}(X_i)-g^0(X_i)\right\}\left\{\frac{1}{\hat{s}^{(-k)}(X_i)}-\frac{1}{s^*(X_i)}\right\}\right]^2=o_P\{1/E(T)\}.$$

Combining with the fact that $(\hat{\theta}_{\mathrm{MAR}}-\theta)^2=o_P\{1/E(T)\}$, we have

$$M^{-1}\sum_{i\in J_k}(\nu_{\theta,i}-\nu^*_{\theta,i})^2=o_P\{1/E(T)\}.$$

Therefore,

$$M^{-1}\sum_{i\in J_k}\nu^2_{\theta,i}=M^{-1}\sum_{i\in J_k}{\nu^*_{\theta,i}}^2+o_P\{1/E(T)\}+2\left[V_{\mathrm{MAR}}\{1+o_P(1)\}o_P\{1/E(T)\}\right]^{1/2}$$
$$=V_{\mathrm{MAR}}\{1+o_P(1)\}+o_P\{1/E(T)\}$$

and hence

$$\frac{\hat{V}_{\mathrm{MAR}}}{V_{\mathrm{MAR}}}=\frac{K^{-1}\sum_{k=1}^K M^{-1}\sum_{i\in J_k}\nu^2_{\theta,i}}{V_{\mathrm{MAR}}}=1+o_P(1).$$

■

Proof of Theorem 1.8. Part 1. We first provide consistency rates of $\hat{\delta}$ and an asymptotic normal result for $n^{1/2}(\hat{\delta}-\delta)$ when some specific rates are satisfied. Recall that, by definition,

$$Y=D\ {\beta_1^*}^{\mathbb{T}}\tilde{X}+(1-D)\ {\beta_0^*}^{\mathbb{T}}\tilde{X}+\varepsilon,\quad D=e(X)+\zeta,\quad E(\zeta\mid X)=0.$$

By the definitions of β_1^* and β_0^* and Lemma 2.1 in [ZBC19],

$$E(Y - \beta_1^{*\mathbb{T}}\tilde{X} \mid D = 1) = 0,\ E(Y - \beta_0^{*\mathbb{T}}\tilde{X} \mid D = 0) = 0,$$

$$E\{(Y - \beta_1^{*\mathbb{T}}\tilde{X})\tilde{X} \mid D = 1\} = 0,\ E\{(Y - \beta_0^{*\mathbb{T}}\tilde{X})\tilde{X} \mid D = 0\} = 0.$$

Hence, $E(D\ \beta_1^{*\mathbb{T}}\tilde{X}\varepsilon) = E\{\ \beta_1^{*\mathbb{T}}\tilde{X}(Y - \beta_1^{*\mathbb{T}}\tilde{X}) \mid D = 1\}E(D) = 0$. Similarly, $E\{(1 - D)\ \beta_0^{*\mathbb{T}}\tilde{X}\varepsilon\} = 0$. Besides,

$$\begin{aligned} E(\varepsilon) &= E\{Y - D\ \beta_1^{*\mathbb{T}}\tilde{X} - (1 - D)\ \beta_0^{*\mathbb{T}}\tilde{X}\} \\ &= E(Y - \beta_1^{*\mathbb{T}}\tilde{X} \mid D = 1)E(D) - E(Y - \beta_0^{*\mathbb{T}}\tilde{X} \mid D = 0)\{1 - E(D)\} = 0. \end{aligned}$$

Therefore,

$$E(Y^2) = E\{D(\ \beta_1^{*\mathbb{T}}\tilde{X})^2\} + E\{(1 - D)(\ \beta_0^{*\mathbb{T}}\tilde{X})^2\} + \sigma_\varepsilon^2,$$

where $\sigma_\varepsilon = \mathrm{var}(\varepsilon)$. Since $P\{c \le e(X) \le 1 - c\} = 1$,

$$E(\ \beta_1^{*\mathbb{T}}\tilde{V})^2 \le E(\ \beta_1^{*\mathbb{T}}\tilde{X})^2 \le c^{-1}E\{e(X)(\ \beta_1^{*\mathbb{T}}\tilde{X})^2\} = c^{-1}E\{D(\ \beta_1^{*\mathbb{T}}\tilde{X})^2\} \le c^{-1}E(Y^2). \tag{1.69}$$

Let

$$r_i^{(-k)} = D_i/\hat{e}^{(-k)}(X_i),\ r^{(-k)} = D/\hat{e}^{(-k)}(X),\ r_i = D_i/e(X_i),\ r = D/e(X).$$

Since both $e(X)$ and $\hat{e}^{(-k)}(X)$ are bounded away from 0 uniformly with probability 1,

$$E_{I_k^c}(r^{(-k)} - r)^2 = E_{I_k^c}\frac{D\{\hat{e}^{(-k)}(X) - e(X)\}^2}{\{\hat{e}^{(-k)}(X)e(X)\}^2} = O_P(b_{m+n,p}^2). \tag{1.70}$$

Here, recall that $E_{I_k^c} g = E\{g \mid (D_i, Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}\}$ and $(D, Y, X) \sim P_{D,Y,X}$ independent of $(D_i, Y_i, X_i)_{i\in\{1,2,\ldots,n\}\setminus I_k}$. By the definition of $\hat{\tau}_1^{(k)}$, we can obtain the following formula

$$\begin{aligned}\hat{\tau}_1^{(k)} = \ & \beta_1^{*\mathbb{T}}\hat{\mu}^{(k)} + N^{-1}\sum_{i\in I_k} r_i(Y_i - \ \beta_1^{*\mathbb{T}}\tilde{X}_i) + (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\hat{\mu}^{(k)} \\ & + N^{-1}\sum_{i\in I_k}(r_i^{(-k)} - r_i)(Y_i - \ \beta_1^{*\mathbb{T}}\tilde{X}_i) \ - N^{-1}\sum_{i\in I_k} r_i(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}_i \\ & - N^{-1}\sum_{i\in I_k}(r_i^{(-k)} - r_i)(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}_i, \end{aligned} \tag{1.71}$$

where recall that $\hat{\mu}^{(k)} = M^{-1}\sum_{i\in J_k}\tilde{X}_i$. Observe that each term of the RHS in (1.71) is an average of (conditional) independent and identically distributed random variables. Hence, by Lemma 1.2, we can obtain the rates of each of the terms by looking at the first and second moments. For the first moments, recall that $r = D/e(X)$ and $E(r \mid X) = E(D \mid X)/e(X) = 1$, we have

$$\begin{aligned} & E(\beta_1^{*\mathbb{T}}\tilde{X}) = \beta_1^{*\mathbb{T}}\tilde{\mu}, \\ & E\{r(Y - \ \beta_1^{*\mathbb{T}}\tilde{X})\} = \tau_1 - \ \beta_1^{*\mathbb{T}}\tilde{\mu}, \\ & E\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\} = (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu}, \\ & E_{I_k^c}\{r(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\} = (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu}, \end{aligned}$$

and by the Holder's inequality,

$$E_{I_k^c}\{(r^{(-k)} - r)(Y - \ \beta_1^{*\mathbb{T}}\tilde{X})\} = E_{I_k^c}\{(r^{(-k)} - r)(E(Y \mid X) - \ \beta_1^{*\mathbb{T}}\tilde{X})\} = O_P(b_{m+n,p}c_p), \tag{1.72}$$

$$E_{I_k^c}\{(r^{(-k)} - r)(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\} = O_P(a_{n,p}b_{m+n,p}). \tag{1.73}$$

We can see that the terms $\beta_1^{*\mathbb{T}}\tilde{\mu}$ and $(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu}$ will cancel out, and the terms (1.72) and (1.73) will be the main contributions of the first moment.

As for the second moments, we have

$$\text{var}(\beta_1^{*\mathbb{T}}\tilde{X}) = E(\beta_1^{*\mathbb{T}}\tilde{V})^2 \leq c^{-1}E(Y^2) = O(1), \tag{1.74}$$

$$E_{I_k^c}\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{V}\}^2 \leq E_{I_k^c}\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2 = O_P(a_{n,p}^2),$$

where (1.74) results from (1.69). By Condition 1.5, $r = D/e(X) \leq c^{-1}$ and $|r^{(-k)} - r| = |D/\hat{e}^{(-k)}(X) - D/e(X)| \leq c^{-1}$ with probability 1. Hence, we have following results for the variance (or second moments) of the terms in (1.71),

$$E\{r^2(Y - \beta_1^{*\mathbb{T}}\tilde{X})^2\} \leq c^{-2}\sigma_\varepsilon^2 = O(1),$$

$$E_{I_k^c}r^2\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{V}\}^2 \leq c_1^{-2}E_{I_k^c}\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2 = O_P(a_{n,p}^2), \tag{1.75}$$

$$E_{I_k^c}(r^{(-k)} - r)^2\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2 \leq c_1^{-2}E_{I_k^c}\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2 = O_P(a_{n,p}^2). \tag{1.76}$$

Besides, by the assumption that $P\{E(\varepsilon^2 \mid X) < C\} = 1$, we have

$$E_{I_k^c}(r^{(-k)} - r)^2(Y - \beta_1^{*\mathbb{T}}\tilde{X})^2 \leq CE_{I_k^c}(r^{(-k)} - r)^2 = O_P(b_{m+n,p}^2). \tag{1.77}$$

Now, by Lemma 1.2, we have asymptotic results for each of the terms in (1.71). The terms $\beta_1^{*\mathbb{T}}\hat{\mu}^{(k)}$ and $(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\hat{\mu}^{(k)}$ are averages of M (conditional) independent and identically distributed random variables, we have

$$\beta_1^{*\mathbb{T}}\hat{\mu}^{(k)} = \beta_1^{*\mathbb{T}}\tilde{\mu} + O_P(M^{-1/2}),$$

$$(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\hat{\mu}^{(k)} = (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu} + O_P(a_{n,p}M^{-1/2}).$$

The other terms in (1.71) are averages of N (conditional) independent and identically dis-

tributed random variables, we have

$$N^{-1}\sum_{i\in I_k} r_i(Y_i - \ \beta_1^{*\mathbb{T}}\tilde{X}_i) = \tau_1 - \ \beta_1^{*\mathbb{T}}\tilde{\mu} + O_P(N^{-1/2}),$$

$$N^{-1}\sum_{i\in I_k} r_i(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}_i = (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu} + O_P(a_{n,p}N^{-1/2}),$$

and

$$N^{-1}\sum_{i\in I_k} (r_i^{(-k)} - r_i)(Y_i - \ \beta_1^{*\mathbb{T}}\tilde{X}_i) = O_P(b_{m+n,p}c_p + b_{m+n,p}N^{-1/2}),$$

$$N^{-1}\sum_{i\in I_k} (r^{(-k)} - r)(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X} = O_P(a_{n,p}b_{m+n,p} + a_{n,p}N^{-1/2}).$$

Combining the previous results, we have

$$\hat{\tau}_1^{(k)} = \tau_1 + O_P(a_{n,p}b_{m+n,p} + b_{m+n,p}c_p + (1 + a_{n,p} + b_{m+n,p})N^{-1/2}),$$

Similarly,

$$\hat{\tau}_2^{(k)} = \tau_2 + O_P(a_{n,p}b_{m+n,p} + b_{m+n,p}c_p + (1 + a_{n,p} + b_{m+n,p})N^{-1/2}).$$

When $K < \infty$, $a_{n,p} = O(1)$, $a_{n,p}b_{m+n,p} = O(n^{-1/2})$ and $b_{m+n,p}c_p = O(n^{-1/2})$,

$$\begin{aligned}\hat{\delta} = \hat{\tau}_1 - \hat{\tau}_2 &= \delta + O_P(a_{n,p}b_{m+n,p} + b_{m+n,p}c_p + (1 + a_{n,p} + b_{m+n,p})n^{-1/2}) \\ &= \delta + O_P(n^{-1/2}).\end{aligned} \tag{1.78}$$

Moreover, if $a_{n,p} = o_P(1)$, $b_{m+n,p} = o_P(1)$, $a_{n,p}b_{m+n,p} = o_P(n^{-1/2})$ and $b_{m+n,p}c_p = o_P(n^{-1/2})$. Then, by the previous results and Lindeberg-Feller Central Limit Theorem,

$$\begin{aligned}n^{1/2}(\hat{\delta} - \delta) = n^{1/2}(m+n)^{-1}\sum_{i=1}^{m+n}(\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{V}_i + n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\zeta_i/[e(X_i)\{1-e(X_i)\}] \\ - E[\varepsilon\zeta/e(X)\{1-e(X)\}] + o_P(1) \\ \to N(0, V_\delta),\end{aligned} \tag{1.79}$$

in distribution, provided that

$$V_\delta = \text{var}\left[\frac{\varepsilon\zeta}{e(X)\{1-e(X)\}}\right] + \tau(\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{C}(\beta_1^* - \beta_0^*) > c > 0.$$

Part 2. Now we provide a consistency result for $\hat{V}_\delta$. Recall the definition of $\nu_{\delta,i}$,

$$\nu_{\delta,i} = r_i^{(-k)}(Y_i - \hat{\beta}_1^{(-k)^{\mathbb{T}}}\tilde{X}_i) - \rho_i^{(-k)}(Y_i - \hat{\beta}_0^{(-k)^{\mathbb{T}}}\tilde{X}_i) - \hat{\delta} + (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathbb{T}}\hat{\mu}^{(k)},$$

where $r_i^{(-k)} = D_i/\hat{e}^{(-k)}(X_i)$ and $\rho_i^{(-k)} = (1-D_i)/\{1-\hat{e}^{(-k)}(X_i)\}$. Define $\nu_{\delta,i}^* = r_i(Y_i - \beta_1^{*\mathbb{T}}\tilde{X}_i) - \rho_i(Y_i - \beta_0^{*\mathbb{T}}\tilde{X}_i)$, where $r_i = D_i/e(X_i)$ and $\rho_i = (1-D_i)/\{1-e(X_i)\}$. Similarly as in (1.55),

$$\begin{aligned}
&\left|N^{-1}\sum_{i\in I_k}\nu_{\delta,i}^2 - N^{-1}\sum_{i\in I_k}{\nu_{\delta,i}^*}^2\right| \\
\le &\left|N^{-1}\sum_{i\in I_k}(\nu_{\delta,i} - \nu_{\delta,i}^*)^2\right| + 2\left\{N^{-1}\sum_{i\in I_k}{\nu_{\delta,i}^*}^2 N^{-1}\sum_{i\in I_k}(\nu_{\delta,i} - \nu_{\delta,i}^*)^2\right\}^{1/2}.
\end{aligned}$$

By Conditions 1.1 and 1.5, $E|r(Y - \beta_1^{*\mathbb{T}}\tilde{X}) - \rho(Y - \beta_0^{*\mathbb{T}}\tilde{X})|^{2+c} < c_1$, where $r = D/e(X)$ and $\rho = (1-D)/\{1-e(X)\}$. By Lemma 1.2,

$$N^{-1}\sum_{i\in I_k}{\nu_{\delta,i}^*}^2 = V_1 + o_P(1),$$

where $V_1 = \text{var}\{r(Y - \beta_1^{*\mathbb{T}}\tilde{X}) - \rho(Y - \beta_0^{*\mathbb{T}}\tilde{X})\}$. Now it remains to show $N^{-1}\sum_{i\in I_k}(\nu_{\delta,i} - \nu_{\delta,i}^*)^2 = o_P(1)$. Observe that $\nu_{\delta,i} - \nu_{\delta,i}^* = A_{1,i} + A_{2,i} + A_3$, where

$$\begin{aligned}
A_{1,i} &= r_i^{(-k)}(Y_i - \hat{\beta}_1^{(-k)^{\mathbb{T}}}\tilde{X}_i) - r_i(Y_i - \beta_1^{*\mathbb{T}}\tilde{X}_i), \\
A_{2,i} &= -\rho_i^{(-k)}(Y_i - \hat{\beta}_0^{(-k)^{\mathbb{T}}}\tilde{X}_i) + \rho_i(Y_i - \beta_0^{*\mathbb{T}}\tilde{X}_i), \\
A_3 &= (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathbb{T}}\hat{\mu}^{(k)} - \hat{\delta}.
\end{aligned}$$

Hence, it suffices to show

$$N^{-1}\sum_{i\in I_k}A_{1,i}^2 = o_P(1), \quad N^{-1}\sum_{i\in I_k}A_{2,i}^2 = o_P(1), \quad A_3 = o_P(1).$$

Observe that

$$A_{1,i} = (r_i^{(-k)} - r_i)\varepsilon_i - r_i(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}_i - (r_i^{(-k)} - r_i)(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}_i$$

From (1.77), (1.75) and (1.76),

$$E_{I_k^c}\{(r^{(-k)} - r)^2\varepsilon^2\} = o_P(1), \quad E_{I_k^c}[r^2\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2] = o_P(1),$$

$$E_{I_k^c}[(r^{(-k)} - r)^2\{(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2] = o_P(1).$$

Hence,

$$E_{I_k^c}\{(r^{(-k)} - r)\varepsilon - r(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X} - (r^{(-k)} - r)(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{X}\}^2 = o_P(1).$$

By Lemma 1.1, $N^{-1}\sum_{i\in I_k} A_{1,i}^2 = o_P(1)$. Similarly, we have $N^{-1}\sum_{i\in I_k} A_{2,i}^2 = o_P(1)$. Besides,

$$\begin{aligned} A_3 = {} & (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu} - (\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}\tilde{\mu} + (\beta_1^* - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) \\ & + (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) - (\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) - (\hat{\delta} - \delta). \end{aligned}$$

Under the condition $a_{n,p} = o(1)$, we have $(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{\mu} = o_P(1)$ and $(\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}\tilde{\mu} = o_P(1)$. By Lemma 1.2, $(\beta_1^* - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$. By Lemma 1.1, $(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$ and $(\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$. Recall (1.78), $\hat{\delta} - \delta = o_P(1)$. Therefore, $A_3 = o_P(1)$. Now, combining all the previous results,

$$N^{-1}\sum_{i\in I_k}(\nu_{\delta,i} - \nu_{\delta,i}^*)^2 = o_P(1), \tag{1.80}$$

and hence

$$N^{-1}\sum_{i\in I_k}\nu_{\delta,i}^2 = N^{-1}\sum_{i\in I_k}{\nu_{\delta,i}^*}^2 + o_P(1) = V_1 + o_P(1).$$

Now recall $\xi_{\delta,i} = (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathbb{T}}(\tilde{X}_i - \hat{\mu}^{(k)})$. Define $\xi^*_{\delta,i} = (\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{V}_i$. Similarly as in (1.55),

$$\begin{aligned}
&\left| N^{-1}\sum_{i\in I_k}\xi_{\delta,i}^2 - N^{-1}\sum_{i\in I_k}{\xi^*_{\delta,i}}^2 \right| \\
\leq &\left| N^{-1}\sum_{i\in I_k}(\xi_{\delta,i} - \xi^*_{\delta,i})^2 \right| + 2\left\{ N^{-1}\sum_{i\in I_k}{\xi^*_{\delta,i}}^2 \cdot N^{-1}\sum_{i\in I_k}(\xi_{\delta,i} - \xi^*_{\delta,i})^2 \right\}^{1/2},
\end{aligned}$$

By Condition 1.1, $E|(\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{V}|^{2+c} < c_1$. By Lemma 1.3,

$$N^{-1}\sum_{i\in I_k}{\xi^*_{\delta,i}}^2 = V_2 + o_P(1),$$

where $V_2 = (\beta_1^* - \beta_0^*)^{\mathbb{T}}\tilde{C}(\beta_1^* - \beta_0^*)$. Now it remains to show $N^{-1}\sum_{i\in I_k}(\xi_{\delta,i} - \xi^*_{\delta,i})^2 = o_P(1)$. Observe that

$$\begin{aligned}
\xi_{\delta,i} - \xi^*_{\delta,i} = &(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{V}_i - (\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}\tilde{V}_i \\
&- (\beta_1^* - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) - (\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) + (\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}).
\end{aligned}$$

By Lemma 1.1, $N^{-1}\sum_{i\in I_k}(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}\tilde{V}_i = o_P(1)$, $N^{-1}\sum_{i\in I_k}(\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}\tilde{V}_i = o_P(1)$, $(\hat{\beta}_1^{(-k)} - \beta_1^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$ and $(\hat{\beta}_0^{(-k)} - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$. By Lemma 1.3, $(\beta_1^* - \beta_0^*)^{\mathbb{T}}(\hat{\mu}^{(k)} - \tilde{\mu}) = o_P(1)$. Therefore,

$$N^{-1}\sum_{i\in I_k}(\xi_{\delta,i} - \xi^*_{\delta,i})^2 = o_P(1),$$

and hence $N^{-1}\sum_{i\in I_k}\xi_{\delta,i}^2 = N^{-1}\sum_{i\in I_k}{\xi^*_{\delta,i}}^2 + o_P(1) = V_2 + o_P(1)$. When $K < \infty$,

$$\hat{V}_\delta = K^{-1}\sum_{k=1}^{K}\left\{ N^{-1}\sum_{i\in I_k}\nu_{\delta,i}^2 + nN^{-1}\sum_{i\in I_k}\xi_{\delta,i}^2/(m+n) \right\} = V_\delta + o_P(1).$$

■

1.9 Acknowledgement

Chaper 1, in full, is a reprint of the material as it appears in Biometrika. Zhang, Yuqian; Bradic, Jelena. High-dimensional semi-supervised learning: in search of optimal inference of the mean, Biometrika, asab042, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap

2.1 Introduction

Inference in semi-supervised (SS) settings has received substantial attention in recent times. Unlike traditional statistical learning settings that are usually either supervised or unsupervised, an SS setting represents a confluence of these two settings. A typical SS setting has two types of available data: apart from a small or moderate-sized *labeled* (or supervised) data $\mathcal{L} = (Y_i, \mathbf{X}_i)_{i=1}^n$, one has access to a *much larger* sized *unlabeled* (or

unsupervised) data $\mathcal{U} = (\mathbf{X}_i)_{i=n+1}^{N}$ with $N \gg n$. Here, $Y_i \in \mathbb{R}$ and $\mathbf{X}_i \in \mathbb{R}^p$ denote the outcome of interest and a covariate vector (possibly high dimensional), respectively. To integrate the notation, we use $R_i \in \{0, 1\}$ to denote the missingness/labeling indicator and use $\mathbb{S} = \mathcal{L} \cup \mathcal{U} = (R_i, R_i Y_i, \mathbf{X}_i)_{i=1}^{N}$ to denote the full data, a collection of N i.i.d. (independent and identically distributed) observations of $(R, RY, \mathbf{X})$, where throughout this chapter, we let $(R, RY, \mathbf{X})$ denote an independent copy of $(R_i, R_i Y_i, \mathbf{X}_i)$.

SS settings arise naturally whenever the covariates are easily available for a large cohort (so that $\mathcal{U}$ is plentiful), but the corresponding response is expensive and/or difficult to obtain due to various practical constraints (thus limiting the size of $\mathcal{L}$), a frequent scenario in modern studies involving large databases in the 'big data' era. Examples of such settings are ubiquitous across various scientific disciplines, including machine learning problems like speech recognition, text mining etc. [Zhu05, CSZ09], as well as more recent (and relevant to our work) biomedical applications, like electronic health records (EHR) and integrative genomic studies [CC18, CG20]. It is important to note that while SS settings can be viewed as a missing data problem of sorts, the fact that $|\mathcal{U}| \gg |\mathcal{L}|$ is a *key distinguishing feature* of SS settings (for instance, $|\mathcal{L}|$ could be of the order of hundreds, while $|\mathcal{U}|$ could be in the order of tens of thousands!). This condition, a natural consequence of the underlying practical situations leading to these data, implies that the proportion of labeled observations in SS settings converges to 0 as the sample sizes $|\mathcal{L}|, |\mathcal{U}| \to \infty$. This makes SS settings unique and fundamentally *different* from any standard missing data problem where this proportion is always assumed to be bounded away from 0, a condition also known as the *positivity* (or *overlap*) assumption in the missing data literature [Imb04, Tsi07], which is *naturally violated* here.

Most of the SS literature, however, implicitly assumes that $\mathbf{X}$ is equally distributed in $\mathcal{L}$ and $\mathcal{U}$ samples, that is, a missing completely at random (MCAR) setting, where $R \perp\!\!\!\perp (Y, \mathbf{X})$, and the goal is to improve efficiency over an (already valid) supervised estimator based on $\mathcal{L}$. A biased, covariate-dependent, missing at random (MAR) type labeling mechanism has not been studied much, although they are much more realistic in practice, especially in biomedical applications (including the examples discussed earlier) where *selection bias* is common. For instance, in EHR data, relatively 'sicker' patients may often be more likely to be labeled, especially if the labeling is for a disease response. We work in this type of a *'decaying'* MAR domain, which we name *MAR-SS* for short, under the typical "ignorability" assumption:

$$R \perp\!\!\!\perp Y \mid \mathbf{X},$$

thereby allowing for a selection bias in the process. It is important to note that the traditional MAR setting amongst the missing data literature is typically studied together with an overlap (positivity) condition that bounds away the propensity score (PS) $E(R|\mathbf{X})$ uniformly from zero [BR05]. Compared to such MAR settings, our MAR-SS setting is significantly more challenging due to the inevitably decaying nature of the PS. We also interchangeably refer to this setting as *decaying overlap*. As $N \gg n$ here, positivity is automatically excluded, thus leading to a *non-standard* asymptotic regime.

Subtleties To work with such unbalanced labeling, we denote the PS as $\pi_N(\mathbf{X}) := E(R|\mathbf{X}) \equiv P(R = 1|\mathbf{X})$ and let $\pi_N := E(R)$. It is important to note that to allow a non-degenerate PS with $E(R) \to 0$ as $N \to \infty$, we *must* allow R, $\pi_N(\mathbf{X})$ and π_N to *depend* on N (otherwise forcing $n/N \to 0$ would lead to a degenerate situation with $E(R) = 0$ and

$E(R|\mathbf{X}) = 0$ almost surely (a.s.)). Hence, both $\{R_{N,i}\}_{N,i}$ and $\{\pi_N(\mathbf{X}_i)\}_{N,i}$ form triangular arrays. We suppress the dependence of R_N on N throughout for notational simplicity.

Under such a *decaying MAR-SS setting*, we study the fundamental problem of estimation and inference towards the mean response, defined as:

$$\theta_0 \ := \ E(Y).$$

The mean estimation problem above is a canonical problem in classical missing data as well as causal inference literature, and we consider it here mainly as a prototype problem. The bigger purpose of this chapter is to provide a deeper understanding of this *MAR-SS* setting and all its subtleties, where the main challenge is to allow for the uniform decay of the PS with the sample size and handle the non-standard asymptotics that arises inevitably. Moreover, unlike "traditional" SS settings (with MCAR), the goal here is *not* to "improve" over a supervised estimator from $\mathcal{L}$ (which is no longer valid under selection bias) but rather *develop from scratch a consistent and rate-optimal estimator along with inferential tools for it.* The contributions of this work therefore constitutes advances both in the literature of classical missing data and causal inference as well as that of traditional SS inference. We first provide an overview of the existing literature(s), followed by a summary of our contributions.

2.1.1 Related Literature

SS-literature on prediction problems is vast, typically under the name of semi-supervised learning; see [Zhu05] and [CSZ09] for a review. SS inference has attracted a lot of recent attention. [ZBC19] and [ZB21] proposed SS mean estimators. The estimators in [ZB21] can be roughly seen as a special (MCAR) case of the MAR-SS setting here. [ABS$^+$21] and [CC18]

tackled the SS linear regression problem, while [KK13] considered likelihood based SS inference. [CG20] studied SS inference of the explained variance in high dimensional linear regression. However, they all require a MCAR assumption, i.e., $R \perp\!\!\!\perp (Y, \mathbf{X})$. MCAR is practically too strong, and these estimators lead to doubtful results once the dependency of R on $\mathbf{X}$ occurs.

Works that remove some of the MCAR restrictions have been proposed recently. A special stratified labeling in a SS framework was studied in [GLTC20] with a focus on prediction performance measures. Stratified labeling was also studied in [HLL20], though their setting is very specific in that their only source of randomness arises from the treatment assignment. [LZC20] consider a covariate shift regression under a SS framework using a semi-nonparametric approach based on density ratio estimations albeit, working only with a non-decaying PS. To our knowledge, only [KM20] have recently considered settings of a similar type as ours. Their main focus, however, was on treatment effects estimation and efficiency theory when surrogate variables occur in the usual MAR setting (with positivity). They do provide some results under a decaying PS setting, including a semiparametric efficiency bound. We provide a *complete characterization* (see Sections 2.3.1-2.3.2) of the asymptotic properties as well as inference based on the estimator (see Section 2.3.3), and under much weaker conditions. For instance, we only require $N\pi_N \to \infty$ (while they require $N\pi_N^2 \to \infty$) and we allow an unbounded support for $\mathbf{X}$, which is essentially violated under the uniformly bounded density ratio condition $\pi_N/\pi_N(\mathbf{X}) < C$ assumed in [KM20]. Moreover, the authors therein did not provide any results and/or methodology on the decaying PS's estimation which is an essential component of the problem here.

Our work is also naturally connected to the rich missing data (and causal infer-

ence) literature on semi-parametric methods, and especially to so-called doubly robust (DR) inference; see [RRZ94], [RR95], [BR05], [Tsi07], [KS07], and [Gra11] for a review. High-dimensional DR equivalents have been presented recently as well; see for example [BCH14, Far15, CCD$^+$18, SRR19, BWZ19]. They work on a low-dimensional parameter estimation problem that involves high-dimensional nuisance parameters. On the other hand, [SC17] and [CLCL19] work on problems where the parameters of interest themselves are high-dimensional. However, the positivity assumption is always assumed. Our work is a direct extension of the above literature where we now include a decaying PS, and therefore a setting of imbalanced treatment mechanisms.

Another related setting to our decaying PS setting is the so-called "limited overlap" setting. A few notable prior works on limited overlap include [CHIM09, KT10, YD17, Rot17, VZ18] among others, where a truncation of the PS is introduced and a restricted analysis to the portions of the treatment groups such that overlap holds is performed. The "limited overlap" condition is also weaker than the usual overlap condition, but very *different* from our decaying PS situation. The limited overlap allows the PS to approach zero on some specific regions in the support of $\mathbf{X}$, while we allow $E(R|\mathbf{X})$ to shrink to zero (with N) uniformly in $\mathbf{X}$. Moreover, they assume that $E(R|\mathbf{X})$ is independent of N. By allowing R to depend on N, we allow P_R and $P_{R|\mathbf{X}}$ to depend on N so that $\pi_N = E(R) \to 0$ is permissible (a necessity under our settings of interest), much unlike the existing limited overlap literature.

2.1.2 Our Contributions

Contributions of our work are three fold: on (i) *double robust estimation with decaying PS*, (ii) *estimation of decaying PS*, and (iii) *average treatment effect (ATE) estimation with imbalanced groups*.

Double robust estimation with decaying propensity We believe this work fills in an important gap in both the SS literature and the missing data literature. A selection bias in the labeling mechanism is allowed, therefore parting with the SS literature. A PS is allowed to decay to zero uniformly, consequently enriching the MAR literature. We propose a *double robust semi-supervised* (DRSS) mean estimator (see Sections 2.3.1-2.3.2), which can be viewed as an adaptation of the standard DR estimator [RRZ94] to our MAR-SS setting. Theorem 2.2, our main result for this part, provides a full characterization of the DRSS estimator and its asymptotic expansion when at least one of the nuisance functions is correctly specified. Throughout, our results bring in a new set of rate-adjusted high-level estimation error conditions on the nuisance estimators that are agnostic to their mode of construction. When both nuisance models are correctly specified, we derive the asymptotic normality of our estimator if a product rate condition for the estimation errors is further assumed, with an asymptotic variance reaching the semi-parametric efficiency bound derived in [KM20]. We also construct a corresponding confidence interval (see Section 2.3.3) that adapts to the rate of decay of the PS. Adaptivity here implies that the confidence sets are wider for the cases of faster decay without changing the estimators themselves. The analyses and the methods are considerably more involved here compared to the standard problems, due to the decaying nature of the PS. For example, we establish that the rate of

convergence is no longer governed by N solely; rather the effective rate is identified to be in terms of Na_N, where $a_N^{-1} = E\{\pi_N^{-1}(\mathbf{X})\}$. In high dimensions, and using standard parametric nuisance models, the product rate condition required for the asymptotic normality is $s_m s_\pi \{\log(p)\}^2 = o(Na_N)$, where s_m and s_π are the sparsity levels of the (linear/logistic) nuisance functions $m(\mathbf{X}) = E(Y|\mathbf{X})$ and $\pi_N(\mathbf{X}) = E(R|\mathbf{X})$, respectively. When $a_N \asymp 1$, such a condition coincides with the usual product condition [CCD$^+$18] where the positivity condition is assumed. However, whenever $a_N \to 0$, the condition is stricter in order to compensate for the decay of the PS.

Estimation of the decaying propensity A key challenge for any methodological development in our MAR-SS setting is the modeling of the decaying PS. We propose several choices and associated results in this regard, including (i) *stratified labeling* (see Section 2.4.4) as well as (ii) a novel *offset based imbalanced logistic regression model* (see Section 2.4.1), under *both* low and high dimensional settings. The first approach, (i), is often practically relevant in the presence of *apriori* information available on a stratifying variable. The second approach, (ii), on the other hand, is applicable quite generally and constitutes a natural extension of logistic models to our case of a decaying PS. Related to the latter model, imbalanced classification in low-dimensions was recently studied by [Owe07] and [Wan20]. Our offset based model is closely related to their diverging intercept model, and yet has distinct methodological advantages; see Remark 2.11 below.

We provide theoretical results about estimation rates and other properties of these models under both high and low dimensional settings. These results may be more generally useful and are of independent interest; for example, our results on estimation of

decaying PS under a logistic model in high dimensions are the *first* such results to our knowledge. We demonstrate that, for a sub-Gaussian $\mathbf{X}$, the estimation error of $\pi_N(\cdot)$ is $O_p(\sqrt{s_\pi \log(p)/(N\pi_N)})$, where s_π is the sparsity level of the logistic model parameter. Such a result is non-trivial as, per Theorem 2.5, an appropriate choice of the regularization parameter is non-standard with $\lambda_N \asymp \sqrt{\pi_N \log(p)/N}$. We also obtain a regular and asymptotically linear (RAL) expansion for the estimator of the logistic regression parameter in the low-dimensional case; see Theorem 2.4. Moreover, we showcase that the estimator reaches the asymptotic variance as established in [Wan20] for low-dimensional problems. For the cases where the outcome model is misspecified, we further construct an adjusted RAL expansion of our DRSS estimator. Lastly, in Section 2.4.5, we also consider the special case of the MCAR model and the corresponding results in that setting.

Average treatment effect (ATE) estimation with imbalanced groups Drawing on a natural connection between the causal inference and missing data settings (see the discussions in Section 1.1 of [CLCL19] for instance) we extend our results to a corresponding ATE estimation problem. Our results allow for an extremely imbalanced treatment or control groups, in that $\pi_N = P(R = 1) \to 0$ (or alternatively, $\pi_N \to 1$) as $N \to \infty$.

We establish a RAL expansion for the proposed ATE estimator with a non-standard consistency rate, $O_p(1/\sqrt{N\pi_N})$, where without loss of generality we assume $\pi_N \to 0$. A sufficient condition for the expansion's validity is correctness of the model for the treatment group's outcome as well as that of the PS model. Notably, the control group's outcome and PS models *can* be (even both) misspecified if $\pi_N \to 0$ fast enough. Such a condition is different from most of the recent results, such as [Far15] and [CCD$^+$18], where the nuisance

functions in both of the groups need to be correctly specified for valid inference results. It is also different from the recent work of [SRR19] and [Tan20a], where they used specific parametric working models and they required at least one of the nuisance functions to be correctly specified for both of the groups.

The PS setting, the parameter of interest, and the methodology are also different from the limited overlap literature, e.g., [CHIM09]. As shown in [KT10], the information bound for the ATE estimation is 0 if only under the ignorability assumption and a.s., $\pi_N(\cdot) \in (0,1)$. As a result, a common approach in the limited overlap literature is to re-target the parameter of interest by considering a "shifted" ATE induced by the truncation of the PS [CHIM09]. In this chapter, we show that it is in fact possible to estimate the ATE directly when we have additional information that the inverse PS has well-behaved tails, e.g., $\pi_N(\cdot)$ follows an offset logistic model and $\mathbf{X}$ is sub-Gaussian; see Theorems 2.2, 2.4, and 2.5.

2.1.3 Notation

We use the following notation throughout. Let $P(\cdot)$ and $E(\cdot)$ denote the probability measure and expectation characterizing the joint distribution of the underlying (possibly unobserved) random vector $\mathbb{Z} := (R, Y, \mathbf{X})$, respectively, where $R \in \{0,1\}$, $Y \in \mathbb{R}$, and $\mathbf{X} \in \mathbb{R}^p$. Let $P_{\mathbf{X}}$ denote the marginal distribution of $\mathbf{X}$. For any $r > 0$, let $\|f(\cdot)\|_{r,P} := \{E|f(\mathbf{Z})|^r\}^{1/r}$ and $\|f(\cdot)\|_{r,P_{\mathbf{X}}} := \{E_{\mathbf{X}}|f(\mathbf{X})|^r\}^{1/r}$. For any vector $\mathbf{z} \in \mathbb{R}^p$, we denote $\mathbf{z}(j)$ as the j-th coordinate of $\mathbf{z}$. For $r \geq 1$, define the l_r-norm of a vector $\mathbf{z}$ with $\|\mathbf{z}\|_r := (\sum_{j=1}^p |\mathbf{z}(j)|^r)^{1/r}$, $\|\mathbf{z}\|_0 := |\{j : \mathbf{z}(j) \neq 0\}|$, and $\|\mathbf{z}\|_\infty := \max_j |\mathbf{z}(j)|$. For a matrix $A \in \mathbb{R}^{p\times p}$, $\|A\|_r := \sup_{\mathbf{z}\neq\mathbf{0}} \|A\mathbf{z}\|_r/\|\mathbf{z}\|_r$ and $\lambda_{\min}(A)$ denotes the smallest eigenvalue of A. For sequences

a_N and b_N, we denote $a_N \asymp b_N$ if there exists constants $c, C, N_0 > 0$ such that $cb_N < a_N < Cb_N$ for all $N > N_0$. Lastly, we define the logit function as $\text{logit}(u) := \log\{u/(1-u)\}$ for any $u \in (0,1)$.

2.2 Problem setup

Let the entire dataset be denoted as: $\mathbb{S} := \{\mathbf{Z}_i = (R_i, R_iY_i, \mathbf{X}_i), i = 1, \ldots, N\}$. The dimension of the covariates p can be either fixed or growing with N in that $p = p_N \to \infty$ as $N \to \infty$. We assume the following ignorability condition throughout.

Assumption 2.1 (Ignorability or MAR condition)**.** *We assume that* $R \perp\!\!\!\perp Y \mid \mathbf{X}$.

The ignorability condition is standard in the missing data literature [BR05, Tsi07]. Let $m(\mathbf{x}) := E(Y|\mathbf{X} = \mathbf{x})$ and $\pi_N(\mathbf{x}) := E(R|\mathbf{X} = \mathbf{x})$ denote the conditional mean of Y and the conditional PS, respectively. We define a_N as:

$$a_N^{-1} \;:=\; E\{\pi_N^{-1}(\mathbf{X})\}, \tag{2.1}$$

which is a natural quantity that appears in all of our results under the MAR-SS setting, and plays a key role in determining the rates of any inverse-probability weighting type estimator. The value a_N shrinks when the distribution of $\pi_N(\cdot)$ has too much mass concentrated around 0; see Remark 2.5 for more details. We consider the case of $a_N \to 0$, although our results hold more broadly. Notice that the usual positivity (overlap) condition, $\pi_N(\mathbf{X}) > c > 0$, is NOT assumed throughout the chapter, and we allow a uniformly decaying PS in that $\pi_N(\mathbf{x}) \to 0$ as $N \to \infty$, for every $\mathbf{x}$ in the support $\mathcal{X}$.

Example 2.1 (Offset based PS model)**.** *Here, as an illustration of such decaying PS models, we introduce a general* offset based PS model *as follows.*

$$\pi_N(\mathbf{X}) \;=\; g(f(\mathbf{X}) + \log(\pi_N)), \quad \textit{with some } \; f : \mathbb{R}^p \to \mathbb{R} \;\; \textit{and} \;\; g(u) \;:=\; \frac{\exp(u)}{1+\exp(u)},$$

where $\log(\pi_N)$ *is an "offset". The model above constitutes a fairly general way of incorporating the naturally decaying nature of the PS in our setting. Further details on the rationale behind and the analysis of such an offset model are discussed in Section 2.4.1. Here we introduce the model mainly to illustrate how* $\pi_N(\mathbf{X})$ *depends on* N*. In our analysis in Section 2.4.1, we allow a linear* f *with any sub-Gaussian* $\mathbf{X}$*, where clearly the positivity condition is easily violated. Moreover, we allow* $\pi_N(\mathbf{X})$ *to be small in a "uniform way": for example, if* $\mathbf{X}$ *has a compact support* $\mathcal{X}$*, then* $c_1\pi_N \le \pi_N(\mathbf{x}) \le c_2\pi_N$ *for all* $\mathbf{x} \in \mathcal{X}$ *with constants* $0 < c_1 < c_2$.

Preliminaries: Identification and alternative representations We have the following three alternative representations or identifications of $\theta_0 = E(Y)$ based on the observable variables and some unknown (but estimable) nuisance functions, i.e., $m(\mathbf{X})$ and $\pi_N(\mathbf{X})$.

(Reg) Regression based representation: $\theta_0 = E\{m(\mathbf{X})\}$.

(IPW) Inverse probability weighting representation: $\theta_0 = E\{\pi_N^{-1}(\mathbf{X})RY\}$.

(DR) Doubly robust representation: $\theta_0 = E[m(\mathbf{X}) + \pi_N^{-1}(\mathbf{X})\{RY - Rm(\mathbf{X})\}]$.

A natural estimator of θ_0 would be the empirical mean of the observed responses, $\bar{Y}_{\text{labeled}} := \sum_{i=1}^N R_i Y_i / \sum_{i=1}^N R_i$. Under a MCAR setting, $\bar{Y}_{\text{labeled}}$ is a consistent estimator. However, under the MAR setting, $\bar{Y}_{\text{labeled}}$ is no longer a consistent estimator; $\bar{Y}_{\text{labeled}} \xrightarrow{p} E(Y|R =$

1) $\neq E(Y)$ in general. According to the above representations, with $\widehat{m}(\cdot)$ and $\widehat{\pi}_N(\cdot)$ estimating $m(\cdot)$ and $\pi_N(\cdot)$, respectively, we could consider $\widehat{\theta}_{\mathrm{Reg}} := N^{-1}\sum_{i=1}^N \widehat{m}(\mathbf{X}_i)$ and $\widehat{\theta}_{\mathrm{IPW}} := N^{-1}\sum_{i=1}^N R_i Y_i \widehat{\pi}_N^{-1}(\mathbf{X}_i)$. For the sake of simplicity, here we consider an ideal case that $\widehat{m}(\cdot)$ and $\widehat{\pi}_N(\cdot)$ are trained on another additional set so that $(\widehat{m}(\cdot), \widehat{\pi}_N(\cdot)) \perp\!\!\!\perp (\mathbf{X}_i)_{i=1}^N$.

It is then not hard to show that

$$\begin{aligned}
\widehat{\theta}_{\mathrm{Reg}} - \theta_0 &= O_p\left(\|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\|_{1,P_{\mathbf{X}}} + N^{-1/2}\right), \\
\widehat{\theta}_{\mathrm{IPW}} - \theta_0 &= O_p\left(\|1 - \pi_N(\mathbf{X})/\widehat{\pi}_N(\mathbf{X})\|_{2,P_{\mathbf{X}}} + N^{-1/2}\right).
\end{aligned}$$

Hence, the Reg and IPW estimators are not even consistent when the corresponding nuisance model is misspecified. Even when the corresponding nuisances are correctly specified, estimators directly depend on the estimation error of $\widehat{m}(\cdot)$ and $\widehat{\pi}_N(\cdot)$, respectively, which are not $\sqrt{N}$-consistent (nor $\sqrt{N\pi_N}$-consistent) in the high-dimensional or non-parametric settings.

The DR representation of θ_0, viewed as a combination of the Reg and IPW representations [Acc74], leads to double robustness. DR estimators are consistent as long as at least one of the models are correctly specified (this property is called "double robustness", see, e.g., Theorem 2 of [Far15]). When both models are correctly specified, the estimation errors of the DR estimators depend on the product of estimation errors of the nuisance functions; this property is called "rate double robustness," as defined in Definition 2 of [SRR19]. Moreover, DR estimators are known to be semi-parametrically optimal when both models are correct [BR05], as well as first order insensitive to the estimation errors of the nuisance functions [CCD$^+$18]; see the discussions in [CLCL19]. In Section 2.3, we propose estimators based on the above DR representation.

2.3 Semi-supervised inference under a MAR-SS setting

2.3.1 Known PS $\pi_N(\cdot)$

We first consider an oracle case where the PS, $\pi_N(\cdot)$, is known. In other words, the missing mechanism is designed and controlled by the researcher. This is also closely related to the randomized controlled trials in causal inference literature. Based on the DR representation, we consider the following SS estimator:

$$\tilde{\theta} \;:=\; N^{-1}\sum_{i=1}^{N}\widehat{m}(\mathbf{X}_i) + N^{-1}\sum_{i=1}^{N}\frac{R_i}{\pi_N(\mathbf{X}_i)}\{Y_i - \widehat{m}(\mathbf{X}_i)\}, \tag{2.2}$$

where $\widehat{m}(\mathbf{X}_i)$ is a *cross-fitted* estimator established as follows: 1) for any fixed $\mathbb{K} \geq 2$, let $\{\mathcal{I}_k\}_{k=1}^{\mathbb{K}}$ be a random partition of $\mathcal{I} := \{1, \ldots, N\}$; 2) for each $k \leq \mathbb{K}$, obtain the estimator $\widehat{m}(\cdot; \mathbb{S}_{-k})$ using the training set $\mathbb{S}_{-k} := \{\mathbf{Z}_i : i \in \mathcal{I} \setminus \mathcal{I}_k\}$, where for typical supervised methods, $\widehat{m}(\cdot; \mathbb{S}_{-k})$ only depends on the labeled observations, $\{\mathbf{Z}_i : i \in \mathcal{I} \setminus \mathcal{I}_k, R_i = 1\}$; 3) for each $i = 1, \ldots, N$, let $\widehat{m}(\mathbf{X}_i) := \widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k(i)})$, where $k(i)$ denotes the unique k such that $i \in \mathcal{I}_k$. The proposed $\tilde{\theta}$ can be seen as a debiased $\widehat{\theta}_{\text{Reg}}$ estimator, where the misspecification or estimation bias of $\widehat{m}(\cdot)$ is removed by the knowledge of $\pi_N(\cdot)$. On the other hand, $\widehat{\theta}_{\text{IPW}}$ is a special case of $\tilde{\theta}$ with $\widehat{\pi}_N(\cdot) = \pi_N(\cdot)$ and $\widehat{m}(\cdot) \equiv 0$. However, $\tilde{\theta}$ with a "good" estimator for the outcome model improves the efficiency of the IPW estimator; see e.g., Remark 2.3. The cross-fitting is vital for the bias correction; see discussions in [CCD$^+$18] and [CLCL19]. By the cross-fitting construction, $\widehat{m}(\cdot; \mathbb{S}_{-k(i)}) \perp\!\!\!\perp \mathbf{Z}_i$ for each $i \leq N$. As a result,

$$E_{\mathbf{X}}\left[\widehat{m}(\mathbf{X}) + \frac{R}{\pi_N(\mathbf{X})}\{Y - \widehat{m}(\mathbf{X})\}\right] \;=\; E_{\mathbf{X}}\left[\widehat{m}(\mathbf{X}) + \frac{\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\{Y - \widehat{m}(\mathbf{X})\}\right] \;=\; \theta_0,$$

and hence the proposed estimator $\tilde{\theta}$ is unbiased for θ_0, even if $m(\cdot)$ is misspecified. We denote $\mu(\cdot)$ as a "limit" (potentially *misspecified*) of $\widehat{m}(\cdot)$, i.e., in general, $\mu(\cdot) \neq m(\cdot)$ is allowed.

Assumption 2.2 (Basic assumption)**.** *(a)* $\mathbf{Z}$ *has finite 2nd moments and* $\mathbf{\Sigma} \equiv Var(\mathbf{X})$ *is positive definite. (b) Let* $E[\{Y - m(\mathbf{X})\}^2|\mathbf{X} = \mathbf{x}] \geq \sigma^2_{\zeta,1} > 0$ *and* $E[\{Y - \mu(\mathbf{X})\}^2|\mathbf{X} = \mathbf{x}] \leq \sigma^2_{\zeta,2} < \infty$ *for all* $\mathbf{x}$ *in the support* $\mathcal{X}$ *of* $\mathbb{P}_{\mathbf{X}}$*. Moreover,* $Var(Y) \leq \sigma^2_{\zeta,2}$*.*

Assumption 2.3 (Tail condition)**.** *Let* $a_N^{-1} E\left[\psi^2_{\mu,\pi}(\mathbf{Z}) \mathbb{1}\left\{|\psi_{\mu,\pi}(\mathbf{Z})| > c\sqrt{N/a_N}\right\}\right] \to 0$*, for any* $c > 0$ *as* $N \to \infty$*, where recall that* a_N *is defined in* (2.1)*, and with* $\psi_{\mu,\pi}(\mathbf{Z})$ *as:*

$$\psi_{\mu,\pi}(\mathbf{Z}) \;:=\; \mu(\mathbf{X}) + \frac{R}{\pi_N(\mathbf{X})}\{Y - \mu(\mathbf{X})\} - \theta_0 \;=\; Y - \theta_0 + \left\{\frac{R}{\pi_N(\mathbf{X})} - 1\right\}\{Y - \mu(\mathbf{X})\}. \tag{2.3}$$

Remark 2.1 (Discussion on Assumptions 2.2 and 2.3)**.** *Assumption 2.2 imposes some mild moment conditions; similar versions can be found in [ZBC19, ZB21]. Assumption 2.3 is needed only for the asymptotic normality and is satisfied if 1)* $\pi_N(\cdot)$ *follows an offset propensity model as in Example 2.1 with sub-Gaussian* $f(\mathbf{X})$ *(see Section 2.4.1 where we analyzed a special case of the offset model); 2)* $E\{|Y - \mu(\mathbf{X})|^{2+\delta}|X\} < C$*,* $E(|Y - \theta_0|^{2+\delta}) < C$ *with constants* $\delta, C > 0$*; and 3)* $N\pi_N \to \infty$ *as* $N \to \infty$*. A sufficient condition for Assumption 2.3 is given in in Assumption 2.4.*

In the result below, we analyze the theoretical properties of $\tilde{\theta}$ including its consistency, convergence rate, asymptotic normality and robustness properties.

Theorem 2.1. *Let Assumptions 2.1 and 2.2 hold. Let* $Na_N \to \infty$ *as* $N \to \infty$*. Let* $\mu(\cdot)$ *be a well-defined limit of the cross-fitted* $\widehat{m}(\cdot)$*, that satisfy:*

$$E_{\mathbf{X}}\left[\frac{a_N}{\pi_N(\mathbf{X})}\{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\}^2\right] \;=\; O_p(c^2_{\mu,N}),\ \textit{with sequence}\ c_{\mu,N} \;=\; o(1), \tag{2.4}$$

for $k \leq \mathbb{K}$. Then,

$$\tilde{\theta} - \theta_0 \;=\; N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}}\right) \;\; \textit{and} \;\; V_N(\mu) \;:=\; Var\{\psi_{\mu,\pi}(\mathbf{Z})\} \;\asymp\; a_N^{-1},$$

where $\psi_{\mu,\pi}(\mathbf{Z}_i)$ is defined in (2.3)*. Alternatively, we also have the following asymptotically linear representation:*

$$\tilde{\theta} - \theta_0 \;=\; N^{-1}\sum_{i=1}^{N}\widetilde{\psi}_{\mu}(\mathbf{Z}_i) + O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}} + \frac{1}{\sqrt{N}}\right) \;\; \textit{and} \;\; \widetilde{V}_N(\mu) := Var\{\widetilde{\psi}_{\mu}(\mathbf{Z})\} \asymp a_N^{-1},$$

where $\widetilde{\psi}_{\mu}(\mathbf{Z}) := R/\pi_N(\mathbf{X})\{Y - \mu(\mathbf{X})\} - E[R/\pi_N(\mathbf{X})\{Y - \mu(\mathbf{X})\}]$ and $E\{\widetilde{\psi}_{\mu}(\mathbf{Z})\} = 0$. Additionally, as long as Assumption 2.3 holds, we have:

$$(Na_N)^{1/2}(\tilde{\theta} - \theta_0) \;=\; O_p(1), \;\; \textit{and} \;\; N^{1/2}V_N^{-1/2}(\mu)(\tilde{\theta} - \theta_0) \;\to\; N(0,1).$$

Moreover, if $a_N \to 0$ as $N \to \infty$, then,

$$N^{1/2}\widetilde{V}_N^{-1/2}(\mu)(\tilde{\theta} - \theta_0) \;\to\; N(0,1), \;\; \textit{and} \;\; \frac{V_N(\mu)}{\widetilde{V}_N(\mu)} \;=\; 1 + O(a_N).$$

Remark 2.2 (Discussion on condition (2.4))**.** *As per Theorem 2.1, consistency and asymptotic normality of $\tilde{\theta}$ depend on* (2.4)*, a condition that involves 1) the convergence rate of $\widehat{m}(\cdot)$ towards some $\mu(\cdot)$, depending on the (expected) labeled sample size ($N\pi_N$), and 2) the tail of $\pi_N^{-1}(\mathbf{X})$, that is, how much of the mass of the distribution of $\pi_N(\mathbf{X})$ concentrates around zero. For a special case of $\pi_N(\mathbf{X}) \equiv \pi_N$, MCAR,* (2.4) *is equivalent to $\|\widehat{m}(\cdot;\mathbb{S}_{-k}) - \mu(\cdot)\|_{2,P_{\mathbf{X}}} = o_p(1)$ coinciding with [ZB21]. On the other hand, when $\pi_N(\cdot)$ follows the offset model (Example 2.1) with sub-Gaussian $f(\mathbf{X})$, we have $a_N \asymp \pi_N$, and* (2.4) *holds once $E_{\mathbf{X}}\{|\widehat{m}(\mathbf{X};\mathbb{S}_{-k}) - \mu(\mathbf{X})|^{2+\delta}\} = o_p(1)$ with $\delta > 0$.*

Remark 2.3 (Efficiency of $\tilde{\theta}$ and the choice of $\mu(\cdot)$)**.** *Although the choice of $\widehat{m}(\cdot)$ is arbitrary as long as it converges to some $\mu(\cdot)$ as in* (2.4)*, the efficiency of $\tilde{\theta}$ does depend on the limit*

$\mu(\cdot)$, and hence also on the choice of $\widehat{m}(\cdot)$. For a simple case of $\widehat{m}(\mathbf{x}) = \mu(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$, $\tilde{\theta}$ can be written as $\tilde{\theta} = N^{-1}\sum_{i=1}^{N} R_i Y_i/\pi_N(\mathbf{X}_i)$, which coincides with the IPW estimator, an estimator independent of $m(\cdot)$. However, an appropriate estimator $\widehat{m}(\cdot)$ will provide a better efficiency for $\tilde{\theta}$. The optimal choice of $\mu(\cdot)$ that minimizes the asymptotic variance $V_N(\mu)$ is $\mu(\cdot) = m(\cdot)$ indicating that the outcome model is correctly specified.

Remark 2.4 (Intuition behind the IFs)**.** *Two separate IFs $\psi_{\mu,\pi}(\mathbf{Z})$ and $\widetilde{\psi}_{\mu}(\mathbf{Z})$ appear in the expansions of $\tilde{\theta}$ in Theorem 2.1. The first IF $\psi_{\mu,\pi}(\mathbf{Z})$ is an "accurate influence function" in that $\tilde{\theta} - \theta_0 = N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + o_p((Na_N)^{-1/2})$ with $N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) \asymp (Na_N)^{-1/2}$. When $a_N \to 0$ as $N \to \infty$, the second IF $\widetilde{\psi}_{\mu}(\mathbf{Z})$ captures the main contribution of $\psi_{\mu,\pi}(\mathbf{Z})$. It only involves the labeled samples and hence one can clearly see that the rate of $\tilde{\theta}$ is effectively determined by the smaller sized, labeled data only. When the outcome model is correctly specified, the second IF $\widetilde{\psi}_{\mu}(\mathbf{Z})$ coincides with the efficient IF of [KM20]; see Theorem 4.1 therein.*

Remark 2.5 (Convergence rate and "effective sample size")**.** *Suppose the conditions in Theorem 2.1 hold, then $\tilde{\theta}$ is a $(Na_N)^{1/2}$-consistent estimator for θ_0. The value Na_N can be seen as an "effective sample size" having a similar role as the sample size in supervised learning. Bellow is a discussion on the value Na_N. By Jensen's inequality, $Na_N \leq N\pi_N$, where the difference between the two rates is related to the tail of $\pi_N^{-1}(\mathbf{X})$. Here, $N\pi_N$ is the expected sample size as $N\pi_N = E(n)$, where $n := \sum_{i=1}^{N} R_i$. Therefore, the effective sample size, Na_N, depends on 1) how much of the mass of the distribution of $\pi_N(\mathbf{X})$ concentrates around 0 and 2) the (expected) size of the labeled sample. MCAR is a special case with $\pi_N(\cdot)$ being a constant and therefore $Na_N = N\pi_N$. In another example, the offset based model in*

Example 2.1 and Section 2.4.1, we have $a_N \asymp \pi_N$ for sub-Gausian $\mathbf{X}$; see Theorems 2.4 and 2.5.

2.3.2 Unknown PS $\pi_N(\cdot)$ and the general version of the DRSS estimator

With $\pi_N(\cdot)$ being unknown in general observational studies, we propose our final estimator, a *doubly robust semi-supervised* (DRSS) estimator of the mean θ_0, given by:

$$\widehat{\theta}_{\text{DRSS}} \;:=\; N^{-1}\sum_{i=1}^{N}\widehat{m}(\mathbf{X}_i) + N^{-1}\sum_{i=1}^{N}\frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i)}\{Y_i - \widehat{m}(\mathbf{X}_i)\}, \tag{2.5}$$

where $\widehat{\pi}_N(\mathbf{X}_i)$ is a cross-fitted estimator of $\pi_N(\mathbf{X}_i)$ constructed similarly as $\widehat{m}(\mathbf{X}_i)$, as discussed below (2.2) in Section 2.3.1. The proposed estimator (2.5) is a plug-in version of (2.2). We denote with $e_N(\cdot)$ a "limit" of $\widehat{\pi}_N(\cdot)$, which is possibly *misspecified*, i.e., $e_N(\cdot)$ is not necessarily the same as $\pi_N(\cdot)$. Define the following generalization of (2.3), i.e., a DR score (influence) function:

$$\psi_{\mu,e}(\mathbf{Z}) \;:=\; \mu(\mathbf{X}) + \frac{R}{e_N(\mathbf{X})}\{Y - \mu(\mathbf{X})\} - \theta_0 \;=\; Y - \theta_0 + \left\{\frac{R}{e_N(\mathbf{X})} - 1\right\}\{Y - \mu(\mathbf{X})\}. \tag{2.6}$$

We have the following asymptotic results under the two cases: (a) both $\pi_N(\cdot)$ and $m(\cdot)$ are correctly specified; (b) one of $\pi_N(\cdot)$ and $m(\cdot)$ is correctly specified.

Theorem 2.2. *Let Assumptions 2.1 and 2.2 hold and let $Na_N \to \infty$, as $N \to \infty$. Suppose the cross-fitted versions of $\widehat{m}(\cdot)$ and $\widehat{\pi}_N(\cdot)$ have well-defined (possibly misspecified) limits $\mu(\cdot)$*

and $e_N(\cdot)$, respectively, such that (2.4) holds for $k \leq \mathbb{K}$ as well as

$$E_{\mathbf{X}}\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1-\frac{e_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}\right\}^2\right] \;=\; O_p(c_{e,N}^2) \;\; \textit{with sequence} \;\; c_{e,N} \;=\; o(1), \tag{2.7}$$

$$E_{\mathbf{X}}\{\widehat{m}(\mathbf{X};\mathbb{S}_{-k})-\mu(\mathbf{X})\}^2 \;=\; O_p(r_{\mu,N}^2) \;\; \textit{with sequence} \;\; r_{\mu,N} \;=\; o(1), \;\; \textit{and} \tag{2.8}$$

$$E_{\mathbf{X}}\left\{1-\frac{e_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}\right\}^2 \;=\; O_p(r_{e,N}^2) \;\; \textit{with sequence} \;\; r_{e,N} \;=\; o(1). \tag{2.9}$$

The properties of $\widehat{\theta}_{\mathrm{DRSS}}$ under different cases are as follows:

(a) Suppose both $\mu(\cdot) = m(\cdot)$ and $e_N(\cdot) = \pi_N(\cdot)$ hold. Then, as $N \to \infty$, $\widehat{\theta}_{\mathrm{DRSS}}$ satisfies the following asymptotic linear expansion:

$$\widehat{\theta}_{\mathrm{DRSS}}-\theta_0 \;=\; N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i)+O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}}+\frac{c_{e,N}}{\sqrt{Na_N}}+r_{\mu,N}r_{e,N}\right),$$

and $V_N(\mu,e) \asymp a_N^{-1}$, where $V_N(\mu,e) := Var\{\psi_{\mu,e}(\mathbf{Z})\}$. Hence, as long as the product rate $r_{\mu,N}r_{e,N}$ from (2.8) and (2.9) further satisfies $r_{\mu,N}r_{e,N} = o(1/\sqrt{Na_N})$, and Assumption 2.3 holds, we have:

$$(Na_N)^{1/2}(\widehat{\theta}_{\mathrm{DRSS}}-\theta_0) \;=\; O_p(1), \;\; \textit{and} \;\; N^{1/2}V_N^{-1/2}(\mu,e)(\widehat{\theta}_{\mathrm{DRSS}}-\theta_0) \;\to\; N(0,1). \tag{2.10}$$

(b) Suppose now that either $\mu(\cdot) = m(\cdot)$ or $e_N(\cdot) = \pi_N(\cdot)$ holds. Moreover, if $e_N(\cdot) \neq \pi_N(\cdot)$, we assume $c \leq \pi_N(\mathbf{X})/e_N(\mathbf{X}) \leq C$ a.s. for some constants $c, C > 0$. Then, as $N \to \infty$, $\widehat{\theta}_{\mathrm{DRSS}}$ satisfies the following asymptotic linear expansion:

$$\widehat{\theta}_{\mathrm{DRSS}}-\theta_0 \;=\; N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i)+O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}}+\frac{c_{e,N}}{\sqrt{Na_N}}+r_{\mu,N}r_{e,N}\right)+\widehat{\Delta}_N,$$

with $\widehat{\Delta}_N$ satisfying (2.11) or (2.12):

$$\widehat{\Delta}_N \;:=\; N^{-1}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi_N(\mathbf{X}_i)}-\frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i)}\right\}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\} \quad \textit{if} \;\; e_N(\cdot) \;=\; \pi_N(\cdot), \tag{2.11}$$

$$\widehat{\Delta}_N \;:=\; N^{-1}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi_N(\mathbf{X}_i)}-\frac{R_i}{e_N(\mathbf{X}_i)}\right\}\{\widehat{m}(\mathbf{X}_i)-m(\mathbf{X}_i)\} \quad \textit{if} \;\; \mu(\cdot) \;=\; m(\cdot). \tag{2.12}$$

Suppose for case (2.11), $\|m(\cdot)-\mu(\cdot)\|_{2,P_{\mathbf{X}}} < C$, *while for case* (2.12), $\|1-\pi_N(\cdot)/e_N(\cdot)\|_{2,P_{\mathbf{X}}} < C$, *with a constant* $C < \infty$. *Then,* $\widehat{\theta}_{\mathrm{DRSS}}$ *satisfies:*

$$\widehat{\theta}_{\mathrm{DRSS}} - \theta_0 = O_p\left(\frac{1 + c_{\mu,N} + c_{e,N}}{\sqrt{N a_N}} + r_{\mu,N} r_{e,N} + r_{e,N}\mathbb{1}\{\mu(\cdot) \neq m(\cdot)\} + r_{\mu,N}\mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\}\right).$$

A few remarks pertaining to the estimation rates are presented next.

Remark 2.6 (Conditions in Theorem 2.2)**.** *Here we discuss the rate conditions* (2.4), (2.7), (2.8), *and* (2.9) *required in Theorem 2.2. The rate* (2.8) *is a standard estimation error of the outcome model; see for example, [ZBC19]. The other rates,* (2.4), (2.7), *and* (2.9), *are rescaled or self-normalized versions of conditions in [CCD+18]. They are needed as the price of violating the positivity condition. The rate* (2.9), *a rescaled version of the usually considered* $E_{\mathbf{X}}\{\widehat{\pi}_N(\mathbf{X}) - e_N(\mathbf{X})\}^2$, *is a change needed to properly address a decaying PS estimator. Then,* (2.4) *and* (2.7) *can be seen as self-normalized versions with the normalization factor being* $\omega(\mathbf{X}) := a_N/\pi_N(\mathbf{X})$. *Notice that* $E\{\omega(\mathbf{X})\} = 1$, *so these weights* $\omega(\cdot)$ *can be viewed as reweighing or redistribution factor. Then, the estimation errors of* $\widehat{\pi}_N(\mathbf{X})$ *and* $\widehat{m}(\mathbf{X})$ *at* $\mathbf{X}$, *with a smaller PS, contribute more to rates* (2.8) *and* (2.9). *The rates of the reweighed versions,* $c_{\mu,N}$ *and* $c_{e,N}$ *in* (2.4) *and* (2.7), *only need to be* $o(1)$*; whereas* $r_{\mu,N}$ *and* $r_{e,N}$ *in* (2.8) *and* (2.9) *appear in the final rate for* $\widehat{\theta}_{\mathrm{DRSS}}$. *In high dimensions, assume* $\pi_N(\cdot)$ *follows an offset based model as in Example 2.1. Suppose* $m(\cdot)$ *and* $f(\cdot)$ *in Example 2.1 are linear with sparsity levels* s_m *and* s_π, *respectively. Then, for sub-Gaussian* $\mathbf{X}$, *we demonstrate in Theorem 2.5 that* $a_N \asymp \pi_N$ *as long as* $r_{e,N} = \sqrt{s_\pi \log(p)/(N\pi_N)}$ *and* $r_{\mu,N} = \sqrt{s_m \log(p)/(N\pi_N)}$, *therefore coming close to the simplest missingness pattern, that of MCAR.*

Remark 2.7 (Double robustness, rates and efficiency)**.** *Here, we discuss the double robustness and the efficiency of the proposed estimator. Whenever* $\pi_N(\cdot)$ *and* $m(\cdot)$ *are correctly*

specified, the asymptotic normality with a rate of consistency $(Na_N)^{-1/2}$ *is guaranteed if a product rate condition* $r_{\mu,N}r_{e,N} = o(1/\sqrt{Na_N})$ *is satisfied. We can see that our product rate condition is an analog of the usual product rate condition in the literature [CCD+18], if the sample size is replaced with* Na_N*, the "effective sample size" in our case; see Remark 2.5. In addition, when the asymptotic normality occurs, our estimator reaches the semi-parametric efficiency bound proposed in [KM20] when* $\pi_N \to 0$ *as* $N \to \infty$*. When one of* $\pi_N(\cdot)$ *and* $m(\cdot)$ *is misspecified, we obtain a consistency rate of* $O_p(r_{e,N})$ *if* $\pi_N(\cdot)$ *is correctly specified, whereas the rate is* $O_p(r_{\mu,N})$ *if* $m(\cdot)$ *is correctly specified. Therefore, the consistency rate of* $\widehat{\theta}_{\text{DRSS}}$ *directly depends on the estimation error rate of the correct model. As a special case,* $\widehat{\theta}_{\text{DRSS}}$ *is consistent as long as the correct model is consistently estimated. Additionally, we can see that* $\widehat{\theta}_{\text{DRSS}}$ *can still be* $(Na_N)^{1/2}$*-consistent as long as the correct model is estimated with an error rate* $O_p(Na_N)^{-1/2}$*, which is reachable in low dimensions. For instance, for a (correctly specified) low-dimensional offset logistic PS model as introduced in Section 2.4.2, as shown in Theorem 2.4, not only do we reach the error rate* $O_p(Na_N)^{-1/2}$ *but are able to construct a RAL expansion for* $\widehat{\theta}_{\text{DRSS}}$*.*

Remark 2.8 (Unbounded support for $\mathbf{X}$)**.** *We do not enforce a bounded support for* $\mathbf{X}$*, which is typically an assumption assumed (implicitly) in missing data and causal inference literature. For instance, suppose* $\pi_N(\cdot)$ *follows an (offset based) logistic model as in Example 2.1. Both the usual positivity condition* $P(\pi_N(\mathbf{X}) > c > 0) = 1$ *in the standard missing data literature [Imb04, Tsi07, IR15b] and the uniform bounded density ratio condition,* $\pi_N/\pi_N(\mathbf{X}) < C$*, in [KM20], which tackles a MAR-SS setting, essentially require a compact support for* $\mathbf{X}$*. However, our results only require a sub-Gaussian* $\mathbf{X}$ *as in Theorems 2.4 and*

2.5.

Remark 2.9 (Asymptotic linearity and $(Na_N)^{1/2}$-consistency under misspecification)**.** *Moreover, in Section 2.4, we demonstrate that* $\widehat{\theta}_{\text{DRSS}}$ *can still be asymptotically normal even if* $m(\cdot)$ *is misspecified. Such an asymptotic normality is constructed based on a careful analysis to obtain the regular and asymptotically linear (RAL) expansion and the IF for the additional error term* $\widehat{\Delta}_N$ *in* (2.11)*, in that*

$$\widehat{\Delta}_N \;=\; N^{-1}\sum_{j=1}^{N} \text{IF}_{\pi}(\mathbf{Z}_j) + o_p\left((Na_N)^{-1/2}\right),$$

for some $\text{IF}_{\pi}(\cdot)$ *with* $E\{\text{IF}_{\pi}(\mathbf{Z})\} = 0$ *and* $E\{\text{IF}^2_{\pi}(\mathbf{Z})\} \asymp a_N^{-1}$*. The final IF of* $\widehat{\theta}_{\text{DRSS}}$ *involves the extra IF contributed from the estimation error of* $\widehat{\pi}_N(\cdot)$*. Consequently, the RAL expansion and the asymptotic normality of* $\widehat{\theta}_{\text{DRSS}}$ *are also affected accordingly. Using the above expansion for* $\widehat{\Delta}_N$ *and the general expansion of* $\widehat{\theta}_{\text{DRSS}}$ *from Theorem 2.2, we have a RAL expansion of* $\widehat{\theta}_{\text{DRSS}}$ *as:*

$$\begin{aligned}\widehat{\theta}_{\text{DRSS}} - \theta_0 \;&=\; N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i) + O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}} + \frac{c_{e,N}}{\sqrt{Na_N}} + r_{\mu,N}r_{e,N}\right) + \widehat{\Delta}_N \\ &=\; N^{-1}\sum_{i=1}^{N}\{\psi_{\mu,e}(\mathbf{Z}_i) + \text{IF}_{\pi}(\mathbf{Z}_i)\} + o_p\left((Na_N)^{-1/2}\right).\end{aligned}$$

The function $\Psi(\mathbf{Z}) := \psi_{\mu,e}(\mathbf{Z}) + \text{IF}_{\pi}(\mathbf{Z})$ *is the final* adjusted *IF of* $\widehat{\theta}_{\text{DRSS}}$ *with* $E\{\Psi(\mathbf{Z})\} = 0$ *and* $Var\{\Psi(\mathbf{Z})\} \asymp a_N^{-1}$*. Consequently, we also have:*

$$N^{1/2}[\,Var\{\Psi(\mathbf{Z})\}]^{-1/2}(\widehat{\theta}_{\text{DRSS}} - \theta_0) \;\to\; N(0,1). \tag{2.13}$$

2.3.3 Asymptotic variance estimation

In this section, we consider the estimation of the asymptotic variances $V_N(\mu)$ in Theorem 2.1 (with $\pi_N(\cdot)$ known) and $V_N(\mu, e)$ in Theorem 2.2 (with $\pi_N(\cdot)$ unknown and

both $m(\cdot)$ and $\pi_N(\cdot)$ are correctly specified). These facilitate inference on θ_0 (via confidence intervals, hypothesis tests etc.) using $\tilde{\theta}$ and $\widehat{\theta}_{\mathrm{DRSS}}$. We assume the following tail condition.

Assumption 2.4 (Tail condition)**.** *With $N \to \infty$, for a constant $\delta > 0$, let*

$$N^{-\delta/2} a_N^{1+\delta/2} E\{|\psi_{\mu,\pi}(\mathbf{Z})|^{2+\delta}\} \ \to \ 0.$$

The Assumption 2.4 is a sufficient condition for Assumption 2.3. Under the setting in Theorem 2.1 and part (a) of Theorem 2.2, we have:

$$N^{1/2} V_N^{-1/2}(\mu)(\tilde{\theta} - \theta_0) \ \to \ N(0,1), \qquad N^{1/2} V_N^{-1/2}(\mu, e)(\widehat{\theta}_{\mathrm{DRSS}} - \theta_0) \ \to \ N(0,1).$$

We propose the plug-in estimates: $\widehat{V}_N(\mu) = \widehat{V}_N(\widehat{m}, \pi_N, \tilde{\theta})$ and $\widehat{V}_N(\mu, e) = \widehat{V}_N(\widehat{m}, \widehat{\pi}_N, \widehat{\theta}_{\mathrm{DRSS}})$, where

$$\widehat{V}_N(a, b, c) \ := \ N^{-1} \sum_{i=1}^{N} \left[a(\mathbf{X}_i) - c + \frac{R_i}{b(\mathbf{X}_i)} \{Y_i - a(\mathbf{X}_i)\} \right]^2.$$

Theorem 2.3. *(a) Let Assumptions in Theorem 2.1 hold. Then, as $N \to \infty$, $\widehat{V}_N(\mu) = V_N(\mu)\{1 + o_p(1)\}$. (b) Let Assumptions (a) of Theorem 2.2 hold. Further let Assumption 2.4 hold and*

$$E\left[\frac{a_N}{\pi_N(\mathbf{X})} \left\{ 1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \right\}^2 \{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - m(\mathbf{X})\}^2 \right] \ = \ o_p(1). \tag{2.14}$$

Then, as $N \to \infty$, $\widehat{V}_N(\mu, e) = V_N(\mu, e)\{1 + o_p(1)\}$.

Notice that we only require a $o_p(1)$ condition in (2.14). Such a condition can be satisfied as long as we have upper bounds for the $(2 + c)$-th moment of the estimation errors and the tail of $\pi_N^{-1}(\mathbf{X})$ is well-behaved. Under a standard positivity condition, when $\mu(\cdot) = m(\cdot)$, (2.14) only requires $r_{\mu,N} = o(1)$, which would have been already assumed for consistency.

Under the conditions in Theorem 2.3, asymptotically valid $100(1-\alpha)\%$ confidence intervals (CIs) for $\tilde{\theta}$ and $\widehat{\theta}_{\text{DRSS}}$ at any significance level α can now be obtained as:

$$\begin{aligned}\text{CI}(\tilde{\theta}) &:= \left(\tilde{\theta} - N^{-1/2}\widehat{V}_N^{1/2}(\mu)z_{1-\alpha/2},\ \tilde{\theta} + N^{-1/2}\widehat{V}_N^{1/2}(\mu)z_{1-\alpha/2}\right), \\ \text{CI}(\widehat{\theta}_{\text{DRSS}}) &:= \left(\widehat{\theta}_{\text{DRSS}} - N^{-1/2}\widehat{V}_N^{1/2}(\mu,e)z_{1-\alpha/2},\ \widehat{\theta}_{\text{DRSS}} + N^{-1/2}\widehat{V}_N^{1/2}(\mu,e)z_{1-\alpha/2}\right), \quad (2.15)\end{aligned}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution. As shown in Theorems 2.1 and 2.2, $V_N(\mu) \asymp a_N^{-1}$ and $V_N(\mu,e) \asymp a_N^{-1}$. Hence, the length of the proposed confidence intervals are of the order $(Na_N)^{-1/2}$.

It is important to note, that these confidence intervals are valid when both the outcome and propensity score models are correctly specified. Whenever the outcome model is misspecified, they need further adjustment based on an adjusted RAL expansion as discussed in Remark 2.9. Based on the adjusted IF, $\Psi(\mathbf{Z})$, therein, one can estimate the asymptotic variance $\text{Var}\{\Psi(\mathbf{Z})\}$ using a plug-in estimate $N^{-1}\sum_{i=1}^{N}\widehat{\Psi}^2(\mathbf{Z}_i;\widehat{\theta}_{\text{DRSS}})$, where $\widehat{\Psi}(\cdot;\widehat{\theta}_{\text{DRSS}})$ is a consistent estimator of $\Psi(\cdot)$, and obtain the corresponding *adjusted* confidence intervals. We also illustrate the numerical performance of these adjusted confidence intervals in Section 2.6.4.

2.4 Decaying PS models

In Section 2.3, we proposed a DR estimator $\widehat{\theta}_{\text{DRSS}}$ of $\theta_0 = E(Y)$. Such an estimator is based on an outcome estimator $\widehat{m}(\cdot)$ and a PS estimator $\widehat{\pi}_N(\cdot)$. Due to the decaying nature of the PS, the estimation of $\pi_N(\cdot)$ itself is also an interesting and challenging problem. In this section, we illustrate three decaying PS models: (i) an *offset logistic model* (Section 2.4.1), (ii) a *stratified labeling model* (Section 2.4.4), and (iii) a *MCAR labeling model* (Section

2.4.5). These are just some natural examples of modeling a decaying PS – our main results are completely general. We propose PS estimators under each of the three models and establish detailed asymptotic results, especially for the offset logistic model (in both low and high dimensions). Moreover, as discussed in Remark 2.9, for a misspecified $m(\cdot)$, based on a case by case study of $\pi_N(\cdot)$, we further construct an adjusted RAL expansion of $\widehat{\theta}_{\text{DRSS}}$ and hence provide an asymptotic normality with an adjusted asymptotic variance.

2.4.1 Offset logistic regression

In this section, we propose a parametric logistic model for extremely unbalanced outcomes, i.e., $\pi_N = P(R=1) \to 0$ as $N \to \infty$, where we let

$$\pi_N(\mathbf{X}) \;=\; \pi_N \frac{\exp(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)}{1+\pi_N \exp(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)} \;=\; \frac{\exp\{\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 + \log(\pi_N)\}}{1+\exp\{\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 + \log(\pi_N)\}}, \tag{2.16}$$

where $\overrightarrow{\mathbf{X}} := (1, \mathbf{X}^T)^T$ and the parameter $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p+1}$ possibly depends on N with $\|\boldsymbol{\gamma}_0\|_2 < C$ for some constant $C > 0$. This model is fairly natural and allows for a general way to incorporate the decaying nature of the labeling fraction. At the same time, it ensures that the dependence of $\pi_N(\mathbf{X})$ on $\mathbf{X}$ is not distorted by the decaying nature of π_N. Model (2.16) could also be viewed as a logistic model with $\log(\pi_N)$ (a diverging negative intercept) as an *offset*. If a standard logistic model is used [Owe07, Wan20] instead, i.e., we let

$$\pi_N(\mathbf{X}) \;=\; g(\overrightarrow{\mathbf{X}}^T \boldsymbol{\beta}), \quad \text{where } g(u) \;:=\; \frac{\exp(u)}{1+\exp(u)}, \tag{2.17}$$

then under some standard conditions whenever an extreme imbalance exists, $\exp(-\boldsymbol{\beta}(1)) \asymp \pi_N^{-1} \to \infty$ whenever $N \to \infty$; see Remark 2.11 for further details. This provides a clear justification for our offset model (2.16) where we precisely extract out $\log(\pi_N)$ as an offset

to be estimated separately and plugged in apriori to the likelihood equation. In this way, we are able to treat the auxiliary intercept and the slope as well-behaved, i.e., finite and independent or bounded in N.

Remark 2.10 (Connections with density ratio estimation)**.** *There is an intricate connection between the offset model* (2.16) *and a model for density ratios usually used in the covariate shift literature where R_is are treated as fixed (or conditioned on) and $P_{\mathbf{X}} \neq P_{\mathbf{X}|R=1}$ is allowed [KK13, LZC20]. Observe that*

$$\textit{logit}\{\pi_N(\mathbf{X})\} \;=\; \log(\pi_N) - \log(1-\pi_N) - \log\{\Lambda_N(\mathbf{X})\},$$

where $\Lambda_N(\mathbf{X}) := f(\mathbf{X}|R=0)/f(\mathbf{X}|R=1)$ and $f(\cdot|R=\cdot)$ is the conditional density of $\mathbf{X}$ given R. However, direct estimation of density ratios is often arduous. The above representation, however, suggests that the same *model can be fitted by a simple logistic regression of $R|\mathbf{X}$, and further using* $\log\{\pi_N/(1-\pi_N)\}$ *as an* offset*. Therefore, missing data literature related to density ratios can now be enriched with an effective estimation of the decaying PS; see Section 4 of [KM20] where semi-parametric efficiency is established but no estimator is discussed. In some sense, such a density ratio estimator is also optimal [Qin98].*

Remark 2.11 (Rationale behind the offset model (2.16) and its connections with a diverging intercept model)**.** *Instead of an offset based model as in* (2.16)*, let us now directly consider a standard logistic model for $P(R=1|\mathbf{X}) = \pi_N(\mathbf{X})$ but allowing (necessarily) for a diverging intercept, given by:*

$$\pi_N(\mathbf{X}) \;=\; g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}) \;=\; \frac{\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}^T\boldsymbol{\beta})}, \quad \textit{where} \tag{2.18}$$

$\boldsymbol{\beta} = (\boldsymbol{\beta}(1), \boldsymbol{\beta}(-1)^T)^T \in \mathbb{R}^{p+1}$ *is a vector allowed to depend on* N*; e.g., see [Owe07] and [Wan20]. For further simplification, let us assume that the slope* $\boldsymbol{\beta}(-1)$*, while allowed to depend on* N*, is finite, i.e.,* $\|\boldsymbol{\beta}(-1)\|_2 < C < \infty$ *for some* C *independent of* N*.*

Under the model (2.18)*, the following holds. Let* $MGF_{\mathbf{X}}(\mathbf{v}) := E\{\exp(\mathbf{v}^T\mathbf{X})\}$ *denote the moment generating function (MGF) of* $\mathbf{X}$ *at* $\mathbf{v} \in \mathbb{R}^p$ *and assume* $MGF_{\mathbf{X}}(\mathbf{v})$ *exists (i.e., finite) at* $\mathbf{v} = \boldsymbol{\beta}(-1)$ *and* $\mathbf{v} = -\boldsymbol{\beta}(-1)$*. Then, the following holds for the intercept* $\boldsymbol{\beta}(1)$*:*

$$\frac{1}{\pi_N}\frac{1-\pi_N}{MGF_{\mathbf{X}}(-\boldsymbol{\beta}(-1))} \;\leq\; \exp(-\boldsymbol{\beta}(1)) \;\leq\; \frac{1}{\pi_N} MGF_{\mathbf{X}}(\boldsymbol{\beta}(-1)), \quad \textit{and consequently,} \tag{2.19}$$

$$\frac{1}{\pi_N}\frac{1-\pi_N}{E\{\exp(\|\boldsymbol{\beta}(-1)\|_2\|\mathbf{X}\|_2)\}} \;\leq\; \exp(-\boldsymbol{\beta}(1)) \;\leq\; \frac{1}{\pi_N} E\{\exp(\|\boldsymbol{\beta}(-1)\|_2\|\mathbf{X}\|_2)\}. \tag{2.20}$$

For the special case of a Gaussian $\mathbf{X}$*, i.e.,* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$*,*

$$MGF_{\mathbf{X}}(-\boldsymbol{\beta}(-1)) = MGF_{\mathbf{X}}(\boldsymbol{\beta}(-1)) \leq \exp\{\|\boldsymbol{\beta}(-1)\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma})\}.$$

Hence, as long as $\|\boldsymbol{\beta}(-1)\|_2^2 < C < \infty$ *and* $\lambda_{\max}(\boldsymbol{\Sigma}) < \infty$*, then using* (2.19)*,* $\exp(-\boldsymbol{\beta}(1)) \asymp \pi_N^{-1} \to \infty$*. More generally, if* $\|\boldsymbol{\beta}(-1)\|_2^2 < C < \infty$ *and* $E\{\exp(C\|\mathbf{X}\|_2)\} < \infty$ *(e.g., if* $\mathbf{X}$ *is sub-Gaussian), then using* (2.20)*, we will have* $\exp(-\boldsymbol{\beta}(1)) \asymp \pi_N^{-1} \to \infty$*.*

Rationale for the offset model (2.16)**.** *The result clearly shows that the intercept* $\boldsymbol{\beta}(1)$ *diverges to* $-\infty$ *and does so* precisely *at a rate of* $\log(\pi_N)$*, i.e.,* $c_1 + \log(\pi_N) \leq \boldsymbol{\beta}(1) \leq C_1 + \log(\pi_N)$*. This provides a clear justification for our offset based model* (2.16) *where we precisely extract out this* $\log(\pi_N)$ *as an offset (to be estimated separately and plugged in apriori to the sample likelihood equation), and then treat the intercept* α_0 *and the slope parameter* $\boldsymbol{\beta}_0$ *to be well-behaved, i.e., finite and independent of* N *(or at least bounded in* N*).*

This makes the parameter space more amenable to theoretical analysis where it is common practice to assume that the truths (the true unconstrained *minimizers) lie as interior points of some compact set. Such assumptions are commonplace in most of empirical process and M-estimation theory, and these results won't be applicable without this assumption, something that has clear justification under the offset model but not under the diverging intercept model.*

Remark 2.12 (Connections with density ratio estimation)**.** *It is interesting (though elementary) to note that the PS is also related to the density ratio of* $\mathbf{X}$ *(given* $R = 0$ *or* 1*), in that*

$$\Lambda_N(\mathbf{X}) \;:=\; \frac{f(\mathbf{X}|R=0)}{f(\mathbf{X}|R=1)} \;=\; \frac{P(R=0|\mathbf{X})P(R=1)}{P(R=1|\mathbf{X})P(R=0)} \;=\; \frac{\{1-\pi_N(\mathbf{X})\}\pi_N}{\pi_N(\mathbf{X})(1-\pi_N)},$$

where $f(\cdot|R=\cdot)$ *is the conditional density function of* $\mathbf{X}$ *given* R*. The density ratio is usually used in the so-called "covariate shift" setting in semi-supervised learning (SSL) and missing data, where* R_i*'s are treated as fixed (or conditioned on) and* $P_{\mathbf{X}} \neq P_{\mathbf{X}|R=1}$ *is allowed; see for example [KK13], [LZC20], and Section 4 of [KM20].*

Here, we discuss a simple and fairly obvious connection of the offset model (2.16) *to a corresponding model for density ratio estimation. The analysis here can actually be seen to be model-free and non-parametric. Observe that*

$$\textit{logit}\{\pi_N(\mathbf{X})\} \;=\; \log(\pi_N) - \log(1-\pi_N) - \log\{\Lambda_N(\mathbf{X})\}.$$

The standard approach to modeling the density ratio is to model $\log\{\Lambda_N(\mathbf{X})\}$ *through basis function expansion based on some basis functions* $\{\phi_j(\mathbf{X})\}_{j=1}^d$ *(e.g., the linear bases will lead to standard parametric forms). But this in general can be difficult to implement in practice. However, the above representation suggests that the* same *model can be fitted by simply using*

a logistic regression model for $R|\mathbf{X}$ *with covariates as the* same *basis functions, and further using* $\log\{\pi_N/(1-\pi_N)\}$ *as an* offset *(which can be estimated separately and plugged in apriori into the likelihood equation). This provides a simple and flexible regression modelling approach to estimate the density ratio. Our offset based model* (2.16) *precisely implements such a model (albeit we had different motivations to consider it), and therefore provides a way to estimate the density ratio as well. This is a key quantity involved (as a nuisance function) in the semi-parametric efficiency bound for our parameter; see Theorem 4.1 of [KM20]. Our approach provides an automated and agnostic way of bypassing its estimation through a theoretically equivalent but practically more flexible regression modeling approach.*

A discussion similar to above can be found in Section 1 of [Qin98] who further proves that the estimation approach as above corresponds to an optimal choice of the estimating equation for estimating the density ratio among the class of all such equations. It is also interesting to note that the semi-supervised (SS) setting actually bears a very close relation to so-called case-control study designs *(which are* retrospective *designs, as opposed to the* prospective *cohort studies that we usually consider), since here the labeling indicator* R *is typically treated as* non-random (or conditioned)*, which is similar in spirit to case-control designs (with* R *being replaced by case/control status). For statistical analyses of these kind of studies, density ratio estimation models are often required and an estimation strategy via a logistic regression model of the PS, similar as above, is often employed; see Section 1 of [Qin98] for more discussions.*

Remark 2.13 (Connection with the maximum likelihood estimate (MLE) of the model (2.18))**.** *In fact, there is an one-one correspondence between* $\widehat{\boldsymbol{\gamma}}$ *(we suppress the dependence*

on k for a moment) and the MLE of the model (2.18)*: if* $(\widehat{\boldsymbol{\beta}}(1), \widehat{\boldsymbol{\beta}}(-1))$ *denotes a sample MLE, i.e., a solution (assuming it exists) to the (sample) likelihood equation for the model* (2.18)*, then* $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\beta}}(1) - \log \widehat{\pi}_N, \widehat{\boldsymbol{\beta}}(-1))$ *is a sample MLE for the model* (2.16)*. Conversely, if* $\widehat{\boldsymbol{\gamma}}$ *is a sample MLE for the model* (2.16)*, then* $(\widehat{\boldsymbol{\beta}}(1), \widehat{\boldsymbol{\beta}}(-1)) = (\widehat{\boldsymbol{\gamma}}(1) + \log(\widehat{\pi}_N), \widehat{\boldsymbol{\gamma}}(-1))$ *is a sample MLE for the model* (2.18)*. All these claims are straightforward to show by means of direct verification.*

Remark 2.14 (Existence and uniqueness of $\widehat{\boldsymbol{\gamma}}$)**.** *The uniqueness of* $\widehat{\boldsymbol{\gamma}}$ *is a direct consequence of the convexity of the sample log-likelihood. As for the existence, we appeal to the one-one correspondence between* $\widehat{\boldsymbol{\gamma}}$ *and the sample MLE of the model* (2.18)*. We further use the results of [Owe07] who demonstrated the existence of the sample MLE for the model* (2.18) *under a fairly mild (sample) overlap condition; see Lemma 5 therein. Note that [Owe07] shows this result for a slightly modified version of the log-likelihood wherein the empirical average over unlabeled data is replaced by an expectation (assuming N is very large). But the same proof technique could be applied to the actual log-likelihood along with a corresponding appropriate modification of the (sample) overlap condition to conclude the existence of the sample MLE for model* (2.18)*. Consequently, this also establishes the existence of the sample MLE for the offset model* (2.16)*.*

Remark 2.15 (Comparison with alternative estimators based on under-sampling)**.** *Under the decaying PS model, a possible alternative to our offset logistic model based estimator could be the "under-sampled" estimators of [Wan20]. Such estimators are constructed based on an under-sampling of the (large sized) unlabeled data. Since the under-sampled data is biased (as the under-sampling is done only for one group, i.e., the unlabeled group), addi-*

tional bias correction or weight adjustment is needed. One can improve the computational efficiency and reduce the storage requirement by considering the under-sampled estimators. However, as discussed in Remarks 3 and 4 of [Wan20], asymptotically, the under-sampled estimators suffer from a loss of efficiency unless the remaining unlabeled data size (after under-sampling) still dominates the labeled data size, which essentially brings us back to the original decaying PS issue. Additionally, the under-sampled estimators have only been discussed in low dimensions, and their high-dimensional alternatives still need to be properly studied.

2.4.2 Low-dimensional offset logistic regression

Let's consider the case of $p < \infty$. We propose a PS estimator $\widehat{\pi}_N(\cdot)$ for the offset model (2.16), based on the full sample $\mathbb{S}$ and use its cross-fitted version (based on a subsample $\mathbb{S}_{-k}$) to construct the DR mean estimator $\widehat{\theta}_{\mathrm{DRSS}}$.

We construct $\widehat{\pi}_N(\cdot)$ based on an apriori chosen estimate $\widehat{\pi}_N := N^{-1}\sum_{i=1}^N R_i$. Let $\widehat{\boldsymbol{\gamma}}$ be the minimizer of $\ell_N(\boldsymbol{\gamma}; \widehat{\pi}_N)$, where

$$\ell_N(\boldsymbol{\gamma}; a) \;:=\; -N^{-1}\sum_{i=1}^N \left[R_i \overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma} - \log\{1 + a\exp(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma})\}\right], \tag{2.21}$$

where recall that $\overrightarrow{\mathbf{X}} = (1, \mathbf{X})^T$. Then, the PS estimate , $\widehat{\pi}_N(\cdot)$, can be obtained by plugging $\widehat{\pi}_N$ into (2.16), as follows:

$$\widehat{\pi}_N(\mathbf{X}) \;:=\; \frac{\widehat{\pi}_N \exp(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\gamma}})}{1 + \widehat{\pi}_N \exp(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\gamma}})}. \tag{2.22}$$

Here, for any $a \in (0, 1]$, $\ell_N(\boldsymbol{\gamma}; a)$ is the negative log-likelihood under the offset based model, up to a term $-N^{-1}\sum_{i=1}^N R_i \log(a)$ that is independent of $\boldsymbol{\gamma}$. Existence and uniqueness of $\widehat{\boldsymbol{\gamma}}$ has been discussed in detail in Remark 2.14. It is worth mentioning that the results

of [Owe07] showcasing the existence of the MLE for the model (2.17) can be extended to guarantee the existence of $\widehat{\boldsymbol{\gamma}}$ as well.

The following theorem provides asymptotic results for $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\pi}_N(\cdot)$, as well as an adjusted RAL expansion of the DRSS estimator $\widehat{\theta}_{\text{DRSS}}$ in low-dimensional setting with p being fixed and when $m(\cdot)$ is possibly misspecified. For this result alone we consider the following conditions on the design: $E\{\exp(t\|\mathbf{X}\|_2)\} < \infty$ for any $t > 0$, $\lambda_{\min}\left[E\{\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\}\right] > 0$, where $g(\cdot)$ was defined in (2.17) and $\dot{g}(\cdot) = g(\cdot)\{1 - g(\cdot)\}$ is the derivative of $g(\cdot)$.

Theorem 2.4. *Let $N\pi_N \to \infty$ as $N \to \infty$, and $\|\boldsymbol{\gamma}_0\|_2 < C < \infty$ where $\boldsymbol{\gamma}_0$ was defined in (2.16). Suppose that $\|[E\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T\}]^{-1}\|_2 < C$ with some constant $C > 0$. Then, as $N \to \infty$,*

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \;=\; N^{-1}\sum_{i=1}^{N}\text{IF}_{\boldsymbol{\gamma}}(\mathbf{Z}_i) + \mathbf{R}_N, \quad \textit{with } \|\mathbf{R}_N\|_2 \;=\; o_p\left((N\pi_N)^{-1/2}\right),$$

$$\text{IF}_{\boldsymbol{\gamma}}(\mathbf{Z}) \;:=\; \mathcal{J}^{-1}(\boldsymbol{\gamma}_0, \pi_N)\{R_i - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_N))\}\overrightarrow{\mathbf{X}} - (\pi_N^{-1}R - 1)\mathbf{e}_1,$$

where $\mathbf{e}_1 := (1, 0, \ldots, 0)^T \in \mathbb{R}^{p+1}$, $\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N) := E\{\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_N))\}$, and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = O_p((N\pi_N)^{-1/2})$. Further, we also have the following error rates:

$$\|\pi_N^{-1}(\mathbf{X})\|_{r,P_{\mathbf{X}}} \;\asymp\; \pi_N^{-1} \quad \forall r > 0, \quad \textit{and hence } a_N \;\asymp\; \pi_N,$$

$$\left\|1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\|_{2,P_{\mathbf{X}}} \;=\; O_p\left((N\pi_N)^{-1/2}\right), \tag{2.23}$$

$$E_{\mathbf{X}}\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2\right] \;=\; O_p\left((N\pi_N)^{-1}\right) \;=\; o_p(1). \tag{2.24}$$

If we further assume that $\|m(\cdot) - \mu(\cdot)\|_{2+c,P_{\mathbf{X}}} < \infty$, then we have a RAL expansion of the

term $\widehat{\Delta}_N$ *defined in* (2.11) *as follows:*

$$\widehat{\Delta}_N \;:=\; N^{-1}\sum_{i=1}^{N}\mathrm{IF}_{\pi}(\mathbf{Z}_i)+o_p\left((N\pi_N)^{-1/2}\right), \quad \textit{where} \tag{2.25}$$

$$\mathrm{IF}_{\pi}(\mathbf{Z}) \;:=\; E\left[\{1-\pi_N(\mathbf{X})\}\{\mu(\mathbf{X})-m(\mathbf{X})\}\overrightarrow{\mathbf{X}}^T\right]\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\overrightarrow{\mathbf{X}}\{R-\pi_N(\mathbf{X})\}. \tag{2.26}$$

Moreover, if we assume $\|\widehat{m}(\cdot)-\mu(\cdot)\|_{2+c,P_{\mathbf{X}}}=o_p(1)$ *(we suppressed the dependency of* $\widehat{m}(\cdot)$ *on* k *as in Theorem 2.2), then we have the following rate:*

$$E_{\mathbf{X}}\left[\frac{a_N}{\pi_N(\mathbf{X})}\{\widehat{m}(\mathbf{X})-\mu(\mathbf{X})\}^2\right] \;=\; o_p(1), \tag{2.27}$$

and with it a RAL expansion of $\widehat{\theta}_{\mathrm{DRSS}}$ *as:*

$$\widehat{\theta}_{\mathrm{DRSS}}-\theta_0 \;=\; N^{-1}\sum_{i=1}^{N}\Psi(\mathbf{Z}_i)+o_p\left(\frac{1}{\sqrt{N\pi_N}}\right), \;\; \textit{where}\; \Psi(\mathbf{Z}) \;:=\; \psi_{\mu,\pi}(\mathbf{Z})+\mathrm{IF}_{\pi}(\mathbf{Z}), \tag{2.28}$$

and $\psi_{\mu,\pi}(\mathbf{Z})$ *is defined in* (2.3). *Lastly,* $E\{\Psi(\mathbf{Z})\}=0$, $E\{\Psi^2(\mathbf{Z})\}=O(\pi_N^{-1})$.

The displays (2.23), (2.24) and (2.27) are the conditions we need to guarantee the assumptions of Theorem 2.2, while the result (2.25) on $\widehat{\Delta}_N$ helps characterize the full RAL expansion of $\widehat{\theta}_{\mathrm{DRSS}}$ under misspecification of $\widehat{m}(\cdot)$. Lastly, notice that we do not assume $\pi_N(\mathbf{X})/\pi_N$ to be bounded bellow a.s., which is a condition required in [KM20].

Remark 2.16 (Necessity of the RAL expansion's modification)**.** *When* $\pi_N\to 0$*, we observe that part of the additional IF,* $\mathrm{IF}_{\pi}(\mathbf{Z})$*, in* (2.26) *has the following property:*

$$E\left[\{1-\pi_N(\mathbf{X})\}\{\mu(\mathbf{X})-m(\mathbf{X})\}\overrightarrow{\mathbf{X}}\right] \;=\; E\left[\{\mu(\mathbf{X})-m(\mathbf{X})\}\overrightarrow{\mathbf{X}}\right]+O_p(\pi_N).$$

If the outcome model is fitted by a linear model whose limit has a linear form $\mu(\mathbf{X})=\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}^*$*, with* $\boldsymbol{\beta}^*:=\{E(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T)\}^{-1}E(\overrightarrow{\mathbf{X}}Y)$*, then,*

$$E\left[\{\mu(\mathbf{X})-m(\mathbf{X})\}\overrightarrow{\mathbf{X}}\right]=\boldsymbol{\beta}^{*T}E(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T)-E(\overrightarrow{\mathbf{X}}Y)=0,$$

indicating that the RAL expansion's modification is unnecessary when $\pi_N \to 0$ *and* $\widehat{m}(\cdot)$ *converges to the linear projection* $\mu(\cdot)$*. Here,* $\mu(\cdot) \neq m(\cdot)$*.The same argument holds if one performs a linear transformation on some basis function* $\{\phi_j(\mathbf{X})\}_{j=1}^d$ *with a fixed* $d < \infty$*. However, when* d *grows with* N *in that* $d/N \to c \in (0, 1)$*, a least squares estimator leads to a latent misspecification i.e., the limit* $\mu(\cdot) \neq m(\cdot)$ *even if* $m(\cdot)$ *is indeed linear on* $\{\phi_j(\mathbf{X})\}_{j=1}^d$*. Hence, an adjusted RAL would be more appropriate if the outcome model is linear with a growing degree of freedom; see Section 2.6.4 for corresponding simulation results.*

A possible alternative to our offset logistic regression model based estimators could be the so called "under-sampled" estimators as studied by [Wan20], where the observations from the large unlabeled data are under-sampled in some way to create a more "balanced" setting. However, such an approach may have several disadvantages; see Remark 2.15 for comparisons and a more detailed discussion of this approach. Moreover, [Wan20] only considered such estimators in low dimensions; whereas, we provide a thorough analysis of our model in both low and high dimensional settings (as in Section 2.4.3 next). The results for the latter case, in particular, are fairly interesting and non-trivial, and possibly the first such results extending existing results on high-dimensional logistic regression.

2.4.3 High-dimensional offset logistic regression

Next, we consider a high-dimensional setting with $p \to \infty$. The problem here is challenging as together with $p \to \infty$, the labels are extremely imbalanced in that $\pi_N = P(R = 1) \to 0$. Unlike before, an adjusted RAL expansion for the case when $m(\cdot)$ is misspecified is now not available, as we are no longer able to obtain a parametric rate for

the PS estimation. In this section, we provide the consistency rate $r_{e,N}$ in (2.9) for an offset, sparse, logistic PS model and establish asymptotic results for $\widehat{\theta}_{\text{DRSS}}$ when both $m(\cdot)$ and $\pi_N(\cdot)$ are correctly specified.

Consider the same parametric offset model (2.16), except here we allow $p \to \infty$ as $N \to \infty$. In this subsection, we assume the parameter $\boldsymbol{\gamma}_0$ to be sparse with $s := \|\boldsymbol{\gamma}_0\|_0$ denoting its sparsity level. Let $\widehat{\pi}_N := N^{-1} \sum_{i=1}^N R_i$ and for every $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ and $a \in (0, 1]$, recall $\ell_N(\boldsymbol{\gamma}; a)$ defined in (2.21). Let $\widehat{\boldsymbol{\gamma}}$ be a minimizer of the convex program:

$$\arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}} \left\{ \ell_N(\boldsymbol{\gamma}; \widehat{\pi}_N) + \lambda_N \|\boldsymbol{\gamma}\|_1 \right\}, \tag{2.29}$$

with a sequence $\lambda_N > 0$. Then, $\pi_N(\mathbf{X})$ can be estimated similarly as in (2.22) by $\widehat{\pi}_N(\mathbf{X}) := g(\overrightarrow{\mathbf{X}}^T \widehat{\boldsymbol{\gamma}} + \log(\widehat{\pi}_N))$. We establish the theoretical properties of our estimators $\widehat{\boldsymbol{\gamma}}$ and $\hat{\pi}_N(\cdot)$ in 3 parts: 1) establish a restricted strong convexity (RSC) property; 2) control the l_∞ norm of the gradient of the loss at the true parameter, i.e., $\|\nabla_{\boldsymbol{\gamma}} \ell_N(\boldsymbol{\gamma}_0; \widehat{\pi}_N)\|_\infty$; and 3) obtain the final probabilistic bounds on the error rates of our estimator.

RSC property for the offset logistic model We first analyze the RSC property of our high dimensional offset logistic model. Under our imbalanced treatment setting, we show that the RSC condition holds with a parameter of the order of $\pi_N \to 0$ (rather than a constant bounded away from 0), once the RSC condition holds for a balanced logistic model with some constant $\kappa > 0$. For any $\boldsymbol{\Delta}, \boldsymbol{\gamma} \in \mathbb{R}^{p+1}$, define the following:

$$\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}) \;:=\; \ell_N(\boldsymbol{\gamma} + \boldsymbol{\Delta}; a) - \ell_N(\boldsymbol{\gamma}; a) - \boldsymbol{\Delta}^T \nabla_{\boldsymbol{\gamma}} \ell_N(\boldsymbol{\gamma}; a). \tag{2.30}$$

We say the restricted strong convexity (RSC) property holds for $\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}_0)$ with parameter κ on a given set A if

$$\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}_0) \ \geq \ \kappa\|\boldsymbol{\Delta}\|_2^2, \qquad \text{for all } \ \boldsymbol{\Delta} \in A. \tag{2.31}$$

We have the following *deterministic* result.

Lemma 2.1. *For any $a \in (0, 1]$,*

$$\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}_0) \ \geq \ a\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0).$$

Hence, for a given set A and for any given realization of the data, if the RSC property holds for $\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0)$ with parameter κ on a set A, then the RSC property also holds for $\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}_0)$ with parameter $a\kappa$ on A.

Notice that

$$\begin{aligned}
\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0) \ &= \ \ell_N(\boldsymbol{\gamma}_0 + \boldsymbol{\Delta}; 1) - \ell_N(\boldsymbol{\gamma}_0; 1) - \boldsymbol{\Delta}^T\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0; 1) \\
&= \ \ell_N^{\text{bal}}(\boldsymbol{\gamma}_0 + \boldsymbol{\Delta}) - \ell_N^{\text{bal}}(\boldsymbol{\gamma}_0) - \boldsymbol{\Delta}^T\nabla_{\boldsymbol{\gamma}}\ell_N^{\text{bal}}(\boldsymbol{\gamma}_0), \quad \text{where} \\
\ell_N^{\text{bal}}(\boldsymbol{\gamma}) \ &:= \ -N^{-1}\sum_{i=1}^{N}[R_i^*\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma} - \log\{1 + \exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma})\}], \ \ \forall\boldsymbol{\gamma} \in \mathbb{R}^{p+1},
\end{aligned}$$

with $(R_i^*)_{i=1}^N$ being i.i.d. random variables generated from Bernoulli$(g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0))$. Here, $\ell_N^{\text{bal}}(\boldsymbol{\gamma})$ is the negative log-likelihood function under a *balanced* logistic model with the true parameter $\boldsymbol{\gamma}_0$. By Lemma 2.1, we relate the RSC property of our imbalanced model to a standard balanced logistic model. The RSC property for a balanced logistic model has been studied in [NRWY10], among others. We also present a more general version in this chapter; see Lemma 2.3.

Gradient control Now, we control the l_∞ norm of the gradient, $\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)\|_\infty$, and the following lemma demonstrates that the rate of $\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)\|_\infty = O_p(\{N^{-1}\pi_N\log(p)\}^{1/2})$.

Lemma 2.2. *Let $\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0$ be a sub-Gaussian random variable and $\overrightarrow{\mathbf{X}}$ a marginal sub-Gaussian random vector, in that $\|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0\|_{\psi_2} \le \sigma_{\boldsymbol{\gamma}_0} < \infty$ and $\max_{1\le j\le p+1}\|\overrightarrow{\mathbf{X}}(j)\|_{\psi_2} \le \sigma < \infty$, respectively. Then, for any $t_1, t_2 \ge 0$ and $t_2 < N\pi_N/9$,*

$$\begin{aligned}\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)\|_\infty \;\le\; & C_1(\pi_N+\pi_N^{1/2})\sqrt{\frac{\{t_1+\log(p+1)\}}{N}} + C_4\left\{\sqrt{\frac{t_2\pi_N}{N}}+\frac{t_2}{N}\right\} \\ & + (C_2+C_3\pi_N)\frac{\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N},\end{aligned}$$

with probability at least $1-6\exp(-t_1)-2\exp(-t_2)$. The constants $C_1, C_2, C_3, C_4 > 0$ independent of N are defined through equations (2.113)-(2.114).

Define $S := \{j \le p+1 : \boldsymbol{\gamma}_0(j) \ne 0\}$, $s = |S|$ and the *cone set*:

$$\mathbb{C}_\delta(S;3) \;:=\; \{\boldsymbol{\Delta}\in\mathbb{R}^{p+1} : \|\boldsymbol{\Delta}_{S^c}\|_1 \le 3\|\boldsymbol{\Delta}_S\|_1,\;\; \|\boldsymbol{\Delta}\|_2 = \delta\}, \tag{2.32}$$

where $\boldsymbol{\Delta}_S = \{\boldsymbol{\Delta}(j)\}_{j\in S}$ and $\boldsymbol{\Delta}_{S^c} = \{\boldsymbol{\Delta}(j)\}_{j\notin S}$. Define the *critical tolerance*:

$$\delta_N := \inf\left\{\delta>0 : \delta \ge 2\lambda_N s^{1/2}\tilde{\kappa}^{-1},\ \text{RSC holds for } \ell_N(\cdot;\widehat{\pi}_N) \text{ with } \tilde{\kappa} \text{ over } \mathbb{C}_\delta(S;3)\right\}.$$

Then, any optimal solution $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_{\lambda_N}$ to the convex program (2.29) satisfies $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2 \le \delta_N$.

Probabilistic bounds Finally, we now obtain the probabilistic bounds and convergence rates for $\widehat{\boldsymbol{\gamma}}$, and subsequently $\widehat{\pi}_N(\cdot)$, in the following result.

Theorem 2.5. *Assume $\log(p)\log(N) = O(N\pi_N)$ and $s\log(p) = o(N\pi_N)$ as $N, p \to \infty$, where $s := \|\boldsymbol{\gamma}_0\|_0$. Assume conditions in Lemma 2.2. Suppose the RSC property holds*

for $\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0)$ *with parameter* $\kappa > 0$ *on the set* $\overline{\mathbb{C}}(S; 3) := \{\boldsymbol{\Delta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\Delta}_S\|_1, \ \|\boldsymbol{\Delta}\|_2 \leq 1\}$, *with probability at least* $1 - \alpha_N$, *where* $\alpha_N = o(1)$. *Let*

$$M_N \ := \ C_5\sqrt{\frac{\pi_N \log(p+1)}{N}} + C_6\frac{\sqrt{\log(2N)}\log(p+1)}{N},$$

with some constants $C_5, C_6 > 0$. *For any* λ_N *satisfying* $2(1+c)M_N \leq \lambda_N \leq 9\kappa\pi_N s^{-1/2}$ *with* $c > 0$, *whenever* $N\pi_N > 9c\log(p+1)$,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \ \leq \ \frac{1}{9}\lambda_N s^{1/2}\pi_N^{-1}\kappa^{-1}, \quad \textit{with probability at least } \ 1 - 8(p+1)^{-c} - \alpha_N.$$

Further assume that $\|\overrightarrow{\mathbf{X}}^T\mathbf{v}\|_{\psi_2} \leq \sigma\|\mathbf{v}\|_2$ *for any* $\mathbf{v} \in \mathbb{R}^{p+1}$. *Then, for any* $r > 0$, *with some* $\lambda_N \asymp \sqrt{\pi_N \log(p)/N}$,

$$\begin{aligned}
&\|\pi_N^{-1}(\mathbf{X})\|_{r,P_\mathbf{X}} \ \asymp \ \pi_N^{-1} \quad \forall r > 0, \quad \textit{and hence } \ a_N \ \asymp \ \pi_N, \\
&\left\|1 - \frac{\pi_N(\cdot)}{\widehat{\pi}_N(\cdot)}\right\|_{r,P_\mathbf{X}} \ = \ O_p\left(\sqrt{\frac{s\log(p)}{N\pi_N}}\right) \quad \forall r > 0, \qquad (2.33) \\
&E_\mathbf{X}\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2\right] \ = \ O_p\left(\frac{s\log(p)}{N\pi_N}\right) \ = \ o_p(1). \qquad (2.34)
\end{aligned}$$

Moreover, if $\|\widehat{m}(\cdot) - m(\cdot)\|_{2+c,P_\mathbf{X}} = o_p(1)$ *with constant* $c > 0$, *then,*

$$E_\mathbf{X}\left[\frac{a_N}{\pi_N(\mathbf{X})}\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}^2\right] \ = \ o_p(1).$$

Remark 2.17 (Non-standard rates)**.** *The implication of Theorem 2.5 is that for some* $\lambda_N \asymp \{N^{-1}\pi_N\log(p)\}^{1/2}$,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \ = \ O_p\left(\sqrt{\frac{s\log(p)}{N\pi_N}}\right). \qquad (2.35)$$

As long as $\pi_N \to 0$, *the rate* $\lambda_N \asymp \{N^{-1}\pi_N\log(p)\}^{1/2}$, *is* faster *than the usual rate of* $\{N^{-1}\log(p)\}^{1/2}$ *used for tuning parameter choice in a standard (i.e., balanced)* ℓ_1*-penalized logistic regression. This in turn implies* slower *than usual rate of convergence in* (2.35), *as*

Na_N is much smaller than N. This is also reflected in the error rates of the conditional propensity score, in (2.33)*. The "effective sample size" here is Na_N rather than N, thus leading to* non-standard *rates. The results above may therefore be seen as a* generalization *of standard high-dimensional logistic regression models (i.e., where positivity holds) to the case of a decaying PS. To our knowledge, these rates are novel for high-dimensional settings.*

Remark 2.18 (Marginal versus "Joint" sub-Gaussianity)**.** *In Theorem 2.5, we obtained a non-asymptotic upper bound for $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2$ that only requires a marginal sub-Gaussianity of $\overrightarrow{\mathbf{X}}$, that is $\max_{1\le j\le p+1} \|\overrightarrow{\mathbf{X}}(j)\|_{\psi_2} \le \sigma < \infty$. Unfortunately, to show* (2.33) *and* (2.34)*, we do require a "joint" sub-Gaussianity of $\overrightarrow{\mathbf{X}}$ in that $\|\overrightarrow{\mathbf{X}}^T\mathbf{v}\|_{\psi_2} \le \sigma\|\mathbf{v}\|_2$ for any $\mathbf{v} \in \mathbb{R}^{p+1}$. In high-dimensions, the joint sub-Gaussianity is stronger than the marginal sub-Gaussianity in that the latter enforces a weaker dependency among the covariates; see Section 4 of [KC18] for more details.*

Note that in Theorem 2.5, we *only* assume the RSC property for a classical balanced logistic regression model, which is standard in the high-dimensional regression (and classification) literature. As shown in Proposition 2 of [NRWY10], with probability at least $1 - 2\exp(-c_1 N)$,

$$\begin{aligned}\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0) &= \ell_N^{\text{bal}}(\boldsymbol{\gamma}_0 + \boldsymbol{\Delta}) - \ell_N^{\text{bal}}(\boldsymbol{\gamma}_0) - \boldsymbol{\Delta}^T\nabla_{\boldsymbol{\gamma}}\ell_N^{\text{bal}}(\boldsymbol{\gamma}_0) \\ &\ge c_2\|\boldsymbol{\Delta}\|_2\left\{\|\Delta\|_2 - c_3\sqrt{\frac{\log(p+1)}{N}}\|\Delta\|_1\right\}, \quad \forall\|\Delta\|_2 \le 1, \end{aligned} \tag{2.36}$$

with some constants $c_1, c_2, c_3 > 0$, and hence, the RSC property holds for $\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0)$ with some $\kappa > 0$ on the set $\overline{\mathbb{C}}(S; 3)$. The conditions required in [NRWY10] essentially amount to: $s\log(p) = o(N)$, the intercept term $\boldsymbol{\gamma}_0(1) = 0$, $\mathbf{X}$ is a jointly sub-Gaussian with mean

zero, and $\lambda_{\min}\{\mathrm{Cov}(\mathbf{X})\} \geq c > 0$. Similar conditions are also required in Example 9.17 and Theorem 9.36 of [Wai19]. In the following Lemma 2.3, we propose a user-friendly version of RSC condition results for a balanced logistic regression problem that only require a marginal sub-Gaussianity of $\mathbf{X}$ and an additional $(2+c)$-th moment condition $\sup_{\|\mathbf{v}\|_2 \leq 1} \|\overrightarrow{\mathbf{X}}^T\mathbf{v}\|_{2+c,P_{\mathbf{X}}} \leq M < \infty$. In addition, we do not enforce a mean zero $\mathbf{X}$, and we do not require a zero intercept term in the logistic model either.

Lemma 2.3. *Assume the smallest eigenvalue* $\lambda_{\min}\{E(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T)\} \geq \kappa_l > 0$, *a* $(2+c)$*-th moment condition* $\sup_{\|\mathbf{v}\|_2 \leq 1} \|\overrightarrow{\mathbf{X}}^T\mathbf{v}\|_{2+c,P_{\mathbf{X}}} \leq M < \infty$, *a c-th moment condition* $\|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0\|_{c,P_{\mathbf{X}}} \leq \mu_c < \infty$ *and the marginal sub-Gaussianity* $\sup_{1\leq j\leq p+1} \|\overrightarrow{\mathbf{X}}(j)\|_{\psi_2} \leq \sigma < \infty$. *Then, with probability at least* $1 - 2\exp(-c_1 N)$, (2.36) *holds, with constants* $c_1, c_2, c_3 > 0$. *If we further assume that* $s\log(p) = o(N)$, *then, for large enough* N, *there exists a constant* $\kappa > 0$ *such that, with probability at least* $1 - 2\exp(-c_1 N)$,

$$\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0) \;\geq\; \kappa\|\Delta\|_2^2, \quad \forall \Delta \in \overline{\mathbb{C}}(S; 3). \tag{2.37}$$

Although Lemma 2.3 is based on the logistic loss function, it in fact applies to any loss function $\ell_N^{\mathrm{bal}}(\cdot)$ based on the maximum likelihood of a balanced generalized linear model.

2.4.4 Stratified labeling

We consider here a stratified labeling mechanism. Here, the labeling indicator R depends on $\mathbf{X}$, but does so only through an intermediate stratification in $\mathbf{X}$. Such mechanisms are often of practical relevance in biomedical studies when *prior* information is available on stratification through another observed variable. Specifically, let $\delta \in \{0, 1\}$ denote an observed random stratum indicator and assume that $R \perp\!\!\!\perp \mathbf{X}|\delta$. Note that nothing changes

if we were to move from binary to finitely many strata, and while we stick to a binary δ here for simplicity, our work can be easily extended to a multiple-stratum situation. Let $\pi_{j,N} := P(R = 1|\delta = j, \mathbf{X}) \equiv P(R = 1|\delta = j)$ for each $j = 0, 1$. We assume δ is a "well behaved" indicator whose distribution is independent of N and itself satisfies the overlap condition

$$c \ < \ p_\delta(\mathbf{x}) \ := \ P(\delta = 1|\mathbf{x}) \ < \ 1 - c, \quad \text{for all } \mathbf{x} \in \mathcal{X},$$

with a constant $0 < c < 1/2$ independent of N. Then, we have:

$$\pi_N(\mathbf{X}) \ = \ \pi_{1,N} p_\delta(\mathbf{X}) + \pi_{0,N}\{1 - p_\delta(\mathbf{X})\}.$$

As long as δ is observed, then $\pi_{j,N}$ for each j can be estimated very easily and at a rate $O_p((N\pi_N)^{-1/2})$. Moreover, when $\pi_N \to 0$ as $N \to \infty$, $p_\delta(\mathbf{X})$ can be estimated at a parametric $N^{-1/2}$ rate if the model is parametric, or at a rate slower than $N^{-1/2}$ but still as a function of N (rather than $N\pi_N$) if a non-parametric estimator is performed. Therefore, we will continue to have a fast enough rate for $\widehat{\pi}_N(\cdot)$ under this setting, so that the error term $\widehat{\Delta}_N$ in (2.11) can potentially have a rate:

$$\widehat{\Delta}_N \ = \ O_p(r_{e,N}) \ = \ O_p\left((N\pi_N)^{-1/2}\right).$$

In this section, we propose a PS estimator based on the stratified labeling model above, and provide a full characterization of its properties as well as a RAL expansion for the error $\widehat{\Delta}_N$.

With a slight abuse of notation, we define $\mathbf{Z}_i = (R_i, R_i Y_i, \delta_i, \mathbf{X}_i)$ in this section, and let $\mathbb{S}$, $\mathbb{S}_{-k}$ be defined accordingly. Suppose $\widehat{p}_\delta(\cdot)$ is an estimator of $p_\delta(\cdot)$, and let $\widehat{p}_\delta(\mathbf{X}_i) := \widehat{p}_\delta(\mathbf{X}_i; \mathbb{S}_{-k(i)})$ be a corresponding cross-fitted version of this estimator.

Define $\widehat{\pi}_1^{-k} =: \sum_{i \notin I_k} \delta_i R_i / \sum_{i \notin I_k} \delta_i$ and $\widehat{\pi}_0^{-k} := \sum_{i \notin I_k} (1 - \delta_i) R_i / \sum_{i \notin I_k} (1 - \delta_i)$ to be

the cross-fitted estimators of $\pi_{1,N}$ and $\pi_{0,N}$, respectively. The PS $\pi_N(\cdot)$ is then estimated by:

$$\widehat{\pi}_N(\mathbf{X}_i) \;:=\; \widehat{\pi}_1^{-k(i)}\widehat{p}_\delta(\mathbf{X}_i) + \widehat{\pi}_0^{-k(i)}\{1-\widehat{p}_\delta(\mathbf{X}_i)\}.$$

Theorem 2.6. *Assume $\pi_N \to 0$ and $N\pi_N \to \infty$ as $N \to \infty$. Suppose*

$$\|\widehat{p}_\delta(\cdot) - p_\delta(\cdot)\|_{2,\mathbb{P}_\mathbf{X}} = O_p(r_{\delta,N}), \qquad \text{for some sequence } \; r_{\delta,N} = o(1). \tag{2.38}$$

Then, for each $k \le \mathbb{K}$,

$$E_\mathbf{X}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2 \;=\; O_p\left(r_{\delta,N}^2 + N\pi_N\right).$$

Besides, assume $\|\mu(\cdot) - m(\cdot)\|_{\infty,\mathbb{P}_\mathbf{X}} < \infty$ and (2.8). Then, the overall RAL expansion of our DRSS estimator $\widehat{\theta}_{\mathrm{DRSS}}$ under a stratified labeling model as above is:

$$\widehat{\theta}_{\mathrm{DRSS}} - \theta_0 \;=\; N^{-1}\sum_{i=1}^N \Psi(\mathbf{Z}_i) + O_p\left((N\pi_N)^{-1} + N^{-1/2} + r_{\mu,N}(N\pi_N)^{-1/2} + r_{\delta,N}\right),$$

where $\Psi(\mathbf{Z}) := \psi_{\mu,\pi}(\mathbf{Z}) + \mathrm{IF}_\pi(\mathbf{Z})$ and $E\{\Psi(\mathbf{Z})\} = 0$ with $\psi_{\mu,\pi}(\mathbf{Z})$ as defined in (2.3) and

$$\begin{aligned}\mathrm{IF}_\pi(\mathbf{Z}) \;:=\; & \left\{\frac{\delta R}{p_\delta} - \pi_{1,N}\right\} E_\mathbf{X}\left[\frac{p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X}) - m(\mathbf{X})\}\right] \\ & + \left\{\frac{(1-\delta)R}{1-p_\delta} - \pi_{0,N}\right\} E_\mathbf{X}\left[\frac{1-p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X}) - m(\mathbf{X})\}\right],\end{aligned}$$

where $p_\delta = E\{p_\delta(\mathbf{X})\} = E(\delta)$. If we further assume $r_{\delta,N} = o((N\pi_N)^{-1/2})$, then

$$\widehat{\theta}_{\mathrm{DRSS}} - \theta_0 \;=\; N^{-1}\sum_{i=1}^N \Psi(\mathbf{Z}_i) + o_p\left((N\pi_N)^{-1/2}\right). \tag{2.39}$$

Note that Theorem 2.6 still holds if π_1 and π_0 are estimated without cross-fitting in that $\widehat{\pi}_1 := \sum_{i=1}^N \delta_i R_i / \sum_{i=1}^N \delta_i$ and $\widehat{\pi}_0 := \sum_{i=1}^N (1-\delta_i) R_i / \sum_{i=1}^N (1-\delta_i)$.

Example 2.2. *Here we illustrate a simple logistic model for $p_\delta(\cdot)$ and investigate the conditions we need for $r_{\delta,N}$ to be $o_p((N\pi_N)^{-1/2})$, so that the RAL expansion (2.39) holds. For*

a fixed dimensional $\mathbf{X}$*, let* $\widehat{p}_\delta(\cdot)$ *be the MLE of the logistic model. Then* $r_{\delta,N} = O(N^{-1/2}) = o((N\pi_N)^{-1/2})$ *as long as* $\pi_N \to 0$*. As for a high-dimensional* $\mathbf{X}$*, consider a sparse logistic model for* $p_\delta(\cdot)$*, and let* $\widehat{p}_\delta(\cdot)$ *be the logistic estimator based on a Lasso penalty. Then,* $r_{\delta,N} = O((s_\delta \log(p)/N)^{1/2})$*, where* s_δ *is the sparsity level of the logistic models parameter. Hence,* $r_{\delta,N} = o_p((N\pi_N)^{-1/2})$ *if* $s_\delta \pi_N \log(p) = o(1)$ *as* $N \to \infty$*.*

Remark 2.19 (Comparisons with other works)**.** *We note that similar, yet different, problems are studied in [GLTC20] and [HLL20]. They both work on decaying stratified labeling propensity score models, but with different types of stratified labeling mechanisms and parameters of interest compared to our setting. [HLL20] assume a deterministic* δ *given* $\mathbf{X}$*. Essentially, they require* $Y \perp\!\!\!\perp R|\delta$*, so that* δ *can be seen as a univariate confounder with a finite support. On the other hand, both [GLTC20] and our work allow additional randomness in* δ*. Besides, [HLL20] work on a "finite-population" ATE estimation problem, where the treatment assignment is the only source of randomness. [GLTC20] focus on the estimation of the regression parameters and prediction performance measures, for low-dimensional covariates and a binary outcome problem; and we are mainly working on the estimation of the mean response while allowing for high-dimensional covariates and real valued outcomes.*

2.4.5 Missing completely at random (MCAR)

Apart from the offset based model and the stratified labeling model discussed in Sections 2.4.1-2.4.4, a simple but commonly used PS model would be a MCAR mechanism. In this section, we consider this special MCAR mechanism with $\pi_N(\cdot) \equiv \pi_N$, and derive the properties of $\widehat{\theta}_{\text{DRSS}}$ including an adjusted regular and asymptotically linear (RAL) expansion

allowing for misspecification of $\widehat{m}(\cdot)$. In this case, a cross-fitted estimator of the PS is proposed as $\widehat{\pi}_N(\mathbf{X}_i) = N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} R_i$ for any $i \in \mathcal{I}_k$, where $N_{-k} := |\mathcal{I}_{-k}|$ and $\mathcal{I}_{-k} := \mathcal{I} \setminus \mathcal{I}_k$. Based on such a MCAR PS estimator, we have the following result on the conditions and conclusions in Theorem 2.2.

Theorem 2.7. *Assume* $\pi_N(\mathbf{X}) \equiv \pi_N$, $N\pi_N \to \infty$ *as* $N \to \infty$, $\|m(\cdot) - \mu(\cdot)\|_{2,P} < \infty$ *and* $\|\widehat{m}(\cdot) - \mu(\cdot)\|_{2,P} = o_p(1)$. *Then,* $a_N = \pi_N$ *and*

$$E\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2\right] \;=\; E\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2 \;=\; O_p\left((N\pi_N)^{-1}\right).$$

Furthermore,

$$\begin{aligned}
\widehat{\theta}_{\mathrm{DRSS}} - \theta_0 \;&=\; N^{-1}\sum_{i=1}^{N} \Psi(\mathbf{Z}_i) + o_p\left((Na_N)^{-1/2}\right), \quad \textit{where } \Psi(\mathbf{Z}) \;:=\; \psi_\mu(\mathbf{Z}) + \mathrm{IF}_\pi(\mathbf{Z}),\\
\psi_\mu(\mathbf{Z}) \;&=\; \frac{R}{\pi_N}\{Y - \mu(\mathbf{X})\} + \mu(\mathbf{X}) - \theta_0,\\
\mathrm{IF}_\pi(\mathbf{Z}) \;&:=\; \left(\frac{R - \pi_N}{\pi_N}\right)\Delta_\mu, \quad \Delta_\mu \;:=\; E\{\mu(\mathbf{X}) - m(\mathbf{X})\}.
\end{aligned}$$

Note that Theorem 2.7 still holds if the PS is estimated without cross-fitting that $\widehat{\pi}_N(\mathbf{X}) \equiv \widehat{\pi}_N = n/N$, where $n = \sum_{i=1}^{N} R_i$.

Remark 2.20. *The modification on the RAL expansion of the mean estimator is needed only when* $\Delta_\mu = E\{\mu(\mathbf{X}) - m(\mathbf{X})\} \neq 0$. *Recall Remark 2.16; if the outcome model is fitted by a linear model that* $\mu(\mathbf{X}) = \overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}^*$, *where* $\overrightarrow{\mathbf{X}} = (1, \mathbf{X}^T)^T$, $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} E\{(Y - \overrightarrow{\mathbf{X}}^T\boldsymbol{\beta})^2\} = \{E(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T)\}^{-1}E(\overrightarrow{\mathbf{X}}Y)$ *is the optimal population slope. Then, we have* $\Delta_\mu = E\{\mu(\mathbf{X}) - m(\mathbf{X})\} = E(\overrightarrow{\mathbf{X}}^T)\{E(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T)\}^{-1}E(\overrightarrow{\mathbf{X}}Y) - E(Y) = 0$. *This suggests that, the RAL modification is unnecessary when* $\widehat{m}(\cdot)$ *converges to the linear projection. In other words, for such cases, the original asymptotic normality* (2.10) *still holds even if* $\mu(\cdot) \neq$

$m(\cdot)$ and it coincides with the results in [ZBC19]. Classical examples for such $\widehat{m}(\cdot)$ include least squares (LS) estimator and regularized least squares such as Lasso and ridge under appropriate conditions.

Reconciliation with "traditional" SS inference literature under MCAR Now, we consider the "traditional" SS setting where all the R_i's are considered *deterministic* (or conditioned) apart from an underlying MCAR assumption. Under this SS setting, we consider the SS mean estimator proposed in [ZB21]. In fact, their estimator is a special case of our double robust SS (DRSS) mean estimator $\widehat{\theta}_{\text{DRSS}}$ except that the PS is estimated without cross-fitting, i.e., $\widehat{\pi}_N(\mathbf{X}) \equiv \widehat{\pi}_N = n/N$. Under this SS setting, we have the following RAL expansion for $\widehat{\theta}_{\text{DRSS}}$:

$$\begin{aligned}\widehat{\theta}_{\text{DRSS}} - \theta_0 &= N^{-1}\sum_{i=1}^{N}\psi_{\mu,\text{SS}}(\mathbf{Z}_i) + o_p(n^{-1/2}), \text{ where} \\ \psi_{\mu,\text{SS}}(\mathbf{Z}) &= \frac{NR_i}{n}[Y_i - \mu(\mathbf{X}_i)] + \mu(\mathbf{X}_i) - \theta_0.\end{aligned}$$

Here, conditional on R_i's, $\{\psi_{\mu,\text{SS}}(\mathbf{Z}_i)\}_{i=1}^N$ are independent and identically distributed, with mean zero.

Now we compare the asymptotic variances for the following three cases: a) R_i's are considered as random (MCAR), π_N is known, and the mean estimator, defined as (2.2), is based on the true PS π_N; b) R_i's are considered as random (MCAR), the mean estimator, defined as (2.5) and studied in Theorem 2.7, is based on the cross-fitted constant estimate that $\widehat{\pi}_N(\mathbf{X}_i) = |\mathbb{S}_{-k}|^{-1}\sum_{i\notin\mathcal{I}_k} R_i$ for $i \in \mathcal{I}_k$; c) R_i's are considered as fixed (SS) and the mean estimator is as defined in [ZB21].

For the above three cases, we have the following asymptotic variances:

$$\begin{aligned}
\mathrm{Var}\{\psi_{\mu,\pi}(\mathbf{Z})\} &= \frac{E\{Y-\mu(\mathbf{X})\}^2}{\pi_N} - \Delta_\mu^2 + \mathrm{Var}\{\mu(\mathbf{X})\} + 2\mathrm{Cov}\{Y-\mu(\mathbf{X}), \mu(\mathbf{X})\}, \\
\mathrm{Var}\{\Psi(\mathbf{Z})\} &= \frac{E\{Y-\mu(\mathbf{X})\}^2}{\pi_N} - \frac{\Delta_\mu^2}{\pi_N} + \mathrm{Var}\{\mu(\mathbf{X})\} + 2\mathrm{Cov}\{Y-\mu(\mathbf{X}), \mu(\mathbf{X})\}, \\
\mathrm{Var}\{\psi_{\mu,\mathrm{SS}}(\mathbf{Z})|R_{\mathcal{I}}\} &= \frac{E\{Y-\mu(\mathbf{X})\}^2}{\widehat{\pi}_N} - \frac{\Delta_\mu^2}{\widehat{\pi}_N} + \mathrm{Var}\{\mu(\mathbf{X})\} + 2\mathrm{Cov}\{Y-\mu(\mathbf{X}), \mu(\mathbf{X})\}.
\end{aligned}$$

We can see that, $\mathrm{Var}\{\psi_\mu(\mathbf{Z})\} = \mathrm{Var}\{\Psi(\mathbf{Z})\} + (\pi_N^{-1} - 1)\Delta_\mu^2 \geq \mathrm{Var}\{\Psi(\mathbf{Z})\}$. It suggests that, under the MCAR setting, even if π_N is known, it is still worth estimating π_N instead of directly plugging in the true value π_N as long as $\Delta_\mu \neq 0$. As for the asymptotic variance under the SS setting, notice the fact that

$$\frac{\pi_N}{\widehat{\pi}_N} - 1 \;=\; O_p\left((N\pi_N)^{-1/2}\right).$$

Hence, $\mathrm{Var}\{\psi_{\mu,\mathrm{SS}}(\mathbf{Z})|R_{\mathcal{I}}\} = \mathrm{Var}\{\Psi(\mathbf{Z})\}\{1 + O_p((N\pi_N)^{-1/2})\} = \mathrm{Var}\{\Psi(\mathbf{Z})\}\{1 + o_p(1)\}$.

2.5 Average treatment effect estimation with imbalanced treatment groups

One important application of our proposed method of Section 2.3 is the popular causal inference problem of ATE estimation and hypothesis testing. Our method is particularly suited when extremely *imbalanced* treatment groups occur. The causal inference literature typically accesses ATE inference by imposing an overlap condition by which $P(c < E(R|\mathbf{X}) < 1 - c) = 1$ for some constant $c > 0$. Here, $R \in \{0, 1\}$ is a binary treatment indicator. In contrast, we show that our method identifies and performs inference about the ATE without requiring an overlap condition. We extend our results for the MAR-SS setting of Section

2.3 to a causal inference setting while allowing a decaying PS in that $\pi_N := E(R) \to 0$ (or alternatively, $\pi_N \to 1$) as $N \to \infty$. To the best of our knowledge, no previous work has addressed such an extremely imbalanced treatment groups setting in the context of ATE estimation.

We formulate the problem setup first. Suppose we have samples $\mathbb{S} := (R_i, Y_i, \mathbf{X}_i)_{i=1}^N$ with $(R, Y, \mathbf{X})$ being an independent copy of $(R_i, Y_i, \mathbf{X}_i)$. Here, $R = R_N \in \{0, 1\}$ is a treatment indicator that, similarly as in Section 2.3, is allowed to depend on N, i.e., $R = R_N$. As before, $\mathbf{X} \in \mathbb{R}^p$ denotes the covariate vector while $Y = \mathrm{Y}(R)$ now denotes the observed potential outcome. Here, $\mathrm{Y}(1)$ denotes the potential outcome if the individual have been treated and $\mathrm{Y}(0)$ denotes the potential outcome if the individual haven't been treated. For each individual, only one of the potential outcomes $\mathrm{Y}(R)$ is observable. Consistency of the potential outcomes is assumed throughout: $Y = \mathrm{Y}(R) = R\mathrm{Y}(1) + (1 - R)\mathrm{Y}(0)$; see [Rub74] and [IR15b].

Now we define the parameter of interest, $\theta_{\mathrm{ATE}} := \theta^1 - \theta^0$ to be the ATE of R on Y, where with a slight abuse of notation we denote with $\theta^1 := E\{\mathrm{Y}(1)\}$ and $\theta^0 := E\{\mathrm{Y}(0)\}$. Moving forward we assume the usual *unconfoundedness condition* [Imb04, Tsi07]:

$$\{\mathrm{Y}(0), \mathrm{Y}(1)\} \perp\!\!\!\perp R \mid \mathbf{X}.$$

Then, $\theta^1 = E\{m_1(\mathbf{X})\}$ and $\theta^0 = E\{m_0(\mathbf{X})\}$, where $m_r(\mathbf{X}) := E(Y|R = r, \mathbf{X}) \equiv E\{Y(r)|\mathbf{X}\}$ denotes the conditional outcome model, and $r \in \{0, 1\}$. With extremely imbalanced groups, without loss of generality, we assume $\pi_N = P(R = 1) \to 0$, i.e., most of the individuals are likely to be in the control group.

The estimation of θ^1 is the same as the mean estimation problem in the MAR-SS

setting, if we set $\mathbf{Z}_i$s to be $(R_i, R_iY_i, \mathbf{X}_i)$. Similarly, θ^0 can be identified as a mean with $\mathbf{Z}_i$s being $(1-R_i, (1-R_i)Y_i, \mathbf{X}_i)$. Now, as in (2.5), θ^1 and θ^0 can be estimated by:

$$\begin{aligned}\widehat{\theta^1} &:= N^{-1}\sum_{i=1}^{N}\left[\widehat{m}_1(\mathbf{X}_i) + \frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i)}\{Y_i - \widehat{m}_1(\mathbf{X}_i)\}\right], \\ \widehat{\theta^0} &:= N^{-1}\sum_{i=1}^{N}\left[\widehat{m}_0(\mathbf{X}_i) + \frac{1-R_i}{1-\widehat{\pi}_N(\mathbf{X}_i)}\{Y_i - \widehat{m}_0(\mathbf{X}_i)\}\right], \end{aligned} \tag{2.40}$$

where, for each $k \leq \mathbb{K}$ and $i \in \mathbb{S}_k$, $\widehat{m}_1(\mathbf{X}_i) = \widehat{m}_1(\mathbf{X}_i; \mathbb{S}_{-k})$, $\widehat{m}_0(\mathbf{X}_i) = \widehat{m}_0(\mathbf{X}_i; \mathbb{S}_{-k})$, and $\widehat{\pi}_N(\mathbf{X}_i) = \widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})$ are cross-fitted estimators of $m_1(\mathbf{X}_i)$, $m_0(\mathbf{X}_i)$, and $\pi_N(\mathbf{X}_i)$, respectively. Here, $\mathbb{S}_{-k} = \{\mathbf{Z}_i : i \in \mathcal{I}\backslash\mathcal{I}_k\}$ is defined analogously as discussed below (2.2) in Section 2.3.1, and for $r \in \{0,1\}$, $\widehat{m}_r(\cdot)$ is constructed based on $\{\mathbf{Z}_i : i \in \mathbb{S}_{-k}, R_i = r\}$. Hence, θ_{ATE} can be estimated by the *DRSS ATE estimator*:

$$\widehat{\theta}_{\text{ATE}} := \widehat{\theta^1} - \widehat{\theta^0}. \tag{2.41}$$

The asymptotic properties of $\widehat{\theta^1}$ follow directly from Theorem 2.2. The following theorem provides the asymptotic results for $\widehat{\theta^0}$. For the sake of a better interpretability, in the following theorem, we suppose $c\pi_N < \pi_N(\mathbf{X})$ with some constant $c > 0$ and hence we have $a_N \asymp \pi_N$.

Theorem 2.8. *Assume $N \to \infty$, $\pi_N \to 0$ and $N\pi_N \to \infty$. Suppose for all $\mathbf{x} \in \mathcal{X}$, $c\pi_N < \pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$, for some $c, C > 0$. Let $\varepsilon := Y - Rm_1(\mathbf{X}) - (1-R)m_0(\mathbf{X})$, assume $\|\varepsilon\|_{2,P} < \infty$, $\|m_0(\cdot) - \mu_0(\cdot)\|_{2,P_{\mathbf{X}}} < C < \infty$, $Var\{m_0(\mathbf{X})\} < \infty$ as well as*

$$\|\widehat{m}_0(\cdot) - \mu_0(\cdot)\|_{2,P_X} = O_p(r_{\mu,0,N}), \quad \sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\pi}_N(\mathbf{x}) - e_N(\mathbf{x})}{\pi_N}\right| = O_p(r_{e,N}), \tag{2.42}$$

for a sequence of positive numbers $r_{\mu,0,N} = o(1)$ *and* $r_{e,N} = o(1)$*. Then,*

$$\begin{aligned}\widehat{\theta}^0 - \theta^0 &= N^{-1}\sum_{i=1}^{N}\psi_0(\mathbf{Z}_i) + O_p(N^{-1/2}\pi_N^{1/2}r_{\mu,0,N} + N^{-1/2}\pi_N r_{e,N} + \pi_N r_{e,N} r_{\mu,0,N}) \\ &\quad + \mathbb{1}\{m_0(\cdot) \neq \mu_0(\cdot)\}O_p(\pi_N r_{e,N}) + \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\}O_p(\pi_N r_{\mu,0,N}),\end{aligned}$$

where

$$\begin{aligned}\psi_0(\mathbf{Z}) &:= \mu_0(\mathbf{X}) - \theta^0 + \frac{1-R}{1-e_N(\mathbf{X})}\{Y - \mu_0(\mathbf{X})\} \\ &= \frac{e_N(\mathbf{X}) - R}{1-e_N(\mathbf{X})}\{m_0(\mathbf{X}) - \mu_0(\mathbf{X})\} + m_0(\mathbf{X}) - \theta^0 + \frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})},\end{aligned} \tag{2.43}$$

with $E\{\psi_0(\mathbf{Z})\} = \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot),\ \mu(\cdot) \neq m(\cdot)\}O_p(\pi_N)$ *and* $Var\{\psi_0(\mathbf{Z})\} = O_p(N^{-1})$*. Hence,*

$$\widehat{\theta}^0 - \theta^0 = N^{-1}\sum_{i=1}^{N}\psi_0(\mathbf{Z}_i) + o_p(N^{-1/2}), \quad \textit{with} \quad E\{\psi_0(\mathbf{Z})\} = 0,$$

once $\pi_N r_{e,N} r_{\mu,0,N} = o(N^{-1/2})$*,* $\pi_N r_{e,N} = o(N^{-1/2})$ *if* $m(\cdot)$ *is misspecified,* $\pi_N r_{\mu,0,N} = o(N^{-1/2})$ *if* $\pi_N(\cdot)$ *is misspecified and at least one of* $m(\cdot)$ *and* $\pi_N(\cdot)$ *is correctly specified. If both* $m(\cdot)$ *and* $\pi_N(\cdot)$ *are misspecified, then we have* $\widehat{\theta}_0 - \theta_0 = O_p(\pi_N + N^{-1/2})$*.*

Remark 2.21 (Comparison with the naive estimator)**.** *Now we consider the comparison of the doubly robust estimator* $\widehat{\theta}^0$ *with the empirical average of the response over the control group* $\bar{Y}_0 := \sum_{i=1}^{N}(1-R_i)Y_i / \sum_{i=1}^{N}(1-R_i)$*. The empirical average* $\bar{Y}_0$ *can be seen as a special case of the estimator* (2.40) *(without cross-fitting) in that* $\widehat{\pi}_N(\mathbf{X}) = N^{-1}\sum_{i=1}^{N}(1-R_i)$ *and* $\widehat{m}_0(\mathbf{X}) = 0$*. Notice that when* $\pi_N \to 0$*,*

$$\begin{aligned}\bar{Y}_0 &= E\{m_0(\mathbf{X})|R=0\} + O_p(N^{-1/2}), \quad \textit{with} \\ \theta_0 - E\{m_0(\mathbf{X})|R=0\} &= [E\{m_0(\mathbf{X})|R=1\} - E\{m_0(\mathbf{X})|R=0\}]\pi_N = O(\pi_N),\end{aligned}$$

and hence $\bar{Y}_0 - \theta_0 = O_p(\pi_N + N^{-1/2})$*, which coincides with the case that both* $m(\cdot)$ *and* $\pi_N(\cdot)$ *are misspecified in Theorem 2.8.*

Corollary 2.1. *Let the assumptions of Theorem 2.8 hold. Assume that at least one of* $e_N(\cdot) = \pi_N(\cdot)$ *and* $\mu_1(\cdot) = m_1(\cdot)$ *holds, and let*

$$\|\widehat{m}_1(\mathbf{X}; \mathbb{S}_{-k}) - \mu_1(\cdot)\|_{2,P_{\mathbf{X}}} \;=\; O_p(r_{\mu,1,N}), \quad \textit{with } r_{\mu,1,N} \;=\; o(1).$$

Then,

$$\begin{aligned}\widehat{\theta}_{\mathrm{ATE}} - \theta_{\mathrm{ATE}} \;=\; & N^{-1}\sum_{i=1}^{N}\psi_1(\mathbf{Z}_i) + \widehat{\Delta}_N + O_p(r_{e,N}r_{\mu,1,N} + \pi_N r_{e,N}r_{\mu,0,N}) \\ & + \mathbb{1}\{m_0(\cdot) \neq \mu_0(\cdot)\}O_p(\pi_N r_{e,N}) + \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\}O_p(\pi_N r_{\mu,0,N}) \\ & + \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot),\ \mu(\cdot) \neq m(\cdot)\}O_p(\pi_N) + o_p\left((N\pi_N)^{-1/2}\right),\end{aligned}$$

where

$$\psi_1(\mathbf{Z}) \;:=\; \mu_1(\mathbf{X}) - \theta_1 + \frac{R}{e_N(\mathbf{X})}\{Y - \mu_1(\mathbf{X})\},$$

with $E\{\psi_1(\mathbf{Z})\} = 0$*,* $E\{\psi_1^2(\mathbf{Z})\} \asymp \pi_N^{-1}$*, and*

$$\widehat{\Delta}_N := \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi_N(\mathbf{X}_i)} - \frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i)}\right\}\{\mu_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)\} = O_p(r_{e,N}) \;\; \textit{if } \; e_N(\cdot) = \pi_N(\cdot),$$

$$\widehat{\Delta}_N := \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi_N(\mathbf{X}_i)} - \frac{R_i}{e_N(\mathbf{X}_i)}\right\}\{\widehat{m}_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)\} = O_p(r_{\mu,1,N}) \;\; \textit{if } \; \mu_1(\cdot) = m_1(\cdot).$$

Moreover, if $r_{e,N}r_{\mu,1,N} = o_p((N\pi_N)^{-1/2})$*,*

$$\widehat{\theta}_{\mathrm{ATE}} - \theta_{\mathrm{ATE}} \;=\; N^{-1}\sum_{i=1}^{N}\psi_1(\mathbf{Z}_i) + o_p\left((N\pi_N)^{-1/2}\right) + \widehat{\Delta}_N,$$

when one of the following holds: (a) both $m(\cdot)$ *and* $\pi_N(\cdot)$ *are correctly specified,* $r_{e,N}r_{\mu,0,N} = o(N^{-1/2}\pi_N^{-3/2})$*; (b)* $\pi_N(\cdot)$ *is correctly specified,* $m(\cdot)$ *is misspecified,* $r_{e,N} = o(N^{-1/2}\pi_N^{-3/2})$*; (c)* $m(\cdot)$ *is correctly specified,* $\pi_N(\cdot)$ *is misspecified,* $r_{\mu,0,N} = o(N^{-1/2}\pi_N^{-3/2})$*; (d) both* $m(\cdot)$ *and* $\pi_N(\cdot)$ *are misspecified,* $N\pi_N^3 = o(1)$*.*

2.6 Simulation studies

We illustrate the performance of our DRSS estimators through extensive simulations under various data generating processes (DGPs). We first provide our main simulation results in Section 2.6.1, where the double robustness (in the sense of consistency or inference) shows up in different misspecification settings. Then, in Section 2.6.2, we show the simulation results under a special stratified labeling PS model that was discussed in Section 2.4.4. We further focus on sparse linear models in high dimensions, and provide results under different sparsity levels in Section 2.6.3. In Section 2.6.4, we further consider the adjusted confidence interval constructed based on the RAL expansion in Remark 2.9.

2.6.1 Main simulation results

We consider the following choices of parameters p, N and π_N:

$$p \in \{10, 500\}, \quad (N, \pi_N) \in \{(10000, 0.01),\ (50000, 0.01),\ (10000, 0.1)\}.$$

We generate i.i.d. Gaussian covariates $\mathbf{X}_i \sim^{\text{iid}} N_p(\mathbf{0}, \boldsymbol{I}_p)$ and residuals $\varepsilon_i \sim^{\text{iid}} N(0,1)$. Given $\mathbf{X}_i$, we generate $R_i|\mathbf{X}_i \sim \text{Bernoulli}(\pi_N(\mathbf{X}_i))$, with the following PS models:

P1. (Constant PS) $\pi_N(\cdot) \equiv \pi_N$.

P2. (Offset logistic PS) $\pi_N(\mathbf{x}) = g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0 + \log(\pi_N))$, where $g(\cdot)$ is defined in (2.17).

We consider the following outcome models for Y_i given $\mathbf{X}_i$:

O1. (Linear outcome) $Y_i = \overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0 + \varepsilon_i$.

O2. (Quadratic outcome) $Y_i = \overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0 + \sum_{j=1}^{p+1}\boldsymbol{\alpha}_0(j)\overrightarrow{\mathbf{X}}_i(j)^2 + \varepsilon_i$.

The parameter values are chosen as:

$$\boldsymbol{\beta}_0 \;=\; (-0.5, 1, 1, 1, \mathbf{0}_{1\times(p-3)})^T, \quad \boldsymbol{\gamma}_0 \;=\; (\boldsymbol{\gamma}_0(1), 1, \mathbf{0}_{1\times(p-1)})^T, \quad \boldsymbol{\alpha}_0 \;=\; (0, 1, 1, 1, \mathbf{0}_{1\times(p-3)})^T,$$

where $\boldsymbol{\gamma}_0(1)$ is chosen so that $E(R) = \pi_N$ for each π_N. The following DGPs are considered: Setting a: P1+O1, Setting b: P1+O2, Setting c: P2+O1, and Setting d: P2+O2. For each DGP, we compare the performance of the following estimators: (1) A naive mean estimator over the labeled samples $\bar{Y}_{\text{labeled}} := \sum_{i=1}^N R_i Y_i / \sum_{i=1}^N R_i$; (2) An oracle case of the mean estimator $\widehat{\theta}_{\text{DRSS}}$ in (2.5) with $\pi_N(\cdot)$ and $m(\cdot)$ treated as known; (3) The proposed DRSS mean estimator $\widehat{\theta}_{\text{DRSS}}$ in (2.5), with $\mathbb{K} = 5$.

We consider several different choices on the outcome and propensity estimators. In low dimensions ($p = 10$), we consider two parametric outcome model estimators: least squares (LS) linear regression and a polynomial (poly) regression with degree 2 (without interaction terms), and two non-parametric outcome estimators: random forest (RF) and reproducing kernel Hilbert space (RKHS) regression using a Gaussian kernel. In high dimensions ($p = 500$), we consider two ℓ_1-regularized parametric outcome model estimators: Lasso and a degree-2 polynomial regression with a Lasso-type penalty (poly-Lasso). As for the PS estimators, we consider a constant estimator that essentially corresponds to a MCAR estimator, and an offset based logistic estimator (or it's ℓ_1-regularized version, log-Lasso, when $p = 500$). The tuning parameters in the regularized estimators are chosen via 5-fold cross-validation. The hyperparameters in the RF are chosen by minimizing the out-of-bag (OOB) error. The bandwidth parameter for the Gaussian kernel in RKHS regression is set to be p.

The simulations are repeated for 500 times and the results are presented in Tables 2.1-

Table 2.1: Simulation setting a with $p = 10$. Bias: empirical bias; RMSE: root mean square error; Length: average length of the 95% confidence intervals; Coverage: average coverage of the 95% confidence intervals; ESD: empirical standard deviation; ASD: average of estimated standard deviations. The **blue** color in the tables denotes the "smallest" and correctly specified parametric model for each of the settings.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
$\bar{Y}_{\text{labeled}}$		0.013	0.204	0.789	0.932	0.203	0.201
oracle		0.003	0.106	0.397	0.942	0.106	0.101
constant	**LS**	0.003	0.115	0.436	0.938	0.115	0.111
	poly	0.002	0.127	0.482	0.934	0.127	0.123
	RF	0.008	0.155	0.604	0.926	0.155	0.154
	RKHS	0.009	0.145	0.547	0.936	0.145	0.139
logistic	LS	0.005	0.127	0.534	0.972	0.127	0.136
	poly	0.003	0.144	0.600	0.972	0.144	0.153
	RF	0.003	0.152	0.762	0.990	0.152	0.194
	RKHS	0.007	0.161	0.698	0.976	0.161	0.178
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
$\bar{Y}_{\text{labeled}}$		0.008	0.092	0.352	0.950	0.092	0.090
oracle		0.002	0.045	0.179	0.948	0.045	0.046
constant	**LS**	0.003	0.045	0.182	0.956	0.045	0.046
	poly	0.003	0.046	0.185	0.952	0.046	0.047
	RF	0.004	0.056	0.218	0.942	0.056	0.056
	RKHS	0.003	0.056	0.213	0.942	0.056	0.054
logistic	LS	0.003	0.046	0.189	0.960	0.046	0.048
	poly	0.003	0.047	0.192	0.962	0.046	0.049
	RF	0.002	0.052	0.227	0.974	0.052	0.058
	RKHS	0.001	0.054	0.223	0.960	0.054	0.057

Table 2.2: Simulation setting b with $p = 10$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ ($N\pi_N = 100$)					
$\bar{Y}_{\text{labeled}}$		0.008	0.334	1.251	0.936	0.334	0.319
oracle		0.001	0.103	0.410	0.946	0.103	0.105
constant	LS	0.009	0.311	1.167	0.938	0.311	0.298
	poly	0.002	0.124	0.496	0.950	0.125	0.127
	RF	0.002	0.253	0.939	0.934	0.253	0.239
	RKHS	0.004	0.245	0.930	0.928	0.245	0.237
logistic	LS	0.154	0.420	1.540	0.950	0.391	0.393
	poly	0.001	0.143	0.615	0.968	0.144	0.157
	RF	0.075	0.319	1.222	0.960	0.310	0.312
	RKHS	0.107	0.327	1.216	0.950	0.310	0.310
		$N = 50000, \pi_N = 0.01$ ($N\pi_N = 500$)					
$\bar{Y}_{\text{labeled}}$		0.005	0.138	0.558	0.958	0.138	0.142
oracle		0.001	0.046	0.184	0.952	0.046	0.047
constant	LS	0.008	0.119	0.478	0.952	0.119	0.122
	poly	0.000	0.047	0.189	0.952	0.047	0.048
	RF	0.001	0.076	0.304	0.942	0.076	0.077
	RKHS	0.001	0.076	0.306	0.952	0.076	0.078
logistic	LS	0.032	0.128	0.505	0.954	0.124	0.129
	poly	0.000	0.048	0.196	0.956	0.048	0.050
	RF	0.009	0.079	0.320	0.958	0.079	0.082
	RKHS	0.014	0.080	0.324	0.960	0.079	0.083
		$N = 10000, \pi_N = 0.1$ ($N\pi_N = 1000$)					
$\bar{Y}_{\text{labeled}}$		0.001	0.107	0.411	0.944	0.107	0.105
oracle		0.000	0.045	0.176	0.946	0.045	0.045
constant	LS	0.000	0.091	0.353	0.958	0.091	0.090
	poly	0.000	0.045	0.177	0.944	0.046	0.045
	RF	-0.001	0.057	0.228	0.952	0.057	0.058
	RKHS	-0.001	0.058	0.232	0.952	0.058	0.059
logistic	LS	0.012	0.093	0.363	0.958	0.092	0.093
	poly	0.000	0.046	0.179	0.944	0.046	0.046
	RF	0.003	0.058	0.233	0.956	0.058	0.059
	RKHS	-0.005	0.058	0.237	0.956	0.058	0.061

Table 2.3: Simulation setting c with $p = 10$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
$\bar{Y}_{\text{labeled}}$		0.980	0.999	0.783	0.002	0.194	0.200
oracle		0.003	0.106	0.397	0.942	0.106	0.101
constant	LS	0.005	0.151	0.434	0.850	0.151	0.111
	poly	-0.004	0.194	0.480	0.792	0.194	0.122
	RF	0.570	0.610	0.596	0.108	0.218	0.152
	RKHS	0.403	0.439	0.543	0.210	0.175	0.139
logistic	**LS**	0.015	0.234	0.884	0.952	0.234	0.226
	poly	-0.002	0.365	1.083	0.922	0.365	0.276
	RF	-0.171	0.705	1.708	0.950	0.685	0.436
	RKHS	-0.140	0.675	1.525	0.938	0.661	0.389
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
$\bar{Y}_{\text{labeled}}$		0.968	0.972	0.350	0.000	0.090	0.089
oracle		0.000	0.070	0.281	0.968	0.070	0.072
constant	LS	0.001	0.063	0.181	0.868	0.063	0.046
	poly	0.001	0.070	0.183	0.812	0.070	0.047
	RF	0.341	0.351	0.215	0.004	0.085	0.055
	RKHS	0.240	0.251	0.211	0.028	0.072	0.054
logistic	**LS**	0.000	0.072	0.297	0.964	0.073	0.076
	poly	0.000	0.076	0.307	0.956	0.076	0.078
	RF	-0.016	0.121	0.491	0.956	0.120	0.125
	RKHS	-0.015	0.123	0.464	0.938	0.122	0.118
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
$\bar{Y}_{\text{labeled}}$		0.820	0.822	0.244	0.000	0.063	0.062
oracle		0.001	0.052	0.200	0.946	0.052	0.051
constant	LS	0.001	0.046	0.142	0.868	0.046	0.036
	poly	0.001	0.051	0.143	0.864	0.051	0.036
	RF	0.241	0.248	0.156	0.006	0.058	0.040
	RKHS	0.149	0.158	0.154	0.102	0.053	0.039
logistic	**LS**	0.001	0.052	0.203	0.936	0.052	0.052
	poly	0.001	0.053	0.206	0.930	0.053	0.053
	RF	-0.009	0.070	0.294	0.970	0.069	0.075
	RKHS	-0.005	0.069	0.277	0.942	0.069	0.071

Table 2.4: Simulation setting d with $p = 10$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ ($N\pi_N = 100$)					
$\bar{Y}_{\text{labeled}}$		1.885	1.934	1.617	0.002	0.432	0.412
oracle		0.008	0.168	0.625	0.952	0.168	0.160
constant	LS	-0.929	1.040	1.148	0.226	0.469	0.293
	poly	0.002	0.189	0.492	0.808	0.189	0.125
	RF	0.317	0.443	1.033	0.752	0.309	0.264
	RKHS	0.392	0.472	1.022	0.664	0.263	0.261
logistic	LS	0.378	1.386	3.996	0.910	1.335	1.019
	poly	0.012	0.255	1.011	0.932	0.255	0.258
	RF	0.026	0.573	1.828	0.954	0.573	0.466
	RKHS	-0.014	0.570	1.727	0.950	0.570	0.441
		$N = 50000, \pi_N = 0.01$ ($N\pi_N = 500$)					
$\bar{Y}_{\text{labeled}}$		1.901	1.910	0.728	0.000	0.192	0.186
oracle		-0.002	0.074	0.287	0.944	0.074	0.073
constant	LS	-0.951	0.972	0.473	0.000	0.202	0.121
	poly	-0.002	0.076	0.189	0.798	0.076	0.048
	RF	0.074	0.123	0.323	0.816	0.099	0.082
	RKHS	0.163	0.191	0.328	0.500	0.100	0.084
logistic	LS	0.078	0.485	1.661	0.924	0.479	0.424
	poly	-0.002	0.082	0.318	0.940	0.082	0.081
	RF	0.002	0.181	0.594	0.924	0.181	0.152
	RKHS	0.003	0.178	0.557	0.936	0.178	0.142
		$N = 10000, \pi_N = 0.1$ ($N\pi_N = 1000$)					
$\bar{Y}_{\text{labeled}}$		1.371	1.377	0.471	0.000	0.120	0.120
oracle		-0.001	0.053	0.225	0.980	0.053	0.057
constant	LS	-0.746	0.756	0.342	0.000	0.122	0.087
	poly	-0.001	0.052	0.172	0.914	0.052	0.044
	RF	0.010	0.065	0.227	0.924	0.064	0.058
	RKHS	0.053	0.084	0.231	0.838	0.065	0.059
logistic	LS	0.019	0.234	0.951	0.944	0.233	0.243
	poly	-0.001	0.054	0.226	0.976	0.054	0.058
	RF	-0.005	0.100	0.379	0.940	0.099	0.097
	RKHS	0.000	0.094	0.358	0.940	0.094	0.091

Table 2.5: Simulation setting a with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
	$\bar{Y}_{\text{labeled}}$	-0.003	0.195	0.788	0.960	0.195	0.201
	oracle	-0.001	0.103	0.400	0.954	0.103	0.102
constant	**Lasso**	0.000	0.119	0.478	0.950	0.119	0.122
	poly-Lasso	-0.002	0.120	0.493	0.950	0.120	0.126
log-Lasso	Lasso	0.000	0.120	0.487	0.960	0.120	0.124
	poly-Lasso	-0.002	0.121	0.502	0.952	0.121	0.128
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	0.003	0.093	0.352	0.944	0.093	0.090
	oracle	0.001	0.044	0.178	0.948	0.044	0.046
constant	**Lasso**	0.001	0.046	0.184	0.952	0.046	0.047
	poly-Lasso	0.001	0.046	0.185	0.952	0.046	0.047
log-Lasso	Lasso	0.001	0.046	0.185	0.948	0.046	0.047
	poly-Lasso	0.001	0.046	0.186	0.952	0.046	0.047
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
	$\bar{Y}_{\text{labeled}}$	-0.003	0.063	0.260	0.958	0.063	0.066
	oracle	-0.002	0.037	0.147	0.956	0.037	0.037
constant	**Lasso**	-0.002	0.038	0.149	0.960	0.038	0.038
	poly-Lasso	-0.002	0.038	0.149	0.960	0.038	0.038
log-Lasso	Lasso	-0.002	0.038	0.149	0.956	0.038	0.038
	poly-Lasso	-0.002	0.038	0.149	0.960	0.038	0.038

2.8. We report the bias, the root mean square error (RMSE), the average length and average coverage of the 95% confidence intervals, the empirical standard error and the averaged estimated standard error for all settings. The **blue** color in the tables denotes the "smallest" (i.e., most parsimonious) and correctly specified parametric model for each of the settings.

We first check the proposed estimator's double robustness in terms of inference. As per Theorem 2.2, the asymptotic normality results hold when the product rate condition is

Table 2.6: Simulation setting b with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
	$\bar{Y}_{\text{labeled}}$	0.003	0.306	1.244	0.948	0.306	0.317
	oracle	-0.003	0.107	0.411	0.934	0.107	0.105
constant	Lasso	0.004	0.296	1.220	0.966	0.296	0.311
	poly-Lasso	0.000	0.174	0.668	0.954	0.175	0.171
log-Lasso	Lasso	0.008	0.297	1.241	0.966	0.298	0.317
	poly-Lasso	0.002	0.175	0.680	0.954	0.175	0.173
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	-0.004	0.147	0.554	0.926	0.147	0.141
	oracle	-0.002	0.054	0.215	0.954	0.054	0.055
constant	Lasso	-0.005	0.129	0.483	0.928	0.129	0.123
	poly-Lasso	-0.002	0.051	0.195	0.934	0.051	0.050
log-Lasso	Lasso	-0.004	0.129	0.484	0.926	0.129	0.124
	poly-Lasso	-0.002	0.051	0.196	0.936	0.051	0.050
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
	$\bar{Y}_{\text{labeled}}$	-0.001	0.104	0.411	0.948	0.104	0.105
	oracle	0.001	0.042	0.175	0.964	0.043	0.045
constant	Lasso	-0.001	0.091	0.357	0.944	0.092	0.091
	poly-Lasso	0.001	0.043	0.178	0.960	0.043	0.046
log-Lasso	Lasso	-0.001	0.091	0.357	0.948	0.091	0.091
	poly-Lasso	0.001	0.043	0.179	0.962	0.043	0.046

satisfied and when both of $\pi_N(\cdot)$ and $m(\cdot)$ are correct, in which case, the proposed estimator is $(N\pi_N)^{1/2}$-consistent with the asymptotic efficiency matching that of the oracle estimator. In low dimensions ($p = 10$), the product rate condition always holds since $\widehat{\pi}_N(\cdot)$ has an estimation error of rate $(N\pi_N)^{-1/2}$ and $\widehat{m}(\cdot)$ (parametric or non-parametric) is consistent. As in Tables 2.1-2.4, the coverage of $\widehat{\theta}_{\text{DRSS}}$ based on correct $\pi_N(\cdot)$ and $m(\cdot)$ is close to 95% even with a small $N\pi_N = 100$. In high dimensions ($p = 500$), the product rate condition

Table 2.7: Simulation setting c with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
	$\bar{Y}_{\text{labeled}}$	0.977	0.997	0.782	0.006	0.198	0.199
	oracle	-0.003	0.160	0.612	0.970	0.160	0.156
constant	Lasso	0.275	0.320	0.480	0.412	0.164	0.123
	poly-Lasso	0.324	0.368	0.496	0.320	0.173	0.127
log-Lasso	**Lasso**	0.099	0.210	0.563	0.782	0.186	0.144
	poly-Lasso	0.117	0.229	0.602	0.778	0.197	0.154
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	0.975	0.979	0.350	0.000	0.087	0.089
	oracle	-0.003	0.071	0.282	0.956	0.071	0.072
constant	Lasso	0.115	0.132	0.184	0.370	0.065	0.047
	poly-Lasso	0.130	0.146	0.185	0.306	0.067	0.047
log-Lasso	**Lasso**	0.022	0.072	0.241	0.884	0.069	0.061
	poly-Lasso	0.026	0.075	0.243	0.886	0.070	0.062
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
	$\bar{Y}_{\text{labeled}}$	0.822	0.824	0.245	0.000	0.065	0.062
	oracle	0.005	0.051	0.198	0.958	0.050	0.050
constant	Lasso	0.077	0.090	0.143	0.438	0.047	0.036
	poly-Lasso	0.086	0.098	0.143	0.386	0.047	0.036
log-Lasso	**Lasso**	0.019	0.053	0.173	0.894	0.049	0.044
	poly-Lasso	0.021	0.054	0.174	0.896	0.050	0.044

depends on the true PS model. When the true PS model is P1 (MCAR), the corresponding $\widehat{\pi}_N(\cdot)$ still has an estimation error of rate $(N\pi_N)^{-1/2}$ and hence the product rate condition holds; see Tables 2.5 and 2.6. When the true PS model is P2 (offset logistic), the product rate condition requires $s_m s_\pi = o(N\pi_N\{\log(p)\}^{-2})$, where $s_m := \|\boldsymbol{\beta}_0\|_0$ and $s_\pi := \|\boldsymbol{\gamma}_0\|_0$. We can see the coverages are slowly growing towards 95% as $N\pi_N$ increases in Tables 2.7 and 2.8. More results with different sparsity levels in the high dimensions can be found in Section

Table 2.8: Simulation setting d with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
	$\bar{Y}_{\text{labeled}}$	1.875	1.917	1.620	0.000	0.401	0.413
	oracle	-0.002	0.157	0.613	0.970	0.157	0.156
constant	Lasso	-0.183	0.535	1.225	0.756	0.503	0.313
	poly-Lasso	0.273	0.360	0.625	0.566	0.235	0.160
log-Lasso	Lasso	-0.229	0.621	1.511	0.736	0.578	0.386
	poly-Lasso	0.090	0.263	0.708	0.824	0.247	0.181
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	1.890	1.900	0.731	0.000	0.191	0.187
	oracle	-0.006	0.074	0.286	0.954	0.074	0.073
constant	Lasso	-0.657	0.686	0.481	0.020	0.200	0.123
	poly-Lasso	0.086	0.113	0.194	0.562	0.074	0.049
log-Lasso	Lasso	-0.254	0.377	0.968	0.684	0.279	0.247
	poly-Lasso	0.010	0.076	0.247	0.888	0.075	0.063
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
	$\bar{Y}_{\text{labeled}}$	1.355	1.360	0.470	0.000	0.118	0.120
	oracle	0.000	0.058	0.224	0.964	0.058	0.057
constant	Lasso	-0.551	0.566	0.343	0.002	0.128	0.088
	poly-Lasso	0.055	0.080	0.175	0.706	0.057	0.045
log-Lasso	Lasso	-0.179	0.250	0.649	0.706	0.175	0.165
	poly-Lasso	0.008	0.058	0.198	0.910	0.058	0.051

2.6.3. Here, in Tables 2.2, 2.4, and 2.6, we can see fairly good coverages *even if* the outcome model is misspecified and the confidence interval is constructed without a modification. This coincides with the Remarks 2.16 and 2.20; see more details in Section 2.6.4.

Regarding efficiency, as in Tables 2.1-2.8, we observe that the proposed estimators based on correct parametric models provide "optimal" RMSEs that are close to the oracle estimator. In Tables 2.1-2.4, the RMSEs based on non-parametric (RF and RKHS) $\widehat{m}(\cdot)$ are

worse than those based on (correctly specified) parametric models with the difference arising from the product rate of the estimation errors of $\widehat{\pi}_N(\cdot)$ and $\widehat{m}(\cdot)$. For a (correctly specified) $\pi_N(\cdot)$ with an estimation error $O_p((N\pi_N)^{-1/2})$, such a difference is not significant and the RMSE is first order insensitive to estimation error of $\widehat{m}(\cdot)$.

We also check the double robustness in terms of consistency of the proposed estimators, when only one of $\pi_N(\cdot)$ and $m(\cdot)$ is correct. As seen in Tables 2.3-2.6, the naive mean estimator, $\bar{Y}_{\text{labeled}}$, is *not* consistent when the selection bias occurs, i.e., the PS is not a constant. Nevertheless, as suggested by Theorem 2.2, the proposed DRSS estimator is still consistent, and it's consistency rate depends on the estimation error rate of the correct one among $\widehat{\pi}_N(\cdot)$ and $\widehat{m}(\cdot)$. The proposed $\widehat{\theta}_{\text{DRSS}}$ can still be $(N\pi_N)^{1/2}$-consistent when the correct estimator has an estimation error of rate $(N\pi_N)^{-1/2}$, which is typically true when the correct model is a low dimensional parametric model: see Tables 2.2, 2.4, and 2.6 for correct $\pi_N(\cdot)$; see Tables 2.3 and 2.4 for correct $m(\cdot)$. If the correct estimator is linear (offset logistic) in high dimensions, $\widehat{\theta}_{\text{DRSS}} - \theta_0$ is of the order $O_p(\{(N\pi_N)^{-1} s \log(p)\}^{1/2})$, where s is the sparsity of the correct model; see results in Tables 2.7 and 2.8. If the correct estimator is non-parametric, we would expect a non-parametric rate on the proposed estimator, matching results in Tables 2.3 and 2.4.

2.6.2 Results under the stratified labeling model

Now we work on a special PS model, that of stratified labeling, as discussed in Section 2.4.4:

P3. (Stratified PS) Suppose we further observe the stratum indicators $\delta_i \in \{0, 1\}$, with the

following model: $\delta_i|\mathbf{X}_i \sim \text{Bernoulli}(p_\delta(\mathbf{X}_i))$ and $R_i|\delta_i \sim \text{Bernoulli}(0.5\pi_N\delta_i + 1.5\pi_N(1-\delta_i))$, where $p_\delta(\mathbf{x}) = g(\mathbf{x}(1))$ and $g(u) = \exp(u)/\{1+\exp(u)\}$.

We consider the same choices of p, N, and π_N as in Section 2.6.1, and we focus on the following DGP of Setting e, i.e., P3+O2. Furthermore, we consider an additional PS estimator based on the stratified labeling of Section 2.4.4, where $p_\delta(\cdot)$ is estimated by a logistic regression (with a Lasso-type regularization when $p = 500$). The results are shown in Tables 2.9 and 2.10 for the case $p = 10$ and $p = 500$, respectively.

By Theorem 2.2, the estimators based on a correctly specified $\pi_N(\cdot)$ (stratified) and $m(\cdot)$ (poly/RF/RKHS) provide $(N\pi_N)^{1/2}$-consistent estimations and valid asymptotic confidence intervals. Additionally, when $\pi_N(\cdot)$ is correctly specified (stratified) and $m(\cdot)$ is misspecified (linear), the proposed DRSS mean estimator has a consistency rate $O_p((N\pi_N)^{-1/2} + r_{\delta,N})$, with $r_{\delta,N}$, defined in (2.38), satisfying $r_{\delta,N} = O(N^{-1/2})$ in low dimensions and $r_{\delta,N} = O(\{s_\delta \log(p)\}^{1/2} N^{-1/2})$, in high dimensions, as discussed in Example 2.2. The simulation results in Tables 2.9 and 2.10 support our theoretical arguments: we can see the stratified+poly/RF/RKHS estimators provide coverages close to 95%, and all the estimators based on a stratified $\widehat{\pi}_N(\cdot)$ provide RMSEs of a similar magnitude.

2.6.3 Results for high dimensional sparse models: Investigating performance under varying sparsity levels

Here we focus on the high dimensional case ($p = 500$) with different sparsity levels. We consider the following PS and outcome models:

P2'. (Offset logistic PS with different sparsity levels) Let $\pi_N(\mathbf{x}) = g(\overrightarrow{\mathbf{x}}^T \boldsymbol{\gamma}_{s_\pi} + \log(\pi_N))$ and

Table 2.9: Simulation setting e with $p = 10$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
$\bar{Y}_{\text{labeled}}$		-0.205	0.360	1.203	0.876	0.297	0.307
oracle		-0.002	0.101	0.415	0.964	0.101	0.106
constant	LS	-0.082	0.318	1.149	0.916	0.308	0.293
	poly	-0.009	0.125	0.493	0.954	0.125	0.126
	RF	-0.100	0.259	0.919	0.908	0.239	0.234
	RKHS	-0.087	0.248	0.909	0.918	0.232	0.232
logistic	LS	0.129	0.419	1.639	0.960	0.399	0.418
	poly	-0.011	0.146	0.633	0.976	0.146	0.162
	RF	0.075	0.328	1.346	0.960	0.320	0.343
	RKHS	0.109	0.341	1.337	0.974	0.323	0.341
stratified	LS	0.009	0.324	1.256	0.948	0.324	0.320
	poly	-0.010	0.126	0.510	0.958	0.126	0.130
	RF	0.021	0.268	1.062	0.956	0.268	0.271
	RKHS	0.016	0.262	1.042	0.960	0.262	0.266
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
$\bar{Y}_{\text{labeled}}$		-0.210	0.251	0.532	0.652	0.138	0.136
oracle		-0.001	0.048	0.187	0.970	0.048	0.048
constant	LS	-0.078	0.150	0.469	0.882	0.128	0.120
	poly	-0.001	0.050	0.188	0.942	0.050	0.048
	RF	-0.053	0.101	0.296	0.832	0.086	0.075
	RKHS	-0.053	0.098	0.299	0.866	0.083	0.076
logistic	LS	-0.014	0.135	0.531	0.952	0.135	0.135
	poly	-0.001	0.051	0.199	0.946	0.051	0.051
	RF	-0.011	0.091	0.349	0.944	0.090	0.089
	RKHS	-0.002	0.090	0.353	0.942	0.090	0.090
stratified	LS	-0.001	0.132	0.508	0.950	0.132	0.130
	poly	-0.001	0.050	0.193	0.944	0.050	0.049
	RF	-0.006	0.089	0.333	0.956	0.089	0.085
	RKHS	-0.005	0.088	0.336	0.940	0.088	0.086

Table 2.10: Simulation setting e with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 10000, \pi_N = 0.01$ $(N\pi_N = 100)$					
	$\bar{Y}_{\text{labeled}}$	-0.189	0.363	1.193	0.880	0.311	0.304
	oracle	0.005	0.098	0.417	0.960	0.098	0.106
constant	Lasso	-0.167	0.352	1.186	0.888	0.310	0.303
	poly-Lasso	-0.103	0.220	0.696	0.894	0.195	0.178
log-Lasso	Lasso	-0.160	0.351	1.215	0.904	0.312	0.310
	poly-Lasso	-0.096	0.218	0.717	0.894	0.196	0.183
stratified	Lasso	0.005	0.339	1.318	0.926	0.339	0.336
	poly-Lasso	0.010	0.200	0.780	0.948	0.200	0.199
		$N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	-0.196	0.239	0.535	0.702	0.136	0.136
	oracle	0.000	0.048	0.187	0.950	0.048	0.048
constant	Lasso	-0.134	0.187	0.478	0.776	0.130	0.122
	poly-Lasso	-0.032	0.061	0.195	0.896	0.052	0.050
log-Lasso	Lasso	-0.101	0.163	0.491	0.844	0.128	0.125
	poly-Lasso	-0.022	0.056	0.196	0.914	0.052	0.050
stratified	Lasso	0.001	0.137	0.530	0.946	0.137	0.135
	poly-Lasso	0.000	0.052	0.202	0.944	0.052	0.051
		$N = 10000, \pi_N = 0.1$ $(N\pi_N = 1000)$					
	$\bar{Y}_{\text{labeled}}$	-0.205	0.226	0.377	0.428	0.095	0.096
	oracle	-0.001	0.044	0.173	0.944	0.044	0.044
constant	Lasso	-0.124	0.151	0.335	0.688	0.087	0.086
	poly-Lasso	-0.024	0.051	0.173	0.900	0.045	0.044
log-Lasso	Lasso	-0.082	0.119	0.347	0.854	0.087	0.088
	poly-Lasso	-0.013	0.047	0.174	0.928	0.045	0.044
stratified	Lasso	-0.013	0.090	0.367	0.956	0.089	0.094
	poly-Lasso	-0.002	0.045	0.176	0.948	0.045	0.045

$R_i|\mathbf{X}_i \sim \text{Bernoulli}(\pi_N(\mathbf{X}_i))$, where $g(u) = \exp(u)/\{1 + \exp(u)\}$.

O1'. (Linear outcome with different sparsity levels) Let $Y_i = \overrightarrow{\mathbf{X}}_i^T \boldsymbol{\beta}_{s_m} + \varepsilon_i$.

The parameter values are:

$$\boldsymbol{\beta}_{s_m} = (-0.5, \sqrt{3/s_m}\mathbf{1}_{1\times s_m}, \mathbf{0}_{1\times(p-s_m)})^T \text{ and } \boldsymbol{\gamma}_{s_\pi} = (\boldsymbol{\gamma}_0(1), \sqrt{1/s_\pi}\mathbf{1}_{1\times s_\pi}, \mathbf{0}_{1\times(p-s_\pi)})^T.$$

We consider a DGP, Setting c': P2'+O1', with the following choices of p, N, π_N, s_m and s_π:

$$p = 500, \quad N \in \{50000, 200000\}, \quad \pi_N = 0.01, \quad (s_m, s_\pi) \in \{(3,15), (15,3)\}.$$

We illustrate the performance of the same estimators that we considered in Section 2.6.1 (for $p = 500$); the results are presented in Table 2.11.

In Table 2.11, we observe that the RMSEs of $\widehat{\theta}_{\text{DRSS}}$ based on log-Lasso PS estimators are smaller than those based on constant PS estimators. This coincides with our Remark 2.6, as well as Theorems 2.2 and 2.5 - if both of the nuisance functions are correctly specified, we have $\widehat{\theta}_{\text{DRSS}} - \theta_0 = O_p((N\pi_N)^{-1/2} + \sqrt{s_m s_\pi}\log(p)/(N\pi_N))$; if only the outcome model is correctly specified, we have a slower upper bound $\widehat{\theta}_{\text{DRSS}} - \theta_0 = O_p(\sqrt{s_m \log(p)/(N\pi_N)})$. For the DRSS estimators based on log-Lasso PS estimators, we can see that the biases of the estimators, originating from the product rate $\sqrt{s_m s_\pi}\log(p)/(N\pi_N)$, are non-ignorable compared to the RMSEs, especially for smaller N. As N grows, however, the coverages of the confidence intervals start getting closer to the desired 95% level. This also coincides with our Remark 2.6 that we expect a valid inference result when $s_m s_\pi \log(p) = o(N\pi_N)$ for correctly specified models.

2.6.4 Inference results based on adjusted confidence intervals

In Sections 2.6.1-2.6.3, we illustrated the simulation performance of the proposed confidence interval (2.15), which requires both the outcome and PS models to be correctly

Table 2.11: Simulation setting c' with $p = 500$. The rest of the caption details remain the same as those in Table 2.1.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$s_m = 3, s_\pi = 15, N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	0.755	0.760	0.351	0.000	0.090	0.090
	oracle	0.000	0.074	0.284	0.944	0.074	0.072
constant	Lasso	0.089	0.103	0.184	0.522	0.052	0.047
constant	poly-Lasso	0.098	0.111	0.185	0.458	0.052	0.047
log-Lasso	**Lasso**	0.038	0.071	0.203	0.860	0.060	0.052
log-Lasso	poly-Lasso	0.042	0.073	0.204	0.848	0.060	0.052
		$s_m = 3, s_\pi = 15, N = 200000, \pi_N = 0.01$ $(N\pi_N = 2000)$					
	$\bar{Y}_{\text{labeled}}$	0.754	0.756	0.175	0.000	0.045	0.045
	oracle	0.000	0.035	0.143	0.960	0.035	0.037
constant	Lasso	0.044	0.051	0.090	0.506	0.025	0.023
constant	poly-Lasso	0.049	0.055	0.090	0.430	0.025	0.023
log-Lasso	**Lasso**	0.012	0.032	0.111	0.914	0.030	0.028
log-Lasso	poly-Lasso	0.013	0.033	0.111	0.914	0.030	0.028
		$s_m = 15, s_\pi = 3, N = 50000, \pi_N = 0.01$ $(N\pi_N = 500)$					
	$\bar{Y}_{\text{labeled}}$	0.752	0.757	0.349	0.000	0.085	0.089
	oracle	0.001	0.068	0.283	0.966	0.068	0.072
constant	Lasso	0.155	0.169	0.196	0.200	0.068	0.050
constant	poly-Lasso	0.184	0.197	0.201	0.118	0.070	0.051
log-Lasso	**Lasso**	0.049	0.086	0.237	0.824	0.071	0.060
log-Lasso	poly-Lasso	0.058	0.094	0.245	0.794	0.074	0.063
		$s_m = 15, s_\pi = 3, N = 200000, \pi_N = 0.01$ $(N\pi_N = 2000)$					
	$\bar{Y}_{\text{labeled}}$	0.753	0.754	0.175	0.000	0.044	0.045
	oracle	0.000	0.036	0.144	0.964	0.036	0.037
constant	Lasso	0.076	0.082	0.091	0.172	0.032	0.023
constant	poly-Lasso	0.089	0.094	0.091	0.094	0.032	0.023
log-Lasso	**Lasso**	0.013	0.037	0.124	0.912	0.034	0.032
log-Lasso	poly-Lasso	0.015	0.038	0.125	0.904	0.035	0.032

specified, as in part (a) of Theorem 2.2. In this section, we compare (2.15) with an adjusted version based on the asymptotic expansion in part (b) of Theorem 2.2 and the RAL expansion

in Remark 2.9. The adjusted confidence interval allows inference via $\widehat{\theta}_{\text{DRSS}}$ even under misspecified outcome models, and the adjusted RAL expansions based on different PS models are provided in Theorems 2.4, 2.6, and 2.7. Here, we only focus on the results for the skewed offset logistic PS model as discussed in Theorem 2.4, and we present numerical results to validate the inference provided by the adjusted RAL expansion (2.28) of $\widehat{\theta}_{\text{DRSS}}$ given therein.

Apart from the settings c and d in Section 2.6.1, an additional DGP, Setting f: P2+O3, is considered. Here, P2 is the offset logistic PS model as in Section 2.6.1, and O3 is a cubic outcome model defined as follows:

O3. (Cubic outcome) $Y_i = \overrightarrow{\mathbf{X}}_i^T \boldsymbol{\beta}_0 + \sum_{j=1}^{p+1} \boldsymbol{\alpha}_0(j) \overrightarrow{\mathbf{X}}_i(j)^2 + \sum_{j=1}^{p+1} \boldsymbol{\zeta}_0(j) \overrightarrow{\mathbf{X}}_i(j)^3 + \varepsilon_i$.

The parameter value is defined as:

$$\boldsymbol{\zeta}_0 \;=\; (0, 0.2, 0.2, 0.2, \mathbf{0}_{1\times(p-3)})^T.$$

We illustrate the behavior of the original confidence interval (2.15) and the adjusted confidence interval based on the RAL expansion (2.28). We consider the Settings c, d, and f, where the outcome models are polynomial (without interaction) with degrees 1, 2, and 3, respectively. Apart from the empirical estimator $\bar{Y}_{\text{labeled}}$ and the oracle estimator as in Section 2.6.1, we also consider the proposed mean estimators $\widehat{\theta}_{\text{DRSS}}$ based on an offset logistic model based PS estimator, and polynomial model based outcome regression estimators with degrees 1, 2, and 3. The simulation results are presented in Table 2.12.

We can clearly see the improvement of the coverage based on the adjusted confidence intervals, especially for polynomial estimators $\widehat{m}(\cdot)$ with degrees 2 and 3. As mentioned in Remark 2.16, a latent misspecification arises here since the effective sample size $N\pi_N = 100$ is comparable with the dimension of the working model: for polynomial regression with

degrees 2 and 3, the dimensions of the design matrix are 21 and 31, respectively. Under such a circumstance, $\widehat{m}(\cdot)$ tends to be a biased estimate and a (latent) misspecification arises, in that its (effective) target (or limit) becomes some $\mu(\cdot) \neq m(\cdot)$.

Such an example suggests that, the adjusted confidence intervals, when $\pi_N(\cdot)$ is correctly specified, allow us to better capture the model complexity of $\widehat{m}(\cdot)$ and improve the efficiency of the DRSS estimator. The modified confidence intervals can *still* provide valid inference even when a degree of freedom of the model becomes comparable with the effective sample size.

In Table 2.12, one can see that neither of the averages of the estimated standard deviations (ASDs) or the adjusted ASDs are close to the empirical standard deviations (ESDs) for the DRSS mean estimators based on polynomial regressions with degrees 2 and 3, while we can still achieve fairly acceptable coverages for the confidence intervals. This is not contradicted with our theory: we only obtain asymptotic results in terms of convergence in distribution or probability, whereas ASD = ESD + $o(1)$ requires a convergence in mean (i.e., L_1 convergence). Such a difference is possibly related to the instability of the LS-type outcome estimator, when the dimension of the working model is comparable with the sample size.

2.7 Application to the NHEFS data

We now apply our imbalanced ATE estimator proposed in Section 2.5 to assess the effect of smoking and alcohol drinking on weight gain, using a subset of data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study

Table 2.12: Simulations under Settings c, d and f, with $p = 10, N = 10000$ and $\pi_N = 0.01$ ($N\pi_N = 100$). Bias: empirical bias; RMSE: root mean square error; Length: average length of the 95% confidence intervals; Coverage: average coverage of the 95% confidence intervals; ESD: empirical standard deviation; ASD: average of estimated standard deviations. The results for adjusted (adj) confidence intervals based on the RAL expansion (2.28) in Theorem 2.4 are provided in parentheses.

$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	Bias	RMSE	Length(adj)	Coverage(adj)	ESD	ASD(adj)
		Setting c					
	$\bar{Y}_{\text{labeled}}$	0.979	0.998	0.784	0.002	0.197	0.200
	oracle	-0.007	0.164	0.607	0.972	0.164	0.155
	poly1(LS)	-0.009	0.227	0.865(0.881)	0.952(0.964)	0.227	0.221(0.225)
logistic	poly2	0.002	0.277	1.017(1.039)	0.940(0.964)	0.277	0.260(0.265)
	poly3	-0.013	0.491	1.436(1.465)	0.922(0.956)	0.492	0.366(0.374)
		Setting d					
	$\bar{Y}_{\text{labeled}}$	1.923	1.968	1.628	0.002	0.418	0.415
	oracle	0.011	0.158	0.615	0.960	0.158	0.157
	poly1(LS)	0.493	1.638	4.352(4.202)	0.908(0.936)	1.564	1.110(1.072)
logistic	**poly2**	-0.006	0.401	1.058(1.080)	0.916(0.954)	0.401	0.270(0.275)
	poly3	-0.013	0.562	1.457(1.483)	0.918(0.942)	0.563	0.372(0.378)
		Setting f					
	$\bar{Y}_{\text{labeled}}$	2.613	2.670	2.182	0.000	0.549	0.557
	oracle	-0.003	0.164	0.623	0.966	0.164	0.159
	poly1(LS)	0.302	1.267	4.406(4.163)	0.914(0.918)	1.232	1.124(1.062)
logistic	poly2	-0.018	0.584	1.752(1.800)	0.862(0.900)	0.584	0.447(0.459)
	poly3	-0.005	0.410	1.279(1.316)	0.894(0.926)	0.410	0.326(0.336)

(NHEFS). As per [HR10], the NHEFS was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the United States Public Health Service. The NHEFS dataset has been studied by [HR10] and [EHvdL20]. The subset of the NHEFS data we consider consists of $N = 1561$ cigarette smokers aged $25 - 74$ years, who had a baseline visit and a follow-up visit approximately 10 years later. We consider two types of product (joint) treatment indicators $R_1^{(1)}, R_1^{(2)} \in \{0, 1\}$: $R^{(1)} = 1$ denotes that the individual has not quit smoking before the follow-up visit and has not drunk alcohol before the baseline visit, and $R^{(1)} = 0$ otherwise; $R^{(2)} = 1$ denotes that

the individual has quit smoking before the follow-up visit and has not drunk alcohol before the baseline visit, and $R^{(2)} = 0$ otherwise. We omit 5 individuals whose alcohol drinking information was missing. The weight gain (in kg), $Y \in \mathbb{R}$, was measured as the body weight at the follow-up visit minus the body weight at the baseline visit. Same as in [HR10] and [EHvdL20], the following 9 confounding variables, $\mathbf{X}$ are considered: sex (0: male, 1: female), age (in years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (number of cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg).

We estimate the ATE of the product (joint) treatments $R^{(1)}$ and $R^{(2)}$ on weight gain. The ATE estimators $\widehat{\theta}^{(1)}_{\mathrm{ATE}}$ and $\widehat{\theta}^{(2)}_{\mathrm{ATE}}$ are constructed using (2.41), based on samples $\mathbb{S}^{(1)} := (Y_i, R_i^{(1)}, \mathbf{X}_i)_{i=1}^N$ and $\mathbb{S}^{(2)} := (Y_i, R_i^{(2)}, \mathbf{X}_i)_{i=1}^N$, respectively. Recall that the sample size is $N = 1561$, and the dimension (after converting categorical variables) is $p = 12$. The estimated proportions of the treated groups are $\widehat{\pi}_N^{(1)} := N^{-1}\sum_{i=1}^N R_i^{(1)} = 0.088$ and $\widehat{\pi}_N^{(2)} := N^{-1}\sum_{i=1}^N R_i^{(2)} = 0.037$. We consider three PS estimators: a constant estimator, an offset based logistic estimator with a Lasso-type penalty (log-Lasso), and a random forest (RF). We consider four outcome models: a Lasso estimator, a degree-2 polynomial estimator without interactions and with a Lasso-type penalty (poly-Lasso), a random forest (RF), and a reproducing kernel Hilbert space (RKHS) estimator. We compare the proposed estimators with naive empirical difference (empdiff) estimators, $(\sum_{i=1}^N R_i^{(j)})^{-1}\sum_{i=1}^N R_i^{(j)}Y_i - \{\sum_{i=1}^N (1 - R_i^{(j)})\}^{-1}\sum_{i=1}^N (1 - R_i^{(j)})Y_i$, for $j = 1, 2$. To reduce the randomness coming from the sample splitting, we repeat the sample splitting for $B = 10$ times and report the median of the ATE estimators based on each split. The asymptotic variance is then estimated by a plugged-in version using the mean estimators as well as the asymptotic variance estimators based on

each split; see more details of this technique in Definition 3.3 of [CCD$^+$18].

We report the ATE estimators, the corresponding 95% confidence intervals, and the length of the confidence intervals in the Table 2.13. We can see negative estimated ATEs for $\theta^{(1)}_{\text{ATE}}$ and positive estimated ATEs for $\theta^{(2)}_{\text{ATE}}$. Moreover, our proposed ATE estimators are close to each other and fairly different from the empirical difference estimator, especially for $\theta^{(2)}_{\text{ATE}}$. Therein, all our confidence intervals do not include 0 while the one based on the empirical difference does. The difference between our proposed ATE estimators and the empirical difference estimator seems to suggest presence of substantial confounding via $\mathbf{X}$, and a significant causal effect of the treatment on the response after adjusting for the confounding.

Table 2.13: Real data analysis: estimation and inference of $\theta^{(1)}_{\text{ATE}}$ and $\theta^{(2)}_{\text{ATE}}$. We compare a naive empirical difference (empdiff) estimator with our proposed estimators based on various choices of nuisance estimators. ATE: the estimated average treatment effect; CI: a 95% confidence interval; Length: length of the 95% confidence interval.

		$\theta^{(1)}_{\text{ATE}}$			$\theta^{(2)}_{\text{ATE}}$		
$\widehat{\pi}_N(\cdot)$	$\widehat{m}(\cdot)$	ATE	CI	Length	ATE	CI	Length
empdiff		-2.003	(-3.282,-0.725)	2.558	1.867	(-0.642,4.377)	5.019
constant	Lasso	-1.935	(-3.219,-0.651)	2.568	4.209	(1.743,6.676)	4.933
	poly-Lasso	-1.865	(-3.152,-0.578)	2.574	3.291	(0.719,5.864)	5.145
	RF	-1.729	(-2.992,-0.466)	2.526	3.095	(0.607,5.584)	4.977
	RHKS	-1.941	(-3.227,-0.655)	2.573	3.183	(0.642,5.723)	5.081
log-Lasso	Lasso	-1.967	(-3.518,-0.416)	3.102	4.780	(1.954,7.605)	5.650
	poly-Lasso	-1.717	(-3.321,-0.113)	3.207	3.890	(0.804,6.976)	6.172
	RF	-1.873	(-3.424,-0.322)	3.102	3.890	(0.955,6.825)	5.870
	RHKS	-2.051	(-3.663,-0.440)	3.223	4.221	(1.068,7.375)	6.307
RF	Lasso	-1.727	(-2.970,-0.484)	2.486	4.932	(1.518,8.345)	6.827
	poly-Lasso	-1.612	(-2.878,-0.346)	2.532	4.693	(0.763,8.622)	6.827
	RF	-1.608	(-2.845,-0.371)	2.474	4.456	(0.942,7.970)	7.027
	RHKS	-1.772	(-3.016,-0.528)	2.487	4.411	(0.621,8.200)	7.579

2.8 Discussion

In this chapter, we study the mean estimation problem in the semi-supervised setting with a decaying PS while allowing for selection bias in the labeling mechanism. To our knowledge this is one of the first full-hearted attempts in extending the SS inference literature to the case of selection bias, and that too in a very general way, as well as the MAR literature to the case of a (uniformly) decaying PS. The proposed DRSS mean estimator is based on estimators of the outcome and the decaying PS models. We establish estimation and inference results under different cases of the correctness of the models, while allowing flexible model choices, including high-dimensional and non-parametric methods. The subtleties of the problem setting and the non-standard asymptotics, among others, make the method and its analyses challenging and our results reveal several novel insights in the process. In particular, we find that the consistency rate of the proposed estimator depends on the (expected) size of the labeled sample and the tail of the PS distribution. Throughout the chapter, Na_N (recall that $a_N = [E\{\pi_N^{-1}(\mathbf{X})\}]^{-1}$) is a crucial value, in that it serves as the "effective sample size" in our MAR-SS setting with a decaying PS. This chapter provides details as to why this happens.

As a necessary component of analyzing the MAR-SS setting, we further propose estimators of the decaying PS under three different models: MCAR, stratified labeling, and a novel offset logistic model, under both high and low dimensional settings. The consistency rates of the PS models are established, which are of independent interest. We also extend our methods to an ATE estimation problem where the treatment groups can be extremely imbalanced. We provide extensive numerical studies to illustrate our results in finite-sample

simulations, as well as a real data analysis using the NHEFS data.

The semi-supervised decaying PS setting is an interesting scenario that occurs in numerous applications in the modern era, and yet has been largely under-studied so far. We provide a detailed analysis of the mean estimation problem under such setting. We hope it serves as a start towards understanding this practically very relevant, and yet technically challenging, scenario in all its subtleties, therefore opening doors to many new questions where inferential results need to be adjusted for the "effective sample size".

2.9 Proofs of main results

Notation Constants $c, C > 0$, independent of N and p, may change values from one line to the other. For any $\tilde{\mathbb{S}} \subseteq \mathbb{S} = (\mathbf{Z}_i)_{i=1}^N$, define $P_{\tilde{\mathbb{S}}}$ as the joint distribution of $\tilde{\mathbb{S}}$ and $E_{\tilde{\mathbb{S}}}(f) = \int f dP_{\tilde{\mathbb{S}}}$. For any $r > 0$, let $\|f(\cdot)\|_{r,P} := \{E|f(\mathbf{Z})|^r\}^{1/r}$. We abbreviate "with probability approaching one" and "almost surely" by "w.p.a. 1" and "a.s.", respectively.

2.9.1 Auxiliary lemmas

The following Lemmas will be useful in the proofs.

Lemma 2.4. *Let $(X_N)_{N\geq 1}, (Y_N)_{N\geq 1}$ be sequences of random variables. If $E(|X_N|^r|Y_N) = O_p(1)$ for any $r \geq 1$, then $X_N = O_p(1)$.*

Proof of Lemma 2.4. For any $c > 0$, there exists $C > 0$ such that, for large enough N,

$$\mathbb{P}\left\{E\left(X_N^r \middle| Y_N\right) > C\right\} \;<\; c/2.$$

Let $\delta = (2C/c)^{1/r}$, then

$$\begin{aligned}
\mathbb{P}(|X_N| > \delta) &= E\left(E\left[\mathbb{1}\{|X_N| > \delta\}\middle|Y_N\right]\right) \\
&= E\left[\mathbb{1}\{E(|X_N|^r\big|Y_N) \leq C\}E\left(\mathbb{1}\{|X_N| > \delta\}\middle|Y_N\right)\right] \\
&\quad + E\left[\mathbb{1}\{E(|X_N|^r\big|Y_N) > C\}E\left(\mathbb{1}\{|X_N| > \delta\}\middle|Y_N\right)\right] \\
&\leq E\left[\mathbb{1}\{E(|X_N|^r\big|Y_N) \leq C\}E\left(\delta^{-r}|X_N|^r\middle|Y_N\right)\right] + E\left(\mathbb{1}\{E(|X_N|^r\big|Y_N) > C\}\right) \\
&= \delta^{-r}E\left[\mathbb{1}\{E(|X_N|^r\big|Y_N) \leq C\}E\left(|X_N|^r\middle|Y_N\right)\right] + \mathbb{P}\left[E(|X_N|^r\big|Y_N) > C\right] \\
&\leq c/2 + c/2 = c.
\end{aligned}$$

That is, $X_N = O_p(1)$. ∎

Lemma 2.5 (Lemma 6.1 of [CCD+18]). *Let $(X_N)_{N\geq 1}$ and $(Y_N)_{N\geq 1}$ be sequences of random variables in $\mathbb{R}$. If for any $c > 0$, $\mathbb{P}(|X_N| > c|Y_N) = o_p(1)$, then $X_N = o_p(1)$.*

In particular, Lemma 2.5 occurs if $E(|X_N|^q|Y_N) = o_p(1)$ for some $q \geq 1$. A typical example we used in our proofs is $X_N = \sum_{i=1}^N Z_{N,i}/N$, where $(Z_{N,i})_{N\geq 1, i\leq N}$ is a row-wise independent and identically distributed triangular array with $E(|Z_{N,i}||Y_N) = o_p(1)$.

Lemma 2.6. *Let $(Z_{N,i})_{N\geq 1, i\leq N}$ be a row-wise independent and identical distributed triangular array, suppose there exists a sequence b_N such that $N^{-r}b_N^{-1-r}E\left(|Z_{N,1}|^{1+r}\right) = o(1)$ with $0 < r < 1$ and $b_N > 0$. Then,*

$$N^{-1}\sum_{i=1}^N Z_{N,i} - E(Z_{N,1}) \;=\; o_p(b_N).$$

Proof of Lemma 2.6. Let $Y_{N,i} = Z_{N,i}\mathbb{1}\{|Z_{N,i}| \leq Nb_N)\}$. For any $c > 0$,

$$\begin{aligned}
&P\left(\left|N^{-1}\sum_{i=1}^{N} Z_{N,i} - E(Y_{N,1})\right| \geq cb_N\right) \\
&\quad \leq P\left(\cup_{i=1}^{N}[Z_{N,i} \neq Y_{N,i}] \cup \left[\left|\sum_{i=1}^{N} Y_{N,i} - E(Y_{N,1})\right| \geq Ncb_N\right]\right) \\
&\quad \leq P\left(\cup_{i=1}^{N}[Z_{N,i} \neq Y_{N,i}]\right) + P\left(\left|\sum_{i=1}^{N} Y_{N,i} - E(Y_{N,1})\right| \geq Ncb_N\right),
\end{aligned}$$

where with a slight abuse of notation, here P denotes the joint distribution of $(Z_{N,i})_{N\geq 1, i\leq N}$.

By Markov's inequality,

$$\begin{aligned}
&P\left(\cup_{i=1}^{N}[Z_{N,i} \neq Y_{N,i}]\right) \leq NP(Z_{N,1} \neq Y_{N,1}) = NP(|Z_{N,1}| > Nb_N) \\
&\quad \leq N(Nb_N)^{-1-r}E\left(|Z_{N,1}|^{1+r}\right) = N^{-r}b_N^{-1-r}E\left(|Z_{N,1}|^{1+r}\right) = o(1),
\end{aligned}$$

where the last equality follows from the assumptions. Moreover, by Chebyshev's inequality

$$\begin{aligned}
&P\left(\left|\sum_{i=1}^{N} Y_{N,i} - E(Y_{N,1})\right| \geq Ncb_N\right) \leq (Ncb_N)^{-2}E\left\{\left|\sum_{i=1}^{N} Y_{N,i} - E(Y_{N,1})\right|^2\right\} \\
&\quad = c^{-2}N^{-1}b_N^{-2}E\{Y_{N,i} - E(Y_{N,i})\}^2 \leq c^{-2}N^{-1}b_N^{-2}E(Y_{N,i}^2) \\
&\quad = c^{-2}N^{-1}b_N^{-2}E[Z_{N,1}^2\mathbb{1}\{|Z_{N,1}| \leq Nb_N\}] \leq c^{-2}N^{-1}b_N^{-2}(Nb_N)^{1-r}E\left(|Z_{N,1}|^{1+r}\right) \\
&\quad = c^{-2}N^{-r}b_N^{-1-r}E\left(|Z_{N,1}|^{1+r}\right) = o(1),
\end{aligned}$$

where in the second to last inequality we used Markov's inequality on $Z_{N,1}$s. Hence,

$$N^{-1}\sum_{i=1}^{N} Z_{N,i} - E(Y_{N,1}) = o_p(b_N).$$

In addition, by similar arguments

$$\begin{aligned}
&E(Z_{N,1}) - E(Y_{N,1}) = E\left[Z_{N,1}\mathbb{1}\{|Z_{N,1}| > Nb_N\}\right] = E\left[|Z_{N,1}|^{1+r}|Z_{N,1}|^{-r}\mathbb{1}\{|Z_{N,1}| > Nb_N\}\right] \\
&\quad \leq (Nb_N)^{-r}E\left(|Z_{N,1}|^{1+r}\right) = b_N N^{-r}b_N^{-1-r}E\left(|Z_{N,1}|^{1+r}\right) = o(b_N).
\end{aligned}$$

Therefore,

$$N^{-1}\sum_{i=1}^{N} Z_{N,i} - E(Z_{N,1}) = N^{-1}\sum_{i=1}^{N} Z_{N,i} - E(Z_{N,1}) + E(Z_{N,1}) - E(Y_{N,1}) = o_p(b_N).$$

■

Lemma 2.7. *For any function $g(\cdot)$ and $\theta \in \mathbb{R}$, define*

$$\psi(\mathbf{Z}, \theta) \;:=\; g(\mathbf{Z}) - \theta.$$

Let $\theta^0 := E\{g(\mathbf{Z})\}$. Assume

$$E\{\psi^2(\mathbf{Z}, \theta^0)\} \;\asymp\; b_N^{-1}, \;\; N^{-r} b_N^{r+1} E\{|\psi(\mathbf{Z}, \theta^0)|^{2+2r}|\} \;=\; o(1), \tag{2.44}$$

for some sequence b_N and $0 < r < 1$. Moreover, let $\widehat{\theta} \in \mathbb{R}$ be such that $\widehat{\theta} - \theta^0 = o_p(b_N^{-1/2})$. Additionally, for $k \leq \mathbb{K}$, and some (possibly random) function $g_{-k}(\cdot) \perp\!\!\!\perp \mathbb{S}_k$, define $\psi_{-k}(\mathbf{Z}, \theta) := g_{-k}(\mathbf{Z}) - \theta$ and suppose that

$$E\{\psi_{-k}(\mathbf{Z}, \theta^0) - \psi(\mathbf{Z}, \theta^0)\}^2 \;=\; o_p(b_N^{-1}).$$

Then, as $N \to \infty$, we have

$$N^{-1}\sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \psi_{-k}^2(\mathbf{Z}_i, \widehat{\theta}) \;=\; E\{\psi^2(\mathbf{Z}, \theta^0)\}\{1 + o_p(1)\}.$$

Proof of Lemma 2.7. By Young's inequality with $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned} &|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \widehat{\theta}) - \psi(\mathbf{Z}_i, \theta^0)\}^2 \\ &\quad \leq 2(\widehat{\theta} - \theta^0)^2 + 2|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \theta^0) - \psi(\mathbf{Z}_i, \theta^0)\}^2 = o_p(b_N^{-1}). \end{aligned} \tag{2.45}$$

In what follows we will use the following equality which is a consequence of Lemma 2.6 and the condition in (2.44):

$$|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \theta^0) - \psi(\mathbf{Z}_i, \theta^0)\}^2 = o_p(b_N^{-1}). \tag{2.46}$$

Using the fact that $a^2 - b^2 = (a+b)(a-b) = (a-b)^2 + 2b(a-b)$, and using the triangle and then Cauchy-Schwarz inequality

$$\begin{aligned}
&\left| |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \psi^2_{-k}(\mathbf{Z}_i, \widehat{\theta}) - |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \psi^2(\mathbf{Z}_i, \theta^0) \right| \\
&= \Big| |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \widehat{\theta}) - \psi(\mathbf{Z}_i, \theta^0)\}^2 + 2|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \widehat{\theta}) - \psi(\mathbf{Z}_i, \theta^0)\} \psi(\mathbf{Z}_i, \theta^0) \Big| \\
&\leq |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \widehat{\theta}) - \psi(\mathbf{Z}_i, \theta^0)\}^2 \\
&\qquad + 2|\mathcal{I}_k|^{-1} \left[\sum_{i \in \mathcal{I}_k} \{\psi_{-k}(\mathbf{Z}_i, \widehat{\theta}) - \psi(\mathbf{Z}_i, \theta^0)\}^2 \sum_{i \in \mathcal{I}_k} \psi^2(\mathbf{Z}_i, \theta^0) \right]^{1/2} \\
&\overset{(i)}{=} o_p(a_N^{-1}) + o_p(a_N^{-1/2}) [E\{\psi^2(\mathbf{Z}, \theta^0)\}\{1 + o_p(1)\}]^{1/2} \overset{(ii)}{=} o_p(a_N^{-1}),
\end{aligned}$$

where (i) follows by (2.45) and (2.46), and in (ii), we utilized the assumption $E\{\psi^2(\mathbf{Z}, \theta^0)\} \asymp a_N^{-1}$ to conclude the asymptotic order of the quantities of interest. Then, by utilizing the result of Lemma 2.6, i.e., (2.46), we have

$$\begin{aligned}
N^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \psi^2_{-k}(\mathbf{Z}_i, \widehat{\theta}) &= |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \psi^2(\mathbf{Z}_i, \theta^0) + o_p(a_N^{-1}) \\
&= E\{\psi^2(\mathbf{Z}, \theta^0)\}\{1 + o_p(1)\} + o_p(a_N^{-1}) = E\{\psi^2(\mathbf{Z}, \theta^0)\}\{1 + o_p(1)\},
\end{aligned}$$

since $E\{\psi^2(\mathbf{Z}, \theta^0)\} \asymp a_N^{-1}$ by assumption. ∎

Lemma 2.8. *The following are some useful properties regarding sub-Gaussian variables.*

(a) If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_2} \leq \|Y\|_{\psi_2}$. If $|X| \leq M$ a.s. for some constant M, then $\|X\|_{\psi_2} \leq \{\log(2)\}^{-1/2} M$.

(b) If $\|X\|_{\psi_2} \leq \sigma$, then $E(|X|^m) \leq 2\sigma^m \Gamma(m/2+1)$, for all $m \geq 1$, where $\Gamma(a) := \int_0^\infty x^{a-1} \exp(-x) dx$ denotes the Gamma function. As a result, $E(|X|) \leq \sigma\sqrt{\pi}$ and $E(|X|^m) \leq 2\sigma^m (m/2)^{m/2}$ for $m \geq 2$.

(c) If $\|X - E(X)\|_{\psi_2} \leq \sigma$, *then* $E(\exp[t\{X - E(X)\}]) \leq \exp(2\sigma^2 t^2)$, *for all* $t \in \mathbb{R}$.

(d) Let $\mathbf{X} \in \mathbb{R}^p$ *be a random vector with* $\sup_{1\leq j\leq p} \|\mathbf{X}(j)\|_{\psi_2} \leq \sigma$. *Then,* $\|\|\mathbf{X}\|_\infty\|_{\psi_2} \leq \sigma\{\log(p) + 2\}^{1/2}$.

(e) Let $(X_i)_{i=1}^N$ *be independent random variables with means* $(\mu_i)_{i=1}^N$ *such that* $\|X_i - \mu_i\|_{\psi_2} \leq \sigma$. *Then,* $\|N^{-1}\sum_{i=1}^N (X_i - \mu_i)\|_{\psi_2} \leq 4\sigma N^{-1/2}$.

Lemma 2.8 is a simple consequence of Lemmas D.1 and D.2 of [CLCL19].

Lemma 2.9. *Assume* $(\mathbf{X}_i)_{i=1}^N$ *are independent and identically distributed,* $\lambda_{\min}\{E(\overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i^T) \geq c > 0$ *and* $\sup_{\|\mathbf{v}\|_2=1} E\{(\overrightarrow{\mathbf{X}}_i^T \mathbf{v})^4\} < C < \infty$, *with constants* c *and* C. *Assume* $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p+1}$ *satisfyes*$\|\boldsymbol{\gamma}_0\|_2 < C < \infty$, $\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0$ *is a sub-Gaussian random variable, and* $\mathbf{X}_i$ *is a marginal sub-Gaussian random vector with*

$$\|\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0\|_{\psi_2} \;=\; \inf\left\{t > 0 : E[\exp\{t^{-2}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0)^2\}] \leq 2\right\} \;<\; \infty, \tag{2.47}$$

$$\sup_{1\leq j\leq p} \|\mathbf{X}_i(j)\|_{\psi_2} \;=\; \inf\left\{t > 0 : E[\exp\{t^{-2}\mathbf{X}_i^2(j)\}] \leq 2\right\} \;<\; \infty. \tag{2.48}$$

Recall that

$$\ell_N^{bal}(\boldsymbol{\gamma}) \;=\; -N^{-1}\sum_{i=1}^N [R_i^* \overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma} - \log\{1 + \exp(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma})\}] \quad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^{p+1},$$

$$\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}) \;=\; \ell_N^{bal}(\boldsymbol{\gamma} + \boldsymbol{\Delta}) - \ell_N^{bal}(\boldsymbol{\gamma}) - \boldsymbol{\Delta}^T \boldsymbol{\nabla}_{\boldsymbol{\gamma}} \ell_N^{bal}(\boldsymbol{\gamma}) \quad \forall\, \boldsymbol{\gamma}, \boldsymbol{\Delta} \in \mathbb{R}^{p+1}.$$

where $(R_i^*)_{i=1}^N$ *are i.i.d. pseudo binary random variables satisfying* $P(R_i^* = 1|\mathbf{X}) = g(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)$. *Then, for some constants* $c_1, c_2, c_3, c_4 > 0$,

$$\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0) \;\geq\; c_1 \|\boldsymbol{\Delta}\|_2 \left\{ \|\boldsymbol{\Delta}\|_2 - c_2 \sqrt{\frac{\log(p+1)}{N}} \|\boldsymbol{\Delta}\|_1 \right\} \;\; \forall\, \boldsymbol{\Delta} \in \mathbb{R}^{p+1}, \;\; \|\boldsymbol{\Delta}\|_2 \leq 1,$$

with probability at least $1 - c_3 \exp(-c_4 N)$.

Lemma 2.9 is a slightly more general version of Proposition 2 of [NRWY10]: instead of assuming $\mathbf{X}$ to be joint sub-Gaussian with mean zero, one can repeat their proof by only requiring $\mathbf{X}$ to be a marginal sub-Gaussian vector and $\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0$ be a sub-Gaussian variable, as well as an additional 4-th moment condition that $\sup_{\|\mathbf{v}\|_2=1} E\{(\overrightarrow{\mathbf{X}}^T\mathbf{v})^4\} < C < \infty$. Unlike [NRWY10], the intercept term is also considered here: since we do not require zero-mean covariates, the intercept term $\overrightarrow{\mathbf{X}}(1) = 1$ can be seen as a sub-gaussian variable.

Lemma 2.10. *(Theorem 3.26 of [Wai19]) Let $\mathcal{F}$ be a class of functions of the form $f : \mathcal{X} \to \mathbb{R}$, and let $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ be drawn from a product distribution $P = \bigotimes_{i=1}^N P_i$, where each P_i is supported on some set $\mathcal{X}_i \subseteq \mathcal{X}$. For each $f \in \mathcal{F}$ and $i = 1, \ldots, N$, assume that there are real numbers $a_{i,f} \leq b_{i,f}$ such that $f(\mathbf{x}) \in [a_{i,f}, b_{i,f}]$ for all $\mathbf{x} \in \mathcal{X}_i$. Let $Z = \sup_{f\in\mathcal{F}} \left\{ N^{-1} \sum_{i=1}^N f(\mathbf{X}_i) \right\}$. Then for all $t \geq 0$, we have $P\{Z \geq E(Z) + t\} \leq \exp\left(-Nt^2/4L^2\right)$, where $L^2 := \sup_{f\in\mathcal{F}} \{N^{-1} \sum_{i=1}^N (b_{i,f} - a_{i,f})^2\}$.*

2.9.2 Proofs of the Main Statements

Proof of Theorem 2.1. We prove Theorem 2.1 by decomposing the estimation error into two terms: $N^{-1}\sum_{i=1}^N \psi_{\mu,\pi}(\mathbf{Z}_i)$ and $\widehat{\Delta}_{N,1,k}$ defined below in (2.49). We use Lemma 2.4 and the Lindeberg-Feller theorem for self-normalized partial sums. Observe that

$$\tilde{\theta} - \theta_0 = N^{-1} \sum_{i=1}^N \left\{ \frac{R_i - \pi_N(\mathbf{X}_i)}{\pi_N(\mathbf{X}_i)} [Y_i - \widehat{m}(\mathbf{X}_i)] + Y_i - \theta_0 \right\} = N^{-1} \sum_{i=1}^N \psi_{\mu,\pi}(\mathbf{Z}_i) + \sum_{k=1}^{\mathbb{K}} \widehat{\Delta}_{N,1,k}, \tag{2.49}$$

where

$$\widehat{\Delta}_{N,1,k} = -N^{-1} \sum_{i\in\mathcal{I}_k} \left\{ \frac{R}{\pi_N(\mathbf{X}_i)} - 1 \right\} \{\widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu(\mathbf{X}_i)\}. \tag{2.50}$$

Consider the remainder term $\widehat{\Delta}_{N,1,k}$. For each $k \leq \mathbb{K}$, notice that $\widehat{\Delta}_{N,1,k}$ is a summation of independent random variables conditional on the training sample $\mathbb{S}_{-k}$:

$$\widehat{\Delta}_{N,1,k} = -N^{-1} \sum_{i \in \mathcal{I}_k} \xi_i, \qquad \xi_i = \left\{ \frac{R_i}{\pi_N(\mathbf{X}_i)} - 1 \right\} \{\widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu(\mathbf{X}_i)\},$$

with $\xi_i \perp \xi_j | \mathbb{S}_{-k}$ for $i, j \in \mathcal{I}_k$. Hence, with

$$\xi = \left\{ \frac{R}{\pi_N(\mathbf{X})} - 1 \right\} \{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\},$$

and recall that $E_{\mathbb{S}_k}$ denotes the expectation with respect to (w.r.t.) the samples in the k-th fold,

$$E_{\mathbb{S}_k}(\widehat{\Delta}_{N,1,k}) = - N^{-1} |\mathcal{I}_k| E\left\{ E\left(\xi \middle| \mathbf{X}\right)\right\} = 0, \tag{2.51}$$

$$E_{\mathbb{S}_k}(\widehat{\Delta}^2_{N,1,k}) = N^{-2} |\mathcal{I}_k| E\left(E\left[\left\{ \frac{R}{\pi_N(\mathbf{X})} - 1 \right\}^2 \{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\}^2 \middle| \mathbf{X} \right] \right) \tag{2.52}$$

$$= N^{-2} |\mathcal{I}_k| E\left\{ \left[\frac{1}{\pi_N(\mathbf{X})} - 1 \right] [\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})]^2 \right\} \tag{2.53}$$

$$= O_p((N a_N)^{-1} c^2_{\mu,N}), \tag{2.54}$$

In the above equations, (2.51) and (2.52) used the fact that $\xi_i \perp \xi_j | \mathbb{S}_{-k}$ for $i, j \in \mathcal{I}_k$; (2.53) used the fact that $R^2 = R$; (2.54) used the fact that $|\mathcal{I}_k| < N$ and the definition of $c_{\mu,N}$; the definition $E(R|\mathbf{X}) = \pi_N(\mathbf{X})$ is also used in (2.51) and (2.53). These techniques will be used for multiple times throughout the proof, and we will not emphasis them in again in the following proofs.

By Lemma 2.4,

$$\widehat{\Delta}_{N,1,k} = O_p((N a_N)^{-1/2} c_{\mu,N}). \tag{2.55}$$

As for the influence function $N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i)$,

$$E_{\mathbb{S}}\left[N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i)\right] = E\left(E\left\{\mu(\mathbf{X})-\theta_0+\frac{R[Y-\mu(\mathbf{X})]}{\pi_N(\mathbf{X})}\big|\mathbf{X}\right\}\right)=0,$$
$$E_{\mathbb{S}}\left[N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i)\right]^2 = N^{-1}E\left[E\left(\left\{\mu(\mathbf{X})-\theta_0+\frac{R[Y-\mu(\mathbf{X})]}{\pi_N(\mathbf{X})}\right\}^2\big|\mathbf{X}\right)\right]=N^{-1}V_N(\mu),$$

where

$$V_N(\mu) = E\left[\mu(\mathbf{X})-\theta_0+\frac{R\{Y-\mu(\mathbf{X})\}}{\pi_N(\mathbf{X})}\right]^2 = E\left[\frac{\{R-\pi_N(\mathbf{X})\}\{Y-\mu(\mathbf{X})\}}{\pi_N(\mathbf{X})}+Y-\theta_0\right]^2$$
$$= E\left[\left\{\frac{1-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\right\}^2\{Y-\mu(\mathbf{X})\}^2\right]+\mathrm{Var}(Y).$$

To control the order of $V_N(\mu)$, we enforce uniform lower and upper bounds for $E[\{Y-\mu(\mathbf{X})\}^2|\mathbf{X}]$ and $\mathrm{Var}(Y)$. Under Assumption 2.2,

$$E[\{Y-m(\mathbf{X})\}^2|\mathbf{X}]\geq\sigma_{\zeta,1}^2,\qquad E[\{Y-\mu(\mathbf{X})\}^2|\mathbf{X}]\leq\sigma_{\zeta,2}^2,\qquad \mathrm{Var}(Y)\leq\sigma_{\zeta,2}^2.$$

Additionally, we have the following lower bounds as $m(\mathbf{X})=E(Y|\mathbf{X})$,

$$E[\{Y-\mu(\mathbf{X})\}^2|\mathbf{X}] = E[\{Y-m(\mathbf{X})\}^2|\mathbf{X}]+E[\{m-\mu(\mathbf{X})\}^2|\mathbf{X}]\geq\sigma_{\zeta,1}^2,$$
$$\mathrm{Var}(Y) = E\left(E[\{Y-m(\mathbf{X})\}^2|\mathbf{X}]+E[\{m(\mathbf{X})-\theta_0\}^2|\mathbf{X}]\right)\geq\sigma_{\zeta,1}^2.$$

Recall that by definition, $a_N=E\{\pi_N^{-1}(\mathbf{X})\}$. Therefore,

$$a_NV_N(\mu)\geq a_N\left\{\sigma_{\xi,1}^2E\left[\frac{1-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\right]+\sigma_{\xi,1}^2\right\}=\sigma_{\xi,1}^2>0, \tag{2.56}$$
$$a_NV_N(\mu)\leq a_N\left\{\sigma_{\xi,2}^2E\left[\frac{1-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\right]+\sigma_{\xi,2}^2\right\}=\sigma_{\xi,2}^2<\infty, \tag{2.57}$$

and $V_N(\mu)\asymp a_N^{-1}$. Since

$$E\left\{(Na_N)^{1/2}N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i)\right\}^2=a_NV_N(\mu)=O(1), \tag{2.58}$$

by Lemma 2.4, $N^{-1}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) = O_p((Na_N)^{-1/2})$. Therefore,

$$\widehat{\theta}_{\mathrm{DRSS}} - \theta_0 = O_p((Na_N)^{-1/2}).$$

In addition, under Assumption 2.3, for any $c > 0$,

$$N^{-1}\sum_{i=1}^{N} E[V_N^{-1}(\mu)\psi_{\mu,\pi}^2(\mathbf{Z}_i)\mathbb{1}\{V_N^{-1/2}(\mu)|\psi_{\mu,\pi}(\mathbf{Z}_i)| > cN^{1/2}\}] = o(1).$$

By Proposition 2.27 (Lindeberg-Feller theorem) of [VdV00],

$$V_N^{-1/2}(\mu)N^{-1/2}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) \to N(0,1). \tag{2.59}$$

Recall that

$$\begin{aligned}
N^{1/2}V_N^{-1/2}(\mu)(\tilde{\theta} - \theta_0) &= V_N^{-1/2}(\mu)N^{-1/2}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + N^{1/2}V_N^{-1/2}(\mu)\sum_{k=1}^{\mathbb{K}}\widehat{\Delta}_{N,1,k} \\
&= V_N^{-1/2}(\mu)N^{-1/2}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + O_p(N^{1/2}a_N^{1/2}(Na_N)^{-1/2}c_{\mu,N}) \\
&= V_N^{-1/2}(\mu)N^{-1/2}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + O_p(c_{\mu,N}) = V_N^{-1/2}(\mu)N^{-1/2}\sum_{i=1}^{N}\psi_{\mu,\pi}(\mathbf{Z}_i) + o_p(1).
\end{aligned}$$

By Lemma 2.8 (Slutsky) of [VdV00],

$$N^{1/2}V_N^{-1/2}(\mu)(\tilde{\theta} - \theta_0) \to N(0,1).$$

■

Proof of Theorem 2.2. We prove Theorem 2.2 by considering two cases: (a) the nuisance models are both correctly specified, and (b) only one of the nuisance models is correctly specified. For case (a), we design a suitable decomposition, (2.60), and apply Lemma 2.4 and the Lindeberg-Feller theorem for self-normalized sums to obtain asymptotic normality. For case (b), we design two different decompositions of the estimation error: one suitable

for the case when PS model is correct (2.68) and the other suitable for the case when the outcome model is correct (2.69).

Case (a): $\mu(\cdot) = m(\cdot)$ and $e_N(\cdot) = \pi_N(\cdot)$. Observe that

$$\begin{aligned}\widehat{\theta}_{\text{DRSS}} - \theta_0 &= N^{-1}\sum_{i=1}^{N}\left[\frac{R_i - \widehat{\pi}_N(\mathbf{X}_i;\mathbb{S}_{-k})}{\widehat{\pi}_N(\mathbf{X}_i;\mathbb{S}_{-k})}\{Y_i - \widehat{m}(\mathbf{X}_i;\mathbb{S}_{-k})\} + Y_i - \theta_0\right] \\ &= N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i) + \sum_{k=1}^{\mathbb{K}}(\widehat{\Delta}_{N,1,k} + \widehat{\Delta}_{N,2,k} + \widehat{\Delta}_{N,3,k}), \end{aligned} \tag{2.60}$$

where $\widehat{\Delta}_{N,1,k}$ is defined as (2.50) and we further define

$$\widehat{\Delta}_{N,2,k} = N^{-1}\sum_{i\in\mathcal{I}_k}\left\{\frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i;\mathbb{S}_{-k})} - \frac{R_i}{e_N(\mathbf{X}_i)}\right\}\{Y_i - m(\mathbf{X}_i)\}, \tag{2.61}$$

$$\widehat{\Delta}_{N,3,k} = -N^{-1}\sum_{i\in\mathcal{I}_k}\left\{\frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i;\mathbb{S}_{-k})} - \frac{R_i}{e_N(\mathbf{X}_i)}\right\}\{\widehat{m}(\mathbf{X}_i;\mathbb{S}_{-k}) - \mu(\mathbf{X}_i)\}. \tag{2.62}$$

Recall from (2.55), $\widehat{\Delta}_{N,1,k} = O_p((Na_N)^{-1/2}c_{\mu,N})$. As for the remainder term $\widehat{\Delta}_{N,2,k}$,

$$\begin{aligned} E_{\mathbb{S}_k}(\widehat{\Delta}_{N,2,k}) =& N^{-1}|\mathcal{I}_k|E\left(E\left[\left\{\frac{R}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{R}{e_N(\mathbf{X})}\right\}\{Y - m(\mathbf{X})\}\Big|\mathbf{X}\right]\right) = 0, \\ E_{\mathbb{S}_k}(\widehat{\Delta}^2_{N,2,k}) =& N^{-2}|\mathcal{I}_k|E\left(E\left[\left\{\frac{R}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{R}{e_N(\mathbf{X})}\right\}^2\{Y - m(\mathbf{X})\}^2\Big|\mathbf{X}\right]\right) \\ =& N^{-2}|\mathcal{I}_k|E\left[\frac{\pi_N(\mathbf{X})}{e_N^2(\mathbf{X})}\left\{1 - \frac{e_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}\right\}^2\{Y - m(\mathbf{X})\}^2\right] \qquad (2.63) \\ \overset{(i)}{\leq}& N^{-1}\sigma^2_{\xi,2}c^2_{e,N}a_N^{-1} = O_p((Na_N)^{-1}c^2_{e,N}), \qquad (2.64) \end{aligned}$$

where (i) holds under the Assumption 2.2 and the condition in (2.9) with $e_N(\cdot) = \pi_N(\cdot)$ and also noting the fact that $|\mathcal{I}_k| \leq N$. By Lemma 2.4,

$$\widehat{\Delta}_{N,2,k} = O_p((Na_N)^{-1/2}c_{e,N}). \tag{2.65}$$

Now, consider the last remainder term $\widehat{\Delta}_{N,3,k}$, by the triangular inequality and the tower

rule,

$$\begin{aligned}E_{\mathbb{S}_k}(|\widehat{\Delta}_{N,3,k}|) \leq & N^{-1}|\mathcal{I}_k|E\left[E\left\{\left|\frac{R}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}-\frac{R}{e_N(\mathbf{X})}\right||\widehat{m}(\mathbf{X}_i;\mathbb{S}_{-k})-\mu(\mathbf{X})|\Big|\mathbf{X}\right\}\right] \\ = & N^{-1}|\mathcal{I}_k|E\left\{\left|1-\frac{e_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}\right||\widehat{m}(\mathbf{X}_i;\mathbb{S}_{-k})-\mu(\mathbf{X})|\right\}=O_p(r_{\mu,N}r_{e,N}).\end{aligned}$$

By Lemma 2.4,

$$\widehat{\Delta}_{N,3,k}=O_p(r_{\mu,N}r_{e,N}). \tag{2.66}$$

Lastly, for the influence function $N^{-1}\sum_{i=1}^N\psi_{\mu,e}(\mathbf{Z}_i)$,

$$\begin{aligned}E_{\mathbb{S}}\left\{N^{-1}\sum_{i=1}^N\psi_{\mu,e}(\mathbf{Z}_i)\right\} &= E\left(E\left[\mu(\mathbf{X})-\theta_0+\frac{R\{Y-m(\mathbf{X})\}}{\pi_N(\mathbf{X})}\Big|\mathbf{X}\right]\right)=0, \\ E_{\mathbb{S}}\left\{N^{-1}\sum_{i=1}^N\psi_{\mu,e}(\mathbf{Z}_i)\right\}^2 &= N^{-1}E\left\{E\left(\left[\mu(\mathbf{X})-\theta_0+\frac{R\{Y-m(\mathbf{X})\}}{\pi_N(\mathbf{X})}\right]^2|\mathbf{X}\right)\right\} \\ &= N^{-1}V_N(\mu,e).\end{aligned}$$

Now we control the rate of the variance, $V_N(\mu,e)$. Under Assumption 2.2, $\mathrm{Var}(Y)\geq E[\{Y-m(\mathbf{X})\}^2|\mathbf{X}]\geq\sigma_{\zeta,1}^2$, $E[\{Y-m(\mathbf{X})\}^2|\mathbf{X}]\leq\sigma_{\zeta,2}^2$ and $\mathrm{Var}(Y)\leq\sigma_{\zeta,2}^2$. Hence,

$$\begin{aligned}a_NV_N(\mu,e) &\geq a_N\left[\sigma_{\xi,1}^2E\left\{\frac{1-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\right\}+\sigma_{\xi,1}^2\right]\geq\sigma_{\zeta,1}^2>0, \\ a_NV_N(\mu,e) &\leq a_N\left[\sigma_{\xi,2}^2E\left\{\frac{1-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\right\}+\sigma_{\xi,2}^2\right]\leq\sigma_{\zeta,2}^2<\infty.\end{aligned}$$

It follows that $V_N(\mu,e)\asymp a_N^{-1}$.

Recall the definition of $\psi_{\mu,e}$ in (2.6). By Lemma 2.4,

$$N^{-1}\sum_{i=1}^N\psi_{\mu,e}(\mathbf{Z}_i)=O_p((Na_N)^{-1/2}).$$

Therefore, $\widehat{\theta}_{\mathrm{DRSS}}-\theta_0=O_p((Na_N)^{-1/2})$. Moreover, by Proposition 2.27 (Lindeberg-Feller theorem) of [VdV00],

$$N^{1/2}V_N^{1/2}(\mu,e)N^{-1}\sum_{i=1}^N\psi_{\mu,e}(\mathbf{Z}_i)\to N(0,1). \tag{2.67}$$

By Lemma 2.8 (Slutsky) of [VdV00],

$$N^{1/2}V_N^{-1/2}(\mu, e)(\widehat{\theta}_{\text{DRSS}} - \theta_0) \to N(0, 1).$$

Case (b.i): $e_N(\cdot) = \pi_N(\cdot)$**.** Observe that

$$\widehat{\theta}_{\text{DRSS}} - \theta_0 = N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i) + \sum_{k=1}^{\mathbb{K}}(\widehat{\Delta}_{N,1,k} + \widehat{\Delta}_{N,2,k} + \widehat{\Delta}_{N,3,k} + \widehat{\Delta}_{N,4,k}), \tag{2.68}$$

where $\widehat{\Delta}_{N,1,k}$, $\widehat{\Delta}_{N,2,k}$, and $\widehat{\Delta}_{N,3,k}$ are defined as (2.50), (2.61), and (2.62), respectively, and we further define

$$\widehat{\Delta}_{N,4,k} = N^{-1}\sum_{i\in\mathcal{I}_k}\left\{\frac{R_i}{\widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})} - \frac{R_i}{e_N(\mathbf{X}_i)}\right\}\{m(\mathbf{X}_i) - \mu(\mathbf{X}_i)\}.$$

As shown in (2.55), (2.65), and (2.66), we have $\widehat{\Delta}_{N,1,k} = O_p((Na_N)^{-1/2}c_{\mu,N})$, $\widehat{\Delta}_{N,2,k} = O_p((Na_N)^{-1/2}c_{e,N})$ and $\widehat{\Delta}_{N,2,k} = O_p(r_{\mu,N}r_{e,N})$. In addition, for the remainder term $\widehat{\Delta}_{N,4,k}$,

$$\begin{aligned}
E_{\mathbb{S}_k}(|\widehat{\Delta}_{N,4,k}|) &\le N^{-1}|\mathcal{I}_k|E\left[E\left\{\left|\frac{R}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} - \frac{R}{e_N(\mathbf{X})}\right| |m(\mathbf{X}_i) - \mu(\mathbf{X})|\Big|\mathbf{X}\right\}\right] \\
&= N^{-1}|\mathcal{I}_k|E\left\{\left|1 - \frac{e_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})}\right| |m(\mathbf{X}_i) - \mu(\mathbf{X})|\right\} \\
&\le \left\|1 - \frac{e_N(\cdot)}{\widehat{\pi}_N(\cdot)}\right\|_{2,\mathbb{P}_{\mathbf{X}}} \|m(\cdot) - \mu(\cdot)\|_{2,\mathbb{P}_{\mathbf{X}}} = O_p(r_{e,N}).
\end{aligned}$$

By Lemma 2.4,

$$\widehat{\Delta}_{N,4,k} = O_p(r_{e,N}).$$

Lastly, for the influence function $N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i)$, similarly as in (2.58) and by Lemma 2.4,

$$N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i) = O_p((Na_N)^{-1/2}).$$

Case (b.ii): $\mu(\cdot) = m(\cdot)$**.** Observe that

$$\widehat{\theta}_{\text{DRSS}} - \theta_0 = N^{-1}\sum_{i=1}^{N}\psi_{\mu,e}(\mathbf{Z}_i) + \sum_{k=1}^{\mathbb{K}}(\widehat{\Delta}_{N,1,k} + \widehat{\Delta}_{N,2,k} + \widehat{\Delta}_{N,3,k} + \widehat{\Delta}_{N,4,k}), \tag{2.69}$$

where $\widehat{\Delta}_{N,1,k}$, $\widehat{\Delta}_{N,2,k}$, and $\widehat{\Delta}_{N,3,k}$ are defined as (2.50), (2.61), and (2.62), respectively, and we further define

$$\widehat{\Delta}_{N,5,k} = N^{-1} \sum_{i \in \mathcal{I}_k} \left\{ \frac{R_i}{\pi_N(\mathbf{X}_i)} - \frac{R_i}{e_N(\mathbf{X}_i)} \right\} \{\widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu(\mathbf{X}_i)\}.$$

Similarly as shown in (2.55), (2.65), and (2.66), we have $\widehat{\Delta}_{N,1,k} = O_p((Na_N)^{-1/2} c_{\mu,N})$, $\widehat{\Delta}_{N,2,k} = O_p((Na_N)^{-1/2} c_{e,N})$ and $\widehat{\Delta}_{N,2,k} = O_p(r_{\mu,N} r_{e,N})$. Here, the only difference from the previous proofs is that, in (2.63), instead of obtaining $\pi_N(\mathbf{X})/e_N^2(\mathbf{X}) = \pi_N^{-1}(\mathbf{X})$ using $e_N(\cdot) = \pi_N(\cdot)$, here we bound $\pi_N(\mathbf{X})/e_N^2(\mathbf{X}) \leq c^{-2}\pi_N^{-1}(\mathbf{X})$ by assuming that, a.s., $\pi_N(\mathbf{X})/e_N(\mathbf{X}) \geq c$. For the remainder term $\widehat{\Delta}_{N,5,k}$,

$$\begin{aligned}
E_{\mathbb{S}_k}(|\widehat{\Delta}_{N,5,k}|) &\leq N^{-1}|\mathcal{I}_k| E\left[E\left\{ \left| \frac{R}{\pi_N(\mathbf{X})} - \frac{R}{e_N(\mathbf{X})} \right| |\widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu(\mathbf{X})| \Big| \mathbf{X} \right\} \right] \\
&= N^{-1}|\mathcal{I}_k| E\left\{ \left| 1 - \frac{\pi_N(\mathbf{X})}{e_N(\mathbf{X})} \right| |\widehat{m}(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu(\mathbf{X})| \right\} \\
&\leq \|1 - \pi_N(\cdot)/e_N(\cdot)\|_{2,\mathbb{P}_\mathbf{X}} \|\widehat{m}(\cdot) - \mu(\cdot)\|_{2,\mathbb{P}_\mathbf{X}} = O_p(r_{\mu,N}).
\end{aligned}$$

By Lemma 2.4,

$$\widehat{\Delta}_{N,5,k} = O_p(r_{\mu,N}).$$

Lastly, for the influence function $N^{-1}\sum_{i=1}^N \psi_{\mu,e}(\mathbf{Z}_i)$, we have

$$\begin{aligned}
E_{\mathbb{S}}\left\{ N^{-1} \sum_{i=1}^N \psi_{\mu,e}(\mathbf{Z}_i) \right\} &= E\left(E\left[m(\mathbf{X}) - \theta_0 + \frac{R\{Y - m(\mathbf{X})\}}{e_N(\mathbf{X})} \Big| \mathbf{X} \right] \right) = 0, \\
E_{\mathbb{S}}\left\{ N^{-1} \sum_{i=1}^N \psi_{\mu,e}(\mathbf{Z}_i) \right\}^2 &= N^{-1} E\left\{ E\left(\left[m(\mathbf{X}) - \theta_0 + \frac{R\{Y - m(\mathbf{X})\}}{e_N(\mathbf{X})} \right]^2 \Big| \mathbf{X} \right) \right\} \\
&= N^{-1} V_N(\mu, e).
\end{aligned}$$

Here,

$$V_N(\mu, e) = \mathrm{Var}\{m(\mathbf{X})\} + E\{\pi_N(\mathbf{X})\{Y - m(\mathbf{X})\}^2/\{\pi_N(\mathbf{X})\}^2\}$$
$$\leq \sigma^2_{\zeta,2}(1 + E[\pi_N(\mathbf{X})/\{\pi_N(\mathbf{X})\}^2]) \leq \sigma^2_{\zeta,2}(1 + C^2 a_N^{-1}) = O(a_N^{-1}).$$

By Lemma 2.4,

$$N^{-1}\sum_{i=1}^{N} \psi_{\mu,e}(\mathbf{Z}_i) = O_p((Na_N)^{-1/2}).$$

■

Proof of Theorem 2.3. We prove the consistency results of the asymptotic variance estimators for the two cases (known PS and unknown PS). The results follows from Lemma 2.7 after we validate the conditions therein.

Case (a). By Lemma 2.7, it is sufficient to show $a_N E(\delta^2_{N,1,k}) = o_p(1)$, where

$$\delta_{N,1,k} = -\left\{\frac{R}{\pi_N(\mathbf{X})} - 1\right\}\{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\}.$$

Recall from (2.54), we have

$$a_N E(\widehat{\delta}^2_{N,1,k}) = O_p(c^2_{\mu,N}) = o_p(1).$$

Case (b). By Lemma 2.7, it is sufficient to show $a_N E(\delta_{N,1,k} + \delta_{N,2,k} + \delta_{N,3,k})^2 = o_p(1)$, where

$$\delta_{N,1,k} = -\left\{\frac{R}{\pi_N(\mathbf{X})} - 1\right\}\{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\},$$
$$\delta_{N,2,k} = \left\{\frac{R}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} - \frac{R}{e_N(\mathbf{X})}\right\}\{Y - m(\mathbf{X})\},$$
$$\delta_{N,3,k} = -\left\{\frac{R}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} - \frac{R}{e_N(\mathbf{X})}\right\}\{\widehat{m}(\mathbf{X}; \mathbb{S}_{-k}) - \mu(\mathbf{X})\}.$$

Recall (2.54) and (2.64), we have

$$a_N E(\widehat{\delta}^2_{N,1,k}) = O_p(c^2_{\mu,N}) = o_p(1), \qquad a_N E(\widehat{\delta}^2_{N,2,k}) = O_p(c^2_{e,N}) = o_p(1).$$

Besides, by the condition (2.14),

$$a_N E(\widehat{\delta}^2_{N,3,k}) = E\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}(\mathbf{X};\mathbb{S}_{-k})}\right\}^2 \{\widehat{m}(\mathbf{X};\mathbb{S}_{-k}) - m(\mathbf{X})\}^2\right] = o_p(1).$$

Therefore,

$$a_N E(\widehat{\Delta}_{N,1,k} + \widehat{\Delta}_{N,2,k} + \widehat{\Delta}_{N,3,k})^2 \leq 3a_N E(\widehat{\Delta}^2_{N,1,k} + \widehat{\Delta}^2_{N,2,k} + \widehat{\Delta}^2_{N,3,k}) = o_p(1).$$

■

Proof of Theorem 2.4. In the proof of Theorem 2.4, we work directly on the cross-fitted version of $\widehat{\boldsymbol{\gamma}}$. The results for a non cross-fitted $\widehat{\boldsymbol{\gamma}}$ can be obtained analogously by repeating the procedure using the full sample $\mathbb{S}$. Here, we first obtain an RAL expansion of the offset logistic regression estimator. Then, we establish the RAL expansion of the DRSS estimator.

For any $k \leq \mathbb{K}$, $a \in (0,1]$, and $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$, let

$$\ell_N(\boldsymbol{\gamma}; a) = -N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} \left[R_i \overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma} - \log\{1 + a\exp(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma})\}\right].$$

Define $g(u) = \exp(u)/\{1+\exp(u)\}$, then $\dot{g}(u) = g(u)\{1-g(u)\}$ and $\ddot{g}(u) = g(u)\{1-g(u)\}\{1-2g(u)\}$. We have

$$g(u + \log(a)) = \frac{a\exp(u)}{1 + a\exp(u)} \geq \frac{a\exp(u)}{1+\exp(u)} = ag(u), \quad \forall u \in \mathbb{R},\ a \in (0,1], \tag{2.70}$$

$$g(u) \leq \exp(u), \quad \dot{g}(u) \leq g(u) \leq \exp(u), \quad |\ddot{g}(u)| \leq g(u) \leq \exp(u), \quad \forall u \in \mathbb{R}. \tag{2.71}$$

For any $\mathbf{u} \in \mathbb{R}^{p+1}$, define

$$\ell_N(\mathbf{u}) = N_{-k}\left\{\ell_N(\boldsymbol{\gamma}_0 + (N_{-k}\pi_N)^{-1/2}\mathbf{u}, \widehat{\pi}_N) - \ell_N(\boldsymbol{\gamma}_0, \pi_N)\right\} - N_{-k}^{-1}\sum_{i \in \mathcal{I}_{-k}} R_i \log(\widehat{\pi}_N/\pi_N).$$

Since $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}$ minimizes $\ell_N(\boldsymbol{\gamma}; \widehat{\pi}_N)$, the terms $\ell_N(\boldsymbol{\gamma}_0, \pi_N)$ and $N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} R_i \log(\widehat{\pi}_N / \pi_N)$ are both independent of $\boldsymbol{\gamma}$, we know that $\mathbf{u}_N = (N_{-k}\pi_N)^{1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ minimizes $\ell_N(\mathbf{u})$. Here, $\widehat{\pi}_N = N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} R_i$ is the cross-fitted estimate of π_N. By Taylor's Theorem,where some $(\tilde{\boldsymbol{\gamma}}_1, \log(\widetilde{\pi}_{N,1}))$ lies between $(\boldsymbol{\gamma}_0, \log(\pi_N))$ and $(\boldsymbol{\gamma}_0 + (N_{-k}\pi_N)^{-1/2}\mathbf{u}, \log(\widehat{\pi}_N))$,

$$\ell_N(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T \mathbf{A}_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})\mathbf{u} + \mathbf{B}_{N,1}^T(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})\mathbf{u} + C_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}),$$

where

$$\begin{aligned}
\mathbf{A}_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) &= (N_{-k}\pi_N)^{-1} \sum_{i \in \mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T \tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1}))\overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i^T, \\
\mathbf{B}_{N,1}(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) &= -(N_{-k}\pi_N)^{-1/2} \sum_{i \in \mathcal{I}_{-k}} \Big\{ R_i - g(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \\
&\qquad - \dot{g}(\overrightarrow{\mathbf{X}}_i^T \tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1})) \log(\widehat{\pi}_N / \pi_N) \Big\} \overrightarrow{\mathbf{X}}_i, \\
C_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) &= \frac{1}{2} \sum_{i \in \mathcal{I}_{-k}} \Big\{ \dot{g}(\overrightarrow{\mathbf{X}}_i^T \tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1})) - \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \Big\} \left\{ \log(\widehat{\pi}_N / \pi_N) \right\}^2.
\end{aligned}$$

Define

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N) &= E\left\{ \overrightarrow{\mathbf{X}} \overrightarrow{\mathbf{X}}^T \dot{g}(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \right\}, \\
\mathbf{B}_{N,2} &= -(N_{-k}\pi_N)^{-1/2} \sum_{i \in \mathcal{I}_{-k}} \Big\{ R_i - g(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \\
&\qquad - \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \log(\widehat{\pi}_N / \pi_N) \Big\} \overrightarrow{\mathbf{X}}_i, \\
\boldsymbol{\zeta}_N &= (N_{-k}\pi_N)^{1/2} \mathcal{J}^{-1}(\boldsymbol{\gamma}_0, \pi_N) N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} \Big\{ R_i - g(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \\
&\qquad - \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \log(\widehat{\pi}_N / \pi_N) \Big\} \overrightarrow{\mathbf{X}}_i. \qquad (2.72)
\end{aligned}$$

Then, $\boldsymbol{\zeta}_N$ is the unique minimizer of

$$Z_N(\mathbf{u}) = \mathbf{u}^T \pi_N^{-1} \mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\mathbf{u}/2 + \mathbf{B}_{N,2}^T \mathbf{u}.$$

By Lemma 2 of [HP11], for each $\delta > 0$,

$$P_{\mathbb{S}_{-k}}(\|\mathbf{u}_N - \boldsymbol{\zeta}_N\|_2 \geq \delta) \leq P\left\{\Delta_N(\delta) \geq \frac{1}{2}h_N(\delta)\right\},$$

where $\mathbb{S}_{-k} = \mathbb{S} \setminus \mathbb{S}_k$ and

$$\Delta_N(\delta) = \sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \leq \delta} |\ell_N(\mathbf{u}) - Z_N(\mathbf{u})|, \quad h_N(\delta) = \inf_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 = \delta} Z_N(\mathbf{u}) - Z_N(\boldsymbol{\zeta}_N).$$

Hence, to prove

$$\left\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 - (N_{-k}\pi_N)^{-1/2}\boldsymbol{\zeta}_N\right\|_2 = (N_{-k}\pi_N)^{-1/2}\|\mathbf{u}_N - \boldsymbol{\zeta}_N\|_2 = o_p\left((N\pi_N)^{-1/2}\right), \tag{2.73}$$

it suffices to show that, for each $\delta > 0$, $\Delta_N(\delta) = o_p(1)$ and $h_N(\delta) > c(\delta)$ with some constant $c(\delta) > 0$ independent of N. First notice that

$$\begin{aligned} h_N(\delta) &= \inf_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 = \delta} \frac{1}{2}(\mathbf{u} - \boldsymbol{\zeta}_N)^T \pi_N^{-1} \mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)(\mathbf{u} - \boldsymbol{\zeta}_N) \\ &\geq \frac{1}{2}\delta^2 \pi_N^{-1} \lambda_{\min}\{\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\} \geq \frac{1}{2}\delta^2 \lambda_{\min}\left[E\{\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T \dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\}\right]. \end{aligned}$$

Now, it remains to show $\Delta_N(\delta) = o_p(1)$. A sufficient condition we would like to show is the following:

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \leq \delta} \left|\mathbf{u}^T\{\mathbf{A}_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) - \pi_N^{-1}\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\}\mathbf{u}\right| = o_p(1), \tag{2.74}$$

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \leq \delta} \left|(\mathbf{B}_{N,1}(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) - \mathbf{B}_{N,2})^T\mathbf{u}\right| = o_p(1), \tag{2.75}$$

$$|C_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})| = o_p(1). \tag{2.76}$$

To prove (2.74)-(2.76), we first analyze some basic properties of $\widetilde{\pi}_{N,1}$ and $\boldsymbol{\zeta}_N$. With (2.122), we have

$$\widetilde{\pi}_N = \pi_N\left\{1 + O_p\left((N\pi_N)^{-1/2}\right)\right\}, \text{ for any } \widetilde{\pi}_N \text{ lies between } \pi_N \text{ and } \widehat{\pi}_N. \tag{2.77}$$

In addition, by the fact that $\log(u) \leq 1-u$ for all $u > 0$ and (2.77) we have

$$|\log(\widetilde{\pi}_N/\pi_N)| \leq |1-\widetilde{\pi}_N/\pi_N| = O_p\left((N\pi_N)^{-1/2}\right), \text{ for any } \widetilde{\pi}_N \text{ lies between } \pi_N \text{ and } \widehat{\pi}_N. \tag{2.78}$$

For any $t < \infty$ and $r < \infty$,

$$\begin{aligned} E\left\{\exp(t\|\overrightarrow{\mathbf{X}}\|_2)\|\overrightarrow{\mathbf{X}}\|_2^r\right\} &\leq r!E\left[\exp\{(t+1)\|\overrightarrow{\mathbf{X}}\|_2\}\right] \\ &\leq r!\exp(t+1)E\left[\exp\{(t+1)\|\mathbf{X}\|_2\}\right] < \infty. \end{aligned} \tag{2.79}$$

Now, to control the supremum over $\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \leq \delta$ in (2.74) and (2.75), we analyse asymptotic properties for $\boldsymbol{\zeta}_N$ defined in (2.72). We consider the following representation:

$$\boldsymbol{\zeta}_N = \boldsymbol{\zeta}_{N,1} - \boldsymbol{\zeta}_{N,2}, \quad \text{where} \tag{2.80}$$

$$\boldsymbol{\zeta}_{N,1} = (N_{-k}\pi_N)^{1/2}\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\left\{R_i - g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0+\log(\pi_N))\right\}\overrightarrow{\mathbf{X}}_i, \tag{2.81}$$

$$\boldsymbol{\zeta}_{N,2} = \log(\widehat{\pi}_N/\pi_N)(N_{-k}\pi_N)^{1/2}\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0+\log(\pi_N))\overrightarrow{\mathbf{X}}_i. \tag{2.82}$$

Moreover, define

$$\begin{aligned} \boldsymbol{\zeta}_{N,3} &= \log(\widehat{\pi}_N/\pi_N)(N_{-k}\pi_N)^{1/2}\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)E\left\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\overrightarrow{\mathbf{X}}\right\} \\ &= (N_{-k}\pi_N)^{1/2}\log(\widehat{\pi}_N/\pi_N)\mathbf{e}_1. \end{aligned} \tag{2.83}$$

In the above we used $\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)E\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\overrightarrow{\mathbf{X}}\} = \mathbf{e}_1$. Note that,

$$\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N)) = \frac{\pi_N\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)}{\{1+\pi_N\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\}^2} \geq \frac{\pi_N\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)}{\{1+\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\}^2} = \pi_N\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0).$$

Hence,

$$\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2 \leq \pi_N^{-1}\left\|\left[E\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}^T\}\right]^{-1}\right\|_2 = O(\pi_N^{-1}). \tag{2.84}$$

Then,

$$
\begin{aligned}
E_{\mathbb{S}_{-k}}\|\boldsymbol{\zeta}_{N,1}\|_2^2 &\overset{(i)}{=} \pi_N E\left\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\overrightarrow{\mathbf{X}}\|_2^2\right\}\\
&\overset{(ii)}{\leq} \pi_N\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2^2 E\left\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\|\overrightarrow{\mathbf{X}}\|_2^2\right\}\\
&\overset{(iii)}{\leq} \pi_N^2\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2^2 E\left\{\exp(\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)\|\overrightarrow{\mathbf{X}}\|_2^2\right\} \overset{(iv)}{=} O(1),
\end{aligned}
$$

where (i) holds by the tower rule with the fact $E[\{R-g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\}^2|\mathbf{X}] = \dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))$, (ii) holds by the fact that $|\boldsymbol{A}\boldsymbol{a}|\leq\|\boldsymbol{A}\|_2\|\boldsymbol{a}\|_2$ for any $\boldsymbol{a}\in\mathbb{R}^{p+1}$ and $\boldsymbol{A}\in\mathbb{R}^{(p+1)\times(p+1)}$, (iii) follows by the fact that $\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\leq g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\leq\pi_N\exp(\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)$, and (iv) holds holds by (2.79) and (2.84). Besides,

$$
\begin{aligned}
&E_{\mathbb{S}_{-k}}\left\|\{\log(\widehat{\pi}_N/\pi_N)\}^{-1}(\boldsymbol{\zeta}_{N,2}-\boldsymbol{\zeta}_{N,3})\right\|_2^2 = \pi_N\mathrm{Var}\left\{\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\overrightarrow{\mathbf{X}}\|_2\right\}\\
&\leq \pi_N E\left\{\dot{g}^2(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\overrightarrow{\mathbf{X}}\|_2^2\right\}\\
&\overset{(i)}{\leq} \pi_N\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2^2 E\left\{\dot{g}^2(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\|\overrightarrow{\mathbf{X}}\|_2^2\right\}\\
&\overset{(ii)}{\leq} \pi_N^3\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2^2 E\left\{\exp(2\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)\|\overrightarrow{\mathbf{X}}\|_2^2\right\} \overset{(iii)}{=} O(\pi_N),
\end{aligned}
$$

where (i) holds by the fact that $|\boldsymbol{A}\boldsymbol{a}|\leq\|\boldsymbol{A}\|_2\|\boldsymbol{a}\|_2$ for any $\boldsymbol{a}\in\mathbb{R}^{p+1}$ and $\boldsymbol{A}\in\mathbb{R}^{(p+1)\times(p+1)}$, (ii) follows by the fact that $\dot{g}(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_N))\leq\pi_N\exp(\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)$, and (iii) holds holds by (2.79) and (2.84). By Chebyshev's Inequality,

$$
\|\boldsymbol{\zeta}_{N,1}\| = O_p(1), \qquad \left\|\{\log(\widehat{\pi}_N/\pi_N)\}^{-1}(\boldsymbol{\zeta}_{N,2}-\boldsymbol{\zeta}_{N,3})\right\|_2 = O_p(\pi_N^{1/2}).
$$

Hence, by (2.78),

$$
\|(\boldsymbol{\zeta}_{N,2}-\boldsymbol{\zeta}_{N,3})\|_2 = \{\log(\widehat{\pi}_N/\pi_N)\}O_p(\pi_N^{1/2}) = O_p(N^{-1/2}), \tag{2.85}
$$

with

$$
\|\boldsymbol{\zeta}_{N,3}\|_2 \leq |\log(\widehat{\pi}_N/\pi_N)|(N_{-k}\pi_N)^{1/2}\|\mathcal{J}^{-1}(\boldsymbol{\gamma}_0,\pi_N)\|_2\pi_N E\left\{\exp(\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)\|\overrightarrow{\mathbf{X}}\|_2\right\} = O_p(1).
$$

Therefore,

$$\|\boldsymbol{\zeta}_N\|_2 \le \|\boldsymbol{\zeta}_{N,1}\|_2 + \|\boldsymbol{\zeta}_{N,2} - \boldsymbol{\zeta}_{N,3}\|_2 + \|\boldsymbol{\zeta}_{N,3}\|_2 = O_p(1).$$

It follows that,

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} \|\mathbf{u}\|_2 \le \sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} \|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 + \|\boldsymbol{\zeta}_N\|_2 \le \delta + \|\boldsymbol{\zeta}_N\|_2 = O_p(1), \tag{2.86}$$

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} \|\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0\|_2 \le \sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} (N_{-k}\pi_N)^{-1/2}\|\mathbf{u}\|_2 = O_p\left((N_{-k}\pi_N)^{-1/2}\right), \tag{2.87}$$

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} \|\tilde{\boldsymbol{\gamma}}_1\|_2 \le \sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta} \|\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0\|_2 + \|\boldsymbol{\gamma}_0\|_2 < M, \text{ w.p.a. 1}, \tag{2.88}$$

where $M > 0$ is a constant independent of N.

Now, we prove (2.74). For any $\mathbf{u}$ satisfying $\|\mathbf{u} - \boldsymbol{\zeta}_N\|_2 \le \delta$,

$$\begin{aligned}
&\left|\mathbf{u}^T\{\mathbf{A}_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1}) - \pi_N^{-1}\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\}\mathbf{u}\right| \\
&\quad\le \left|(N_{-k}\pi_N)^{-1}\|\mathbf{u}\|_2^2 \sum_{i\in\mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T\tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1})) - \dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\|\overrightarrow{\mathbf{X}}_i\|_2^2\right| \\
&\qquad + \|\mathbf{u}\|_2^2 \left\|(N_{-k}\pi_N)^{-1} \sum_{i\in\mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}_i^T - \pi_N^{-1}\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\right\|_2.
\end{aligned}$$

By Taylor's Theorem, with some $(\tilde{\boldsymbol{\gamma}}_2, \widetilde{\pi}_{N,2})$ lies between $(\boldsymbol{\gamma}_0, \pi_N)$ and $(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})$, uniformly on $\|\mathbf{u} - \boldsymbol{\zeta}_N\|_2 \le \delta$,

$$\begin{aligned}
&\left|N_{-k}^{-1} \sum_{i\in\mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T\tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1}))\|\overrightarrow{\mathbf{X}}_i\|_2^2 - N_{-k}^{-1} \sum_{i\in\mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\|\overrightarrow{\mathbf{X}}_i\|_2^2\right| \\
&\quad\overset{(i)}{=} \left|N_{-k}^{-1} \sum_{i\in\mathcal{I}_{-k}} \ddot{g}(\overrightarrow{\mathbf{X}}_i^T\tilde{\boldsymbol{\gamma}}_2 + \log(\widetilde{\pi}_{N,2}))\left\{\overrightarrow{\mathbf{X}}_i^T(\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0) + \log(\widetilde{\pi}_{N,1}/\pi_N)\right\}\|\overrightarrow{\mathbf{X}}_i\|_2^2\right| \\
&\quad\overset{(ii)}{\le} \widetilde{\pi}_{N,2}N_{-k}^{-1} \sum_{i\in\mathcal{I}_{-k}} \exp(\overrightarrow{\mathbf{X}}_i^T\tilde{\boldsymbol{\gamma}}_2)\left|\overrightarrow{\mathbf{X}}_i^T(\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0) + \log(\widetilde{\pi}_{N,1}/\pi_N)\right|\|\overrightarrow{\mathbf{X}}_i\|_2^2 \\
&\quad\overset{(iii)}{\le} \widetilde{\pi}_{N,2}N_{-k}^{-1} \sum_{i\in\mathcal{I}_{-k}} \exp(\|\overrightarrow{\mathbf{X}}_i\|_2 M)\left\{\|\overrightarrow{\mathbf{X}}_i\|_2\|\tilde{\gamma}_1 - \boldsymbol{\gamma}_0\|_2 + |\log(\widetilde{\pi}_{N,1}/\pi_N)|\right\}\|\overrightarrow{\mathbf{X}}_i\|_2^2,
\end{aligned} \tag{2.89}$$

with probability approaching 1. Here, (i) holds by Taylor's Theorem, (ii) holds by (2.71), (iii) holds by (2.88). Recall (2.79), by Markov's Inequality,

$$N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} \exp(\|\overrightarrow{\mathbf{X}}_i\|_2 M) \|\overrightarrow{\mathbf{X}}_i\|_2^r = O_p(1). \tag{2.90}$$

Hence,

$$\begin{aligned}
&\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \le \delta} (N_{-k}\pi_N)^{-1} \|\mathbf{u}\|_2^2 \sum_{i \in \mathcal{I}_{-k}} \left| \dot{g}(\overrightarrow{\mathbf{X}}_i^T \tilde{\boldsymbol{\gamma}}_1 + \log(\widetilde{\pi}_{N,1})) - \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \right| \|\overrightarrow{\mathbf{X}}_i\|_2^2 \\
&\overset{(i)}{\le} \sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2 \le \delta} \pi_N^{-1} \|\mathbf{u}\|_2^2 \widetilde{\pi}_{N,2} \left\{ \|\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0\|_2 O_p(1) + |\log(\widetilde{\pi}_{N,1}/\pi_N)| O_p(1) \right\} \\
&\overset{(ii)}{=} O_p\left((N\pi_N)^{-1/2} \right) = o_p(1).
\end{aligned} \tag{2.91}$$

where (i) holds by (2.89) and (2.90), (ii) holds by (2.77), (2.78), (2.86) and (2.87). Notice that

$$E\{\dot{g}(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \|\overrightarrow{\mathbf{X}}\|_2^2\} \le \pi_N E\{\exp(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) \|\overrightarrow{\mathbf{X}}\|_2^2\}.$$

$$\|\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\|_2 \le \pi_N \|E\{\exp(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) \overrightarrow{\mathbf{X}} \overrightarrow{\mathbf{X}}^T\}\|_2 \le \pi_N E\left\{ \exp(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) \|\overrightarrow{\mathbf{X}}\|_2^2 \right\}.$$

Recall that p is fixed, by Theorem 5.48 of [Ver10], with some constant $C > 0$,

$$\begin{aligned}
&E_{\mathbb{S}_{-k}} \left\| N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i^T - \mathcal{J}(\boldsymbol{\gamma}_0, \pi_N) \right\|_2 \\
&\le \max\left[\|\mathcal{J}(\boldsymbol{\gamma}_0, \pi_N)\|_2^{1/2} C \sqrt{\frac{\pi_N \log\{\min(N, p+1)\}}{N}}, \frac{C^2 \pi_N \log\{\min(N, p+1)\}}{N} \right] \\
&= O\left(\max\left(N^{-1/2}\pi_N, N^{-1}\pi_N \right) \right) = O\left(N^{-1/2}\pi_N \right).
\end{aligned}$$

By Markov's Inequality,

$$\left\| N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} \dot{g}(\overrightarrow{\mathbf{X}}_i^T \boldsymbol{\gamma}_0 + \log(\pi_N)) \overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i^T - \mathcal{J}(\boldsymbol{\gamma}_0, \pi_N) \right\|_2 = O_p\left(N^{-1/2}\pi_N \right).$$

It follows that

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta}\|\mathbf{u}\|_2^2\left\|(N_{-k}\pi_N)^{-1}\sum_{i\in\mathcal{I}_{-k}}\dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0+\log(\pi_N))\overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}_i^T-\pi_N^{-1}\mathcal{J}(\boldsymbol{\gamma}_0,\pi_N)\right\|_2$$
$$=O_p\left((N\pi_N)^{-1/2}\right)=o_p(1). \tag{2.92}$$

Hence, by (2.91) and (2.92),

$$\sup_{\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta}\left|\mathbf{u}^T\{\mathbf{A}_N-\pi_N^{-1}\mathcal{J}(\boldsymbol{\gamma}_0,\pi_N)\}\mathbf{u}\right|=O_p\left((N\pi_N)^{-1/2}\right)=o_p(1). \tag{2.93}$$

Now, we show (2.75). By Taylor's Theorem, where some $(\tilde{\boldsymbol{\gamma}}_3,\widetilde{\pi}_{N,3})$ lies between $(\boldsymbol{\gamma}_0,\pi_N)$ and $(\tilde{\boldsymbol{\gamma}}_1,\widetilde{\pi}_{N,1})$,

$$\begin{aligned}
&|(\mathbf{B}_{N,1}-\mathbf{B}_{N,2})^T\mathbf{u}|\\
&\overset{(i)}{=}\left|\log\left(\frac{\widehat{\pi}_N}{\pi_N}\right)(N_{-k}\pi_N)^{-1/2}\sum_{i\in\mathcal{I}_{-k}}\ddot{g}(\overrightarrow{\mathbf{X}}_i^T\tilde{\boldsymbol{\gamma}}_3+\log(\widetilde{\pi}_{N,3}))\left\{\overrightarrow{\mathbf{X}}_i^T(\tilde{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_0)+\log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right)\right\}\overrightarrow{\mathbf{X}}_i^T\mathbf{u}\right|\\
&\overset{(ii)}{\le}\left|\log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right)\right|(N_{-k}\pi_N)^{-1/2}\widetilde{\pi}_{N,3}\sum_{i\in\mathcal{I}_{-k}}\exp(\|\overrightarrow{\mathbf{X}}_i\|_2M)\|\overrightarrow{\mathbf{X}}_i\|_2^2\|\tilde{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_0\|_2\|\mathbf{u}\|_2\\
&\quad+\left|\log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right)\right|(N_{-k}\pi_N)^{-1/2}\widetilde{\pi}_{N,3}\sum_{i\in\mathcal{I}_{-k}}\exp(\|\overrightarrow{\mathbf{X}}_i\|_2M)\left|\log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right)\right|\|\overrightarrow{\mathbf{X}}_i\|_2\|\mathbf{u}\|_2\ \text{ w.p.a. 1}\\
&\overset{(iii)}{\le}\left|\log\left(\frac{\widehat{\pi}_N}{\pi_N}\right)\right|N_{-k}^{1/2}\pi_N^{-1/2}\|\mathbf{u}\|_2\widetilde{\pi}_{N,3}\left\{\|\tilde{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_0\|_2O_p(1)+\left|\log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right)\right|O_p(1)\right\}\ \text{ w.p.a. 1}\\
&\overset{(iv)}{=}O_p\left((N\pi_N)^{-1/2}\right)=o_p(1),\qquad\text{uniformly on }\|\mathbf{u}-\boldsymbol{\zeta}_N\|_2\le\delta,
\end{aligned}\tag{2.94}$$

where (i) holds by Taylor's Theorem, (ii) holds by (2.71) and (2.88), (iii) holds by (2.90), (iv) holds by (2.77), (2.78) (2.86) and (2.87).

As for (2.76), by Taylor's Theorem, with some $(\tilde{\boldsymbol{\gamma}}_4,\widetilde{\pi}_{N,4})$ lies between $(\boldsymbol{\gamma}_0,\pi_N)$ and

$(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})$,

$$
\begin{aligned}
&|C_N(\tilde{\boldsymbol{\gamma}}_1, \widetilde{\pi}_{N,1})| \qquad (2.95)\\
&\overset{(i)}{=} \left| \frac{1}{2} \sum_{i \in \mathcal{I}_{-k}} \ddot{g}(\overrightarrow{\mathbf{X}}_i^T \tilde{\boldsymbol{\gamma}}_4 + \log(\widetilde{\pi}_{N,4})) \left\{ \overrightarrow{\mathbf{X}}^T (\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0) + \log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right) \right\} \left\{ \log\left(\frac{\widehat{\pi}_N}{\pi_N}\right) \right\}^2 \right| \\
&\overset{(ii)}{\leq} \frac{1}{2} \widetilde{\pi}_{N,4} \sum_{i \in \mathcal{I}_{-k}} \exp(\|\overrightarrow{\mathbf{X}}_i\| M) \left\{ \|\overrightarrow{\mathbf{X}}\|_2 \|\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0\|_2 + \left| \log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right) \right| \right\} \left\{ \log\left(\frac{\widehat{\pi}_N}{\pi_N}\right) \right\}^2 \quad \text{w.p.a. 1} \\
&\overset{(iii)}{\leq} \frac{1}{2} \left| \log\left(\frac{\widehat{\pi}_N}{\pi_N}\right) \right|^2 \widetilde{\pi}_{N,4} N_{-k} \left\{ \|\tilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_0\|_2 O_p(1) + \left| \log\left(\frac{\widetilde{\pi}_{N,1}}{\pi_N}\right) \right| O_p(1) \right\} \quad \text{w.p.a. 1} \\
&\overset{(iv)}{=} O_p\left((N\pi_N)^{-1/2}\right) = o_p(1). \qquad (2.96)
\end{aligned}
$$

where (i) holds by Taylor's Theorem, (ii) holds by (2.71) and (2.88), (iii) holds by (2.90), (iv) holds by (2.77), (2.78) (2.86) and (2.87).

Combining (2.93), (2.94) and (2.96), we have

$$\Delta_N(\delta) = o_p(1), \qquad \text{for any } \delta > 0,$$

and hence (2.73) holds. Recall the definition of $\boldsymbol{\zeta}_{N,3}$ in (2.83), we have

$$
\begin{aligned}
&\left\| \boldsymbol{\zeta}_{N,3} - N_{-k}^{-1/2} \pi_N^{1/2} \sum_{i \in \mathcal{I}_{-k}} (\pi_N^{-1} R_i - 1) \mathbf{e}_1 \right\|_2 = \left\| \boldsymbol{\zeta}_{N,3} - (N_{-k}\pi_N)^{1/2} \frac{\widehat{\pi}_N - \pi_N}{\pi_N} \mathbf{e}_1 \right\|_2 \\
&\quad \overset{(i)}{=} \left\| (N_{-k}\pi_N)^{1/2} \frac{(\widehat{\pi}_N - \pi_N)^2}{\widetilde{\pi}_{N,5}^2} \mathbf{e}_1 \right\|_2 = (N_{-k}\pi_N)^{1/2} \frac{(\widehat{\pi}_N - \pi_N)^2}{\widetilde{\pi}_{N,5}^2} = O_p\left((N\pi_N)^{-1/2}\right) = o_p(1).
\end{aligned}
$$

where (i) follows from the Taylor's Theorem with some $\widetilde{\pi}_{N,5}$ lying between π_N and $\widehat{\pi}_N$.

Hence, with $\mathrm{IF}_{\boldsymbol{\gamma}}(\mathbf{Z}) = \mathcal{J}^{-1}(\boldsymbol{\gamma}_0, \pi_N)\{R - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_N))\}\overrightarrow{\mathbf{X}} - (\pi_N^{-1}R - 1)\mathbf{e}_1$,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 - N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_{\boldsymbol{\gamma}}(\mathbf{Z}_i)\right\|_2 &= (N_{-k}\pi_N)^{-1/2}\left\|\boldsymbol{\zeta}_{\boldsymbol{N}} - N_{-k}^{-1/2}\pi_N^{1/2}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_{\boldsymbol{\gamma}}(\mathbf{Z}_i)\right\|_2 \\
&= (N_{-k}\pi_N)^{-1/2}\left\|N_{-k}^{-1/2}\pi_N^{1/2}\sum_{i\in\mathcal{I}_{-k}}(\pi_N^{-1}R_i - 1)\mathbf{e}_1 - \boldsymbol{\zeta}_{N,3} + (\boldsymbol{\zeta}_{N,3} - \boldsymbol{\zeta}_{N,2})\right\|_2 \\
&\leq (N_{-k}\pi_N)^{-1/2}\left\|N_{-k}^{-1/2}\pi_N^{1/2}\sum_{i\in\mathcal{I}_{-k}}(\pi_N^{-1}R_i - 1)\mathbf{e}_1 - \boldsymbol{\zeta}_{N,3}\right\|_2 + (N_{-k}\pi_N)^{-1/2}\left\|\boldsymbol{\zeta}_{N,3} - \boldsymbol{\zeta}_{N,2}\right\|_2 \\
&= (N_{-k}\pi_N)^{-1/2}O_p\left((N\pi_N)^{-1/2} + N^{-1/2}\right) = o_p\left((N_{-k}\pi_N)^{-1/2}\right).
\end{aligned}$$

Now, it remains to analyze the IF of the PS $\widehat{\pi}_N(\mathbf{X}) = g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\gamma}} + \log(\widehat{\pi}_N))$. For this, define

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\gamma}} + \log(\widehat{\pi}_N)\mathbf{e}_1, \quad \boldsymbol{\beta}_0 = \boldsymbol{\gamma}_0 + \log(\pi_N)\mathbf{e}_1, \tag{2.97}$$

$$\mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}) = \mathcal{J}^{-1}(\boldsymbol{\gamma}_0, \pi_N)\{R - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_N))\}\overrightarrow{\mathbf{X}}. \tag{2.98}$$

Then,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}_i)\right\|_2 &\overset{(i)}{=} \left\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 + \log(\widehat{\pi}_N/\pi_N)\mathbf{e}_1 - (N_{-k}\pi_N)^{-1/2}\boldsymbol{\zeta}_{N,1}\right\|_2 \\
&\overset{(ii)}{=} \left\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 - (N_{-k}\pi_N)^{-1/2}\boldsymbol{\zeta}_N + (N_{-k}\pi_N)^{-1/2}(\boldsymbol{\zeta}_{N,3} - \boldsymbol{\zeta}_{N,1} - \boldsymbol{\zeta}_N)\right\|_2 \\
&\overset{(iii)}{\leq} \left\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 - (N_{-k}\pi_N)^{-1/2}\boldsymbol{\zeta}_N\right\|_2 + (N_{-k}\pi_N)^{-1/2}\left\|\boldsymbol{\zeta}_{N,3} - \boldsymbol{\zeta}_{N,2}\right\|_2 \\
&\overset{(iv)}{=} o_p\left((N\pi_N)^{-1/2}\right) + (N_{-k}\pi_N)^{-1/2}O_p(N^{-1/2}) = o_p\left((N\pi_N)^{-1/2}\right),
\end{aligned} \tag{2.99}$$

where (i) holds by (2.81), (4.13) and (2.98), (ii) holds by (2.83), (iii) holds by (2.80) and the triangular inequality, (iv) holds by (2.73) and (2.85). It follows that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_2 &\overset{(i)}{\leq} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}_i)\right\|_2 + (N_{-k}\pi_N)^{-1/2}\left\|\boldsymbol{\zeta}_{N,1}\right\|_2 \\
&\overset{(ii)}{=} o_p\left((N\pi_N)^{-1/2}\right) + O_p\left((N\pi_N)^{-1/2}\right) = O_p\left((N\pi_N)^{-1/2}\right),
\end{aligned}$$

where (i) holds by (2.81) and the triangular inequality, (ii) holds by (2.99) and (2.85). Furthermore,

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2 &\le \left\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0-(N_{-k}\pi_N)^{-1/2}\boldsymbol{\zeta}_N\right\|_2+(N_{-k}\pi_N)^{-1/2}\|\boldsymbol{\zeta}_N\|_2\\ &= o_p\left((N\pi_N)^{-1/2}\right)+O_p\left((N\pi_N)^{-1/2}\right)=O_p\left((N\pi_N)^{-1/2}\right)=o_p(1),\end{aligned}$$

and hence $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2 < 1$ w.p.a. 1. By Taylor's Theorem, for any $\mathbf{x}\in\mathcal{X}$, where some $(\tilde{\boldsymbol{\gamma}}_6,\widetilde{\pi}_{N,6})$ (depending on $\mathbf{x}$) lies between $(\boldsymbol{\gamma}_0,\pi_N)$ and $(\widehat{\boldsymbol{\gamma}},\widehat{\pi}_N)$ and $\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{\gamma}}_6+\log(\widetilde{\pi}_{N,6})\mathbf{e}_1$,

$$\begin{aligned}&1-\frac{g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0+\log(\pi_N))}{g(\overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}}+\log(\widehat{\pi}_N))}-\left\{1-g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\\ &\quad= g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\left\{g^{-1}(\overrightarrow{\mathbf{x}}^T\tilde{\boldsymbol{\beta}})-1\right\}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\right\}^2\\ &\quad= g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\exp(-\overrightarrow{\mathbf{x}}^T\tilde{\boldsymbol{\beta}})\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\right\}^2\\ &\quad= \widetilde{\pi}_{N,6}^{-1}g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\exp(-\overrightarrow{\mathbf{x}}^T\tilde{\boldsymbol{\gamma}})\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\right\}^2\\ &\quad\le \max(\pi_N^{-1},\widehat{\pi}_N^{-1})g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\exp\left\{\|\overrightarrow{\mathbf{x}}\|_2(\|\boldsymbol{\gamma}_0\|_2+\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2)\right\}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\right\}^2\\ &\quad\le \max(\pi_N^{-1},\widehat{\pi}_N^{-1})g(\overrightarrow{\mathbf{x}}^T\boldsymbol{\beta}_0)\exp\left\{\|\overrightarrow{\mathbf{x}}\|_2(\|\boldsymbol{\gamma}_0\|_2+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2)\right\}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{x}}\right\}^2.\end{aligned}$$

Therefore, for any fixed $r>0$, with $\mathbf{X}$ independent of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\pi}_N$,

$$\begin{aligned}&\left\|1-\frac{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\beta}})}-\left\{1-g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{X}}\right\|_{r,P}\\ &\quad\le \max(\pi_N^{-1},\widehat{\pi}_N^{-1})\left\|g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\exp\left\{\|\overrightarrow{\mathbf{X}}\|_2(\|\boldsymbol{\gamma}_0\|_2+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2)\right\}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{X}}\right\}^2\right\|_{r,P}\\ &\quad\le \max(\pi_N^{-1},\widehat{\pi}_N^{-1})\left\|g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\exp\left\{\|\overrightarrow{\mathbf{X}}\|_2(\|\boldsymbol{\gamma}_0\|_2+1)\right\}\left\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{X}}\right\}^2\right\|_{r,P}\ \text{w.p.a. 1}\\ &\quad\le \max(1,\pi_N\widehat{\pi}_N^{-1})\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2\left\|\exp\left\{\|\overrightarrow{\mathbf{X}}\|_2(2\|\boldsymbol{\gamma}_0\|_2+1)\right\}\|\overrightarrow{\mathbf{X}}\|_2^2\right\|_{r,P}\ \text{w.p.a. 1}\\ &\quad= \{1+o_p(1)\}O_p\left((N\pi_N)^{-1}\right)O(1)=O_p\left((N\pi_N)^{-1}\right)=o_p\left((N\pi_N)^{-1/2}\right).\end{aligned}\tag{2.100}$$

Define

$$\mathrm{IF}_\pi(\mathbf{Z}; S_{-k}) = \left\{1 - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\right\} \overrightarrow{\mathbf{X}}^T N_{-k}^{-1} \sum_{i\in\mathcal{I}_k} \mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}_i), \tag{2.101}$$

where $\mathbf{Z}$ is independent of $(\mathbf{Z}_i)_{i\in\mathcal{I}_k}$. Then,

$$\begin{aligned}
&\left\|\left\{1 - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\right\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{X}} - \mathrm{IF}_\pi(\mathbf{Z}; S_{-k})\right\|_{r,P} \\
&\quad\le \left\|\|\overrightarrow{\mathbf{X}}\|_2\right\|_{r,P} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}} \mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}_i)\right\|_2 = o_p\left((N\pi_N)^{-1/2}\right),
\end{aligned} \tag{2.102}$$

and

$$\left\|\left\{1 - g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)\right\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\overrightarrow{\mathbf{X}}\right\|_{r,P} \le \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \left\|\|\overrightarrow{\mathbf{X}}\|_2\right\|_{r,P} = O_p\left((N\pi_N)^{-1/2}\right).$$

Hence,

$$\left\|1 - \frac{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\beta}})}\right\|_{r,P} = O_p\left((N\pi_N)^{-1/2}\right).$$

For any fixed $r > 0$,

$$\left\|\pi_N^{-1}(\mathbf{X})\right\|_{r,P} = \left\|1 + \pi_N^{-1}\exp(-\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\right\|_{r,P} \le 1 + \pi_N^{-1}\left\|\exp(\|\overrightarrow{\mathbf{X}}\|_2\|\boldsymbol{\gamma}_0\|_2)\right\|_{r,P} = O(\pi_N^{-1}).$$

Additionally, by Jensen's Inequality,

$$\left\|\pi_N^{-1}(\mathbf{X})\right\|_{r,P} = \left[E\{\pi_N^{-r}(\mathbf{X})\}\right]^{1/r} \ge E\{\pi_N^{-1}(\mathbf{X})\} = \pi_N^{-1}, \tag{2.103}$$

and hence $\left\|\pi_N^{-1}(\mathbf{X})\right\|_{r,P} \asymp \pi_N^{-1}$, which implies that $a_N \asymp \pi_N$. It follows that, with $r, s > 0$ satisfying $1/r + 1/s = 1$ and $2s = 2 + c$,

$$E\left[\frac{a_N}{\pi_N(\mathbf{X})}\{\widehat{m}(\mathbf{X}) - \mu(\mathbf{X})\}^2\right] \le a_N\left\|\pi_N^{-1}(\mathbf{X})\right\|_{r,P}\|\widehat{m}(\cdot) - \mu(\cdot)\|_{2s,P}^2 = o_p(1), \tag{2.104}$$

$$\begin{aligned}
E\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2\right] &\le a_N\left\|\pi_N^{-1}(\mathbf{X})\right\|_{r,P}\left\|1 - \frac{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\beta}})}\right\|_{2s,P}^2 \\
&= O_p\left((N\pi_N)^{-1}\right) = o_p(1),
\end{aligned} \tag{2.105}$$

where (2.104) requires an additional assumption that $\|\hat{m}(\cdot) - \mu(\cdot)\|_{2+c,P} = o_p(1)$.

Now, we analyze $\widehat{\theta}_{\text{DRSS}} - \theta_0$, where we further assume that $\|m(\cdot) - \mu(\cdot)\|_{2+c,P} < \infty$. Applying part (b) of Theorem 2.2, we have

$$(\widehat{\theta}_{\text{DRSS}} - \theta_0) = \frac{1}{N}\sum_{i=1}^{N} \psi_\mu(\mathbf{Z}_i) + o_p\left((N\pi_N)^{-1/2}\right) + \widehat{\Delta}_N,$$

where $\psi_{\mu,e}(\mathbf{Z}) = \mu(\mathbf{X}) - \theta_0 + R/\pi_N(\mathbf{X})\{Y - \mu(\mathbf{X})\}$ and

$$\begin{aligned}
\widehat{\Delta}_N &= N^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\left\{1 - \frac{\pi_N(\mathbf{X}_i)}{\widehat{\pi}_N(\mathbf{X}_i)}\right\}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\} \\
&= N^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}\text{IF}_\pi(\mathbf{Z}; S_{-k}) \\
&\quad + N^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}\left\{1 - \frac{g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}; S_{-k})\right\}.
\end{aligned}$$

For each $k \le \mathbb{K}$,

$$\begin{aligned}
&E_{\mathbb{S}_k}\left|\,|\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}\left\{1 - \frac{g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}; S_{-k})\right\}\right| \\
&\quad \le E\left|\{\mu(\mathbf{X}) - m(\mathbf{X})\}\left\{1 - \frac{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}; S_{-k})\right\}\right| \\
&\quad \le \|\mu(\cdot) - m(\cdot)\|_{2,P}\left\|1 - \frac{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}; S_{-k})\right\|_{2,P} \overset{(i)}{=} o_p\left((N\pi_N)^{-1/2}\right),
\end{aligned}$$

where (i) holds by (2.100) and (2.102). By Lemma 2.5,

$$|\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}\left\{1 - \frac{g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}_i; S_{-k})\right\} = o_p\left((N\pi_N)^{-1/2}\right),$$

and hence

$$N^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}\left\{1 - \frac{g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\beta}_0)}{g(\overrightarrow{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}})} - \text{IF}_\pi(\mathbf{Z}_i; S_{-k})\right\} = o_p\left((N\pi_N)^{-1/2}\right).$$

Besides, for each $k \leq \mathbb{K}$, with $r, s > 0$ satisfying $1/r + 1/s = 1$ and $2s = 2 + c$, and recalling the definition of $\mathrm{IF}_\pi(\mathbf{Z}; S_{-k})$ in (2.101), we have

$$\begin{aligned}
&\mathrm{Var}_{\mathbb{S}_k}\left[|\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k}\frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}\mathrm{IF}_\pi(\mathbf{Z}_i;S_{-k})\right]\\
&\quad\overset{(i)}{\leq}|\mathcal{I}_k|^{-1}\left\|E\left[\pi_N^{-1}(\mathbf{X})\{\mu(\mathbf{X})-m(\mathbf{X})\}^2\|\overrightarrow{\mathbf{X}}\|_2\right]\right\|_2\left\|N_{-k}^{-1}\sum_{j\in\mathcal{I}_{-k}}\mathrm{IF}_{\boldsymbol{\beta}}(\mathbf{Z}_j)\right\|_2^2\\
&\quad\leq|\mathcal{I}_k|^{-1}\|\pi_N^{-1}(\mathbf{X})\|_{2r,P}\|\mu(\mathbf{X})-m(\mathbf{X})\|_{2s,P}^2\left\|\|\overrightarrow{\mathbf{X}}\|_2\right\|_{2r,P}(N_{-k}\pi_N)^{-1}\left\|\boldsymbol{\zeta}_{N,1}\right\|_2^2\\
&=O\left((N\pi_N)^{-2}\right).
\end{aligned}$$

By Lemma 2.4 and recalling the definition (2.101),

$$\begin{aligned}
&|\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k}\frac{R_i}{\pi_N(\mathbf{X}_i)}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}\mathrm{IF}_\pi(\mathbf{Z}_i;S_{-k})\\
&\quad=E_{\mathbf{X}}\left[\{\mu(\mathbf{X})-m(\mathbf{X})\}\mathrm{IF}_\pi(\mathbf{Z};S_{-k})\right]+O_p\left((N\pi_N)^{-1}\right)\\
&\quad=N_{-k}^{-1}\sum_{j\in\mathcal{I}_{-k}}\mathrm{IF}_\pi(\mathbf{Z}_j)+O_p\left((N\pi_N)^{-1}\right)=N_{-k}^{-1}\sum_{j\in\mathcal{I}_{-k}}\mathrm{IF}_\pi(\mathbf{Z}_j)+o_p\left((N\pi_N)^{-1/2}\right),
\end{aligned}$$

where $\mathrm{IF}_\pi(\mathbf{Z}) = E\left[\{1-\pi_N(\mathbf{X})\}\{\mu(\mathbf{X})-m(\mathbf{X})\}\overrightarrow{\mathbf{X}}^T\right]J^{-1}(\pi_N,\boldsymbol{\gamma}_0)\overrightarrow{\mathbf{X}}\{R-\pi_N(\mathbf{X})\}$. Therefore,

$$\widehat{\Delta}_N=\mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}N_{-k}^{-1}\sum_{j\in\mathcal{I}_{-k}}\mathrm{IF}_\pi(\mathbf{Z}_j)+o_p\left((N\pi_N)^{-1/2}\right)=N^{-1}\sum_{i=1}^{N}\mathrm{IF}_\pi(\mathbf{Z}_i)+o_p\left((N\pi_N)^{-1/2}\right).$$

■

Proof of Lemma 2.1. For any $a \in (0,1]$, we have the corresponding Jacobian (or Hessian) matrices of $\ell_N(\boldsymbol{\gamma};a)$ and $\ell_N(\boldsymbol{\gamma},1)$ w.r.t. $\boldsymbol{\gamma}\in\mathbb{R}^{p+1}$ satisfy the following (analytical) inequality:

$$\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}\{\ell_N(\boldsymbol{\gamma};a)\}&=N^{-1}\sum_{i=1}^{N}\dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}+\log(a))\overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}_i^T\\
&\succeq aN^{-1}\sum_{i=1}^{N}\dot{g}(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma})\overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}_i^T=a\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}\{\ell_N(\boldsymbol{\gamma};1)\},
\end{aligned}$$

since for any $a \in (0,1]$ and $u \in \mathbb{R}$,

$$\dot{g}(u + \log(a)) = \frac{a\exp(u)}{\{1 + a\exp(u)\}^2} \geq \frac{a\exp(u)}{\{1+\exp(u)\}^2} = a\dot{g}(u).$$

Let $\mathcal{G}(\boldsymbol{\gamma}; a) := \ell_N(\boldsymbol{\gamma}; a) - a\ell(\boldsymbol{\gamma}; 1)$. Then,

$$\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}\{\mathcal{G}(\boldsymbol{\gamma}; a)\} \succeq \mathbf{0}, \qquad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^{p+1}.$$

That is, the function $\mathcal{G}(\boldsymbol{\gamma}, a)$ is convex in $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$, and hence by the basic properties of convex functions, we have: for any $\boldsymbol{\gamma}, \boldsymbol{\Delta} \in \mathbb{R}^{p+1}$,

$$\mathcal{G}(\boldsymbol{\gamma} + \boldsymbol{\Delta}; a) - \mathcal{G}(\boldsymbol{\gamma}; a) - \boldsymbol{\Delta}^T\{\nabla_{\boldsymbol{\gamma}}\mathcal{G}(\boldsymbol{\gamma}; a)\} \geq 0$$

and hence

$$\begin{aligned}\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}) &= \ell_N(\boldsymbol{\gamma} + \boldsymbol{\Delta}; a) - \ell_N(\boldsymbol{\gamma}; a) - \boldsymbol{\Delta}^T\{\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}; a)\} \\ &\geq a\left[\ell_N(\boldsymbol{\gamma} + \boldsymbol{\Delta}; 1) - \ell_N(\boldsymbol{\gamma}; 1) - \boldsymbol{\Delta}^T\{\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}; 1)\}\right] = a\left\{\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma})\right\}.\end{aligned}$$

Therefore,

$$\delta\ell(\boldsymbol{\Delta}; a; \boldsymbol{\gamma}) \geq a\kappa\|\boldsymbol{\Delta}\|_2^2, \qquad \forall\, \boldsymbol{\Delta} \in A,$$

if $\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}) \geq \kappa\|\boldsymbol{\Delta}\|_2^2$ for all $\boldsymbol{\Delta} \in A$. ∎

Proof of Lemma 2.2. We first derive the following useful properties: define $\mu_{\boldsymbol{\gamma}_0} = E(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)$,

then by Lemma 2.8 and some calculation, for all $t \in \mathbb{R}$, $j \leq p+1$, and $r \geq 1$,

$$
\begin{aligned}
|\mu_{\boldsymbol{\gamma}_0}| \leq E(|\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0|) \leq \sigma_{\boldsymbol{\gamma}_0}\sqrt{\pi} < 2\sigma_{\boldsymbol{\gamma}_0}, \\
\|\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 - \mu_{\boldsymbol{\gamma}_0}\|_{\psi_2} \leq \|\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0\|_{\psi_2} + \|\mu_{\boldsymbol{\gamma}_0}\|_{\psi_2} \leq \sigma_{\boldsymbol{\gamma}_0} + \{\log(2)\}^{-1/2}|\mu_{\boldsymbol{\gamma}_0}| < 4\sigma_{\boldsymbol{\gamma}_0}, \\
E\{\exp(t\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)\} = \exp(t\mu_{\boldsymbol{\gamma}_0}) E[\exp\{t(\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0 - \mu_{\boldsymbol{\gamma}_0})\}] \leq \exp\{2\sigma_{\boldsymbol{\gamma}_0}|t| + 20\sigma_{\boldsymbol{\gamma}_0}^2 t^2\},
\end{aligned}
\tag{2.106}
$$

$$
\begin{aligned}
\|\overrightarrow{\mathbf{X}}(j) - E\{\overrightarrow{\mathbf{X}}(j)\}\|_{\psi_2} \leq \|\overrightarrow{\mathbf{X}}(j)\|_{\psi,2} + \|E\{\overrightarrow{\mathbf{X}}(j)\}\|_{\psi_2} \leq \sigma + \{\log(2)\}^{-1/2}\sqrt{\pi}\sigma, \\
\max_{1\leq j\leq p+1} E\{|\overrightarrow{\mathbf{X}}(j)|^r\} \leq r! \max_{1\leq j\leq p+1} E[\exp\{|\overrightarrow{\mathbf{X}}(j)|\}] \leq r!\exp(20\sigma^2).
\end{aligned}
$$

Notice that $\||\cdot|\|_{\psi_2}$ is a monotone increasing function leading to $\|X_2\|_{\psi_2} \geq \|X_1\|_{\psi_2}$ if $|X_2| \geq |X_1|$. Hence,

$$
\max_{1\leq j\leq p+1} \|\{R_i - \pi_N(\mathbf{X}_i)\}\overrightarrow{\mathbf{X}}_{ij}\|_{\psi_2} \leq \max_{1\leq j\leq p+1} \|\overrightarrow{\mathbf{X}}_{ij}\|_{\psi_2} \leq \sigma.
$$

In addition,

$$
\begin{aligned}
&\max_{1\leq j\leq p+1} E\left[\{R - \pi_N(\mathbf{X})\}^2 \overrightarrow{\mathbf{X}}^2(j)\right] \\
&\quad = \max_{1\leq j\leq p+1} E\left[\pi_N(\mathbf{X})\{1-\pi_N(\mathbf{X})\}\overrightarrow{\mathbf{X}}^2(j)\right] \leq \max_{1\leq j\leq p+1} \pi_N E\left\{\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\overrightarrow{\mathbf{X}}^2(j)\right\} \\
&\quad \leq \pi_N\left[E\{\exp(2\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\} \max_{1\leq j\leq p+1} E\{\overrightarrow{\mathbf{X}}^4(j)\}\right]^{1/2} \leq 2\exp(2\sigma_{\boldsymbol{\gamma}_0} + 40\sigma_{\boldsymbol{\gamma}_0}^2 + 10\sigma^2)\pi_N.
\end{aligned}
$$

Now, apply Theorem 3.4 of [KC18], for any $t_1 \geq 0$, with probability at least $1 - 3\exp(-t_1)$,

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0; \pi_N)\|_\infty &= \left\| N_{-k} \sum_{i\in\mathcal{I}_{-k}} \{R_i - \pi_N(\mathbf{X}_i)\}\overrightarrow{\mathbf{X}}_i \right\|_\infty \\
&\leq 7\sqrt{\frac{2\exp(2\sigma_{\boldsymbol{\gamma}_0} + 40\sigma_{\boldsymbol{\gamma}_0}^2 + 10\sigma^2)\pi_N\{t_1 + \log(p+1)\}}{N}} \qquad (2.107) \\
&\quad + \frac{c_6\sigma\sqrt{\log(2N)}\{t_1 + \log(p+1)\}}{N}, \qquad (2.108)
\end{aligned}
$$

with some constant c_6 independent of N. Define $\mathcal{B} = \mathcal{B}(t_1)$ to be an event that (2.108) holds, then $P(\mathcal{B}) \geq 1 - 3\exp(-t_1)$.

Now, we consider the error that originated from the first step estimation $\widehat{\pi}_N$. By Taylor's Theorem, for each $i \leq N$, there exists π'_N (depends on i) lies between π_N and $\widehat{\pi}_N$, such that

$$\left|g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\widehat{\pi}_N)) - g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\right| = \frac{|\widehat{\pi}_N - \pi_N|}{\pi'_N}|\phi(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma} + \log(\pi'_N))| \tag{2.109}$$

$$\leq \frac{|\widehat{\pi}_N - \pi_N|}{\pi'_N} g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma} + \log(\pi'_N)) \leq \frac{|\widehat{\pi}_N - \pi_N|}{\min(\pi_N, \widehat{\pi}_N)} g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma} + \log(\max\{\pi_N, \widehat{\pi}_N\})), \tag{2.110}$$

since function $g(\cdot)$ is monotone increasing. Observe that, for each $r \geq 2$,

$$\begin{aligned} E|R - \pi_N|^r &= E\left[|1 - \pi_N|^r \pi_N(\mathbf{X}) + |-\pi_N|^r\{1 - \pi_N(\mathbf{X})\}\right] \\ &= (1-\pi_N)^r\pi_N + \pi_N^r(1-\pi_N) \leq 2\pi_N \leq \frac{r!}{2}1^{r-2}\cdot 2\pi_N. \end{aligned}$$

By Theorem 1 of [vdGL13], for any $t_2 > 0$,

$$P_{\mathbb{S}}\left(|\widehat{\pi}_N - \pi_N| \geq 2\sqrt{\frac{t_2\pi_N}{N}} + \frac{t_2}{N}\right) \leq 2\exp(-t_2).$$

Define event

$$\mathcal{A} = \mathcal{A}(t_2) := \{|\widehat{\pi}_N - \pi_N| < 2\sqrt{t_2\pi_N/N} + t_2/N\}. \tag{2.111}$$

Then, $P_{\mathbb{S}}(\mathcal{A}) \geq 1 - 2\exp(-t_2)$. Define

$$\pi_{N,\min} = \pi_N - 2\sqrt{t_2\pi_N/N} - t_2/N, \quad \pi_{N,\max} = \pi_N + 2\sqrt{t_2\pi_N/N} + t_2/N.$$

Suppose $t_2 < N\pi_N/9$, then $2\sqrt{t_2\pi_N/N} + t_2/N < 7\pi_N/9$, $\pi_{N,\min} > 2\pi_N/9 > 0$ and $\pi_{N,\max} < 16\pi_N/9 < 16/9$. Recall (2.110), on event $\mathcal{A}$, we have for each $i \leq N$,

$$\left|g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\widehat{\pi}_N)) - g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\right| \leq \frac{|\widehat{\pi}_N - \pi_N|}{\pi_{N,\min}} g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max})),$$

and

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N) - \nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\pi_N)\|_\infty \\
&\quad= \left\| N^{-1}\sum_{i=1}^{N}\{g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\widehat{\pi}_N)) - g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_N))\}\overrightarrow{\mathbf{X}}_i \right\|_\infty \\
&\quad\le \frac{\widehat{\pi}_N - \pi_N}{\pi_{N,\min}} \left\| N^{-1}\sum_{i=1}^{N} g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}_i \right\|_\infty \\
&\quad\le \frac{\widehat{\pi}_N - \pi_N}{\pi_{N,\min}} \left\| N^{-1}\sum_{i=1}^{N}\mathbf{V}_i \right\|_\infty + \frac{\widehat{\pi}_N - \pi_N}{\pi_{N,\min}} \left\| E\left\{ g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}} \right\} \right\|_\infty,
\end{aligned}
$$

where

$$
\mathbf{V}_i = g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}_i - E\left\{ g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}} \right\}.
$$

For any vector $\mathbf{v}$, let $\mathbf{v}(j)$ denotes the j-th element of the vector $\mathbf{v}$. Notice that, on event $\mathcal{A}$,

$$
\begin{aligned}
\max_{1\le j\le p+1}\left| E\left\{ g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}(j) \right\} \right| &\le \max_{1\le j\le p+1}\pi_{N,\max}E\{\exp(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)|\overrightarrow{\mathbf{X}}(j)|\} \\
&\le \pi_{N,\max}\left[E\{\exp(2\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\} \max_{1\le j\le p+1} E\{\overrightarrow{\mathbf{X}}^2(j)\} \right]^{1/2} \\
&\le \pi_{N,\max}\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0} + 40\sigma_{\boldsymbol{\gamma}_0}^2 + 10\sigma^2),
\end{aligned}
$$

and hence

$$
\begin{aligned}
\max_{1\le j\le p+1}\|\mathbf{V}(j)\|_{\psi_2} &\le \max_{1\le j\le p+1}\left\| g(\overrightarrow{\mathbf{X}}_i^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}_i(j) \right\|_{\psi_2} \\
&\quad + \max_{1\le j\le p+1}\left\| E\left\{ g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0 + \log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}(j) \right\} \right\|_{\psi_2} \\
&\le \max_{1\le j\le p+1}\left\| \overrightarrow{\mathbf{X}}_i(j) \right\|_{\psi_2} + \pi_{N,\max}\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0} + 40\sigma_{\boldsymbol{\gamma}_0}^2 + 10\sigma^2) \\
&\le \sigma + \frac{16\pi_N}{9}\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0} + 40\sigma_{\boldsymbol{\gamma}_0}^2 + 10\sigma^2).
\end{aligned}
$$

Additionally,

$$\begin{aligned}\max_{1\le j\le p+1} E(\mathbf{V}_i^2) &\le \max_{1\le j\le p+1} E\left\{g^2(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}^2(j)\right\}\\ &\le \pi_{N,\max}^2 \max_{1\le j\le p+1} E\left\{\exp(2\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\overrightarrow{\mathbf{X}}^2(j)\right\}\\ &\le \pi_{N,\max}^2\left[E\{\exp(4\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\}\max_{1\le j\le p+1} E\{\overrightarrow{\mathbf{X}}^4(j)\}\right]^{1/2}\\ &\le \pi_{N,\max}^2 2\exp(4\sigma_{\boldsymbol{\gamma}_0}+160\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2).\end{aligned}$$

Define

$$\mathcal{C}=\left\{\left\|N^{-1}\sum_{i=1}^N\mathbf{V}_i\right\|_\infty \le 7c_8\pi_{N,\max}\sqrt{\frac{t_1+\log(p+1)}{N}}+\frac{c_9\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N}\right\}, \tag{2.112}$$

where $c_8=\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0}+80\sigma_{\boldsymbol{\gamma}_0}^2+5\sigma^2)$, $c_9=c_6\{\sigma+16\pi_N\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2)/9\}$. By Theorem 3.4 of [KC18], $P(\mathcal{C})\ge 1-3\exp(-t_1)$. It follows that, on events $\mathcal{A}$ and $\mathcal{C}$,

$$\begin{aligned}&\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)-\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\pi_N)\|_\infty\\ &\quad\le \frac{|\widehat{\pi}_N-\pi_N|}{\pi_{N,\min}}\left\|N\sum_{i=1}^N\mathbf{V}_i\right\|_\infty+\frac{|\widehat{\pi}_N-\pi_N|}{\pi_{N,\min}}\left\|E\left\{g(\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0+\log(\pi_{N,\max}))\overrightarrow{\mathbf{X}}\right\}\right\|_\infty\\ &\quad\le \frac{2\sqrt{t_2\pi_N/N}+t_2/N}{\pi_{N,\min}}\left\{7c_8\pi_{N,\max}\sqrt{\frac{t_1+\log(p+1)}{N}}+\frac{c_9\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N}\right\}\\ &\qquad+\frac{2\sqrt{t_2\pi_N/N}+t_2/N}{\pi_{N,\min}}\pi_{N,\max}\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2).\end{aligned}$$

Recall that, when $t_2<N\pi_N/9$,

$$\frac{2\sqrt{t_2\pi_N/N}+t_2/N}{\pi_{N,\min}}<\frac{7}{2},\qquad \pi_{N,\min}>\frac{2}{9}\pi_N,\qquad \pi_{N,\max}<\frac{16}{9}\pi_N.$$

Hence, when $t_2 < N\pi_N/9$, on events $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$,

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)\|_\infty \le \|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\pi_N)\|_\infty + \|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N) - \nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\pi_N)\|_\infty \\
&\le 7\sqrt{\frac{2\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2)\pi_N\{t_1+\log(p+1)\}}{N}} + \frac{c_6\sigma\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N} \\
&\quad + \frac{2\sqrt{t_2\pi_N/N}+t_2/N}{\pi_{N,\min}}\left\{7c_8\pi_{N,\max}\sqrt{\frac{t_1+\log(p+1)}{N}} + \frac{c_9\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N}\right\} \\
&\quad + \frac{2\sqrt{t_2\pi_N/N}+t_2/N}{\pi_{N,\min}}\pi_{N,\max}\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2) \\
&\le C_1(\pi_N+\pi_N^{1/2})\sqrt{\frac{\{t_1+\log(p+1)\}}{N}} + (C_2+C_3\pi_N)\frac{\sqrt{\log(2N)}\{t_1+\log(p+1)\}}{N} \\
&\quad + C_4\left\{\sqrt{\frac{t_2\pi_N}{N}}+\frac{t_2}{N}\right\}.
\end{aligned}
$$

where $P_{\mathbb{S}}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}) \ge 1-6\exp(-t_1)-2\exp(-t_2)$,

$$C_1 = 62\exp(2\sigma_{\boldsymbol{\gamma}_0}+80\sigma_{\boldsymbol{\gamma}_0}^2+5\sigma^2), \qquad C_2 = \frac{9}{2}c_6\sigma, \tag{2.113}$$

$$C_3 = \frac{56}{9}\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2), \qquad C_4 = 16\sqrt{2}\exp(2\sigma_{\boldsymbol{\gamma}_0}+40\sigma_{\boldsymbol{\gamma}_0}^2+10\sigma^2). \tag{2.114}$$

■

Proof of Theorem 2.5. Here, we establish a non-asymptotic property of the offset logistic regression estimator based on the full sample $\mathbf{S}$. The result follows from the Lemmas 2.1 and 2.2, where we obtained the RSC property and controlled the gradient $\|\nabla_{\boldsymbol{\gamma}}\ell_N(\boldsymbol{\gamma}_0;\widehat{\pi}_N)\|_\infty$, respectively. After that, we validate the conditions required in Theorem 2.2 for the proposed offset logistic PS estimator.

For any $t \in \mathbb{R}$, set $t_1 = t_2 = t\log(p+1)$. By Lemma 2.2, on events $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, with

$P_{\mathbb{S}}(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}) \geq 1 - 8(p+1)^{-t}$,

$$
\begin{aligned}
&\left\|\nabla_{\boldsymbol{\gamma}} \ell_N(\boldsymbol{\gamma}_0; \widehat{\pi}_N)\right\|_\infty \\
&\leq (t+1)\left\{C_1(\pi_N + \pi_N^{1/2})\sqrt{\frac{\log(p+1)}{N}} + (C_2 + C_3\pi_N)\frac{\sqrt{\log(2N)}\log(p+1)}{N}\right\} \\
&\quad + C_4 t^{1/2}\pi_N^{1/2}\sqrt{\frac{\log(p+1)}{N}} + C_4 t\frac{\log(p+1)}{N} \\
&\leq (t+1)\left\{(C_1 + C_4)(\pi_N + \pi_N^{1/2})\sqrt{\frac{\log(p+1)}{N}} + (C_2 + C_4 + C_3\pi_N)\frac{\sqrt{\log(2N)}\log(p+1)}{N}\right\} \\
&\leq (t+1)M_N,
\end{aligned}
$$

where

$$
M_N = 2(C_1 + C_4)\pi_N^{1/2}\sqrt{\frac{\log(p+1)}{N}} + (C_2 + C_3 + C_4)\frac{\sqrt{\log(2N)}\log(p+1)}{N}.
$$

Hence, for any $\lambda_N \geq 2(1+c)M_N$ with constant $c > 0$,

$$
2\left\|\nabla_{\boldsymbol{\gamma}} \ell_N(\boldsymbol{\gamma}_0; \widehat{\pi}_N)\right\|_\infty \leq \lambda_N, \qquad \text{on events } \mathcal{A}, \mathcal{B} \text{ and } \mathcal{C}.
$$

Define event

$$
\mathcal{D} := \left\{\delta\ell(\boldsymbol{\Delta}; \widehat{\pi}_N; \boldsymbol{\gamma}) \geq \widehat{\pi}_N\kappa\|\boldsymbol{\Delta}\|_2^2, \qquad \forall\, \boldsymbol{\Delta} \in \mathbb{C}_\delta(S;3) \text{ and } \delta \leq 1\right\}. \tag{2.115}
$$

By Lemma 2.1, $P(\mathcal{D}) \geq 1 - \alpha_N$. Let $\delta_N^* = 2\lambda_N s^{1/2}(\widehat{\pi}_N\kappa)^{-1}$. Then, the RSC condition holds for $\ell_N(\cdot; \widehat{\pi}_N)$ with parameter $\widehat{\pi}_N\kappa$ over $\mathbb{C}_{\delta_N^*}(S;3)$. By Theorem 1 of [NRWY10], when $\lambda_N \geq 2\left\|\nabla_{\boldsymbol{\gamma}} \ell_N(\boldsymbol{\gamma}_0; \widehat{\pi}_N)\right\|_\infty$, $2\lambda_N s^{1/2}(\widehat{\pi}_N\kappa)^{-1} \leq 1$ and on event $\mathcal{D}$,

$$
\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \leq \delta_N^* \leq 2\lambda_N s^{1/2}(\widehat{\pi}_N\kappa)^{-1}.
$$

Recall (2.111), for any $t > 0$, on event $\mathcal{A} = \mathcal{A}(t)$,

$$
\begin{aligned}
\widehat{\pi}_N &\geq \pi_N - 2\sqrt{\frac{t\log(p+1)\pi_N}{N}} - \frac{t\log(p+1)}{N} \geq \frac{2}{9}\pi_N, \\
2\lambda_N s^{1/2}(\widehat{\pi}_N\kappa)^{-1} &\leq \frac{1}{9}\lambda_N s^{1/2}\pi_N^{-1}\kappa^{-1} \leq 1,
\end{aligned}
$$

when $t < N\pi_N\{\log(p+1)\}^{-1}/9$ and $\lambda_N \leq 9\kappa\pi_N s^{-1/2}$. Hence, if $N\pi_N > 9c\log(p+1)$,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \leq \frac{1}{9}\lambda_N s^{1/2}\pi_N^{-1}\kappa^{-1}, \quad \text{on events } \mathcal{A}, \mathcal{B}, \mathcal{C} \text{ and } \mathcal{D},$$

where $P_{\mathbb{S}}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap\mathcal{D}) \geq 1 - 8(p+1)^{-c} - \alpha_N$.

Now, consider the asymptotic performance that as $N \to \infty$, $\log(p)\log(N) = O(N\pi_N)$ and $s\log(p) = o(N\pi_N)$. Then,

$$M_N \asymp \pi_N^{1/2}\sqrt{\frac{\log(p)}{N}}.$$

Hence, with some $\lambda_N \asymp \{N^{-1}\pi_N\log(p)\}^{1/2}$,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 = O_p\left(\sqrt{\frac{s\log(p)}{N\pi_N}}\right) = o_p(1). \tag{2.116}$$

Now we analyze the consistency rate of the PS estimator $\widehat{\pi}_N(\cdot)$. For any $r > 0$,

$$\left\|1 - \frac{\pi_N(\cdot)}{\widehat{\pi}_N(\cdot)}\right\|_{r,P} \leq \|\widehat{\pi}_N(\cdot) - \pi_N(\cdot)\|_{2r,P}\,\|\widehat{\pi}_N^{-1}(\cdot)\|_{2r,P}.$$

Let $u_0 = \overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0 + \log(\pi_N)$ and $\Delta_u = \overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}} + \log(\widehat{\pi}_N) - \{\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0 + \log(\pi_N)\}$. By mean value theorem, and notice that $g'(u) = g(u)\{1-g(u)\}$, for some $v' \in (0,1)$,

$$\begin{aligned}|g(u_0+\Delta_u) - g(u_0)| &= g'(u_0 + v'\Delta_u)|\Delta_u| \leq g(u_0+v'\Delta_u)|\Delta_u| \\ &\leq \max\{g(u_0), g(u_0+\Delta_u)\}|\Delta_u| \leq \{g(u_0)+g(u_0+\Delta_u)\}|\Delta_u|,\end{aligned}$$

since the function $g(\cdot) > 0$ is monotone increasing. Besides, notice that, on $\mathcal{A}$ and $\mathcal{E} := \{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2 \leq 1\}$,

$$\begin{aligned}|\log(\widehat{\pi}_N) - \log(\pi_N)| \leq \frac{|\widehat{\pi}_N - \pi_N|}{\min(\widehat{\pi}_N, \pi_N)} &\leq \frac{2\sqrt{t_2\pi_N/N} + t_2/N}{2\pi_N/9} \leq \frac{21}{2}\sqrt{\frac{t_2}{N\pi_N}}, \\ g(u_0) = \frac{\pi_N\exp(-\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0)}{1+\pi_N\exp(-\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0)} &\leq \pi_N\exp(-\overrightarrow{\mathbf{x}}^T\boldsymbol{\gamma}_0), \\ g(u_0+\Delta_u) = \frac{\widehat{\pi}_N\exp(-\overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}})}{1+\widehat{\pi}_N\exp(-\overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}})} &\leq \widehat{\pi}_N\exp(-\overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}}) \leq \frac{16}{9}\pi_N\exp(-\overrightarrow{\mathbf{x}}^T\widehat{\boldsymbol{\gamma}}),\end{aligned}$$

when $t_2 < N\pi_N/9$. Hence, on $\mathcal{A} \cap \mathcal{E}$,

$$\begin{aligned}
&\|\widehat{\pi}_N(\cdot) - \pi_N(\cdot)\|_{2r,P} \\
&\quad \le \pi_N \left\| \left\{ \exp(-\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) + \frac{16}{9} \exp(-\overrightarrow{\mathbf{X}}^T \widehat{\boldsymbol{\gamma}}) \right\} \left\{ |\overrightarrow{\mathbf{X}}^T (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)| + \frac{21}{2} \sqrt{\frac{t_2}{N\pi_N}} \right\} \right\|_{2r,P} \\
&\quad \le \left[\left\| \exp(-\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) \right\|_{4r,P} + \frac{16}{9} \left\| \exp(-\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0) \right\|_{8r,P} \left\| \exp\{-\overrightarrow{\mathbf{X}}^T (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\} \right\|_{8r,P} \right] \\
&\qquad \cdot \left\{ \left\| \overrightarrow{\mathbf{X}}^T (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \right\|_{4r,P} + \frac{21}{2} \sqrt{\frac{t_2}{N\pi_N}} \right\} \\
&\quad \le C \left\{ \|\widehat{\boldsymbol{\gamma}} - \gamma_0\|_2 + \sqrt{\frac{t_2}{N\pi_N}} \right\},
\end{aligned}$$

with some constant $C > 0$. Here, $P(\mathcal{A}) \ge 1 - \exp(-t_2)$ and recall (2.116). Hence,

$$\|\widehat{\pi}_N(\cdot) - \pi_N(\cdot)\|_{2r,P} = O_p\left(\sqrt{\frac{s \log(p)}{N\pi_N}} \right). \tag{2.117}$$

Additionally, observe that

$$\begin{aligned}
\|\widehat{\pi}_N^{-1}(\cdot)\|_{2r,P} = \|1 + \widehat{\pi}_N^{-1} \exp(-\overrightarrow{\mathbf{X}}^T \widehat{\boldsymbol{\gamma}})\|_{2r,P} &\le 1 + \widehat{\pi}_N^{-1} \| \exp(-\overrightarrow{\mathbf{X}}^T \widehat{\boldsymbol{\gamma}})\|_{2r,P} \\
&\le 1 + \widehat{\pi}_N^{-1} \| \exp(-\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)\|_{4r,P} \| \exp(-U)\|_{4r,P}.
\end{aligned}$$

By (2.106), $\| \exp(-\overrightarrow{\mathbf{X}}^T \boldsymbol{\gamma}_0)\|_{4r,P} = O(1)$. By (2.122), $\widehat{\pi}_N^{-1} = \pi_N^{-1}\{1 + o_p(1)\}$. By Lemma part (b) of 2.8,

$$|E(U)| \le E(|U|) \le \sigma\sqrt{\pi} \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2.$$

Hence, by triangular inequality and part (a) of 2.8,

$$\|U - E(U)\|_{\psi_2} \le \|U\|_{\psi_2} + \|E(U)\|_{\psi_2} \le \{1 + \sqrt{\pi/\log(2)}\}\sigma \|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 \le 4\sigma \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2.$$

By part (c) of Lemma 2.8,

$$\| \exp\{-U + E(U)\}\|_{4r,P} = (E \exp[-4r\{U - E(U)\}])^{1/(4r)} \le \exp\left(128 r \sigma^2 \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2\right).$$

Hence,

$$\begin{aligned}\|\exp(-U)\|_{4r,P} &= \|\exp\{-U+E(U)\}\|_{4r,P}\exp\{-E(U)\}\\ &\le \exp\left(128r\sigma^2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2^2\right)\exp(\sigma\sqrt{\pi}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2) = 1+o_p(1).\end{aligned}$$

It follows that

$$\|\widehat{\pi}_N^{-1}(\cdot)\|_{2r,P} \le 1+\pi_N^{-1}\{1+o_p(1)\}\cdot O(1)\cdot\{1+o_p(1)\} = O_p(\pi_N^{-1}). \tag{2.118}$$

Therefore, by (2.117) and (2.118),

$$\left\|1-\frac{\pi_N(\cdot)}{\widehat{\pi}_N(\cdot)}\right\|_{r,P} \le \|\widehat{\pi}_N(\cdot)-\pi_N(\cdot)\|_{2r,P}\,\|\widehat{\pi}_N^{-1}(\cdot)\|_{2r,P} = O_p\left(\sqrt{\frac{s\log(p)}{N\pi_N}}\right).$$

Besides, recall (2.103) and notice that

$$\|\pi_N^{-1}(\mathbf{X})\|_{r,P} \le 1+\pi_N^{-1}\|\exp(-\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0)\|_{r,P} \le 1+\pi_N^{-1}\exp(2\sigma_{\boldsymbol{\gamma}_0}+20\sigma_{\boldsymbol{\gamma}_0}r) = O(\pi_N^{-1}).$$

Therefore,

$$E\left[\frac{a_N}{\pi_N(\mathbf{X})}\left\{1-\frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X})}\right\}^2\right] \le a_N\left\|\pi_N^{-1}(\mathbf{X})\right\|_{2,P}\left\|1-\frac{\pi_N(\cdot)}{\widehat{\pi}_N(\cdot)}\right\|_{4,P}^2 = O_p\left(\frac{s\log(p)}{N\pi_N}\right).$$

If further assume $\|\widehat{m}(\cdot)-m(\cdot)\|_{2+c,P} = o_p(1)$, then

$$E\left[\frac{a_N}{\pi_N(\mathbf{X})}\{\widehat{m}(\mathbf{X})-m(\mathbf{X})\}^2\right] \le a_N\left\|\pi_N^{-1}(\mathbf{X})\right\|_{1+c/2,P}\|\widehat{m}(\cdot)-m(\cdot)\|_{2+c,P}^2 = o_p(1).$$

■

Proof of Lemma 2.3. The proof of Lemma 2.3 is based on the proof of Proposition 2 in [NRWY10]. Here, we only provide the details that are different from their proof and we will use our notations in the folloing proof. As a reminder, N denotes the number of samples, $\overrightarrow{\mathbf{X}} \in \mathbb{R}^{p+1}$ is the covariate containing the intercept term and $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p+1}$ is the coefficient

of the balanced logistic model (the dimension p in their proof will be replaced by $p+1$ everywhere because of the usage of the intercept term).

The proof consists of 3 main steps: 1) show that (71) of [NRWY10] holds under our assumptions and the parameter K_3 we choose, 2) prove a slightly different version of (72) in [NRWY10], 3) conclude the RSC property result.

Step 1. For the inequality (71), similarly as in their proof, we have $E\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2\} \geq \kappa_l\|\boldsymbol{\Delta}\|_2^2 = \kappa_l$ for any $\|\boldsymbol{\Delta}\|_2 = 1$. Hence, it suffices to show their inequality (73). Instead of assuming $\overrightarrow{\mathbf{X}}$ to be a zero-mean jointly sub-Gaussian random vector (which is not ture since we have the intercept term here), we only assume a $(2+c)$-th moment condition that $\sup_{\|\mathbf{v}\|_2\leq 1}\|\overrightarrow{\mathbf{X}}^T\mathbf{v}\|_{2+c,P} \leq M < \infty$ and a c-th moment condition that $\|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0\|_{c,P} \leq \mu_c < \infty$, with our choice on the constant K_3. We have

$$\sup_{\|\boldsymbol{\Delta}\|_2\leq 1} P(|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}| \geq \tau/2) \leq (\tau/2)^{-2-c} \sup_{\|\mathbf{v}\|_2\leq 1} E|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}|^{2+c} \leq M^{2+c}(\tau/2)^{-2-c},$$
$$P(|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0| \geq T) \leq E|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0|^c T^{-c} = \mu_c^c T^{-c}.$$

Hence, by Hölder's Inequality, for any $\|\boldsymbol{\Delta}\|_2 \leq 1$,

$$E\left\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2 1_{|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0|\geq T}\right\} \leq \|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}\|_{2+c,P}^2\{P(|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0| \geq T)\}^{\frac{c}{2+c}} \leq M^2\mu_c^{\frac{c^2}{2+c}}T^{-\frac{c^2}{2+c}},$$
$$E\left\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2 1_{|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}|\geq \tau/2}\right\} \leq \|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}\|_{2+c,P}^2\{P(|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}| \geq \tau/2)\}^{\frac{c}{2+c}} \leq M^{2+c}(\tau/2)^{-c}.$$

It follows that, for $\tau^2 = T^2 = K_3 \geq 1$,

$$E\left\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2 - g_{\boldsymbol{\Delta}}(\mathbf{X})\right\} \leq E\left\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2 1_{|\overrightarrow{\mathbf{X}}^T\boldsymbol{\gamma}_0|\geq T}\right\} + E\left\{(\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta})^2 1_{|\overrightarrow{\mathbf{X}}^T\boldsymbol{\Delta}|\geq \tau/2}\right\}$$
$$\leq M^2\mu_c^{\frac{c^2}{2+c}}T^{-\frac{c^2}{2+c}} + M^{2+c}(\tau/2)^{-c} \leq (M^2\mu_c^{\frac{c^2}{2+c}} + M^{2+c}2^c)K_3^{-\frac{c^2}{4+2c}}.$$

Hence, (73) of [NRWY10] holds when we set

$$K_3 = \max\left[1, \left\{2\kappa_l^{-1}(M^2\mu_c^{\frac{c^2}{2+c}} + M^{2+c}2^c)\right\}^{\frac{4+2c}{c}}\right].$$

Step 2. We will demonstrate a slightly different version of (72) in [NRWY10] that

$$P_{\mathbb{S}}\left\{Z(t) \geq \frac{\kappa_l}{4} + 66K_3\sigma\sqrt{\frac{\log(p+1)}{N}}t\right\} \leq \exp\left\{-\frac{N\kappa_l^2}{64K_3^2} - \sigma^2t^2\log(p+1)\right\}. \tag{2.119}$$

Set $z^*(t) = \kappa_l/4 + 2K_3\sigma\sqrt{\log(p+1)/N}t$ and let

$$\mathcal{F} := \{\pm f(\cdot) : f(\mathbf{u}) = g_{\boldsymbol{\Delta}}(\mathbf{u}) - E\{g_{\boldsymbol{\Delta}}(\mathbf{X})\}, \|\boldsymbol{\Delta}\|_2 = 1, \|\boldsymbol{\Delta}\|_1 = t\}.$$

Since $0 \leq g_{\boldsymbol{\Delta}}(\mathbf{u}) \leq K_3$ for all $\mathbf{u}$, we have $|f(\mathbf{X}_i)| \leq K_3$ for all $f \in \mathcal{F}$. By Lemma 2.10, we have a slightly different version of their (76):

$$P_{\mathbb{S}}[Z(t) \geq E\{Z(t)\} + z^*(t)] \leq \exp\left[-\frac{N\{z^*(t)\}^2}{4K_3^2}\right] \leq \exp\left\{-\frac{N\kappa_l^2}{64K_3^2} - \sigma^2t^2\log(p+1)\right\}. \tag{2.120}$$

Now, we need to obtain an upper bound for $E_{\mathbb{S}}\|N^{-1}\sum_{i=1}^N \varepsilon_i\mathbf{u}_i\|_\infty$ only using the marginal sub-Gaussianity of $\overrightarrow{\mathbf{X}}$. Firstly, since $|\varepsilon_i\mathbf{u}_i(j)| \leq |X_i(j)|$, by part (a) of Lemma 2.8,

$$\sup_{1\leq j\leq p+1}\|\varepsilon_i\mathbf{u}_i(j)\|_{\psi_2} \leq \sup_{1\leq j\leq p+1}\|\overrightarrow{\mathbf{X}}_i(j)\|_{\psi_2} \leq \sigma.$$

Notice that $E(\varepsilon\mathbf{u}) = 0$ since ε is independent with $\mathbf{u}$ and $E(\varepsilon) = 0$, by part (e) of Lemma 2.8, for any $1 \leq j \leq p+1$,

$$\left\|N^{-1}\sum_{i=1}^N \varepsilon_i\mathbf{u}_i(j)\right\|_{\psi_2} \leq 4\sigma/\sqrt{N}.$$

By part (d) of Lemma 2.8,

$$\left\|\left\|N^{-1}\sum_{i=1}^N \varepsilon_i\mathbf{u}_i\right\|_\infty\right\|_{\psi_2} \leq 4\{\log(p+1)+2\}^{1/2}\sigma/\sqrt{N} \leq 8\{\log(p+1)\}^{1/2}\sigma/\sqrt{N},$$

for any $p \geq 1$. Hence, by part (b) of Lemma 2.8,

$$E_{\mathbb{S}} \left\| N^{-1} \sum_{i=1}^{N} \varepsilon_i \mathbf{u}_i \right\|_{\infty} \leq 8\sqrt{\pi}\sigma \sqrt{\frac{\log(p+1)}{N}}.$$

Combining the upper bound with (78) of [NRWY10], we have a slightly different version of their inequality (77):

$$E_{\mathbb{S}}\{Z(t)\} \leq 64 K_3 t \sqrt{\pi}\sigma \sqrt{\frac{\log(p+1)}{N}}.$$

and recall (2.120), hence (2.119) follows. Notice that the statements in Step 2 are all independent of the choice of K_3, so our choice on the constant K_3 does not affect the validity of the resutlts.

Step 3. By inequality (71) of [NRWY10] and our (2.119), we conclude that: for any $t > 0$,

$$P_{\mathbb{S}} \left[\widehat{E}_N\{g_{\boldsymbol{\Delta}}(\mathbf{X})\} < \frac{\kappa_l}{4} - 66 K_3 \sigma \sqrt{\frac{\log(p+1)}{N}} t, \ \ \exists \boldsymbol{\Delta} \in \mathbb{R}^{p+1}, \text{ with } \|\boldsymbol{\Delta}\|_1 = t, \|\boldsymbol{\Delta}\|_2 = 1 \right]$$
$$\leq \exp\left\{ -\frac{N\kappa_l^2}{64 K_3^2} - \sigma^2 t^2 \log(p+1) \right\}.$$

Let $\mathbb{S}(1,t) = \{\boldsymbol{\Delta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\Delta}\|_2 \leq 1, \|\boldsymbol{\Delta}\|_1 / \|\boldsymbol{\Delta}\|_2 = t\}$. By their inequality (66) and the technique in (69),

$$P_{\mathbb{S}} \left[\delta\ell(\boldsymbol{\Delta}; 1; \boldsymbol{\gamma}_0) < L_\psi(K_3^{1/2}) \left\{ \frac{\kappa_l}{4} \|\boldsymbol{\Delta}\|_2^2 - 66 K_3 \sigma \sqrt{\frac{\log(p+1)}{N}} \|\boldsymbol{\Delta}\|_2 t \right\}, \ \ \exists \boldsymbol{\Delta} \in \mathbb{S}(1,t) \right]$$
$$\leq \exp\left\{ -\frac{N\kappa_l^2}{64 K_3^2} - \sigma^2 t^2 \log(p+1) \right\},$$

where for a logistic model, $L_\psi(K_3^{1/2}) = \dot{g}(2K_3^{1/2}) > 0$.

By a peeling argument as in [RWY10], (2.36) holds. If further assume that $s\log(p) = o(N)$. For any $\boldsymbol{\Delta} \in \mathcal{C}(S,3)$,

$$\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1 \leq 4\sqrt{s}\|\boldsymbol{\Delta}_S\|_2 \leq 4\sqrt{s}\|\boldsymbol{\Delta}\|_2.$$

and hence (2.37) holds. ∎

Proof of Theorem 2.6. We establish the asymptotic properties of the stratified PS estimator and the DRSS estimator based on the stratified PS estimator.

Let $\pi_N = E(R)$, then $\pi_{1,N} + \pi_{0,N} \in (\pi_N/(1-C), \pi_N/C)$ and $\pi(\mathbf{X}) \in (C\pi_N/(1-C), (1-C)\pi_N/C)$ for all $\mathbf{X} \in \mathcal{X}$. Let $N_1 = \sum_{i\in\mathcal{I}_{-k}} \delta_i$ and $N_0 = \sum_{i\in\mathcal{I}_{-k}}(1-\delta_i)$. Similarly as in (2.122), for $j \in \{0,1\}$, $N_j^{-1} = O_p(N^{-1})$, $\widehat{\pi}_j(\mathbb{S}_{-k}) - \pi_{j,N} = O_p(\sqrt{\pi_N/N})$ and $1 - \pi_{j,N}/\hat{\pi}_j(\mathbb{S}_{-k}) = O_p(1/\sqrt{N\pi_N})$. Hence, $\widehat{\pi}_j^{-1}(\mathbb{S}_{-k}) = \pi_{j,N}^{-1}\{1 + O_p(1/\sqrt{N\pi_N})\}$. It follows that there exists $c > 0$ such that

$$P_{\mathbb{S}_{-k}}(\widehat{\pi}_j(\mathbb{S}_{-k}) > c\pi_N) \to 1, \quad \text{for } \ j \in \{0,1\}.$$

Hence, with probability approaching 1,

$$\begin{aligned}\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k}) &= \widehat{\pi}_1(\mathbb{S}_{-k})\widehat{p}_\delta(\mathbf{X};\mathbb{S}_{-k}) + \widehat{\pi}_0(\mathbb{S}_{-k})\{1 - \widehat{p}_\delta(\mathbf{X};\mathbb{S}_{-k})\} \\ &\geq c\pi_N\widehat{p}_\delta(\mathbf{X};\mathbb{S}_{-k}) + c\pi_N\{1 - \widehat{p}_\delta(\mathbf{X};\mathbb{S}_{-k})\} = c\pi_N.\end{aligned}$$

Observe that

$$\begin{aligned}\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k}) - \pi_N(\mathbf{X}) &= \{\widehat{\pi}_1(\mathbb{S}_{-k}) - \widehat{\pi}_0(\mathbb{S}_{-k})\}\{\widehat{p}_\delta(\mathbf{X};\mathbb{S}_{-k}) - p_\delta(\mathbf{X})\} \\ &\quad + \{\widehat{\pi}_1(\mathbb{S}_{-k}) - \pi_{1,N}\}p_\delta(\mathbf{X}) + \{\widehat{\pi}_0(\mathbb{S}_{-k}) - \pi_{0,N}\}\{1 - p_\delta(\mathbf{X})\}.\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\frac{\widehat{\pi}_N(\cdot;\mathbb{S}_{-k}) - \pi_N(\cdot)}{\pi_N(\cdot)}\right\|_{2,\mathbb{P}_\mathbf{X}} &= O_p\left(r_{p_\delta,N} + (N\pi_N)^{-1/2}\right), \\ \left\|\frac{\widehat{\pi}_N(\cdot;\mathbb{S}_{-k}) - \pi_N(\cdot)}{\widehat{\pi}_N(\cdot;\mathbb{S}_{-k})}\right\|_{2,\mathbb{P}_\mathbf{X}} &= O_p\left(r_{p_\delta,N} + (N\pi_N)^{-1/2}\right).\end{aligned}$$

Following the case (b) in Theorem 2.2 that $\pi_N(\cdot) = e_N(\cdot)$ being correctly specified,

$$\widehat{\theta}_{\text{DRSS}} - \theta_0 = \frac{1}{N}\sum_{i=1}^N \psi_{\mu,e}(\mathbf{Z}_i) + \widehat{\Delta}_N + O_p\left(\frac{c_{\mu,N}}{\sqrt{Na_N}} + \frac{c_{e,N}}{\sqrt{Na_N}} + r_{\pi,m.N}\right),$$

where $\psi_{\mu,e}(\mathbf{Z}) = \mu(X) - \theta_0 + R/\pi_N(X)[Y - \mu(X)]$, $\widehat{\Delta}_N = \sum_{k=1}^{\mathbb{K}} \widehat{\Delta}_{N,k}$ and

$$\widehat{\Delta}_{N,k} = N^{-1} \sum_{i \in \mathcal{I}_k} \frac{R_i}{\pi_N(\mathbf{X}_i)} \left\{ 1 - \frac{\pi_N(\mathbf{X}_i)}{\widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})} \right\} \{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\}.$$

With a slight abuse of notation, let $\mathbf{Z} = (\delta, R, \mathbf{X})$, $\mathbf{Z}_i = (\delta_i, R_i, \mathbf{X}_i)$ and $\mathbb{S}_k = \{\mathbf{Z}_i : i \in \mathcal{I}_k\}$. Then,

$$\begin{aligned}
\mathrm{Var}_{\mathbb{S}_k}(\widehat{\Delta}_{N,k}) &= N^{-2}|\mathcal{I}_k|\mathrm{Var}\left[\frac{R}{\pi_N(\mathbf{X})} \left\{ 1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \right\} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&\leq N^{-1} E\left[\frac{1}{\pi_N(\mathbf{X})} \left\{ 1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \right\}^2 \{\mu(\mathbf{X}) - m(\mathbf{X})\}^2 \right] \\
&\leq \frac{1 - C}{C} (N\pi_N)^{-1} \left\| \frac{\widehat{\pi}_N(\cdot; \mathbb{S}_{-k}) - \pi_N(\cdot)}{\widehat{\pi}_N(\cdot; \mathbb{S}_{-k})} \right\|_{2,\mathbb{P}_\mathbf{X}}^2 \|\mu(\cdot) - m(\cdot)\|_{\infty,\mathbb{P}_\mathbf{X}}^2 \\
&= O_p\left((N\pi_N)^{-1} r_{p_\delta,N}^2 + (N\pi_N)^{-2} \right),
\end{aligned}$$

and by Lemma 2.4,

$$\widehat{\Delta}_{N,k} = E_{\mathbb{S}_k}(\widehat{\Delta}_{N,k}) + O_p\left((N\pi_N)^{-1/2} r_{p_\delta,N} + (N\pi_N)^{-1} \right).$$

In addition,

$$\begin{aligned}
N|\mathcal{I}_k|^{-1} E_{\mathbb{S}_k}(\widehat{\Delta}_{N,k}) &= E\left[\frac{R}{\pi_N(\mathbf{X})} \left\{ 1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \right\} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&= E\left[\left\{ 1 - \frac{\pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \right\} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&= E\left[\frac{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) - \pi_N(\mathbf{X})}{\pi_N(\mathbf{X})} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&\quad - E\left[\frac{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) - \pi_N(\mathbf{X})}{\pi_N(\mathbf{X})} \cdot \frac{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) - \pi_N(\mathbf{X})}{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k})} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&= E\left[\frac{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) - \pi_N(\mathbf{X})}{\pi_N(\mathbf{X})} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] \\
&\quad + O_p\left(\left\| \frac{\widehat{\pi}_N(\cdot; \mathbb{S}_{-k}) - \pi_N(\cdot)}{\pi_N(\cdot)} \right\|_{2,\mathbb{P}_\mathbf{X}} \left\| \frac{\widehat{\pi}_N(\cdot; \mathbb{S}_{-k}) - \pi_N(\cdot)}{\widehat{\pi}_N(\cdot; \mathbb{S}_{-k})} \right\|_{2,\mathbb{P}_\mathbf{X}} \|\mu(\cdot) - m(\cdot)\|_\infty \right) \\
&= E\left[\frac{\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) - \pi_N(\mathbf{X})}{\pi_N(\mathbf{X})} \{\mu(\mathbf{X}) - m(\mathbf{X})\} \right] + O_p\left(\|\widehat{p}_\delta(\cdot) - p_\delta(\cdot)\|_{2,\mathbb{P}_\mathbf{X}}^2 + (N\pi_N)^{-1} \right).
\end{aligned}$$

Let $\widetilde{\pi}_N(\cdot;\mathbb{S}_{-k}) = \widehat{\pi}_1(\mathbb{S}_{-k})p_\delta(\cdot) + \widehat{\pi}_0(\mathbb{S}_{-k})\{1-p_\delta(\cdot)\}$. Then,

$$
\begin{aligned}
&E\left[\frac{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] \\
&\quad= E\left[\frac{\widetilde{\pi}_N(\mathbf{X};\mathbb{S}_{-k})-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] \\
&\qquad+ E\left[\frac{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})-\widetilde{\pi}_N(\mathbf{X};\mathbb{S}_{-k})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] \\
&\quad= E\left[\frac{\widetilde{\pi}_N(\mathbf{X};\mathbb{S}_{-k})-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] + O_p\left(r_{p_\delta,N}\|\mu(\cdot)-m(\cdot)\|_{2,\mathbb{P}_\mathbf{X}}\right) \\
&\quad= \{\widehat{\pi}_1(\mathbb{S}_{-k})-\pi_{1,N}\}E\left[\frac{p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] \\
&\qquad+ \{\widehat{\pi}_0(\mathbb{S}_{-k})-\pi_{0,N}\}E\left[\frac{1-p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] + O_p\left(r_{p_\delta,N}\right).
\end{aligned}
$$

Let $\widehat{p}_\delta(\mathbb{S}_{-k}) = N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\delta_i$ and $p_\delta = E[p_\delta(\mathbf{X})]$. Then, similarly as (2.122), $\widehat{p}_\delta^{-1}(\mathbb{S}_{-k}) = p_\delta^{-1}\{1+O_p(1/\sqrt{N})\}$. Hence,

$$
\begin{aligned}
\widehat{\pi}_1(\mathbb{S}_{-k})-\pi_{1,N} &= N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{\delta_i R_i}{\widehat{p}_\delta(\mathbb{S}_{-k})}-\pi_{1,N} \\
&= N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{\delta_i R_i}{p_\delta}-\pi_{1,N}+O_p(N^{-1/2})N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{\delta_i R_i}{p_\delta} \\
&= N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{\delta_i R_i}{p_\delta}-\pi_{1,N}+O_p(N^{-1/2}\pi_N).
\end{aligned}
$$

Similarly,

$$
\widehat{\pi}_0(\mathbb{S}_{-k})-\pi_{0,N} = N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{(1-\delta_i)R_i}{1-p_\delta}-\pi_{0,N}+O_p(N^{-1/2}\pi_N).
$$

Let

$$
\begin{aligned}
\mathrm{IF}_\pi(\mathbf{Z}) &= \left\{\frac{\delta R}{p_\delta}-\pi_{1,N}\right\}E\left[\frac{p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right] \\
&\quad+\left\{\frac{(1-\delta)R}{1-p_\delta}-\pi_{0,N}\right\}E\left[\frac{1-p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right].
\end{aligned}
$$

Then,

$$E\left[\frac{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})-\pi_N(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right]=N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_\pi(\mathbf{Z}_i)+O_p\left(N^{-1/2}\right).$$

Hence,

$$\begin{aligned}\widehat{\Delta}_N=\sum_{k=1}^{\mathbb{K}}\widehat{\Delta}_{N,k}&=(\mathbb{K}N_{-k})^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_{-k}}\mathrm{IF}_\pi(\mathbf{Z}_i)+O_p\left(\|\widehat{p}_\delta(\cdot)-p_\delta(\cdot)\|_{2,\mathbb{P}_\mathbf{X}}^2+(N\pi_N)^{-1}\right)\\&\quad+O_p\left(r_{p_\delta,N}\right)+O_p\left(N^{-1/2}\right)+O_p\left((N\pi_N)^{-1/2}r_{p_\delta,N}+(N\pi_N)^{-1}\right)\\&=N^{-1}\sum_{i=1}^{N}\mathrm{IF}_\pi(\mathbf{Z}_i)+O_p\left(r_{p_\delta,N}+(N\pi_N)^{-1}+N^{-1/2}\right).\end{aligned}$$

By part (b) of Theorem 2.2,

$$\begin{aligned}\widehat{\theta}_{\mathrm{DRSS}}-\theta_0&=\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{Z}_i)+O_p\left(r_{p_\delta,N}+(N\pi_N)^{-1}+N^{-1/2}\right)\\&\quad+O_p\left(r_{\mu,N}(N\pi_N)^{-1/2}+\left\{r_{p_\delta,N}+(N\pi_N)^{-1/2}\right\}\left\{(N\pi_N)^{-1/2}+r_{\mu,N}\right\}\right),\\&=\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{Z}_i)+O_p\left((N\pi_N)^{-1}+N^{-1/2}+r_{\mu,N}(N\pi_N)^{-1/2}+r_{p_\delta,N}\right),\end{aligned}$$

where $\Psi(\mathbf{Z}):=\psi_\mu(\mathbf{Z})+\mathrm{IF}_\pi(\mathbf{Z})$ and $E\{\Psi(\mathbf{Z})\}=0$ with

$$\begin{aligned}\psi_\mu(\mathbf{Z})&=\frac{R}{\pi_N(\mathbf{X})}\{Y-\mu(\mathbf{X})\}+\mu(\mathbf{X})-\theta_0,\\\mathrm{IF}_\pi(\mathbf{Z})&=\left\{\frac{\delta R}{p_\delta}-\pi_{1,N}\right\}E\left[\frac{p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right]\\&\quad+\left\{\frac{(1-\delta)R}{1-p_\delta}-\pi_{0,N}\right\}E\left[\frac{1-p_\delta(\mathbf{X})}{\pi_N(\mathbf{X})}\{\mu(\mathbf{X})-m(\mathbf{X})\}\right].\end{aligned}$$

If further $r_{p_\delta,N}=o_p((N\pi_N)^{-1/2})$,

$$\widehat{\theta}_{\mathrm{DRSS}}-\theta_0=\frac{1}{N}\sum_{i=1}^{N}\Psi(\mathbf{Z}_i)+o_p\left((N\pi_N)^{-1/2}\right).$$

■

Proof of Theorem 2.7. Here we provide asymptotic results of the DRSS estimator based on a MCAR PS.

Under MCAR, neither $\pi_N(\mathbf{X}) \equiv \pi_N$ nor $\widehat{\pi}_N(\mathbf{X}; \mathbb{S}_{-k}) \equiv \widehat{\pi}_N(\mathbb{S}_{-k})$ depend on $\mathbf{X}$, and we recall that $\pi_N = P(R = 1)$. For each $k \leq \mathbb{K}$, $\widehat{\pi}_N(\mathbb{S}_{-k}) = N_{-k}^{-1} \sum_{i \in \mathcal{I}_{-k}} R_i$, where $N_{-k} = N - N/\mathbb{K}$. Notice that

$$E_{\mathbb{S}_{-k}} \left[\left\{ \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} \right\}^2 \right] = \pi_N^{-2} N_{-k}^{-1} E(R - \pi_N)^2 = N_{-k}^{-1} \pi_N^{-1} (1 - \pi_N) = O\left((N\pi_N)^{-1} \right).$$

By Lemma 2.4,

$$\frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} = O_p\left((N\pi_N)^{-1/2} \right). \tag{2.121}$$

By the fact that

$$\frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})} \left\{ 1 + \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} \right\} = \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N},$$

we have

$$\frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})} = \left\{ 1 + \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} \right\}^{-1} \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} = O_p\left((N\pi_N)^{-1/2} \right). \tag{2.122}$$

Hence,

$$\frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})} - \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} = \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\pi_N} \cdot \frac{\widehat{\pi}_N(\mathbb{S}_{-k}) - \pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})} = O_p\left((N\pi_N)^{-1} \right).$$

Additionally, notice that

$$E_{\mathbb{S}_k} \left[|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \frac{R_i}{\pi_N} \{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\} \right] = E\{\mu(\mathbf{X}) - m(\mathbf{X})\},$$

$$E_{\mathbb{S}_k} \left[|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \frac{R_i}{\pi_N} \{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\} \right]^2 = |\mathcal{I}_k|^{-1} \pi_N^{-1} E\left[\{\mu(\mathbf{X}) - m(\mathbf{X})\}^2 \right] = O\left((N\pi_N)^{-1} \right).$$

Let $\Delta_\mu = E\{\mu(\mathbf{X}) - m(\mathbf{X})\}$. By Lemma 2.4,

$$|\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \frac{R_i}{\pi_N} \{\mu(\mathbf{X}_i) - m(\mathbf{X}_i)\} - \Delta_\mu = O_p\left((N\pi_N)^{-1/2} \right). \tag{2.123}$$

Using the definition of $\widehat{\Delta}_N$ from Theorem 2.2 and adapting it to the MCAR setting,

$$\begin{aligned}\widehat{\Delta}_N &= N^{-1}\sum_{i=1}^{N}\frac{R_i}{\pi_N(\mathbf{X}_i)}\left\{\frac{\widehat{\pi}_N(\mathbb{S}_{-k})-\pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})}\right\}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}\\ &= \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\frac{\widehat{\pi}_N(\mathbb{S}_{-k})-\pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})}\left[|\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k}\frac{R_i}{\pi_N}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}\right].\end{aligned}$$

Recall (2.122) and (2.123), since $\mathbb{K}<\infty$ is a fixed number, we have

$$\begin{aligned}\sup_{k\le\mathbb{K}}\left|\frac{\widehat{\pi}_N(\mathbb{S}_{-k})-\pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})}\right| &= O_p\left((N\pi_N)^{-1/2}\right),\\ \sup_{k\le\mathbb{K}}\left||\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k}\frac{R_i}{\pi_N}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}-\Delta_\mu\right| &= O_p\left((N\pi_N)^{-1/2}\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\left|\widehat{\Delta}_N-\mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\frac{\widehat{\pi}_N(\mathbb{S}_{-k})-\pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})}\Delta_\mu\right|\\ &\quad\le\sup_{k\le\mathbb{K}}\left|\frac{\widehat{\pi}_N(\mathbb{S}_{-k})-\pi_N}{\widehat{\pi}_N(\mathbb{S}_{-k})}\right|\sup_{k\le\mathbb{K}}\left||\mathcal{I}_k|^{-1}\sum_{i\in\mathcal{I}_k}\frac{R_i}{\pi_N}\{\mu(\mathbf{X}_i)-m(\mathbf{X}_i)\}-\Delta_\mu\right|\\ &\quad= O_p\left((N\pi_N)^{-1}\right).\end{aligned}$$

It follows that,

$$\begin{aligned}\widehat{\Delta}_N &= \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\left(\frac{\widehat{\pi}_N(S_{-k})-\pi_N}{\pi_N}\right)\Delta_\mu+O_p\left((N\pi_N)^{-1}\right)\\ &\overset{(i)}{=}\mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}\frac{R_i-\pi_N}{\pi_N}\Delta_\mu+O_p\left((N\pi_N)^{-1}\right)\\ &= \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}N_{-k}^{-1}\left(\sum_{i=1}^{N}\frac{R_i-\pi_N}{\pi_N}-\sum_{i\in\mathcal{I}_k}\frac{R_i-\pi_N}{\pi_N}\right)\Delta_\mu+O_p\left((N\pi_N)^{-1}\right)\\ &= \left\{N_{-k}^{-1}-(\mathbb{K}N_{-k})^{-1}\right\}\sum_{i=1}^{N}\frac{R_i-\pi_N}{\pi_N}\Delta_\mu+O_p\left((N\pi_N)^{-1}\right)\\ &\overset{(ii)}{=}N^{-1}\sum_{i=1}^{N}\mathrm{IF}_\pi(\mathbf{Z}_i)+O_p\left((N\pi_N)^{-1}\right),\end{aligned}$$

where $\mathrm{IF}_\pi(\mathbf{Z})=(\pi_N^{-1}R-1)\Delta_\mu$. Here, (i) holds by definition that $\widehat{\pi}_N(S_{-k})=N_{-k}^{-1}\sum_{i\in\mathcal{I}_{-k}}R_i$ and (ii) follows by the fact that $N_{-k}^{-1}-(\mathbb{K}N_{-k})^{-1}=\mathbb{K}/\{(\mathbb{K}-1)N\}-1/\{(\mathbb{K}-1)N\}=N^{-1}$. ∎

Proof of Theorem 2.8. Since $\pi_N(\mathbf{X}) > c\pi_N$, we have

$$a_N = 1/E\{\pi_N^{-1}(\mathbf{X})\} \geq c^{-1}\pi_N.$$

Additionally, by Jensen's inequality, $a_N \leq \pi_N$. Hence, $a_N \asymp \pi_N$. For each $k \leq \mathbb{K}$, define the following event

$$\mathcal{E}_{-k} := \{\widehat{\pi}_N(\mathbf{x}; \mathbb{S}_{-k}) < 2C\pi_N, \ \ \forall \mathbf{x} \in \mathcal{X}\}.$$

Then, under conditions $e_N(\mathbf{X}) < C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$, (2.42), and $r_{e,N} = o(1)$, we have

$$P_{\mathbb{S}_{-k}}(\mathcal{E}_{-k}) \geq P\left(\sup_{x\in\mathcal{X}} \left| \frac{\widehat{\pi}_N(\mathbf{x}; \mathbb{S}_{-k}) - e_N(\mathbf{X})}{\pi_N} \right| > C\right) = 1 - o(1). \tag{2.124}$$

Recall that $\varepsilon = Y - Rm_1(\mathbf{X}) - (1-R)m_0(\mathbf{X})$. We have $E(\varepsilon|R, \mathbf{X}) = 0$. Observe that

$$\hat{\theta}^0 - \theta^0 = N^{-1}\sum_{i=1}^{N} \psi_0(\mathbf{Z}_i) + \sum_{k=1}^{\mathbb{K}} (\widehat{\Delta}'_{N,1,k} + \widehat{\Delta}'_{N,2,k} + \widehat{\Delta}'_{N,3,k} + \widehat{\Delta}'_{N,4,k} + \widehat{\Delta}'_{N,5,k}),$$

where

$$\begin{aligned}
\psi_0(\mathbf{Z}) &= \mu_0(\mathbf{X}) - \theta_0 + \frac{1-R}{1-e_N(\mathbf{X})}\{Y - \mu_0(\mathbf{X})\} \\
&= \frac{e_N(\mathbf{X}) - R}{1-e_N(\mathbf{X})}\{m_0(\mathbf{X}) - \mu_0(\mathbf{X})\} + m_0(\mathbf{X}) - \theta^0 + \frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}, \\
\widehat{\Delta}'_{N,1,k} &= -N^{-1}\sum_{i\in\mathcal{I}_k} \left\{ \frac{1-R_i}{1-\pi_N(\mathbf{X}_i)} - 1 \right\} \{\widehat{m}_0(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu_0(\mathbf{X}_i)\}, \\
\widehat{\Delta}'_{N,2,k} &= N^{-1}\sum_{i\in\mathcal{I}_k} \left\{ \frac{1-R_i}{1-\widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})} - \frac{1-R_i}{1-e_N(\mathbf{X}_i)} \right\} \{Y_i - m_0(\mathbf{X}_i)\}, \\
\widehat{\Delta}'_{N,3,k} &= -N^{-1}\sum_{i\in\mathcal{I}_k} \left\{ \frac{1-R_i}{1-\widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})} - \frac{1-R_i}{1-e_N(\mathbf{X}_i)} \right\} \{\widehat{m}_0(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu_0(\mathbf{X}_i)\}, \\
\widehat{\Delta}'_{N,4,k} &= N^{-1}\sum_{i\in\mathcal{I}_k} \left\{ \frac{1-R_i}{1-\widehat{\pi}_N(\mathbf{X}_i; \mathbb{S}_{-k})} - \frac{1-R_i}{1-e_N(\mathbf{X}_i)} \right\} \{m_0(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)\}, \\
\widehat{\Delta}'_{N,5,k} &= N^{-1}\sum_{i\in\mathcal{I}_k} \left\{ \frac{1-R_i}{1-\pi_N(\mathbf{X}_i)} - \frac{1-R_i}{1-e_N(\mathbf{X}_i)} \right\} \{\widehat{m}_0(\mathbf{X}_i; \mathbb{S}_{-k}) - \mu_0(\mathbf{X}_i)\}.
\end{aligned}$$

We first obtain the rates for the terms $N^{-1}\sum_{i=1}^{N}\psi_0(\mathbf{Z}_i)$, $\widehat{\Delta}'_{N,1,k}$, and $\widehat{\Delta}'_{N,2,k}$ for each $k \leq \mathbb{K}$. Observe the following properties for the first moments:

$$\begin{aligned} E\{\psi_0(\mathbf{Z})\} &= E\left[\frac{e_N(\mathbf{X})-\pi_N(\mathbf{X})}{1-e_N(\mathbf{X})}\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}\right] \\ &= \mathbb{1}\{e_N(\cdot)\neq\pi_N(\cdot),\ \mu(\cdot)\neq m(\cdot)\}O_p(\pi_N), \\ E_{\mathbb{S}_k}(\widehat{\Delta}'_{N,1,k}) &= E_{\mathbb{S}_k}(\widehat{\Delta}'_{N,2,k}) = 0. \end{aligned}$$

For the second moments, we have

$$\begin{aligned} \mathrm{Var}\{\psi_0(\mathbf{Z})\} &= \mathrm{Var}\left[\frac{\{e_N(\mathbf{X})-R\}\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}}{1-e_N(\mathbf{X})}+m_0(\mathbf{X})-\theta_0+\frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}\right] \\ &\overset{(i)}{=} \mathrm{Var}\left[\frac{\{e_N(\mathbf{X})-R\}\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}}{1-e_N(\mathbf{X})}+m_0(\mathbf{X})-\theta_0\right]+\mathrm{Var}\left\{\frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}\right\} \\ &\leq \left\|\frac{\{e_N(\mathbf{X})-R\}\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}}{1-e_N(\mathbf{X})}+m_0(\mathbf{X})-\theta_0\right\|_{2,P}^2+\left\|\frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}\right\|_{2,P}^2 \\ &\leq 2\left\|\frac{\{e_N(\mathbf{X})-R\}\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}}{1-e_N(\mathbf{X})}\right\|_{2,P}+2\left\|m_0(\mathbf{X})-\theta_0\right\|_{2,P}^2+\left\|\frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}\right\|_{2,P}^2 \\ &\overset{(ii)}{=} 2E\left(\frac{[\{e_N(\mathbf{X})-\pi_N(\mathbf{X})\}^2+\pi_N(\mathbf{X})\{1-\pi_N(\mathbf{X})\}]\{m_0(\mathbf{X})-\mu_0(\mathbf{X})\}^2}{\{1-e_N(\mathbf{X})\}^2}\right) \\ &\quad +2\left\|m_0(\mathbf{X})-\theta_0\right\|_{2,P}^2+\left\|\frac{\varepsilon(1-R)}{1-e_N(\mathbf{X})}\right\|_{2,P}^2 \\ &\overset{(iii)}{\leq} 2\left[(1-C\pi_N)^{-2}\{(2C\pi_N)^2+C\pi_N\}+1\right]\left\|m_0(\mathbf{X})-\theta_0\right\|_{2,P}^2+(1-C\pi_N)^{-2}\|\varepsilon\|_{2,P}^2 \\ &= O(1). \end{aligned}$$

where (i) holds by the fact that $E(\varepsilon|R,\mathbf{X})=0$, (ii) holds by the tower rule with $E(R|\mathbf{X})=\pi_N(\mathbf{X})$, and (iii) follows by the assumption that $\pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$ for all $\mathbf{x}\in\mathcal{X}$. Besides,

$$\begin{aligned} E_{\mathbb{S}_k}(\widehat{\Delta}'^2_{N,1,k}) &= N^{-2}|\mathcal{I}_k|E\left[\left\{\frac{\pi_N(\mathbf{X})-R}{1-\pi_N(\mathbf{X})}\right\}^2\{\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k})-\mu_0(\mathbf{X})\}^2\right] \\ &= N^{-2}|\mathcal{I}_k|E\left[\frac{\pi_N(\mathbf{X})}{1-\pi_N(\mathbf{X})}\{\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k})-\mu_0(\mathbf{X})\}^2\right] \\ &\overset{(i)}{\leq} N^{-1}(1-C\pi_N)^{-1}C\pi_N\|\widehat{m}_0(\cdot;\mathbb{S}_{-k})-\mu_0(\cdot)\|_{2,P_X}^2 = O_p(N^{-1}\pi_N r_{\mu,0,N}^2), \end{aligned}$$

where (i) holds by the fact that $\pi_N(\mathbf{x}) < C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$. On the event $\mathcal{E}_{-k}$, with (2.124), we have

$$
\begin{aligned}
E_{\mathbb{S}_k}(\widehat{\Delta}'^2_{N,2,k}) &= N^{-2}|\mathcal{I}_k| E\left[\left\{\frac{1-R}{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{1-R}{1-e_N(\mathbf{X})}\right\}^2 \{Y - m_0(\mathbf{X})\}^2\right] \\
&= N^{-2}|\mathcal{I}_k| E\left[\frac{\{1-\pi_N(\mathbf{X})\}\{\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k}) - e_N(\mathbf{X})\}^2\varepsilon^2}{\{1-e_N(\mathbf{X})\}^2\{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})\}^2}\right] \\
&\overset{(i)}{\leq} N^{-1}(1-2C\pi_N)^{-4}(1-c\pi_N)\pi_N^2 \sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\pi}_N(\mathbf{x};\mathbb{S}_{-k}) - e_N(\mathbf{x})}{\pi_N}\right|^2 \|\varepsilon\|^2_{2,P} = O_p(N^{-1}\pi_N^2 r_{e,N}^2),
\end{aligned}
$$

where (i) holds by the fact that $c\pi_N < \pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$ and $\widehat{\pi}_N(\mathbf{x};\mathbb{S}_{-k}) < 2C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$ on $\mathcal{E}_{-k}$. Here, if we fix (or conditional on) $\mathbb{S}_{-k}$, on the event $\mathcal{E}_{-k}$, the inequality $\widehat{\pi}_N(\mathbf{x};\mathbb{S}_{-k}) < 2C\pi_N$ holds almost surely, w.r.t. the probability measure P; if $\mathbb{S}_{-k}$ is treated as random, recall (2.124), the inequality holds w.p.a. 1 , w.r.t. the joint probability measre of P and $P_{\mathbb{S}_{-k}}$. As a result, we have $E_{\mathbb{S}_k}(\widehat{\Delta}'^2_{N,2,k}) = O_p(N^{-1}\pi_N^2 r_{e,N}^2)$ w.r.t. the joint probability measure of P and $P_{\mathbb{S}_{-k}}$. By Lemma 2.4,

$$
\begin{aligned}
N^{-1}\sum_{i=1}^N \psi_0(\mathbf{Z}_i) &= \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot),\ \mu(\cdot) \neq m(\cdot)\} O_p(\pi_N) + O_p(N^{-1/2}), \\
\widehat{\Delta}'_{N,1,k} &= O_p(N^{-1/2}\pi_N^{1/2} r_{\mu,0,N}), \\
\widehat{\Delta}'_{N,2,k} &= O_p(N^{-1/2}\pi_N r_{e,N}).
\end{aligned}
$$

Now we consider the terms $\widehat{\Delta}_{N,3,k}$, $\widehat{\Delta}_{N,4,k}$, and $\widehat{\Delta}_{N,5,k}$. On the event $\mathcal{E}_{-k}$, we have

$$
\begin{aligned}
E_{\mathbb{S}_k}|\widehat{\Delta}_{N,3,k}| &\overset{(i)}{\leq} N^{-1}|\mathcal{I}_k| E\left\{\left|\frac{1-R}{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{1-R}{1-e_N(\mathbf{X})}\right| |\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k}) - \mu_0(\mathbf{X})|\right\} \\
&\overset{(ii)}{=} N^{-1}|\mathcal{I}_k| E\left\{\frac{\{1-\pi_N(\mathbf{X})\}|\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k}) - e_N(\mathbf{X})|}{\{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})\}\{1-e_N(\mathbf{X})\}} |\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k}) - \mu_0(\mathbf{X})|\right\} \\
&\overset{(iii)}{\leq} \frac{1-c\pi_N}{(1-2C\pi_N)^2}\pi_N \sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\pi}_N(\mathbf{x};\mathbb{S}_{-k}) - e_N(\mathbf{x})}{\pi_N}\right| \|\widehat{m}_0(\cdot;\mathbb{S}_{-k}) - \mu_0(\cdot)\|_{2,P_X} \\
&= O_p(\pi_N r_{e,N} r_{\mu,0,N}),
\end{aligned}
$$

where (i) holds by the triangular inequality, (ii) follows by the tower rule with the fact that $E(R|\mathbf{X}) = \pi_N(\mathbf{X})$, and (iii) holds by the fact that $c\pi_N < \pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$ and $\widehat{\pi}_N(\mathbf{x}; \mathbb{S}_{-k}) < 2C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$ on $\mathcal{E}_{-k}$. Similarly, on the event $\mathcal{E}_{-k}$,

$$\begin{aligned}
E_{\mathbb{S}_k}|\widehat{\Delta}_{N,4,k}| &\leq N^{-1}|\mathcal{I}_k| E\left\{\left|\frac{1-R}{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{1-R}{1-\pi_N(\mathbf{X})}\right| |m_0(\mathbf{X}) - \mu_0(\mathbf{X})|\right\} \\
&= N^{-1}|\mathcal{I}_k| E\left\{\frac{\{1-\pi_N(\mathbf{X})\}|\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k}) - \pi_N(\mathbf{X})|}{\{1-\widehat{\pi}_N(\mathbf{X};\mathbb{S}_{-k})\}\{1-e_N(\mathbf{X})\}} |m_0(\mathbf{X}) - \mu_0(\mathbf{X})|\right\} \\
&\overset{(i)}{\leq} \frac{1-c\pi_N}{(1-2C\pi_N)^2}\pi_N \sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\widehat{\pi}_N(\mathbf{x};\mathbb{S}_{-k}) - e_N(\mathbf{x})}{\pi_N}\right| \|m_0(\cdot) - \mu_0(\cdot)\|_{2,P_X} \\
&= \mathbb{1}\{m_0(\cdot) \neq \mu_0(\cdot)\} O_p(\pi_N r_{e,N}),
\end{aligned}$$

where (i) holds by the assumption that $c\pi_N < \pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$ and $\widehat{\pi}_N(\mathbf{x}; \mathbb{S}_{-k}) < 2C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$ on $\mathcal{E}_{-k}$. Additionally, we also have

$$\begin{aligned}
E_{\mathbb{S}_k}|\widehat{\Delta}_{N,5,k}| &\leq N^{-1}|\mathcal{I}_k| E\left\{\left|\frac{1-R}{1-\pi_N(\mathbf{X};\mathbb{S}_{-k})} - \frac{1-R}{1-\pi_N(\mathbf{X})}\right| |\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k}) - \mu_0(\mathbf{X})|\right\} \\
&= N^{-1}|\mathcal{I}_k| E\left\{\frac{|\pi_N(\mathbf{X}) - e_N(\mathbf{X})|}{1-e_N(\mathbf{X})} |\widehat{m}_0(\mathbf{X};\mathbb{S}_{-k}) - \mu_0(\mathbf{X})|\right\} \\
&\leq (1-C\pi_N)^{-1} \sup_{\mathbf{x}\in\mathcal{X}} |\pi_N(\mathbf{x}) - e_N(\mathbf{x})| \, \|\widehat{m}_0(\cdot;\mathbb{S}_{-k}) - \mu_0(\cdot)\|_{2,P_X} \\
&= \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\} O_p(\pi_N r_{\mu,0,N}),
\end{aligned}$$

since $\pi_N(\mathbf{x}), e_N(\mathbf{x}) < C\pi_N$ for all $\mathbf{x} \in \mathcal{X}$ by assumption. By Lemma 2.4,

$$\begin{aligned}
\widehat{\Delta}'_{N,3,k} &= O_p(\pi_N r_{e,N} r_{\mu,N}), \\
\widehat{\Delta}'_{N,4,k} &= \mathbb{1}\{m_0(\cdot) \neq \mu_0(\cdot)\} O_p(\pi_N r_{e,N}), \\
\widehat{\Delta}'_{N,5,k} &= \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\} O_p(\pi_N r_{\mu,0,N}).
\end{aligned}$$

Therefore,

$$\widehat{\theta}^0_{\text{DRSS}} - \theta^0 = N^{-1}\sum_{i=1}^{N}\psi_0(\mathbf{Z}_i) + O_p(N^{-1/2}\pi_N^{1/2}r_{\mu,0,N} + N^{-1/2}\pi_N r_{e,N} + \pi_N r_{e,N}r_{\mu,0,N})$$
$$+ \mathbb{1}\{m_0(\cdot) \neq \mu_0(\cdot)\}O_p(\pi_N r_{e,N}) + \mathbb{1}\{e_N(\cdot) \neq \pi_N(\cdot)\}O_p(\pi_N r_{\mu,0,N}).$$

■

Corollary 2.1 is a direct consequence of Theorems 2.2 and 2.8.

2.10 Acknowledgement

Chaper 2, in full, has been submitted for publication of the material. Zhang, Yuqian; Chakrabortty, Abhishek; Bradic, Jelena. Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. The dissertation author was one of the primary investigators and authors of this material.

Chapter 3

High-dimensional inference for dynamic treatment effects

3.1 Introduction

The complexity of a certain disease or economic policy is often reflected by the diversity and the size of the personal characteristics of each individual or economy at hand, consequently inducing strong heterogeneity in the observations. On the other hand, access to randomized control trials, especially over time, has become overly restrictive, often due to various costs or ethical concerns. Access to time-varying observational studies has, however, exploded recently. Data-driven decisions span daily life or almost every individual: from continuous measurements of individuals' health on mobile devices and medical decisions made as a result of that to the monitoring of individuals' online presence or daily measuring of the economic and social policies introduced to better the public health of each individual. Studying the true treatment or policy effect has therefore become that much more compli-

cated. This chapter brings to the literature a way to construct confidence intervals about dynamic treatment effects in the presence of high-dimensional observations.

Given a sequence of binary treatment assignments or policy interventions, $A_1, A_2, \dots$, and an outcome of interest, $Y \in \mathbb{R}$, a collection of possibly high-dimensional, sequential (pre-treatment) covariates $\mathbf{S}_1, \mathbf{S}_2, \dots$ is also observed. We seek to estimate how these covariates regulate and modify the effect of the multiple time-varying treatments on the outcome of interest. Covariates, collected over multiple exposure times, are not required to have the same variables observed at each exposure: $\mathbf{S}_1 \in \mathbb{R}^{d_1}, \mathbf{S}_2 \in \mathbb{R}^{d_2}, \dots$. Potential or counterfactual outcomes, $Y(a_1, a_2, \dots)$, denote participant's outcome had he or she followed a specific treatment (sequence), $a_1, a_2, \dots$, which is possibly different from the treatment he or she was observed with. For a given treatment path of interest $a = (a_1, a_2, \dots)$ and its corresponding control $a' = (a'_1, a'_2, \dots)$, we are interested in understanding $E[Y(a) - Y(a')]$.

Average treatment effects (ATE) in the presence of multiple exposure times have been a longstanding problem of interest. Difficulties with studying treatment effects over time are numerous. Previous treatments may affect the distribution of future confounders, mediators, and treatment choices. In these settings, more traditional approaches, such as generalized estimating equations or random effects models, are not guaranteed to lead to a consistent estimation. Here, adjustment for confounders may have no causal interpretation, even if all confounders are measured, and the regression is correctly specified; see, e.g., [DCDS$^+$13]. Mimicking sequential [**?**] and sequential multiple randomized control trials (SMART, e.g., see [CM14]) became the gold standard; see, e.g., [HSHD$^+$16]. [CRL$^+$10] exemplified the need for inverse probability weighting (IPW) even if treatment probabilities are constants; the effects of the past treatment probabilities needed to be accounted for. Structural nested

mean (SNM) models and marginal structural mean (MSM) models have been developed to handle these particular challenges, see, e.g., [Rob97] and [MvdLRG01] among others. G-computation [Rob86] has been used for the estimation and a vast literature has contributed to this topic; see, e.g., [HBR01, JYF10, vdLPJ05, VG03].

In this chapter, we focus on MSM models with continuous outcomes, binary treatments, and continuous covariates. Binary covariates are also possible, albeit their presence would indicate that one or more of the models are misspecified; see, e.g., the discussion in Section 4 of [BRR19]. We work under the sequential ignorability assumption and formalize the problem of a root-N confidence interval construction for identifying the presence of ATE for multi-stage observational experiments with time-varying treatment assignments and high-dimensional covariates. Here, due to the high-dimensional nature of the problem, unbiased estimation of the effects of the confounders at the root-N rate is not possible. Despite that, we are able to achieve a root-N consistent and asymptotically normal estimation of the average treatment effect where we would allow for Lasso shrinkage effects but do not assume standard asymptotics, i.e., the number of samples, N is much smaller than the number of the confounders (at any given time or in total).

We achieve this result by establishing a new, dynamic rate double robustness (RDR) suitable for dynamic treatment effects. RDR weakens reliance on stringent sparsity assumptions by offering an opportunity to avoid committing to two extremely sparse modeling assumptions – assumptions restricting the sparsity to be at a root-N level. This is, for a single treatment, reflected in a "product-rate" condition that is sufficient condition for guaranteeing asymptotic normality with high-dimensional confounding; see, e.g., Theorem 3.1 of [CCD$^+$18] or Theorem 1 of [SRR19]. For a setting with two exposure times, we iden-

tify two product rate conditions, each ensuring the RDR property of a single time period. This, in turn, results in three product rate conditions for the sparsity parameter of our high-dimensional models. The first two products correspond to the products of the sparsity of the outcome and its matching propensity at the same exposure time, whereas the third product considers the cross product between the exposures: sparsity of the propensity at the first exposure and sparsity of the outcome at the second exposure. More generally, if t denotes the exposure time and $s_{\mathrm{o},t}$ and $s_{\mathrm{p},t}$ denote the sparsity of the outcome and propensity model at the exposure time t, our product rate conditions are $s_{\mathrm{o},t}s_{\mathrm{p},t} = o(N/\log^2(d))$ and for every $1 \leq k \leq t-1$, $\sum_{j=k}^{t} s_{\mathrm{o},j}s_{\mathrm{p},k} = o(N/\log^2(d))$.

The dynamic treatment effect estimation with MSM models has also been studied recently in [BHL20]. They proposed a general RDR estimator, which requires three product rate conditions for the nuisance estimators. In contrast, we identify that only two of those are sufficient. Moreover, they did not provide any valid nuisance estimators, nor did they verify when their required consistency conditions hold. In fact, the estimation of one of the nuisance models, the outcome at the first exposure, is a non-trivial problem; see Remark 3.1. The theoretical advancements in this work hinder upon developing new estimation error bounds of independent interest for a Lasso estimator with imputed outcomes. We allow the imputation error to be dependent on the covariates and to be dependent across individuals. Some results on imputed Lasso have appeared previously [SFSL18, ZZS19]; however, with more restrictive settings and vastly different conditions. These results apply broadly across many different problems; see Section Theorem 3.1. Additionally, [LS20] provided estimators for the counterfactual mean (3.1) by relying on SNM models and g-estimation. However, the authors require the blip functions to be correctly specified at all times. Even when the

blip functions are linear, the authors therein obtain valid inference only in low dimensions. In contrast, Theorem 3.2 provides inference guarantees with high-dimensional confounders; Theorems 3.3 and 3.5 provide consistency as long as one, and not necessarily both, of the nuisance models is correctly specified at each time spot.

3.1.1 Related work

Our work fits into a growing literature on static average treatment effect estimation and inference, including but not limited to [CCD$^+$18, Tan20a, BWZ19, SRR19, DV20, DAV20]. Dynamic treatments should not be confused with static ones. The most common method of handling confounders of treatment effect is to adjust for them or by including all the variables in a regression model. In single-time treatment studies, such an adjustment may have causal interpretation in the absence of unmeasured confounding. In multiple time treatment studies (dynamic settings), the treatment changes over time, possibly in response to a change in the observed confounders. Here, regression adjustment will no longer have causal interpretation even if all confounders are observed, and the regression model is correctly specified. In addition, if one adjusts for the covariates by including them in traditional one-time models, even causal ones, the resulting estimate of the causal effect of treatment will not include the component of the causal effect mediated by the dynamic changes.

MSMs of [Rob97] emerged as a powerful tool in addressing the above concerns. Theoretical advancements of MSMs with low-dimensional confounders culminated in a seminal work of [TS12]. However, in the presence of high-dimensional covariates, inferential double robust questions are yet to be studied to the best of our knowledge. Some approaches towards

covariate balancing in MSMs have been discussed in [ZW20, VB21, RS19]. However, the approach strongly depends on the validity of the sequential mean models that we specifically relax in this work. We should also mention the IPW approaches of [BS19, BRS21].

A closely related literature is that of optimal treatment allocation and methods based on Q, A, or R -learning, including [Mur03, Rob04, ORR10, ZTD+12, CZW21]. These approaches are helpful when dealing with dynamic treatments, however, the authors' primary concern is not confidence interval construction or efficient estimation of the treatment effect itself. Confidence intervals on the selected treatment rule have also been considered; see, e.g., [CMS10, LLQ+14]. A form of a doubly robust property has been studied in the context of A-learning; see, e.g., [SFSL18]. The contrast function's estimator is consistent as long as either the baseline mean or the propensity score function is correctly specified. However, to consistently estimate the first-stage contrast, the second-stage contrast needs to be correctly specified – such a condition is not required in our work.

Lastly, our work has a connection to the ever-expanding work on high-dimensional inference; see, e.g., [ZZ14,VdGBRD14,BCK15,RWG19,ZB18]. Although they bare similarity in treating sparsity and regularization, the authors estimate a very different parameter of interest – a coefficient in the regression model. To that end, they utilize distinct approaches to resolve the bias issue induced by the regularization and nominal shrinkage effects.

3.1.2 Notation

For any $\alpha > 0$, let $\psi_\alpha(\cdot)$ denote the function given by $\psi_\alpha(x) := \exp(\alpha^2) - 1$, $\forall x > 0$. Then, the ψ_α-Orlicz norm $\|\cdot\|_{\psi_\alpha}$ of a random variable X is defined as $\|X\|_{\psi_\alpha} := \inf\{c >$

$0 : E[\psi_\alpha(|X|/c)] \leq 1\}$. Two special cases of finite $\psi_\alpha-$Orlicz norm are given by $\psi_2(x) = \exp(x^2) - 1$ and $\psi_1(x) = \exp(x) - 1$, which correspond to sub-Gaussian and sub-exponential random variables, respectively. The notation $a_N \ll b_N$ denotes $a_N = o(b_N)$, and $a_N \gg b_N$ denotes $b_N \ll a_N$ as $N \to \infty$. The notation $a_N \asymp b_N$ denotes $cb_N \leq a_N \leq Cb_N$ for all $N \geq 1$ and with some constants $c, C > 0$. The notation $\mathbf{X}[j]$ denotes the j-th element of vector $\mathbf{X}$.

3.2 Causal effects in the interactive model

3.2.1 Model setting

Suppose that we have access to N i.i.d. observations $\{W_i\}_{i=1}^N = (Y_i, A_{1i}, A_{2i}, \mathbf{S}_{1i}, \mathbf{S}_{2i})_{i=1}^N$ following a distribution P. Let $W = (Y, A_1, A_2, \mathbf{S}_1, \mathbf{S}_2)$ be an independent copy of W_i; if $\{W_i\}_{i=1}^N$ are training data, then W is a single, new test data. Let $\mathbf{S}_t \in \mathbb{R}^{d_t}$ denote the covariates of the subject at the exposure time t, and $A_t \in \{0, 1\}$ denote the binary treatment taken at time t. At any time t, we assume that any treatment-specific variable can only be affected by the past treatments or past covariates; and not the future. This is sometimes called temporal ordering. Due to notational complications, we exemplify our ideas and results for two-stage trials, with observables $(\mathbf{S}_1, A_1, \mathbf{S}_2, A_2, Y)$, although the same theory and methods developed herein apply more broadly to multiple-stage trials.

A dynamic treatment assignment, denoted with $a = (a_1, a_2)$, $a_1, a_2 \in \{0, 1\}$ is a sequence of treatment rules applied to each treatment exposure time. We use the potential outcome framework to define the causal effect. $Y(a_1, a_2)$ denotes the potential outcome we would have obtained if the individual was exposed to the treatment sequence (a_1, a_2).

Throughout this work, we assume a "no interference" setting.

Our parameter of interest $\theta = E[Y(a)] - E[Y(a')]]$, with $a \neq a'$ and

$$\theta_a = E[Y(a)], \tag{3.1}$$

resulting in $\theta = \theta_a - \theta_{a'}$, is characterized by two population means and would have been identified had we observed both the outcome under treatment a as well as the one under treatment a'. In order to identify causal effects above, we make the standard assumptions of sequential ignorability, consistency, and overlap; see, e.g., [IR15a, LM05, Mur03, Rob00a, Rob87].

Assumption 3.1. *(i) (Sequential Ignorability)* $Y(a_1, a_2) \perp\!\!\!\perp A_1 \mid \mathbf{S}_1$ *and* $Y(a_1, a_2) \perp\!\!\!\perp A_2 \mid \mathbf{S}_1, \mathbf{S}_2, A_1 = a_1$. *(ii) (Consistency of potential outcomes)* $Y = Y(A_1, A_2)$. *(iii) (Overlap) Let* $c_0 \in (0, 1)$ *be a positive constant, such that*

$$P(c_0 \leq \pi(\mathbf{S}_1) \leq 1 - c_0) = 1, \quad P(c_0 \leq \rho_a(\mathbf{S}_1, \mathbf{S}_2) \leq 1 - c_0) = 1,$$

where the treatment assignments (propensity scores) are defined as

$$\pi(\mathbf{s}_1) := P[A_1 = a_1 | \mathbf{S}_1 = \mathbf{s}_1], \tag{3.2}$$

$$\rho_a(\mathbf{s}_1, \mathbf{s}_2) := P[A_2 = a_2 | \mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2 = \mathbf{s}_2, A_1 = a_1]. \tag{3.3}$$

Assumption 3.1 (i) is also known as "exchangeability" or "sequential randomization" or "no unmeasured confounding". It states that the observed treatment at time t is independent of the potential outcomes given all the data observed prior to the exposure time t. Assumptions are standard and sufficient to identify the parameter of interest based on the observed data. Under Assumption 3.1 (i) and (ii), we have

$$\theta_a = E\left[\frac{\mathbb{1}_{\{A_1=a_1, A_2=a_2\}} Y}{\pi(\mathbf{S}_1)\rho_a(\mathbf{S}_1, \mathbf{S}_2)}\right].$$

3.2.2 Doubly Robust Estimator

We estimate $\theta_a = E[Y(a)]$, (3.1), by utilizing a doubly robust score $\psi_a(\cdot;\cdot)$ defined as

$$\psi_a(W;\eta_a) := \mu_a(\mathbf{S}_1) + \tau_a(\mathbf{S}_1)\Big(\nu_a(\mathbf{S}_1,\mathbf{S}_2) - \mu_a(\mathbf{S}_1)\Big) + \omega_a(\mathbf{S}_1,\mathbf{S}_2)\Big(Y - \nu_a(\mathbf{S}_1,\mathbf{S}_2)\Big), \quad (3.4)$$

as seen in, e.g., [NBW21, TYWK$^+$19, vdLG11, ORR10, MvdLRG01]. With a slight abuse of notation, we denote with $\eta_a(\cdot) := (\mu_a(\cdot), \nu_a(\cdot), \pi(\cdot), \rho_a(\cdot))$ the true nuisance parameters. Additionally, $\tau_a(\mathbf{s}_1)$ and $\omega_a(\mathbf{s}_1,\mathbf{s}_2)$ denote the population inverse probability weights, where

$$\tau_a(\mathbf{s}_1) := \mathbb{1}_{\{A_1=a_1\}}\pi^{-1}(\mathbf{s}_1), \qquad \omega_a(\mathbf{s}_1,\mathbf{s}_2) := \mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\pi^{-1}(\mathbf{s}_1)\rho_a^{-1}(\mathbf{s}_1,\mathbf{s}_2). \quad (3.5)$$

Double robust representation $\theta_a = E[\psi_a(W;\eta_a)]$ hinders upon two outcome models,

$$\nu_a(\mathbf{s}_1,\mathbf{s}_2) := E[Y|\mathbf{S}_1=\mathbf{s}_1,\mathbf{S}_2=\mathbf{s}_2,A_1=a_1,A_2=a_2], \quad (3.6)$$

$$\mu_a(\mathbf{s}_1) := E[\nu_a(\mathbf{S}_1,\mathbf{S}_2)|\mathbf{S}_1=\mathbf{s}_1,A_1=a_1]. \quad (3.7)$$

Here, $\nu_a(\mathbf{s}_1,\mathbf{s}_2)$ represents the conditional mean outcome model at the second exposure time, and $\mu_a(\mathbf{s}_1)$ is a nested conditional mean outcome model at the first exposure time. It follows from Theorem 3.2 of [Rob97] that, under the Sequential Ignorability and Consistency of the potential outcomes (see Assumption 3.1) the above nested outcome models can be identified as

$$\nu_a(\mathbf{s}_1,\mathbf{s}_2) = E[Y(a_1,a_2)|\mathbf{S}_1=\mathbf{s}_1,\mathbf{S}_2=\mathbf{s}_2,A_1=a_1], \quad \mu_a(\mathbf{s}_1) = E[Y(a_1,a_2)|\mathbf{S}_1=\mathbf{s}_1].$$

The idea of nested models is not new; see, e.g., [BRR19] for a review. With $\theta_a = E[\psi_a(W;\eta_a)]$, we estimate θ_a as

$$\widehat{\theta}_a := \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\mu}_a(\mathbf{S}_{1i}) + \widehat{\tau}_a(\mathbf{S}_{1i})\left(\widehat{\nu}_a(\mathbf{S}_{1i},\mathbf{S}_{2i}) - \widehat{\mu}_a(\mathbf{S}_{1i})\right) + \widehat{\omega}_a(\mathbf{S}_{1i},\mathbf{S}_{2i})\left(Y_i - \widehat{\nu}_a(\mathbf{S}_{1i},\mathbf{S}_{2i})\right)\right],$$

Algorithm 1 Dynamic ATE

Require: Observations $\{Y_i, \mathbf{S}_{1i}, A_{1i}, \mathbf{S}_{2i}, A_{2i}\}_{i=1}^N$.

Require: Treatment path $a = (a_1, a_2)$ and a control path $a' = (a_1', a_2')$.

For any fixed integer $K \geq 2$, split the indices $I = \{1, 2, ..., N\}$ into K equal-sized parts $\{I_k\}_{k=1}^K$ randomly, such that the size of each fold I_k is $n := N/K$. Define $I_{-k} := I \backslash I_k$.

for $c \in \{a, a'\}$ **do**

  for $k \in \{1, \cdots, K\}$ **do**

    Let $\mathcal{I}$ be a subset of indices of I_{-k} with the same treatment path as $c = (c_1, c_2)$.

    Let $\mathcal{I}_1$ be a subset of indices of I_{-k} with the same treatment path as c_1 only;

    Construct $\widehat{\nu}_c$ using $\mathcal{I}$ samples. ▷ *Outcome for time two*

    Construct $\widehat{\mu}_c$ using $\mathcal{I}_1$ samples. ▷ *Outcome for time one*

    Construct $\widehat{\rho}_c$ using $\mathcal{I}_1$ samples. ▷ *Propensity for time two*

    Construct $\widehat{\pi}$ using I_{-k} samples. ▷ *Propensity for time one*

    Let $\widehat{\eta}_c := (\widehat{\mu}_c, \widehat{\nu}_c, \widehat{\pi}, \widehat{\rho}_c)$, $\widehat{\tau}_c = \mathbb{1}_{\{A_1 = a_1\}} \widehat{\pi}^{-1}$, and $\widehat{\omega}_c = \mathbb{1}_{\{A_1 = a_1, A_2 = a_2\}} \widehat{\pi}^{-1} \widehat{\rho}_c^{-1}$.

    For $\psi_c(W; \eta_c)$, (3.4), construct a cross-fitted estimator $\check{\theta}_c^{(k)}$ as $\check{\theta}_c^{(k)} = \frac{1}{n} \sum_{i \in I_k} \psi_c\big(W_i; \widehat{\eta}_c\big)$.

  end for

  $\widehat{\theta}_c = \sum_{k=1}^K \check{\theta}_c^{(k)} / K$.

end for

return The dynamic treatment effect estimator $\widehat{\theta} = \widehat{\theta}_a - \widehat{\theta}_{a'}$.

where $\widehat{\nu}_a(\cdot)$, $\widehat{\mu}_a(\cdot)$, $\widehat{\tau}_a(\cdot)$, $\widehat{\omega}_a(\cdot)$ are estimators of $\nu_a(\cdot)$, $\mu_a(\cdot)$, $\tau_a(\cdot)$, $\omega_a(\cdot)$ as defined in (3.6), (3.7), and (3.5), respectively.

The above equation avoids complicated notations needed for a cross-fitting procedure

we propose; see Algorithm 1 for more details. The above estimator is an innate generalization of the augmented inverse propensity score estimator of [RRZ94] for the static case. In this chapter, we study its properties in the presence of high-dimensional confounders.

3.3 Dynamic Treatment Lasso (DTL)

To simplify the exposition, we begin by listing some shorthand notations used throughout the following sections of the chapter. We define the dimension of all of the observed covariates at the second exposure time with d, i.e., $d := d_1 + d_2$. We let $\mathbf{U} := (1, \mathbf{S}_1^T, \mathbf{S}_2^T)^T$ denote $(d+1)$-dimensional observed covariates collecting both time one and time two. We denote with $\mathbf{V} := (1, \mathbf{S}_1^T)^T$ $(d_1 + 1)$-dimensional observed covariates of the first exposure time. In the following it is important to follow the individuals with pre-specified treatment plan. For that purpose we introduce the following shorthand notation: $\widetilde{Y}_a := Y \mathbb{1}_{\{(A_1,A_2)=a\}}$, $\widetilde{\mathbf{U}}_a := \mathbf{U}\mathbb{1}_{\{(A_1,A_2)=a\}}$ denote the outcome and the covariates of those individuals which have taken the treatment path a, i.e., whose $(A_1, A_2) = a$. Additionally, we use $\bar{Y} := Y \mathbb{1}_{\{A_1=a_1\}}$, $\bar{\mathbf{U}}_a := \mathbf{U}\mathbb{1}_{\{A_1=a_1\}}$, $\bar{\mathbf{V}}_a := \mathbf{V}\mathbb{1}_{\{A_1=a_1\}}$ to denote the outcome and the covariates at time two and time one, respectively, of those individuals which have taken the treatment a_1, i.e., whose $A_1 = a_1$, regardless of which treatment they have received in the second time period. Where possible, we suppress the sub-index a.

3.3.1 Outcome Models

Below we discuss estimation of the two outcome models ν_a, (3.6), and μ_a, (3.7), and we proceed sequentially; estimation at the latter exposure time, ν_a, is discussed first and

later used for the estimation at the earlier exposure, μ_a.

A linear working model is used to estimate ν_a, (3.6), i.e., $E[Y|\mathbf{S}_1, \mathbf{S}_2, A_1 = a_1, A_2 = a_2]$. The best linear working model, or the best linear approximation, is denoted with

$$\nu_a^*(\mathbf{s}_1, \mathbf{s}_2) = \mathbf{u}^T \boldsymbol{\alpha}_a^*, \tag{3.8}$$

where, for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$, $\mathbf{s}_2 \in \mathbb{R}^{d_2}$, $\mathbf{u} = (1, \mathbf{s}_1^T, \mathbf{s}_2^T)^T$. To motivate the proposed working model, we define

$$\boldsymbol{\alpha}_a^* := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{d+1}} E\left[\widetilde{Y} - \widetilde{\mathbf{U}}^T \boldsymbol{\alpha}\right]^2 = \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E[\widetilde{\mathbf{U}}\widetilde{Y}]. \tag{3.9}$$

The corresponding population residual, ζ_a, can be defined as

$$\zeta_a := \widetilde{Y} - \widetilde{\mathbf{U}}^T \boldsymbol{\alpha}^*. \tag{3.10}$$

It should be noted that, under the misspecified setting, there is no independence assumption between $\widetilde{\mathbf{U}}$ and ζ_a, and $E(\zeta_a|\widetilde{\mathbf{U}}) \neq 0$ is allowed.

Similarly, a linear working model is used to estimate the nested mean μ_a, (3.7). First, we observe that $\mu_a(\mathbf{S}_1) = E[\bar{\mathbf{U}}^T \boldsymbol{\alpha}_a^* | \mathbf{S}_1]$ and henceforth denote the best linear model for μ_a as

$$\mu_a^*(\mathbf{s}_1) = \mathbf{v}^T \boldsymbol{\beta}_a^*, \tag{3.11}$$

where for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$, $\mathbf{v} = (1, \mathbf{s}_1^T)^T$. To motivate the proposed working model, we define

$$\boldsymbol{\beta}_a^* = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{d_1+1}} E[\bar{\mathbf{U}}^T \boldsymbol{\alpha}_a^* - \bar{\mathbf{V}}^T \boldsymbol{\beta}]^2 = [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1} E[\bar{\mathbf{V}}\bar{\mathbf{U}}^T] \boldsymbol{\alpha}_a^* \tag{3.12}$$

as the best population slope for $E[\bar{\mathbf{U}}^T \boldsymbol{\alpha}_a^* | \bar{\mathbf{V}}]$. Note that the definition of $\boldsymbol{\beta}_a^*$ only depends on a_1, and is independent of a_2. To simplify the notation, we use $\boldsymbol{\beta}_a^*$ instead of $\boldsymbol{\beta}_{a_1}^*$. See

Remark 3.1 below on the reasons why we cannot use $\bar{Y}$ and $\bar{\mathbf{V}}$ directly to estimate μ_a. The corresponding population residual ε_a can be defined as

$$\varepsilon_a := \bar{\mathbf{U}}^T \boldsymbol{\alpha}_a^* - \bar{\mathbf{V}}^T \boldsymbol{\beta}_a^*. \tag{3.13}$$

Lastly, under model misspecification, we consider the case of $E[\varepsilon_a|\bar{\mathbf{V}}] \neq 0$.

Identification As nested models may be difficult to interpret, we provide a set of examples and discussions illustrating their correctness and identification. More has been said about this throughout the literature; see, e.g., [BRR19].

Remark 3.1 (Estimation of μ_a)**.** *To estimate the nuisance function $\mu_a(\mathbf{S}_1) = E[Y(a)|\mathbf{S}_1]$, the most natural method would be to regress $Y(a)$ on $\mathbf{S}_1$ for those observed $Y(a)$ whose $(A_1, A_2) = a$. However, under the Sequential Ignorability of Assumption 3.1,*

$$E[Y(A_1, A_2)|\mathbf{S}_1, A_1 = a_1, A_2 = a_2] = E[Y(a)|\mathbf{S}_1, A_1 = a_1, A_2 = a_2] \neq E[Y(a)|\mathbf{S}_1],$$

since in general, $Y(a) \not\perp\!\!\!\perp A_2|\mathbf{S}_1$.

Remark 3.2 (Model Misspecification)**.** *We illustrate when will the two working outcome models, $\nu_a^*(\cdot)$ and $\mu_a^*(\cdot)$, be correctly specified. If model $\nu_a^*(\cdot)$ is misspecified, then the model $\mu_a^*(\cdot)$ is also very likely to be misspecified, but there are no guarantees either way. A few comments are in order as the relationship between the two nested models is often masked. The following four cases are of potential interest.*

(i) If we assume that the true outcome model, $\nu_a(\cdot)$ is linear in that

$$\nu_a(\mathbf{S}_1, \mathbf{S}_2) = E[Y(a)|\mathbf{S}_1, \mathbf{S}_2, A_1 = a_1, A_2 = a_2] = \mathbf{U}^T \boldsymbol{\alpha}_a \tag{3.14}$$

holds for some vector $\boldsymbol{\alpha}_a \in \mathbb{R}^{d+1}$*, then it follows that* $\boldsymbol{\alpha}_a^* = \boldsymbol{\alpha}_a$ *and hence* $\nu_a^*(\cdot) = \nu_a(\cdot)$*, i.e.,* $\nu_a^*(\cdot)$ *is correctly specified.*

(ii) Otherwise, if we assume that (only) the true outcome model, $\mu_a(\cdot)$*, is linear in that*

$$\mu_a(\mathbf{S}_1) = E[Y(a)|\mathbf{S}_1, A_1 = a_1] = \mathbf{V}^T\boldsymbol{\beta}_a \tag{3.15}$$

holds for some vector $\boldsymbol{\beta}_a \in \mathbb{R}^{d_1+1}$*, then it is possible that the working model is still not linear, i.e.,* $\mu_a^*(\cdot) \neq \mu_a(\cdot)$ *making* $\mu_a^*(\cdot)$ *potentially misspecified.*

(iii) Now, if the true outcome model (3.15) *holds and in addition* $\boldsymbol{\alpha}_a^*$*,* (3.9)*, is equal to* $\bar{\boldsymbol{\alpha}}_a^*$*, with* $\bar{\boldsymbol{\alpha}}_a^*$ *defined as*

$$\bar{\boldsymbol{\alpha}}_a^* := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{d+1}} E\left[(Y(a) - \mathbf{U}^T\boldsymbol{\alpha})^2|A_1 = a_1\right] = \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1} E[\bar{\mathbf{U}}Y(a)],$$

then, we have $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}_a$ *and* $\mu_a^*(\cdot) = \mu_a(\cdot)$*, i.e.,* $\mu_a^*(\cdot)$ *is correctly specified.*

(iv) Lastly, if both of the true outcome models are linear, i.e., (3.14) *and* (3.15) *hold simultaneously, then, both* $\nu_a^*(\cdot)$ *and* $\mu_a^*(\cdot)$ *are correctly specified. Case (iv) is equivalent to requiring* $E(\mathbf{S}_2^T\boldsymbol{\alpha}_{a,2}|\boldsymbol{S}_1)$ *to be linear in* $\boldsymbol{S}_1$*; here,* $\boldsymbol{\alpha}_a = (\boldsymbol{\alpha}_{a,1}, \boldsymbol{\alpha}_{a,2})^T$ *where* $\boldsymbol{\alpha}_{a,1} \in \mathbb{R}^{d_1+1}$ *and* $\boldsymbol{\alpha}_{a,2} \in \mathbb{R}^{d_2}$*. This, in turn, occurs for any closed class of spherical distributions, including normal and Student-t distributions, or any linear time-series models of covariate dependence.*

Some discussions are provided below. We can see that the correctness of the model $\mu_a^*(\cdot)$ *also depends on* $\boldsymbol{\alpha}_a^*$*, the slope parameter of* $\nu_a^*(\cdot)$*. A true linear outcome model* $\mu_a(\cdot)$ *does not guarantee a correctly specified* $\mu_a^*(\cdot)$*; however, if the true outcome model* $\nu_a(\cdot)$ *is also linear, then* $\mu_a^*(\cdot)$ *is correctly specified. Moreover, a linear* $\nu_a(\cdot)$ *and* $\mu_a(\cdot)$ *are sufficient for a*

correctly specified $\nu_a^*(\cdot)$, but they are not required. Case (iii) provides an illustration where a correctly specified $\mu_a^*(\cdot)$ does not require a correctly specified $\nu_a^*(\cdot)$. This occurs, for example, whenever $\boldsymbol{\alpha}_a^* = \bar{\boldsymbol{\alpha}}_a^*$.

For an illustration, consider $a = (1,1)$ and $S_1, S_2, Z \sim^{\text{iid}} \text{Unif}(-1,1)$ with a nonlinear outcome model $\nu_a(\cdot)$, $Y(1,1) = S_1 + S_2^3 + Z$. Let the treatment assignments satisfy

$$\pi(s_1) = |s_1|, \text{ and } \rho_a(s_1, s_2) = \exp(s_1 + s_2)/\{1 + \exp(s_1 + s_2)\},$$

for all $s_1, s_2 \in \mathbb{R}$. Then, $\boldsymbol{\alpha}_a^* = \bar{\boldsymbol{\alpha}}_a^*$ and therefore guaranteeing correctness of the linear working model $\mu_a^*(\cdot)$. Here, $\pi^*(\cdot)$ and $\nu_a^*(\cdot)$ are misspecified, $\rho_a^*(\cdot)$ and $\mu_a^*(\cdot)$ are correctly specified.

Justifications Below are the justifications of the cases (i)-(iv) in Remark 3.2.

(i) Under Assumption 1 and by the law of iterated expectations, we have

$$\begin{aligned}
\boldsymbol{\alpha}_a^* &= \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E[\widetilde{\mathbf{U}}\widetilde{Y}] = \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1, A_2=a_2\}}\mathbf{U}Y(a)\right] \\
&= \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E\left[\mathbf{U}E\left[Y(a)|\mathbf{U}, A_1 = a_1, A_2 = a_2\right] P\left[A_1 = a_1, A_2 = a_2|\mathbf{U}\right]\right] \\
&= \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E\left[\mathbf{U}\mathbf{U}^T\boldsymbol{\alpha}_a E\left[\mathbb{1}_{\{A_1=a_1, A_2=a_2\}}|\mathbf{U}\right]\right] \\
&= \left[E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\right]^{-1} E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]\boldsymbol{\alpha}_a = \boldsymbol{\alpha}_a.
\end{aligned}$$

It follows that

$$\nu_a(\mathbf{S}) = \mathbf{U}^T\boldsymbol{\alpha}_a = \mathbf{U}^T\boldsymbol{\alpha}_a^* = \nu_a^*(\mathbf{S}).$$

Therefore, if the model (3.14) holds, the model for $\nu_a^*(\mathbf{S})$ is correctly specified.

(ii) It suffices to prove a counterexample. We refer to example M10 in the Simulation Experiments; see Section 6.2.

(iii) If we assume that $\bar{\boldsymbol{\alpha}}_a^* = \boldsymbol{\alpha}_a^*$, under Assumption 1 and by the law of iterated expectations, we have

$$\begin{aligned}
\boldsymbol{\beta}_a^* &= [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E[\bar{\mathbf{V}}\bar{\mathbf{U}}^T]\boldsymbol{\alpha}_a^* = [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E[\bar{\mathbf{V}}\bar{\mathbf{U}}^T]\bar{\boldsymbol{\alpha}}_a^* \\
&= [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E[\bar{\mathbf{V}}\bar{\mathbf{U}}^T][E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]]^{-1}E[\bar{\mathbf{U}}Y(a)] = [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E[\bar{\mathbf{V}}Y(a)] \qquad (3.16) \\
&= [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{V}Y(a)\right] \\
&= [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E\left[\mathbf{V}E\left[Y(a)|\mathbf{V}, A_1 = a_1\right]E\left[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{V}\right]\right] \\
&= [E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]]^{-1}E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{V}\mathbf{V}^T\boldsymbol{\beta}_a\right] = \boldsymbol{\beta}_a.
\end{aligned}$$

In (3.16), we used the fact that $\mathbf{U} = (\mathbf{V}^T, \mathbf{S}_2^T)^T$, i.e.,

$$\mathbf{V} = \mathbf{Q}\mathbf{U} \text{ where } \mathbf{Q} = \begin{pmatrix} \boldsymbol{I}_{d_1+1} & \mathbf{0}_{(d_1+1)\times d_2} \end{pmatrix}, \qquad (3.17)$$

and hence $\bar{\mathbf{V}} = \mathbf{Q}\bar{\mathbf{U}}$, which implies that

$$\begin{aligned}
E[\bar{\mathbf{V}}\bar{\mathbf{U}}^T][E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]]^{-1}E[\bar{\mathbf{U}}Y(a)] &= \mathbf{Q}E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T][E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]]^{-1}E[\bar{\mathbf{U}}Y(a)] \\
&= \mathbf{Q}E[\bar{\mathbf{U}}Y(a)] = E[\bar{\mathbf{V}}Y(a)].
\end{aligned}$$

(iv) Based on the results in (i), we have $\boldsymbol{\alpha}_a^* = \boldsymbol{\alpha}_a$. Under Assumption 1 and (3.14), we have

$$\nu_a(\mathbf{S}) = E\left[Y(a)|\mathbf{S}, A_1 = a_1, A_2 = a_2\right] = \mathbf{U}^T\boldsymbol{\alpha}_a.$$

Hence, we also have

$$\begin{aligned}
\bar{\boldsymbol{\alpha}}_a^* &= \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1}E[\bar{\mathbf{U}}Y(a)] = \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1}E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{U}Y(a)\right] \\
&= \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1}E\left[\mathbf{U}E\left[Y(a)|\mathbf{U}, A_1 = a_1\right]P\left[A_1 = a_1|\mathbf{U}\right]\right] \\
&= \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1}E\left[\mathbf{U}\mathbf{U}^T\boldsymbol{\alpha}_a E\left[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{U}\right]\right] \\
&= \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\right]^{-1}E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\boldsymbol{\alpha}_a = \boldsymbol{\alpha}_a.
\end{aligned}$$

Therefore,

$$\boldsymbol{\alpha}_a^* = \bar{\boldsymbol{\alpha}}_a^* = \boldsymbol{\alpha}_a.$$

Together with the results in (iii), we conclude that $\mu_a^*(\cdot)$ is correctly specified.

Estimation Estimation of the linear working models in the presence of high-dimensional covariates can be achieved with many regularizations. Throughout this work, we focus on Lasso regularization, albeit the theoretical developments apply more broadly. Recall the notation of I_{-k} introduced in the Dynamic ATE Algorithm 1.

The estimation is performed sequentially backward in time. We first obtain an estimator of (3.9) and, with it, an estimator of ν_a^* and ν_a, (3.6). We do so by regressing $\widetilde{Y}$ onto $\widetilde{\mathbf{U}}$ while utilizing a sparsity regularizing penalty, Lasso. That is, the Lasso estimator $\widehat{\boldsymbol{\alpha}}_a$ is defined as

$$\widehat{\boldsymbol{\alpha}}_a := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{|I_{-k}|} \sum_{i\in I_{-k}} \left(\widetilde{Y}_i - \widetilde{\mathbf{U}}_i^T\boldsymbol{\alpha}\right)^2 + \widetilde{\lambda}_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \right\}, \tag{3.18}$$

where $\widetilde{\lambda}_{\boldsymbol{\alpha}} = \widetilde{\lambda}_{\boldsymbol{\alpha}_a} > 0$ is some tuning parameter. In the above, we are considering a Lasso regularized regression among the individuals with the treatment plan a. For example, for $\theta = E[Y(a)] - E[Y(a')]$, we are interested in $a = (1,1)$ or $a' = (0,0)$ only. Let the corresponding estimators be named $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\alpha}}_0$, respectively.

The second step is to regress $\bar{\mathbf{U}}^T\widehat{\boldsymbol{\alpha}}_a$ onto $\bar{\mathbf{V}}$, in order to obtain an estimator of μ_a^* and, with it, μ_a, (3.7). Recall that $\bar{\mathbf{U}} = \mathbf{U}\mathbb{1}_{\{A_1=a_1\}}$ and that now we have to consider $a \in \{(1,0),(1,1)\}$ corresponding to $\widehat{\boldsymbol{\alpha}}_1$ and similarly $a \in \{(0,0),(0,1)\}$ corresponding to $\widehat{\boldsymbol{\alpha}}_0$. In other words, we need to consider individuals following the treatment paths of $\{(1,0),(1,1)\}$ when estimating $\widehat{\boldsymbol{\beta}}_1$ and individuals following the treatment paths $\{(0,0),(0,1)\}$ when esti-

mating $\widehat{\boldsymbol{\beta}}_0$. Notice that each of these estimators are an imputed, high-dimensional estimators, as the correct outcome for the problem should be $\bar{\mathbf{U}}^T\boldsymbol{\alpha}_a^*$. In other words, we define a Lasso estimator $\widehat{\boldsymbol{\beta}}_a$ as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_a &:= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}} \left\{ \frac{1}{|I_{-k}|} \sum_{i\in I_{-k}} \left(\bar{\mathbf{U}}_i^T\widehat{\boldsymbol{\alpha}}_a - \bar{\mathbf{V}}_i^T\boldsymbol{\beta}\right)^2 + \bar{\lambda}_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}} \left\{ \frac{1}{|I_{-k}|} \sum_{i\in I_{-k}, A_{1i}=a_1, A_{2i}\in\{0,1\}} \left(\mathbf{U}_i^T\widehat{\boldsymbol{\alpha}}_a - \mathbf{V}_i^T\boldsymbol{\beta}\right)^2 + \bar{\lambda}_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1 \right\},\end{aligned} \tag{3.19}$$

where $\bar{\lambda}_{\boldsymbol{\beta}} = \bar{\lambda}_{\boldsymbol{\beta}_a} > 0$ is a tuning parameter. For convience of expression, we use $\widehat{\boldsymbol{\beta}}_1$ to denote $\widehat{\boldsymbol{\beta}}_a$ for $a = (1,1)$ and similarly $\widehat{\boldsymbol{\beta}}_0$ for $a = (0,0)$. See Figure 3.1 for a representation.

Now, based on the estimated parameters, $\widehat{\boldsymbol{\alpha}}_a$ and $\widehat{\boldsymbol{\beta}}_a$, we propose corresponding nuisance function estimators as

$$\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2) = \mathbf{U}^T\widehat{\boldsymbol{\alpha}}_a, \tag{3.20}$$

$$\widehat{\mu}_a(\mathbf{S}_1) = \mathbf{V}^T\widehat{\boldsymbol{\beta}}_a. \tag{3.21}$$

Since the above is done for each individual in the sample, notice that we are, in turn, therefore, estimating the counterfactual outcomes for all those individuals not following the treatment path a.

3.3.2 Propensity Models

The estimation of the propensity models is also characterized by their working model class. We consider logistic regression model as a working model for both the propensity score at time one, $\pi(\mathbf{S}_1)$ as well as the one at time two, $\rho_a(\mathbf{S}_1, \mathbf{S}_2)$. Naturally, the logistic regression model is a particular case of generalized linear model, based on the binary response variable

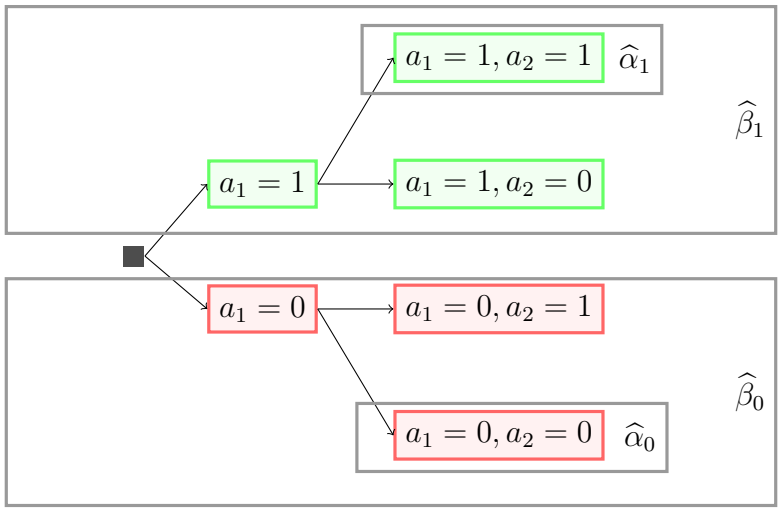


Figure 3.1: Treatment path utilization for the estimation of the nuisances. Each observation belongs to one of the four treatment paths depending on the treatment assignment in the first and the second exposure time. Gray boxes denote which treatment paths and, therefore, which samples are utilized to estimate the corresponding parameter.

A_1 and the link function $\phi(u) = \log(1 + \exp(u))$. The population minimizer of the loss function for the logistic model (3.2) is defined as

$$\boldsymbol{\gamma}^* := \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{d_1+1}} E\Big[-A_1 \mathbf{V}^T \boldsymbol{\gamma} + \log\big(1 + \exp(\mathbf{V}^T \boldsymbol{\gamma})\big)\Big]. \tag{3.22}$$

We define $\pi^*(\mathbf{s}_1)$ as

$$\pi^*(\mathbf{s}_1) = \frac{\exp(\mathbf{v}^T \boldsymbol{\gamma}^*)}{1 + \exp(\mathbf{v}^T \boldsymbol{\gamma}^*)}, \tag{3.23}$$

where for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$, $\mathbf{v} = (1, \mathbf{s}_1^T)^T$. Here, $\pi^*(\mathbf{s}_1)$ is a proxy of $\pi(\mathbf{s}_1)$, (3.2). We use the sample I_{-k} to construct the estimator $\widehat{\pi}(\mathbf{S}_1)$ as

$$\widehat{\pi}(\mathbf{S}_1) = \frac{\exp(\mathbf{V}^T \widehat{\boldsymbol{\gamma}})}{1 + \exp(\mathbf{V}^T \widehat{\boldsymbol{\gamma}})}, \tag{3.24}$$

where $\widehat{\boldsymbol{\gamma}}$ is defined as

$$\widehat{\boldsymbol{\gamma}} := \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{d_1+1}} \left\{ \frac{1}{|I_{-k}|} \sum_{i \in I_{-k}} \Big[-A_{1i} \mathbf{V}_i^T \boldsymbol{\gamma} + \log(1 + \exp(\mathbf{V}_i^T \boldsymbol{\gamma}))\Big] + \lambda_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_1 \right\}, \tag{3.25}$$

Algorithm 2 Dynamic Treatment Lasso (DTL)

Require: Observations $\{Y_i, \mathbf{S}_{1i}, A_{1i}, \mathbf{S}_{2i}, A_{2i}\}_{i=1}^N$. Treatment path $a = (1,1)$, $a' = (0,0)$.

For any fixed integer $K \geq 2$, split the indices $I = \{1, 2, ..., N\}$ into K equal-sized parts $\{I_k\}_{k=1}^K$ randomly such that the size of each fold I_k is $n := N/K$. Define $I_{-k} := I \backslash I_k$.

for $k = 1, 2, ..., K$ **do**

  while in I_{-k} **do**

    for $c \in \{a, a'\}$ **do**

      Set $\widehat{\nu}_c(\mathbf{S}_1, \mathbf{S}_2) = \boldsymbol{U}^T \widehat{\boldsymbol{\alpha}}_c$ with $\widehat{\boldsymbol{\alpha}}_c$ as in (3.18), using samples from the "small boxes" of Figure 3.1. Set $\widehat{\mu}_c(\mathbf{S}_1) = \mathbf{V}^T \widehat{\boldsymbol{\beta}}_c$ with $\widehat{\boldsymbol{\beta}}_c$ as in (3.19), using samples from the "large boxes" of Figure 3.1.

      Construct estimators of $\pi(\mathbf{S}_1)$ and $\rho_c(\mathbf{S}_1, \mathbf{S}_2)$, using (3.25) and (3.29).

    end for

  end while

  Compute $\check{\theta}^{(k)}$ as

$$\check{\theta}^{(k)} = \frac{1}{n} \sum_{i \in I_k} \Bigg[\mathbf{V}_i^T (\widehat{\boldsymbol{\beta}}_a - \widehat{\boldsymbol{\beta}}_{a'}) + \frac{A_{1i}}{\widehat{\pi}(\mathbf{S}_{1i})} (\mathbf{U}_i^T \widehat{\boldsymbol{\alpha}}_a - \mathbf{V}_i^T \widehat{\boldsymbol{\beta}}_a) - \frac{1 - A_{1i}}{1 - \widehat{\pi}(\mathbf{S}_{1i})} (\mathbf{U}_i^T \widehat{\boldsymbol{\alpha}}_{a'} - \mathbf{V}_i^T \widehat{\boldsymbol{\beta}}_{a'})$$
$$+ \frac{A_{1i} A_{2i}}{\widehat{\pi}(\mathbf{S}_{1i}) \widehat{\rho}_a(\mathbf{S}_{1i}, \mathbf{S}_{2i})} (Y_i - \mathbf{U}_i^T \widehat{\boldsymbol{\alpha}}_a) - \frac{(1 - A_{1i})(1 - A_{2i})}{(1 - \widehat{\pi}(\mathbf{S}_{1i}))(1 - \widehat{\rho}_{a'}(\mathbf{S}_{1i}, \mathbf{S}_{2i}))} (Y_i - \mathbf{U}_i^T \widehat{\boldsymbol{\alpha}}_{a'}) \Bigg].$$

end for

return The final estimator is obtained as

$$\widehat{\theta} = \frac{1}{K} \sum_{k=1}^K \check{\theta}^{(k)}. \tag{3.26}$$

with some tuning parameter $\lambda_{\boldsymbol{\gamma}} > 0$. Observe that, for this estimator, we utilize all of the observations at hand, regardless of its treatment path.

The population minimizer of the loss function for the logistic model (3.3) is defined as

$$\boldsymbol{\delta}^* := \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^{d+1}} E[-A_2\bar{\mathbf{U}}^T\boldsymbol{\delta} + \log(1+\exp(\bar{\mathbf{U}}^T\boldsymbol{\delta}))]. \tag{3.27}$$

With it, we define $\rho_a^*(\mathbf{s}_1, \mathbf{s}_2)$ as

$$\rho_a^*(\mathbf{s}_1, \mathbf{s}_2) = \frac{\exp(\mathbf{u}^T\boldsymbol{\delta}_a^*)}{1+\exp(\mathbf{u}^T\boldsymbol{\delta}_a^*)}, \tag{3.28}$$

where, for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$, $\mathbf{s}_2 \in \mathbb{R}^{d_2}$, $\mathbf{u} = (1, \mathbf{s}_1^T, \mathbf{s}_2^T)^T$. We use the sample I_{-k} to construct the estimator $\widehat{\boldsymbol{\delta}}_a$ as follows

$$\widehat{\boldsymbol{\delta}}_a := \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{|I_{-k}|} \sum_{i\in I_{-k}} \left[-A_{2i}\bar{\mathbf{U}}_i^T\boldsymbol{\delta} + \log(1+\exp(\bar{\mathbf{U}}_i^T\boldsymbol{\delta}))\right] + \bar{\lambda}_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_1 \right\}, \tag{3.29}$$

where $\bar{\lambda}_{\boldsymbol{\delta}} = \bar{\lambda}_{\boldsymbol{\delta}_a} > 0$ is some tuning parameter. In contrast to $\widehat{\boldsymbol{\gamma}}$, we are now utilizing only observations whose treatment path matches a_1 regardless of what is a_2; in Figure 3.1, it corresponds to the samples of $\widehat{\boldsymbol{\beta}}_a$. Then, the propensity score at the second time point can be naturally defined as

$$\widehat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2) = \frac{\exp(\mathbf{U}^T\widehat{\boldsymbol{\delta}}_a)}{1+\exp(\mathbf{U}^T\widehat{\boldsymbol{\delta}}_a)}. \tag{3.30}$$

3.3.3 Doubly Robust Lasso Estimator

From the previous subsection, we know the expressions for the estimators $\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2)$, $\widehat{\mu}_a(\mathbf{S}_1)$, $\widehat{\pi}(\mathbf{S}_1)$, and $\widehat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2)$ are (3.20), (3.21), (3.24), and (3.30) respectively. The corresponding estimators $\widehat{\boldsymbol{\alpha}}_a$, $\widehat{\boldsymbol{\beta}}_a$, $\widehat{\boldsymbol{\gamma}}$, and $\widehat{\boldsymbol{\delta}}_a$ are constructed based on the sample I_{-k} for each

$k = 1, 2, ..., K$. The final estimator is obtained as an average over I_k samples. Here, we only focus on the treatment paths $a = (1, 1)$ and $a' = (0, 0)$. Let $\eta := (\eta_a, \eta_{a'})$. For binary treatments, $\theta = E[Y(1,1) - Y(0,0)] = E[\psi(W; \eta)]$ and the score is defined as

$$\psi(W; \eta) = \psi_a(W; \eta_a) - \psi_{a'}(W; \eta_{a'}), \tag{3.31}$$

where we recall the definitions of η_a and $\psi_a(\cdot; \cdot)$ from (3.4). Details are presented in the Dynamic Treatment Lasso (DTL) Algorithm 2.

3.4 Theoretical characteristics of DTL

Before we discuss our main theoretical findings, we introduce a sequence of assumptions necessary for our analysis. These are related to the distribution of covariates $\mathbf{U}$ as well as errors ζ and ε defined below.

Assumption 3.2. *Let* $\mathbf{U}$ *be a sub-Gaussian vector that* $\|\boldsymbol{x}^T\mathbf{U}\|_{\psi_2} \leq \sigma_u \|\boldsymbol{x}\|_2$ *for any vector* $\boldsymbol{x} \in \mathbb{R}^{d+1}$, *with some constant* $\sigma_u > 0$. *In addition, let the smallest eigenvalue of the matrix* $E[\mathbf{U}\mathbf{U}^T]$ *satisfies* $\lambda_{\min}(E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1 = a_1\}}]) \geq \kappa_l$ *for each* $a_1 \in \{0, 1\}$, *with some constant* $\kappa_l > 0$.

Assumption 3.2 is standard and general in the literature. We note that it also contains an upper bound on the largest eigenvalue of $E[\mathbf{U}\mathbf{U}^T]$, as

$$\lambda_{\max}(E[\mathbf{U}\mathbf{U}^T]) = \max_{\|\boldsymbol{x}\|_2 = 1} E[(\boldsymbol{x}^T\mathbf{U})^2] \leq \max_{\|\boldsymbol{x}\|_2 = 1} 2\sigma_u^2 \|\boldsymbol{x}\|_2^2 = 2\sigma_u^2 < \infty.$$

Recall the definition of the true score function, $\psi_a(W; \eta_a)$ from (3.4). Recall the definition of the estimands collected as $\eta_a^*(\cdot) := (\mu_a^*(\cdot), \nu_a^*(\cdot), \pi^*(\cdot), \rho_a^*(\cdot))$, where the working

models are defined in (3.8), (3.11), (3.23) and (3.28), respectively. Let $\eta^* := (\eta_a^*, \eta_{a'}^*)$. With that in mind, we define the "working" score as

$$\psi(W; \eta^*) = \psi_a(W; \eta_a^*) - \psi_{a'}(W; \eta_{a'}^*),$$

where similar to (3.31),

$$\psi_a(W; \eta_a^*) := \mu_a^*(\mathbf{S}_1) + \tau_a^*(\mathbf{S}_1)\Big(\nu_a^*(\mathbf{S}_1, \mathbf{S}_2) - \mu_a^*(\mathbf{S}_1)\Big) + \omega_a^*(\mathbf{S}_1, \mathbf{S}_2)\Big(Y - \nu_a^*(\mathbf{S}_1, \mathbf{S}_2)\Big).$$

In the above, we have used inverse weights $\tau_a^*(\mathbf{s}_1) := \mathbb{1}_{\{A_1=a_1\}}\{\pi^*\}^{-1}(\mathbf{s}_1)$, $\omega_a^*(\mathbf{s}_1, \mathbf{s}_2) := \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}\{\rho_a^*\}^{-1}(\mathbf{s}_1, \mathbf{s}_2)$. Let

$$\sigma^2 := E[\psi(W; \eta^*) - \theta]^2. \tag{3.32}$$

By Lemma 3.8, we know $\theta = E[\psi(W; \eta^*)]$ when at least one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and at least one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Then, $\sigma^2 := E[\psi(W; \eta^*) - \theta]^2 = \mathrm{Var}[\psi(W; \eta^*)]$ denotes the variance of the "working score".

Assumption 3.3. *Define $\zeta := \zeta_a + \zeta_{a'}$ and $\varepsilon := \varepsilon_a + \varepsilon_{a'}$, where ζ_a and ε_a are defined in* (3.10) *and* (3.13)*, respectively. There exist some positive $\sigma_\zeta < \infty$ and $\sigma_\varepsilon < \infty$, such that ζ and ε are sub-Gaussian, with $\|\zeta\|_{\psi_2} \leq \sigma\sigma_\zeta$ and $\|\varepsilon\|_{\psi_2} \leq \sigma\sigma_\varepsilon$.*

Assumption 3.3 is fairly general even among the high-dimensional literature. As the number of samples N tends to infinity, $N \to \infty$, we allow the ψ_2-norm bound of ζ and ε to diverge or to shrink to zero. Consider treatment paths $a = (1, 1)$ and $a' = (0, 0)$. When all the nuisance models are correctly specified, under the overlap condition in Assumption 3.1, we have

$$\sigma^2 \asymp E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \geq \max\{E[\zeta^2], E[\varepsilon^2]\}.$$

where $\xi := \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta$. Hence, a sufficient condition for Assumption 3.3, while Assumption 3.1 holds, is $\|\zeta/\sqrt{E[\zeta^2]}\|_{\psi_2} \leq \sigma_\zeta$ and $\|\varepsilon/\sqrt{E[\varepsilon^2]}\|_{\psi_2} \leq \sigma_\varepsilon$, i.e., the "normalized" residuals have constant $\|\cdot\|_{\psi_2}$ norms. Note that, we allow $\sigma = \sigma_N$ to be dependent on N with assuming σ_ζ and σ_ε to be constants independent of N; $\sigma \to 0$ and $\sigma \to \infty$ are both allowed as $N \to \infty$. Besides, the variances $E[\zeta^2]$, $E[\varepsilon^2]$, and $\mathrm{Var}[\mathbf{U}^T(\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^*)] \asymp \|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^*\|_2^2$, $E[\xi^2]$, are all allowed to dependent on N and they are NOT necessarily of the same order; see more discussions in Remark 3.6.

3.4.1 Convergence rates of the nuisance parameters

The major difficulty in obtaining error of estimation regarding the outcome model estimates arises from the non-i.i.d. structure of the imputed outcomes used in the construction of $\widehat{\boldsymbol{\beta}}_a$. Here, we provide a general theory which establishes error bounds for the imputed least-squares Lasso estimators: estimators of the form (3.33), where $\widehat{Y}_i$ can be seen as an approximation of some Y_i or its conditional mean $E(Y_i|\mathbf{X}_i)$.

Imputed Lasso estimator Suppose $\mathbb{S} := (Y_i^*, \mathbf{X}_i)_{i=1}^M$ are *i.i.d.* observations and let $(Y^*, \mathbf{X})$ be an independent copy of $\mathbb{S}$, with $Y^* \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^d$. Suppose there exists, possibly random, $\widehat{Y}_i \in \mathbb{R}$ $(i = 1, \ldots, M)$. With a little abuse in notation, the true population slope as if all of the outcomes Y^* have been observed, is defined as

$$\boldsymbol{\beta}^* := \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} E\left[Y^* - \mathbf{X}^T\boldsymbol{\beta}\right]^2.$$

Then, its estimator is

$$\widehat{\boldsymbol{\beta}} := \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ M^{-1} \sum_{i=1}^{M} [\widehat{Y}_i - \mathbf{X}_i^T\boldsymbol{\beta}]^2 + \lambda_M \|\boldsymbol{\beta}\|_1 \right\}, \tag{3.33}$$

for $\lambda_M > 0$. Note that for some, and possibly all observations, outcomes Y^* are imputed, i.e., estimated using $\hat{Y}_i$. The following result delineates property of such imputed-Lasso, $\widehat{\boldsymbol{\beta}}$ estimator.

Theorem 3.1. *Let $\varepsilon_i := Y_i^* - \mathbf{X}_i^T\boldsymbol{\beta}^*$ with $s = \|\boldsymbol{\beta}^*\|_0$. Suppose that $\|\boldsymbol{a}^T\mathbf{X}\|_{\psi_2} \leq \sigma_{\mathbf{X}}\|\boldsymbol{a}\|_2$ for $\boldsymbol{a} \in \mathbb{R}^d$, $\lambda_{\min}(E[\mathbf{X}\mathbf{X}^T]) > \lambda_{\mathbf{X}}$, and $\|\varepsilon\|_{\psi_2} \leq \sigma$ with some constants $\sigma_{\mathbf{X}}, \lambda_{\mathbf{X}} > 0$ and a positive $\sigma = \sigma_M > 0$ potentially dependent on M. For some $\delta_M > 0$, define*

$$\mathcal{E}_1 := \Big\{ M^{-1} \sum_{i=1}^{M} [\widehat{Y}_i - Y_i^*]^2 < \delta_M^2 \Big\}.$$

For any $t > 0$, let $\lambda_M := 16\sigma\sigma_{\mathbf{X}}(\sqrt{\log(d)/M} + t)$. Then, on the event $\mathcal{E}_1$, when $M > \max\{\log(d), 100\kappa_2^2 s\log(d)\}$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &\leq \max\left(\frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M, 8\kappa_1^{-1}\sqrt{s}\lambda_M \right), \\
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &\leq \max\left(20\lambda_M^{-1}\delta_M^2, 40\kappa_1^{-1}s\lambda_M \right), \\
\frac{1}{M}\sum_{i=1}^{M} [\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 &\leq \max\left(16\delta_M^2, 32\kappa_1^{-1}s\lambda_M^2 \right),
\end{aligned}$$

with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2 M)$, wher $\kappa_1, \kappa_2, c_1, c_2 > 0$ are some constants independent of M and d. Moreover, if $\delta_M = o(\sigma)$, $P(\mathcal{E}_1) = 1 - o(1)$, and $M \gg s\log(d)$, then, with some $\lambda_M \asymp \sigma\sqrt{\frac{\log(d)}{M}}$, as $M \to \infty$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\Big(\sigma\sqrt{\frac{s\log(d)}{M}} + \delta_M\Big), \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\Big(\sigma s\sqrt{\frac{\log(d)}{M}} + \sigma^{-1}\delta_M^2\sqrt{\frac{M}{\log(d)}}\Big),$$

$$\frac{1}{M}\sum_{i=1}^{M} [\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 = O_p\left(\delta_M^2 + \sigma^2\frac{s\log(d)}{M} \right).$$

A few comments are essential. The above result contributes to the literature in three specific aspects: 1) The "imputation error", $\widehat{Y}_i - Y_i^*$, is dependent on and even possibly

correlated with covariates $\mathbf{X}_i$; 2) We allow $\widehat{Y}_i$, $\forall i \in \{1, \ldots, M\}$, to be fitted using the same set of observations $(X_i, Y_i)_{i=1}^M$, i.e., $\widehat{Y}_i$s are also possibly dependent on each other; 3) The tuning parameter λ_M is of the same order as the one chosen for the fully observed data and is independent of any sparsity parameter. As a result, Theorem 3.1 leads to better rates of estimation. [ZZS19] require rate of $o(n/\log(p))$ on the product of sparsities at the time of exposures. Our results rely on the sum instead; see Corrolary 3.2 below.

The result requires developing new techniques: the standard Lasso inequality followed by the cone-set reduction are not valid in this instance. In fact, the error vector, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, no longer belongs to the accustomed cone set, $\mathcal{C}(S, 4) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1\}$. We identify a new cone set, $\widetilde{\mathcal{C}}(S, 4, 1) = \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 16\lambda_M^{-1}\delta_M^2, \|\boldsymbol{\Delta}_S\|_1 \leq 4\lambda_M^{-1}\delta_M^2\}$, and show that the error vector belongs to the union of the above two sets. As shown in Theorem 3.1, the rate of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ consists of two components: 1) the standard (non-imputed) estimation rate $\sigma\sqrt{s\log(d)}/M$; 2) the imputation error δ_M. When there is no imputation, i.e., $\delta_M = 0$, our results reaches the standard consistency rate in the high-dimensional statistics literature, e.g., [BRT09, NRWY12, Wai19].

Nuisance estimation Based on Theorem 3.1, we provide theoretical properties of our nuisance parameters in the following Corollaries. As is typical in high-dimensional models, our analysis will rely on certain sparsity assumptions of the underlying models. In fact, only the approximate models will be considered. To that end, we let $S_{\boldsymbol{\alpha}_a} = \{j : \boldsymbol{\alpha}_a^*[j] \neq 0\}$ and $S_{\boldsymbol{\beta}_a} = \{j : \boldsymbol{\beta}_a^*[j] \neq 0\}$ be the sets of nonzero coordinates of $\boldsymbol{\alpha}_a^*$, (3.9) and $\boldsymbol{\beta}_a^*$, (3.12), respectively. Let $s_{\boldsymbol{\alpha}_a} = |S_{\boldsymbol{\alpha}_a}|$ and $s_{\boldsymbol{\beta}_a} = |S_{\boldsymbol{\beta}_a}|$ denote the numbers of nonzero coordinates of $\boldsymbol{\alpha}_a^*$ and $\boldsymbol{\beta}_a^*$. Let $S_{\boldsymbol{\gamma}} = \{j : \boldsymbol{\gamma}^*[j] \neq 0\}$ be the set of nonzero coordinates of $\boldsymbol{\gamma}^*$, (3.22) and let

$s_{\boldsymbol{\gamma}} = |S_{\boldsymbol{\gamma}}|$ denote the number of nonzero coordinates of $\boldsymbol{\gamma}^*$. Similarly, $S_{\boldsymbol{\delta}_a} = \{j : \boldsymbol{\delta}_a^*[j] \neq 0\}$ be the set of nonzero coordinates of $\boldsymbol{\delta}_a^*$, (3.27), and $s_{\boldsymbol{\delta}_a} = |S_{\boldsymbol{\delta}_a}|$ be the number of nonzero coordinates of $\boldsymbol{\delta}_a^*$. Throughout this section we denote with M the size of the set I_{-k}, i.e., $M = |I_{-k}| = \frac{(K-1)N}{K}$ with K denoting the number of folds used in Algorithm 1.

Corollary 3.1. *Let Assumptions 3.1, 3.2, and 3.3 hold. For any $t > 0$, let $\widetilde{\lambda}_{\boldsymbol{\alpha}} := 32\sigma\sigma_u\sigma_\zeta(t + \sqrt{\frac{\log(d+1)}{M}})$. Let $M > \max\{\log(d+1), 100\kappa_2^2 s_{\boldsymbol{\alpha}_a}\log(d+1)\}$. Then, $\widehat{\boldsymbol{\alpha}}_a$, (3.18), satisfies*

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 &\leq 8\kappa_1^{-1}\widetilde{\lambda}_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\alpha}_a}}, \quad \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_1 \leq 40\kappa_1^{-1}\widetilde{\lambda}_{\boldsymbol{\alpha}} s_{\boldsymbol{\alpha}_a}, \\ \frac{1}{M}\sum_{i=1}^{M}[\widetilde{\mathbf{U}}_i^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 &\leq 32\kappa_1^{-1}\widetilde{\lambda}_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}, \end{aligned} \tag{3.34}$$

with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2 M)$ and constants $c_1, c_2, \kappa_1, \kappa_2 > 0$. Therefore, if $N \gg s_{\boldsymbol{\alpha}_a}\log(d)$, then with some $\widetilde{\lambda}_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\frac{\log(d)}{N}}$, as $N \to \infty$,

$$\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\right), \quad \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_1 = O_p\left(\sigma s_{\boldsymbol{\alpha}_a}\sqrt{\frac{\log(d)}{N}}\right), \tag{3.35}$$

$$E[\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2) - \nu_a^*(\mathbf{S}_1, \mathbf{S}_2)]^2 = O_p\left(\sigma^2\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}\right). \tag{3.36}$$

In the above, the left-hand side of (3.36) is denoting expectation with respect to the distribution of the new observation's covariates $\mathbf{S}_1, \mathbf{S}_2$. The results in Corollary 3.1 can be seen as a special (degenerate) case of Theorem 3.1. The asymptotic results in (3.35) coincide with the high-dimensional linear regression literature, e.g., [NRWY12] and [Wai19].

Now we discuss the results for the estimation of $\boldsymbol{\beta}_a^*$. The estimator $\widehat{\boldsymbol{\beta}}_a$ proposed in (3.19) is constructed based on $\widehat{\boldsymbol{\alpha}}_a$ and hence we need to first control the estimation error of $\widehat{\boldsymbol{\alpha}}_a$. Note that, $\widehat{\boldsymbol{\alpha}}_a$ and $\widehat{\boldsymbol{\beta}}_a$ in (3.18) and (3.19) are actually obtained based on overlapping but different sample groups. For $\widehat{\boldsymbol{\alpha}}_a$, we only utilize the samples satisfying $A_{1i} = a_1$ and

$A_{2i} = a_2$; as for $\widehat{\boldsymbol{\beta}}_a$, we are using the samples such that $A_{1i} = a_1$ and there is no constraint on A_{2i}. As a result, the in-sample error (3.34) is not enough for our analysis. Instead, we require an upper bound for a "partially in-sample" error. We show the prerequisite results in the following lemma.

Lemma 3.1. *Let Assumptions of Corollary 3.1 hold. In addition, let* $M \geq \max\{\log(d+1), (c_3 + 100\kappa_2^2)s_{\boldsymbol{\alpha}_a}\log(d+1)\}$, *with some constant* $c_3 > 0$. *Then,*

$$\frac{1}{M}\sum_{i=1}^{M}[\bar{\mathbf{U}}_i^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \leq 288\sigma_u\kappa_1^{-2}\tilde{\lambda}_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a},$$

with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2M) - 2\exp(-c_4M)$ *and constants* $c_1, c_2, c_4 > 0$.

Now, based on Theorem 3.1 and Lemma 3.1, we are ready to obtain the estimation and prediction quality of the estimator $\widehat{\boldsymbol{\beta}}_a$.

Corollary 3.2. *Let Assumptions 3.1-3.3 hold. Define* $\widehat{\boldsymbol{\beta}}_a$ *as in* (3.19). *For any* $t > 0$, *let* $\widetilde{\lambda}_{\boldsymbol{\alpha}} := 32\sigma\sigma_u\sigma_\zeta(\sqrt{\frac{\log(d+1)}{M}} + t)$ *and* $\bar{\lambda}_{\boldsymbol{\beta}} := 32\sigma\sigma_u\sigma_\varepsilon(\sqrt{\frac{\log(d_1+1)}{M}} + t)$. *Suppose that* $M \geq \max\{\log(d+1), (c_3 + 100\kappa_2^2)s_{\boldsymbol{\alpha}_a}\log(d+1), 100\kappa_2^2 s_{\boldsymbol{\beta}_a}\log(d_1+1)\}$. *Let* $\delta_M^2 = 288\sigma_u\kappa_1^{-2}\tilde{\lambda}_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}$. *Then,*

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2 &\leq \max\left(\frac{5\kappa_2\delta_M^2}{8\sigma\sigma_u\sigma_\varepsilon} + 4\kappa_1^{-1/2}\delta_M, 8\kappa_1^{-1}\bar{\lambda}_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}_a}}\right), \\
\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_1 &\leq \max\left(20\bar{\lambda}_{\boldsymbol{\beta}}^{-1}\delta_M^2, 40\kappa_1^{-1}\bar{\lambda}_{\boldsymbol{\beta}} s_{\boldsymbol{\beta}_a}\right), \\
\frac{1}{M}\sum_{i=1}^{M}[\bar{\mathbf{V}}_i^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)]^2 &\leq \max\left(16\delta_M^2, 32\kappa_1^{-1}\bar{\lambda}_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_a}\right),
\end{aligned}$$

with probability at least $1 - 4\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - 2c_1\exp(-c_2M) - 2\exp(-c_4M)$ *and some constants* $c_1, c_2, c_3, c_4, \kappa_1, \kappa_2 > 0$. *Moreover, assume* $N \gg \max\{s_{\boldsymbol{\alpha}_a}\log(d), s_{\boldsymbol{\beta}_a}\log(d_1)\}$. *Then,*

with some $\widetilde{\lambda}_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\frac{\log(d)}{N}}$ *and* $\bar{\lambda}_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\frac{\log(d_1)}{N}}$, *as* $N \to \infty$,

$$\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2 = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d) + s_{\boldsymbol{\beta}_a}\log(d_1)}{N}}\right),$$

$$\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_1 = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}^2\log^2(d)}{N\log(d_1)}} + \sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}^2\log(d_1)}{N}}\right),$$

$$\frac{1}{M}\sum_{i=1}^{M}[\bar{\mathbf{V}}_i^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)]^2 = O_p\left(\sigma^2\frac{s_{\boldsymbol{\alpha}_a}\log(d) + s_{\boldsymbol{\beta}_a}\log(d_1)}{N}\right),$$

and it follows that $E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^2 = O_p\left(\sigma^2\frac{s_{\boldsymbol{\alpha}_a}\log(d)+s_{\boldsymbol{\beta}_a}\log(d_1)}{N}\right)$.

Note that the left-hand side of the very last equation is considering an expectation with respect to a distribution of a new, test data, i.e., its covariate $\mathbf{S}_1$, only.

Remark 3.3 (Model misspecifications)**.** *In the estimation of* $\mu_a(\cdot)$, *we allow two types of model misspecifications:* $\nu_a^*(\cdot) \neq \nu_a(\cdot)$ *and/ or* $\mu_a^*(\cdot) \neq \mu_a(\cdot)$. *When the model is misspecified, in that* $\mu_a(\cdot)$ *is non-linear, the estimator* $\widehat{\mu}_a(\cdot)$ *converges to some* $\mu_a^*(\cdot) \neq \mu_a(\cdot)$. *Here, the target function* $\mu_a^*(\cdot)$ *can be seen as an "optimal" linear function approximating* $\mu_a(\cdot)$ *and the target parameter* $\boldsymbol{\beta}_a^*$ *can be seen as an "optimal" linear slope in the population level. The nuisance function,* $\nu_a(\cdot)$, *is also allowed to be misspecified, although the estimation of* $\mu_a(\cdot)$ *does depend on the estimator* $\widehat{\nu}_a(\cdot)$. *The results in Corollary 3.2 are valid as long as the assumptions in Corollary 3.1 hold:* $\widehat{\boldsymbol{\alpha}}_a$ *estimates well the target "optimal" slope,* $\boldsymbol{\alpha}_a^*$.

When misspecification occurs in the propensity score models, we need an extra "overlap condition" for the "target" propensity score functions:

Assumption 3.4. *Let* c *be fixed positive constant.* $\pi^*(\mathbf{S}_1)$ *and* $\rho_a^*(\mathbf{S}_1, \mathbf{S}_2)$ *satisfy the following conditions for* $a \in \{0, 1\}$*:*

$$P(c_0 \leq \pi^*(\mathbf{S}_1) \leq 1 - c_0) = 1, \quad P(c_0 \leq \rho_a^*(\mathbf{S}_1, \mathbf{S}_2) \leq 1 - c_0) = 1.$$

Lemma 3.2. *Let Assumptions 3.1-3.4 hold. Assume* $N \gg \max\{s_{\boldsymbol{\alpha}_a}\log(d), s_{\boldsymbol{\beta}_a}\log(d_1)\}$. *Then, with some* $\widetilde{\lambda}_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\frac{\log(d)}{N}}$ *and* $\bar{\lambda}_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\frac{\log(d_1)}{N}}$, *as* $N \to \infty$, *we obtain*

$$\{E|\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2) - \nu_a^*(\mathbf{S}_1, \mathbf{S}_2)|^r\}^{1/r} = O_p\Big(\sigma\sqrt{s_{\boldsymbol{\alpha}_a}\log(d)/N}\Big),$$
$$\{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^r\}^{1/r} = O_p\Big(\sigma\sqrt{s_{\boldsymbol{\alpha}_a}\log(d) + s_{\boldsymbol{\beta}_a}\log(d_1)/N}\Big).$$

Additionally, let $N \gg \max\{s_{\boldsymbol{\gamma}}\log(d_1), s_{\boldsymbol{\delta}_a}\log(d)\}$. *Consider some* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log(d_1)}{N}}$ *and* $\bar{\lambda}_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log(d)}{N}}$. *Define the event* $\mathcal{A} := \{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1\}$. *Then, as* $N \to \infty$, $P(\mathcal{A}) = 1 - o(1)$. *Moreover, on the event* $\mathcal{A}$, *as* $N \to \infty$, $\{E|\widehat{\pi}(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}}$ *and* $\{E|\widehat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2)|^{-r}\}^{\frac{1}{r}}$ *are both bounded uniformly by some constants independent of* N *and for* $r > 2$,

$$\big\{E\,\big|\widehat{\pi}^{-1}(\mathbf{S}_1) - \pi^{*-1}(\mathbf{S}_1)\big|^r\big\}^{1/r} = O_p\Big(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log(d_1)}{N}}\Big),$$
$$\big\{E\,\big|\widehat{\rho}_a^{-1}(\mathbf{S}_1, \mathbf{S}_2) - \rho_a^{*-1}(\mathbf{S}_1, \mathbf{S}_2)\big|^r\big\}^{1/r} = O_p\Big(\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\Big),$$
$$\big\{E\,\big|\widehat{\pi}^{-1}(\mathbf{S}_1)\widehat{\rho}_a^{-1}(\mathbf{S}_1, \mathbf{S}_2) - \pi^{*-1}(\mathbf{S}_1)\rho_a^{*-1}(\mathbf{S}_1, \mathbf{S}_2)\big|^r\big\}^{1/r} = O_p\Big(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log(d_1) + s_{\boldsymbol{\delta}_a}\log(d)}{N}}\Big).$$

In the above, the left-hand side of the first equation denotes the expectation w.r.t. the distribution of the new observation's covariate at time 1, $\mathbf{S}_1$. The left-hand sides of the last two equations denote the expectation w.r.t. the distribution of the new observation's covariates at both times, $\mathbf{S}_1, \mathbf{S}_2$.

3.4.2 Dynamic Treatment: Estimation and Inference

To provide valid inference result, we assume the following conditions on the sparsity levels:

Assumption 3.5. *Let* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$ *together with the following product rate condition*

$$\max\{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}\}\log^2(d) = o(N), \tag{3.37}$$

where, for the sake of simplicity, let $d_1 \asymp d_2 \asymp d$.

The first of the above two conditions is a simple condition requiring consistency of estimation of the nuisance parameters. The second of the two conditions, (3.37), is an equivalent of a product-rate condition required for double-robust estimation, but now it is in the context of dynamic treatment. Instead of one product rate, the above condition requires three product rate conditions to hold.

Theorem 3.2 (Rate double robustness)**.** *Suppose that the models* $\nu_a^*(\mathbf{S}_1, \mathbf{S}_2)$, $\mu_a^*(\mathbf{S}_1)$, $\pi^*(\mathbf{S}_1)$ *and* $\rho_a^*(\mathbf{S}_1, \mathbf{S}_2)$ *are all correctly specified. Let Assumptions 3.1-3.3 and 3.5 be satisfied. Then, as* $N \to \infty$, $\widehat{\theta}$, (3.26)*, is asymptotically normal with*

$$\sigma^{-1}\sqrt{N}(\widehat{\theta} - \theta) \to N(0, 1),$$

where σ^2 *is defined in* (3.32)*. The result continues to hold if* σ^2 *is replaced by* $\widehat{\sigma}^2 := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}[\psi(W_i; \widehat{\eta}) - \widehat{\theta}]^2$, *with* $\widehat{\eta} := (\widehat{\eta}_a, \widehat{\eta}_{a'})$.

Remark 3.4. *We compare the sparsity conditions of Assumption 3.5 with the double robust static ATE estimation literature. The ATE estimation problem can be seen as a special (degenerate) case of the dynamic ATE estimation, where we assume* $\mathbf{S}_1$ *and* A_1 *are completely random. In other words, the nuisance functions* $\mu_a(\cdot)$ *and* $\pi(\cdot)$ *are both constants, and hence can be estimated with a root-N rate. Then, Assumption 3.5 requires* $s_{\boldsymbol{\alpha}_a} + s_{\boldsymbol{\delta}_a} = o(N/\log(d))$

and $s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a} = o(N/\log^2(d))$ *coinciding with the sparsity conditions in [CCD+18, SRR19] and being weaker than [Far15, Tan20a, DV20, DAV20, AV21].*

We also provide the following Theorem that characterizes the consistency rate of the proposed estimator, $\widehat{\theta}$, in the presence of model misspecifications.

Table 3.1: Consistency rate of $\widehat{\theta}$ under various misspecification settings under Theorem 3.3. Misspecified and well-specified models are denoted with ✗ and ✓, respectively.

Nuisance model correctness				Consistency rate of $\widehat{\theta}$
$\rho_a^*(\cdot)$	$\pi^*(\cdot)$	$\mu_a^*(\cdot)$	$\nu_a^*(\cdot)$	
✓	✓	✓	✓	$O_p\left(\frac{\sigma}{\sqrt{N}}\left(1+\frac{\max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}}}, \sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\}\log(d)}{\sqrt{N}}\right)\right)$
✗	✓	✓	✓	$O_p\left(\sigma \max\left\{\frac{\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\log(d)}{N}, \sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\right\}\right)$
✓	✗	✓	✓	$O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}\}\log(d)}{N}}\right)$
✓	✓	✗	✓	$O_p\left(\sigma \max\left\{\frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}\log(d)}{N}, \sqrt{\frac{s_{\boldsymbol{\gamma}}\log(d)}{N}}\right\}\right)$
✓	✓	✓	✗	$O_p\left(\sigma \max\left\{\frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}}}\log(d)}{N}, \frac{\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\log(d)}{N}, \sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right\}\right)$
✗	✗	✓	✓	$O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}\}\log(d)}{N}}\right)$
✗	✓	✗	✓	$O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\gamma}}\}\log(d)}{N}}\right)$
✓	✗	✓	✗	$O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right)$
✓	✓	✗	✗	$O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right)$

Theorem 3.3 (Consistency rate)**.** *Suppose that one of the models* $\mu_a^*(\mathbf{S}_1)$ *and* $\pi^*(\mathbf{S}_1)$ *is correctly specified, and one of the models* $\nu_a^*(\mathbf{S}_1, \mathbf{S}_2)$ *and* $\rho_a^*(\mathbf{S}_1, \mathbf{S}_2)$ *is correctly specified. Let Assumptions 3.1-3.4 hold. Let* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$*. Then, with some tuning parameters* $\widetilde{\lambda}_{\boldsymbol{\alpha}} \asymp \bar{\lambda}_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\frac{\log(d)}{N}}$ *and* $\lambda_{\boldsymbol{\gamma}} \asymp \bar{\lambda}_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log(d)}{N}}$*, as* $N \to \infty$*, the estimator* $\widehat{\theta}$*,* (3.26)*,*

satisfies

$$\widehat{\theta} - \theta = O_p\bigg(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\bigg), \tag{3.38}$$

with $s_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}}}, \sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\}$ *and*

$$s_2 := \max\left\{s_{\boldsymbol{\alpha}_a}\mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot) \text{ or } \rho_a^*(\cdot)\neq\rho_a(\cdot)\}}, s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}}, s_{\boldsymbol{\gamma}}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right\}.$$

Remark 3.5 (Consistency rate under various misspecification settings)**.** *Below we discuss the consistency rate of* $\widehat{\theta}$ *under different misspecification settings. Therefore, when all the nuisance functions are correctly specified, we have* $s_2 = 0$ *and hence* $\widehat{\theta} - \theta = O_p\left(\sigma\left(1 + s_1\log(d)/\sqrt{N}\right)/\sqrt{N}\right)$*. However, when one of the models is misspecified at each exposure time, we have* $\widehat{\theta} - \theta = O_p\left(\sigma\sqrt{s_2\log(d)/N}\right)$*. More specifically, in Table 3.1, we illustrate the consistency rate of* $\widehat{\theta}$ *under all the considered model misspecification cases. We observe that, the consistency rate is asymmetric w.r.t. the sparsity levels. For instance, when all the models are correctly specified, the consistency rate of* $\widehat{\theta}$ *depends on three product rates:* $s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\gamma}}$*,* $s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\delta}_a}$*, and* $s_{\boldsymbol{\beta}_a}s_{\boldsymbol{\gamma}}$*. We can see that the sparsity levels* $s_{\boldsymbol{\alpha}_a}$ *and* $s_{\boldsymbol{\gamma}}$ *seem to be more "important" than* $s_{\boldsymbol{\beta}_a}$ *and* $s_{\boldsymbol{\delta}_a}$*: both* $s_{\boldsymbol{\alpha}_a}$ *and* $s_{\boldsymbol{\gamma}}$ *appear twice in the three product rates, whereas* $s_{\boldsymbol{\beta}_a}$ *and* $s_{\boldsymbol{\delta}_a}$ *only appear once. We can see that the consistency rate of* $\widehat{\theta}$ *depends on* σ*. Note that, we allow the dependency of* $\sigma = \sigma_N$ *on* N*;* $\sigma \to 0$ *and* $\sigma \to \infty$ *are both allowed as* $N \to \infty$*.*

3.5 Inference with general high-dimensional nuisances

Consider the general dynamic treatment effect estimator $\widehat{\theta}$ proposed in Algorithm 1. Let $\widehat{\mu}_a$, $\widehat{\nu}_a$, $\widehat{\pi}$, and $\widehat{\rho}_a$ denote any reasonable machine learning or nonparametric estimators

of the nuisance parameters η. Here, model misspecification is allowed. Let μ_a^*, ν_a^*, π^*, and ρ_a^* denote the 'target' functions of $\widehat{\mu}_a$, $\widehat{\nu}_a$, $\widehat{\pi}$, and $\widehat{\rho}_a$ respectively. In this Section, unless specified differently, E denotes an expectation only with respect to a probability measure of a new, test observation W.

Assumption 3.6. *There exist $\mu_a^*(\mathbf{S}_1)$, $\nu_a^*(\mathbf{S}_1, \mathbf{S}_2)$, $\pi^*(\mathbf{S}_1)$, and $\rho_a^*(\mathbf{S}_1, \mathbf{S}_2)$ such that $\widehat{\mu}_a(\mathbf{S}_1)$, $\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2)$, $\widehat{\pi}(\mathbf{S}_1)$, and $\widehat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2)$, computed on a subset I_{-k} obey the following conditions for all $a = (a_1, a_2)$ and $a_1, a_2 \in \{0, 1\}$. (i) Consistency for μ_a^* and ν_a^*: $E[\widehat{\nu}_a(\mathbf{S}_1, \mathbf{S}_2) - \nu_a^*(\mathbf{S}_1, \mathbf{S}_2)]^2 = O_p(a_N^2)$, $E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^2 = O_p(b_N^2)$, with sequences $a_N = o(\sigma)$ and $b_N = o(\sigma)$. (ii) Consistency for π^* and ρ_a^*: $E[\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1)]^2 = O_p(c_N^2)$, $E[\widehat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2) - \rho_a^*(\mathbf{S}_1, \mathbf{S}_2)]^2 = O_p(d_N^2)$, with sequences $c_N = o(1)$ and $d_N = o(1)$.*

Assumption 3.7. *Let c_0 be a fixed positive constant. Suppose that $\hat{\pi}(\mathbf{S}_1)$ and $\hat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2)$ satisfy $P(c_0 \leq \hat{\pi}(\mathbf{S}_1) \leq 1 - c_0) = 1$, $P(c_0 \leq \hat{\rho}_a(\mathbf{S}_1, \mathbf{S}_2) \leq 1 - c_0) = 1$, for $a \in \{0, 1\}$, with probability approaching one.*

With a little abuse of notation, in this section, we define $\zeta := \zeta_1 + \zeta_0$ and $\varepsilon := \varepsilon_1 + \varepsilon_0$, where for any general treatment path a,

$$\zeta_a := \mathbb{1}_{\{A_1 = a_1, A_2 = a_2\}} \left(Y(a) - \nu_a^*(\mathbf{S}_1, \mathbf{S}_2) \right), \quad \varepsilon_a := \mathbb{1}_{\{A_1 = a_1\}} \left(\nu_a^*(\mathbf{S}_1, \mathbf{S}_2) - \mu_a^*(\mathbf{S}_1) \right). \tag{3.39}$$

We also define, $\xi := \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta = E[Y(1,1) - Y(0,0)|\mathbf{S}_1] - E[Y(1,1) - Y(0,0)]$ as the centered conditional dynamic treatment effect at the first exposure. We impose the following assumptions on the distribution of ζ, ε, and ξ.

Assumption 3.8. *Suppose that, there exists some fixed constants $C > 0$ and $q > 2$, such that $\max \left\{ \frac{E|\zeta|^q}{[E|\zeta|^2]^{\frac{q}{2}}}, \frac{E|\varepsilon|^q}{[E|\varepsilon|^2]^{\frac{q}{2}}}, \frac{E|\xi|^q}{[E|\xi|^2]^{\frac{q}{2}}} \right\} \leq C$ as well as $P(E[\zeta^2|\mathbf{S}_1, \mathbf{S}_2] \leq CE[\zeta^2]) = 1$ and $P(E[\varepsilon^2|\mathbf{S}_1] \leq CE[\varepsilon^2]) = 1$.*

The max condition above is a moment condition that controls the tails of the distributions of ζ, ε, and ξ. For example, this condition holds if ζ, ε, and ξ are sub-Gaussian random variables. The last two conditions require that the "normalized" conditional second moments are almost surely bounded, assumed only for the interpretability of the obtained results. One can also replace these with some moment conditions on ζ and ε; however, we would then need to require upper bounds on higher moments on the estimation error rates instead of the second moments as used in Assumption 3.6.

3.5.1 Main results

The main result is presented below. We establish asymptotic normality of the general dynamic treatment effect estimator $\widehat{\theta}$ proposed in Algorithm 1, when all the nuisance functions are correctly specified but estimated using high-dimensional, machine learning or modern nonparametrics estimators.

Theorem 3.4. *(Rate double robustness) Assume that the models $\nu_a^*(\mathbf{S}_1, \mathbf{S}_2)$, $\mu_a^*(\mathbf{S}_1)$, $\pi^*(\mathbf{S}_1)$, and $\rho_a^*(\mathbf{S}_1, \mathbf{S}_2)$ are all correctly specified. Let Assumptions 3.1, and 3.6 - 3.8 hold. Moreover, assume that the rates of estimation satisfy the following product condition*

$$b_N c_N = o(\sigma N^{-1/2}), \qquad a_N d_N = o(\sigma N^{-1/2}). \tag{3.40}$$

Then, the estimator $\widehat{\theta}$ is approximately unbiased and normally distributed

$$\sigma^{-1}\sqrt{N}(\widehat{\theta} - \theta) \to N(0, 1),$$

with σ defined in (3.32)*. The result continues to hold when σ^2 is replaced with $\widehat{\sigma}^2$ as defined in Theorem 3.2.*

The notion of rate double robustness, although previously established in earlier works, has been named in [SRR19]. It stands to illustrate conditions termed "product rate conditions " needed when the models are correctly specified but the estimators of the nuisance parameters are not root-N consistent; see, e.g., Theorem 5.1 in [CCD$^+$18]. To the best of our knowledge, for the case of multiple time exposures, product rate conditions as identified in (3.40) are new. For a special case of one-time exposure, the above result matches those obtained in [CCD$^+$18].

Remark 3.6 (Signal-to-noise ratios)**.** *Suppose all the nuisance models are correctly specified and let Assumption 3.1 holds. For each a, we introduce the signal-to-noise ratios (SNRs) of the models,* (3.39)*, as*

$$\begin{aligned}
\mathrm{SNR}_{\nu,a} :=& \frac{Var[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\nu_a(\mathbf{S}_1,\mathbf{S}_2)]}{E[\zeta_a^2]} \asymp \frac{E[\varepsilon^2]+\ Var[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\mu_a(\mathbf{S}_1)]}{E[\zeta^2]},\\
\mathrm{SNR}_{\mu,a} :=& \frac{Var[\mathbb{1}_{\{A_1=a_1\}}\mu_a(\mathbf{S}_1)]}{E[\varepsilon_a^2]} \asymp \frac{Var[\mathbb{1}_{\{A_1=a_1\}}\mu_a(\mathbf{S}_1)]}{E[\varepsilon^2]},
\end{aligned}$$

under the overlap condition in Assumption 3.1. In our Theorem 3.2, both $\mathrm{SNR}_{\nu,a}$ *and* $\mathrm{SNR}_{\mu,a}$ *are allowed shrink to zero or diverge, as $N \to \infty$.*

Remark 3.7 (Rate double robustness)**.** *Rate double robustness in the presence of multiple exposures is discussed in [BHL20], however, the authors therein require three product rate conditions. In addition to the two product rates* (3.40)*, they require $a_N c_N = o(N^{-1/2})$; see Assumption 4 therein. Therefore, the case of high a_N and c_N is not permitted, although, our setting allows it. An example where $a_N \asymp N^{-1/10}$, $b_N \asymp N^{-2/5}$, $c_N \asymp N^{-1/10}$ and $d_N \asymp N^{-2/5}$ satisfies* (3.40) *but violates $a_N c_N = o(N^{-1/2})$ of [BHL20]. We introduce some specific nonparametric examples that satisfy such conditions for a_N and c_N. In low dimensions, if the*

multilayer perceptrons are utilized for the estimation of $\widehat{\nu}_a(\cdot)$ and $\widehat{\pi}(\cdot)$, Theorem 1 of [FLM21] guarantees $a_N \asymp N^{-1/10}$ and $c_N \asymp N^{-1/10}$ as long as $\beta_\nu > d/4$ and $\beta_\pi > d/4$, for $\nu(\cdot)$ and $\pi(\cdot)$ lying in the Hölder ball with smoothness β_ν and β_π, respectively. In high dimensional sparse settings, the guess-and-check forests proposed by [WW15] also achieve the desirable rates for a_N and c_N as long as the outcome Y is only dependent on at most 4 covariates; see Theorem 4 therein.

Remark 3.8 (Comparison with low-dimensional DR dynamic ATE estimators)**.** *DR dynamic ATE estimation with low-dimensional parametric nuisance models has been studied by [Rob00b, MvdLRG01, BR05, YvdL06]. Their proposed estimators for the dynamic ATE are consistent and asymptotically normal (CAN) when either 1) all the OR models are correctly specified or 2) all the PS models are correctly specified. Recently, [BRR19] proposed a new multiple robust (MR) estimator that further allows another model misspecification situation that only the OR model at time one and the PS model at time two are correctly specified. However, all of the mentioned work requires parametric nuisance estimators with low-dimensional covariates. Such nuisance estimators are $\sqrt{N}$-consistent.*

In this chapter, we allow 1) non-parametric nuisance models and 2) high-dimensional parametric nuisance models. For low and moderate dimensional covariates, we allow non-parametric nuisance estimators. Such nuisance estimators are known to be consistent to the true nuisance functions under some mild smoothness conditions. In other words, all the nuisance models can be seen as correctly specified. Unlike the previously mentioned work, no parametric assumption is needed for all the nuisance models, and our results are much more robust in the sense of model correctness.

Remark 3.9 (Comparison with an "oracle" IPW estimator)**.** *Suppose all the nuisance functions are correctly specified and all the other assumptions of Theorem 1. We compare the proposed DR estimator, $\widehat{\theta}$, with an "oracle" IPW estimator defined as follows:*

$$\widehat{\theta}_{\text{IPW}} := N^{-1}\sum_{i=1}^{N}\omega_1(\mathbf{S}_{1i},\mathbf{S}_{2i})Y_i - N^{-1}\sum_{i=1}^{N}\omega_0(\mathbf{S}_{1i},\mathbf{S}_{2i})Y_i,$$

where recall that $\omega_1(\cdot)$ defined in (3.5) is based on the true propensity score functions. Under mild conditions, we have $\sigma_{\text{IPW}}^{-1}\sqrt{N}(\widehat{\theta}_{\text{IPW}}-\theta)\to N(0,1)$ and the same asymptotic normality also holds for an "oracle" IPW estimator. When all the nuisance models are correctly specified, we can see that $\sigma_{\text{IPW}}^2 := Var\left[\frac{A_1A_2Y}{\pi(\mathbf{S}_1)\rho_a(\mathbf{S}_1,\mathbf{S}_2)} - \frac{(1-A_1)(1-A_2)Y}{(1-\pi(\mathbf{S}_1))(1-\rho_{a'}(\mathbf{S}_1,\mathbf{S}_2))}\right]$ satisfies

$$\begin{aligned}\sigma_{\text{IPW}}^2 = \sigma^2 &+ E\left[\frac{A_1}{\pi^2(\mathbf{S}_1)}\left(1-\frac{A_2}{\rho_a(\mathbf{S}_1,\mathbf{S}_2)}\right)^2\nu_a^2(\mathbf{S}_1,\mathbf{S}_2)\right]\\ &+ E\left[\frac{1-A_1}{(1-\pi(\mathbf{S}_1))^2}\left(1-\frac{1-A_2}{1-\rho_{a'}(\mathbf{S}_1,\mathbf{S}_2)}\right)^2\nu_{a'}^2(\mathbf{S}_1,\mathbf{S}_2)\right]\\ &+ E\left[\left(1-\frac{A_1}{\pi(\mathbf{S}_1)}\right)\mu_a(\mathbf{S}_1) - \left(1-\frac{1-A_1}{1-\pi(\mathbf{S}_1)}\right)\mu_{a'}(\mathbf{S}_1)\right]^2 \geq \sigma^2.\end{aligned}$$

That is, $\widehat{\theta}$ is asymptotically more efficient than the "oracle" IPW estimator. This seems to be an important corollary in itself: estimating unknown outcome models is beneficial for the inferential guarantees when comparing the size of the asymptotic variance.

We also provide the following consistency result that allows model misspecifications.

Theorem 3.5. *(Consistency rate) Suppose that one of the models $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of the models $\nu_a^*(\mathbf{S}_1,\mathbf{S}_2)$ and $\rho_a^*(\mathbf{S}_1,\mathbf{S}_2)$ is correctly specified. Let Assumptions 3.1, 3.4, 3.6, 3.7 hold. Additionally, assume that $E[\mathbb{1}_{\{A_1=a_1\}}(\mu_a(\mathbf{S}_1)-$*

$\mu_a^*(\mathbf{S}_1))^2] \leq C_\mu \sigma^2$, *with some constant* $C_\mu > 0$. *Then, the estimator* $\widehat{\theta}$ *satisfies*

$$\widehat{\theta} - \theta = O_p\Bigg(b_N c_N + a_N d_N + b_N \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}} + a_N \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \\ + c_N \sigma \mathbb{1}_{\{\mu_a^*(\cdot) \neq \mu_a(\cdot)\}} + d_N \sigma \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} + \frac{1}{\sqrt{N}} \sigma \Bigg). \tag{3.41}$$

From Theorem 3.5, we can further conclude that $\widehat{\theta} - \theta = o_p(\sigma)$ following Assumption 3.6. That is, $\widehat{\theta}$ is a consistent estimator as long as $\sigma = O(1)$ and at least one of the nuisance models is correctly specified at each exposure time. If all the nuisance models are correctly specified, we have $\widehat{\theta} - \theta = O_p(b_N c_N + a_N d_N + \sigma N^{-1/2})$. Hence, $\widehat{\theta}$ is $\sqrt{N}$-consistent as long as $b_N c_N + a_N d_N = O(N^{-1/2})$ and $\sigma = O(1)$.

Model misspecification presents here with asymmetric form in terms of the rates of estimation: (3.41) is symmetric in the rates themselves, but as b_N potentially depends on a_N, it leads to inherent asymmetries. Similar asymmetries, albeit in the low-dimensional inferential context, appear in the recent work [BRR19], where the authors allow $\mu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ to be misspecified simultaneously, but do not allow $\nu_a^*(\cdot)$ and $\pi^*(\cdot)$ being misspecified simultaneously; Theorem 3.5, however, allows for such case.

If only one of the nuisance functions is misspecified, then the consistency rate of $\widehat{\theta}$ mainly depends on 1) the estimation rate of the other nuisance function at the same time spot and 2) the product estimation rates at the other time spot. For instance, if only $\pi^*(\cdot)$ is misspecified and all the other models are correctly specified, we have $\widehat{\theta} - \theta = O_p(b_N + a_N d_N + \sigma N^{-1/2})$.

If two of the nuisance functions are misspecified at two different time spots, then the consistency rate of $\widehat{\theta}$ mainly depends on the estimation rates of the other two correctly specified nuisance models. For instance, if only $\pi^*(\cdot)$ and $\nu_a^*(\cdot)$ are misspecified, we have

$\widehat{\theta} - \theta = O_p(b_N + d_N\sigma + \sigma N^{-1/2})$.

3.6 Numerical Experiments

We illustrate the finite sample properties of the introduced estimator on a number of simulated experiments. We focus on the estimation of $\theta = \theta_a - \theta_{a'}$ where $a = (1, 1)$ and $a' = (0, 0)$. We first consider data generating processes (DGPs) where all the models are correctly specified; see Section 3.6.1. Then, in Section 3.6.2, we consider DGPs where one of the nuisance functions is possibly misspecified.

3.6.1 Correctly specified models

We consider models that all the nuisance functions are correctly specified. Generate covariates at time $t = 1$: for each $i \leq N$,

$$\mathbf{S}_{1i} \sim^{\text{iid}} N_{d_1}(\mathbf{0}, \mathbf{I}_{d_1}).$$

The treatment indicators at time $t = 1$ are generated as

$$A_{1i}|\mathbf{S}_{1i} \sim \text{Bernoulli}(\pi(\mathbf{S}_{1i})), \text{ with } \pi(\mathbf{S}_{1i}) = g(\mathbf{V}_i^T\boldsymbol{\gamma})$$

and $g(u) = \exp(u)/\{1 + \exp(u)\}$ is the logistic function. The noise variables are $\delta_{1i} \sim^{\text{iid}} N(0, 1)$, $\boldsymbol{\delta}_{1i} \sim^{\text{iid}} N_{d_1}(0, \mathbf{I}_{d_1})$ and $\boldsymbol{\delta}_{2i} \sim^{\text{iid}} N_{d_2}(0, \mathbf{I}_{d_2})$. The following models on $\mathbf{S}_{2i}|(\mathbf{S}_{1i}, A_{1i})$ are considered.

M1. (Shifting model) $\mathbf{S}_{2i} = \mathbf{S}_{1i} + A_{1i}(1 + \delta_{1i})\mathbf{1}_{d_1\times 1} + \boldsymbol{\delta}_{1i}$, where $\mathbf{1}_{d_1\times 1} = (1, \ldots, 1)^T$.

M2. (Sparse linear) $\mathbf{S}_{2i} = W_s(A_{1i})\mathbf{S}_{1i} + A_{1i}(1 + \delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$.

M3. (Dense linear) $\mathbf{S}_{2i} = W_d(A_{1i})\mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$.

M4. (Dense quadratic) $\mathbf{S}_{2i} = 0.5\widetilde{W}_d(A_{1i})(\mathbf{S}_{1i}^2 - 1) + W_d(A_{1i})\mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$, where $\mathbf{S}_{1i}^2 \in \mathbb{R}^{d_1}$ is the coordinate-wise square of $\mathbf{S}_{1i}$.

For each $c = (c_1, c_2) \in \{a, a'\}$, the matrices $W_s(c), W_d(c), \widetilde{W}_d(c) \in \mathbb{R}^{d_2\times d_1}$ are defined as the following: for each $i \le d_2$ and $j \le d_1$,

$$\{W_s(a)\}_{i,j} = 0.8^{|i-j|}\mathbb{1}\{|i-j| \le 1\}, \quad \{W_d(a)\}_{i,j} = 0.8^{|i-j|},$$

$$\{W_s(a')\}_{i,j} = 0.7^{|i-j|}\mathbb{1}\{|i-j| \le 2\}, \quad \{W_d(a')\}_{i,j} = 0.7^{|i-j|},$$

$$\{\widetilde{W}_d(c)\}_{i,j} = \{W_d(c)\}_{i,j}\mathbb{1}\{j > 3\} \text{ for each } c \in \{a, a'\}.$$

The treatment indicators at time $t = 2$ are generated as

$$A_{2i}|(\mathbf{S}_{1i}, \mathbf{S}_{2i}, A_{1i} = c_1) \sim \text{Bernoulli}(\rho_c(\mathbf{S}_{1i}, \mathbf{S}_{2i})), \text{ with}$$

$$\rho_c(\mathbf{S}_{1i}, \mathbf{S}_{2i}) = g(c_1\mathbf{U}_i^T\boldsymbol{\eta}_a + (1-c_1)\mathbf{U}_i^T\boldsymbol{\eta}_{a'}), \text{ for each } c = (c_1, c_2) \in \{a, a'\}.$$

The outcome variables are generated as

$$Y_i(c) = \mathbf{U}_i^T\boldsymbol{\alpha}_c + \zeta_i, \text{ for each } c \in \{a, a'\}, \text{ where } \zeta_i \sim^{\text{iid}} N(0,1),$$

$$Y_i = A_{1i}A_{2i}Y_i(a) + (1-A_{1i})(1-A_{2i})Y_i(a').$$

Let $\mathbf{0}_q := (0, \ldots, 0) \in \mathbb{R}^q$ for any $q \ge 1$. The parameter values are $\boldsymbol{\alpha}_c = (\boldsymbol{\alpha}_{c,1}^T, \boldsymbol{\alpha}_{c,2}^T)^T$, for each $c \in \{a, a'\}$, where $\boldsymbol{\alpha}_{a,1} = (-1, -1, 1, -1, \mathbf{0}_{(d_1-3)})^T$, $\boldsymbol{\alpha}_{a,2} = (-1, -1, 1, \mathbf{0}_{(d_2-3)})^T$, $\boldsymbol{\alpha}_{a',1} = (1, 1, 1, -1, \mathbf{0}_{(d_1-3)})^T$, $\boldsymbol{\alpha}_{a',2} = (1, 1, 1, \mathbf{0}_{(d_2-3)})^T$. Additionally, $\boldsymbol{\gamma} = (0, 1, 1, 1, \mathbf{0}_{(d_1-3)})^T$, $\boldsymbol{\eta}_a = (0, 1, 1, \mathbf{0}_{(d_1-2)}, 1, -1, \mathbf{0}_{(d_2-2)})^T$, and $\boldsymbol{\eta}_{a'} = (0, 0.5, 0, -0.5, \mathbf{0}_{(d_1-3)}, 0.5, 0, 0.5, \mathbf{0}_{(d_2-3)})^T$,

Under the above DGPs, we have the following nuisance functions: for each $c \in \{a, a'\}$,

$$\nu_c(\mathbf{S}_1, \mathbf{S}_2) = E[Y(c)|\mathbf{S}_1, \mathbf{S}_2, A_1 = c_1] = \mathbf{U}^T \boldsymbol{\alpha}_c, \tag{3.42}$$

$$\mu_c(\mathbf{S}_1) = E[Y(c)|\mathbf{S}_1, A_1 = c_1] = \mathbf{V}^T \boldsymbol{\alpha}_{c,1} + E[\mathbf{S}_2^T \boldsymbol{\alpha}_{c,2}|\mathbf{S}_1, A_1 = c_1] = \mathbf{V}^T \boldsymbol{\beta}_c, \tag{3.43}$$

where $\boldsymbol{\beta}_c$ varies for different models on $\mathbf{S}_{2i}|(\mathbf{S}_{1i}, A_{1i})$ as follows:

M1. $\boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2} \boldsymbol{\alpha}_{a',2} \mathbb{1}\{c = a'\}, \boldsymbol{\alpha}_{c,2}^T)^T$ with $\|\boldsymbol{\beta}_c\|_0 = 4$.

M2. $\boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2} \boldsymbol{\alpha}_{a',2} \mathbb{1}\{c = a'\}, (W_s(c)\boldsymbol{\alpha}_{c,2})^T)^T$ with $\|\boldsymbol{\beta}_a\|_0 = 4$ and $\|\boldsymbol{\beta}_{a'}\|_0 = 5$.

M3-4. $\boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2} \boldsymbol{\alpha}_{a',2} \mathbb{1}\{c = a'\}, (W_d(c)\boldsymbol{\alpha}_{c,2})^T)^T$ is weakly sparse in that $\|\boldsymbol{\beta}_a\|_0 = \|\boldsymbol{\beta}_{a'}\|_0 = d_1 + 1$, $\|\boldsymbol{\beta}_a\|_1 < 5.23$, and $\|\boldsymbol{\beta}_{a'}\|_1 < 7.24$.

In addition, we consider another DGP:

M5. (Dense $\nu_a(\cdot)$ and $\pi(\cdot)$) Everything is the same as in M1-M4, except the following:

$$\{\mathbf{S}_{1i,j}\}_{i \le N, j \le d_1} \sim^{\text{iid}} \text{Uniform}(-1, 1), \quad \{\boldsymbol{\delta}_{i,j}\}_{i \le N, j \le d_2} \sim^{\text{iid}} \text{Uniform}(-1, 1),$$

$$\mathbf{S}_{2i,1} = \boldsymbol{\delta}_{i1} + 3A_{1i}\mathbf{S}_{1i,1} - 2(1 - A_{1i})\mathbf{S}_{1i,1} \text{ for } 1 \le i \le N, \text{ and}$$

$$\mathbf{S}_{2i,j} = \boldsymbol{\delta}_{i,j} \text{ for } 1 \le i \le N \text{ and } 2 \le j \le d_2,$$

with $\boldsymbol{\alpha}_a = (-1, \mathbf{a}_3, \mathbf{0}_{(d_1-3)}, \mathbf{a}_{20}, \mathbf{0}_{(d_2-20)})^T$, $\boldsymbol{\alpha}_{a'} = (1, -\mathbf{a}_3, \mathbf{0}_{(d_1-3)}, \mathbf{a}_{20}, \mathbf{0}_{(d_2-20)})^T$, $\boldsymbol{\gamma} = (0, \mathbf{a}_{20}, \mathbf{0}_{(d_1-20)})^T$, $\boldsymbol{\eta}_a = (0, \mathbf{a}_3, \mathbf{0}_{(d_1-3)}, \mathbf{a}_3, \mathbf{0}_{(d_2-3)})^T$, $\boldsymbol{\eta}_{a'} = -(0, \mathbf{a}_3, \mathbf{0}_{(d_1-3)}, \mathbf{a}_3, \mathbf{0}_{(d_2-3)})^T$, where $\mathbf{a}_3 := \frac{1}{\sqrt{3}}(1, 1, 1) \in \mathbb{R}^3$ and $\mathbf{a}_{20} := \frac{1}{\sqrt{20}}(1, \ldots, 1) \in \mathbb{R}^{20}$. Under M5, we have the nuisance functions (3.42) and (3.43) with

$$\boldsymbol{\beta}_a = \left(-1, \frac{4}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \mathbf{0}_{(d_1-3)}\right)^T \text{ and } \boldsymbol{\beta}_{a'} = \left(1, -\frac{3}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \mathbf{0}_{(d_1-3)}\right)^T.$$

Table 3.2: Simulation under M1. Bias: empirical bias; RMSE: root mean square error; Length: average length of the 95% confidence intervals; Coverage: average coverage of the 95% confidence intervals; ESD: empirical standard deviation; ASD: average of estimated standard deviations. All the reported values (except Coverage) are based on robust (median-type) estimates. Denote N_1 and N_0 as the expected number of observations in the treatment groups $(1,1)$ and $(0,0)$, respectively.

$\widehat{\rho}_a(\cdot)$	$\widehat{\mu}_a(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 1000, N_1 = 294, N_0 = 282, d_1 = 100, d_2 = 100$					
empdiff		0.734	0.734	0.957	0.138	0.234	0.244
oracle		0.003	0.220	1.091	0.954	0.325	0.278
log-Lasso	Lasso	0.130	0.203	0.882	0.882	0.264	0.225
	gLasso	0.128	0.197	0.876	0.890	0.265	0.224
	elasticnet	0.152	0.208	0.881	0.868	0.268	0.225
log-gLasso	Lasso	0.130	0.202	0.868	0.888	0.264	0.221
	gLasso	0.124	0.196	0.860	0.890	0.261	0.219
	elasticnet	0.152	0.206	0.867	0.864	0.267	0.221
log-elasticnet	Lasso	0.136	0.200	0.878	0.886	0.262	0.224
	gLasso	0.137	0.197	0.869	0.888	0.260	0.222
	elasticnet	0.157	0.212	0.874	0.868	0.260	0.223
		$N = 4000, N_1 = 1178, N_0 = 1128, d_1 = 100, d_2 = 100$					
empdiff		0.731	0.731	0.478	0.000	0.111	0.122
oracle		-0.006	0.121	0.602	0.956	0.178	0.153
log-Lasso	Lasso	0.035	0.097	0.490	0.932	0.139	0.125
	gLasso	0.036	0.098	0.489	0.928	0.136	0.125
	elasticnet	0.041	0.096	0.490	0.926	0.139	0.125
log-gLasso	Lasso	0.038	0.095	0.485	0.928	0.136	0.124
	gLasso	0.037	0.095	0.484	0.924	0.133	0.123
	elasticnet	0.042	0.095	0.484	0.924	0.136	0.123
log-elasticnet	Lasso	0.036	0.096	0.487	0.930	0.138	0.124
	gLasso	0.038	0.097	0.485	0.928	0.137	0.124
	elasticnet	0.041	0.094	0.486	0.926	0.138	0.124

Table 3.3: Simulation under M2. The rest of the caption details remain the same as those in Table 3.2.

$\widehat{\rho}_a(\cdot)$	$\widehat{\mu}_a(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 1000, N_1 = 279, N_2 = 312, d_1 = 100, d_2 = 50$					
empdiff		2.485	2.485	1.258	0.000	0.318	0.321
oracle		-0.035	0.243	1.305	0.972	0.350	0.333
log-Lasso	Lasso	0.063	0.218	1.121	0.934	0.326	0.286
	gLasso	0.093	0.215	1.114	0.928	0.322	0.284
	elasticnet	0.094	0.223	1.118	0.920	0.316	0.285
log-gLasso	Lasso	0.083	0.220	1.131	0.936	0.324	0.289
	gLasso	0.093	0.219	1.127	0.936	0.333	0.288
	elasticnet	0.100	0.224	1.132	0.924	0.331	0.289
log-elasticnet	Lasso	0.063	0.220	1.118	0.930	0.322	0.285
	gLasso	0.092	0.214	1.111	0.924	0.319	0.283
	elasticnet	0.094	0.219	1.116	0.920	0.307	0.285
		$N = 4000, N_1 = 1115, N_0 = 1248, d_1 = 100, d_2 = 50$					
empdiff		2.484	2.484	0.627	0.000	0.162	0.160
oracle		0.003	0.125	0.706	0.946	0.185	0.180
log-Lasso	Lasso	0.029	0.119	0.600	0.928	0.171	0.153
	gLasso	0.032	0.122	0.599	0.922	0.170	0.153
	elasticnet	0.038	0.122	0.600	0.926	0.171	0.153
log-gLasso	Lasso	0.030	0.122	0.606	0.930	0.173	0.155
	gLasso	0.033	0.123	0.604	0.922	0.176	0.154
	elasticnet	0.040	0.122	0.605	0.930	0.174	0.154
log-elasticnet	Lasso	0.029	0.119	0.597	0.924	0.167	0.152
	gLasso	0.031	0.121	0.596	0.924	0.170	0.152
	elasticnet	0.038	0.121	0.597	0.928	0.172	0.152

Notice that, in M4, although $E(\mathbf{S}_2|\mathbf{S}_1, A_1 = c_1)$ is quadratic in $\mathbf{S}_1$, $E(\mathbf{S}_2^T\alpha_{c,2}|\mathbf{S}_1, A_1 = c_1)$ is still linear on $\mathbf{S}_1$ and hence the linear model $\mu_a^*(\cdot)$ is correctly specified. The following choices of parameters are implemented: $(N, d_1) \in \{(1000, 100), (4000, 100)\}$. For M1, we set $d_2 = d_1 = 100$; for the other models (M2-M5), we set $d_2 = d_1/2 = 50$. For each of the DGPs, we repeat the simulation for 500 times. For each replication, we construct the proposed estimator $\widehat{\theta}$ based on the following estimators: for $\widehat{\nu}_a(\cdot)$ and $\widehat{\pi}(\cdot)$, we use a Lasso and a logistic estimator with a Lasso penalty (log-Lasso), respectively; for $\widehat{\rho}_a(\cdot)$, we consider logistic estimators with a Lasso penalty (log-Lasso), a grouped Lasso penalty (log-gLasso), and an elasticnet penalty (log-elasticnet); for $\widehat{\mu}_a(\cdot)$, we consider Lasso, grouped Lasso (gLasso), and elasticnet. The regularization parameters are chosen from 10-fold cross validations, the α parameter for elasticnet is chosen as 0.7. For comparison purposes, we also consider a naive empirical difference estimator (empdiff), $\widehat{\theta}_{\text{empdiff}} = \sum_{i=1}^N A_{1i}A_{2i}Y_i / \sum_{i=1}^N A_{1i}A_{2i} - \sum_{i=1}^N (1 - A_{1i})(1 - A_{2i})Y_i / \sum_{i=1}^N (1 - A_{1i})(1 - A_{2i})$, as well as an oracle estimator, $\widehat{\theta}_{\text{oracle}}$, which is constructed based on the correct nuisance functions. The results are reported in Tables 3.2-3.5.

In this section, we consider DGPs M1-M5, where the DGPs M1-M4 only different on the procedure of generating $\mathbf{S}_2$ based on $\mathbf{S}_1$ and A_1. In M1, we consider a simple shifting model that $\mathbf{S}_2$ and $\mathbf{S}_1$ can be understood as a same set of features evaluated at different time points. In M2, we consider a sparse linear dependence that $\mathbf{S}_2$ is linearly dependent on $\mathbf{S}_1$ through a sparse and dense matrix operator, where the corresponding coefficient $\boldsymbol{\beta}_a$ is a sparse vector. In M3, we consider a dense linear dependence that the corresponding coefficient $\boldsymbol{\beta}_a$ is only weakly sparse that it's $\|\cdot\|_1$ norm is bounded. In M4, we consider a dense quadratic dependence between $\mathbf{S}_2$ and $\mathbf{S}_1$ but the nuisance function $\mu_a(\mathbf{S}_1)$ is still

linear - we can see that the nuisance function can be linear even when $\mathbf{S}_2$ is not linearly dependent on $\mathbf{S}_1$. As for in M5, we consider relatively dense models for $\nu_a(\cdot)$ and $pi(\cdot)$: the density levels of $\mu_a(\cdot)$, $\nu_a(\cdot)$, $\pi(\cdot)$, and $\rho_a(\cdot)$ are 4, 24, 20, and 6, respectively.

We first consider the inference results. As demonstrated in Theorem 3.2, we should expect good coverages when $\max\{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}\} \log^2(d)/N$ is small enough. Indeed, as shown in Tables 3.2, 3.3, and 3.6, the coverages are relatively acceptable when $N = 4000$. The coverages in Tables 3.2 and 3.6 with $N = 1000$ are relatively poor. Note that, we have $d = 201$ and $d = 151$ for models M1 and M5, respectively; the expected sample sizes for estimating $\nu_a(\cdot)$ and $\mu_a(\cdot)$ are $0.4N_{a_1}$, where $a = (a_1, a_2)$ and $a_1 = a_2 \in \{0, 1\}$. In addition, we can also see relatively good covarages in Tables 3.4 and 3.5, where $\boldsymbol{\beta}_a$ is only weakly sparse.

As for the estimation performance, as illustrated in Tables 3.2-3.6, all the proposed estimators provide RMSEs close to (or even slightly better than) the RMSE of the oracle estimators. This observation coincides with our Theorems 3.2-3.5, when all the nuisance functions are correctly specified, we expect that our estimators should provide $\sqrt{N}$-consistent estimations when N is large enough that the product rate conditions are satisfied. On the other hand, the naive empirical difference estimator, $\widehat{\theta}_{\text{empdiff}}$, is not even consistent because of the appearance of confounders.

3.6.2 Misspecified models

Now, we consider misspecified nuisance functions, $\pi^*(\cdot)$, $\rho_c^*(\cdot)$, $\nu_c^*(\cdot)$, and $\mu_c^*(\cdot)$ for each $c \in \{a, a'\}$. The following DGPs are considered:

Table 3.4: Simulation under M3. The rest of the caption details remain the same as those in Table 3.2.

$\widehat{\rho}_a(\cdot)$	$\widehat{\mu}_a(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 1000, N_1 = 296, N_0 = 310, d_1 = 100, d_2 = 50$					
empdiff		2.921	2.921	1.239	0.000	0.317	0.316
oracle		0.002	0.245	1.346	0.962	0.364	0.343
log-Lasso	Lasso	0.084	0.219	1.139	0.920	0.322	0.291
	gLasso	0.084	0.227	1.137	0.920	0.315	0.290
	elasticnet	0.102	0.226	1.136	0.912	0.336	0.290
log-gLasso	Lasso	0.083	0.226	1.142	0.916	0.322	0.291
	gLasso	0.090	0.223	1.139	0.922	0.318	0.291
	elasticnet	0.105	0.220	1.140	0.914	0.320	0.291
log-elasticnet	Lasso	0.092	0.223	1.135	0.916	0.318	0.290
	gLasso	0.093	0.221	1.132	0.920	0.318	0.289
	elasticnet	0.114	0.226	1.132	0.914	0.320	0.289
		$N = 4000, N_1 = 1184, N_0 = 1240, d_1 = 100, d_2 = 50$					
empdiff		2.922	2.922	0.619	0.000	0.159	0.158
oracle		-0.006	0.137	0.710	0.946	0.202	0.181
log-Lasso	Lasso	0.019	0.113	0.608	0.934	0.166	0.155
	gLasso	0.026	0.114	0.607	0.930	0.166	0.155
	elasticnet	0.028	0.114	0.609	0.930	0.165	0.155
log-gLasso	Lasso	0.016	0.114	0.610	0.940	0.165	0.156
	gLasso	0.026	0.116	0.609	0.934	0.164	0.155
	elasticnet	0.030	0.115	0.610	0.934	0.166	0.156
log-elasticnet	Lasso	0.019	0.114	0.607	0.934	0.164	0.155
	gLasso	0.023	0.112	0.605	0.930	0.162	0.154
	elasticnet	0.029	0.113	0.607	0.930	0.162	0.155

Table 3.5: Simulation under M4. The rest of the caption details remain the same as those in Table 3.2.

$\widehat{\rho}_a(\cdot)$	$\widehat{\mu}_a(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 1000, N_1 = 296, N_0 = 310, d_1 = 100, d_2 = 50$					
empdiff		2.921	2.921	1.239	0.000	0.317	0.316
oracle		0.002	0.245	1.346	0.962	0.364	0.343
log-Lasso	Lasso	0.083	0.225	1.141	0.924	0.318	0.291
	gLasso	0.098	0.229	1.136	0.918	0.324	0.290
	elasticnet	0.102	0.226	1.139	0.916	0.321	0.291
log-gLasso	Lasso	0.081	0.222	1.144	0.918	0.326	0.292
	gLasso	0.088	0.228	1.143	0.926	0.322	0.292
	elasticnet	0.103	0.227	1.145	0.918	0.323	0.292
log-elasticnet	Lasso	0.087	0.215	1.136	0.924	0.312	0.290
	gLasso	0.098	0.232	1.135	0.922	0.318	0.290
	elasticnet	0.107	0.223	1.137	0.914	0.324	0.290
		$N = 4000, N_1 = 1184, N_0 = 1240, d_1 = 100, d_2 = 50$					
empdiff		2.922	2.922	0.619	0.000	0.159	0.158
oracle		-0.006	0.137	0.710	0.946	0.202	0.181
log-Lasso	Lasso	0.019	0.114	0.610	0.936	0.166	0.156
	gLasso	0.025	0.114	0.609	0.932	0.166	0.155
	elasticnet	0.030	0.113	0.609	0.932	0.164	0.155
log-gLasso	Lasso	0.017	0.113	0.611	0.934	0.164	0.156
	gLasso	0.023	0.115	0.609	0.930	0.166	0.155
	elasticnet	0.027	0.116	0.611	0.930	0.164	0.156
log-elasticnet	Lasso	0.017	0.112	0.608	0.934	0.163	0.155
	gLasso	0.025	0.115	0.607	0.930	0.164	0.155
	elasticnet	0.030	0.112	0.608	0.930	0.161	0.155

Table 3.6: Simulation under M5. The rest of the caption details remain the same as those in Table 3.2.

$\widehat{\rho}_a(\cdot)$	$\widehat{\mu}_a(\cdot)$	Bias	RMSE	Length	Coverage	ESD	ASD
		$N = 1000, N_1 = 296, N_0 = 310, d_1 = 100, d_2 = 50$					
empdiff		0.418	0.418	0.475	0.072	0.114	0.121
oracle		0.004	0.096	0.500	0.952	0.144	0.128
log-Lasso	Lasso	0.065	0.106	0.487	0.900	0.139	0.124
	gLasso	0.059	0.102	0.489	0.910	0.139	0.125
	elasticnet	0.057	0.105	0.490	0.928	0.136	0.125
log-gLasso	Lasso	0.080	0.114	0.500	0.890	0.140	0.128
	gLasso	0.065	0.107	0.505	0.910	0.143	0.129
	elasticnet	0.067	0.106	0.505	0.908	0.142	0.129
log-elasticnet	Lasso	0.066	0.105	0.482	0.904	0.141	0.123
	gLasso	0.057	0.105	0.485	0.912	0.140	0.124
	elasticnet	0.056	0.107	0.485	0.918	0.136	0.124
		$N = 4000, N_1 = 1184, N_0 = 1240, d_1 = 100, d_2 = 50$					
empdiff		0.416	0.416	0.237	0.000	0.059	0.061
oracle		-0.001	0.041	0.258	0.946	0.061	0.066
log-Lasso	Lasso	0.015	0.043	0.239	0.934	0.066	0.061
	gLasso	0.012	0.042	0.239	0.928	0.064	0.061
	elasticnet	0.012	0.042	0.239	0.932	0.065	0.061
log-gLasso	Lasso	0.016	0.043	0.243	0.936	0.068	0.062
	gLasso	0.012	0.043	0.244	0.940	0.066	0.062
	elasticnet	0.011	0.043	0.244	0.942	0.065	0.062
log-elasticnet	Lasso	0.015	0.043	0.237	0.928	0.066	0.061
	gLasso	0.013	0.042	0.238	0.930	0.064	0.061
	elasticnet	0.013	0.042	0.238	0.930	0.065	0.061

M6. Non-logistic $\pi(\cdot)$ and $\rho_c(\cdot)$. Let $\pi(\mathbf{S}_{1i}) = \tilde{g}(\mathbf{V}_i^T\boldsymbol{\gamma})$ and $\rho_c(\mathbf{S}_{1i}, \mathbf{S}_{2i}) = \tilde{g}(c_1\mathbf{U}_i^T\boldsymbol{\eta}_a + (1 - c_1)\mathbf{U}_i^T\boldsymbol{\eta}_{a'})$, where $\tilde{g}(u) = (|u+1| + 0.1)/(|u+1| + 1)$. All the other processes are the same as in M2 in Section 3.6.1.

M7. Non-linear $\mu_c(\cdot)$ and $\nu_c(\cdot)$. Let $Y_i(c) = \mathbf{U}_i^T\boldsymbol{\alpha}_c + 0.5(\mathbf{S}_{1i}^T\boldsymbol{\alpha}_{c,1}[-1])^2 + \zeta_i$, where $\boldsymbol{\alpha}_{c,1} = (\boldsymbol{\alpha}_{c,1}[1], \boldsymbol{\alpha}_{c,1}[-1]^T)^T$. All the other processes are the same as in M2 in Section 3.6.1. It follows that

$$\nu_c(\mathbf{S}_1, \mathbf{S}_2) = E[Y(c)|\mathbf{S}_1, \mathbf{S}_2, A_1 = c_1] = \mathbf{U}^T\boldsymbol{\alpha}_c + 0.5(\mathbf{S}_1^T\boldsymbol{\alpha}_{c,1}[-1])^2,$$

$$\mu_c(\mathbf{S}_1) = E[Y(c)|\mathbf{S}_1] = \mathbf{V}^T\boldsymbol{\beta}_c + 0.5(\mathbf{S}_1^T\boldsymbol{\alpha}_{c,1}[-1])^2.$$

M8. Non-linear $\mu_c(\cdot)$ and $\nu_c(\cdot)$ with some bivariate features. Generate $\mathbf{W}_{2i} = W_s(A_{1i})\mathbf{S}_{1i} + A_{1i}\mathbf{1}_{d_2\times 1}$, $\mathbf{S}_{2i}[j]|\mathbf{W}_{2i} \sim$ Bernoulli$(g(\mathbf{W}_{2i}[j]))$ for each $j \leq 2$, and $\mathbf{S}_{2i}[j] = \mathbf{W}_{2i}[j] + A_{1i}\delta_{1i}\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$ for each $j \geq 3$. Let $Y_i(c) = \boldsymbol{\alpha}_{c,1}[1] + (2\mathbf{S}_{2i}[1] - 1)\mathbf{S}_{1i}^T\boldsymbol{\alpha}_{c,1}[-1] + \mathbf{S}_{2i}^T\boldsymbol{\alpha}_{c,2} + \zeta_i$. All the other processes are the same as in M2 in Section 3.6.1. It follows that

$$\nu_c(\mathbf{S}_1, \mathbf{S}_2) = E[Y(c)|\mathbf{S}_1, \mathbf{S}_2, A_1 = c_1] = \boldsymbol{\alpha}_{c,1}[1] + (2\mathbf{S}_{2i}[1] - 1)\mathbf{S}_{1i}^T\boldsymbol{\alpha}_{c,1}[-1] + \mathbf{S}_{2i}^T\boldsymbol{\alpha}_{c,2},$$

$$\begin{aligned}\mu_c(\mathbf{S}_1) = E[Y(c)|\mathbf{S}_1] = \boldsymbol{\alpha}_{c,1}[1] &+ \left(2g(\mathbf{S}_{1i}^T W_{s,1}(c) + c_1) - 1\right)\mathbf{S}_{1i}^T\boldsymbol{\alpha}_{c,1}[-1] \\ &+ \sum_{j=1}^{2}\boldsymbol{\alpha}_{c,2}[j]g(\mathbf{S}_{1i}^T W_{s,j}(c) + c_1) + \sum_{j=3}^{d_2}\boldsymbol{\alpha}_{c,2}[j]\left(\mathbf{S}_{1i}^T W_{s,j}(c) + c_1\right),\end{aligned}$$

where $W_{s,j}(c)$ is the j-th row of the matrix $W_s(c)$.

M9. Non-linear $\mu_c(\cdot)$ and non-logistic $\rho_c(\cdot)$. Let $\rho_c(\mathbf{S}_{1i}, \mathbf{S}_{2i}) = \tilde{g}(c_1\mathbf{U}_i^T\boldsymbol{\eta}_a + (1 - c_1)\mathbf{U}_i^T\boldsymbol{\eta}_{a'})$ and generate $\mathbf{S}_{2i} = 0.5W_s(A_{1i})(\mathbf{S}_{1i}^2 - 1) + W_s(A_{1i})\mathbf{S}_{1i} + A_{1i}(1 + \delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$, where $\mathbf{S}_{1i}^2 \in \mathbb{R}^{d_1}$ is the coordinate-wise square of $\mathbf{S}_{1i}$. All the other processes are the same as in Section 3.6.1.

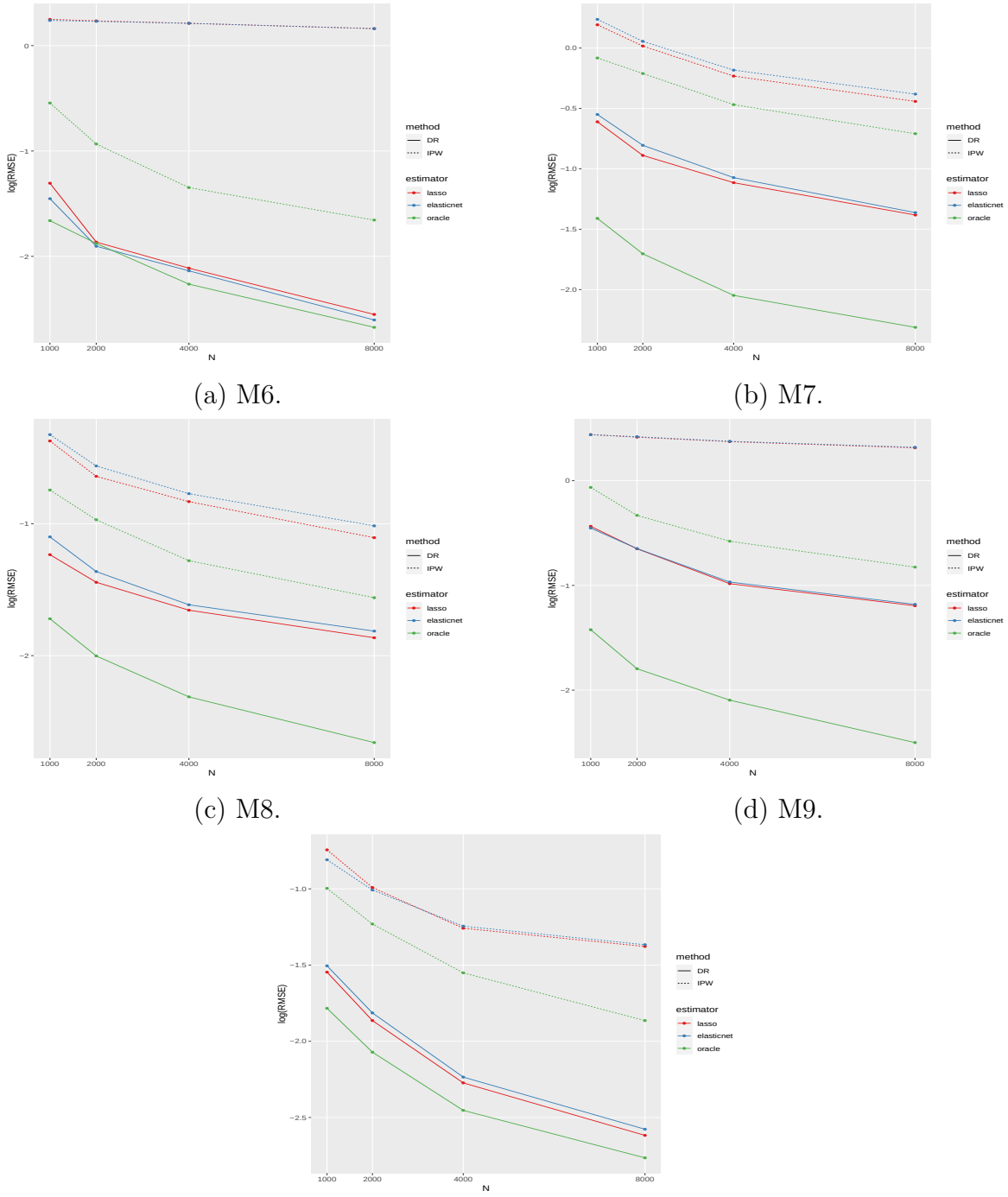
M10. Non-linear $\nu_c(\cdot)$ and non-logistic $\pi(\cdot)$. Let $\pi(\mathbf{S}_{1i}) = \tilde{g}(\mathbf{V}_i^T\boldsymbol{\gamma})$ and generate $\mathbf{W}_{2i} = \mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$ and $\mathbf{S}_{2i}[j] = \text{sgn}(\mathbf{W}_{2i}[j])|\mathbf{W}_{2i}[j]|^{1/2}$ for each $j \le d_2$. Let $Y_i(c) = \mathbf{V}_i^T\boldsymbol{\alpha}_{c,1} + \sum_{j=1}^{d_2}\alpha_{c,2}[j]\text{sgn}(\mathbf{S}_{2i}[j])\mathbf{S}_{2i}^2[j] + \zeta_i$. All the other processes are the same as in Section 3.6.1. It follows that

$$\nu_c(\mathbf{S}_1, \mathbf{S}_2) = E[Y(a)|\mathbf{S}_1, \mathbf{S}_2, A_1 = c_1] = \mathbf{V}^T\boldsymbol{\alpha}_{c,1} + \sum_{j=1}^{d_2}\alpha_{c,2}[j]\text{sgn}(\mathbf{S}_2[j])\mathbf{S}_2^2[j],$$

$$\mu_c(\mathbf{S}_1) = E[Y(a)|\mathbf{S}_1] = \mathbf{V}^T\boldsymbol{\beta}_c, \quad \text{where } \boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2}\boldsymbol{\alpha}_{a',2}\mathbb{1}\{c = a'\}, \boldsymbol{\alpha}_{c,2}^T)^T.$$

For M6-M9, we set $d_1 = 100$, $d_2 = 50$; for M10, we set $d_1 = d_2 = 100$. The sample size N varies from $\{1000, 2000, 4000, 8000\}$. We repeat the simulation 500 times for each of the DGPs. For each replication, we construct the proposed estimator $\widehat{\theta}$ based on the following estimators: a Lasso based estimator that all the nuisance functions are estimated using a linear (or logistic) regression with a Lasso penalty; an elasticnet-based estimator that all the nuisance functions are estimated using a linear (or logistic) regression with an elasticnet penalty, where $\alpha = 0.7$; an oracle estimator that all the nuisance functions are based on the true nuisance functions. We also implement IPW-based estimators for comparison purposes, which are special types of our proposed DR estimator with the outcome nuisance functions forced to be zeros. We report the root mean squared errors (RMSEs) of the estimators as N varies; see Figure 3.2.

All the DGPs M6-M10 are under the situation that two nuisance functions are correctly specified, and the other two nuisance functions are misspecified. Based on Theorem 3.5, we should expect that our proposed DR estimators to have consistency rates at most $O_p(\sigma\sqrt{s\log(d)/N})$, where the sparsity level $s \le 3$ under our DGPs. Such an upper bound is, in general, slower than the consistency rate of the DR-oracle estimator, $O_p(\sigma/\sqrt{N})$. For



(a) M6.

(b) M7.

(c) M8.

(d) M9.

(e) M10.

Figure 3.2: The root mean square errors of the proposed estimators in M6-M10 as N varies.

the WIPW-lasso and IPW-elasticnet estimators, in M6, M9, and M10, where at least one of the propensity score models is misspecified, we do not expect a consistent result; in M7

and M9, where both the propensity score models are correctly specified, we should expect consistent results with rates $O_p(\sigma_{\text{WIPW}}\sqrt{s\log(d)/N})$, where $\sigma_{\text{WIPW}} \geq \sigma$ is defined in (3.9). Lastly, we expect the WIPW-oracle estimator to be consistent with rate $O_p(\sigma_{\text{WIPW}}/\sqrt{N})$, as discussed in Remark 3.9.

3.7 Discussion

This work breaks new ground in understanding the intricate details of double robust estimation in the presence of multiple time exposures. We identify new conditions for achieving rate double robustness using Lasso type estimators for evaluating the nuisance components. We showcase that three product rate conditions are necessary to guarantee root-n inference with high-dimensional confounders. When interested in using more general nuisance estimators, we identify two global conditions needed for rate double robustness: product rates between propensity and outcome at different time exposures need to be controlled at the correct rate.

This chapter identifies new theoretical ingredients leading to the new study of the robustness of dynamical treatments. Unlike classical results, we see the impact of imputation is significant and leads to certain asymmetries in the obtained results. Naturally, this leads to a need to understand whether imputation itself can be avoided or altered in a way to remove some of the undesired effects.

Our results facilitate the theory of any Lasso-type estimators with imputed outcomes; see Theorem 3.1. Typical examples appear in high-dimensional optimal dynamic treatment regimes and policy learning, e.g., [SFSL18, ZZS19, NBW21]. We develop new techniques to

show the estimators' consistency with tuning parameters of the rate $\sqrt{\log(d)/N}$, which is standard for non-imputed lasso in the high-dimensional statistics literature. Additionally, our work also potentially promotes the development of new theoretical foundations of non-stationary reinforcement learning. Our results suggest that if the reward model varies across time, the estimation error accumulates among the time periods.

Inferential questions allowing model misspecification are now understood to be significantly different in low and high-dimensional settings. Naturally, further open questions remain unanswered: can model misspecification be allowed in high-dimensional inferential tasks? Our results would imply that possibly a new type of nuisance estimators would be required. Lastly, we would like to further understand the impact of sparsity on inference with multiple exposures.

3.8 Proofs of main results

3.8.1 Convergence rates for nuisance parameters

The following lemmas will be helpful in our proofs.

Lemma 3.3. *Let $X \in \mathbb{R}$ be a random variable. If $E(|X|^{2k}) \leq 2\sigma^{2k}\Gamma(k+1)$ for any $k \in \mathbb{N}$, then $\|X\|_{\psi_2} \leq 2\sigma$. Here, $\Gamma(a) := \int_0^\infty x^{a-1}\exp(-x)dx \ \forall a > 0$ denotes the Gamma function.*

The following lemma provides the same type of results as used in the Assumption 3.2 but now for covariates at different exposure time and different treatment paths.

Lemma 3.4. *Let Assumption 3.2 and the overlap conditions of Assumption 3.1 hold. Consider the constants c_0, κ_l, σ_u defined as in Assumptions 3.1 and 3.2. Then, the following*

statements hold:

a) $0 < c_0\kappa_l \leq \lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) \leq \lambda_{\max}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) \leq 2\sigma_u^2 < \infty$ *and* $\widetilde{\mathbf{U}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^T\widetilde{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d+1}$*;*

b) $0 < \kappa_l \leq \lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \leq \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \leq 2\sigma_u^2 < \infty$ *and* $\bar{\mathbf{U}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^T\bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d+1}$*;*

c) $0 < \kappa_l \leq \lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) \leq \lambda_{\max}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) \leq 2\sigma_u^2 < \infty$ *and* $\bar{\mathbf{V}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^T\bar{\mathbf{V}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$*;*

d) $0 < \kappa_l \leq \lambda_{\min}(E[\mathbf{V}\mathbf{V}^T]) \leq \lambda_{\max}(E[\mathbf{V}\mathbf{V}^T]) \leq 2\sigma_u^2 < \infty$ *and* $\mathbf{V}$ *is sub-Gaussian with* $\|\boldsymbol{x}^T\mathbf{V}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$*.*

The following lemma provides an asymptotic upper bounds on the estimation errors of the propensity score models, $\pi^*(\cdot)$ and $\rho_a^*(\cdot)$.

Lemma 3.5. *Let Assumption 3.2 holds and the overlap conditions of Assumption 3.1 hold. Let the sample size be such that* $N \gg \max\{s_{\boldsymbol{\gamma}}\log(d_1), s_{\boldsymbol{\delta}_a}\log(d)\}$*. Then, as* $N \to \infty$*, a) the logistic Lasso* (3.25) *with* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log(d_1)}{N}}$ *satisfies*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log(d_1)}{N}}\right), \tag{3.44}$$

$$E[\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1)]^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log(d_1)}{N}\right), \tag{3.45}$$

whereas b) the logistic Lasso (3.29) *with* $\bar{\lambda}_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log(d)}{N}}$ *satisfies*

$$\|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right), \tag{3.46}$$

$$E[\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S})]^2 = O_p\left(\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}\right). \tag{3.47}$$

In the left-hand side of (3.45) *and* (3.47)*, the expectations are only taken w.r.t. the distribution of the new observations* $\mathbf{S}_1$ *and* $(\mathbf{S}_1, \mathbf{S}_2)$*, respectively.*

Proof of Theorem 3.1. By definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\frac{1}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}}]^2 + \lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^T\boldsymbol{\beta}^*]^2 + \lambda_M\|\boldsymbol{\beta}^*\|_1,$$

or, expanding and rearranging,

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 + \lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \\
&\quad\le \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^T\boldsymbol{\beta}^*]\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda_M\|\boldsymbol{\beta}^*\|_1 \\
&\quad= \frac{2}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda_M\|\boldsymbol{\beta}^*\|_1. \qquad (3.48)
\end{aligned}$$

For any $t > 0$, let $\lambda_M := 16\sigma\sigma_{\mathbf{X}}(\sqrt{\frac{\log(d)}{M}} + t)$. Define the event

$$\mathcal{E}_2 := \left\{\max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \le \frac{\lambda_M}{4}\right\},$$

where $\mathbf{X}_{i,j}$ represents the j-th component of $\mathbf{X}_i$. Note that

$$\begin{aligned}
P\left(\max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right) &= P\left(\bigcup_{j=1}^{d}\left\{\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right\}\right) \\
&\le \sum_{j=1}^{d}P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right). \qquad (3.49)
\end{aligned}$$

Let $\boldsymbol{e}_j \in \mathbb{R}^d$ be the vector whose j-th element is 1 and other elements are 0s, for each $1 \le j \le d$. Since $\|\boldsymbol{e}_j^T\mathbf{X}\|_{\psi_2} \le \sigma_{\mathbf{X}}$ and $\|\varepsilon\| \le \sigma$, by Lemma D.1 (v) of [CLCL19],

$$\|\boldsymbol{e}_j^T\mathbf{X}\varepsilon\|_{\psi_1} \le \|\boldsymbol{e}_j^T\mathbf{X}\|_{\psi_2}\cdot\|\varepsilon\|_{\psi_2} \le \sigma\sigma_{\mathbf{X}}.$$

Note that, here we do not make any assumption on the sample gram matrix $\hat{\boldsymbol{\Sigma}} := M^{-1}\sum_{i=1}^{M}\mathbf{X}_i\mathbf{X}_i^T$, e.g., $\sup_{1\le j\le d}\hat{\boldsymbol{\Sigma}}_{j,j} \le 1$ as required in [Wai19, NRWY12]. Instead, we consider $\boldsymbol{e}_j^T\mathbf{X}\varepsilon$ as a sub-exponential random variable, and the Bernstein's inequality is applied

in the following to control (3.49). Recall the definition of $\boldsymbol{\beta}^*$, we have $E[\mathbf{X}\varepsilon] = 0$. By Lemma D.4 of [CLCL19], for each $1 \le j \le d$,

$$P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge 2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2\right) \le 2\exp\left(-M\epsilon^2\right), \quad \text{for any } \epsilon > 0. \tag{3.50}$$

Set $\epsilon = \sqrt{\frac{\log(d)}{M}} + \frac{\sqrt{1+8t}-1}{2}$ for any $t > 0$. When $M > \log(d)$, we have

$$\begin{aligned}
2\epsilon + \epsilon^2 &\le 2\sqrt{\frac{\log(d)}{M}} + \sqrt{1+8t} - 1 + \left(\sqrt{\frac{\log(d)}{M}} + \frac{\sqrt{1+8t}-1}{2}\right)^2 \\
&\le 2\sqrt{\frac{\log(d)}{M}} + \sqrt{1+8t} - 1 + \frac{2\log(d)}{M} + 2\left(\frac{\sqrt{1+8t}-1}{2}\right)^2 \\
&= 2\sqrt{\frac{\log(d)}{M}} + \sqrt{1+8t} - 1 + 2\sqrt{\frac{\log(d)}{M}}\cdot\sqrt{\frac{\log(d)}{M}} + 1 + 4t - \sqrt{1+8t} \\
&\le 4\sqrt{\frac{\log(d)}{M}} + 4t,
\end{aligned}$$

and hence

$$2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2 \le 4\sigma\sigma_{\mathbf{X}}\left(\sqrt{\frac{\log(d)}{M}} + t\right) = \frac{\lambda_M}{4}. \tag{3.51}$$

Additionally, we also have

$$\begin{aligned}
\epsilon^2 &= \left(\sqrt{\frac{\log(d)}{M}} + \frac{\sqrt{1+8t}-1}{2}\right)^2 \ge \frac{\log(d)}{M} + \frac{1+4t-\sqrt{1+8t}}{2} \\
&= \frac{\log(d)}{M} + \frac{8t^2}{1+4t+\sqrt{1+8t}} \ge \frac{\log(d)}{M} + \frac{4t^2}{1+2t+\sqrt{2t}}
\end{aligned}$$

Together with (3.50) and (3.51), we have, for each $1 \le j \le d$,

$$\begin{aligned}
P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right) &\le P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge 2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2\right) \\
&\le 2\exp\left(-M\epsilon^2\right) \le \frac{2}{d}\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right).
\end{aligned}$$

Together with (3.49),

$$P(\mathcal{E}_2) = P\left(\max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \le \frac{\lambda_M}{4}\right) \ge 1 - 2\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right). \tag{3.52}$$

On the event $\mathcal{E}_2$, we have

$$\left|\frac{2}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right| \le 2\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \le \lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1/2. \tag{3.53}$$

As for the second term of (3.48), by the fact that $2ab \le a^2+b^2$ for any $a, b \in \mathbb{R}$, and we set $a = \sqrt{2}[\widehat{Y}_i - Y_i^*]$, $b = \mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/\sqrt{2}$, we have

$$\begin{aligned}\left|\frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right| &\le \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]^2 + \frac{1}{2M}\sum_{i=1}^{M}\left[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2 \\ &\le 2\delta_M^2 + \frac{1}{2M}\sum_{i=1}^{M}\left[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2, \end{aligned} \tag{3.54}$$

on the event $\mathcal{E}_1 = \{M^{-1}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]^2 < \delta_M^2\}$. Multiplying the left-hand side and right-hand side of (3.48) by 2, we have

$$\begin{aligned}&\frac{2}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \\ &\quad\le \frac{4}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + \frac{4}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + 2\lambda_M\|\boldsymbol{\beta}^*\|_1.\end{aligned}$$

Together with (3.53) and (3.54), on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned}&\frac{2}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \\ &\quad\le \lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 + \frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 4\delta_M^2 + 2\lambda_M\|\boldsymbol{\beta}^*\|_1.\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \le \lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 + 2\lambda_M\|\boldsymbol{\beta}^*\|_1 + 4\delta_M^2 \\ &\quad= \lambda_M\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\|_1 + \lambda_M\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 + 2\lambda_M\|\boldsymbol{\beta}_S^*\|_1 + 4\delta_M^2,\end{aligned} \tag{3.55}$$

where $S := \{j \le d : \boldsymbol{\beta}_j^* \ne 0\}$ and note that $s = |S|$, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 = \|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}-\boldsymbol{\beta}_{S^c}^*\|_1 = \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$, and $\|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}_S^*\|_1$. By the triangle inequality,

$$\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \ge \|\boldsymbol{\beta}_S^*\|_1 - \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1. \tag{3.56}$$

By (3.55) and (3.56), on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we get that

$$\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 + \lambda_M \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \leq 3\lambda_M \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 + 4\delta_M^2. \tag{3.57}$$

By Lemma 4.5 of [ZCB21] there exist constants $\kappa_1, \kappa_2 > 0$, such that

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\Delta})^2 \geq \kappa_1\|\boldsymbol{\Delta}\|_2\left\{\|\boldsymbol{\Delta}\|_2 - \kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\Delta}\|_1\right\} \quad \text{for all } \|\boldsymbol{\Delta}\|_2 \leq 1, \tag{3.58}$$

with probability at least $1 - c_1\exp(-c_2 M)$ and some constants $c_1, c_2 > 0$. Lemma 4.5 of [ZCB21] discusses logistic loss but applies more broadly and does include the least squares loss as well.

Let $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and define

$$\mathcal{E}_3 := \left\{\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \geq \kappa_1\|\boldsymbol{\delta}\|_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1\|\boldsymbol{\delta}\|_2\right\}. \tag{3.59}$$

Let $\boldsymbol{\Delta} = \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$. Then, $\|\boldsymbol{\Delta}\|_2 = 1$ and hence by (3.58),

$$P(\mathcal{E}_3) \geq 1 - c_1\exp(-c_2 M).$$

We now condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and introduce two cases need to be separately analyzed.

Case 1. Case of $\|\boldsymbol{\delta}_S\|_1 < 4\lambda_M^{-1}\delta_M^2$. Then, by (3.57),

$$\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1 + 4\lambda_M^{-1}\delta_M^2 \leq 16\lambda_M^{-1}\delta_M^2.$$

Hence,

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \leq 20\lambda_M^{-1}\delta_M^2,$$

and

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \leq 3\lambda_M\|\boldsymbol{\delta}_S\|_1 + 4\delta_M^2 \leq 16\delta_M^2.$$

In addition, on the event $\mathcal{E}_3$,

$$\kappa_1\|\boldsymbol{\delta}\|_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1\|\boldsymbol{\delta}\|_2 \le \frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \le 16\delta_M^2.$$

It follows that,

$$\begin{aligned}\|\boldsymbol{\delta}\|_2 &\le \frac{\kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1 + \sqrt{\kappa_1^2\kappa_2^2\frac{\log(d)}{M}\|\boldsymbol{\delta}\|_1^2 + 64\kappa_1\delta_M^2}}{2\kappa_1}\\ &\le \kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1 + 4\kappa_1^{-1/2}\delta_M \le 20\kappa_2\sqrt{\frac{\log(d)}{M}}\lambda_M^{-1}\delta_M^2 + 4\kappa_1^{-1/2}\delta_M\\ &\le \frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M,\end{aligned}$$

since $\lambda_M = 16\sigma\sigma_{\mathbf{X}}(\sqrt{\frac{\log(d)}{M}} + t) \ge 16\sigma\sigma_{\mathbf{X}}\sqrt{\frac{\log(d)}{M}}$.

Case 2. Case of $\|\boldsymbol{\delta}_S\|_1 \ge 4\lambda_M^{-1}\delta_M^2$. Then, by (3.57),

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 + \lambda_M\|\boldsymbol{\delta}_{S^c}\|_1 \le \lambda_M(3\|\boldsymbol{\delta}_S\|_1 + 4\lambda_M^{-1}\delta_M^2) \le 4\lambda_M\|\boldsymbol{\delta}_S\|_1, \tag{3.60}$$

and hence

$$\|\boldsymbol{\delta}_{S^c}\|_1 \le 4\|\boldsymbol{\delta}_S\|_1. \tag{3.61}$$

Notice that, $\|\boldsymbol{\delta}_S\|_1 \le \sqrt{s}\|\boldsymbol{\delta}_S\|_2$. It follows that

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \le 5\|\boldsymbol{\delta}_S\|_1 \le 5\sqrt{s}\|\boldsymbol{\delta}_S\|_2 \le 5\sqrt{s}\|\boldsymbol{\delta}\|_2.$$

Hence, under the event $\mathcal{E}_3$, when $M > 100{\kappa_2}^2 s\log(d)$,

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 &\ge \kappa_1\|\boldsymbol{\delta}\|_2^2 - 5\kappa_1\kappa_2\sqrt{\frac{s\log(d)}{M}}\|\boldsymbol{\delta}\|_2^2\\ &\ge \frac{\kappa_1}{2}\|\boldsymbol{\delta}\|_2^2 \ge \frac{\kappa_1}{2}\|\boldsymbol{\delta}_S\|_2^2 \ge \frac{\kappa_1}{2s}\|\boldsymbol{\delta}_S\|_1^2.\end{aligned} \tag{3.62}$$

Together with (3.60), we have

$$\frac{\kappa_1}{2s}\|\boldsymbol{\delta}_S\|_1^2 \le \frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \le 4\lambda_M\|\boldsymbol{\delta}_S\|_1.$$

Hence, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|\boldsymbol{\delta}_S\|_1 \leq 8\kappa_1^{-1} s\lambda_M. \tag{3.63}$$

By (3.61),

$$\|\boldsymbol{\delta}\|_1 \leq \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \leq 5\|\boldsymbol{\delta}_S\|_1 \leq 40\kappa_1^{-1} s\lambda_M.$$

Besides, by (3.60) and (3.63),

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \leq 4\lambda_M\|\boldsymbol{\delta}_S\|_1 \leq 32\kappa_1^{-1} s\lambda_M^2.$$

Additionally, by (3.62), when $M > 100{\kappa_2}^2 s\log(d)$,

$$\|\boldsymbol{\delta}\|_2 \leq \sqrt{\frac{2}{\kappa_1 M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2} \leq 8\kappa_1^{-1}\sqrt{s}\lambda_M.$$

To sum up, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and when $M > \max\{\log(d), 100{\kappa_2}^2 s\log(d)\}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \max\left(\frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M, 8\kappa_1^{-1}\sqrt{s}\lambda_M\right), \tag{3.64}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \max\left(20\lambda_M^{-1}\delta_M^2, 40\kappa_1^{-1} s\lambda_M\right), \tag{3.65}$$

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^T\boldsymbol{\delta})^2 \leq \max\left(16\delta_M^2, 32\kappa_1^{-1} s\lambda_M^2\right). \tag{3.66}$$

Here,

$$P(\mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - P(\mathcal{E}_2^c) - P(\mathcal{E}_3^c) = 1 - 2\exp\left(-\frac{4Mt^2}{1 + 2t + \sqrt{2t}}\right) - c_1\exp(-c_2 M).$$

The remaining claims follow by noticing that for some $\lambda_M \asymp \sigma\sqrt{\frac{\log(d)}{M}}$ and $\delta_M = o(\sigma)$, $P(\mathcal{E}_1) = 1 - o(1)$, and with $M \gg s\log(d)$ as $M \to \infty$,

$$P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - 2\exp\left(-\frac{4Mt^2}{1 + 2t + \sqrt{2t}}\right) - c_1\exp(-c_2 M) - o(1).$$

■

Proof of Corollary 3.1. Now, we consider the Lasso estimator $\widehat{\boldsymbol{\alpha}}_a$ defined as (3.18), which is constructed using the outcome $\widetilde{Y}$, covariates $\widetilde{\mathbf{U}}$ and training samples I_{-k}. Note that $\widehat{\boldsymbol{\alpha}}_a$ is a special case of $\widehat{\boldsymbol{\beta}}$, (3.33).

Let $\widehat{Y} = Y^* = \widetilde{Y}$, $\mathbf{X} = \widetilde{\mathbf{U}}$, $\mathbb{S} = (\mathbf{X}_i)_{i\in J}$, $M = \frac{(K-1)N}{K}$, and $\delta_M = 0$. By Lemma 3.4, $\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) \geq c_0\kappa_l$ and $\widetilde{\mathbf{U}}$ is sub-Gaussian with $\|\boldsymbol{x}^T\widetilde{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$. Additionally, under Assumption 3.3, $\|\zeta\|_{\psi_2} \leq \sigma\sigma_\zeta$. Here, c_0, κ_l, σ_u, σ_ζ, and σ, defined in Assumptions 3.1-3.3 and (3.32), are positive constants independent of N and d. Hence, the estimation rates of $\widehat{\boldsymbol{\alpha}}_a$ in Corollary 3.1 follows from Theorem 3.1. To show the estimation rate of $\widehat{\nu}_a(\cdot)$, (3.36), by Lemma D (iv) of [CLCL19],

$$E[\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})]^2 = E[\mathbf{U}^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \leq 2\sigma_u^2\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2^2 = O_p\left(\sigma^2\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}\right),$$

since $\|\mathbf{U}^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\|_{\psi_2} \leq \sigma_u\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2$ under Assumption 3.2. Here, the expectation is only taken w.r.t. the joint distribution of the new observations $(\mathbf{S}_1, \mathbf{S}_2)$.

■

Proof of Lemma 3.1. Let $\widehat{Y} = Y^* = \widetilde{Y}$, $\mathbf{X} = \widetilde{\mathbf{U}}$, $\mathbb{S} = (\mathbf{X}_i)_{i\in J}$, $M = \frac{(K-1)N}{K}$, and $\delta_M = 0$. Following the proof of Theorem 3.1, since $\delta_M = 0$, we have $\|\boldsymbol{\delta}_S\|_1 \geq 4\lambda^{-1}\delta_M^2$. That is, we are under Case 2. Hence, $\boldsymbol{\delta}$ is in the cone set as in (3.61). By Lemma 3.4, $\|\boldsymbol{a}^T\bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{a}\|_2$ for any $\boldsymbol{a} \in \mathbb{R}^{d+1}$ and $\lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \geq \kappa_l$. Here, σ_u and κ_l, defined in Assumption 3.2, are positive constants independent of N and d. By Theorem 15 of [RZ12], with some constants $c_3, c_4 > 0$, when $M \geq c_3 s_{\boldsymbol{\alpha}_a}\log(d+1)$,

$$\frac{1}{M}\sum_{i=1}^{M}\left\{\bar{\mathbf{U}}_i^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\right\}^2 \leq 1.5^2\lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T])\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2^2 \leq 4.5\sigma_u\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2^2,$$

with probability at least $1 - 2\exp(-c_4 M)$. In addition, by Corollary 3.1, we have

$$\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 \leq 8\kappa_1^{-1}\tilde{\lambda}_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\alpha}_a}},$$

with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2 M)$. Therefore, with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2 M) - 2\exp(-c_4 M)$,

$$\frac{1}{M}\sum_{i=1}^{M}[\bar{\mathbf{U}}_i^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \leq 288\sigma_u\kappa_1^{-2}\tilde{\lambda}_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}.$$

■

Proof of Corollary 3.2. Let $\widehat{Y} = \bar{\mathbf{U}}^T\widehat{\boldsymbol{\alpha}}_a$, $Y^* = \bar{\mathbf{U}}^T\boldsymbol{\alpha}_a^*$, $\mathbf{X} = \bar{\mathbf{V}}$, $\mathbb{S} = (\bar{\mathbf{V}}_i)_{i\in J}$, $M = \frac{(K-1)N}{K}$, and $\delta_M^2 = 288\sigma_u\kappa_1^{-2}\tilde{\lambda}_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}$. Now, for the event $\mathcal{E}_1 := \{M^{-1}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]^2 < \delta_M^2\}$, by Lemma 3.1, we have

$$P(\mathcal{E}_1) \geq 1 - 2\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right) - c_1\exp(-c_2 M) - 2\exp(-c_4 M).$$

By Lemma 3.4, $\lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) \geq \kappa_l$ and $\bar{\mathbf{V}}$ is sub-Gaussian with $\|\boldsymbol{x}^T\bar{\mathbf{V}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$, for any $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$. Additionally, under Assumption 3.3, $\|\varepsilon\|_{\psi_2} \leq \sigma\sigma_\varepsilon$. Here, κ_l, σ_u, σ_ε, and σ, defined in Assumptions 3.2, 3.3, and (3.32), are positive constants independent of N and d. Hence, the estimation rates of $\widehat{\boldsymbol{\beta}}_a$ in Corollary 3.2 follow from Theorem 3.1. To show the esitmation rate of $\widehat{\mu}_a(\cdot)$, by Lemma D (iv) of [CLCL19],

$$\begin{aligned} E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^2 &= E[\mathbf{V}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)]^2 \leq 2\sigma_u^2\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2^2 \\ &= O_p\left(\sigma^2\frac{s_{\boldsymbol{\alpha}_a}\log(d) + s_{\boldsymbol{\beta}_a}\log(d_1)}{N}\right), \end{aligned}$$

since $\|\mathbf{V}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)\|_{\psi_2} \leq \sigma_u\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2$ under Assumption 3.2. Here, the expectation is only taken w.r.t. the distribution of the new observation $\mathbf{S}_1$.

■

Proof of Lemma 3.2. In this proof, the expectations are only taken w.r.t. the distribution of the new observations $\mathbf{S}_1, \mathbf{S}_2$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). By Assumption 3.2, we have $\|\mathbf{U}^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\|_{\psi_2} \leq \sigma_u \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2$. Together with (3.17),

$$\|\mathbf{V}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)\|_{\psi_2} = \|\mathbf{U}^T\mathbf{Q}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)\|_{\psi_2} \leq \sigma_u \|\mathbf{Q}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)\|_2 = \sigma_u \|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2. \quad (3.67)$$

Note that, the ψ_2 norm here is defined through the expectation taken w.r.t. the distribution of the new observations $\mathbf{S}_1, \mathbf{S}_2$ (or only $\mathbf{S}_1$). It follows that, for any constant $r > 2$,

$$\{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^r\}^{\frac{1}{r}} = \{E|\mathbf{U}^T(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)|^r\}^{\frac{1}{r}}, \leq 2^{1/r}(r/2)^{1/2}\sigma_u \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2,$$

$$\{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^r\}^{\frac{1}{r}} = \{E|\mathbf{V}^T(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)|^r\}^{\frac{1}{r}} \leq 2^{1/r}(r/2)^{1/2}\sigma_u \|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2,$$

which follows from Lemma D.1 (iv) of [CLCL19]. From Corollary 3.1 and 3.2, we obtain that

$$\{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^r\}^{\frac{1}{r}} = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\right),$$

$$\{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^r\}^{\frac{1}{r}} = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d) + s_{\boldsymbol{\beta}_a}\log(d_1)}{N}}\right).$$

Recall the definition $\mathcal{A} := \{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1\}$. By Lemma 3.5, we have $P(\mathcal{A}) = 1 - o(1)$. By Minkowski's inequality, we have

$$\{E|\widehat{\pi}(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}} = \{E|1 + \exp(-\mathbf{V}^T\widehat{\boldsymbol{\gamma}})|^r\}^{\frac{1}{r}} \leq 1 + \{E|\exp(-\mathbf{V}^T\widehat{\boldsymbol{\gamma}})|^r\}^{\frac{1}{r}}.$$

Under Assumption 3.4, we know that

$$P\left(\frac{c_0}{1-c_0} \leq \exp(-\mathbf{V}^T\boldsymbol{\gamma}^*) \leq \frac{1-c_0}{c_0}\right) = 1. \quad (3.68)$$

which implies that

$$\begin{aligned}\{E|\exp(-\mathbf{V}^T\widehat{\boldsymbol{\gamma}})|^r\}^{\frac{1}{r}} &= \{E|\exp(-\mathbf{V}^T\boldsymbol{\gamma}^*)\exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))|^r\}^{\frac{1}{r}} \\ &\leq \frac{1-c_0}{c_0}\{E|\exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))|^r\}^{\frac{1}{r}}.\end{aligned}$$

Hence, to prove $\{E|\widehat{\pi}(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly, i.e., bounded by a constant independent of N, it suffices to show $\{E|\exp(-r\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*))|\}^{\frac{1}{r}}$ is bounded uniformly.

Let $\mu = E[\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)]$. By Assumption 3.2 and (3.17), similarly as in (3.67), we have

$$\|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_{\psi_2} \leq \sigma_u \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2. \tag{3.69}$$

Now, conditional on the event $\mathcal{A}$, we have

$$\mu \leq \sqrt{\pi}\sigma_u, \quad \|\mu\|_{\psi_2} \leq (\log 2)^{-1/2}\sqrt{\pi}\sigma_u, \tag{3.70}$$

which follows from Lemma D.1 (iv) and (ii) of [CLCL19]. Note that, in the above, the ψ_2-norm is defined through the probability measure of a new observation $\mathbf{S}_1$. By basic properties of Orlicz norm $\|X+Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}$, we have

$$\|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)-\mu\|_{\psi_2} \leq \|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|_{\psi_2} + \|\mu\|_{\psi_2} \leq [1+(\log 2)^{-1/2}\sqrt{\pi}]\sigma_u.$$

Then it follows Lemma D.1 (vii) of [CLCL19] that

$$E[\exp\{-r(\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)-\mu)\}] \leq \exp\{2r^2[1+(\log 2)^{-1/2}\sqrt{\pi}]^2\sigma_u^2\}.$$

Using (3.70), we get that

$$\{E|\exp(-r\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*))|\}^{\frac{1}{r}} \leq \exp\{-\sqrt{\pi}\sigma_u + 2r[1+(\log 2)^{-1/2}\sqrt{\pi}]^2\sigma_u^2\}, \tag{3.71}$$

which is bounded and hence $\{E|\widehat{\pi}(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly. Repeating the same procedure above, we can obtain that $\{E|\widehat{\pi}(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}$ is also bounded uniformly, which will be used later on in the proof. By (3.68), we have

$$\begin{aligned}\left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)}-\frac{1}{\pi^*(\mathbf{S}_1)}\right|^r\right\}^{\frac{1}{r}} &= \{E|\exp(-\mathbf{V}^T\boldsymbol{\gamma}^*)[\exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*))-1]|^r\}^{\frac{1}{r}} \\ &\leq \frac{1-c_0}{c_0}\{E|\exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*))-1|^r\}^{\frac{1}{r}}. \end{aligned} \tag{3.72}$$

For any $u \in \mathbb{R}$, by Taylor's Theorem, $\exp(u) = 1 + \exp(tu)u$ with some $t \in (0,1)$. Hence, with some $t \in (0,1)$

$$\begin{aligned} |\exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)) - 1| &= \exp(-t\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)| \\ &\overset{(i)}{\leq} [1 + \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))]|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)|, \end{aligned} \tag{3.73}$$

where (i) holds since for any $t \in (0,1)$ and $u \in \mathbb{R}$, $\exp(tu) \leq \exp(u)$ when $u > 0$ and $\exp(tu) \leq \exp(0) = 1$ when $u \leq 0$, and it follows that $\exp(tu) \leq 1 + \exp(u)$.

Combining (3.72) and (3.73), we have

$$\begin{aligned} \left\{ E \left| \frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)} \right|^r \right\}^{\frac{1}{r}} &\leq \frac{1 - c_0}{c_0} \{ E| \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)) - 1|^r \}^{\frac{1}{r}} \\ &\leq \frac{1 - c_0}{c_0} \left\{ E \left| [1 + \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))] \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right|^r \right\}^{\frac{1}{r}} \\ &\overset{(i)}{\leq} \frac{1 - c_0}{c_0} \left\{ E \left| \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right|^r \right\}^{\frac{1}{r}} + \frac{1 - c_0}{c_0} \left\{ E \left| \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)) \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right|^r \right\}^{\frac{1}{r}} \\ &\overset{(ii)}{\leq} \frac{1 - c_0}{c_0} \left\{ E \left| \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right|^r \right\}^{\frac{1}{r}} \\ &\quad + \frac{1 - c_0}{c_0} \left\{ E \left| \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)) \right|^{2r} \right\}^{\frac{1}{2r}} \left\{ E \left| \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \right|^{2r} \right\}^{\frac{1}{2r}}, \end{aligned}$$

where (i) holds by the Minkowski inequality; (ii) holds by the Hölder's inequality.

Recall the equation (3.71), we know that $\{E| \exp(-\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*))|^{2r}\}^{\frac{1}{2r}}$ is bounded uniformly. In addition, recall the equation (3.69), by Lemma D.1 (iv) of [CLCL19], we have

$$\{E|\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)|^r\}^{\frac{1}{r}} \leq 2^{1/r}(r/2)^{1/2}\sigma_u \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log(d_1)}{N}}\right).$$

Therefore, we obtain that

$$\left\{ E \left| \frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)} \right|^r \right\}^{\frac{1}{r}} = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log(d_1)}{N}}\right). \tag{3.74}$$

Repeating the same procedure, we obtain that $\{E|\widehat{\rho}_a(\mathbf{S})|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly and

$$\left\{E\left|\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right|^r\right\}^{\frac{1}{r}}=O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right). \tag{3.75}$$

Therefore,

$$\begin{aligned}
&\left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\pi^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right|^r\right\}^{\frac{1}{r}}\\
&\quad\overset{(i)}{\leq}\left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)}\left(\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right)\right|^r\right\}^{\frac{1}{r}}+\left\{E\left|\frac{1}{\rho_a^*(\mathbf{S})}\left(\frac{1}{\widehat{\pi}(\mathbf{S}_1)}-\frac{1}{\pi^*(\mathbf{S}_1)}\right)\right|^r\right\}^{\frac{1}{r}}\\
&\quad\overset{(ii)}{\leq}\{E|\widehat{\pi}(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}\left\{E\left|\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right|^{2r}\right\}^{\frac{1}{2r}}+\frac{1}{c_0}\left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)}-\frac{1}{\pi^*(\mathbf{S}_1)}\right|^r\right\}^{\frac{1}{r}}\\
&\quad\overset{(iii)}{=}O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log(d_1)+s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right).
\end{aligned}$$

where (i) holds by the Minkowski inequality; (ii) holds by the Hölder's inequality; (iii) holds by (3.74), (3.75), and the fact that $\{E|\widehat{\pi}(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}$ is bounded uniformly. ∎

3.8.2 Asymptotic theory for Dynamic Treatment Lasso (DTL)

Below we introduce some shorthand notations that increase the readability of the proofs. We only focus on the treatment paths $a=(1,1)$ and $a'=(0,0)$. Let $\widehat{\eta}:=(\widehat{\eta}_a,\widehat{\eta}_{a'})$, where $\widehat{\eta}_c=(\widehat{\mu}_c,\widehat{\nu}_c,\widehat{\pi},\widehat{\rho}_c)$ for each $c\in\{a,a'\}$. Here, $\widehat{\eta}=\widehat{\eta}(\{W_i\}_{i\in I_{-k}})$ are the cross-fitted nuisance estimators. Define $\check{\theta}^{(k)}:=\check{\theta}_a^{(k)}-\check{\theta}_{a'}^{(k)}$ and $\psi(W_i;\widehat{\eta}):=\psi_a(W_i;\widehat{\eta}_a)-\psi_{a'}(W_i;\widehat{\eta}_{a'})$, where $\psi_c(W;\eta_c)$ is defined as (3.4). Then,

$$\check{\theta}^{(k)}=\frac{1}{n}\sum_{i\in I_k}\psi(W_i;\widehat{\eta}),\quad\widehat{\theta}=\frac{1}{K}\sum_{k=1}^{K}\check{\theta}^{(k)},$$

where $n:=N/K=|I_k|$ for each $k\leq K$. Let $\eta^*:=(\eta_a^*,\eta_{a'}^*)$ and $\eta:=(\eta_a,\eta_{a'})$, where $\eta_c^*:=(\mu_c^*(\cdot),\nu_c^*(\cdot),\pi^*(\cdot),\rho_c^*(\cdot))$ and $\eta_c:=(\mu_c(\cdot),\nu_c(\cdot),\pi(\cdot),\rho_c(\cdot))$ for each $c\in\{a,a'\}$. When possible, we abbreviate the subscripts $(1,1)$ and $(0,0)$ by 1 and 0. For instance, $\eta_1(\cdot)=\eta_{1,1}(\cdot)$.

For each $k = 1, ..., K$, we divide $\check{\theta}^{(k)} - \theta$ into four terms T_1, T_2, T_3, T_4,

$$\check{\theta}^{(k)} - \theta = \frac{1}{n} \sum_{i \in I_k} \psi(W_i; \widehat{\eta}) - \theta := T_1 + T_2 + T_3 + T_4, \tag{3.76}$$

where

$$T_1 := E[\psi(W; \eta^*)] - \theta, \tag{3.77}$$

$$T_2 := T_2^{(k)} := E[\psi(W; \widehat{\eta}) - \psi(W; \eta^*)], \tag{3.78}$$

$$T_3 := T_3^{(k)} := \frac{1}{n} \sum_{i \in I_k} \psi(W_i; \eta^*) - E[\psi(W; \eta^*)], \tag{3.79}$$

$$T_4 := T_4^{(k)} := \frac{1}{n} \sum_{i \in I_k} [\psi(W_i; \widehat{\eta}) - \psi(W_i; \eta^*)] - E[\psi(W; \widehat{\eta}) - \psi(W; \eta^*)]. \tag{3.80}$$

We suppress the dependence on k when possible.

In this section, we consider the following nuisance estimators: $\widehat{\nu}_a(\mathbf{S})$, $\widehat{\mu}_a(\mathbf{S}_1)$, $\widehat{\pi}(\mathbf{S}_1)$ and $\widehat{\rho}_a(\mathbf{S})$, defined as (3.20), (3.21), (3.24) and (3.30), respectively. Consider the following target nuisance functions: $\nu_a^*(\mathbf{S})$, $\mu_a^*(\mathbf{S}_1)$, $\pi^*(\mathbf{S}_1)$, $\rho_a^*(\mathbf{S})$, defined as (3.8), (3.11), (3.23) and (3.28), respectively.

Lemma 3.6. *a) Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let the Assumptions in Lemma 3.2 hold. Then,*

$$T_2 = O_p\bigg(\sigma \frac{s_1 \log(d)}{N} + \sigma \sqrt{\frac{s_2 \log(d)}{N}}\bigg), \tag{3.81}$$

where T_2 is defined as (3.78) *and*

$$s_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}}}, \sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\},$$

$$s_2 := \max\{s_{\boldsymbol{\alpha}_a}(\mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}} + \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}}), s_{\boldsymbol{\beta}_a} \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}},$$

$$s_{\boldsymbol{\gamma}} \mathbb{1}_{\{\mu_a^*(\cdot) \neq \mu_a(\cdot)\}}, s_{\boldsymbol{\delta}_a} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}}\}.$$

b) Further, assume that all the nuisance models are correctly specified. Then, we have

$$T_2 = O_p\bigg(\sigma\frac{s_1\log(d)}{N}\bigg). \tag{3.82}$$

Lemma 3.7. *Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let the assumptions in Lemma 3.2 hold. Then,*

$$[E(\psi(W;\widehat{\eta}) - \psi(W;\eta^*))^2]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right), \tag{3.83}$$

$$T_4 = O_p\bigg(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\bigg), \tag{3.84}$$

where T_4 is defined as (3.80).

Proof of Theorem 3.2. In this theorem, we consider the setting where all the nuisance models are correctly specified, i.e., $\eta^* = \eta$. Note that, Assumption 3.4 is implied by Assumption 3.1 when all the nuisance models are correct.

Consistency Let $\xi := \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta$. Recall the representation (3.76), by Lemmas 3.8, 3.6, 3.10, and 3.7 in that order we have

$$\begin{aligned}
T_1 &= 0,\\
T_2^{(k)} &= O_p\bigg(\sigma\frac{s_1\log(d)}{N}\bigg),\\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right),\\
T_4^{(k)} &= O_p\bigg(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\bigg).
\end{aligned}$$

for each $k \le K$. Therefore, by Lemma 3.13 and under Assumption 3.5, we obtain that

$$\widehat{\theta} - \theta = K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) = O_p\left(\frac{1}{\sqrt{N}}\sigma\right) \tag{3.85}$$

Asymptotic Normality By Assumption 3.5, we have $s_1 \log(d) = o(\sqrt{N})$, $s_2 \log(d) = o(N)$ and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$. Together with Lemmas 3.6, 3.7 and 3.8, we have

$$\sqrt{n}\sigma^{-1}(T_1 + T_2^{(k)} + T_4^{(k)}) = o_p(1)$$

for each $k \leq K$. Hence, to demonstrate

$$\sqrt{N}\sigma^{-1}(\widehat{\theta} - \theta) = \sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \rightsquigarrow N(0,1),$$

we need to show

$$\sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K} T_3^{(k)} = \sqrt{N}\left(N^{-1}\sum_{i=1}^{N}\psi(W_i;\eta) - \theta\right) \rightsquigarrow N(0,1),$$

where $T_3^{(k)}$ is defined as (3.79). Here, $\psi_{N,i} := \psi(W_i, \eta)$ is possibly dependent with N since both W_i and η potentially depend on (d_1, d_2), and $(d_1, d_2) = (d_{1,N}, d_{2,N})$ are allowed to grow with N. Hence, $\{\psi_{N,i}\}_{N,i}$ forms a triangular array. By Lyapunov's central limit theorem, it suffices to show that, for some $t > 0$, the following Lyapunov's condition holds:

$$\lim_{n\to\infty} \frac{E|\psi(W;\eta) - \theta|^{2+t}}{n^{\frac{t}{2}}\sigma^{2+t}} = 0. \tag{3.86}$$

Step 1 In order to check Lyapunov's condition, we show that for some constant C',

$$\frac{E|\psi(W;\eta) - \theta|^{2+t}}{\sigma^{2+t}} < C'. \tag{3.87}$$

By Lemma 3.13, we have, for some constants $t > 0$ and $C_t > 0$,

$$\frac{E|\psi(W;\eta) - \theta|^{2+t}}{\sigma^{2+t}} \leq \frac{2C_t}{c_0^{4+2t}}\left(\frac{E[|\zeta|^{2+t}}{\sigma^{2+t}} + \frac{E[|\varepsilon|^{2+t}]}{\sigma^{2+t}} + \frac{E|\xi|^{2+t}}{[E|\xi|^2]^{1+\frac{t}{2}}}\right).$$

Let $\boldsymbol{e}_1 = (1, \mathbf{0}_{1\times d_1})^T$, then we write $\xi = \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta = \mathbf{V}^T(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta)$. By Assumption 3.2 and (3.17), similarly as in (3.67), we have

$$\|\xi\|_{\psi_2} = \|(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta)^T\mathbf{V}\|_{\psi_2} \leq \sigma_u\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2.$$

It follows from Lemma D.1 (iv) of [CLCL19] that

$$E[|\xi|^{2+t}] \leq 2\sigma_u^{2+t}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2^{2+t}\Gamma(2+t/2). \tag{3.88}$$

Similarly, by Assumption 3.3, we have

$$E[|\zeta|^{2+t}] \leq 2\sigma^{2+t}\sigma_\zeta^{2+t}\Gamma(2+t/2), \tag{3.89}$$

$$E[|\varepsilon|^{2+t}] \leq 2\sigma^{2+t}\sigma_\varepsilon^{2+t}\Gamma(2+t/2). \tag{3.90}$$

By Assumption 3.2 and (3.17), we also have

$$\begin{aligned} E[|\xi|^2] = E[|\mathbf{V}^T(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta)|^2] &\geq \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2^2 \cdot \lambda_{\min}(E[\mathbf{V}\mathbf{V}^T]) \\ &\geq \kappa_l\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2^2. \end{aligned} \tag{3.91}$$

Using (3.88) and (3.91), we get that

$$\frac{E|\xi|^{2+t}}{[E|\xi|^2]^{1+\frac{t}{2}}} \leq \frac{2\sigma_u^{2+t}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^* - \boldsymbol{e}_1\theta\|_2^{2+t}} = \frac{2\sigma_u^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}}. \tag{3.92}$$

Using (3.89), (3.90) and (3.92), then we obtain that

$$\frac{E|\psi(W;\eta) - \theta|^{2+t}}{\sigma^{2+t}} \leq \frac{2C_t}{c_0^{4+2t}}\left(2\sigma_\zeta^{2+t}\Gamma(2+t/2) + 2\sigma_\varepsilon^{2+t}\Gamma(2+t/2) + \frac{2\sigma_u^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}}\right)$$

Taking $C' = \frac{2C_t}{c_0^{4+2t}}\left(2\sigma_\zeta^{2+t}\Gamma(2+t/2) + 2\sigma_\varepsilon^{2+t}\Gamma(2+t/2) + \frac{2\sigma_u^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}}\right)$, we get (3.87) and hence the Lyapunov's condition is satisfied.

Step 2 In this step, the expecations are taken w.r.t. the joint distribution of $(W_i)_{i\in I_k}$.

By (3.85), we have $\widehat{\theta} - \theta = O_p(\sigma/\sqrt{N})$. Then, we show, for each $k \leq K$,

$$\left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta}) - \psi(W_i;\eta)|^2\right]^{\frac{1}{2}} = o_p(\sigma). \tag{3.93}$$

It follows from Jensen's inequality that

$$E\left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta})-\psi(W_i;\eta)|^2\right]^{\frac{1}{2}} \le \left\{E\left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta})-\psi(W_i;\eta)|^2\right]\right\}^{\frac{1}{2}}$$
$$= [E|\psi(W;\widehat{\eta})-\psi(W;\eta)|^2]^{\frac{1}{2}} = O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right),$$

where the last assertion follows from (3.83) in Lemma 3.7 with correctly specified nuisance models $\eta = \eta^*$. By Markov's inequality, we have

$$\left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta})-\psi(W_i;\eta)|^2\right]^{\frac{1}{2}} = O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right) = o_p(\sigma).$$

Therefore, using (3.85), (3.87) and (3.93), we get $\widehat{\sigma}^2 - \sigma^2 = o_p(\sigma^2)$ by Lemma 3.14.

■

Proof of Theorem 3.3. Now, we consider the case that model misspecification is allowed potentially. Suppose one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Recall the representation (3.76). By Lemmas 3.8, 3.6, 3.10, and 3.7, we have

$$\begin{aligned}
T_1 &= 0,\\
T_2^{(k)} &= O_p\left(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}}\right),\\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right),\\
T_4^{(k)} &= O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).
\end{aligned}$$

for each $k \le K$. Therefore, by Lemma 3.12, we obtain that

$$\begin{aligned}
\widehat{\theta} - \theta &= K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)})\\
&= O_p\left(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\right),
\end{aligned}$$

where

$$s_1 := \max\{\sqrt{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}_a}}, \sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}}}\},$$

$$s_2 := \max\{s_{\boldsymbol{\alpha}_a}(\mathbb{1}_{\{\pi^*(\boldsymbol{S}_1)\neq\pi(\boldsymbol{S}_1)\}} + \mathbb{1}_{\{\rho_a^*(\boldsymbol{S}_1,\boldsymbol{S}_2)\neq\rho_a(\boldsymbol{S}_1,\boldsymbol{S}_2)\}}), s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi^*(\boldsymbol{S}_1)\neq\pi(\boldsymbol{S}_1)\}},$$

$$s_{\boldsymbol{\gamma}}\mathbb{1}_{\{\mu_a^*(\boldsymbol{S}_1)\neq\mu_a(\boldsymbol{S}_1)\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\}.$$

■

3.8.3 Asymptotic theory for general dynamic treatment effect

In this section, we consider general nuisance estimators and general working models.

Lemma 3.8. *Suppose that at least one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and at least one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let Assumption 3.1 hold. Then,*

$$T_1 = 0, \tag{3.94}$$

where T_1 is defined as (3.77).

Lemma 3.9. *a) Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let Assumptions 3.1, 3.4, 3.6 and 3.7 hold. Then,*

$$\begin{aligned} T_2 = O_p\Bigg(& b_N c_N + b_N d_N + b_N \mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}} + a_N \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}} \\ & + c_N\sqrt{E[\zeta^2+\varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}} + d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\Bigg), \end{aligned} \tag{3.95}$$

where T_2 is defined as (3.78).

b) Suppose all the nuisance models are correctly specified and Assumptions 3.1, 3.6 and 3.7 hold, then we have

$$T_2 = O_p\left(b_N c_N + a_N d_N\right), \tag{3.96}$$

Lemma 3.10. *a) Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let Assumptions 3.1, 3.4 hold. Then,*

$$T_3 = O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \tag{3.97}$$

where $\xi := \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta$ and T_3 is defined as (3.79).

b) Suppose all the models are correctly specified and Assumption 3.1 holds, then we also have (3.97).

Lemma 3.11. *a) Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let Assumptions 3.1, 3.4, 3.6 and 3.7 hold. Then,*

$$T_4 = O_p\left(\frac{1}{\sqrt{N}}\left[a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right]\right), \tag{3.98}$$

where T_4 is defined as (3.80).

b) Suppose all the models are correctly specified and and Assumptions 3.1, 3.6, 3.7 and 3.8 hold, then we have

$$T_4 = O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right). \tag{3.99}$$

Lemma 3.12. *Suppose that one of $\mu_a^*(\mathbf{S}_1)$ and $\pi^*(\mathbf{S}_1)$ is correctly specified, and one of $\nu_a^*(\mathbf{S})$ and $\rho_a^*(\mathbf{S})$ is correctly specified. Let Assumption 3.1 holds. Then,*

$$\psi(W;\eta^*) - \theta = \sum_{i=1}^{8} O_i, \quad \text{and} \quad \sigma^2 := E(\psi(W;\eta^*) - \theta)^2 = \sum_{i=1}^{8} E[O_i^2],$$

where $\{O_i\}_{i=1}^8$ are defined as (3.171)-(3.178).

a) Assume that $E[\mathbb{1}_{\{A_1=a_1\}}(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2] \leq C_\mu\sigma^2$, with some constant $C_\mu > 0$. Then,

$$E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \leq \left(\frac{4}{c_0^2} + 6C_\mu\right)\sigma^2,$$

where $\sigma^2 := E(\psi(W;\eta^*) - \theta)^2$.

b) Let Assumption 3.3 holds. Then,

$$E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \leq \left(\frac{1}{c_0^2} + 2\sigma_\varepsilon^2\right)\sigma^2.$$

Lemma 3.13. *Suppose all the models are correctly specified that* $\eta^* = \eta$ *and let Assumption 3.1 holds, then we have for some constants* $t > 0$ *and* $C_t > 0$ *possibly dependent with* t*, such that*

$$\sigma^2 := E(\psi(W;\eta^*) - \theta)^2 = E(\psi(W;\eta) - \theta)^2 \geq E[\zeta^2] + E[\varepsilon^2] + E[\xi^2], \tag{3.100}$$

$$E|\psi(W;\eta) - \theta|^{2+t} \leq \frac{2C_t}{c_0^{4+2t}} E\left[|\zeta|^{2+t} + |\varepsilon|^{2+t} + |\xi|^{2+t}\right]. \tag{3.101}$$

Lemma 3.14. *Suppose all the nuisance models are correctly specified that* $\eta^* = \eta$ *and let Assumption 3.1 holds. Define* $\widehat{\sigma}_k^2 := \frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \widehat{\theta})^2$ *and* $\widehat{\sigma}^2 = \frac{1}{K}\sum_{k=1}^K \widehat{\sigma}_k^2$*. Let* $\sigma^2 := E(\psi(W;\eta^*) - \theta)^2 = E(\psi(W;\eta) - \theta)^2$*. If*

$$\widehat{\theta} - \theta = O_p(\sigma/\sqrt{N}),\ [\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta}) - \psi(W_i;\eta)|^2]^{\frac{1}{2}} = o_p(\sigma)$$

for each $k \leq K$*, and* $[E|(\psi(W;\eta) - \theta)|^{2+t}]^{\frac{2}{2+t}} < C\sigma^2$ *for some constant* C*, we have*

$$\widehat{\sigma}^2 - \sigma^2 = o_p(\sigma^2). \tag{3.102}$$

Proof of Theorem 3.4. In this theorem, we consider correctly specified nuisance models, in that $\eta^* = \eta$.

Consistency Recall the representation (3.76), by Lemmas 3.8, 3.9, 3.10, and 3.11, we have

$$T_1 = 0, \tag{3.103}$$

$$T_2^{(k)} = O_p\left(b_N c_N + a_N d_N\right), \tag{3.104}$$

$$T_3^{(k)} = O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right),$$

$$T_4^{(k)} = O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right). \tag{3.105}$$

By assumption, $b_N c_N + a_N d_N = o(\sigma N^{-1/2})$. Together with Lemma 3.13, we obtain that

$$\begin{aligned}
\widehat{\theta} - \theta &= K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \\
&= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right] + b_N c_N + a_N d_N\right) \qquad (3.106) \\
&\quad + O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right) \\
&= O_p\left(\frac{1}{\sqrt{N}}\sigma\right). \qquad (3.107)
\end{aligned}$$

Asymptotic Normality Now, we demonstrate that $\sqrt{N}\sigma^{-1}(\widehat{\theta}-\theta) \rightsquigarrow N(0,1)$. By (3.103), (3.104), and (3.105), under Assumption 3.6 and $b_N c_N + a_N d_N = o(\sigma N^{-1/2})$, we have

$$\sqrt{n}\sigma^{-1}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) = o_p(1)$$

for each $k \leq K$. Hence, we only need to show

$$\sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K} T_3^{(k)} = \sqrt{N}\left(N^{-1}\sum_{i=1}^{N}\psi(W_i;\eta) - \theta\right) \rightsquigarrow N(0,1),$$

where $T_3^{(k)}$ is defined as (3.79). By Lyapunov's central limit theorem, it suffices to show the following Lyapunov's condition holds: with some $t > 0$,

$$\lim_{n\to\infty}\frac{E|\psi(W;\eta) - \theta|^{2+t}}{n^{\frac{t}{2}}\sigma^{2+t}} = 0. \tag{3.108}$$

Step 1 To check Lyapunov's condition, it suffices to show that for some constant $C' > 0$,

$$\frac{E|\psi(W;\eta) - \theta|^{2+t}}{\sigma^{2+t}} < C'. \tag{3.109}$$

By Lemma 3.13, we have, for some constants $t > 0$ and $C_t > 0$,

$$\begin{aligned}\frac{E|\psi(W;\eta) - \theta|^{2+t}}{\sigma^{2+t}} &\leq \frac{2C_t}{c_0^{4+2t}} \frac{E[|\zeta|^{2+t} + |\varepsilon|^{2+t} + |\xi|^{2+t}]}{(E[\zeta^2] + E[\varepsilon^2] + E[\xi^2])^{1+\frac{t}{2}}} \\ &\leq \frac{2C_t}{c_0^{4+2t}} \left(\frac{E[|\zeta|^{2+t}]}{(E[\zeta^2])^{1+\frac{t}{2}}} + \frac{E[|\varepsilon|^{2+t}]}{(E[\varepsilon^2])^{1+\frac{t}{2}}} + \frac{E[|\xi|^{2+t}]}{(E[\xi^2])^{1+\frac{t}{2}}} \right) \leq \frac{2CC_t}{c_0^{4+2t}},\end{aligned} \tag{3.110}$$

where the last inequality follows from Assumption 3.8. Taking $C' = \frac{2CC_t}{c_0^{4+2t}}$, we get (3.108) so that Lyapunov's condition is satisfied.

Step 2 By (3.107), we have $\widehat{\theta} - \theta = O_p(\sigma/\sqrt{N})$. Here, we show that, for each $k \leq K$,

$$\left[\frac{1}{n} \sum_{i \in I_k} |\psi(W_i; \widehat{\eta}) - \psi(W_i; \eta)|^2 \right]^{\frac{1}{2}} = o_p(\sigma). \tag{3.111}$$

Note that

$$E\left[\frac{1}{n} \sum_{i \in I_k} |\psi(W_i; \widehat{\eta}) - \psi(W_i; \eta)|^2 \right]^{\frac{1}{2}} \overset{(i)}{\leq} \left\{ E\left[\frac{1}{n} \sum_{i \in I_k} |\psi(W_i; \widehat{\eta}) - \psi(W_i; \eta)|^2 \right] \right\}^{\frac{1}{2}} \tag{3.112}$$

$$\overset{(ii)}{=} [E|\psi(W; \widehat{\eta}) - \psi(W; \eta)|^2]^{\frac{1}{2}} \tag{3.113}$$

$$\overset{(iii)}{=} O_p\left(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N \sqrt{E[\zeta^2]} \right),$$

where in (3.112), the expectations are taken w.r.t. the joint distribution of $(W_i)_{i \in I_k}$; in (3.113), the expectation is taken w.r.t. the joint distribution of a new W. In the above, (i) holds by Jensen's inequality; (ii) holds since $\widehat{\eta}$ is independent of $\{W_i\}_{i \in I_k}$ based on cross-fitting, $\{W_i\}_{i \in I_k}$ are i.i.d. distributed and W is an independent copy of them; (iii) holds by

Lemma 3.11. By Markov's inequality, we have

$$\left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta})-\psi(W_i;\eta)|^2\right]^{\frac{1}{2}}$$
$$=O_p\left(a_N+b_N+c_N(\sqrt{E[\zeta^2]}+\sqrt{E[\varepsilon^2]})+d_N\sqrt{E[\zeta^2]}\right)=o_p(\sigma).$$

Together with (3.107), (3.108), (3.111), and Lemma 3.14, we conclude that

$$\widehat{\sigma}^2-\sigma^2=o_p(\sigma^2).$$

■

Proof of Theorem 3.5. Recall the representation (3.76). By Lemmas 3.8, 3.9, 3.10, and 3.11, we have

$$\begin{aligned}
T_1&=0,\\
T_2^{(k)}&=O_p\Big(b_Nc_N+b_Nd_N+b_N\mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}}+a_N\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\\
&\qquad+c_N\sqrt{E[\zeta^2+\varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}}+d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\Big),\\
T_3^{(k)}&=O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]}+\sqrt{E[\varepsilon^2]}+\sqrt{E[\xi^2]}\right]\right),\\
T_4^{(k)}&=O_p\left(\frac{1}{\sqrt{N}}\left[a_N+b_N+\sqrt{E[\zeta^2]}+\sqrt{E[\varepsilon^2]}\right]\right).
\end{aligned}$$

Together with Lemma 3.12 and further assume that $E(\mu_a^*(\mathbf{S}_1)-\mu_a(\mathbf{S}_1))^2\leq C_\mu\sigma^2$ with some constant $C_\mu>0$, we obtain

$$\begin{aligned}
\widehat{\theta}-\theta&=K^{-1}\sum_{k=1}^{K}(T_1+T_2^{(k)}+T_3^{(k)}+T_4^{(k)})\\
&=O_p\Big(b_Nc_N+a_Nd_N+b_N\mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}}+a_N\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\\
&\qquad+c_N\sigma\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}}+d_N\sigma\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}+\frac{1}{\sqrt{N}}\sigma\Big).
\end{aligned}$$

■

3.8.4 Proofs of Auxiliary Lemmas

Proof of Lemma 3.3. By the definition of $\|X\|_{\psi_2} = \inf\{c > 0 : E[\exp(X^2/c^2)] \leq 2\}$ and

$$E\left[\exp\left(\frac{X^2}{4\sigma^2}\right)\right] = E\left[\sum_{k=0}^{\infty} \frac{X^{2k}}{k!(4\sigma^2)^k}\right] \leq \sum_{k=0}^{\infty} \frac{2^k \sigma^{2k} \Gamma(k+1)}{k!(4\sigma^2)^k} = \sum_{k=0}^{\infty} \frac{1}{2^k} = 2,$$

therefore, leading to $\|X\|_{\psi_2} \leq 2\sigma$. ∎

Proof of Lemma 3.4. a) we observe that

$$\begin{aligned}
\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}]\boldsymbol{x} \\
&= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[E[(\mathbf{U}^T\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1, A_2=a_2\}} | \mathbf{U}, A_1 = a_1] P[A_1 = a_1|\mathbf{U}]] \\
&= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^T\boldsymbol{x})^2 \cdot P[A_2 = a_2|\mathbf{U}, A_1 = a_1] E[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{U}]] \\
&= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^T\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1\}} \cdot P[A_2 = a_2|\mathbf{U}, A_1 = a_1]]. \qquad (3.114)
\end{aligned}$$

Under the overlap conditions of Assumption 3.1,

$$P(c_0 \leq P[A_2 = a_2|\mathbf{U}, A_1 = a_1] \leq 1 - c_0) = 1.$$

Together with (3.114), under Assumption 3.2, we obtain

$$\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) \geq c_0 \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^T\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1\}}] \geq c_0\kappa_l > 0.$$

Additionally, we also have

$$\begin{aligned}
\lambda_{\max}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}]\boldsymbol{x} \\
&\leq \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T]\boldsymbol{x} = \lambda_{\max}(E[\mathbf{U}\mathbf{U}^T]) \overset{(i)}{\leq} 2\sigma_u^2,
\end{aligned}$$

where (i) holds since, by Lemma D.1 (iv) of [CLCL19],

$$\lambda_{\max}(E[\mathbf{U}\mathbf{U}^T]) = \max_{\|\boldsymbol{x}\|_2=1} E[(\boldsymbol{x}^T\mathbf{U})^2] \leq \max_{\|\boldsymbol{x}\|_2=1} 2\sigma_u^2\|\boldsymbol{x}\|_2^2 = 2\sigma_u^2. \qquad (3.115)$$

Besides, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^T \widetilde{\mathbf{U}}|^{2k}] = E[|\boldsymbol{x}^T \mathbf{U}|^{2k} \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}] \leq E[|\boldsymbol{x}^T \mathbf{U}|^{2k}] \overset{(i)}{\leq} 2(\sigma_u \|\boldsymbol{x}\|_2)^{2k} \Gamma(k+1),$$

where (i) holds by Lemma D.1 (iv) of [CLCL19]. By Lemma 3.3, we have

$$\|\boldsymbol{x}^T \widetilde{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u \|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d+1}.$$

b) Under Assumption 3.2, we also have

$$\lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) = \min_{\boldsymbol{x} \in \mathbb{R}^{d+1}: \|\boldsymbol{x}\|_2 = 1} E[(\mathbf{U}^T \boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1\}}] \geq \kappa_l > 0, \tag{3.116}$$

and by (3.115),

$$\begin{aligned} \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) &= \max_{\boldsymbol{x} \in \mathbb{R}^{d+1}: \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1=a_1\}}] \boldsymbol{x} \\ &\leq \max_{\boldsymbol{x} \in \mathbb{R}^{d+1}: \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T] \boldsymbol{x} \leq 2\sigma_u^2 < \infty. \end{aligned} \tag{3.117}$$

In addition, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^T \bar{\mathbf{U}}|^{2k}] = E[|\boldsymbol{x}^T \mathbf{U}|^{2k} \mathbb{1}_{\{A_1=a_1\}}] \leq E[|\boldsymbol{x}^T \mathbf{U}|^{2k}] \overset{(i)}{\leq} 2(\sigma_u \|\boldsymbol{x}\|_2)^{2k} \Gamma(k+1), \tag{3.118}$$

where (i) holds by Lemma D.1 (iv) of [CLCL19]. By Lemma 3.3, we have

$$\|\boldsymbol{x}^T \bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u \|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d+1}.$$

c) Recall the representation (3.17), we also have

$$\begin{aligned} \lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) &= \min_{\boldsymbol{x} \in \mathbb{R}^{d_1+1}: \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T E[\mathbf{V}\mathbf{V}^T \mathbb{1}_{\{A_1=a_1\}}] \boldsymbol{x} \\ &= \min_{\boldsymbol{x} \in \mathbb{R}^{d_1+1}: \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T E[\mathbf{Q}\mathbf{U}\mathbf{U}^T\mathbf{Q}^T \mathbb{1}_{\{A_1=a_1\}}] \boldsymbol{x} \\ &\geq \min_{\boldsymbol{x} \in \mathbb{R}^{d+1}: \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1=a_1\}}] \boldsymbol{x} = \lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \overset{(i)}{\geq} \kappa_l, \end{aligned} \tag{3.119}$$

where (i) follows from (3.116). Similarly,

$$\begin{aligned}\lambda_{\max}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{V}\mathbf{V}^T \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{Q}\mathbf{U}\mathbf{U}^T\mathbf{Q}^T \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &\le \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \overset{(i)}{\le} 2\sigma_u^2,\end{aligned}$$

where (i) follows from (3.117). In addition, for any $k \in \mathbb{N}$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\bar{\mathbf{V}}|^{2k}] &= \sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\mathbf{Q}\bar{\mathbf{U}}|^{2k}] \\ &\overset{(i)}{\le} \sup_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\bar{\mathbf{U}}|^{2k}] \overset{(ii)}{\le} 2\sigma_u^{2k}\Gamma(k+1),\end{aligned}$$

where (i) holds since, for every $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$, $\mathbf{Q}^T\boldsymbol{x} = (\boldsymbol{x}^T, 0, \ldots, 0)^T \in \mathbb{R}^{d+1}$ and hence $\|\mathbf{Q}^T\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 = 1$; (ii) follows from (3.118). Hence, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^T\bar{\mathbf{V}}|^{2k}] \le 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1).$$

By Lemma 3.3, we have $\bar{\mathbf{V}}$ is sub-Gaussian with

$$\|\boldsymbol{x}^T\bar{\mathbf{V}}\|_{\psi_2} \le 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d_1+1}.$$

d) Lastly, note that

$$\begin{aligned}\lambda_{\min}(E[\mathbf{V}\mathbf{V}^T]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{V}\mathbf{V}^T]\boldsymbol{x} \\ &\ge \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{V}\mathbf{V}^T \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^T]) \overset{(i)}{\ge} \kappa_l,\end{aligned}$$

where (i) holds by (3.119). Besides,

$$\begin{aligned}\lambda_{\max}(E[\mathbf{V}\mathbf{V}^T]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{V}\mathbf{V}^T]\boldsymbol{x} = \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{Q}\mathbf{U}\mathbf{U}^T\mathbf{Q}^T]\boldsymbol{x} \\ &\le \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T E[\mathbf{U}\mathbf{U}^T]\boldsymbol{x} = \lambda_{\max}(E[\mathbf{U}\mathbf{U}^T]) \overset{(i)}{\le} 2\sigma_u^2,\end{aligned}$$

where (i) follows from (3.117). In addition, for any $k \in \mathbb{N}$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\mathbf{V}|^{2k}] &= \sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\mathbf{Q}\mathbf{U}|^{2k}] \\ &\overset{(i)}{\leq} \sup_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^T\mathbf{U}|^{2k}] \overset{(ii)}{\leq} 2\sigma_u^{2k}\Gamma(k+1),\end{aligned}$$

where (i) holds since, for every $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$, $\|\mathbf{Q}^T\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 = 1$; (ii) follows from (3.118). Hence, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^T\mathbf{V}|^{2k}] \leq 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1).$$

By Lemma 3.3, we have $\mathbf{V}$ is also sub-Gaussian with

$$\|\boldsymbol{x}^T\mathbf{V}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d_1+1}.$$

■

Proof of Lemma 3.5. In this Lemma, we provide estimation rates for $\widehat{\boldsymbol{\gamma}}$, $\widehat{\pi}(\cdot)$, $\widehat{\boldsymbol{\delta}}_a$, and $\widehat{\rho}_a(\cdot)$. We allow model misspecifications that $\pi^*(\cdot) \neq \pi(\cdot)$ and $\rho_a^*(\cdot) \neq \rho_a(\cdot)$. Note that, classical results for generalized linear models only consider correrctly specified cases; see, e.g., Corollary 9.26 of [Wai19] and Section 4.4 of [NRWY12].

a) We first show (3.44) and (3.45). In part a), the expectations are only taken w.r.t. the distribution of the new observation $\mathbf{S}_1$.

Consider the link function $\Psi(u) = \log(1 + \exp(u))$, we have

$$\Psi''(\mathbf{V}^T\boldsymbol{\gamma}^*) = \frac{\exp(\mathbf{V}^T\boldsymbol{\gamma}^*)}{(1+\exp(\mathbf{V}^T\boldsymbol{\gamma}^*))^2} = \pi(\mathbf{S}_1)(1-\pi(\mathbf{S}_1)).$$

Under Assumption 3.4, we have $P(c_0^2 \leq \Psi''(\mathbf{V}^T\boldsymbol{\gamma}^*) \leq (1-c_0)^2) = 1$. By Lemma 3.4,

$$\lambda_{\min}(E[\mathbf{V}\mathbf{V}^T]) \geq \kappa_l > 0, \quad \lambda_{\max}(E[\mathbf{V}\mathbf{V}^T]) \leq 2\sigma_u^2 < \infty, \tag{3.120}$$

and $\mathbf{V}$ is sub-Gaussian with $\|\boldsymbol{x}^T\mathbf{V}\|_{\psi_2} \le 2\sigma_u\|\boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$.

Next, we control the gradient at the potentially misspecified location: recall that the underlying model may be misspecified and $\pi^*(\cdot)$ not necessarily equal to $\pi(\cdot)$; The true $\boldsymbol{\gamma}$ may not exists such that $\widehat{\pi}(\cdot)$ has a logistic form Below we ensure and discuss the Restricted Strong Convexity (RSC) as well as the properties of the gradient.

We first consider the RSC property. Note that, the RSC property (3.122) below only depends on the distribution of $\mathbf{S}_1$ and does not depend on the distribution of $A_1|\mathbf{S}_1$. This is because $\delta\ell_M(\boldsymbol{\Delta}, \boldsymbol{\gamma}^*)$ defined in (3.121) can be written as

$$\delta\ell_M(\boldsymbol{\Delta}, \boldsymbol{\gamma}^*) = M^{-1} \sum_{i \in I_{-k}} \left[\Psi(\mathbf{V}_i^T(\boldsymbol{\gamma}^* + \boldsymbol{\Delta})) - \Psi(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - \boldsymbol{\Delta}^T\mathbf{V}_i\Psi'(\mathbf{V}_i^T\boldsymbol{\gamma}^*)\right],$$

which is function of $\mathbf{S}_{1i}$s, and A_{1i}s are not involved above. As a result, the model misspecification for $\pi(\mathbf{S}_1) = E(A_1|\mathbf{S}_1)$ does not affect the RSC property. In below, we consider the RSC property studied by [ZCB21].

For any $\boldsymbol{\gamma}, \boldsymbol{\Delta} \in \mathbb{R}^{d_1+1}$, define

$$\ell_M(\boldsymbol{\gamma}) := M^{-1} \sum_{i \in I_{-k}} \left[-A_{1i}\mathbf{V}_i^T\boldsymbol{\gamma} + \log(1 + \exp(\mathbf{V}_i^T\boldsymbol{\gamma}))\right],$$

$$\delta\ell_M(\boldsymbol{\Delta}, \boldsymbol{\gamma}^*) := \ell_M(\boldsymbol{\gamma}^* + \boldsymbol{\Delta}) - \ell_M(\boldsymbol{\gamma}^*) - \boldsymbol{\Delta}^T\nabla\ell_M(\boldsymbol{\gamma}^*). \tag{3.121}$$

By Lemma 4.5 of [ZCB21], we have the following RSC property holds:

$$\begin{aligned}\delta\ell_M(\boldsymbol{\Delta}, \boldsymbol{\gamma}^*) &\ge \kappa_1\|\boldsymbol{\Delta}\|_2 \left\{\|\boldsymbol{\Delta}\|_2 - \kappa_2\sqrt{\frac{\log(d_1+1)}{M}}\|\boldsymbol{\Delta}\|_1\right\} \\ &\ge \frac{\kappa_1}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{\kappa_1\kappa_2^2\log(d_1+1)}{2M}\|\boldsymbol{\Delta}\|_1^2 \quad \text{for all } \|\boldsymbol{\Delta}\|_2 \le 1,\end{aligned} \tag{3.122}$$

with probability at least $1 - c_1\exp(-c_2M)$, where $c_1, c_2, \kappa_1, \kappa_2 > 0$ are some constants.

Additionally, the gradient $\|\nabla\ell_M(\boldsymbol{\gamma}^*)\|_\infty$ is controlled in the following. We allow a possibly misspecified model that $\pi^*(\cdot) \ne \pi(\cdot)$. Note that, even under model misspecification,

we still have (3.124) below. Hence, $\|\nabla \ell_M(\boldsymbol{\gamma}^*)\|_\infty$ is the maximum of zero-mean random variables.

By the union bound, we have

$$P\left(\|\nabla \ell_M(\boldsymbol{\gamma}^*)\|_\infty \geq \frac{\lambda_\gamma}{2}\right) = P\left(\max_{1\leq j\leq d_1+1}\left|M^{-1}\sum_{i\in I_{-k}}(f(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - A_{1i})\mathbf{V}_{i,j}\right| \geq \frac{\lambda_\gamma}{2}\right)$$
$$\leq \sum_{j=1}^{d_1+1} P\left(\left|M^{-1}\sum_{i\in I_{-k}}(f(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - A_{1i})\mathbf{V}_{i,j}\right| \geq \frac{\lambda_\gamma}{2}\right), \tag{3.123}$$

where $f(u) = \frac{\exp(u)}{1+\exp(u)}$ is the logistic function. By definition, $\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_1+1}} E[\ell(\boldsymbol{\gamma})]$, where for any $\boldsymbol{\gamma} \in \mathbb{R}^{d_1+1}$,

$$\ell(\boldsymbol{\gamma}) := E\left[-A_1\mathbf{V}^T\boldsymbol{\gamma} + \log(1+\exp(\mathbf{V}^T\boldsymbol{\gamma}))\right].$$

By the first-order optimality condition, we know that

$$\nabla E[\ell(\gamma^*)] = E\left[(f(\mathbf{V}^T\boldsymbol{\gamma}^*) - A_1)\mathbf{V}\right] = \mathbf{0} \in \mathbb{R}^{d_1+1}. \tag{3.124}$$

Additionally, since $|f(\mathbf{V}^T\boldsymbol{\gamma}^*) - A_1| \leq 1$, by Lemma D.1 (ii) of [CLCL19] and under Assumption 3.2, for any $i \in I_{-k}$ and $j \leq d_1 + 1$,

$$\|(f(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - A_{1i})\mathbf{V}_{i,j}\|_{\psi_2} \leq \|\mathbf{V}_{i,j}\|_{\psi_2} \leq \sigma_u.$$

That is, $(f(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - A_{1i})\mathbf{V}_{i,j}$ is a zero-mean sub-Gaussian random variable. Hence, by Lemma D.2 of [CLCL19], for each $j \leq d_1 + 1$,

$$P\left(\left|M^{-1}\sum_{i\in I_{-k}}(f(\mathbf{V}_i^T\boldsymbol{\gamma}^*) - A_{1i})\mathbf{V}_{i,j}\right| \geq \frac{\lambda_\gamma}{2}\right) \leq 2\exp\left(\frac{-M\lambda_\gamma^2}{32\sigma_u^2}\right)$$
$$\leq 2\exp\left(\frac{-M\lambda_\gamma^2}{32\sigma_u^2}\right) \leq 2\exp\left(-\log(d_1+1) - Mt^2\right) = \frac{2\exp(-Mt^2)}{d_1+1},$$

where for any $t > 0$, we set $\lambda_{\boldsymbol{\gamma}} := 4\sqrt{2}\sigma_u(\sqrt{\frac{\log(d_1+1)}{M}} + t)$. Together with (3.123), it follows that

$$P\left(\|\ell_M(\boldsymbol{\gamma}^*)\|_\infty \leq \frac{\lambda_{\boldsymbol{\gamma}}}{2}\right) \leq 1 - 2\exp(-Mt^2).$$

Together with (3.122), when $M \geq 64\kappa_2^2 s_{\boldsymbol{\gamma}} \log(d_1 + 1)$ and $9s_{\boldsymbol{\gamma}}\lambda_{\boldsymbol{\gamma}}^2 \leq \kappa_1^2$, by Corollary 9.20 of [Wai19], we conclude that

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq \frac{3\sqrt{s_{\boldsymbol{\gamma}}}\lambda_{\boldsymbol{\gamma}}}{\kappa_1}, \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 \leq \frac{6s_{\boldsymbol{\gamma}}\lambda_{\boldsymbol{\gamma}}}{\kappa_1},$$

with probability at least $1 - 2\exp(-Mt^2) - c_1\exp(-c_2 M)$. Hence, when $M \gg s_{\boldsymbol{\gamma}}\log(d_1)$, with some $\lambda_M \asymp \sqrt{\frac{\log(d_1)}{M}}$,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log(d_1)}{N}\right). \tag{3.125}$$

Now, we show the estimation rate for $\widehat{\pi}(\cdot)$. In the following, we will use Taylor's Theorem to control the estimation error of $\widehat{\pi}(\cdot)$ by the estimation error of $\widehat{\boldsymbol{\gamma}}$ as in (3.127). Then, we apply the estimation rate (3.125) proved above to obtain the rate for $\widehat{\pi}(\cdot)$.

Let $f(u) := \frac{\exp(u)}{1+\exp(u)} = \Psi'(u)$ for any $u \in \mathbb{R}$. Note that, for any $u^*, \Delta \in \mathbb{R}$,

$$\begin{aligned}
\frac{d(f(u^* + t\Delta) - f(u^*))^2}{dt} &= 2(f(u^* + t\Delta) - f(u^*))f'(u^* + t\Delta)\Delta, \\
\frac{d^2(f(u^* + t\Delta) - f(u^*))^2}{dt^2} &= 2(f'(u^* + t\Delta))^2\Delta^2 + 2(f(u^* + t\Delta) - f(u^*))f''(u^* + t\Delta)\Delta^2,
\end{aligned}$$

where, for any $u \in \mathbb{R}$, since $f(u) \in (0, 1)$, we have

$$f'(u) = f(u)(1 - f(u)) \in (0, 1), \quad f''(u) = f(u)(1 - f(u))(1 - 2f(u)) \in (-1, 1). \tag{3.126}$$

Set $u^* = \mathbf{V}^T\boldsymbol{\gamma}^*$ and $\Delta = \mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$. By Taylor's Theorem, with some $\widetilde{t} \in (0,1)$,

$$\begin{aligned}
E[f(\mathbf{V}^T\widehat{\boldsymbol{\gamma}}) - f(\mathbf{V}^T\boldsymbol{\gamma}^*)]^2 &= E[f(u^* + 1\cdot\Delta) - f(u^*)]^2 \\
&= E[f(u^* + 0\cdot\Delta) - f(u^*)]^2 + \frac{dE(f(u^*+t\Delta) - f(u^*))^2}{dt}\bigg|_{t=0}\cdot 1 \\
&\quad + \frac{d^2E(f(u^*+t\Delta) - f(u^*))^2}{2dt^2}\bigg|_{t=\widetilde{t}}\cdot 1^2 \\
&= 0 + E\left[2(f(u^*+0\cdot\Delta) - f(u^*))f'(u^*+0\cdot\Delta)\Delta\right] \\
&\quad + E\left[(f'(u^*+\widetilde{t}\Delta))^2\Delta^2 + (f(u^*+\widetilde{t}\Delta) - f(u^*))f''(u^*+\widetilde{t}\Delta)\Delta^2\right] \\
&= E\left[(f'(u^*+\widetilde{t}\Delta))^2\Delta^2 + (f(u^*+\widetilde{t}\Delta) - f(u^*))f''(u^*+\widetilde{t}\Delta)\Delta^2\right] \\
&\overset{(i)}{\leq} 2E[\Delta^2] = 2E[\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)]^2,
\end{aligned}$$

where (i) holds since, by (3.126), $(f'(u^*+\widetilde{t}\Delta))^2 \leq 1$ and $(f(u^*+\widetilde{t}\Delta) - f(u^*))f''(u^*+\widetilde{t}\Delta) \leq 1$. Hence,

$$E[\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1)]^2 = E[f(\mathbf{V}^T\widehat{\boldsymbol{\gamma}}) - f(\mathbf{V}^T\boldsymbol{\gamma}^*)]^2 \leq 2E[\mathbf{V}^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)]^2. \tag{3.127}$$

Then, from (3.120) and (3.125), we have

$$E[\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1)]^2 \leq 2\|E[\mathbf{V}\mathbf{V}^T]\|_2\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log(d_1)}{N}\right). \tag{3.128}$$

b) Now, we show (3.46) and (3.47). In part b), the expectations are only taken w.r.t. the distribution of the new observations $\mathbf{S}_1, \mathbf{S}_2$.

By Lemma 3.4, we know that the minimum and maximum eigenvalues of covariance matrix $E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]$ satisfy

$$\lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \geq \kappa_l > 0, \;\; \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]) \leq 2\sigma_u^2 < \infty,$$

and $\bar{\mathbf{U}}$ is sub-Gaussian with $\|\boldsymbol{x}^T\bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$. Additionally, we also have $P(c_0^2 \leq \Psi''(\bar{\mathbf{U}}^T\boldsymbol{\delta}_a) \leq (1-c_0)^2) = 1$. Repeating the same procedure as in part a), we

also have

$$\|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}\right),$$

and

$$\begin{aligned} E[\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S})]^2 = E[f(\bar{\mathbf{U}}^T\widehat{\boldsymbol{\delta}}_a) - f(\bar{\mathbf{U}}^T\boldsymbol{\delta}_a^*)]^2 &\leq 2E[\bar{\mathbf{U}}^T(\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*)]^2 \\ &\leq 2\|E[\bar{\mathbf{U}}\bar{\mathbf{U}}^T]\|_2\|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}\right). \end{aligned}$$

■

Proof of Lemma 3.6. In this proof, the expectations are taken w.r.t. the distribution of a new observation W. We only focus on the treatment paths $a = (1,1)$ and $a' = (0,0)$. Hence, when possible, we abbreviate the subscripts $(1,1)$ and $(0,0)$ by 1 and 0. For instance, $\rho_1(\cdot) = \rho_{1,1}(\cdot)$, $\rho_1^*(\cdot) = \rho_{1,1}^*(\cdot)$ and $\widehat{\rho}_1(\cdot) = \widehat{\rho}_{1,1}(\cdot)$.

We begin by decomposing T_2, (3.78), as a sum of six terms

$$\psi(W;\widehat{\eta}) - \psi(W;\eta^*) = \sum_{i=1}^{6} Q_i, \tag{3.129}$$

where

$$Q_1 := \frac{A_1 A_2}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}(Y - \widehat{\nu}_1(\mathbf{S})) - \frac{A_1 A_2}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}(Y - \nu_1^*(\mathbf{S})), \tag{3.130}$$

$$Q_2 := \frac{A_1}{\widehat{\pi}(\mathbf{S}_1)}(\widehat{\nu}_1(\mathbf{S}) - \widehat{\mu}_1(\mathbf{S}_1)) - \frac{A_1}{\pi^*(\mathbf{S}_1)}(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1)), \tag{3.131}$$

$$Q_3 := \widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1), \tag{3.132}$$

$$\begin{aligned} Q_4 := &-\frac{(1-A_1)(1-A_2)}{(1-\widehat{\pi}(\mathbf{S}_1))(1-\widehat{\rho}_0(\mathbf{S}))}(Y - \widehat{\nu}_0(\mathbf{S})) \\ &+ \frac{(1-A_1)(1-A_2)}{(1-\pi^*(\mathbf{S}_1))(1-\rho_0^*(\mathbf{S}))}(Y - \nu_0^*(\mathbf{S})), \end{aligned} \tag{3.133}$$

$$Q_5 := -\frac{1-A_1}{1-\widehat{\pi}(\mathbf{S}_1)}(\widehat{\nu}_0(\mathbf{S}) - \widehat{\mu}_0(\mathbf{S}_1)) + \frac{1-A_1}{1-\pi^*(\mathbf{S}_1)}(\nu_0^*(\mathbf{S}) - \mu_0^*(\mathbf{S}_1)), \tag{3.134}$$

$$Q_6 := -\widehat{\mu}_0(\mathbf{S}_1) + \mu_0^*(\mathbf{S}_1). \tag{3.135}$$

Hence, we have the following representation for T_2:

$$T_2 = E[\psi(W; \widehat{\eta}) - \psi(W; \eta^*)] = \sum_{i=1}^{6} E[Q_i], \tag{3.136}$$

where the expecatations are only taken w.r.t. the distribution of the new obseravtion W.

a) Recall the representation (3.136). Here, we first obtain an upper bound for $E[Q_1 + Q_2 + Q_3]$. By the law of iterated expectations,

$$E[Q_1] = E\left[\frac{A_1\rho_1(\mathbf{S})}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}(\nu_1(\mathbf{S}) - \widehat{\nu}_1(\mathbf{S})) - \frac{A_1\rho_1(\mathbf{S})}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}(\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))\right].$$

Through rearranging, we have the following representation:

$$E[Q_1 + Q_2 + Q_3] = \sum_{i=1}^{8} R_i, \tag{3.137}$$

where

$$R_1 := E\left[\frac{A_1\rho_1^*(\mathbf{S})(\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S}))}{\widehat{\pi}(\mathbf{S}_1)}\left(\frac{1}{\rho_1^*(\mathbf{S})}-\frac{1}{\widehat{\rho}_1(\mathbf{S})}\right)\right], \tag{3.138}$$

$$R_2 := E\left[\pi^*(\mathbf{S}_1)(\widehat{\mu}_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))\left(\frac{1}{\pi^*(\mathbf{S}_1)}-\frac{1}{\widehat{\pi}(\mathbf{S}_1)}\right)\right], \tag{3.139}$$

$$R_3 := E\left[\frac{A_1(\rho_1^*(\mathbf{S})-\rho_1(\mathbf{S}))(\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S}))}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}\right], \tag{3.140}$$

$$\begin{aligned} R_4 :=& E\left[\frac{(\pi^*(\mathbf{S}_1)-A_1)(\widehat{\mu}_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))}{\widehat{\pi}(\mathbf{S}_1)}\right] \\ \overset{(i)}{=}& E\left[\frac{(\pi^*(\mathbf{S}_1)-\pi(\mathbf{S}_1))(\widehat{\mu}_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))}{\widehat{\pi}(\mathbf{S}_1)}\right], \end{aligned} \tag{3.141}$$

$$R_5 := E\left[\frac{A_1\rho_1^*(\mathbf{S})(\nu_1^*(\mathbf{S})-\nu_1(\mathbf{S}))}{\widehat{\pi}(\mathbf{S}_1)}\left(\frac{1}{\rho_1^*(\mathbf{S})}-\frac{1}{\widehat{\rho}_1(\mathbf{S})}\right)\right], \tag{3.142}$$

$$R_6 := E\left[A_1(\mu_1^*(\mathbf{S}_1)-\mu_1(\mathbf{S}_1))\left(\frac{1}{\pi^*(\mathbf{S}_1)}-\frac{1}{\widehat{\pi}(\mathbf{S}_1)}\right)\right], \tag{3.143}$$

$$R_7 := E\left[\left(\frac{A_1}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}-\frac{A_1}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}\right)(\rho_1^*(\mathbf{S})-\rho_1(\mathbf{S}))(\nu_1^*(\mathbf{S})-\nu_1(\mathbf{S}))\right] \overset{(ii)}{=} 0, \tag{3.144}$$

$$R_8 := E\left[\frac{A_1(\widehat{\pi}(\mathbf{S}_1)-\pi^*(\mathbf{S}_1))(\mu_1(\mathbf{S}_1)-\nu_1(\mathbf{S}))}{\widehat{\pi}(\mathbf{S}_1)\pi^*(\mathbf{S}_1)}\right] \overset{(iii)}{=} 0. \tag{3.145}$$

Here, (i) holds by the law of iterated expectations; (ii) holds since either $\rho_1^*(\cdot) = \rho_1(\cdot)$ or $\mu_1^*(\cdot) = \mu_1(\cdot)$ by assumption; (iii) holds by the law of iterated expectations and the fact that, under Assumption 3.1,

$$\begin{aligned} E[\nu_1(\mathbf{S})|\mathbf{S}_1, A_1=1] &= E[E[Y|\mathbf{S}_1,\mathbf{S}_2,A_1=1,A_2=1]|\mathbf{S}_1,A_1=1] \\ &= E[E[Y(1,1)|\mathbf{S}_1,\mathbf{S}_2,A_1=1,A_2=1]|\mathbf{S}_1,A_1=1] \\ &= E[E[Y(1,1)|\mathbf{S}_1,\mathbf{S}_2,A_1=1]|\mathbf{S}_1,A_1=1] \\ &= E[Y(1,1)|\mathbf{S}_1,A_1=1] = \mu_1(\mathbf{S}_1). \end{aligned} \tag{3.146}$$

Now, we obtain an upper bound for R_i ($i \in \{1,\ldots,6\}$). For $R_1 + R_2$, since $|A_1| \leq 1$,

$|\pi^*(\mathbf{S}_1)| \leq 1$ and $|\rho_1^*(\mathbf{S})| \leq 1$, we have

$$\begin{aligned} R_1 + R_2 &\overset{(i)}{\leq} \{E|\widehat{\pi}(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \left\{ E \left| \frac{1}{\widehat{\rho}_1(\mathbf{S})} - \frac{1}{\rho_1^*(\mathbf{S})} \right|^3 \right\}^{\frac{1}{3}} \{E|\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S})|^3\}^{\frac{1}{3}} \\ &\quad + \left\{ E \left| \frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} \{E|\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \\ &\overset{(ii)}{=} O_p\left(\sigma \frac{s_1 \log(d)}{N} \right), \end{aligned}$$

where (i) holds by Hölder's inequality; (ii) follows from Lemma 3.2. Similarly, for $R_3 + R_4$, since $|A_1| \leq 1$, $|\rho_1^*(\mathbf{S}) - \rho_1(\mathbf{S})| \leq 1$, $|\pi^*(\mathbf{S}_1) - \pi(\mathbf{S}_1)| \leq 1$, and together with Lemma 3.2,

$$\begin{aligned} R_3 + R_4 \leq & \{E|\widehat{\pi}(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\rho}_1(\mathbf{S})|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S})|^3\}^{\frac{1}{3}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \\ & + \{E|\widehat{\pi}(\mathbf{S}_1)|^{-2}\}^{\frac{1}{2}} \{E|\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}} \\ = & O_p\left(\sigma \sqrt{\frac{(s_{\boldsymbol{\alpha}_a} + s_{\boldsymbol{\beta}_a}) \log(d)}{N}} \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}} + \sigma \sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log(d)}{N}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \right). \end{aligned}$$

For $R_5 + R_6$, since $|\rho_1^*(\mathbf{S})| \leq 1$,

$$\begin{aligned} R_5 + R_6 &\leq \{E|\widehat{\pi}(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}} \left\{ E \left| \frac{1}{\widehat{\rho}_1(\mathbf{S})} - \frac{1}{\rho_1^*(\mathbf{S})} \right|^4 \right\}^{\frac{1}{4}} \{E[A_1|\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S})|^2]\}^{\frac{1}{2}} \\ &\quad + \left\{ E \left| \frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} \{E[A_1|\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)|^2]\}^{\frac{1}{2}} \\ &= O_p\left(\sigma \sqrt{\frac{s_{\boldsymbol{\gamma}} \log(d)}{N}} \mathbb{1}_{\{\mu_a^*(\cdot) \neq \mu_a(\cdot)\}} + \sigma \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \right). \end{aligned}$$

where the last assertion follows from Lemma 3.2, (3.156), (3.158), and Lemma 3.12.

Combining all the previous results, we have

$$E[Q_1 + Q_2 + Q_3] = \sum_{i=1}^{6} R_i = O_p\left(\sigma \frac{s_1 \log(d)}{N} + \sigma \sqrt{\frac{s_2 \log(d)}{N}} \right).$$

Analogously to $E[Q_1 + Q_2 + Q_3]$, we have the same result for $E[Q_4 + Q_5 + Q_6]$. Theorefore, (3.81) follows.

b) When all the models are correctly specified, we have $s_2 = 0$. Hence, by part a), (3.82) holds. ∎

Proof of Lemma 3.7. In this proof, the expecations are taken w.r.t. a new observation W, unless stated otherwise.

We first show that (3.83) holds. Recall the representation (3.129), by Minkowski inequality, we have

$$[E(\psi(W;\widehat{\eta}) - \psi(W;\eta^*))^2]^{\frac{1}{2}} \leq \sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}}, \tag{3.147}$$

where Q_i ($i \in \{1,\ldots,6\}$) are defined as(3.130)-(3.135). In the following, we show that

$$\sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right).$$

By Minkowski's inequality,

$$\begin{aligned}
[E(Q_1^2)]^{\frac{1}{2}} &\leq \left\{E\left[\frac{A_1A_2}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))\right]^2\right\}^{\frac{1}{2}} \\
&\quad + \left\{E\left[\left(\frac{A_1A_2}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})} - \frac{A_1A_2}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}\right)(Y - \nu_1^*(\mathbf{S}))\right]^2\right\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \left\{E\left[\frac{1}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})}(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))\right]^2\right\}^{\frac{1}{2}} \\
&\quad + \left\{E\left[\left(\frac{1}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})} - \frac{1}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}\right)\zeta\right]^2\right\}^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} \{E|\widehat{\pi}(\mathbf{S}_1)|^{-6}\}^{\frac{1}{6}}\{E|\widehat{\rho}_1(\mathbf{S})|^{-6}\}^{\frac{1}{6}}\{E|\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S})|^6\}^{\frac{1}{6}} \\
&\quad + \{E|\zeta|^4\}^{\frac{1}{4}}\left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\pi^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right|^4\right\}^{\frac{1}{4}} \\
&\overset{(iii)}{=} O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right),
\end{aligned}$$

where (i) holds by the fact that $|A_1| \leq 1$, $|A_2| \leq 1$ and $A_1A_2\zeta = \zeta_1 = A_1A_2(Y - \nu_1^*(\mathbf{S}))$; (ii) holds byHölder's inequality; (iii) follows from Lemma 3.2, and under Assumption 3.3, by

Lemma D.1 (iv) of [CLCL19],

$$E[|\zeta|^4] \leq 8\sigma^4\sigma_\zeta^4, \quad E[|\varepsilon|^4] \leq 8\sigma^4\sigma_\varepsilon^4. \tag{3.148}$$

Then, similarly as above, we obtain

$$\begin{aligned}
[E(Q_2^2)]^{\frac{1}{2}} &\leq \left\{E\left[\frac{A_1}{\widehat{\pi}(\mathbf{S}_1)}(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))\right]^2\right\}^{\frac{1}{2}} + \left\{E\left[\frac{A_1}{\widehat{\pi}(\mathbf{S}_1)}(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))\right]^2\right\}^{\frac{1}{2}} \\
&\quad + \left\{E\left[\left(\frac{A_1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{A_1}{\pi^*(\mathbf{S}_1)}\right)(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))\right]^2\right\}^{\frac{1}{2}} \\
&\leq \left\{E\left[\frac{1}{\widehat{\pi}(\mathbf{S}_1)}(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))\right]^2\right\}^{\frac{1}{2}} + \left\{E\left[\frac{1}{\widehat{\pi}(\mathbf{S}_1)}(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))\right]^2\right\}^{\frac{1}{2}} \\
&\quad + \left\{E\left[\left(\frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)}\right)\varepsilon\right]^2\right\}^{\frac{1}{2}} \\
&\leq \{E|\widehat{\pi}(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}}\{E|\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S})|^4\}^{\frac{1}{4}} + \left\{E\left|\frac{1}{\widehat{\pi}(\mathbf{S}_1)} - \frac{1}{\pi^*(\mathbf{S}_1)}\right|^4\right\}^{\frac{1}{4}}\{E|\varepsilon|^4\}^{\frac{1}{4}} \\
&\quad + \{E|\widehat{\pi}(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}}\{E|\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1)|^4\}^{\frac{1}{4}} \\
&= O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}\}\log(d)}{N}}\right),
\end{aligned}$$

where the last assertion follows from the Lemma 3.2 and (3.148). Also, by Lemma 3.2,

$$[E(Q_3^2)]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}\}\log(d)}{N}}\right).$$

Hence, we have

$$[E(Q_1^2)]^{\frac{1}{2}} + [E(Q_2^2)]^{\frac{1}{2}} + [E(Q_3^2)]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right).$$

Repeating the procedure above, we obtain the same result for $[E(Q_4^2)]^{\frac{1}{2}} + [E(Q_5^2)]^{\frac{1}{2}} + [E(Q_6^2)]^{\frac{1}{2}}$. Therefore, (3.83) holds.

Now, we show (3.84). Recall the definition (3.80), by Chebyshev's inequality, we have

for any $t > 0$,

$$P(|T_4| > t) \leq \frac{1}{t^2}\mathrm{Var}\left[\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \psi(W_i;\eta^*))\right] \tag{3.149}$$
$$\leq \frac{1}{nt^2}E[\psi(W;\widehat{\eta}) - \psi(W;\eta^*)]^2.$$

In the righ-hand side of (3.149), the variance is taken over the joint distribution of $(W_i)_{i\in\mathcal{I}_k}$. Note that, based on the sample-splitting, $\widehat{\eta}$ is independent of $(W_i)_{i\in\mathcal{I}_k}$. Together with (3.83), we conclude that (3.84) holds. ∎

Proof of Lemma 3.8. Recall the definition (3.77). Since $\theta = E[\mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1)]$, we have

$$T_1 = E[\psi_a(W;\eta_a^*) - \mu_a(\mathbf{S}_1)] - E[\psi_{a'}(W;\eta_{a'}^*) - \mu_{a'}(\mathbf{S}_1)].$$

It suffices to show $E[\psi_c(W;\eta_c^*) - \mu_c(\mathbf{S}_1)] = 0$ for each $c \in \{a, a'\}$. Without loss of generality, we consider $c = a = (1,1)$. Observe that,

$$\begin{aligned}
&E[\psi_a(W;\eta_a^*) - \mu_a(\mathbf{S}_1)] \\
&\quad= E\left[\frac{A_1A_2(Y-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})} + \frac{A_1(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)} + \mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)\right] \\
&\quad\overset{(i)}{=} E\left[\frac{A_1\rho_1(\mathbf{S})(\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})} + \frac{A_1(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)} + \mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)\right] \\
&\quad\overset{(ii)}{=} T_{1,1} + T_{1,2} + T_{1,3},
\end{aligned}$$

where

$$\begin{aligned}
T_{1,1} :=& E\left[\frac{A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))}{\pi^*(\mathbf{S}_1)}\left(1 - \frac{\rho_1(\mathbf{S})}{\rho_1^*(\mathbf{S})}\right)\right], \\
T_{1,2} :=& E\left[(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))\left(1 - \frac{A_1}{\pi^*(\mathbf{S}_1)}\right)\right], \\
T_{1,3} :=& E\left[\frac{A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)}\right].
\end{aligned}$$

In the above, (i) holds by the law of iterated expectations and under Assumption 3.1 since

$$\begin{aligned} E\left[\frac{A_1A_2(Y-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}\right] &= E\left[E\left[\frac{A_1A_2(Y(1,1)-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}|\mathbf{S},A_1=1\right]P(A_1=1|\mathbf{S})\right] \\ &= E\left[\frac{E[A_2|\mathbf{S},A_1=1](E[Y(1,1)|\mathbf{S},A_1=1]-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}E[A_1|\mathbf{S}]\right] \\ &= E\left[\frac{\rho_1(\mathbf{S})(\nu_1(\mathbf{S})-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}E[A_1|\mathbf{S}]\right] = E\left[\frac{A_1\rho_1(\mathbf{S})(\nu_1(\mathbf{S})-\nu_1^*(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}\right]. \end{aligned}$$

Additionally, (ii) holds by rearranging the terms after the following decomposition

$$(\nu_1^*(\mathbf{S})-\mu_1^*(\mathbf{S}_1)) = (\nu_1^*(\mathbf{S})-\nu_1(\mathbf{S})) + (\nu_1(\mathbf{S})-\mu_1(\mathbf{S}_1)) + (\mu_1(\mathbf{S})-\mu_1^*(\mathbf{S}_1)).$$

By assumption, either $\nu_1^*(\cdot)=\nu_1(\cdot)$ or $\rho_1^*(\cdot)=\rho_1(\cdot)$. Hence, $T_{1,1}=0$. By the law of iterated expectations, under Assumption 3.1,

$$T_{1,2} = E\left[(\mu_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))\left(1-\frac{\pi(\mathbf{S}_1)}{\pi^*(\mathbf{S}_1)}\right)\right] = 0,$$

since, by assumption, either $\mu_1^*(\cdot)=\mu_1(\cdot)$ or $\pi^*(\cdot)=\pi(\cdot)$. Besides, as in (3.146), we have $E[\nu_1(\mathbf{S})|\mathbf{S}_1,A_1=1]=\mu_1(\mathbf{S}_1)$. Hence, by the law of iterated expectations,

$$\begin{aligned} T_{1,3} &= E\left[E\left[\frac{A_1(\nu_1(\mathbf{S})-\mu_1(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)}|\mathbf{S}_1,A_1=1\right]P(A_1=1|\mathbf{S}_1)\right] \\ &= E\left[\frac{\pi(\mathbf{S}_1)}{\pi^*(\mathbf{S}_1)}\left[E\left[\nu_1(\mathbf{S})|\mathbf{S}_1,A_1=1\right]-\mu_1(\mathbf{S}_1)\right]\right] = 0. \end{aligned}$$

Combining the previous results, we have

$$E[\psi_a(W;\eta_a^*)-\mu_a(\mathbf{S}_1)] = T_{1,1}+T_{1,2}+T_{1,3} = 0.$$

Repeating the same procedure, we also have $E[\psi_{a'}(W;\eta_{a'}^*)-\mu_{a'}(\mathbf{S}_1)]=0$, and hence (3.94) follows. ∎

Proof of Lemma 3.9. In this proof, the expectations are taken w.r.t. the distribution of new observations $\mathbf{S}_1, \mathbf{S}_2$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). We condition on the following event

$$\mathcal{E}_4 := \left\{ P(c_0 \leq \widehat{\pi}(\mathbf{S}_1) \leq 1 - c_0) = 1, \;\; P(c_0 \leq \widehat{\rho}_1(\mathbf{S}) \leq 1 - c_0) = 1 \right\}. \tag{3.150}$$

Under Assumption 3.7, the event $\mathcal{E}_4$ occurs with probability approaching one.

Recall the representation (3.136). Here, we first upper bound $E[Q_1 + Q_2 + Q_3]$. Same as in the proof of Lemma 3.6, we also have (3.137) holds, with R_is defined in (3.138)-(3.145). Same as in (3.144) and (3.145), we have $R_7 = R_8 = 0$. In the following, we use Cauchy-Schwarz inequality to upper bound R_i ($i \in \{1, \ldots, 6\}$). For $R_1 + R_2$, on the event $\mathcal{E}_4$, we have

$$\begin{aligned} R_1 + R_2 \leq & \frac{1}{c_0^2} [E(\widehat{\rho}_1(\mathbf{S}) - \rho_1^*(\mathbf{S}))^2]^{\frac{1}{2}} [E(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} \\ & + \frac{1}{c_0} [E(\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1))^2]^{\frac{1}{2}} [E(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\ = & \, O_p\left(b_N c_N + a_N d_N\right), \end{aligned} \tag{3.151}$$

under Assumption 3.6. For $R_3 + R_4$, on the event $\mathcal{E}_4$, we have

$$\begin{aligned} R_3 + R_4 \leq & \frac{1}{c_0^2} [E(\rho_1^*(\mathbf{S}) - \rho_1(\mathbf{S}))^2]^{\frac{1}{2}} [E(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} \\ & + \frac{1}{c_0} [E(\pi^*(\mathbf{S}_1) - \pi(\mathbf{S}_1))^2]^{\frac{1}{2}} [E(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\ \leq & \frac{\mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}}}{c_0^2} [E(\widehat{\nu}_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{\mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}}}{c_0} [E(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2]^{\frac{1}{2}}, \end{aligned}$$

since

$$\begin{aligned} E(\rho_1^*(\mathbf{S}) - \rho_1(\mathbf{S}))^2 &= \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} E(\rho_1^*(\mathbf{S}) - \rho_1(\mathbf{S}))^2 \overset{(i)}{\leq} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}}, \\ E(\pi^*(\mathbf{S}_1) - \pi(\mathbf{S}_1))^2 &= \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}} E(\pi^*(\mathbf{S}_1) - \pi(\mathbf{S}_1))^2 \overset{(ii)}{\leq} \mathbb{1}_{\{\pi^*(\cdot) \neq \pi(\cdot)\}}, \end{aligned}$$

where (i) and (ii) hold because $\rho_1(\mathbf{S}) = E(A_2|\mathbf{S}, A_1 = 1) \in (0,1)$, $\pi(\mathbf{S}_1) = E(A_1|\mathbf{S}_1) \in (0,1)$, and, under Assumption 3.4, $\rho_1(\mathbf{S}), \pi^*(\mathbf{S}_1) \in (0,1)$ with probability one. Hence, under Assumption 3.6, we have

$$R_3 + R_4 = O_p\left(b_N \mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}} + a_N \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\right). \tag{3.152}$$

As for $R_5 + R_6$, similarly, we have

$$\begin{aligned} R_5 + R_6 \leq \frac{1}{c_0^2}[E(\widehat{\rho}_1(\mathbf{S}) - \rho_1^*(\mathbf{S}))^2]^{\frac{1}{2}} \left[E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2]\right]^{\frac{1}{2}} \\ + \frac{1}{c_0^2}[E(\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \left[E[A_1(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2]\right]^{\frac{1}{2}} \end{aligned} \tag{3.153}$$

Here, we need upper bound for $[E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2]]^{\frac{1}{2}}$ and $[E[A_1(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2]]^{\frac{1}{2}}$. By definition,

$$\zeta = \zeta_1 + \zeta_0, \quad \varepsilon = \varepsilon_1 + \varepsilon_0, \quad Y = Y(1,1)A_1A_2 + Y(0,0)(1-A_1)(1-A_2),$$

where

$$\zeta_1 = A_1A_2\left(Y(1,1) - \nu_1^*(\mathbf{S})\right), \quad \varepsilon_1 = A_1\left(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1)\right).$$

Hence, we have

$$E[\zeta^2] \geq E[A_1A_2\zeta^2] = E[\zeta_1^2] = E[A_1A_2(Y - \nu_1^*(\mathbf{S}))^2] \tag{3.154}$$

Note that

$$\begin{aligned} &E[A_1A_2(Y - \nu_1(\mathbf{S}))(\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))] \\ &\overset{(i)}{=} E[E[A_1A_2(Y(1,1) - \nu_1(\mathbf{S}))(\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))|\mathbf{S}, A_1 = 1]P(A_1 = 1|\mathbf{S})] \\ &\overset{(ii)}{=} E[E[A_2|\mathbf{S}, A_1 = 1](E[Y(1,1)|\mathbf{S}, A_1 = 1] - \nu_1(\mathbf{S}))(\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))P(A_1 = 1|\mathbf{S})] \\ &\overset{(iii)}{=} 0, \end{aligned}$$

where (i) holds by the law of iterated expectations and the fact that $A_1A_2Y = A_1A_2Y(1,1)$; (ii) holds under Assumption 3.1; (iii) holds since $\nu_1(\mathbf{S}) = E[Y(1,1)|\mathbf{S}, A_1 = 1, A_2 = 1] = E[Y(1,1)|\mathbf{S}, A_1 = 1]$ under Assumption 3.1. Therefore,

$$\begin{aligned} E[A_1A_2(Y - \nu_1^*(\mathbf{S}))^2] &= E[A_1A_2[(Y - \nu_1(\mathbf{S}))^2 + (\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))^2]] \\ &\geq E[A_1A_2(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2] = E[A_1\rho_1(\mathbf{S})(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2] \\ &\geq c_0 E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2], \end{aligned} \tag{3.155}$$

under Assumption 3.1. Together with (3.154), we have

$$E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2] \leq \frac{1}{c_0} E[\zeta^2]. \tag{3.156}$$

Besides, note that

$$\begin{aligned} &E[A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))(\mu_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))] \\ &\quad = E[(\mu_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))E[(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))|\mathbf{S}_1, A_1 = 1]P(A_1 = 1|\mathbf{S})] = 0, \end{aligned}$$

since $E[\nu_1(\mathbf{S})|\mathbf{S}_1, A_1 = 1] = \mu_1(\mathbf{S}_1)$ as shown in (3.146). Therefore, we have

$$\begin{aligned} E[A_1(\nu_1(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))^2] &= E[A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))^2] + E[A_1(\mu_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2] \\ &\geq E[A_1(\mu_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2]. \end{aligned} \tag{3.157}$$

Additionally, observe that

$$\begin{aligned} E[A_1(\nu_1(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))^2] &\leq 2E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2] + 2E[\varepsilon_1^2] \\ &\overset{(i)}{\leq} \frac{2}{c_0} E[\zeta^2] + 2E[A_1\varepsilon^2] \leq \frac{2}{c_0} E[\zeta^2] + 2E[\varepsilon^2], \end{aligned}$$

where (i) holds by (3.156) and the fact that $\varepsilon_1^2 = A_1\varepsilon^2$. Together with (3.157), we obtain

$$E[A_1(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2] \leq \frac{2}{c_0} E[\zeta^2] + 2E[\varepsilon^2]. \tag{3.158}$$

Therefore, under Assumption 3.6,

$$R_5 + R_6 \leq \frac{1}{c_0^2}[E(\widehat{\rho}_1(\mathbf{S}) - \rho_1^*(\mathbf{S}))^2]^{\frac{1}{2}}[E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2]]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\pi}(\mathbf{S}_1) - \pi^*(\mathbf{S}_1))^2]^{\frac{1}{2}}[E[A_1(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2]]^{\frac{1}{2}}$$
$$= O_p\left(c_N\sqrt{E[\zeta^2+\varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}} + d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right). \tag{3.159}$$

Pluging (3.144), (3.145), (3.151),(3.152), and (3.159) into (3.137), we obtain

$$E[Q_1+Q_2+Q_3] = O_p\bigg(b_N c_N + a_N d_N + b_N\mathbb{1}_{\{\pi^*(\cdot)\neq\pi(\cdot)\}} + a_N\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}} + c_N\sqrt{E[\zeta^2+\varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}} + d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\bigg).$$

By repeating all the previous steps, we can obtain the same result for $E[Q_4+Q_5+Q_6]$. Therefore, (3.95) follows.

b) When all the nuisance models are correct, Assumption 3.4 holds under Assumption 3.1. Hence, by part a), we also have (3.95). Since all the nuisance models are correct, we further conclude that (3.96) holds. ∎

Proof of Lemma 3.10. a) Recall the definition (3.79). By Chebyshev's inequality, we have for any $t > 0$,

$$P(|T_3| > t) \leq \frac{1}{t^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i\in I_k}\psi(W_i;\eta^*)\right) = \frac{1}{nt^2}E[\psi(W;\eta^*)]^2,$$

where $n = N/K = |I_k|$. To prove (3.97), we only need to show $[E(\psi(W;\eta^*))^2]^{\frac{1}{2}} = O(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]})$. By Minkowski inequality, we have

$$[E(\psi(W;\eta^*))^2]^{\frac{1}{2}} \leq \sum_{i=1}^{5} T_{3,i}, \tag{3.160}$$

where

$$T_{3,1} := \left[E\left(\frac{A_1 A_2}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})} (Y - \nu_1^*(\mathbf{S})) \right)^2 \right]^{\frac{1}{2}},$$

$$T_{3,2} := \left[E\left(\frac{A_1}{\pi^*(\mathbf{S}_1)} (\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1)) \right)^2 \right]^{\frac{1}{2}},$$

$$T_{3,3} := \left[E\left(\frac{(1-A_1)(1-A_2)}{(1-\pi^*(\mathbf{S}_1))(1-\rho_0^*(\mathbf{S}))} (Y - \nu_0^*(\mathbf{S})) \right)^2 \right]^{\frac{1}{2}},$$

$$T_{3,4} := \left[E\left(\frac{1-A_1}{1-\pi^*(\mathbf{S}_1)} (\nu_0^*(\mathbf{S}) - \mu_0^*(\mathbf{S}_1)) \right)^2 \right]^{\frac{1}{2}},$$

$$T_{3,5} := \left[E\left(\mu_1^*(\mathbf{S}_1) - \mu_0^*(\mathbf{S}_1) - \theta \right)^2 \right]^{\frac{1}{2}}.$$

We bound each of the above terms in turn. Under Assumption 3.4 and recall the equation (3.154), we have

$$T_{3,1} \leq \frac{1}{c_0^2} [E(A_1 A_2 (Y - \nu_1^*(\mathbf{S}))^2)]^{\frac{1}{2}} \leq \frac{1}{c_0^2} \sqrt{E[\zeta^2]}. \tag{3.161}$$

Similarly, since $E[\varepsilon^2] \geq E[A_1 \varepsilon^2] = E[\varepsilon_1^2] = E[A_1(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))^2]$, we have

$$T_{3,2} \leq \frac{1}{c_0} [E(A_1 (\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))^2)]^{\frac{1}{2}} \leq \frac{1}{c_0} \sqrt{E[\varepsilon^2]}. \tag{3.162}$$

Repeating the same process for $T_{3,3}$ and $T_{3,4}$, we also have

$$T_{3,3} \leq \frac{1}{c_0^2} \sqrt{E[\zeta^2]}, \quad T_{3,4} \leq \frac{1}{c_0} \sqrt{E[\varepsilon^2]}. \tag{3.163}$$

Additionally,

$$\begin{aligned} \frac{2}{c_0} E[\zeta^2] + 2E[\varepsilon^2] &\overset{(i)}{\geq} E[A_1(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2] \overset{(ii)}{=} E[\pi(\mathbf{S}_1)(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2] \\ &\overset{(iii)}{\geq} c_0 E[(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2], \end{aligned}$$

where (i) holds by (3.158); (ii) holds by the law of iterated expectations; (iii) holds under Assumption 3.1. Similarly, we also have

$$\frac{2}{c_0}E[\zeta^2] + 2E[\varepsilon^2] \geq c_0 E[(\mu_0^*(\mathbf{S}_1) - \mu_0(\mathbf{S}_1))^2].$$

By Minkowski inequality,

$$\begin{aligned} T_{3,5} &\leq [E(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E(\mu_0^*(\mathbf{S}_1) - \mu_0(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E[\xi^2]]^{\frac{1}{2}} \\ &\leq 2\sqrt{\frac{2}{c_0^2}E[\zeta^2] + \frac{2}{c_0}E[\varepsilon^2]} + \sqrt{E[\xi^2]} \leq \frac{2\sqrt{2}}{c_0}\sqrt{E[\zeta^2]} + \frac{2\sqrt{2}}{\sqrt{c_0}}\sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}. \quad (3.164) \end{aligned}$$

Plugging (3.161)-(3.164) into (3.160), we have

$$[E(\psi(W;\eta^*))^2]^{\frac{1}{2}} = O\bigg(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\bigg).$$

b) When all the models are correctly specified, Assumption 3.1 implies Assumption 3.4. Hence, by part a), we also have (3.97). ∎

Proof of Lemma 3.11. In this proof, the expectations are taken w.r.t. the distribution of new observations $\mathbf{S}_1, \mathbf{S}_2$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). Additionally, we condition on the event $\mathcal{E}_4$, defined as (3.150). Under Assumption 3.7, such an event occurs with probability approaching one.

a) We first show (3.98). Same as in the proof of Lemma 3.7, we also have (3.147) here. Then, by Chebyshev's inequality, it suffices to show

$$\sum_{i=1}^{6}[E(Q_i^2)]^{\frac{1}{2}} = O_p\left(a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right),$$

where Q_i ($i \in \{1, \ldots, 6\}$) are defined as (3.130)-(3.135). Additionally, under Assumption 3.4, we also have

$$P(c_0 \leq \pi^*(\mathbf{S}_1) \leq 1 - c_0) = 1, \quad P(c_0 \leq \rho_1^*(\mathbf{S}) \leq 1 - c_0) = 1.$$

For the first term $[E(Q_1^2)]^{\frac{1}{2}}$, under Assumptions 3.4 and on the event $\mathcal{E}_4$,

$$
\begin{aligned}
&[E(Q_1^2)]^{\frac{1}{2}} \\
&\leq \frac{1}{c_0^4}\{E[A_1A_2\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})(Y-\widehat{\nu}_1(\mathbf{S})) - A_1A_2\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})(Y-\nu_1^*(\mathbf{S}))]^2\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \frac{1}{c_0^4}\{E[\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})(\nu_1^*(\mathbf{S})+\zeta-\widehat{\nu}_1(\mathbf{S})) - \widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})\zeta]^2\}^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} \frac{1}{c_0^4}\{E[\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S})]^2\}^{\frac{1}{2}} + \frac{1}{c_0^4}\{E[(\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S})-\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))\zeta]^2\}^{\frac{1}{2}}, \qquad (3.165)
\end{aligned}
$$

where (i) holds by the fact that $|A_1| \leq 1$, $|A_2| \leq 1$ and $A_1A_2Y = A_1A_2\nu_1^*(\mathbf{S}) + A_1A_2\zeta$; (ii) holds from Minkowski inequality and the fact that $P(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}) \leq 1) = 1$. Since $P(0 \leq \pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}) \leq 1) = 1$ and $P(0 \leq \widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S}) \leq 1) = 1$ under $\mathcal{E}_4$, we have

$$
[E(Q_1^2)]^{\frac{1}{2}} \leq \frac{1}{c_0^4}[E(\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^4}[E(\zeta^2)]^{\frac{1}{2}} = O_p\left(b_N + \sqrt{E[\zeta^2]}\right). \qquad (3.166)
$$

Similarly, for the second term $[E(Q_2^2)]^{\frac{1}{2}}$, under Assumptions 3.4 and on the event $\mathcal{E}_4$,

$$
\begin{aligned}
[E(Q_2^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^2}\{E[A_1\pi^*(\mathbf{S}_1)(\widehat{\nu}_1(\mathbf{S})-\widehat{\mu}_1(\mathbf{S}_1)) - A_1\widehat{\pi}(\mathbf{S}_1)(\nu_1^*(\mathbf{S})-\mu_1^*(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \frac{1}{c_0^2}\{E[\pi^*(\mathbf{S}_1)(\widehat{\nu}_1(\mathbf{S})-\widehat{\mu}_1(\mathbf{S}_1)) - \widehat{\pi}(\mathbf{S}_1)\varepsilon]^2\}^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} \frac{1}{c_0^2}[E(\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\
&\quad + \frac{1}{c_0^2}\{E[(\widehat{\pi}(\mathbf{S}_1)-\pi^*(\mathbf{S}_1))\varepsilon]^2\}^{\frac{1}{2}} \qquad (3.167) \\
&\overset{(iii)}{\leq} \frac{1}{c_0^2}[E(\widehat{\nu}_1(\mathbf{S})-\nu_1^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_1(\mathbf{S}_1)-\mu_1^*(\mathbf{S}_1))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}\{E[\varepsilon^2]\}^{\frac{1}{2}} \\
&= O_p\left(a_N + b_N + \sqrt{E[\varepsilon^2]}\right), \qquad (3.168)
\end{aligned}
$$

where (i) holds from the fact that $|A_1| \leq 1$ and $A_1\nu_1^*(\mathbf{S}) = A_1\mu_1^*(\mathbf{S}_1) + A_1\varepsilon$; (ii) holds from Minkowski inequality and $P(\pi^*(\mathbf{S}_1) \leq 1) = 1$; (iii) holds by the fact that $P(0 \leq \pi^*(\mathbf{S}_1) \leq$

1) = 1 and $P(0 \leq \widehat{\pi}(\mathbf{S}_1) \leq 1) = 1$ on $\mathcal{E}_4$. For the third term $[E(Q_3^2)]^{\frac{1}{2}}$, we have

$$[E(Q_3^2)]^{\frac{1}{2}} = O_p\left(b_N\right), \tag{3.169}$$

under Assumption 3.6. Combining (3.166), (3.168) and (3.169), we obtain that

$$[E(Q_1^2)]^{\frac{1}{2}} + [E(Q_2^2)]^{\frac{1}{2}} + [E(Q_3^2)]^{\frac{1}{2}} = O_p\left(a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right).$$

Repeating the same procedure above, we also have the same result for $[E(Q_4^2)]^{\frac{1}{2}} + [E(Q_5^2)]^{\frac{1}{2}} + [E(Q_6^2)]^{\frac{1}{2}}$. Then, (3.98) follows.

b) Now, we show (3.99). By (3.165), under Assumption 3.8, we have

$$\begin{aligned}
[E(Q_1^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^4}[E(\widehat{\nu}_1(\mathbf{S}) - \nu_1(\mathbf{S}))^2]^{\frac{1}{2}} \\
&\quad + \frac{1}{c_0^4}\{E[\zeta^2|\mathbf{S}]\}^{\frac{1}{2}}\{E[(\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S}) - \pi(\mathbf{S}_1)\rho_1(\mathbf{S}))]^2\}^{\frac{1}{2}} \\
&\leq \frac{1}{c_0^4}[E(\widehat{\nu}_1(\mathbf{S}) - \nu_1(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{\sqrt{CE[\zeta^2]}}{c_0^4}\{E[(\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S}) - \pi(\mathbf{S}_1)\rho_1(\mathbf{S}))]^2\}^{\frac{1}{2}}
\end{aligned}$$

By Minkowski inequality and under $\mathcal{E}_4$, we have

$$\begin{aligned}
&\{E[\widehat{\pi}(\mathbf{S}_1)\widehat{\rho}_1(\mathbf{S}) - \pi(\mathbf{S}_1)\rho_1(\mathbf{S})]^2\}^{\frac{1}{2}} \\
&\quad \leq \{E[(\widehat{\pi}(\mathbf{S}_1) - \pi(\mathbf{S}_1))\widehat{\rho}_1(\mathbf{S})]^2\}^{\frac{1}{2}} + \{E[\pi(\mathbf{S}_1)(\widehat{\rho}_1(\mathbf{S}) - \rho_1(\mathbf{S}))]^2\}^{\frac{1}{2}} \\
&\quad \leq [E(\widehat{\pi}(\mathbf{S}_1) - \pi(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E(\widehat{\rho}_1(\mathbf{S}) - \rho_1(\mathbf{S}))^2]^{\frac{1}{2}} = O_p\left(c_N + d_N\right).
\end{aligned}$$

Hence,

$$[E(Q_1^2)]^{\frac{1}{2}} = O_p\left(a_N + (c_N + d_N)\sqrt{E[\zeta^2]}\right).$$

In addition, by (3.167), we have

$$\begin{aligned}
[E(Q_2^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^2}[E(\widehat{\nu}_1(\mathbf{S}) - \nu_1(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)]^2]^{\frac{1}{2}} \\
&\quad + \frac{1}{c_0^2}\{E[\varepsilon^2|\mathbf{S}_1]\}^{\frac{1}{2}}\{E[(\widehat{\pi}(\mathbf{S}_1) - \pi(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\
&\leq \frac{1}{c_0^2}[E(\widehat{\nu}_1(\mathbf{S}) - \nu_1(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_1(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)]^2]^{\frac{1}{2}} \\
&\quad + \frac{\sqrt{CE[\varepsilon^2]}}{c_0^2}\{E[(\widehat{\pi}(\mathbf{S}_1) - \pi(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\
&= O_p\left(a_N + b_N + c_N\sqrt{E[\varepsilon^2]}\right).
\end{aligned}$$

Besides, by Assumption 3.6,

$$[E(Q_3^2)]^{\frac{1}{2}} = O_p\left(b_N\right).$$

Repeating the same procedure above, we also have

$$\begin{aligned}
[E(Q_4^2)]^{\frac{1}{2}} =& O_p\left(a_N + (c_N + d_N)\sqrt{E[\zeta^2]}\right), \\
[E(Q_5^2)]^{\frac{1}{2}} =& O_p\left(a_N + b_N + c_N\sqrt{E[\varepsilon^2]}\right), \\
[E(Q_6^2)]^{\frac{1}{2}} =& O_p\left(b_N\right).
\end{aligned}$$

Now, we have

$$[E(\psi(W;\widehat{\eta}) - \psi(W;\eta))^2]^{\frac{1}{2}} = O_P\left(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]}\right).$$

By Chebyshev's inequality, we conclude that (3.99) holds. ∎

Proof of Lemma 3.12. a) We notice the following representation:

$$\psi(W;\eta^*) - \theta = \sum_{i=1}^{8} O_i, \tag{3.170}$$

where

$$O_1 := \frac{A_1 A_2 (Y - \nu_1(\mathbf{S}))}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})}, \tag{3.171}$$

$$O_2 := \frac{A_1}{\pi^*(\mathbf{S}_1)}\left(1 - \frac{A_2}{\rho_1^*(\mathbf{S})}\right)(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S})), \tag{3.172}$$

$$O_3 := \frac{A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)}, \tag{3.173}$$

$$O_4 := -\frac{(1 - A_1)(1 - A_2)(Y - \nu_0(\mathbf{S}))}{(1 - \pi^*(\mathbf{S}_1))(1 - \rho_0^*(\mathbf{S}))}, \tag{3.174}$$

$$O_5 := -\frac{1 - A_1}{1 - \pi^*(\mathbf{S}_1)}\left(1 - \frac{1 - A_2}{1 - \rho_0^*(\mathbf{S})}\right)(\nu_0^*(\mathbf{S}) - \nu_0(\mathbf{S})), \tag{3.175}$$

$$O_6 := -\frac{(1 - A_1)(\nu_0(\mathbf{S}) - \mu_0(\mathbf{S}))}{1 - \pi^*(\mathbf{S}_1)}, \tag{3.176}$$

$$\begin{aligned} O_7 := &\left(1 - \frac{A_1}{\pi^*(\mathbf{S}_1)}\right)(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)) \\ &- \left(1 - \frac{1 - A_1}{1 - \pi^*(\mathbf{S}_1)}\right)(\mu_0^*(\mathbf{S}_1) - \mu_0(\mathbf{S}_1)), \end{aligned} \tag{3.177}$$

$$O_8 := \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta = \xi. \tag{3.178}$$

In the following, we demonstrate that

$$\sigma^2 = E(\psi(W; \eta^*) - \theta)^2 = \sum_{i=1}^{8} E[O_i^2]. \tag{3.179}$$

It suffices to show that $E[O_i O_j] = 0$ for all $i \neq j$. Firstly, since $A_1(1 - A_1) = 0$, we have

$$O_i O_j = 0, \text{ for each } i \in \{1, 2, 3\}, \text{ and } j \in \{4, 5, 6\}. \tag{3.180}$$

Step 1 We show $E[O_1 O_i] = 0$ for each $i \geq 2$. By (3.180), we know that $O_1 O_i = 0$ for $i \in \{4, 5, 6\}$. Note that, O_3, O_7, O_8 are all functions of $(\mathbf{S}, A_1)$. Hence, for each $i \in \{3, 7, 8\}$,

$$E[O_1 O_i] = E[O_i E[O_1 | \mathbf{S}, A_1 = 1] P(A_1 = 1 | \mathbf{S})] = 0,$$

since

$$E[O_1 | \mathbf{S}, A_1 = 1] \overset{(i)}{=} \frac{E[A_2 | \mathbf{S}, A_1 = 1] E[Y(1,1) - \mu_1(\mathbf{S}_1) | \mathbf{S}, A_1 = 1]}{\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S})} \overset{(ii)}{=} 0,$$

where (i) holds under Assumption 3.1; (ii) holds because $E[Y(1,1)|\mathbf{S}, A_1 = 1] = \mu_1(\mathbf{S}_1)$. Besides, we note that

$$\begin{aligned}
E[O_1 O_2] &= E\left[\frac{A_1 A_2 (Y - \nu_1(\mathbf{S}))(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))(\rho_1^*(\mathbf{S}) - 1)}{(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))^2}\right] \\
&\overset{(i)}{=} E\left[\frac{E[A_2(Y(1,1) - \nu_1(\mathbf{S}))|\mathbf{S}, A_1 = 1](\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))(\rho_1^*(\mathbf{S}) - 1)}{(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))^2} P(A_1 = 1|\mathbf{S})\right] \\
&\overset{(ii)}{=} E\left[\frac{\rho_1(\mathbf{S}) E[Y(1,1) - \nu_1(\mathbf{S})|\mathbf{S}, A_1 = 1](\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))(\rho_1^*(\mathbf{S}) - 1)}{(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))^2} P(A_1 = 1|\mathbf{S})\right] \\
&\overset{(iii)}{=} 0,
\end{aligned}$$

where (i) holds by the law of iterated expectations; (ii) holds under Assumption 3.1; (iii) holds because $E[Y(1,1)|\mathbf{S}, A_1 = 1] = \mu_1(\mathbf{S}_1)$.

Step 2 We show $E[O_2 O_i] = 0$ for each $i \geq 3$. By (3.180), we know that $O_2 O_i = 0$ for $i \in \{4, 5, 6\}$. Since O_3, O_7, O_8 are all functions of $(\mathbf{S}, A_1)$, it follows that, for each $i \in \{3, 7, 8\}$,

$$E[O_2 O_i] = E[O_i E[O_2|\mathbf{S}, A_1 = 1] P(A_1 = 1|\mathbf{S})] = 0,$$

since

$$\begin{aligned}
E[O_2|\mathbf{S}, A_1 = 1] &= \frac{\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S})}{\pi^*(\mathbf{S}_1)} \left(1 - \frac{E[A_2|\mathbf{S}, A_1 = 1]}{\rho_1^*(\mathbf{S})}\right) \\
&= \frac{\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S})}{\pi^*(\mathbf{S}_1)} \left(1 - \frac{\rho_1(\mathbf{S})}{\rho_1^*(\mathbf{S})}\right) \overset{(i)}{=} 0,
\end{aligned}$$

where (i) holds because either $\nu_1^*(\cdot) = \nu_1(\cdot)$ or $\rho_1^*(\cdot) = \rho_1(\cdot)$ by assumption.

Step 3 We show $E[O_3 O_i] = 0$ for each $i \geq 4$. By (3.180), we know that $O_3 O_i = 0$ for $i \in \{4, 5, 6\}$. Since O_7, O_8 are all functions of $(\mathbf{S}_1, A_1)$, it follows that, for each $i \in \{7, 8\}$,

$$E[O_3 O_i] = E[O_i E[O_3|\mathbf{S}_1, A_1 = 1] P(A_1 = 1|\mathbf{S}_1)] = 0,$$

since

$$E[O_3|\mathbf{S}_1, A_1 = 1] = \frac{E[\nu_1(\mathbf{S})|\mathbf{S}_1, A_1 = 1] - \mu_1(\mathbf{S}_1)}{\pi^*(\mathbf{S}_1)} \overset{(i)}{=} 0,$$

where (i) holds because $E[\nu_1(\mathbf{S})|\mathbf{S}_1, A_1 = 1] = \mu_1(\mathbf{S}_1)$ as shown in (3.146).

Step 4 By repeating the same procedure as in Steps 1-3, we also have $E[O_i O_j] = 0$ for each $i \in \{4, 5, 6\}$ and $j \geq i + 1$.

Step 5 We show $E[O_7 O_8] = 0$. Since O_8 is a function of $\mathbf{S}_1$, we have

$$E[O_7 O_8] = E[O_8 E[O_7|\mathbf{S}_1]] = 0,$$

since

$$\begin{aligned} E[O_7|\mathbf{S}_1] &= \left(1 - \frac{\pi(\mathbf{S})}{\pi^*(\mathbf{S}_1)}\right)(\mu_1^*(\mathbf{S}_1) - \mu_1(\mathbf{S}_1)) - \left(1 - \frac{1 - \pi(\mathbf{S})}{1 - \pi^*(\mathbf{S}_1)}\right)(\mu_0^*(\mathbf{S}_1) - \mu_0(\mathbf{S}_1)) \\ &\overset{(i)}{=} 0, \end{aligned}$$

where (i) holds because, by assumption, 1) either $\pi^*(\cdot) = \pi(\cdot)$ or $\mu_1^*(\cdot) = \mu_1(\cdot)$, and 2) either $\pi^*(\cdot) = \pi(\cdot)$ or $\mu_0^*(\cdot) = \mu_0(\cdot)$.

Based on all Steps 1-5, we conclude that (3.179) holds. Now, note that

$$\begin{aligned} E[O_1^2] &\geq E[A_1 A_2 (Y(1,1) - \nu_1(\mathbf{S}))^2], \\ E[O_2^2] &= E\left[\frac{A_1((\rho_1^*(\mathbf{S}))^2 - 2A_2\rho_1^*(\mathbf{S}) + A_2)}{(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))^2}(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2\right] \\ &= E\left[\frac{A_1((\rho_1^*(\mathbf{S}) - \rho_1(\mathbf{S}))^2 + \rho_1(\mathbf{S})(1 - \rho_1(\mathbf{S})))}{(\pi^*(\mathbf{S}_1)\rho_1^*(\mathbf{S}))^2}(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2\right] \\ &\geq c_0^2 E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2], \\ E[O_3^2] &= E\left[\frac{A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))^2}{(\pi^*(\mathbf{S}_1))^2}\right] \geq E[A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))^2] \end{aligned}$$

Hence,

$$E[A_1A_2\zeta^2] = E[\zeta_1^2] = E[A_1A_2(Y(1,1) - \nu_1^*(\mathbf{S}))^2]$$
$$\overset{(i)}{=} E[A_1A_2((Y(1,1) - \nu_1(\mathbf{S}))^2 + (\nu_1(\mathbf{S}) - \nu_1^*(\mathbf{S}))^2)] \leq E[O_1^2] + \frac{1}{c_0^2}E[O_2^2], \tag{3.181}$$

where (i) holds as in (3.155). Additionally,

$$E[A_1\varepsilon^2] = E[\varepsilon_1^2] = E[A_1(\nu_1^*(\mathbf{S}) - \mu_1^*(\mathbf{S}_1))^2]$$
$$\leq 3\left[E[A_1(\nu_1^*(\mathbf{S}) - \nu_1(\mathbf{S}))^2] + E[A_1(\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1))^2] + E[A_1(\mu_1(\mathbf{S}_1) - \mu_1^*(\mathbf{S}_1))^2]\right]$$
$$\leq \frac{3}{c_0^2}E[O_2^2] + 3E[O_3^2] + 3C_\mu\sigma^2.$$

Repeating the process above, we also have

$$E[(1 - A_1)(1 - A_2)\zeta^2] \leq E[O_4^2] + \frac{1}{c_0^2}E[O_5^2], \tag{3.182}$$
$$E[(1 - A_1)\varepsilon^2] \leq \frac{3}{c_0^2}E[O_5^2] + 3E[O_6^2] + 3C_\mu\sigma^2.$$

Besides, we also have

$$E[\xi^2] = E[O_8^2]. \tag{3.183}$$

Therefore, we conclude that

$$E[\zeta^2] + E[\varepsilon^2] + E[\xi^2]$$
$$= E[A_1A_2\zeta^2] + E[(1 - A_1)(1 - A_2)\zeta^2] + E[A_1\varepsilon^2] + E[(1 - A_1)\varepsilon^2] + E[\xi^2]$$
$$\leq E[O_1^2 + \frac{4}{c_0^2}O_2^2 + 3O_3^2 + O_4^2 + \frac{4}{c_0^2}O_5^2 + 3O_6^2 + O_8^2] + 6C_\mu\sigma^2 \leq \left(\frac{4}{c_0^2} + 6C_\mu\right)\sigma^2,$$

since $c < 1$ and (3.179) holds.

b) Now, we assume Assumption 3.3 holds. Same as in part a), we also have (3.179), (3.181), (3.182), and (3.183) hold. Additionally, under Assumption 3.3, by Lemma D.1 (iv)

of [CLCL19], we also have

$$E[\varepsilon^2] \leq 2\sigma_\varepsilon^2 \sigma^2.$$

Therefore,

$$\begin{aligned}
&E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \\
&\quad = E[A_1 A_2 \zeta^2] + E[(1-A_1)(1-A_2)\zeta^2] + E[\varepsilon^2] + E[\xi^2] \\
&\quad \leq E[O_1^2 + \frac{1}{c_0^2} O_2^2 + O_4^2 + \frac{1}{c_0^2} O_5^2 + O_8^2] + 2\sigma_\varepsilon^2 \sigma^2 \leq \left(\frac{1}{c_0^2} + 2\sigma_\varepsilon^2 \right) \sigma^2.
\end{aligned}$$

■

Proof of Lemma 3.13. We first show that (3.100) holds. By Lemma 3.12, we have

$$\psi(W; \eta^*) - \theta = \sum_{i=1}^{8} O_i, \quad \sigma^2 = E(\psi(W; \eta^*) - \theta)^2 = \sum_{i=1}^{8} E[O_i^2],$$

where $\{O_i\}_{i=1}^8$ are defined as (3.171)-(3.178). Since now we assume $\eta^* = \eta$ that all the models are correctly specified, we have $O_i = 0$ for $i \in \{2, 5, 7\}$ and hence

$$\psi(W; \eta^*) - \theta = O_1 + O_3 + O_4 + O_6 + O_8, \tag{3.184}$$

$$\sigma^2 = E[O_1^2] + E[O_3^2] + E[O_4^2] + E[O_6^2] + E[O_8^2] = \sum_{i=1}^{5} V_i,$$

where

$$\begin{aligned}
V_1 :=& E\left[\left(\frac{A_1 A_2}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})} (Y - \nu_1(\mathbf{S})) \right)^2 \right], \\
V_2 :=& E\left[\left(\frac{A_1}{\pi(\mathbf{S}_1)} (\nu_1(\mathbf{S}) - \mu_1(\mathbf{S}_1)) \right)^2 \right], \\
V_3 :=& E\left[\left(\frac{(1-A_1)(1-A_2)}{(1-\pi(\mathbf{S}_1))(1-\rho_0(\mathbf{S}))} (Y - \nu_0(\mathbf{S})) \right)^2 \right], \\
V_4 :=& E\left[\left(\frac{1-A_1}{1-\pi(\mathbf{S}_1)} (\nu_0(\mathbf{S}) - \mu_0(\mathbf{S}_1)) \right)^2 \right], \\
V_5 :=& E\left[(\mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta)^2 \right].
\end{aligned}$$

We lower bound each terms above:

$$V_1 \overset{(i)}{=} E\left[\left(\frac{\zeta_1}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})}\right)^2\right] \overset{(ii)}{=} E\left[\left(\frac{A_1A_2}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})}\zeta\right)^2\right] \overset{(iii)}{\geq} E[A_1A_2\zeta^2],$$
$$V_2 \overset{(iv)}{=} E\left[\left(\frac{\varepsilon_1}{\pi(\mathbf{S}_1)}\right)^2\right] \overset{(v)}{=} E\left[\left(\frac{A_1}{\pi(\mathbf{S}_1)}\varepsilon\right)^2\right] \overset{(vi)}{\geq} E[A_1\varepsilon^2],$$

where (i) and (iv) hold since $\nu_1^*(\cdot) = \nu_1(\cdot)$ and $\mu_1^*(\cdot) = \mu_1(\cdot)$; (ii) and (v) hold since $\zeta_1 = A_1A_2\zeta$ and $\varepsilon_1 = A_1\varepsilon$; (iii) and (vi) hold since $A_1, A_2 \in \{0,1\}$, $\pi(\mathbf{S}_1) \leq 1$ and $\rho_1(\mathbf{S}) \leq 1$ with probability 1 under Assumption 3.1. Similarly,

$$V_3 \geq E[(1-A_1)(1-A_2)\zeta^2], \quad V_4 \geq E[(1-A_1)\varepsilon^2].$$

Additionally, by definition, $\xi = \mu_1(\mathbf{S}_1) - \mu_0(\mathbf{S}_1) - \theta$. Hence,

$$V_5 = E[\xi^2].$$

Combining all the previous results, we have

$$\begin{aligned}
\sigma^2 :&= E[\psi(W;\eta^*) - \theta]^2 = E[\psi(W;\eta) - \theta]^2 \\
&\geq E[A_1A_2\zeta^2 + (1-A_1)(1-A_2)\zeta^2] + E[A_1\varepsilon^2 + (1-A_1)\varepsilon^2] + E[\xi^2] \\
&= E[\zeta^2] + E[\varepsilon^2] + E[\xi^2].
\end{aligned}$$

Next, we show that (3.101) holds. Recall the representation (3.184). By the finite form of Jensen's inequality, and note that the function $u \mapsto |u|^{2+t}$ is convex for any $t > 0$, we have

$$\begin{aligned}
\left|\frac{\psi(W;\eta) - \theta}{5}\right|^{2+t} &= \left|\frac{O_1 + O_3 + O_4 + O_6 + O_8}{5}\right|^{2+t} \\
&\leq \frac{|O_1|^{2+t} + |O_3|^{2+t} + |O_4|^{2+t} + |O_6|^{2+t} + |O_8|^{2+t}}{5}
\end{aligned}$$

Therefore,

$$E|\psi(W;\eta)-\theta|^{2+t} \le 5^{1+t}E[|O_1|^{2+t}+|O_3|^{2+t}+|O_4|^{2+t}+|O_6|^{2+t}+|O_8|^{2+t}]$$
$$= C_t\sum_{i=1}^{5} V_i',$$

where $C_t = 5^{1+t}$ and

$$V_1' := E\left[\left|\frac{A_1A_2}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})}(Y-\nu_1(\mathbf{S}))\right|^{2+t}\right],$$
$$V_2' := E\left[\left|\frac{A_1}{\pi(\mathbf{S}_1)}(\nu_1(\mathbf{S})-\mu_1(\mathbf{S}_1))\right|^{2+t}\right],$$
$$V_3' := E\left[\left|\frac{(1-A_1)(1-A_2)}{(1-\pi(\mathbf{S}_1))(1-\rho_0(\mathbf{S}))}(Y-\nu_0(\mathbf{S}))\right|^{2+t}\right],$$
$$V_4' := E\left[\left|\frac{1-A_1}{1-\pi(\mathbf{S}_1)}(\nu_0(\mathbf{S})-\mu_0(\mathbf{S}_1))\right|^{2+t}\right],$$
$$V_5' := E\left[|\mu_1(\mathbf{S}_1)-\mu_0(\mathbf{S}_1)-\theta|^{2+t}\right].$$

Now, we upper bound each of the terms above.

$$V_1' \overset{(i)}{=} E\left[\left|\frac{\zeta_1}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})}\right|^{2+t}\right] \overset{(ii)}{=} E\left[\left|\frac{A_1A_2}{\pi(\mathbf{S}_1)\rho_1(\mathbf{S})}\zeta\right|^{2+t}\right] \overset{(iii)}{\le} \frac{1}{c_0^{4+2t}}E[|\zeta|^{2+t}],$$
$$V_2' \overset{(iv)}{=} E\left[\left|\frac{\varepsilon_1}{\pi(\mathbf{S}_1)}\right|^{2+t}\right] \overset{(v)}{=} E\left[\left|\frac{A_1}{\pi(\mathbf{S}_1)}\varepsilon\right|^{2+t}\right] \overset{(vi)}{\le} \frac{1}{c_0^{4+2t}}E[|\varepsilon|^{2+t}],$$

where (i) and (iv) hold since $\nu_1^*(\cdot)=\nu_1(\cdot)$ and $\mu_1^*(\cdot)=\mu_1(\cdot)$; (ii) and (v) hold since $\zeta_1 = A_1A_2\zeta$ and $\varepsilon_1 = A_1\varepsilon$; (iii) and (vi) hold since $A_1, A_2 \in \{0,1\}$, $\pi(\mathbf{S}_1), \rho_1(\mathbf{S}) \in [c_0, 1-c_0]$ with probability 1 under Assumption 3.1. Similarly, we also have

$$V_3' \le \frac{1}{c_0^{4+2t}}E[|\zeta|^{2+t}], \quad V_4' \le \frac{1}{c_0^{2+t}}E[|\varepsilon|^{2+t}].$$

In addition, by definition, $\xi = \mu_1(\mathbf{S}_1)-\mu_0(\mathbf{S}_1)-\theta$. Hence,

$$V_5' = E[|\xi|^{2+t}].$$

Therefore, we conclude that

$$E|\psi(W;\eta)-\theta|^{2+t} \le C_t\left[\frac{2}{c_0^{4+2t}}E[|\zeta|^{2+t}]+\frac{2}{c_0^{2+t}}E[|\varepsilon|^{2+t}]+E[|\xi|^{2+t}]\right]$$
$$\le \frac{2C_t}{c_0^{4+2t}}E[|\zeta|^{2+t}+|\varepsilon|^{2+t}+|\xi|^{2+t}],$$

since $0<c<1$ and $t>0$. ∎

Proof of Lemma 3.14. We show that for each $k=1,...,K$,

$$\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\eta)-\theta)^2-\sigma^2=o_p(\sigma^2), \tag{3.185}$$

$$\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta})-\widehat{\theta})^2-\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\eta)-\theta)^2=o_p(\sigma^2), \tag{3.186}$$

We first show (3.185). Let $Z_{N,i}:=\sigma^{-1}(\psi(W_i;\eta)-\theta)^2-1$, note that both W_i and η are possibly dependent with $(d_1,d_2)=(d_{N,1},d_{N,2})$. Hence, $(Z_{N,i})_{N,i}$ forms a row-wise independent and identically distributed triangular array, and (3.185) is equivalent to

$$\frac{1}{n}\sum_{i\in I_k}Z_i=o(1).$$

By Lemma 3 of [ZB21], it suffices to show that $E(Z_{d,1})=0$ and $E|Z_{d,1}|^q<C'$ with some constants $q>1$ and $C'>0$. By definition,

$$E(Z_{d,1})=E\left[\frac{(\psi(W;\eta)-\theta)^2}{\sigma^2}-1\right]=\frac{\sigma^2}{\sigma^2}-1=0.$$

In addition, by Minkowski inequality,

$$\left[E\left|\frac{(\psi(W;\eta)-\theta)^2}{\sigma^2}-1\right|^{\frac{2+t}{2}}\right]^{\frac{2}{2+t}} \le \left[\frac{E|(\psi(W;\eta)-\theta)|^{2+t}}{\sigma^{2+t}}\right]^{\frac{2}{2+t}}+1<C+1.$$

It follows that

$$E|Z_{d,1}|^{\frac{2+t}{2}}=E\left|\frac{(\psi(W;\eta)-\theta)^2}{\sigma^2}-1\right|^{\frac{2+t}{2}}<(C+1)^{\frac{2+t}{2}},$$

with $(2+t)/2 > 1$. Therefore, by Lemma 3 of [ZB21], we conclude that (3.185) holds.

Next, we show (3.186). Let $a_i = \psi(W_i;\widehat{\eta}) - \psi(W_i;\eta) - (\widehat{\theta} - \theta)$ and $b_i = \psi(W_i;\eta) - \theta$. Then, it follows from the triangle inequality that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \widehat{\theta})^2 - \frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\eta) - \theta)^2\right| \\
&\quad\leq \frac{1}{n}\sum_{i\in I_k}|a_i|\cdot|a_i + 2b_i| \overset{(i)}{\leq} \left[\frac{1}{n}\sum_{i\in I_k}a_i^2\right]^{\frac{1}{2}}\cdot\left[\frac{1}{n}\sum_{i\in I_k}(a_i+2b_i)^2\right]^{\frac{1}{2}} \\
&\quad\overset{(ii)}{\leq} \left[\frac{1}{n}\sum_{i\in I_k}a_i^2\right]^{\frac{1}{2}}\cdot\left[\left(\frac{1}{n}\sum_{i\in I_k}a_i^2\right)^{\frac{1}{2}} + 2\left(\frac{1}{n}\sum_{i\in I_k}b_i^2\right)^{\frac{1}{2}}\right],
\end{aligned}$$

where (i) follows from Cauchy-Schwarz inequality; (ii) follows from Minkowski inequality. Recall the equation (3.185), we have

$$\frac{1}{n}\sum_{i\in I_k}b_i^2 = \frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\eta)-\theta)^2 = \sigma^2(1+o_p(1)).$$

Since, by assumption, $\widehat{\theta} - \theta = O_p(\sigma/\sqrt{N})$ and $[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta}) - \psi(W_i;\eta)|^2]^{\frac{1}{2}} = o_p(\sigma)$, we have

$$\left[\frac{1}{n}\sum_{i\in I_k}a_i^2\right]^{\frac{1}{2}} \leq \left[\frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\widehat{\eta}) - \psi(W_i;\eta)|^2\right]^{\frac{1}{2}} + |\widehat{\theta} - \theta| = o_p(\sigma).$$

Therefore,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \widehat{\theta})^2 - \frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\eta) - \theta)^2\right| \\
&\quad= o_p(\sigma)\cdot[o_p(\sigma) + \sigma(1+o_p(1))] = o_p(\sigma^2).
\end{aligned}$$

Now, by (3.185) and (3.186), we have

$$\begin{aligned}\widehat{\sigma}^2 - \sigma^2 &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \widehat{\theta})^2 - \sigma \\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i\in I_k}(\psi(W_i;\widehat{\eta}) - \widehat{\theta})^2 - (\psi(W_i;\eta) - \theta)^2 + (\psi(W_i;\eta) - \theta)^2 - \sigma\right) \\ &= o_p(\sigma^2).\end{aligned}$$

■

3.9 Acknowledgement

Chaper 3, in full, has been submitted for publication of the material. Bradic, Jelena; Ji, Weijie; Zhang, Yuqian. High-dimensional inference for dynamic treatment effects. The dissertation author was one of the primary investigators and authors of this material.

Chapter 4

Dynamic treatment effects: high-dimensional inference under model misspecification

4.1 Introduction

Statistical inference and estimation for causal relationships has a long tradition and has attracted significant attention as the emerging of large and complex datasets and the need for new statistical tools to handle such challenging datasets. In many applications, data is collected dynamically over time, and individuals are exposed to treatments at multiple stages. Typical examples include mobile health datasets, electronic health records, and many other biomedical studies and political science datasets. This work considers statistical inference of causal effects for longitudinal and observational data with high-dimensional covariates (confounders). We aim to establish valid statistical inference for dynamic treatment effects

under possible *model misspecifications*.

For the sake of simplicity, we consider dynamic settings with two exposure times. Suppose that we collect independent and identically distributed (i.i.d.) samples $\mathbb{S} := (\mathbf{W}_i)_{i=1}^N := (Y_i, A_{1i}, A_{2i}, \mathbf{S}_{1i}, \mathbf{S}_{2i})_{i=1}^N$, and let $\mathbf{W} := (Y, A_1, A_2, \mathbf{S}_1, \mathbf{S}_2)$ be an independent copy of $\mathbf{W}_i$. Here, $Y \in \mathbb{R}$ denotes the observed outcome variable at the final stage, $Y(a_1, a_2)$ is the (unobservable) potential outcome for $a_1, a_2 \in \{0, 1\}$, and we assume the standard consistency condition $Y = Y(A_1, A_2)$ throughout; $A_1, A_2 \in \{0, 1\}$ denote the treatment indicator variable at time 1 and time 2, respectively; $\mathbf{S}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{S}_2 \in \mathbb{R}^{d_2}$ denote the covariate (or confounder) variables at time 1 and time 2, respectively. Suppose the first coordinate of $\mathbf{S}_1$ is 1. Let $\bar{\mathbf{S}}_{2i} := (\mathbf{S}_{1i}^T, \mathbf{S}_{2i}^T)^T$, $\bar{\mathbf{S}}_2 := (\mathbf{S}_1^T, \mathbf{S}_2^T)^T$, $d := d_1 + d_2$, and possibly, $d \gg N$ as $N \to \infty$. Define the following counterfactual mean parameters:

$$\theta_{a_1,a_2} := E\{Y(a_1, a_2)\}, \quad \text{for any} \quad a_1, a_2 \in \{0, 1\}.$$

The dynamic treatment effect of the given a treatment sequence (a_1, a_2) compared with the control sequence (a_1', a_2') can then be defined as the difference: DTE $:= \theta_{a_1,a_2} - \theta_{a_1',a_2'}$. To estimate the dynamic treatment effect, it suffices to estimate θ_{a_1,a_2} and $\theta_{a_1',a_2'}$ seperately. Without loss of generality, we focus on the inference of the counterfactual mean $\theta_{1,1}$. Similar results also hold analogously for $\theta_{0,0}$, $\theta_{0,1}$, $\theta_{1,0}$, and statistical inference for the dynamic treatment effect can be further provided by combining the results for the two counterfactual means.

To identify the parameter of interest under dynamic settings, we consider the marginal structural mean (MSM) models. Three different approaches are studied under the MSM models: the inverse probability weighting (IPW) method, the covariate balancing method,

and the doubly robust method. The IPW method has been well-studied by, e.g., [Rob86, Rob00a, HBR01, Rob04, BAWM18]. The consistencies of IPW estimators require correctly specified propensity score (PS) models and statistical inference is usually only valid in low dimensions. In addition, [ZW18, KS18, YS18, VB21] proposed covariate balancing dynamic treatment effect estimators, where correctly specified outcome regression (OR) models are required. To achieve $\sqrt{N}$-inference in high dimensions, they also need very strong sparsity conditions for the OR models' parameters, e.g., with a sparsity level $o(N^{1/8}\log^{-3/4}(Nd))$ as in [VB21]. The doubly robust method is studied using a doubly robust representation for the parameter of interest, which involves both the PS and the OR models [IR15a, LM05, Mur03, Rob00a, Rob87]. With the presence of high-dimensional covariates, doubly robust estimators for the dynamic treatment effects have been recently studied by [BHL20, BJZ21], and $\sqrt{N}$-inference is provided when all the nuisance models are correctly specified.

We consider the doubly robust approach, but with carefully constructed "moment targeted" nuisance estimators. Based on such nuisance estimators, we achieve $\sqrt{N}$ inference for the parameter of interest even model misspecification occurs. This is the first result establishing statistical inference for dynamic treatment effects under high-dimensional settings that allows model misspecifications. Specifically, our proposed estimator is *sequentially doubly robust*. That is, the estimator is consistent and asymptotically normal (CAN) as long

as one of the following cases occurs:

The OR models at time 1 and 2 are correctly specified; (4.1)

The PS models at time 1 and 2 are correctly specified; (4.2)

The OR model at time 1 and the PS model at time 2 are correctly specified; (4.3)

The OR model at time 2 and the PS model at time 1 are correctly specified. (4.4)

In other words, we require at least one of the models to be correctly specified at each time spot; see results in Theorem 4.1 and Assumption 4.2 required therein. We reach the best model robustness so far, even containing the existing results in low dimensions.

In low-dimensional settings, doubly robust estimators for dynamic treatment effects have been studied by [Rob00b, MvdLRG01, BR05, YvdL06]. Their proposed estimators are CAN when either (4.1) or (4.2) holds, but (4.3) and (4.4) are not allowed. Recently, [BRR19] proposed a "multiple robust" estimator (also in low-dimensions), which reaches the best model robustness so far in the existing literature. Their estimator is CAN when any of (4.1), (4.2), or (4.3) holds, but case (4.4) is not allowed. Additionally, our estimator is also more robust than the IPW and covariate balancing estimators. The IPW estimators always require correctly specified PS models, and the covariate balancing estimators always require correctly specified OR models. Whereas we do not enforce any single model to be correctly specified, and hence weaker conditions on the model correctness are assumed in our work.

Apart from the MSM models, the dynamic treatment effect can also be identified through the structural nested mean (SNM) models, and G-computation has been used to estimate the parameter of interest [Rob86]. Recently, [LS20] proposed estimators for some "blip functions" under SNM models in high dimensions. However, they always require cor-

rectly specified blip functions, and they provided statistical inference for the counterfactual mean estimator only in low dimensions.

Furthermore, even when all the models are correctly specified, we require weaker sparsity conditions than the existing doubly robust literature in high dimensions [BHL20, BJZ21], where three "product sparsity" conditions are required. Whereas, we only need two "product sparsity" conditions; see Theorem 4.2 and comparisons in Remark 4.8. This is because, based on a special "doubly robust type" estimator for the OR model at time 1, we can achieve a better consistency rate and hence result in a weaker condition on the counterfactual mean estimator; see discussion in Remark 4.12.

The average treatment effect estimation problem is closely related to the dynamic treatment effect estimation problem – it can be seen as a special (degenerated) problem with non-longitudinal data. The average treatment effect estimation has a long tradition [Rub74], and it has attracted a significant amount of recent attention with the appearance of high-dimensional covariates; e.g., [Far15, AIW18, CCD$^+$18, SRR19, BWZ19, Tan20a]. Statistical inference for the average treatment effect under model misspecifications has been studied by [SRR19, Tan20a, DV20, DAV20, AV21]. They have proposed "model doubly robust" estimators, which are shown to be CAN as long as either the OR model or the PS model is correctly specified. Their estimators are constructed based on a doubly robust representation for the average treatment effect and some special nuisance estimators. Among these work, [SRR19] required the weakest sparsity conditions, and our results reach the same conditions if a degenerated non-longitudinal case is considered; see discussion in Remark 4.7. In addition, [BWZ19] also proposed another average treatment effect estimator based on the same type of nuisance estimators but with a special type of cross-fitting. They only focused

on correctly specified models, and CAN has been achieved requiring sparsity conditions different from the literature; see more details in Remark 4.8.

The optimal dynamic treatment regime [Mur03] is another related field, where their goal is to provide optimal individualized treatments over time under dynamic settings. Estimators for the optimal dynamic treatment regime have been proposed using reinforcement learning algorithms, such as Q-learning [Wat89, WD92, Mur05, CMS10, MR10, SWZK15, LLL$^+$19, FWXY20], A-learning [Mur03, BMZ04, Rob04, SFSL18], and some other recent methods, e.g., IQ-learning [LLS14, LLS17] and V-learning [LLK$^+$20]. In addition, [NBW21] also tackled another related problem recently, where their purpose is to learn when a treatment should perform if the treatment is only allowed to act once.

Notation We use the following notation throughout. Let $P(\cdot)$ and $E(\cdot)$ denote the probability measure and expectation characterizing the joint distribution of the underlying random vector $\mathbb{W} := (\{Y(a_1, a_2)\}_{a_1,a_2\in\{0,1\}}, A_1, A_2, \mathbf{S}_1, \mathbf{S}_2)$, respectively. For any $\alpha > 0$, let $\psi_\alpha(\cdot)$ denote the function given by $\psi_\alpha(x) := \exp(\alpha^2) - 1$, $\forall x > 0$. The ψ_α-Orlicz norm $\|\cdot\|_{\psi_\alpha}$ of a random variable $X \in \mathbb{R}$ is defined as $\|X\|_{\psi_\alpha} := \inf\{c > 0 : E[\psi_\alpha(|X|/c)] \leq 1\}$. Two special cases of finite $\psi_\alpha-$Orlicz norm are given by $\psi_2(x) = \exp(x^2) - 1$ and $\psi_1(x) = \exp(x) - 1$, which correspond to sub-Gaussian and sub-exponential random variables, respectively. The notation $a_N \asymp b_N$ denotes $cb_N \leq a_N \leq Cb_N$ for all $N \geq 1$ and with some constants $c, C > 0$. For any $\tilde{\mathbb{S}} \subseteq \mathbb{S} = (\mathbf{Z}_i)_{i=1}^N$, define $P_{\tilde{\mathbb{S}}}$ as the joint distribution of $\tilde{\mathbb{S}}$ and $E_{\tilde{\mathbb{S}}}(f) = \int f dP_{\tilde{\mathbb{S}}}$. For $r \geq 1$, define the l_r-norm of a vector $\mathbf{z}$ with $\|\mathbf{z}\|_r := (\sum_{j=1}^p |\mathbf{z}(j)|^r)^{1/r}$, $\|\mathbf{z}\|_0 := |\{j : \mathbf{z}(j) \neq 0\}|$, and $\|\mathbf{z}\|_\infty := \max_j |\mathbf{z}(j)|$.

4.2 Doubly robust representation and working models

In Section 4.2.1, we first introduce a doubly robust representation for the counterfactual mean, $\theta_{1,1}$. Such a representation has already been studied by the literature and is constructed based on a doubly robust score, (4.11), satisfying the "Neyman orthogonality" [CCD$^+$18]. Estimators based on doubly robust scores are known to be asymptotically normal when all nuisance models are correctly specified, and some product rate conditions are satisfied. Such a property is also known as "rate double robustness" in non-dynamic settings; see, e.g., [SRR19]. However, when model misspecification occurs, the doubly robust score does not guarantee a $\sqrt{N}$-inference. To reduce the bias originating from model misspecification, we propose novel working models for the nuisance functions based on special loss functions in Section 4.2.2. Further discussions about the model correctness of the proposed working nuisance models are then provided in Section 4.2.3.

4.2.1 A doubly robust representation in dynamic settings

To identify the counterfactual mean $\theta_{1,1}$, we assume the standard sequential ignorability, consistency, and overlap conditions; see, e.g., [IR15a, LM05, Mur03, Rob00a, Rob87].

Assumption 4.1 (Basic assumptions)**.** *(a) Sequential ignorability: for each $a \in \{0, 1\}$,*

$$Y(a_1, a_2) \perp\!\!\!\perp A_1 \mid \mathbf{S}_1, \quad Y(a_1, a_2) \perp\!\!\!\perp A_2 \mid (\mathbf{S}_1, \mathbf{S}_2, A_1 = a_1).$$

(b) Consistency:

$$Y = Y(A_1, A_2).$$

(c) Overlap: define the (true) PS functions at time 1 and time 2 as

$$\pi(\mathbf{s}_1) := E(A_1 \mid \mathbf{S}_1 = \mathbf{s}_1), \;\; \rho(\bar{\mathbf{s}}_2) := E(A_2 \mid \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, A_1 = 1), \;\; \forall \mathbf{s}_1 \in \mathbb{R}^{d_1}, \bar{\mathbf{s}}_2 \in \mathbb{R}^d, \quad (4.5)$$

respectively. Let

$$P(c_0 < \pi(\mathbf{S}_1) < 1 - c_0) = 1, \quad P(c_0 < \rho(\bar{\mathbf{S}}_2) < 1 - c_0) = 1, \quad (4.6)$$

with some constant $c_0 \in (0, 1)$. Additionally, let $\pi^*(\cdot)$ and $\rho^*(\cdot)$ be some functions satisfying

$$P(c_0 < \pi^*(\mathbf{S}_1) < 1 - c_0) = 1, \quad P(c_0 < \rho^*(\bar{\mathbf{S}}_2) < 1 - c_0) = 1. \quad (4.7)$$

Here, $\pi^*(\cdot)$ and $\rho^*(\cdot)$ are the working models for the true PS functions $\pi(\cdot)$ and $\rho(\cdot)$, respectively. We consider possible model misspecifications that $\pi^*(\cdot) \neq \pi(\cdot)$ and $\rho^*(\cdot) \neq \rho(\cdot)$ are allowed. The condition (4.7) is an overlap condition for the working PS models, which reaches the usual overlap condition (4.6) when the PS models are correctly specified.

In addition, we denote the (true) OR functions as

$$\mu(\mathbf{s}_1) := E\{Y(1,1) \mid \mathbf{S}_1 = \mathbf{s}_1\}, \;\; \nu(\bar{\mathbf{s}}_2) := E\{Y(1,1) \mid \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, A_1 = 1\}, \;\; \forall \mathbf{s}_1 \in \mathbb{R}^{d_1}, \bar{\mathbf{s}}_2 \in \mathbb{R}^d.$$

Similarly, let $\mu^*(\cdot)$ and $\nu^*(\cdot)$ be the working models for the true OR functions $\mu(\cdot)$ and $\nu(\cdot)$, respectively, and misspecified OR models are also allowed.

We let the following condition holds:

Assumption 4.2 (Sequential model double robustness)**.** *Let (a) either $\pi(\cdot) = \pi^*(\cdot)$ or $\mu(\cdot) = \mu^*(\cdot)$ holds, but not necessarily both; and (b) either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$, but not necessarily both.*

Under Assumptions 4.1 and 4.2, the following doubly robust representation holds:

$$\theta_{1,1} = E\left[\mu^*(\mathbf{S}_1) + \frac{A_1\{\nu^*(\bar{\mathbf{S}}_2) - \mu^*(\mathbf{S}_1))}{\pi^*(\mathbf{S}_1)} + \frac{A_1 A_2(Y - \nu^*(\bar{\mathbf{S}}_2))}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right]. \quad (4.8)$$

The above doubly robust representation has been also studied by, e.g., [NBW21, TYWK+19, vdLG11, ORR10, MvdLRG01, BHL20, BJZ21].

With the presence of high-dimensional covariates, [BHL20, BJZ21] proposed mean estimators based on the doubly robust representation (4.8). Their proposed estimators are asymptotically normally distributed when all the nuisance models are correctly specified. However, when model misspecification occurs, their estimators are only shown to be consistent under Assumption 4.2, and statistical inference is not valid under such scenarios.

In this chapter, we also consider the standard doubly robust representation (4.8). However, as discussed in the following Section 4.2.2, using carefully chosen nuisance parameters, we can achieve asymptotic normality and hence construct valid inference even when model misspecification occurs.

4.2.2 Construction of the sequential model double robustness

In this chapter, we consider linear working models for the OR functions and logistic working models for the PS functions: for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$ and $\bar{\mathbf{s}}_2 \in \mathbb{R}^d$, let

$$\pi^*(\mathbf{s}_1) := g(\mathbf{s}_1^T \boldsymbol{\gamma}^*), \quad \rho^*(\bar{\mathbf{s}}_2) := g(\bar{\mathbf{s}}_2^T \boldsymbol{\delta}^*), \quad \nu^*(\bar{\mathbf{s}}_2) := \bar{\mathbf{s}}_2^T \boldsymbol{\alpha}^*, \quad \mu^*(\mathbf{s}_1) := \mathbf{s}_1^T \boldsymbol{\beta}^*, \tag{4.9}$$

where $\boldsymbol{\eta}^{*T} := (\boldsymbol{\gamma}^{*T}, \boldsymbol{\delta}^{*T}, \boldsymbol{\alpha}^{*T}, \boldsymbol{\beta}^{*T})^T$ are some carefully chosen population nuisance parameters defined later in (4.12)-(4.15), and $g(\cdot)$ is the logistic function that

$$g(u) = \frac{\exp(u)}{1+\exp(u)} = \frac{1}{1+\exp(-u)}, \quad \text{for all } \ u \in \mathbb{R}. \tag{4.10}$$

Based on the doubly robust representation (4.8) with linear and logistic working models, we consider the following (uncentered) doubly robust score function:

$$\psi(\mathbf{W};\boldsymbol{\eta}) := \mathbf{S}_1^T\boldsymbol{\beta} + \frac{A_1(\bar{\mathbf{S}}_2^T\boldsymbol{\alpha} - \mathbf{S}_1^T\boldsymbol{\beta})}{g(\mathbf{S}_1^T\boldsymbol{\gamma})} + \frac{A_1A_2(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha})}{g(\mathbf{S}_1^T\boldsymbol{\gamma})g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})}, \tag{4.11}$$

for any arbitrary $\boldsymbol{\eta} := (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$. Then, under Assumptions 4.1 and 4.2, we have $E\{\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \theta_{1,1}$. Let $\widehat{\boldsymbol{\eta}} := (\widehat{\boldsymbol{\gamma}}^T, \widehat{\boldsymbol{\delta}}^T, \widehat{\boldsymbol{\alpha}}^T, \widehat{\boldsymbol{\beta}}^T)^T$ be an estimate of $\boldsymbol{\eta}^*$. Then, a doubly robust mean estimator has the following form:

$$\widehat{\theta}_{1,1} = N^{-1}\sum_{i=1}^{N}\psi(\mathbf{W}_i;\widehat{\boldsymbol{\eta}}).$$

For the sake of simplicity, let $\widehat{\boldsymbol{\eta}} \perp\!\!\!\perp (\mathbf{W}_i)_{i=1}^N$. Then, by Taylor's theorem, we have the following expression for the bias:

$$\begin{aligned} E(\widehat{\theta}_{1,1}) - \theta_{1,1} =& E\{\psi(\mathbf{W};\widehat{\boldsymbol{\eta}}) - \psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \Delta_1 + \Delta_2, \text{ where} \\ \Delta_1 :=& E\{\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\}^T(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*), \end{aligned}$$

and Δ_2 is some remainder term potentially of the order $o(N^{-1/2})$. Since the score function (4.11) satisfies the "Neyman orthogonality" [CCD+18], we have $E\{\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ and hence $\Delta_1 = 0$ when all the nuisance models are correctly specified. Such a fact is only originated from the construction of the score (4.11) and is independent of the choice of the target nuisance paramters, $\boldsymbol{\eta}^*$. However, when model misspecification occurs, the "Neyman orthogonality" does not guarantee that $E\{\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ and the bias term Δ_1 is unignorable since, in high dimensions, the convergence rate of $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*$ is typically slower than the paramteric rate $N^{-1/2}$.

To reduce the bias, we construct the target population nuisance parameters $\boldsymbol{\eta}^*$ in a

way such that $E\{\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ always holds even when misspecification occurs: define

$$\begin{aligned}
\boldsymbol{\gamma}^* &:= \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_1}} E\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\}, && (4.12)\\
\boldsymbol{\delta}^* &:= \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^{d}} E\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\}, && (4.13)\\
\boldsymbol{\alpha}^* &:= \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{d}} E\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha})\}, && (4.14)\\
\boldsymbol{\beta}^* &:= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} E\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta})\}, && (4.15)
\end{aligned}$$

where, for any $\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathbb{R}^{d_1}$ and $\boldsymbol{\alpha},\boldsymbol{\delta}\in\mathbb{R}^{d}$, the loss functions are defined as

$$\begin{aligned}
\ell_1(\mathbf{W};\boldsymbol{\gamma}) &:= (1-A_1)\mathbf{S}_1^T\boldsymbol{\gamma} + A_1\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}), && (4.16)\\
\ell_2(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta}) &:= \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma})}\left\{(1-A_2)\bar{\mathbf{S}}_2^T\boldsymbol{\delta} + A_2\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\}, && (4.17)\\
\ell_3(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}) &:= \frac{A_1A_2\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta})}{g(\mathbf{S}_1^T\boldsymbol{\gamma})}(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha})^2, && (4.18)\\
\ell_4(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}) &:= A_1\exp(-\mathbf{S}_1^T\boldsymbol{\gamma})\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha} + \frac{A_2(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha})}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})} - \mathbf{S}_1^T\boldsymbol{\beta}\right\}^2. && (4.19)
\end{aligned}$$

The loss functions (4.16)-(4.19) are carefully chosen such that

$$\begin{aligned}
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -E\left[A_1\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* + \frac{A_2(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)} - \mathbf{S}_1^T\boldsymbol{\beta}^*\right\}\mathbf{S}_1\right]\\
&= \nabla_{\boldsymbol{\beta}}E\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\}/2,\\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -E\left[\frac{A_1A_2\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\bar{\mathbf{S}}_2\right]\\
&= \nabla_{\boldsymbol{\alpha}}E\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\}/2,\\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= E\left[\frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\left\{1-\frac{A_2}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right\}\bar{\mathbf{S}}_2\right] = \nabla_{\boldsymbol{\delta}}E\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\},\\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= E\left[\left\{1-\frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right\}\mathbf{S}_1\right] = \nabla_{\boldsymbol{\gamma}}E\{\ell_1(\mathbf{W};\boldsymbol{\gamma}^*)\}.
\end{aligned}$$

By the constructions (4.12)-(4.15) and the first-order optimality conditions, the left-hand sides of every equalities above are all zero vectors, and hence $E\{\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ is guar-

anteed even when the nuisance models are misspecified. Now, it follows that $\Delta_1 = 0$ and, when Δ_2 is small enough, the bias is ignorable. Hence, valid $\sqrt{N}$-inference is possible under model misspecifications.

Remark 4.1 (Discussion for the loss functions)**.** *The loss function designed for the PS model at time 1,* (4.16)*, coincides with the PS model's loss function studied by [SRR19, Tan20a, AV21, BWZ19], where non-longitudinal data was considered therein. The second loss function,* (4.17)*, can be seen as a weighted version of* (4.16)*. The loss function for the OR model at time 2,* (4.18)*, is a weighted squre loss, and it can also be viewed as a weighted version of the OR model's loss function in [SRR19, Tan20a, AV21, BWZ19]. Lastly, the loss function* (4.19) *can also be seen as a weighted square loss, with a "doubly robust" type outcome* $\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y(1,1) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)$*; see more discussion in Remark 4.3.*

4.2.3 Correctness of the nuisance models

The nested models are hard to interpret, especially the OR model at time spot 1, $\mu(\cdot)$. Such difficulty has been discussed by, e.g., [BRR19]. In below, we also provide discussion and illustrations on the correctness of the carefully designed nuisance models, (4.9).

Remark 4.2 (Model correctness)**.** *We discuss when will the two PS models, $\pi^*(\cdot)$ and $\rho^*(\cdot)$, and the two OR models, $\nu^*(\cdot)$ and $\mu^*(\cdot)$ be correctly specified.*

(a) We say $\pi^*(\cdot)$ is correctly specified when $\pi^*(\cdot) = \pi(\cdot)$, which occurs if and only if there exists some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$, such that $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^T\boldsymbol{\gamma}^0)$ holds. Additionally, we have $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^0$.

(b) We say $\rho^*(\cdot)$ is correctly specified when $\rho^*(\cdot) = \rho(\cdot)$, which occurs if and only if there exists some $\boldsymbol{\delta}^0 \in \mathbb{R}^d$, such that $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^T\boldsymbol{\delta}^0)$ holds. Additionally, we have $\boldsymbol{\delta}^* = \boldsymbol{\delta}^0$.

(c) We say $\nu^*(\cdot)$ is correctly specified when $\nu^*(\cdot) = \nu(\cdot)$, which occurs if and only if there exists some $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$, such that $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^T \boldsymbol{\alpha}^0$ holds. Additionally, we have $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0$.

(d) We say $\mu^*(\cdot)$ is correctly specified when $\mu^*(\cdot) = \mu(\cdot)$, which occurs if there exists some $\boldsymbol{\beta}^0 \in \mathbb{R}^d$, such that $\mu(\mathbf{s}_1) = \mathbf{s}_1^T \boldsymbol{\beta}^0$ and, furthermore, either case (b) or (c) holds. Additionally, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$.

Note that, $\boldsymbol{\delta}^*$, (4.13), is constructed based on $\boldsymbol{\gamma}^*$. However, from the case (b), we can see that the correctness of $\rho^*(\cdot)$ does not depend on $\boldsymbol{\gamma}^*$. That is, whether the model $\pi^*(\cdot)$ is correctly specified does not affect the correctness of $\rho^*(\cdot)$. Analogous result for $\nu^*(\cdot)$ can be found in case (c). From the cases (a)-(c), we conclude that the correctness of the three models $\pi^*(\cdot)$, $\rho^*(\cdot)$ and $\nu^*(\cdot)$ has no effects on each other.

However, unlike in cases (a)-(c), $\mu(\cdot)$ is linear that $\mu(\mathbf{s}_1) = \mathbf{s}_1^T \boldsymbol{\beta}^0$ with some $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$ does not *imply $\mu^*(\cdot)$ is correctly specified, since $\boldsymbol{\beta}^*$ defined in* (4.15) *may not reach the true parameter $\boldsymbol{\beta}^0$. As in case (d), we can see that $\mu^*(\cdot)$ is correctly specified if we additionally assume either $\rho^*(\cdot)$ or $\nu^*(\cdot)$ is (or both are) correctly specified. Such a condition is always assumed throughout the chapter (as in in Assumption 4.2), since it is also required in the doubly robust representation* (4.8).

Based on the results in cases (a)-(d), to estimate the counterfactual mean $\theta_{1,1}$, the required Assumption 4.2 is equivalent to the following: (a) either $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^T \boldsymbol{\gamma}^0)$ with some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$ or $\mu(\mathbf{s}_1) = \mathbf{s}_1^T \boldsymbol{\beta}^0$ with some $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$, but not necessarily both; and (b) either $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^T \boldsymbol{\delta}^0)$ with some $\boldsymbol{\delta}^0 \in \mathbb{R}^d$ or $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^T \boldsymbol{\alpha}^0$ with some $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$, but not necessarily both.

Justifications Below are the justifications of the cases (a)-(d) of Remark 1.

By the construction of $\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$, we have

$$E\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1},$$

$$E\left[A_1 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^{d},$$

$$E\left\{A_1 A_2 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\bar{\mathbf{S}}_2\right\} = \mathbf{0} \in \mathbb{R}^{d},$$

$$E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^* + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

For (a), (b) and (c), by the tower rule and the corresponding model $E(A_1|\mathbf{S}_1) = g(\mathbf{S}_1^T\boldsymbol{\gamma}^0)$, $E(A_2 \mid \bar{\mathbf{S}}_2, A_1 = 1) = g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^0)$ and $E\{Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1\} = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0$, we have

$$E\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^0)\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1},$$

$$E\left[A_1 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^0)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^{d},$$

$$E\left\{A_1 A_2 g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0)\bar{\mathbf{S}}_2\right\} = \mathbf{0} \in \mathbb{R}^{d},$$

which implies $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^0$, $\boldsymbol{\delta}^* = \boldsymbol{\delta}^0$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0$.

For (d), if we assume that $E\{Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1\} = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0$ and $E\{Y(1,1) \mid \mathbf{S}_1\} = \mathbf{S}_1^T\boldsymbol{\beta}^0$, we have

$$\begin{aligned}
&E\left[A_1 A_2 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1\right] \\
&\overset{(i)}{=} E\left[E\left[A_1 A_2 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y(1,1) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1 \mid \bar{\mathbf{S}}_2, A_1 = 1\right] \cdot P(A_1 = 1 \mid \bar{\mathbf{S}}_2)\right] \\
&\overset{(ii)}{=} E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\rho(\bar{\mathbf{S}}_2)g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(E[Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1] - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1\right] \\
&\overset{(iii)}{=} \mathbf{0} \in \mathbb{R}^{d_1}
\end{aligned}$$

where (i) holds by the tower rule and $A_1A_2Y = A_1A_2Y(1,1)$; (ii) holds by Assumption 4.1(a); (iii) holds by $E\{Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1\} = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0 = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*$, since $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0$ by (c). Also,

by the tower rule, we have

$$
\begin{aligned}
&E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^0\right\}\mathbf{S}_1\right] \\
&\quad = E\left[E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^0\right\}\mathbf{S}_1 \mid \mathbf{S}_1, A_1 = 1\right]\pi(\mathbf{S}_1)\right] \\
&\quad \overset{(i)}{=} E[\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\pi(\mathbf{S}_1)\left\{E(\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0 \mid \mathbf{S}_1, A_1 = 1) - \mathbf{S}_1^T\boldsymbol{\beta}^0\right\}\mathbf{S}_1] = \mathbf{0} \in \mathbb{R}^{d_1},
\end{aligned}
$$

where (i) holds by $E[E\{Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1\} \mid \mathbf{S}_1, A_1 = 1] = E\{Y(1,1) \mid \mathbf{S}_1\}$. Hence,

$$
E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^0 + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1},
$$

which implies $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$.

If we assume that $E(A_2 \mid \bar{\mathbf{S}}_2, A_1 = 1) = g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^0)$ and $E\{Y(1,1) \mid \mathbf{S}_1\} = \mathbf{S}_1^T\boldsymbol{\beta}^0$, we have

$$
\begin{aligned}
&E\left[A_1 A_2 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*) g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1\right] \\
&\quad \overset{(i)}{=} E\left[E\left[A_1 A_2 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*) g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y(1,1) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1 \mid \bar{\mathbf{S}}_2, A_1 = 1\right] \cdot P(A_1 = 1 \mid \bar{\mathbf{S}}_2)\right] \\
&\quad \overset{(ii)}{=} E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*) E[A_2 \mid \bar{\mathbf{S}}_2, A_1 = 1] g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(E[Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1] - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1\right] \\
&\quad \overset{(iii)}{=} E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)(\nu(\bar{\mathbf{S}}_2) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\mathbf{S}_1\right]
\end{aligned}
$$

where (i) holds by the tower rule and $A_1 A_2 Y = A_1 A_2 Y(1,1)$; (ii) holds by Assumption 4.1(a); (iii) holds by $E\{Y(1,1) \mid \bar{\mathbf{S}}_2, A_1 = 1\} = \nu(\bar{\mathbf{S}}_2)$ and $E(A_2 \mid \bar{\mathbf{S}}_2, A_1 = 1) = g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^0) = g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)$, since $\boldsymbol{\delta}^* = \boldsymbol{\delta}^0$ by (b). Hence, by the tower rule,

$$
\begin{aligned}
&E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^0 + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)\right\}\mathbf{S}_1\right] \\
&\quad = E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\nu(\bar{\mathbf{S}}_2) - \mu(\mathbf{S}_1)\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1},
\end{aligned}
$$

which implies $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$.

Remark 4.3 (Nuisance parameters comparison with [BJZ21])**.** *Here, we compare the nuisance parameters* $\boldsymbol{\gamma}^*$*,* $\boldsymbol{\delta}^*$*,* $\boldsymbol{\alpha}^*$*, and* $\boldsymbol{\beta}^*$ *with the nuisance parameters* $\boldsymbol{\gamma}_1^*$*,* $\boldsymbol{\delta}_1^*$*,* $\boldsymbol{\alpha}_1^*$*, and* $\boldsymbol{\beta}_1^*$ *proposed therein:*

$$\boldsymbol{\gamma}_1^* := \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_1}} E\left[-A_1\mathbf{S}_1^T\boldsymbol{\gamma} + \log\{1+\exp(\mathbf{S}_1^T\boldsymbol{\gamma})\}\right], \tag{4.20}$$

$$\boldsymbol{\delta}_1^* := \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^{d}} E\left(A_1\left[-A_2\bar{\mathbf{S}}_2^T\boldsymbol{\delta} + \log\{1+\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\}\right]\right), \tag{4.21}$$

$$\boldsymbol{\alpha}_1^* := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{d}} E\left\{A_1A_2(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha})^2\right\}, \tag{4.22}$$

$$\boldsymbol{\beta}_1^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d}} E\left\{A_1(\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}_1^* - \mathbf{S}_1^T\boldsymbol{\beta})^2\right\}. \tag{4.23}$$

In general, the nuisance parameters proposed in (4.12)-(4.15) *are not the same as the above nuisance parameters* (4.20)-(4.23)*. However, we have* $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^0 = \boldsymbol{\gamma}_1^*$ *under case (a) of Remark 4.2;* $\boldsymbol{\delta}^* = \boldsymbol{\delta}^0 = \boldsymbol{\delta}_1^*$ *under case (b); and* $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0 = \boldsymbol{\alpha}_1^*$ *under case (c). In addition, any of* $\pi^*(\cdot), \rho^*(\cdot), \nu^*(\cdot)$ *is correctly specified if and only if the corresponding nuisance model of [BJZ21] is also correctly specified.*

Now, we compare $\boldsymbol{\beta}^*$ *with* $\boldsymbol{\beta}_1^*$ *and discuss the conditions required for the correctness of* $\mu^*(\cdot)$*. Let* $\mu(\mathbf{s}_1) = \mathbf{s}_1^T\boldsymbol{\beta}^0$ *holds with some* $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$*. Then, observe that*

$$\boldsymbol{\beta}^0 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} E[A_1\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T\boldsymbol{\beta}\}^2] = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} E[A_1\{\nu(\bar{\mathbf{S}}_2) - \mathbf{S}_1^T\boldsymbol{\beta}\}^2].$$

Hence, $\boldsymbol{\beta}_1^*$*,* (4.23)*, can be seen as an approximate of* $\boldsymbol{\beta}^0$ *where* $\nu(\bar{\mathbf{S}}_2)$ *is approximated by a "regression" representation* $\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}_1^*$*. As discussed in Remark 2 of [BJZ21],* $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}^0$ *requires additional restrictive constraint on* $\boldsymbol{\alpha}_1^*$ *(the OR working model at time 2), which is typically satisfied when case (c) holds.*

On the other hand, when $\mu(\mathbf{s}_1) = \mathbf{s}_1^T\boldsymbol{\beta}^0$*, we also have*

$$\begin{aligned}\boldsymbol{\beta}^0 &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T\boldsymbol{\beta}\}^2] \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\nu(\bar{\mathbf{S}}_2) - \mathbf{S}_1^T\boldsymbol{\beta}\}^2].\end{aligned}$$

Recall the definitions (4.15) *and* (4.19)*,* $\boldsymbol{\beta}^*$ *can be seen as an approximate of* $\boldsymbol{\beta}^0$ *where* $\nu(\bar{\mathbf{S}}_2)$ *is approximated by a doubly robust representation* $\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y(1,1) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)$*. Note that,*

$$\nu(\bar{\mathbf{S}}_2) = E\left\{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* + A_2 g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y(1,1) - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*) \mid \bar{\mathbf{S}}_2, A_1 = 1\right\}, \tag{4.24}$$

as long as either $\rho^*(\cdot)$ *or* $\nu^*(\cdot)$ *is correctly specified, which is a condition always required to ensure the doubly robust representation* (4.8)*. Then, as in case (d) of Remark 4.2,* $\mu^*(\cdot)$ *is correctly specified as long as we further assume that* $\mu(\cdot)$ *is truly linear. Based on the doubly robust representation, we can see that there is no need for any additional constraint on the OR working model at time 2. Our nuisance model* $\mu^*(\cdot)$ *constructed based on* $\boldsymbol{\beta}^*$ *is more likely to be correctly specified.*

4.3 Sequential model doubly robust estimation

Now we propose estimators for the working nuisance models' parameters introduced in Section 4.2.2 and develop a novel sequential model doubly robust estimator for $\theta_{1,1}$; see Section 4.3.1. The asymptotic properties are provided in Section 4.3.2, where we show that $\sqrt{N}$-inference is possible even under model misspecifications.

4.3.1 Construction of the sequential model doubly robust estimator

Targeted bias reducing nuisance estimators We introduce estimators for each of the nuisance parameters defined in (4.12)-(4.15). Let $\mathbb{S}_{\boldsymbol{\gamma}}, \mathbb{S}_{\boldsymbol{\delta}}, \mathbb{S}_{\boldsymbol{\alpha}}, \mathbb{S}_{\boldsymbol{\beta}}$ be disjoint subsets of $\mathbb{S}$ with equal sizes $M \asymp N$, indexed by $\mathcal{I}_{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\delta}}, \mathcal{I}_{\boldsymbol{\alpha}}, \mathcal{I}_{\boldsymbol{\beta}}$, respectively. For any $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\delta}, \boldsymbol{\alpha} \in \mathbb{R}^d$, define

$$\bar{\ell}_1(\boldsymbol{\gamma}) := M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\gamma}}} \ell(\mathbf{W}_i; \boldsymbol{\gamma}), \quad \bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta}) := M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} \ell(\mathbf{W}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}),$$

$$\bar{\ell}_3(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}) := M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} \ell(\mathbf{W}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}), \quad \bar{\ell}_4(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} \ell(\mathbf{W}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

where the loss functions are defined as (4.16)-(4.19). We propose the following *moment targeted nuisance estimators*:

$$\widehat{\boldsymbol{\gamma}} := \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{d_1}} \left\{ \bar{\ell}_1(\boldsymbol{\gamma}) + \lambda_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_1 \right\}, \tag{4.25}$$

$$\widehat{\boldsymbol{\delta}} := \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^{d}} \left\{ \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}) + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1 \right\}, \tag{4.26}$$

$$\widehat{\boldsymbol{\alpha}} := \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{d}} \left\{ \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}) + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \right\}, \tag{4.27}$$

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d_1}} \left\{ \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \right\}. \tag{4.28}$$

Remark 4.4 (Comparison with the nuisance estimators of [BRR19])**.** *Similarly as discussed in Remarks 4.1 and 4.3, $\widehat{\boldsymbol{\beta}}$ itself is also constructed based on a doubly robust representation. Another doubly robust estimator for the first OR model, $\mu(\cdot)$, has been studied by [BRR19], where low-dimensional covariates are considered. They used the usual maximum likelihood estimators for the PS models. They did not consider the doubly robust representation for $\nu(\bar{\mathbf{S}}_2)$, (4.24). Instead, they achieved the double robustness for the estimation of $\mu(\cdot)$ using*

Algorithm 3 Sequential model doubly robust counterfactual mean estimator

Require: Observations $\mathbb{S} = (\mathbf{W}_i)_{i=1}^N = (Y_i, A_{1i}, A_{2i}, \mathbf{S}_{1i}, \mathbf{S}_{2i})_{i=1}^N$ and the treatment path of interest $(a_1, a_2) = (1, 1)$.

1: Split the sample $\mathbb{S}$ into $\mathbb{K} \geq 2$ folds that $\mathbb{S} = \cup_{k=1}^{\mathbb{K}} \mathbb{S}_k$, indexed by $(\mathcal{I}_k)_{k=1}^{\mathbb{K}}$ and with equal sizes $n := N/\mathbb{K}$.

2: **for** $k = 1, 2, ..., \mathbb{K}$ **do**

3: Define $\mathcal{I}_{-k} := \mathcal{I} \backslash \mathcal{I}_k$ and $\mathbb{S}_{-k} := (\mathbf{W}_i)_{i \in \mathcal{I}_{-k}}$. Let $(\mathbb{S}_{\boldsymbol{\gamma}}, \mathbb{S}_{\boldsymbol{\delta}}, \mathbb{S}_{\boldsymbol{\alpha}}, \mathbb{S}_{\boldsymbol{\beta}})$ be a disjoint partition of $\mathbb{S}_{-k}$ with equal sizes $M = |\mathcal{I}_{-k}|/4 = N(\mathbb{K} - 1)/(4\mathbb{K})$.

4: Construct $\widehat{\boldsymbol{\gamma}}_{-k}$, (4.25), using the sub-sample $\mathbb{S}_{\boldsymbol{\gamma}}$. ▷ *Propensity for time one*

5: Construct $\widehat{\boldsymbol{\delta}}_{-k}$, (4.26), using $\widehat{\boldsymbol{\gamma}}_{-k}$ and the sub-sample $\mathbb{S}_{\boldsymbol{\delta}}$. ▷ *Propensity for time two*

6: Construct $\widehat{\boldsymbol{\alpha}}_{-k}$, (4.27), using $\widehat{\boldsymbol{\gamma}}_{-k}$, $\widehat{\boldsymbol{\delta}}_{-k}$, and the sub-sample $\mathbb{S}_{\boldsymbol{\alpha}}$. ▷ *Outcome for time two*

7: Construct $\widehat{\boldsymbol{\beta}}_{-k}$, (4.28), using $\widehat{\boldsymbol{\gamma}}_{-k}$, $\widehat{\boldsymbol{\delta}}_{-k}$, $\widehat{\boldsymbol{\alpha}}_{-k}$, and the sub-sample $\mathbb{S}_{\boldsymbol{\beta}}$. ▷ *Outcome for time one*

8: **end for**

9: **return** The *sequential model doubly robust counterfactual mean estimator* is proposed as

$$\widehat{\theta}_{1,1} = N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}), \tag{4.29}$$

where $\psi(\cdot;\cdot)$ is defined as (4.11) and let $\widehat{\boldsymbol{\eta}}_{-k} := (\widehat{\boldsymbol{\gamma}}_{-k}^T, \widehat{\boldsymbol{\delta}}_{-k}^T, \widehat{\boldsymbol{\alpha}}_{-k}^T, \widehat{\boldsymbol{\beta}}_{-k}^T)^T$ for each $k \leq \mathbb{K}$.

some weighted least squares estimators for the OR models, where the weights are different from ours in (4.18) *and* (4.19)*. However, they only considered low-dimensional settings, and they required stronger model correctness conditions to achieve CAN for the counterfactual*

mean estimator than us, as discussed in Section 4.1 and Remark 4.6.

Bias-reduced doubly robust counterfactual mean estimator Based on the moment targeted nuisance estimators, we propose a *sequential model doubly robust estimator* for the counterfactual mean $\theta_{1,1} = E\{Y(1,1)\}$. We consider the doubly robust score (4.11) and cross-fitted versions of the moment targeted nuisance estimators; see details in Algorithm 3.

4.3.2 Inference under model misspecification

We study the asymptotic properties of the proposed counterfactual mean estimator $\widehat{\theta}_{1,1}$. The results in this section are based on the nuisance estimators' theoretical properties studied later in Section 4.4.

We first make the following mild assumption on the population nuisance parameters' sparsity levels:

Assumption 4.3 (Sparse signals)**.** *Let the sparsity levels of the population nuisance parameters $\boldsymbol{\gamma}^*$, $\boldsymbol{\delta}^*$, $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ satisfy*

$$s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\beta}} = o\left(\frac{N}{\log d_1}\right), \quad s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = o\left(\frac{N}{\log d}\right), \quad s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O\left(\frac{N}{\log d_1 \log d}\right).$$

The sparsity conditions of the type $s = o\left(N/\log d\right)$ is very common in the high-dimensional Statistics literature. Here we also need a slightly stonger additional condition $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$.

Remark 4.5 (Bounded covariates)**.** *Similarly as later discussed in Remark 4.10, if we further assume that $\|\bar{\mathbf{S}}_2\|_\infty < C$, which is a condition assumed in, e.g., [BWZ19], [Tan20a], and*

[SRR19], then, the condition $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$ *in Assumption 4.3 is no longer needed.*

We establish the following asymptotic results for the proposed sequential model doubly robust counterfactual mean estimator under possible model misspecification.

Theorem 4.1 (Inference under model misspecifications)**.** *Let Assumptions 4.2-4.3 hold. Choose some* $\lambda_{\boldsymbol{\gamma}}, \lambda_{\boldsymbol{\delta}}, \lambda_{\boldsymbol{\alpha}}, \lambda_{\boldsymbol{\beta}} > 0$ *with* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$, $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\frac{\log d}{N}}$, $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\frac{\log d_1}{N}}$. *Let the following product sparsity conditions hold*

$$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o\left(\frac{N}{(\log d_1)^2}\right), \quad s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right). \tag{4.30}$$

We assume the following additional conditions if model misspecification occurs:

$$\text{if } \rho(\cdot) \neq \rho^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{\log d_1 \log d}\right); \tag{4.31}$$

$$\text{if } \nu(\cdot) \neq \nu^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} = o\left(\frac{N}{\log d_1 \log d}\right), \ s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right); \tag{4.32}$$

$$\text{if } \mu(\cdot) \neq \mu^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d_1}\right), \ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{\log d_1 \log d}\right). \tag{4.33}$$

Then, as $N \to \infty$,

$$\sigma^{-1} N^{-1/2} (\widehat{\theta}_{1,1} - \theta_{1,1}) \ \to \ \mathcal{N}(0,1)$$

in distribution, where

$$\sigma^2 \ := \ E\left\{\psi(\mathbf{W}; \boldsymbol{\eta}^*) - \theta_{1,1}\right\}^2. \tag{4.34}$$

In addition, define

$$\widehat{\sigma}^2 \ := \ N^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \left\{\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \widehat{\theta}_{1,1}\right\}^2. \tag{4.35}$$

Then, as $N \to \infty$, $\widehat{\sigma}^2 = \sigma^2\{1 + o_p(1)\}$.

Remark 4.6 (Sequential model double robustness)**.** *In Theorem 4.1, we demonstrate the "sequential model double robustness" of our proposed estimator: $\sqrt{N}$-inference is provided as long as at least one nuisance model is correctly specified at each time spot; see Assumption 4.2. In the sense of model robustness, our results outperform all the existing results, even containing those who only considered low-dimensional covariates.*

In the presence of high-dimensional covariates, the regularized nuisance estimators are known to be biased and with a consistency rate slower than $N^{-1/2}$. Hence, valid inference results are more difficult to obtain than in low dimensions. We are the first to establish statistical inference for the dynamic counterfactual mean (and hence dynamic treatment efect) when any model misspecification occurs.

Among the literature who considered low-dimensional covariates, the recent work of [BRR19] provided the best results so far on model robustness. Their proposed estimator is CAN when either (4.1), (4.2) *or* (4.3) *holds. However, as in Assumption 4.2, we allow an additional case* (4.4)*, which is not considered in [BRR19].*

The model double robustness for average treatment effect estimation without dynamic settings has been studied recently by [SRR19, Tan20a, AV21]. They provided valid inference as long as one of the nuisance models is correctly specified, and their results can be seen as special cases of ours where only one exposure time is considered.

Remark 4.7 (Required sparsity conditions under model misspecification)**.** *Here we discuss the sparsity conditions required in Theorem 4.1 for $\sqrt{N}$-inference. We can see that the correctness of $\pi^*(\cdot)$ does not affect the sparsity conditions; in addition, the more model misspecification occurs among $\rho^*(\cdot)$, $\nu^*(\cdot)$, and $\mu^*(\cdot)$, the more sparsity conditions we require.*

When $\rho^*(\cdot)$, $\nu^*(\cdot)$, and $\mu^*(\cdot)$ are all correctly specified, we require Assumption 4.3 and (4.30)*. Whenever a model at time $t \in \{1, 2\}$ is misspecified, we require a product condition between 1) the sparsity level of the other (correctly specified) model at the same time t and 2) the summation of sparsity levels corresponds to all the nuisance estimators that such a misspecified estimator is constracted based on. Recall that we constrct the nuisance estimators sequentally in the order: $\widehat{\boldsymbol{\gamma}} \to \widehat{\boldsymbol{\delta}} \to \widehat{\boldsymbol{\alpha}} \to \widehat{\boldsymbol{\beta}}$. For instance, when $\mu^*(\cdot)$ is misspecified, as shown in* (4.33)*, we need a product condition between 1) $s_{\boldsymbol{\gamma}}$ and 2) $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}}$. Moreover, consider the cases that the OR model at time t is misspecified. Since the OR estimators are constructed after the PS estimators, based on the pattern we discussed above, we always require an ultra-sparse PS parameter at time t. More details for the required sparsity conditions are listed in Table 4.1.*

In addition, consider the degenerated case that only the first exposure time is involved. Then, we require $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o(N/(\log d_1)^2)$ when $\nu(\cdot) = \nu^*(\cdot)$; or, $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o(N/(\log d_1)^2)$ and $s_{\boldsymbol{\gamma}} = o(\sqrt{N}/\log d_1)$ when $\nu(\cdot) \neq \nu^*(\cdot)$. Such conditions coincide with [SRR19] and are weaker than the sparsity conditions in [Tan20a, AV21], where both $s_{\boldsymbol{\gamma}} = o(\sqrt{N}/\log d_1)$ and $s_{\boldsymbol{\beta}} = o(\sqrt{N}/\log d_1)$ are required since cross-fitting was not performed therein.

If all the nuisance models are correctly specified, we have the following result:

Theorem 4.2 (Inference under correctly specified models)**.** *Suppose all the nuisance models are correctly specified. Let Assumptions 4.1-4.3 and the product sparsity conditions* (4.30) *hold. Choose some $\lambda_{\boldsymbol{\gamma}}, \lambda_{\boldsymbol{\delta}}, \lambda_{\boldsymbol{\alpha}}, \lambda_{\boldsymbol{\beta}} > 0$ with $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$, $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\frac{\log d}{N}}$, $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\frac{\log d_1}{N}}$. Then, as $N \to \infty$,*

$$\sigma^{-1} N^{-1/2} (\widehat{\theta}_{1,1} - \theta_{1,1}) \ \to \ \mathcal{N}(0, 1)$$

Table 4.1: Let $\|\bar{\mathbf{S}}_2\|_\infty < C$, $d_1 \asymp d$, and $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\beta}} = o(N/\log d)$. Sparsity conditions required for the sequential model doubly robust counterfactual mean estimator to be consistent and asymptotically normal

Model correctness				Required sparsity conditions
$\pi^*(\cdot)$	$\rho^*(\cdot)$	$\nu^*(\cdot)$	$\mu^*(\cdot)$	
✓	✓	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✓	✗	$s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✓	$s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✓	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✗	$s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✗	✗	✓	$s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✓	✗	✓	$s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✗	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$

in distribution, where σ^2 *is defined in* (4.34)*. In addition, define* $\widehat{\sigma}^2$ *as in* (4.35)*. Then, as* $N \to \infty$, $\widehat{\sigma}^2 = \sigma^2\{1 + o_p(1)\}$.

Remark 4.8 (Sequential rate double robustness)**.** *Consider the case that all the nuisance models are correctly specified. Then, as shown in Theorem 4.2,* $\sqrt{N}$*-inference requires two product sparsity conditions,* (4.30)*. We name such a property as "sequential rate double robustness". Note that, we only require product sparsity conditions between the nuisance paramters' sparsity levels at each time spot. Such conditions are weaker than [BHL20, BJZ21], where they require an additional product sparsity condition* $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o(\frac{N}{\log d_1 \log d})$*. For instance, when* $s_{\boldsymbol{\gamma}} \asymp N^{0.8}$*,* $s_{\boldsymbol{\delta}} \asymp N^{0.1}$*,* $s_{\boldsymbol{\alpha}} \asymp N^{0.8}$*,* $s_{\boldsymbol{\beta}} \asymp N^{0.1}$*, our Assumption 4.3 and* (4.30) *holds but* $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o(\frac{N}{\log d_1 \log d})$ *fails (omitting the logarithmic terms). We are able to provide* $\sqrt{N}$*-inference under weaker sparsity conditions since we achieve better consistency*

result for the estimation of $\boldsymbol{\beta}^*$; see Remark 4.12.

Besides, similarly as discussed in Remark 4.7, consider the degenerated case that only the first exposure time is involved. Then, we require $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o(N/(\log d_1)^2)$. Such a condition coincides with the "rate double robustness" of [CCD+18, SRR19] and is weaker than the sparsity conditions in [Far15, Tan20a, AV21] since cross-fitting was not performed therein. In addition, based on a special type of cross-fitting, [BWZ19] imposed either 1) $s_{\boldsymbol{\beta}} = o(\sqrt{N}/\log d_1)$ and $s_{\boldsymbol{\gamma}} = o(N/\log d_1)$ or 2) $s_{\boldsymbol{\beta}} = o(N^{3/4}/\log d_1)$ and $s_{\boldsymbol{\gamma}} = o(\sqrt{N}/\log d_1)$. Such a condition is different from (not stronger nor weaker than) the "rate double robustness" condition $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o(N/(\log d_1)^2)$.

4.4 Theoretical results for the nuisance estimators

We develop theoretical properties of the proposed moment targeted nuisance estimators. In Section 4.4.1, we demonstrate the consistency of the nuisance estimators allowing all the models to be misspecified. In Section 4.4.2, we provide faster consistency rates for the nuisance estimators assuming correctly specified models.

4.4.1 Results with misspecified models

Define $s_{\boldsymbol{\gamma}} := \|\boldsymbol{\gamma}^*\|_0$, $s_{\boldsymbol{\delta}} := \|\boldsymbol{\delta}^*\|_0$, $s_{\boldsymbol{\alpha}} := \|\boldsymbol{\alpha}^*\|_0$, and $s_{\boldsymbol{\beta}} := \|\boldsymbol{\beta}^*\|_0$ as the sparsity levels of the population nuisance parameters. The following assumption imposes some standard moment conditions:

Assumption 4.4 (Sub-Gaussianity)**.** *Let $\bar{\mathbf{S}}_2$ be a sub-Gaussian random vector, i.e., for all $\mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{v}^T \bar{\mathbf{S}}_2\|_{\psi_2} \leq \sigma_{\mathbf{S}} \|\mathbf{v}\|_2$. Define $\varepsilon := Y(1,1) - \bar{\mathbf{S}}_2^T \boldsymbol{\alpha}^*$ and $\zeta := \bar{\mathbf{S}}_2^T \boldsymbol{\alpha}^* - \mathbf{S}_1^T \boldsymbol{\beta}^*$. Let*

ε and ζ be sub-Gaussian random variables, i.e., $\|\varepsilon\|_{\psi_2} \le \sigma_\varepsilon$ and $\|\zeta\|_{\psi_2} \le \sigma_\zeta$. In addition, let $Var\{Y(1,1)\} > c_Y$ and the smallest eigenvalue of $E(A_1 \bar{\mathbf{S}}_2 \bar{\mathbf{S}}_2^T)$ is bounded bellow by $c_{\min}$. Here, $\sigma_{\mathbf{S}}, \sigma_\varepsilon, \sigma_\zeta, c_Y, c_{\min}$ are some positive constants.

Furthermore, for any $\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Delta} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha}, \boldsymbol{\delta} \in \mathbb{R}^d$, define

$$\delta\bar{\ell}_1(\boldsymbol{\gamma}, \boldsymbol{\Delta}) := \bar{\ell}_1(\boldsymbol{\gamma} + \boldsymbol{\Delta}) - \bar{\ell}_1(\boldsymbol{\gamma}) - \nabla_{\boldsymbol{\gamma}} \bar{\ell}_1(\boldsymbol{\gamma})^T \boldsymbol{\Delta}, \tag{4.36}$$

$$\delta\bar{\ell}_4(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta}) := \bar{\ell}_4(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta} + \boldsymbol{\Delta}) - \bar{\ell}_4(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\beta})^T \boldsymbol{\Delta}. \tag{4.37}$$

Similarly, for any $\boldsymbol{\gamma} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\Delta} \in \mathbb{R}^d$, define

$$\delta\bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\Delta}) := \bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta} + \boldsymbol{\Delta}) - \bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta}) - \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta})^T \boldsymbol{\Delta}, \tag{4.38}$$

$$\delta\bar{\ell}_3(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Delta}) := \bar{\ell}_3(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha} + \boldsymbol{\Delta}) - \bar{\ell}_3(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}) - \nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha})^T \boldsymbol{\Delta}. \tag{4.39}$$

We begin by demonstrating the following restricted strong convexity (RSC) conditions. Note that, the nuisance estimators are constructed based on different samples, and the probability measures in (4.41)-(4.43) are also different.

Lemma 4.1. *Let Assumptions 4.1 and 4.4 hold. Define $f_{M,d_1}(\boldsymbol{\Delta}) := \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d_1}{M} \|\boldsymbol{\Delta}\|_1^2$ for any $\boldsymbol{\Delta} \in \mathbb{R}^{d_1}$ and $f_{M,d}(\boldsymbol{\Delta}) := \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d}{M} \|\boldsymbol{\Delta}\|_1^2$ for any $\boldsymbol{\Delta} \in \mathbb{R}^d$. Then, with some constants $\kappa_1, \kappa_2, c_1, c_2 > 0$ and recall that $M \asymp N$, we have*

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}} \left(\delta\bar{\ell}_1(\boldsymbol{\gamma}^*, \boldsymbol{\Delta}) \ge f_{M,d_1}(\boldsymbol{\Delta}), \ \ \forall \|\boldsymbol{\Delta}\|_2 \le 1 \right) \ge 1 - c_1 \exp(-c_2 M). \tag{4.40}$$

Further, let $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1$. Then,

$$P_{\mathbb{S}_{\boldsymbol{\delta}}} \left(\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) \ge f_{M,d}(\boldsymbol{\Delta}), \ \ \forall \|\boldsymbol{\Delta}\|_2 \le 1 \right) \ge 1 - c_1 \exp(-c_2 M), \tag{4.41}$$

$$P_{\mathbb{S}_{\boldsymbol{\beta}}} \left(\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \ge f_{M,d_1}(\boldsymbol{\Delta}), \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^{d_1} \right) \ge 1 - c_1 \exp(-c_2 M). \tag{4.42}$$

Note that, in (4.41)*, we only consider the randomness in* $\mathbb{S}_{\boldsymbol{\delta}}$*, and* $\widehat{\boldsymbol{\gamma}}$ *is treated as fixed (or conditional on). Similarly, in* (4.43)*,* $\widehat{\boldsymbol{\gamma}}$*,* $\widehat{\boldsymbol{\delta}}$*, and* $\widehat{\boldsymbol{\alpha}}$ *are all treated as fixed.*

Moreover, let $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1$*. Then,*

$$P_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \geq f_{M,d}(\boldsymbol{\Delta}),\ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right) \geq 1 - c_1 \exp(-c_2 M), \tag{4.43}$$

where $\widehat{\boldsymbol{\gamma}}$ *and* $\widehat{\boldsymbol{\delta}}$ *are treated as fixed.*

Additionally, we upper bound the gradients of the loss functions evaluated at the target population parameter values. By construction, the gradients are averages of i.i.d. random vectors with zero means even under model misspecifications. Hence, we can use the union bound techniques to control the infinite norms by the usual rate $O_p(\sqrt{\log d/M})$ or $O_p(\sqrt{\log d_1/M})$.

Lemma 4.2. *Let Assumption 4.4 holds. Let* $\sigma_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\delta}}, \sigma_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\beta}} > 0$ *be some constants and recall that* $M \asymp N$*. Then, for any* $t > 0$*,*

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left\|\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)\right\|_\infty \leq \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d}{M}}\right) \geq 1 - 2\exp(-t).$$

Further, let the Assumption 4.1 holds. Then, for any $t > 0$*,*

$$P_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty \leq \sigma_{\boldsymbol{\delta}}\sqrt{\frac{t + \log d}{M}}\right) \geq 1 - 2\exp(-t),$$

$$P_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\|_\infty \leq \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t + \log d}{M}} + \frac{t + \log d}{M}\right)\right) \geq 1 - 2\exp(-t),$$

$$P_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty \leq \sigma_{\boldsymbol{\beta}}\left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M}\right)\right) \geq 1 - 2\exp(-t).$$

We then demonstrate the asymptotic results for the moment targeted nuisance estimators when all the nuisance models are possibly misspecified. Note that the estimators $\widehat{\boldsymbol{\delta}}$,

$\widehat{\boldsymbol{\alpha}}$, and $\widehat{\boldsymbol{\beta}}$ are constructed based on some previous nuisance estimators, we carefully control the errors originated from the previous steps' estimation. Among the results in Theorem 4.3, part (b) is the most challenging to show. This is because $\widehat{\boldsymbol{\delta}}$ is constructed based on $\widehat{\boldsymbol{\gamma}}$ and the loss function $\bar{\ell}_2$ is not constructed based on a (weighted) square loss. Instead of considering the usual cone set $\mathbb{C}(S,k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq k\|\boldsymbol{\Delta}_S\|_1\}$, we show that $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$ belongs to another cone set $\widetilde{C}(s,k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \leq k\sqrt{s}\|\boldsymbol{\Delta}\|_2\}$ with high probability and some constant $k > 0$, as well as some $s > 0$ depending on both $s_{\boldsymbol{\gamma}}$ and $s_{\boldsymbol{\delta}}$; see details in Lemma 4.12.

Theorem 4.3. *Let Assumptions 4.1 and 4.4 hold.*

(a) *Let $s_{\boldsymbol{\gamma}} = o(\frac{N}{\log d_1})$. Choose some $\lambda_{\boldsymbol{\gamma}} > 0$ with $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$. Then, as $N \to \infty$,*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right), \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{\log d_1}{N}}\right).$$

(b) *In addition to part (a), let $s_{\boldsymbol{\delta}} = o(\frac{N}{\log d})$. Choose some $\lambda_{\boldsymbol{\delta}} > 0$ with $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$. Then, as $N \to \infty$,*

$$\begin{aligned}\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 =& O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),\\ \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_1 =& O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{(\log d_1)^2}{N \log d}} + s_{\boldsymbol{\delta}}\sqrt{\frac{\log d}{N}}\right).\end{aligned}$$

(c) *In addition to parts (a) and (b), let $s_{\boldsymbol{\alpha}} = o(\frac{N}{\log d})$. Choose some $\lambda_{\boldsymbol{\alpha}} > 0$ with $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\frac{\log d}{N}}$. Then, as $N \to \infty$,*

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 =& O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}}\right),\\ \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 =& O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{(\log d_1)^2}{N \log d}} + s_{\boldsymbol{\delta}}\sqrt{\frac{\log d}{N}} + s_{\boldsymbol{\alpha}}\sqrt{\frac{\log d}{N}}\right).\end{aligned}$$

(d) *In addition to parts (a), (b), and (c), let* $s_{\boldsymbol{\beta}} = o(\frac{N}{\log d_1})$. *Choose some* $\lambda_{\boldsymbol{\beta}} > 0$ *with* $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\frac{\log d_1}{N}}$. *Then, as* $N \to \infty$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d + s_{\boldsymbol{\beta}} \log d_1}{N}}\right),$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{\log d_1}{N}} + s_{\boldsymbol{\delta}}\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\alpha}}\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

Remark 4.9 (Consistency of the nuisance estimators under model misspecification)**.** *By Theorem 4.3, we can see that the nuisance estimators are consistent even when the models are misspecified. Since the nuisance estimators* $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ *are constructed sequentially that the later estimators depend on all the previous estimators, the estimation errors of the nuisance parameters are cumulative. That is, the consistency rate of a nuisance estimator depends on the sparsity levels of all the nuisance parameters up to the current one.*

4.4.2 Results with correctly specified models

In Theorem 4.3, we have provided consistency results under model misspecifications. In fact, if we have additional information that some of the nuisance models are correctly specified, we are able to achieve better consistency results than Theorem 4.3.

Assuming correctly specified models, we control the gradients in Lemma 4.3 below (approximately) by the usual rate $O_p(\sqrt{\log d/N})$ or $O_p(\sqrt{\log d_1/N})$. Note that, different from Lemma 4.2, we can upper bound the gradients involving the estimated nuisance parameters. For instance, in part (a) of Lemma 4.3 below, we can control $\|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\|_\infty$ and the estimation error of $\widehat{\boldsymbol{\gamma}}$ is ignorable as long as $s_{\boldsymbol{\gamma}} = O_p(N/(\log d_1 \log d))$.

Lemma 4.3. ***(a)*** *Let* $\rho(\cdot) = \rho^*(\cdot)$. *Let the assumptions in part (a) of Theorem 4.3 hold.*

Then, as $N \to \infty$,

$$\left\|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 \log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

(b) *Let* $\nu(\cdot) = \nu^*(\cdot)$. *Let the assumptions in part (b) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d) \log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

(c) *Let* $\nu(\cdot) = \nu^*(\cdot)$ *and* $\mu(\cdot) = \mu^*(\cdot)$. *Let the assumptions in part (c) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d) \log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

(d) *Let* $\rho(\cdot) = \rho^*(\cdot)$ *and* $\mu(\cdot) = \mu^*(\cdot)$. *Let the assumptions in part (c) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\|_\infty \\ &\quad = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d) \log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).\end{aligned}$$

(e) *Let* $\rho(\cdot) = \rho^*(\cdot)$, $\nu(\cdot) = \nu^*(\cdot)$, *and* $\mu(\cdot) = \mu^*(\cdot)$. *Let the assumptions in part (c) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{s_{\boldsymbol{\gamma}} (\log d_1)^2}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

Then, with additional assumptions on the model correctness, we provide better consistency results for the moment targeted nuisance estimators than in Theorem 4.3.

Theorem 4.4. ***(a)*** *Let* $\rho(\cdot) = \rho^*(\cdot)$. *Let the assumptions in part (b) of Theorem 4.3 hold. Additionally, let* $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$. *Then, as* $N \to \infty$,

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right), \quad \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_1 = O_p\left(s_{\boldsymbol{\delta}}\sqrt{\frac{\log d}{N}}\right).$$

(b) *Let $\nu(\cdot) = \nu^*(\cdot)$. Let the assumptions in part (c) of Theorem 4.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ and $s_{\boldsymbol{\delta}} = O(\frac{N}{(\log d)^2})$. Then, as $N \to \infty$,*

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\right), \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 = O_p\left(s_{\boldsymbol{\alpha}}\sqrt{\frac{\log d}{N}}\right).$$

(c) *Let $\nu(\cdot) = \nu^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 4.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ and $s_{\boldsymbol{\delta}} = O(\frac{N}{(\log d)^2})$. Then, as $N \to \infty$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d + s_{\boldsymbol{\beta}} \log d_1}{N}}\right),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(s_\gamma\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

(d) *Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 4.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(\frac{N}{\log d_1 \log d})$. Then, as $N \to \infty$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\beta}} \log d_1}{N}}\right),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(s_{\boldsymbol{\delta}}\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

(e) *Let $\rho(\cdot) = \rho^*(\cdot)$, $\nu(\cdot) = \nu^*(\cdot)$, and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 4.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ and $s_{\boldsymbol{\delta}} = O(\frac{N}{(\log d)^2})$. Then, as $N \to \infty$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\frac{\sqrt{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}}\right),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(\frac{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} \log d}{N}\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

Further, let $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o(\frac{N}{(\log d)^2})$. Then, as $N \to \infty$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}}\right).$$

Remark 4.10 (Bounded covariates)**.** *If we further assume that* $\|\bar{\mathbf{S}}_2\|_\infty < C < \infty$*, then, the following conditions can be omitted:* $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ *in case (a);* $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ *and* $s_{\boldsymbol{\delta}} = O(\frac{N}{(\log d)^2})$ *in cases (b), (c), and (e);* $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(\frac{N}{\log d_1 \log d})$ *in case (d).*

Remark 4.11 (Consistency of the nuisance estimators with correctly specified models)**.** *Here we summarize the* ℓ_2 *convergence rates of the nuisance estimators when the corresponding models are correctly specified:*

- *For the "first" nuisance estimator* $\widehat{\boldsymbol{\gamma}}$*, as shown in case (a) of Theorem 4.3, we have* $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{s_{\boldsymbol{\gamma}} \log d_1 / N})$ *no matter whether* $\pi^*(\cdot)$ *is correctly specified or not.*

- *When* $\rho^*(\cdot)$ *is correctly specified, the convergence rate of* $\widehat{\boldsymbol{\delta}}$ *depends only on* $s_{\boldsymbol{\delta}}$ *as shown in cases (a) of Theorem 4.4. This is different from part (b) of Theorem 4.3 and Remark 4.9, where* $\rho^*(\cdot)$ *is possibly misspecified.*

- *When* $\nu^*(\cdot)$ *is correctly specified, the convergence rate of* $\widehat{\boldsymbol{\alpha}}$ *depends only on* $s_{\boldsymbol{\alpha}}$ *as shown in cases (b) of Theorem 4.4. This is different from part (c) of Theorem 4.3 and Remark 4.9, where* $\nu^*(\cdot)$ *is possibly misspecified.*

- *As for the convergence rate of* $\widehat{\boldsymbol{\beta}}$*, apart from* $\mu^*(\cdot)$*, it also depends on the correctness of* $\rho^*(\cdot)$ *and* $\nu^*(\cdot)$*. If only one of* $\rho^*(\cdot)$ *and* $\nu^*(\cdot)$ *is correctly specified, as shown in cases (c) and (d), the consistency rate of* $\widehat{\boldsymbol{\beta}}$ *depends on* $s_{\boldsymbol{\beta}}$ *and also the nuisance paramter's sparsity level of the correct model among* $\rho^*(\cdot)$ *and* $\nu^*(\cdot)$*. If both of* $\rho^*(\cdot)$ *and* $\nu^*(\cdot)$ *are correctly specified, as in case (e), the consistency rate of* $\widehat{\boldsymbol{\beta}}$ *depends on* $s_{\boldsymbol{\beta}}$ *and a product sparsity* $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}$*. When a product sparsity condition,* $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o(\frac{N}{(\log d)^2})$*, is assumed as in* (4.30) *of Theorem 4.1, the product sparsity* $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}$ *can also be omitted.*

Remark 4.12 (Comparison of nuisance estimators' consistency rates with [BJZ21])**.** *We compare the consistency rates of* $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ *with some* $\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\delta}}_1, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1$*, the nuisance parameters proposed therein. Note that, it is only reasonable to do such comparisons when the target nuisance parameters are the same. As discussed in Remarks 4.2 and 4.3, we have (a)* $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_1^*$ *when* $\pi(\mathbf{S}_1) = g(\mathbf{S}_1^T\boldsymbol{\gamma}^0)$*; (b)* $\boldsymbol{\delta}^* = \boldsymbol{\delta}_1^*$ *when* $\rho(\mathbf{S}_1) = g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^0)$*; (c)* $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_1^*$ *when* $\nu(\bar{\mathbf{S}}_2) = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0$*; (d)* $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^*$ *when* $\nu(\bar{\mathbf{S}}_2) = \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^0$ *and* $\mu(\mathbf{S}_1) = \mathbf{S}_1^T\boldsymbol{\beta}^0$*. Under each case of (a)-(d), we can see that both our proposed nuisance estimator and the estimator proposed by [BJZ21] converges to the true (and the same) nuisance parameter.*

By Theorems 4.3 and 4.4, we can see that $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}$ *reach the same consistency rates as* $\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\delta}}_1, \widehat{\boldsymbol{\alpha}}_1$ *shown in [BJZ21], under cases (a), (b), (c), respectively. As for the consistency rates of* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\beta}}_1$*, we can see that they also have the same consistency rate under case (d). However, if we further assume case (b) also happens simultaneously, i.e.,* $\rho^*(\cdot)$*,* $\nu^*(\cdot)$*, and* $\mu^*(\cdot)$ *are all correctly specified, we have* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}})$ *as shown in Theorem 4.4. Whereas,* $\widehat{\boldsymbol{\beta}}_1$ *only satisfies* $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}})$*. Since* $s_{\boldsymbol{\delta}} = o(N/\log d)$*, we can see that* $\widehat{\boldsymbol{\beta}}$ *outperforms* $\widehat{\boldsymbol{\beta}}_1$ *in the sence of* ℓ_2*-consistency rates. Because of the better consistency rate we obtain, we require weaker sparsity conditions to establish statistical inference for* $\theta_{1,1}$ *as discussed in Remark 4.8.*

4.5 Proof of the main results

4.5.1 Auxiliary lemmas

The following Lemmas will be useful in the proofs.

Lemma 4.4 (Lemma S4.1 of [ZCB21]). *Let $(X_N)_{N\geq 1}$ and $(Y_N)_{N\geq 1}$ be sequences of random variables in $\mathbb{R}$. If $E(|X_N|^r|Y_N) = O_p(1)$ for any $r \geq 1$, then $X_N = O_p(1)$.*

Lemma 4.5 (Lemma S.2 of [BJZ21]). *Let Assumptions 4.1 and 4.4 hold. Then, the smallest eigenvalues of $E(A_1\mathbf{S}_1\mathbf{S}_1^T)$ and $E(A_1A_2\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^T)$ are both lower bounded by some constant $c'_{\min} > 0$. Additionally, $\|\mathbf{v}^T\mathbf{S}_1\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$, $\|A_1\mathbf{v}^T\mathbf{S}_1\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^{d_1}$ and $\|A_1A_2\mathbf{v}^T\bar{\mathbf{S}}_2\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^d$, with some constant $\sigma'_{\mathbf{S}} > 0$.*

Lemma 4.6 (Lemma D.1 (iv) and (vi) of [CLCL19]). *Let $X \in \mathbb{R}$ be a random variable. If $\|X\|_{\psi_2} \leq \sigma$, then $E(X) \leq \sigma\sqrt{\pi}$ and $E(|X|^m) \leq 2\sigma^m(m/2)^{m/2}$ $\forall$ $m \geq 2$. Let $\{X_i\}_{i=1}^n (n > 1)$ be random variables (possibly dependent) with $\max_{1\leq i\leq n}\|X_i\|_{\psi_2} \leq \sigma$. Then $\|\max_{1\leq i\leq n}|X_i|\|_{\psi_2} \leq \sigma(\log n + 2)^{1/2}$.*

Lemma 4.7 (Corollary 2.3 of [DVDGVW10]). *Let $\{X_i\}_{i=1}^n (n > 1)$ be identically distributed, then*

$$E\left[\|n^{-1}\sum_{i=1}^n X_i\|_\infty^2\right] \leq n^{-1}(2e\log d - e)E\left[\|X_i\|_\infty^2\right]$$

Lemma 4.8. *Suppose that $\mathbb{S}' = (\mathbf{U}_i)_{i\in\mathcal{J}}$ are independent and identically distributed (i.i.d.) sub-Gaussian random vectors, i.e., $\|\mathbf{a}^T\mathbf{U}\|_{\psi_2} \leq \sigma_{\mathbf{U}}\|\mathbf{a}\|_2$ for all $\mathbf{a} \in \mathbb{R}^d$ with some constant $\sigma_{\mathbf{U}} > 0$. Additionally, suppose the smallest eigenvalue of $E(\mathbf{U}\mathbf{U}^T)$ is bounded bellow by some constant $\lambda_{\mathbf{U}} > 0$. Let $M = |\mathcal{J}|$. For any continuous function $\phi : \mathbb{R} \to (0, \infty)$, $v \in [0, 1]$, and $\boldsymbol{\eta} \in \mathbb{R}^d$ satisfying $E\{|\mathbf{U}^T\boldsymbol{\eta}|^c\} < C$ with some constants $c, C > 0$, there exists constants $\kappa_1, \kappa_2, c_1, c_2 > 0$, such that*

$$P_{\mathbb{S}'}\left(M^{-1}\sum_{i\in\mathcal{J}}\phi(\mathbf{U}_i^T(\boldsymbol{\eta} + v\boldsymbol{\Delta}))(\mathbf{U}_i^T\boldsymbol{\Delta})^2 \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2,\ \ \forall\|\boldsymbol{\Delta}\|_2 \leq 1\right)$$
$$\geq 1 - c_1\exp(-c_2M). \tag{4.44}$$

Lemma 4.8 follows directly by repeating the proof of Lemma 4.5 of [ZCB21]; see also for other slightly different versions in Proposition 2 of [NRWY10] and Theorem 9.36 and Example 9.17 of [Wai19]. Note that, instead of (4.44), the lower bound in [ZCB21] is

$$\kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\sqrt{\frac{\log d}{M}}\|\boldsymbol{\Delta}\|_2\|\boldsymbol{\Delta}\|_1,$$

with some constants $\kappa_1', \kappa_2' > 0$. Here, by the fact that $2ab \le a^2 + b^2$, we get

$$\begin{aligned}\kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\sqrt{\frac{\log d}{M}}\|\boldsymbol{\Delta}\|_2\|\boldsymbol{\Delta}\|_1 &= \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \sqrt{\kappa_1'}\|\boldsymbol{\Delta}\|_2 \cdot \frac{1}{\sqrt{\kappa_1'}}\kappa_2'\sqrt{\frac{\log d}{M}}\|\boldsymbol{\Delta}\|_1 \\ &\ge \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \frac{\kappa_1'}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{{\kappa_2'}^2}{2\kappa_1'}\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2 = \frac{\kappa_1'}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{{\kappa_2'}^2}{2\kappa_1'}\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2.\end{aligned}$$

Lemma 4.9. *Suppose $(\mathbf{X}_i)_{i=1}^m$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$ and $\mathbf{X}$ is an independent copy of $\mathbf{X}_i$. Let $S \subseteq \{1, \ldots, d_1\}$ and $s = |S|$. Then, as $m \to \infty$,*

$$\sup_{\boldsymbol{\Delta} \in \{\boldsymbol{\Delta}_{S^c} = \mathbf{0}, \|\boldsymbol{\Delta}\|_2 = 1\}} \left| m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2 - E\{(\mathbf{X}^T\boldsymbol{\Delta})^2\} \right| = O_p\left(\sqrt{\frac{s}{m}}\right).$$

If we further assume that $S \subset \{1, \ldots, d\}$. Then, as $m \to \infty$,

$$\sup_{\boldsymbol{\Delta} \in \mathbb{C}(S,3) \cap \|\boldsymbol{\Delta}\|_2 = 1} \left| m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2 - E\{(\mathbf{X}^T\boldsymbol{\Delta})^2\} \right| = O_p\left(\sqrt{\frac{s}{m}}\right).$$

Lemma 4.9 is an analog of Lemmas 15 and 16 of [BWZ19]. It can be shown by repeating the proof of [BWZ19], with replacing $\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T$ therein by $E(\mathbf{X}\mathbf{X}^T)$.

Lemma 4.10. *Suppose $(\mathbf{X}_i)_{i=1}^m$ are i.i.d. sub-Gaussian random vectors. Then, for any (possibly random) $\boldsymbol{\Delta} \in \mathbb{R}^d$, as $m \to \infty$,*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

For any $s, k > 0$, define $\widetilde{C}(s,k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \le k\sqrt{s}\|\boldsymbol{\Delta}\|_2\}$ and $\widetilde{K}(s,k,1) := \widetilde{C}(s,k) \cap \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_2 = 1\}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, define

$$\mathcal{F}(\boldsymbol{\Delta}) := \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta} - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1.$$

The following Lemmas 4.11 and 4.13 are anologs of Lemma 9.21 and the proof of Theorem 9.19 of [Wai19], respectively. By Lemma 4.12, we can see that $\boldsymbol{\Delta}_{\boldsymbol{\delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$ with high probability. Instead of the usual cone set $\mathbb{C}(S,k) = \{\|\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq k\|\boldsymbol{\Delta}_S\|_1\}$, we work on another cone set $\widetilde{C}(s,k)$ (and $\widetilde{K}(s,k,1)$) defined above.

Lemma 4.11. *Let Assumptions 4.1 and 4.4 hold,* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$, $s_{\boldsymbol{\gamma}} = o(\frac{N}{\log d_1})$, *and* $s_{\boldsymbol{\delta}} = o(\frac{N}{\log d})$. *For any* $t > 0$, *suppose that* $\lambda_{\boldsymbol{\delta}} > 2\sigma_{\boldsymbol{\delta}}\sqrt{\frac{t+\log d}{M}}$. *Define*

$$\mathcal{A}_1 := \{\|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\|_\infty \leq \lambda_{\boldsymbol{\delta}}/2\}, \tag{4.45}$$

$$\mathcal{A}_2 := \left\{|R_1(\boldsymbol{\Delta})| \leq c\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}} \left(\frac{\|\boldsymbol{\Delta}\|_1}{\sqrt{N}} + \|\boldsymbol{\Delta}\|_2\right), \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right\}, \tag{4.46}$$

$$\mathcal{A}_3 := \left\{\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_2 \leq 1\right\}, \tag{4.47}$$

where $R_1(\boldsymbol{\Delta}) := \left\{\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\}^T \boldsymbol{\Delta}$ *and* $c > 0$ *is some constant. Let* $\bar{s}_{\boldsymbol{\delta}} := \frac{s_{\boldsymbol{\gamma}} \log d_1}{\log d} + s_{\boldsymbol{\delta}}$. *Then, on the event* $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, *for all* $\boldsymbol{\Delta} \in \widetilde{K}(\bar{s}_{\boldsymbol{\delta}}, k_0, 1)$, *we have* $\mathcal{F}(\boldsymbol{\Delta}) > 0$, *when* $N > N_1$ *with some constant* $N_1 > 0$, *and* $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - t - 2\exp(-t)$.

Lemma 4.12. *Let Assumptions 4.1 and 4.4 hold and* $s_{\boldsymbol{\gamma}} = o(\frac{N}{\log d_1})$. *Define* $\boldsymbol{\Delta}_{\boldsymbol{\delta}} := \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$. *Let* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$. *For any* $t > 0$, *suppose that* $\lambda_{\boldsymbol{\delta}} > 2\sigma_{\boldsymbol{\delta}}\sqrt{\frac{t+\log d}{M}}$. *Events* $\mathcal{A}_1$ *and* $\mathcal{A}_2$ *are defined in* (4.45) *and* (4.46). *Then, on the event* $\mathcal{A}_1 \cap \mathcal{A}_2$, *when* $N > N_0$,

$$4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \leq \left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} + 4c\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right)\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2,$$

$$\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \leq k_0\sqrt{\bar{s}_{\boldsymbol{\delta}}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2,$$

where N_0, k_0 *and* $c > 0$ *are some constants and* $\bar{s}_{\boldsymbol{\delta}} := \frac{s_{\boldsymbol{\gamma}} \log d_1}{\log d} + s_{\boldsymbol{\delta}}$.

Lemma 4.13. *Let the assumptions in Lemma 4.11 hold and also that* $\boldsymbol{\Delta}_{\boldsymbol{\delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$. *Then, on the event* $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, *we have* $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$.

Lemma 4.14. *Suppose $a, b, c, x \in \mathbb{R}$, $a > 0$, and $b, c > 0$. Let $ax^2 - bx - c \leq 0$. Then,*

$$x \leq \frac{b}{a} + \sqrt{\frac{c}{a}}.$$

Lemma 4.15. *Let the assumptions in part (a) of Theorem 4.3 hold. Let $r > 0$ be any positive constant. Then, as $N \to \infty$,*

$$\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r} &= \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r} \\ &= O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right). \end{aligned} \tag{4.48}$$

Define

$$\mathcal{E}_1 := \left\{ \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1 \text{ and } \left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,12} \leq C,\ \forall \boldsymbol{\gamma} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\} \right\}. \tag{4.49}$$

Then, as $N \to \infty$,

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1 - o(1).$$

On the event $\mathcal{E}_1$, for any $r' \in [1, 12]$ and $\boldsymbol{\gamma} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\}$, we also have

$$\left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} \leq C, \quad \left\|\exp(-\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} \leq C, \quad \left\|\exp(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} \leq C',$$

with some constant $C' > 0$.

Lemma 4.16. *Let $r > 0$ be any positive constant.*

(a) *Let the assumptions in part (b) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} &= \left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} \\ &= O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right). \end{aligned} \tag{4.50}$$

(b) *Let the assumptions in part (a) of Theorem 4.4 hold. Then, as* $N \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} &= \left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} \\ &= O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right). \end{aligned} \tag{4.51}$$

Let either (a) or (b) holds. Let $C > 0$ *be some constant, define*

$$\mathcal{E}_2 := \left\{ \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1 \text{ and } \left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,6} \le C,\ \forall \boldsymbol{\delta} \in \left\{ w\boldsymbol{\delta}^* + (1-w)\widehat{\boldsymbol{\delta}} : w \in [0,1] \right\} \right\}. \tag{4.52}$$

Then, as $N \to \infty$,

$$P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{E}_2) = 1 - o(1).$$

On the event $\mathcal{E}_2$, *for any* $r' \in [1, 12]$ *and* $\boldsymbol{\delta} \in \{w\boldsymbol{\delta}^* + (1-w)\widehat{\boldsymbol{\delta}} : w \in [0,1]\}$, *we also have*

$$\left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} \le C, \quad \left\|\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} \le C, \quad \left\|\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} \le C',$$

with some constant $C' > 0$.

Lemma 4.17. *Let* $r > 0$ *be any positive constant.*

(a) *Let the assumptions in part (c) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}}\right).$$

(b) *Let the assumptions in part (b) of Theorem 4.4 hold. Then, as* $N \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\right).$$

Let either (a) or (b) holds. For any $v_1 \in [0,1]$, *let* $\widetilde{\boldsymbol{\alpha}} = v_1\boldsymbol{\alpha}^* + (1 - v_1)\widehat{\boldsymbol{\alpha}}$. *Define* $\widetilde{\varepsilon} := Y(1,1) - \bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\alpha}}$. *Then, for any constant* $r > 0$, $\|\widetilde{\varepsilon}\|_{P,r} = O_p(1)$.

Lemma 4.18. *Let* $r > 0$ *be any positive constant.*

(a) *Let the assumptions in part (d) of Theorem 4.3 hold. Then, as* $N \to \infty$,

$$\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right) = O_p\left(\sqrt{\frac{(s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\beta}}) \log d_1 + (s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}}) \log d}{N}}\right).$$

(b) *Let the assumptions in part (c) or part (d) or part (e) of Theorem 4.4 hold. Then, as* $N \to \infty$,

$$\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right)$$

Let either (a) or (b) holds, and let either (a) or (b) of 4.17 holds. For any $v_1, v_2 \in [0,1]$, *let* $\widetilde{\boldsymbol{\alpha}} = v_1\boldsymbol{\alpha}^* + (1 - v_1)\widehat{\boldsymbol{\alpha}}$ *and* $\widetilde{\boldsymbol{\beta}} = v_1\boldsymbol{\beta}^* + (1 - v_1)\widehat{\boldsymbol{\beta}}$. *Define* $\widetilde{\zeta} := \bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\alpha}} - \mathbf{S}_1^T\widetilde{\boldsymbol{\beta}}$. *Then, for any constant* $r > 0$, $\|\widetilde{\zeta}\|_{P,r} = O_p(1)$.

4.5.2 Proof of the main theorems

Proof of Theorem 4.1. Recall the definition of the score function, (4.11). Observe that

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}) =& - A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}) \left\{ \frac{A_2(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha})}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})} + \bar{\mathbf{S}}_2^T\boldsymbol{\alpha} - \mathbf{S}_1^T\boldsymbol{\beta} \right\} \mathbf{S}_1, \\
\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}) =& - \frac{A_1 A_2 \exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta})(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha})}{g(\mathbf{S}_1^T\boldsymbol{\gamma})} \bar{\mathbf{S}}_2, \\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}) =& \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma})} \left\{ 1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})} \right\} \bar{\mathbf{S}}_2, \\
\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}) =& \left\{ 1 - \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma})} \right\} \mathbf{S}_1.
\end{aligned}$$

By the constructions in (4.12)-(4.15), we have

$$\begin{aligned}
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*) \left\{ \frac{A_2(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)} + \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^* \right\} \mathbf{S}_1 \right] \\
&= \mathbf{0} \in \mathbb{R}^{d_1}, \qquad (4.53) \\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -E\left[\frac{A_1 A_2 \exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)(Y-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)} \bar{\mathbf{S}}_2 \right] = \mathbf{0} \in \mathbb{R}^d, \qquad (4.54) \\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= E\left[\frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)} \left\{ 1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)} \right\} \bar{\mathbf{S}}_2 \right] = \mathbf{0} \in \mathbb{R}^d, \qquad (4.55) \\
E\left\{\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= E\left[\left\{ 1 - \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)} \right\} \mathbf{S}_1 \right] = \mathbf{0} \in \mathbb{R}^{d_1}, \qquad (4.56)
\end{aligned}$$

Note that,

$$\begin{aligned}
\widehat{\theta}_{1,1} - \theta_{1,1} &= N^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \theta_{1,1} \\
&= \mathbb{K}^{-1} \sum_{k=1}^{\mathbb{K}} n^{-1} \sum_{i \in \mathcal{I}_k} \{\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}_i; \boldsymbol{\eta}^*)\} + N^{-1} \sum_{i=1}^{N} \psi(\mathbf{W}_i; \boldsymbol{\eta}^*) - \theta_{1,1} \\
&= N^{-1} \sum_{i=1}^{N} \psi(\mathbf{W}_i; \boldsymbol{\eta}^*) - \theta_{1,1} + \mathbb{K}^{-1} \sum_{k=1}^{\mathbb{K}} (\Delta_{k,1} + \Delta_{k,2}),
\end{aligned}$$

where

$$\Delta_{k,1} = n^{-1} \sum_{i \in \mathcal{I}_k} \{\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}_i; \boldsymbol{\eta}^*)\} - E\{\psi(\mathbf{W}; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\},$$

$$\Delta_{k,2} = E\{\psi(\mathbf{W}; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\}.$$

Step 1 We demonstrate that

$$E\{\psi(\mathbf{W}; \boldsymbol{\eta}^*)\} - \theta_{1,1} = 0. \tag{4.57}$$

Here, (4.57) can be shown under the Assumption 4.2:

$$\begin{aligned} E\{\psi(\mathbf{W}; \boldsymbol{\eta}^*)\} - \theta_{1,1} &= E\left[\left\{1 - \frac{A_1}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^T \boldsymbol{\beta}^* - Y(1,1)\}\right] \\ &\quad + E\left[\frac{A_1}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^T \boldsymbol{\alpha}^* - Y(1,1)\}\right] \\ &\overset{(i)}{=} E\left[\left\{1 - \frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^T \boldsymbol{\beta}^* - \mu(\mathbf{S}_1)\}\right] \\ &\quad + E\left[\frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\left\{1 - \frac{\rho(\bar{\mathbf{S}}_2)}{g(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^T \boldsymbol{\alpha}^* - \nu(\bar{\mathbf{S}}_2)\}\right] \overset{(ii)}{=} 0, \end{aligned}$$

where (i) holds by the tower rule, (ii) holds under the Assumption 4.2.

Step 2 We demonstrate that, for each $k \leq \mathbb{K}$ and any $\theta \in \mathbb{R}$, as $N \to \infty$,

$$\Delta_{k,2} = o_p(N^{-1/2}). \tag{4.58}$$

Note that,

$$\Delta_{k,2} = \Delta_{k,3} + \Delta_{k,4} + \Delta_{k,5} + \Delta_{k,6} + \Delta_{k,7} + \Delta_{k,8} + \Delta_{k,9},$$

where

$$\Delta_{k,3} = E\left[\frac{A_1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}\left\{1-\frac{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})}\right\}\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*)\right],$$

$$\Delta_{k,4} = E\left[\left\{1-\frac{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}\right\}\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*)\right],$$

$$\Delta_{k,5} = E\left[\frac{1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}\left\{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)-A_1\right\}\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*)\right],$$

$$\Delta_{k,6} = E\left[\frac{A_1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})}\left\{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)-A_2\right\}\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*)\right],$$

$$\Delta_{k,7} = E\left[\frac{A_1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}\left\{\frac{A_2}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})}-\frac{B}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right\}\{Y(1,1)-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*\}\right],$$

$$\Delta_{k,8} = E\left[\left\{\frac{A_1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}-\frac{A_2}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right\}\{Y(1,1)-\mathbf{S}_1^T\boldsymbol{\beta}^*\}\right],$$

$$\Delta_{k,9} = E\left[\left\{\frac{A_1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}-\frac{A_2}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right\}\left\{\frac{A_2}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}-1\right\}\{Y(1,1)-\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*\}\right].$$

By the tower rule, $\Delta_{k,5}=0$ when $\pi(\mathbf{S}_1)=\pi^*(\mathbf{S}_1):=g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)$; $\Delta_{k,6}=0$ when $\rho(\cdot)=\rho^*(\cdot):=g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)$; $\Delta_{k,7}=0$ when $\nu(\cdot)=\nu^*(\cdot):=\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*$; $\Delta_{k,8}=0$ when $\mu(\cdot)=\mu^*(\cdot):=\mathbf{S}_1^T\boldsymbol{\beta}^*$; $\Delta_{k,9}=0$ since either $\rho(\cdot)=\rho^*(\cdot)$ or $\nu(\cdot)=\nu^*(\cdot)$. Hence,

$$\Delta_{k,2}=\Delta_{k,3}+\Delta_{k,4}+\Delta_{k,5}\mathbb{1}_{\pi\neq\pi^*}+\Delta_{k,6}\mathbb{1}_{\rho\neq\rho^*}+\Delta_{k,7}\mathbb{1}_{\nu\neq\nu^*}+\Delta_{k,8}\mathbb{1}_{\mu\neq\mu^*},$$

Now, we condition on the event $\mathcal{E}_1\cap\mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ are defined as (4.49) and (4.52), respectively. By Lemmas 4.15 and 4.16, $\mathcal{E}_1\cap\mathcal{E}_2$ occurs with probability $1-o(1)$. Then, by Lemmas 4.16 and 4.17,

$$\begin{aligned}|\Delta_{k,3}| &\le \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})\right\|_{P,4}\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})\right\|_{P,4}\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})-g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,4}\\ &\quad\cdot\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*)\right\|_{P,4}\\ &= O_p\left(\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*\|_2\|\widehat{\boldsymbol{\alpha}}_{-k}-\gamma^*\|_2\right).\end{aligned}$$

Similarly, by Lemmas 4.15 and 4.18,

$$
\begin{aligned}
|\Delta_{k,4}| &\leq \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})\right\|_{P,4} \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,4} \left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{P,2} \\
&= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right).
\end{aligned}
$$

In addition, recall (4.56), we have

$$
\begin{aligned}
|\Delta_{k,5}| &= \left|E\left[\left\{\frac{1}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})} - \frac{1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right\}\left\{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*) - A\right\}\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right]\right| \\
&\leq \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,2} \left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{P,2} \\
&= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right),
\end{aligned}
$$

by Lemmas 4.15 and 4.18. Recall (4.55), we have

$$
\begin{aligned}
|\Delta_{k,6}| &= \left|E\left(\left[\frac{A_1\left\{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*) - A_2\right\}}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})} - \frac{A_1\left\{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*) - A_2\right\}}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right]\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right)\right| \\
&\leq \left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,2} \\
&\quad \cdot \left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right\|_{P,2} \\
&\overset{(i)}{=} O_p\left(\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\right)\|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2\right),
\end{aligned}
$$

where (i) holds by Lemma 4.17 and also note that, using Lemmas 4.15 and 4.16, we have

$$
\begin{aligned}
&\left\| g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,2} \\
&\quad\le \left\| g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})\left\{ g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\}\right\|_{P,2} \\
&\qquad + \left\| g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\left\{ g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}\right\|_{P,2} \\
&\qquad + \left\| \left\{ g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}\left\{ g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\}\right\|_{P,2} \\
&\quad\le \left\| g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})\right\|_{P,4}\left\| g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,4} \\
&\qquad + \left\| g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,4}\left\| g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,4} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\right).
\end{aligned}
$$

With some $\widetilde{\boldsymbol{\gamma}}_1$ lies between $\boldsymbol{\gamma}^*$ and $\widehat{\boldsymbol{\gamma}}_{-k}$, some $\widetilde{\boldsymbol{\delta}}$ lies between $\boldsymbol{\delta}^*$ and $\widehat{\boldsymbol{\delta}}_{-k}$, we have

$$
\begin{aligned}
|\Delta_{k,7}| &\overset{(i)}{=} \left| E\left[\frac{A_2}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}_{-k})}\left\{ \frac{A_1}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}_{-k})} - \frac{A_1}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right\}\varepsilon\right]\right| \\
&\overset{(ii)}{\le} \left| E\left\{ \frac{A_1A_2}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^T\right\}(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right| \\
&\quad + \left| E\left\{ A_1A_2\exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_1)\exp(-\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\varepsilon\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}\right| \\
&\quad + \left| E\left[\frac{A_1A_2}{g(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_1)}\exp(-\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\varepsilon\left\{\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\}^2\right]\right| \\
&\overset{(iii)}{\le} \left\|\exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_1)\right\|_{P,4}\left\|\exp(-\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,4}\|\varepsilon\|_{P,4}\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{P,8}\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{P,8} \\
&\quad + \left\| g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_1)\right\|_{P,4}\left\|\exp(-\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,4}\|\varepsilon\|_{P,4}\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{P,8}^2 \\
&\overset{(iv)}{=} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2^2\right),
\end{aligned}
$$

where (i) holds since either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$; (ii) holds by Taylor's theorem; (iii) holds by (4.54) and Hölder's inequality; (iv) holds by Lemmas 4.15 and 4.16. Similarly, by

Taylor's theorem, with some $\widetilde{\boldsymbol{\gamma}}_2$ lies between $\boldsymbol{\gamma}^*$ and $\widehat{\boldsymbol{\gamma}}_{-k}$, we have

$$
\begin{aligned}
|\Delta_{k,8}| \leq & \left|E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{Y(1,1) - \mathbf{S}_1^T\boldsymbol{\beta}^*\}\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right]\right| \\
& + \left|E\left[A_1 \exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_2)\{Y(1,1) - \mathbf{S}_1^T\boldsymbol{\beta}^*\}\left\{\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}^2\right]\right| \\
\overset{(i)}{=} & \left|E\left[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\frac{A_2(Y - \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)} + \bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - \mathbf{S}_1^T\boldsymbol{\beta}^*\right\}\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right]\right| \\
& + \left|E\left[A_1 \exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_2)(\varepsilon + \zeta)\left\{\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}^2\right]\right| \\
\overset{(ii)}{\leq} & 0 + \left\|\exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}}_2)\right\|_{P,4}\|\varepsilon + \zeta\|_{P,4}\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{P,4}^2 \\
\overset{(iii)}{=} & O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2^2\right),
\end{aligned}
$$

where (i) holds since either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$; (ii) holds by (4.53) and Hölder's inequality; (iii) holds by Lemma 4.15.

To sum up, we have

$$
\begin{aligned}
\Delta_{k,2} = & O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2\right) \\
& + \mathbb{1}_{\rho\neq\rho^*}O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2\right) \\
& + \mathbb{1}_{\nu\neq\nu^*}O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2^2\right) \\
& + \mathbb{1}_{\mu\neq\mu^*}O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2^2\right).
\end{aligned}
$$

Define

$$
r_{\boldsymbol{\gamma}} := \sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}},\ \ r_{\boldsymbol{\delta}} := \sqrt{\frac{s_{\boldsymbol{\delta}}\log d}{N}},\ \ r_\gamma := \sqrt{\frac{s_\gamma \log d}{N}},\ \ r_{\boldsymbol{\beta}} := \sqrt{\frac{s_{\boldsymbol{\beta}}\log d_1}{N}}.
$$

By Theorems 4.3 and 4.4,

$$
\begin{aligned}
&\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 = O_p\left(r_{\boldsymbol{\gamma}}\right), \\
&\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 = \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\mathbb{1}_{\rho=\rho^*} + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\mathbb{1}_{\rho\neq\rho^*} = O_p\left(r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}}\mathbb{1}_{\rho\neq\rho^*}\right), \\
&\|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2 = \|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2\mathbb{1}_{\nu=\nu^*} + \|\widehat{\boldsymbol{\alpha}}_{-k} - \gamma^*\|_2\mathbb{1}_{\nu\neq\nu^*} = O_p\left(r_\gamma + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}})\mathbb{1}_{\nu\neq\nu^*}\right).
\end{aligned}
$$

Additionally, note that either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$ (or both) holds. By Theorems 4.3 and 4.4, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 &= \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 \mathbb{1}_{\rho=\rho^*,\nu=\nu^*,\mu=\mu^*} + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 \mathbb{1}_{\rho=\rho^*,\nu\neq\nu^*,\mu=\mu^*} \\ &\quad + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 \mathbb{1}_{\rho\neq\rho^*,\nu=\nu^*,\mu=\mu^*} + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 \mathbb{1}_{\mu\neq\mu^*} \\ &= O_p\left(r_{\boldsymbol{\beta}} + r_{\boldsymbol{\delta}} \mathbb{1}_{\nu\neq\nu^*} + r_{\boldsymbol{\alpha}} \mathbb{1}_{\rho\neq\rho^*} + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}} + r_{\boldsymbol{\alpha}}) \mathbb{1}_{\mu\neq\mu^*}\right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\Delta_{k,2} &= O_p\left(r_{\boldsymbol{\gamma}} \left\{r_{\boldsymbol{\beta}} + r_{\boldsymbol{\delta}} \mathbb{1}_{\nu\neq\nu^*} + r_{\boldsymbol{\alpha}} \mathbb{1}_{\rho\neq\rho^*} + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}} + r_{\boldsymbol{\alpha}}) \mathbb{1}_{\mu\neq\mu^*}\right\}\right) \\ &\quad + O_p\left(\left(r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} \mathbb{1}_{\rho\neq\rho^*}\right)\left(r_\gamma + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}}) \mathbb{1}_{\nu\neq\nu^*}\right)\right) \\ &\quad + \mathbb{1}_{\rho\neq\rho^*} O_p\left(r_{\boldsymbol{\gamma}} \left\{r_\gamma + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}}) \mathbb{1}_{\nu\neq\nu^*}\right\}\right) \\ &\quad + \mathbb{1}_{\nu\neq\nu^*} O_p\left(\left(r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} \mathbb{1}_{\rho\neq\rho^*}\right)\left(r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} \mathbb{1}_{\rho\neq\rho^*}\right)\right) + \mathbb{1}_{\mu\neq\mu^*} O_p\left(r_{\boldsymbol{\gamma}}^2\right) \\ &\overset{(i)}{=} O_p\left(r_{\boldsymbol{\gamma}} r_{\boldsymbol{\beta}} + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\delta}} \mathbb{1}_{\nu\neq\nu^*} + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\alpha}} \mathbb{1}_{\rho\neq\rho^*} + r_{\boldsymbol{\gamma}}(r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}} + r_{\boldsymbol{\alpha}}) \mathbb{1}_{\mu\neq\mu^*}\right) \\ &\quad + O_p\left(r_{\boldsymbol{\delta}} r_\gamma + r_{\boldsymbol{\gamma}} r_\gamma \mathbb{1}_{\rho\neq\rho^*} + r_{\boldsymbol{\delta}}(r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}}) \mathbb{1}_{\nu\neq\nu^*}\right) \\ &\quad + \mathbb{1}_{\rho\neq\rho^*} O_p(r_{\boldsymbol{\gamma}} r_\gamma) + \mathbb{1}_{\nu\neq\nu^*} O_p\left(\left(r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}}\right) r_{\boldsymbol{\delta}}\right) + \mathbb{1}_{\mu\neq\mu^*} O_p\left(r_{\boldsymbol{\gamma}}^2\right) \\ &= O_p(r_{\boldsymbol{\gamma}} r_{\boldsymbol{\beta}} + r_{\boldsymbol{\delta}} r_\gamma) + \mathbb{1}_{\rho\neq\rho^*} O_p(r_{\boldsymbol{\gamma}} r_{\boldsymbol{\alpha}}) + \mathbb{1}_{\nu\neq\nu^*} O_p(r_{\boldsymbol{\gamma}} r_{\boldsymbol{\delta}} + r_{\boldsymbol{\delta}}^2) \\ &\quad + \mathbb{1}_{\mu\neq\mu^*} O_p(r_{\boldsymbol{\gamma}}^2 + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\alpha}}).\end{aligned}$$

where (i) holds since $\mathbb{1}_{\rho\neq\rho^*} \mathbb{1}_{\nu\neq\nu^*} = 0$ that either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$ (or both) holds. Note that when all the nuisance models are correctly specified,

$$r_{\boldsymbol{\gamma}} r_{\boldsymbol{\beta}} + r_{\boldsymbol{\delta}} r_\gamma = \frac{\sqrt{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}}} \log d_1}{N} + \frac{\sqrt{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}} \log d}{N} = o(N^{-1/2}),$$

and (a) when $\rho(\cdot) \neq \rho^*(\cdot)$,

$$r_{\boldsymbol{\gamma}} r_\gamma = \frac{\sqrt{s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} \log d_1 \log d}}{N} = o(N^{-1/2}),$$

b) when $\nu(\cdot) \neq \nu^*(\cdot)$,

$$r_{\boldsymbol{\gamma}} r_{\boldsymbol{\delta}} + r_{\boldsymbol{\delta}}^2 = \frac{\sqrt{s_{\boldsymbol{\delta}} \log d (s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d)}}{N} = o(N^{-1/2}),$$

c) when $\mu(\cdot) \neq \mu^*(\cdot)$,

$$r_{\boldsymbol{\gamma}}^2 + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\alpha}} = \frac{\sqrt{s_{\boldsymbol{\gamma}} \log d_1 (s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d)}}{N} = o(N^{-1/2}).$$

Hence, we conclude that

$$\Delta_{k,2} = o(N^{-1/2}).$$

Step 3 We demonstrate that, for each $k \leq \mathbb{K}$ and any $\theta \in \mathbb{R}$, as $N \to \infty$,

$$\Delta_{k,1} = o_P(N^{-1/2}).$$

By construction, we have

$$E_{\mathbb{S}_k}(\Delta_{k,1}) = 0.$$

By Taylor's theorem, with some $\widetilde{\boldsymbol{\eta}} = (\widetilde{\boldsymbol{\gamma}}^T, \widetilde{\boldsymbol{\delta}}^T, \widetilde{\boldsymbol{\alpha}}^T, \widetilde{\boldsymbol{\beta}}^T)^T$ lies between $\boldsymbol{\eta}^*$ and $\widehat{\boldsymbol{\eta}}_{-k}$,

$$\begin{aligned}
E_{\mathbb{S}_k}(\Delta_{k,1}^2) &= n^{-1} E\left[\{\psi(\mathbf{W}; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\}^2\right] \\
&= 2n^{-1} E\left[\{\psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\} \nabla_{\boldsymbol{\eta}} \psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}})^T (\widehat{\boldsymbol{\eta}}_{-k} - \boldsymbol{\eta}^*)\right] \\
&\leq 2n^{-1} \left\{ \left\| \psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}}) - \mathbf{S}_1^T \boldsymbol{\beta}^* \right\|_{P,2} + \left\| \psi(\mathbf{W}; \boldsymbol{\eta}^*) - \mathbf{S}_1^T \boldsymbol{\beta}^* \right\|_{P,2} \right\} \left\| \nabla_{\boldsymbol{\eta}} \psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}})^T (\widehat{\boldsymbol{\eta}}_{-k} - \boldsymbol{\eta}^*) \right\|_{P,2}.
\end{aligned}$$

Note that, with probability 1,

$$\left\| \psi(\mathbf{W}; \boldsymbol{\eta}^*) - \mathbf{S}_1^T \boldsymbol{\beta}^* \right\|_{P,2} \leq c^{-1} \left\| \zeta \right\|_{P,2} + c^{-2} \left\| \varepsilon \right\|_{P,2} = O(1).$$

Define

$$\widetilde{\varepsilon} := Y(1,1) - \bar{\mathbf{S}}_2^T \widetilde{\boldsymbol{\alpha}}, \quad \widetilde{\zeta} := \bar{\mathbf{S}}_2^T \widetilde{\boldsymbol{\alpha}} - \mathbf{S}_1^T \widetilde{\boldsymbol{\beta}}.$$

Condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$. By Lemmas 4.17 and 4.18, we also have

$$\begin{aligned}
&\left\|\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}}) - \mathbf{S}_1^T\boldsymbol{\beta}^*\right\|_{P,2} \\
&\quad\le \left\|g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,4}\left\|\widetilde{\zeta}\right\|_{P,4} + \left\|g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,6}\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,6}\|\widetilde{\varepsilon}\|_{P,6} + \left\|\mathbf{S}_1^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{P,2} \\
&\quad= O_p\left(1 + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right) = O_p\left(1 + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right) = O_p(1).
\end{aligned}$$

In addition,

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\eta}}_{-k} - \boldsymbol{\eta}^*)\right\|_{P,2} \\
&\quad\le \left\|\nabla_{\boldsymbol{\gamma}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{P,2} + \left\|\nabla_{\boldsymbol{\delta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{P,2} \\
&\qquad + \left\|\nabla_{\boldsymbol{\alpha}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right\|_{P,2} + \left\|\nabla_{\boldsymbol{\beta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{P,2}.
\end{aligned}$$

Here, by Lemma 4.15,

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\gamma}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{P,2} \\
&\quad\le \left\|\exp(-\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,6}\left\{\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,6}\|\widehat{\varepsilon}\|_{P,6} + \left\|\widehat{\zeta}\right\|_{P,3}\right\}\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{P,6} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\right).
\end{aligned}$$

Similarly, for the second term, by Lemmas 4.16 and 4.17,

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\delta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{P,2} \\
&\quad\le \left\|g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,6}\left\|\exp(-\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,6}\|\widetilde{\varepsilon}\|_{P,12}\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{P,12} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\right).
\end{aligned}$$

For the third term, by Lemma 4.17,

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\alpha}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*)\right\|_{P,2}\\
&\quad\le \left\|g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,6}\left\{1+\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widetilde{\boldsymbol{\delta}})\right\|_{P,6}\right\}\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*)\right\|_{P,6}\\
&\quad= O_p\left(\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*\|_2\right).
\end{aligned}$$

Lastly, by Lemma 4.18,

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\beta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*)\right\|_{P,2}\\
&\quad\le\left\{1+\left\|g^{-1}(\mathbf{S}_1^T\widetilde{\boldsymbol{\gamma}})\right\|_{P,4}\right\}\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*)\right\|_{P,4}=O_p\left(\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*\|_2\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^T(\widehat{\boldsymbol{\eta}}_{-k}-\boldsymbol{\eta}^*)\right\|_{P,2}\\
&\quad= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*\|_2+\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*\|_2+\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*\|_2\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
E_{\mathbb{S}_k}(\Delta_{k,1}^2)&=n^{-1}E\left[\{\psi(\mathbf{W};\widehat{\boldsymbol{\eta}}_{-k})-\psi(\mathbf{W};\boldsymbol{\eta}^*)\}^2\right]\\
&=N^{-1}O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*\|_2+\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*\|_2+\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*\|_2\right). \quad (4.59)
\end{aligned}$$

By Lemma 4.4,

$$\begin{aligned}
\Delta_{k,1}&=O_p\left(N^{-1/2}\sqrt{\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*\|_2+\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^*\|_2+\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^*\|_2}\right)\\
&=o_p(N^{-1/2}).
\end{aligned}$$

Step 4 We show that, as $N \to \infty$,

$$\sigma^{-1} N^{-1/2} \sum_{i=1}^{N} \psi(\mathbf{W}_i; \boldsymbol{\eta}^*) - \theta_{1,1} \to \mathcal{N}(0,1). \tag{4.60}$$

By Lyapunov's central limit theorem, it suffices to show that, for some $t > 2$,

$$\sigma^{-t} E\left\{|\psi(\mathbf{W}; \boldsymbol{\eta}^*) - \theta_{1,1}|^t\right\} < C, \tag{4.61}$$

with some constant $C > 0$. Note that

$$\begin{aligned}
\sigma^2 &= E\left[\{Y(1,1) - \theta_{1,1}\}\right]^2 + E\left(\left[\left\{1 - \frac{A_1}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^T \boldsymbol{\beta}^* - Y(1,1)\}\right]^2\right) \\
&\quad + E\left(\left[\frac{A_1}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^T \boldsymbol{\alpha}^* - Y(1,1)\}\right]^2\right) \\
&\geq E\left[\{Y(1,1) - \theta_{1,1}\}^2\right] \overset{(i)}{\geq} E\left[\{Y(1,1) - \theta_{1,1}\}^2\right]/2 + E[\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2]/2 \\
&\overset{(ii)}{\geq} c_Y/2 + c_0(1-c_0)^{-1} E[A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2]/2,
\end{aligned}$$

where (i) holds since $E\left[\{Y(1,1) - \theta_{1,1}\}^2\right] = E\left[\{Y(1,1) - \mu(\mathbf{S}_1)\}^2\right] + E\left[\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2\right]$; (ii) holds since $\exp(\mathbf{S}_1^T \boldsymbol{\gamma}^*) > c_0(1-c_0)^{-1}$ under Assumption 4.1, $E[\{Y(1,1) - \theta_{1,1}\}]^2 \geq c_Y$ under Assumption 4.4, and $A_1 \leq 1$. Based on the construction of $\boldsymbol{\beta}^*$ as in (4.15), and since either $\rho^*(\cdot)$ or $\nu^*(\cdot)$ is correctly specified, we have

$$\begin{aligned}
\boldsymbol{\beta}^* &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d_1}} E\left[A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*)\left\{\nu(\bar{\mathbf{S}}_2) - \mathbf{S}_1^T \boldsymbol{\beta}\right\}^2\right] \\
&= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d_1}} E\left[A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*)\left\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T \boldsymbol{\beta}\right\}^2\right],
\end{aligned}$$

which implies

$$E\left[A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*)\left\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T \boldsymbol{\beta}^*\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

Under Assumptions 4.1 and 4.4, it follows that

$$
\begin{aligned}
&E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2] \\
&\quad \overset{(i)}{=} E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T\boldsymbol{\beta}^*\}^2] + E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mathbf{S}_1^T\boldsymbol{\beta}^* - \theta_{1,1}\}^2] \\
&\quad \geq E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mathbf{S}_1^T\boldsymbol{\beta}^* - \theta_{1,1}\}^2] = E[\pi(\mathbf{S}_1) \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mathbf{S}_1^T\boldsymbol{\beta}^* - \theta_{1,1}\}^2] \\
&\quad \geq c_0(c_0^{-1} - 1)E[\{\mathbf{S}_1^T\boldsymbol{\beta}^* - \theta_{1,1}\}^2] \geq (1 - c_0)c_{\min}\|\boldsymbol{\beta}^*\|_2^2,
\end{aligned}
$$

where (i) holds since

$$
\begin{aligned}
&E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T\boldsymbol{\beta}^*\}\{\mathbf{S}_1^T\boldsymbol{\beta}^* - \theta_{1,1}\}] \\
&\quad = E[A_1 \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^T\boldsymbol{\beta}^*\}\mathbf{S}_1^T\{\boldsymbol{\beta}^* - \theta_{1,1}\mathbf{e}_1\}] = 0.
\end{aligned}
$$

Therefore, we have

$$
\sigma^2 \geq c_Y/2 + c_0 c_{\min}\|\boldsymbol{\beta}^*\|_2^2/2.
$$

Additionally, for any $r > 0$,

$$
\begin{aligned}
&\|\psi(\mathbf{W}; \boldsymbol{\eta}^*) - \theta_{1,1}\|_{P,r} \\
&\quad \leq \|Y(1,1) - \theta_{1,1}\|_{P,r} + \left\| \left\{ 1 - \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)} \right\} \{\mathbf{S}_1^T\boldsymbol{\beta}^* - Y(1,1)\} \right\|_{P,r} \\
&\qquad + \left\| \frac{A_1}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)} \left\{ 1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)} \right\} \{\bar{\mathbf{S}}_2^T\boldsymbol{\alpha}^* - Y(1,1)\} \right\|_{P,r} \\
&\quad \leq \|\mathbf{S}_1^T\boldsymbol{\beta}^*\|_{P,r} + \|\varepsilon\|_{P,r} + \|\zeta\|_{P,r} + |\theta_{1,1}| + (1 + c_0^{-1}) \|\varepsilon + \zeta\|_{P,r} \\
&\qquad + c_0^{-1}(1 + c_0^{-1}) \|\varepsilon\|_{P,r} \\
&\quad \overset{(i)}{=} O\left(\|\boldsymbol{\beta}^*\|_2 + 1\right).
\end{aligned}
$$

where (i) holds by $|\theta_{1,1}| = |E(\mathbf{S}_1^T\boldsymbol{\beta}^*)| \leq \|\mathbf{S}_1^T\boldsymbol{\beta}^*\|_{P,1}$. Therefore,

$$
\sigma \asymp \|\boldsymbol{\beta}^*\|_2 + 1, \tag{4.62}
$$

and

$$\sigma^{-t}E\left\{|\psi(\mathbf{W};\boldsymbol{\eta}^*)-\theta_{1,1}|^t\right\}=\left\{\frac{\|\psi(\mathbf{W};\boldsymbol{\eta}^*)-\theta_{1,1}\|_{P,t}}{\sigma}\right\}^t$$
$$=O\left(\frac{\|\boldsymbol{\beta}^*\|_2+1}{c_Y/2+c_{\min}\|\boldsymbol{\beta}^*\|_2^2/2}\right)=O(1),$$

and (4.61) follows.

Step 5 Finally, we prove that, as $N\to\infty$,

$$\widehat{\sigma}^2=\sigma^2\{1+o_p(1)\}. \tag{4.63}$$

Note that

$$E_{\mathbb{S}}\left[\left\{N^{-1}\sum_{i=1}^N\psi(\mathbf{W}_i;\boldsymbol{\eta}^*)-\theta_{1,1}\right\}^2\right]=N^{-1}\sigma^2\asymp\frac{\|\boldsymbol{\beta}^*\|_2^2+1}{N}.$$

By Lemma 4.4,

$$N^{-1}\sum_{i=1}^N\psi(\mathbf{W}_i;\boldsymbol{\eta}^*)-\theta_{1,1}=O_p\left(\frac{\|\boldsymbol{\beta}^*\|_2+1}{\sqrt{N}}\right).$$

By (4.59), (4.61), (4.62), and Lemma S4.4 of [ZCB21], we have (4.63) holds. ∎

Proof of Theorem 4.2. Theorem 4.2 follows directly from Theorem 4.1. ∎

Proof of Lemma 4.1. We show that, with high probability, the RSC property holds for each of the loss functions. By Taylor's theorem, with some $v_1,v_2\in(0,1)$,

$$\delta\bar{\ell}_1(\boldsymbol{\gamma}^*,\boldsymbol{\Delta})=(2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}A_{1i}\exp\{-\mathbf{S}_{1i}^T(\boldsymbol{\gamma}^*+v_1\boldsymbol{\Delta})\}(\mathbf{S}_{1i}^T\boldsymbol{\Delta})^2,$$

$$\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta})=(2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}A_{1i}A_{2i}g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\exp\{-\bar{\mathbf{S}}_{2i}^T(\boldsymbol{\delta}^*+v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2, \tag{4.64}$$

$$\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\Delta})=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}A_{1i}A_{2i}g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\exp(-\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2, \tag{4.65}$$

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*,\boldsymbol{\Delta})=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}A_{1i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^T\boldsymbol{\Delta})^2. \tag{4.66}$$

Part 1 Let $\mathbf{U} = A_1\mathbf{S}_1$, $\mathbb{S}' = (A_{1i}\mathbf{S}_{1i})_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}$, $\phi(u) = \exp(-u)$, $v = v_1$, and $\boldsymbol{\eta} = \boldsymbol{\gamma}^*$. Under Assumption 4.1, $|\mathbf{U}^T\boldsymbol{\eta}| \le |\mathbf{S}_1^T\boldsymbol{\gamma}^*| < C$ with some constant $C > 0$. By Lemmas 4.5 and 4.8, we have (4.40) holds.

Part 2 Now we treat $\widehat{\boldsymbol{\gamma}}$ as fixed (or conditional on) and suppose that $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1$. Note that $g^{-1}(u) = 1 + \exp(-u)$ and $\mathbf{S} = (\mathbf{S}_1^T, \mathbf{S}_2^T)^T$. Hence,

$$\begin{aligned}\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) = (2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\boldsymbol{\delta}^* + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 \\ + (2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}} + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2,\end{aligned}$$

where $\check{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}^T, 0, \ldots, 0)^T \in \mathbb{R}^d$. Let $\mathbf{U} = A_1A_2\mathbf{S}$, $\mathbb{S}' = (A_{1i}A_{2i}\bar{\mathbf{S}}_{2i})_{i\in\mathcal{I}_{\boldsymbol{\delta}}}$, $\phi(u) = \exp(-u)$, $v = v_2$, and $\boldsymbol{\eta} = \boldsymbol{\delta}^*$. Note that, under Assumption 4.1, we have $|\mathbf{U}^T\boldsymbol{\eta}| \le |\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*| < C$ with some constant $C > 0$. By Lemmas 4.5 and 4.8, we have

$$\begin{aligned}(2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\boldsymbol{\delta}^* + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 \\ \ge \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2 \le 1,\end{aligned} \tag{4.67}$$

with probability $P_{\mathbb{S}_{\boldsymbol{\delta}}}$ at least $1 - c_1'\exp(-c_2'M)$ and some constants $\kappa_1', \kappa_2', c_1', c_2' > 0$.

Similarly, let $\mathbf{U} = A_1A_2\mathbf{S}$, $\mathbb{S}' = (A_{1i}A_{2i}\bar{\mathbf{S}}_{2i})_{i\in\mathcal{I}_{\boldsymbol{\delta}}}$, $\phi(u) = \exp(-u)$, $v = v_2$, and $\boldsymbol{\eta} = \boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}}$. On the event $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1$, under Assumptions 4.1 and 4.4, we have $E\{|\mathbf{U}^T\boldsymbol{\eta}|\} \le E(|\mathbf{S}_1^T\boldsymbol{\gamma}^*|) + E(|\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*|) + E\{|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)|\} < C$ with some constant $C > 0$. By Lemmas 4.5 and 4.8, we have

$$\begin{aligned}(2M)^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}} + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 \\ \ge \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2 \le 1,\end{aligned} \tag{4.68}$$

with probability $P_{\mathbb{S}_{\boldsymbol{\delta}}}$ at least $1-c_1'\exp(-c_2'M)$. Hence, (4.41) follows from (4.67) and (4.68).

Part 3 We treat both $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\delta}}$ as fixed (or conditional on) and suppose that $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 \le 1$, $\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*\|_2 \le 1$. Note that

$$\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\Delta}) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp(-\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 + M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\delta}}+\check{\boldsymbol{\gamma}})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2. \quad (4.69)$$

Let $\mathbf{U}=A_1A_2\mathbf{S}$, $\mathbb{S}'=(A_{1i}A_{2i}\bar{\mathbf{S}}_{2i})_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}$, $\phi(u)=\exp(-u)$, $v=0$, and $\boldsymbol{\eta}=\widehat{\boldsymbol{\delta}}$. Here, $E(|\mathbf{U}^T\boldsymbol{\eta}|) \le E(|\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*|)+E\{|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*)|\} < C$ with some constant $C>0$. By Lemmas 4.5 and 4.8, we have

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp(-\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 \ge \kappa_1'\|\boldsymbol{\Delta}\|_2^2-\kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2\le 1, \quad (4.70)$$

with probability $P_{\mathbb{S}_{\boldsymbol{\alpha}}}$ at least $1-c_1'\exp(-c_2'M)$.

Similarly, let $\mathbf{U}=A_1A_2\mathbf{S}$, $\mathbb{S}'=(A_{1i}A_{2i}\bar{\mathbf{S}}_{2i})_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}$, $\phi(u)=\exp(-u)$, $v=0$, and $\boldsymbol{\eta}=\widehat{\boldsymbol{\delta}}+\check{\boldsymbol{\gamma}}$. Then, $E(|\mathbf{U}^T\boldsymbol{\eta}|) \le E(|\mathbf{S}_1^T\boldsymbol{\gamma}^*|)+E(|\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*|)+E\{|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)|\}+E\{|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*)|\} < C$ with some constant $C>0$. By Lemmas 4.5 and 4.8, we have

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\delta}}+\check{\boldsymbol{\gamma}})\}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2 \ge \kappa_1'\|\boldsymbol{\Delta}\|_2^2-\kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2\le 1, \quad (4.71)$$

with probability $P_{\mathbb{S}_{\boldsymbol{\alpha}}}$ at least $1-c_1'\exp(-c_2'M)$. Note that, the function $\delta\ell_N(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\Delta})$ is based on a weighted squared loss, and hence the lower bounds in (4.70) and (4.71) can be extended to any $\boldsymbol{\Delta}\in\mathbb{R}^d$. For any $\boldsymbol{\Delta}'\in\mathbb{R}^d$, we let $\boldsymbol{\Delta}=\boldsymbol{\Delta}'/\|\boldsymbol{\Delta}'\|_2$. Then, $\|\boldsymbol{\Delta}\|_2=1$. The lower bounds in (4.70) and (4.71) hold if we multiply the LHS and RHS by a factor $\|\boldsymbol{\Delta}'\|_2^2$. Therefore, (4.43) holds by combining the lower bounds with (4.69) .

Part 4 Lastly, treat $\widehat{\boldsymbol{\gamma}}$ as fixed (or conditional on) and suppose that $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$. Let $\mathbf{U} = A_1\mathbf{S}_1$, $\mathbb{S}' = (A_{1i}\mathbf{S}_{1i})_{i\in\mathcal{I}_{\boldsymbol{\beta}}}$, $\phi(u) = \exp(-u)$, $v = 0$, and $\boldsymbol{\eta} = \widehat{\boldsymbol{\gamma}}$. Here, $E\{|\mathbf{U}^T\boldsymbol{\eta}|\} \leq E(|\mathbf{S}_1^T\boldsymbol{\gamma}^*|) + E\{|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)|\} < C$ with some constant $C > 0$. Then, (4.42) holds by Lemmas 4.5 and 4.8. Here, similarly as in part 3, the lower bound can be extended to any $\boldsymbol{\Delta} \in \mathbb{R}^d$, since $\delta\ell_N(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})$ is also constructed based on a weighted squared loss. ∎

Proof of Lemma 4.2. Now, we control the gradients of the loss functions.

Part 1 Note that

$$\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}\{1 - A_{1i}g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\}\mathbf{S}_{1i}.$$

By the construction of $\boldsymbol{\gamma}^*$, we have

$$E\left[\{1 - A_1g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

Also, for each $1 \leq j \leq d_1$, $|\{1 - A_1g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\}\mathbf{S}_1^T\mathbf{e}_j| \leq (1 + c_0^{-1})|\mathbf{S}_1^T\mathbf{e}_j|$ and hence, by Lemma D.1 (ii) of [CLCL19],

$$\|\{1 - A_1g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\}\mathbf{S}_1^T\mathbf{e}_j\|_{\psi_2} \leq (1 + c_0^{-1})\|\mathbf{S}_1^T\mathbf{e}_j\|_{\psi_2} \leq (1 + c_0^{-1})\sigma_{\mathbf{S}}.$$

Let $\sigma_{\boldsymbol{\gamma}} := \sqrt{8}(1 + c_0^{-1})\sigma_{\mathbf{S}}$. By Lemma D.2 of [CLCL19], for each $1 \leq j \leq d_1$ and any $t > 0$,

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left|\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)^T\mathbf{e}_j\right| > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) \leq 2\exp(-t - \log d_1).$$

It follows that,

$$\begin{aligned} P_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left\|\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)\right\|_\infty > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) &\leq \sum_{j=1}^{d_1} P_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left|\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)^T\mathbf{e}_j\right| > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) \\ &\leq 2d_1\exp(-t - \log d_1) = 2\exp(-t). \end{aligned}$$

Part 2 Note that

$$\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i} g^{-1}(\mathbf{S}_{1i}^T \boldsymbol{\gamma}^*) \{1 - A_{2i} g^{-1}(\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*)\} \bar{\mathbf{S}}_{2i}.$$

By the construction of $\boldsymbol{\delta}^*$, we have

$$E\left[A_1 g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)\} \bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d.$$

Under Assumption 4.1, we have

$$|A_1 g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)\} \bar{\mathbf{S}}_2^T \mathbf{e}_j| \leq c_0^{-1}(1 + c_0^{-1}) |\bar{\mathbf{S}}_2^T \mathbf{e}_j|,$$

for each $1 \leq j \leq d$. By Lemma D.1 (i) and (ii),

$$\|A_1 g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)\} \bar{\mathbf{S}}_2^T \mathbf{e}_j\|_{\psi_2} \leq (c_0^{-2} + c_0^{-1}) \|\bar{\mathbf{S}}_2^T \mathbf{e}_j\|_{\psi_2} \leq (c_0^{-2} + c_0^{-1}) \sigma_{\mathbf{S}}.$$

Let $\sigma_{\boldsymbol{\delta}} := \sqrt{8}(c_0^{-2} + c_0^{-1}) \sigma_{\mathbf{S}}$. By Lemma D.2 of [CLCL19], for each $1 \leq j \leq d$ and any $t > 0$,

$$P_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T \mathbf{e}_j\right| > \sigma_{\boldsymbol{\delta}} \sqrt{\frac{t + \log d}{M}}\right) \leq 2 \exp(-t - \log d).$$

It follows that,

$$\begin{aligned} P_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty > \sigma_{\boldsymbol{\delta}} \sqrt{\frac{t + \log d}{M}}\right) &\leq \sum_{j=1}^d P_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T \mathbf{e}_j\right| > \sigma_{\boldsymbol{\delta}} \sqrt{\frac{t + \log d}{M}}\right) \\ &\leq 2d \exp(-t - \log d) = 2 \exp(-t). \end{aligned}$$

Part 3 Note that

$$\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*) = -2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} A_{1i} A_{2i} g^{-1}(\mathbf{S}_{1i}^T \boldsymbol{\gamma}^*) \exp(-\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*) \varepsilon_i \bar{\mathbf{S}}_{2i}.$$

By the construction of $\boldsymbol{\alpha}^*$, we have

$$E\left\{-2A_1 A_2 g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \varepsilon \bar{\mathbf{S}}_2\right\} = \mathbf{0} \in \mathbb{R}^d.$$

Under Assumption 4.1, we have

$$|-2A_1A_2g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^T\mathbf{e}_j| \le 2c_0^{-1}(c_0^{-1}-1)|\varepsilon\bar{\mathbf{S}}_2^T\mathbf{e}_j|,$$

for each $1 \le j \le d$. By Lemma D.1 (i), (ii), and (v),

$$\begin{aligned}&\|-2A_1A_2g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^T\mathbf{e}_j\|_{\psi_1}\\ &\quad\le 2c_0^{-1}(c_0^{-1}-1)\|\varepsilon\|_{\psi_2}\|\bar{\mathbf{S}}_2^T\mathbf{e}_j\|_{\psi_2} \le 2c_0^{-1}(c_0^{-1}-1)\sigma_\varepsilon\sigma_{\mathbf{S}}.\end{aligned}$$

Let

$$\sigma_{\boldsymbol{\alpha}} := 2c_0^{-1}(c_0^{-1}-1)\sigma_\varepsilon\sigma_{\mathbf{S}}.$$

By Lemmas D.1 (iv) and D.4 of [CLCL19], for each $1 \le j \le d$ and any $t > 0$,

$$P_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)^T\mathbf{e}_j\right| > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right) \le 2\exp(-t-\log d).$$

It follows that,

$$\begin{aligned}&P_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\right\|_\infty > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right)\\ &\quad\le \sum_{j=1}^d P_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)^T\mathbf{e}_j\right| > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right)\\ &\quad\le 2d\exp(-t-\log d) = 2\exp(-t).\end{aligned}$$

Part 4 Note that

$$\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = -2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\left\{\zeta_i + A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\varepsilon_i\right\}\mathbf{S}_{1i}.$$

By the construction of $\boldsymbol{\beta}^*$, we have

$$E\left[-2A_1\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left\{\zeta + A_2g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\varepsilon\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

Under Assumption 4.1, we have $| - 2A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*) \left\{ \zeta + A_2 g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \varepsilon \right\} \mathbf{S}_1^T \mathbf{e}_j| \leq 2(c_0^{-1} - 1)(|\zeta| + c_0^{-1}|\varepsilon|)|\mathbf{S}_1^T \mathbf{e}_j|$ for each $1 \leq j \leq d$. By Lemma D.1 (i), (ii), and (v),

$$\begin{aligned}
&\| - 2A_1 \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*) \left\{ \zeta + A_2 g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \varepsilon \right\} \mathbf{S}_1^T \mathbf{e}_j \|_{\psi_1} \\
&\quad \leq 2(c_0^{-1} - 1)(\|\zeta\|_{\psi_2} + c_0^{-1} \|\varepsilon\|_{\psi_2}) \|\mathbf{S}_1^T \mathbf{e}_j\|_{\psi_2} \leq 2(c_0^{-1} - 1)(\sigma_\zeta + c_0^{-1} \sigma_\varepsilon) \sigma_{\mathbf{S}}.
\end{aligned}$$

Let $\sigma_{\boldsymbol{\beta}} := 2(c_0^{-1} - 1)(\sigma_\zeta + c_0^{-1} \sigma_\varepsilon) \sigma_{\mathbf{S}}$. By Lemmas D.1 (iv) and D.4 of [CLCL19], for each $1 \leq j \leq d_1$ and any $t > 0$,

$$\begin{aligned}
&P_{\mathbb{S}_{\boldsymbol{\beta}}} \left(\left| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \mathbf{e}_j \right| > \sigma_{\boldsymbol{\beta}} \left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M} \right) \right) \\
&\quad \leq 2 \exp(-t - \log d_1).
\end{aligned}$$

It follows that,

$$\begin{aligned}
&P_{\mathbb{S}_{\boldsymbol{\beta}}} \left(\left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty > \sigma_{\boldsymbol{\beta}} \left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M} \right) \right) \\
&\quad \leq \sum_{j=1}^{d_1} P_{\mathbb{S}_{\boldsymbol{\beta}}} \left(\left| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \mathbf{e}_j \right| > \sigma_{\boldsymbol{\beta}} \left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M} \right) \right) \\
&\quad \leq 2d_1 \exp(-t - \log d_1) = 2 \exp(-t).
\end{aligned}$$

■

Proof of Theorem 4.3. We proof the consistency rates of the nuisance parameter estimators when the models are possibly misspecified.

(a) By Lemmas 4.1 and 4.2, as well as Corollary 9.20 of [Wai19], we have

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p \left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{M}} \right), \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = O_p \left(s_{\boldsymbol{\gamma}} \sqrt{\frac{\log d_1}{M}} \right).$$

(b) By Lemma 4.11, $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - t - 2\exp(-t)$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined in (4.45) and (4.46), respectively. By Lemma 4.12, condition on $\mathcal{A}_1 \cap \mathcal{A}_2$, we have $\boldsymbol{\Delta}_{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^* \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0) = \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \leq k_0 \sqrt{\bar{s}_{\boldsymbol{\delta}}} \|\boldsymbol{\Delta}\|_2\}$, where $\bar{s}_{\boldsymbol{\delta}} = \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{\log d} + s_{\boldsymbol{\delta}}}$ and $k_0 > 0$ is a constant. Additionally, by Lemma 4.13, we also have $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$. By (a), we have $P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1\}) = 1 - o(1)$. Then, by (4.41) in Lemma 4.1, $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_3) \geq 1 - o(1) - c_1 \exp(-c_2 M) = 1 - o(1)$, where $\mathcal{A}_3$ is defined in (4.47). Now, also condition on $\mathcal{A}_3$. Then, we have, for large enough N,

$$
\begin{aligned}
\left(2\lambda_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}} + c\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right) \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 &\overset{(i)}{\geq} \delta \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \frac{\lambda_{\boldsymbol{\delta}}}{4} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \\
&\overset{(ii)}{\geq} \kappa_1 \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2^2 - \kappa_2 \frac{\log d}{M} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1^2 + \frac{\lambda_{\boldsymbol{\delta}}}{4} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \\
&\overset{(iii)}{\geq} \left(\kappa_1 - \kappa_2 k_0^2 \frac{\bar{s}_{\boldsymbol{\delta}} \log d}{M}\right) \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2^2 \overset{(iv)}{\geq} \frac{\kappa_1}{2} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2^2,
\end{aligned}
$$

where (i) holds by Lemma 4.12; (ii) holds by the construction of $\mathcal{A}_3$ and also that $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$; (iii) holds since $\boldsymbol{\Delta}_{\boldsymbol{\delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$ and $\frac{\lambda_{\boldsymbol{\delta}}}{4} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \geq 0$; (iv) holds for large enough N, since $\frac{\bar{s}_{\boldsymbol{\delta}} \log d}{M} = \frac{s_{\boldsymbol{\gamma}} \log d_1}{M} + \frac{s_{\boldsymbol{\delta}} \log d}{M} = o(1)$. Therefore, condition on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$
\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq \frac{4\lambda_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}}}{\kappa_1} + \frac{2c}{\kappa_1} \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}} = O\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),
$$

and since $\boldsymbol{\Delta}_{\boldsymbol{\delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$, it follows that

$$
\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \leq k_0 \sqrt{\bar{s}_{\boldsymbol{\delta}}} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 = O\left(s_{\boldsymbol{\gamma}} \sqrt{\frac{(\log d_1)^2}{N \log d}} + s_{\boldsymbol{\delta}} \sqrt{\frac{\log d}{N}}\right).
$$

Therefore, we conclude that

$$
\begin{aligned}
\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 &= O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right), \\
\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_1 &= O_p\left(s_{\boldsymbol{\gamma}} \sqrt{\frac{(\log d_1)^2}{N \log d}} + s_{\boldsymbol{\delta}} \sqrt{\frac{\log d}{N}}\right).
\end{aligned}
$$

(c) For any $t > 0$, with some $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$ and $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\frac{\log d}{N}}$, define

$$\mathcal{A}_4 := \{\|\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\|_\infty \le \lambda_{\boldsymbol{\alpha}}/2\}, \tag{4.72}$$

$$\mathcal{A}_5 := \left\{\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \ge \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d}{M} \|\boldsymbol{\Delta}\|_1^2, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right\}. \tag{4.73}$$

Let $\lambda_{\boldsymbol{\alpha}} > 2\sigma_{\boldsymbol{\alpha}} \left(2\sqrt{\frac{t+\log d}{M}} + \frac{t+\log d}{M}\right)$ with some $t > 0$. By Lemma 4.2, we have $P_{\mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_4) \ge 1 - 2\exp(-t)$. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$. Similar to the proof of Lemma 4.12 for obtaining (4.96), we have, on the event $\mathcal{A}_4$,

$$2\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_1 + 2|R_2|. \tag{4.74}$$

where

$$\begin{aligned} R_2 &= \left\{\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*) - \nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\}^T \boldsymbol{\Delta} \\ &= 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} A_{1i} A_{2i} \left\{g^{-1}(\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}}) \exp(-\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\mathbf{S}_{1i}^T \boldsymbol{\gamma}^*) \exp(-\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*))\right\} \varepsilon_i \bar{\mathbf{S}}_{2i}^T \boldsymbol{\Delta}. \end{aligned}$$

By the fact that $2ab \le \frac{1}{2}a^2 + 2b^2$,

$$|R_2| \le \frac{1}{2}\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + 2R_3,$$

where

$$R_3 = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} \left(\frac{\exp(-\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*)}{g(\mathbf{S}_{1i}^T \boldsymbol{\gamma}^*)}\right)^2 \frac{g(\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}})} \varepsilon_i^2.$$

By (a) and (b) of Theorem 4.3, we have $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1\}) = 1 - o(1)$.

Note that

$$\begin{aligned} E_{\mathbb{S}_{\boldsymbol{\alpha}}}[R_3] &= E\left[\left(\frac{\exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\right)^2 \frac{g(\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}})} \varepsilon^2\right] \\ &\le \left\|\frac{\exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)}\right\|_{P,6}^2 \left\|\frac{g(\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}})}\right\|_{P,3} \|\varepsilon\|_{P,6}^2 \\ &\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}\right). \end{aligned}$$

where (i) holds by Lemma D.1 (iv) of [CLCL19], as well as the fact that

$$\begin{aligned}
&\left\| \frac{\exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T \boldsymbol{\gamma}^*)} \right\|_{P,6} \\
&\quad \le \left\| g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \left\{ \exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \right\} \right\|_{P,6} \\
&\qquad + \left\| \exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \left\{ g^{-1}(-\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}}) - g^{-1}(-\mathbf{S}_1^T \boldsymbol{\gamma}^*) \right\} \right\|_{P,6} \\
&\qquad + \left\| \exp(-\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \right\|_{P,12} \left\| g^{-1}(-\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}}) - g^{-1}(-\mathbf{S}_1^T \boldsymbol{\gamma}^*) \right\|_{P,12} \\
&\quad = O_p \left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}} \right)
\end{aligned} \tag{4.75}$$

using Lemmas 4.15 and 4.16. Hence,

$$R_3 = O_p \left(\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N} \right). \tag{4.76}$$

Recall (4.74), we have

$$\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_1 + 2R_3.$$

Note that $\|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_1 \le \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_2 \le \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2$. Hence,

$$\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 + 2R_3.$$

Recall (4.65), we have $\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \ge 0$. Then,

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 + \frac{2R_3}{\lambda_{\boldsymbol{\alpha}}} \tag{4.77}$$

Then, by Lemma 4.1, $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}} \cup \mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_5) \ge 1 - o(1) - c_1 \exp(-c_2 M) = 1 - o(1)$, where $\mathcal{A}_5$ is defined in (4.73). Now, condition on $\mathcal{A}_4 \cap \mathcal{A}_5$, for large enough N,

$$\begin{aligned}
4\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 + 2R_3 &\overset{(i)}{\ge} \delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \overset{(ii)}{\ge} \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d}{M} \|\boldsymbol{\Delta}\|_1^2 \\
&\overset{(iii)}{\ge} \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - 2\kappa_2 \frac{\log d}{M} \left(16 s_{\boldsymbol{\alpha}} \|\boldsymbol{\Delta}\|_2^2 + \frac{4R_3^2}{\lambda_{\boldsymbol{\alpha}}^2} \right) \\
&\overset{(iv)}{\ge} \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 - 8\kappa_2 R_3^2 \frac{\log d}{M \lambda_{\boldsymbol{\alpha}}^2}
\end{aligned}$$

where (i) holds by $\|\boldsymbol{\Delta}\|_1 \geq 0$; (ii) holds by the construction of $\mathcal{A}_5$; (iii) holds by (4.77) and the fact that $(a+b)^2 \leq 2a^2+2b^2$; (iv) holds for large enough N, since $\frac{s_{\boldsymbol{\alpha}} \log d}{M} = o(1)$. Hence, on the event $\mathcal{A}_4 \cap \mathcal{A}_5$, for large enough N,

$$\kappa_1 \|\boldsymbol{\Delta}\|_2^2 - 8\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 - 16\kappa_2 R_3^2 \frac{\log d}{M\lambda_{\boldsymbol{\alpha}}^2} - 4R_3 \leq 0.$$

It follows from Lemma 4.14 that

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_2 &\leq \frac{8\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}}}{\kappa_1} + \sqrt{16R_3^2 \frac{\kappa_2 \log d}{\kappa_1 M \lambda_{\boldsymbol{\alpha}}^2} + \frac{4R_3}{\kappa_1}} \\
&\overset{(i)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}} + \frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}} \right) \\
&= O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}} \right)
\end{aligned}$$

where (i) holds by $\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}} \asymp \sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}$ and (4.76). Recall (4.77), we have

$$\|\boldsymbol{\Delta}\|_1 = O_p\left(s_{\boldsymbol{\gamma}} \sqrt{\frac{(\log d_1)^2}{N \log d}} + s_{\boldsymbol{\delta}} \sqrt{\frac{\log d}{N}} + s_{\boldsymbol{\alpha}} \sqrt{\frac{\log d}{N}} \right).$$

(d) For any $t > 0$, with some $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\frac{\log d_1}{N}}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$, $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\frac{\log d}{N}}$ and $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\frac{\log d}{N}}$, define

$$\mathcal{A}_6 := \left\{ \left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty \leq \lambda_{\boldsymbol{\beta}}/2 \right\}, \tag{4.78}$$

$$\mathcal{A}_7 := \left\{ \delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d_1}{M} \|\boldsymbol{\Delta}\|_1^2, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^{d_1} \right\}. \tag{4.79}$$

Let $\lambda_{\boldsymbol{\beta}} > 2\sigma_{\boldsymbol{\beta}} \left(2\sqrt{\frac{t+\log d_1}{M}} + \frac{t+\log d_1}{M} \right)$ with some $t > 0$. By Lemma 4.2, we have $P_{\mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_6) \geq 1 - 2\exp(-t)$. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Similar to the proof of Lemma 4.12 for obtaining (4.96), we have, on the event $\mathcal{A}_6$,

$$2\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}\|_1 \leq 4\lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_1 + 2|R_4|. \tag{4.80}$$

where

$$R_4 = \left\{\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\}^T \boldsymbol{\Delta}$$
$$= 2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\Bigg\{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\alpha}} - \mathbf{S}_{1i}^T\boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\alpha}})}{g(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})}\right)$$
$$- \exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\left(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\alpha}^* - \mathbf{S}_{1i}^T\boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^T\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)}\right)\Bigg\}\mathbf{S}_{1i}^T\boldsymbol{\Delta}.$$

By the fact that $2ab \le \frac{1}{2}a^2 + 2b^2$,

$$|R_4| \le \frac{1}{2}\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + 2R_5,$$

where

$$R_5 = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} \frac{1}{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})}\Bigg\{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\alpha}} - \mathbf{S}_{1i}^T\boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\alpha}})}{g(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})}\right)$$
$$- \exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\left(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\alpha}^* - \mathbf{S}_{1i}^T\boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^T\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)}\right)\Bigg\}^2$$

Note that

$$E_{\mathbb{S}_{\boldsymbol{\beta}}}[R_5] = E\left[\frac{1}{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}(Q_1 + Q_2 + Q_3)^2\right]$$

where

$$Q_1 = \exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\left(1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}\right)\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*),$$

$$Q_2 = \left\{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}\zeta,$$

$$Q_3 = B\left\{\frac{\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\gamma}})}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^T\boldsymbol{\delta}^*)}\right\}\varepsilon.$$

By (a) and (b) of Theorem 4.3, we have $P_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1\}) = 1 - o(1)$.

Then by Hölder's inequality,

$$E_{\mathbb{S}_{\boldsymbol{\beta}}}[R_5] \le \left\|\frac{1}{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}\right\|_{P,2}(\|Q_1\|_{P,4} + \|Q_2\|_{P,4} + \|Q_3\|_{P,4})^2$$

and

$$\begin{aligned}
\|Q_1\|_{P,4} &\le \left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\right\|_{P,12}\left\|\left(1-\frac{B}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}\right)\right\|_{P,12}\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\right\|_{P,12} \\
&\overset{(i)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d}{N}}\right), \\
\|Q_2\|_{P,4} &\le \left\|\left\{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}\right\|_{P,8}\|\zeta\|_{P,8} \overset{(ii)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right), \\
\|Q_3\|_{P,4} &\le \left\|\frac{\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\gamma}})}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\delta}})}-\frac{\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^T\boldsymbol{\delta}^*)}\right\|_{P,8}\|\varepsilon\|_{P,8} \overset{(iii)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d}{N}}\right),
\end{aligned}$$

where (i) and (ii) hold by Lemmas 4.15, 4.16, 4.17 and Lemma 4.6; (iii) holds analogously as in (4.75). Hence,

$$R_5 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d}{N}\right) \tag{4.81}$$

Recall (4.80), we have

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*,\boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_1 + 2R_5.$$

Note that $\|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_1 \le \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_2 \le \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2$. Hence,

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*,\boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_1 + 2|R_5|.$$

Recall (4.66), we have $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*,\boldsymbol{\Delta}) \ge 0$. Then,

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + \frac{2R_3}{\lambda_{\boldsymbol{\beta}}} \tag{4.82}$$

Then, by Lemma 4.1, $P_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}\cup\mathbb{S}_{\boldsymbol{\beta}}}(\mathcal{A}_7) \ge 1 - o(1) - c_1\exp(-c_2M) = 1 - o(1)$, where $\mathcal{A}_7$ is defined in (4.79). The remaining parts of the proof can be shown analogously as (c) of Theorem 4.3. Now, condition on $\mathcal{A}_6 \cap \mathcal{A}_7$, for large enough N,

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_2 &\le \frac{8\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{16R_3^2\frac{\kappa_2\log d_1}{\kappa_1 M\lambda_{\boldsymbol{\beta}}^2} + \frac{4R_3}{\kappa_1}} \\
&= O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d + s_{\boldsymbol{\beta}}\log d_1}{N}}\right).
\end{aligned}$$

Recall (4.82), we have

$$\|\boldsymbol{\Delta}\|_1 = O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{\log d_1}{N}} + s_{\boldsymbol{\delta}}\sqrt{\frac{(\log d)^2}{N\log d_1}} + s_{\boldsymbol{\alpha}}\sqrt{\frac{(\log d)^2}{N\log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

■

Proof of Lemma 4.3. By Lemma 4.2, we have

$$\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right), \tag{4.83}$$

$$\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right), \tag{4.84}$$

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right). \tag{4.85}$$

(a) Let $\rho(\cdot) = \rho^*(\cdot)$. Note that

$$\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\mathbf{W}_{\boldsymbol{\delta},i},$$

where

$$\mathbf{W}_{\boldsymbol{\delta},i} := A_{1i}\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\}\{1 - A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_{2i}.$$

Let $\mathbf{W}_{\boldsymbol{\delta}}$ be an indepenent copy of $\mathbf{W}_{\boldsymbol{\delta},i}$. Then, by the tower rule,

$$E(\mathbf{W}_{\boldsymbol{\delta}}) = \mathbf{0} \in \mathbb{R}^d.$$

By Corollary 2.3 of Lemma 4.7, we have

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\mathbf{W}_{\boldsymbol{\delta},i}\right\|_\infty^2\right) &\le M^{-1}(2e\log d - e)E(\|\mathbf{W}_{\boldsymbol{\delta}}\|_\infty^2) \\
&\le (1 + c_0^{-1})M^{-1}(2e\log d - e)E\left\{\left|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right|^2\|\bar{\mathbf{S}}_2\|_\infty^2\right\} \\
&\le (1 + c_0^{-1})M^{-1}(2e\log d - e)\left\|g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,4}^2\left\|\|\bar{\mathbf{S}}_2\|_\infty\right\|_{P,4}^2 \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d)^2}{N^2}\right),
\end{aligned}$$

where (i) holds by Lemma 4.15 and Lemma 4.6. By Lemma 4.4,

$$\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*)-\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\right\|_\infty=\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\mathbf{W}_{\boldsymbol{\delta},i}\right\|_\infty=O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1}\log d}{N}\right).$$

Together with (4.83), we have

$$\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*)\right\|_\infty=O_p\left(\left(1+\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1\log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

The remaining parts of the proof can be shown analogously as in (a).

(b) Let $\nu(\cdot)=\nu^*(\cdot)$. Note that

$$\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*)-\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}\mathbf{W}_{\boldsymbol{\alpha},i},$$

where

$$\mathbf{W}_{\boldsymbol{\alpha},i}:=-2A_{1i}A_{2i}\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\exp(-\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})-g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\}\varepsilon_i\bar{\mathbf{S}}_{2i}.$$

Let $\mathbf{W}_{\boldsymbol{\alpha}}$ be an indepenent copy of $\mathbf{W}_{\boldsymbol{\alpha},i}$. Then, by the tower rule,

$$E(\mathbf{W}_{\boldsymbol{\alpha}})=\mathbf{0}\in\mathbb{R}^d.$$

By Lemma 4.7, we have

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}\mathbf{W}_{\boldsymbol{\alpha},i}\right\|_\infty^2\right)&\le M^{-1}(2e\log d-e)E(\|\mathbf{W}_{\boldsymbol{\alpha}}\|_\infty^2)\\
&\le 2M^{-1}(2e\log d-e)E\left\{\left|\frac{\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}-\frac{\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right|^2\varepsilon^2\|\bar{\mathbf{S}}_2\|_\infty^2\right\}\\
&\le 2M^{-1}(2e\log d-e)\left\|\frac{\exp(-\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}-\frac{\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}{g(\mathbf{S}_1^T\boldsymbol{\gamma}^*)}\right\|_{P,6}^2\left\|\|\varepsilon\|_{P,6}^2\|\|\bar{\mathbf{S}}_2\|_\infty\right\|_{P,6}^2\\
&\overset{(i)}{=}O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d)(\log d)^2}{N^2}\right),
\end{aligned}$$

where (i) holds by Lemma 4.6 and (4.75). By Lemma 4.4,

$$\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*)-\nabla_{\boldsymbol{\delta}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1}\log d}{N}\right).$$

Together with (4.84), we have

$$\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\left(1+\sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d)\log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

(c) Let $\nu(\cdot)=\nu^*(\cdot)$ and $\mu(\cdot)=\mu^*(\cdot)$. Note that

$$\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}(\mathbf{W}_{\boldsymbol{\beta},1,i}+\mathbf{W}_{\boldsymbol{\beta},2,i}),$$

where

$$\begin{aligned}\mathbf{W}_{\boldsymbol{\beta},1,i} :=& -2A_{1i}\{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})-\exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\}\zeta_i\mathbf{S}_{1i},\\ \mathbf{W}_{\boldsymbol{\beta},2,i} :=& -2A_{1i}A_{2i}\{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})g^{-1}(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})-\exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\}\varepsilon_i\mathbf{S}_{1i}.\end{aligned}$$

Let $\mathbf{W}_{\boldsymbol{\beta},1}$ and $\mathbf{W}_{\boldsymbol{\beta},2}$ be indepenent copies of $\mathbf{W}_{\boldsymbol{\beta},1,i}$ and $\mathbf{W}_{\boldsymbol{\beta},1,i}$, respectively. Then, by the tower rule,

$$E(\mathbf{W}_{\boldsymbol{\beta},1})=E(\mathbf{W}_{\boldsymbol{\beta},2})=\mathbf{0}\in\mathbb{R}^{d_1}.$$

By Lemma 4.7, we have

$$\begin{aligned}E_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},1,i}\right\|_\infty^2\right) &\le M^{-1}(2e\log d_1-e)E(\|\mathbf{W}_{\boldsymbol{\beta},1}\|_\infty^2)\\ &\le 4M^{-1}(2e\log d_1-e)E\left\{\left|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-\exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right|^2\zeta^2\|\mathbf{S}_1\|_\infty^2\right\}\\ &\le 4M^{-1}(2e\log d_1-e)\left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-\exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,6}^2\left\|\zeta\right\|_{P,6}^2\|\|\mathbf{S}_1\|_\infty\|_{P,6}^2\\ &\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d_1)^2}{N^2}\right),\end{aligned}$$

where (i) holds by Lemma 4.15 and Lemma 4.6. Similarly, we also have

$$
\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},2,i}\right\|_{\infty}^{2}\right) &\leq M^{-1}(2e\log d_1 - e)E(\|\mathbf{W}_{\boldsymbol{\beta},2}\|_{\infty}^{2}) \\
&\leq 4M^{-1}(2e\log d_1 - e)E\left\{\left|\frac{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right|^2 \varepsilon^2\|\mathbf{S}_1\|_{\infty}^2\right\} \\
&\leq 4M^{-1}(2e\log d_1 - e)\left\|\frac{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right\|_{P,6}^2 \|\varepsilon\|_{P,6}^2\|\|\mathbf{S}_1\|_{\infty}\|_{P,6}^2 \\
&\overset{(i)}{=} O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d)(\log d_1)^2}{N^2}\right),
\end{aligned}
$$

where (i) holds by Lemma 4.6, and analogously as in (4.75),

$$
\left\|\frac{\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)}\right\|_{P,6} = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d}{N}}\right).
$$

Hence, it follows that

$$
\begin{aligned}
&E_{\mathbb{S}_{\boldsymbol{\beta}}}\left\{\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\alpha}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_{\infty}^2\right\} \\
&\quad = O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d)(\log d_1)^2}{N^2}\right).
\end{aligned}
$$

By Lemma 4.4,

$$
\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_{\infty} = O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d}\log d_1}{N}\right).
$$

Together with (4.85), we have

$$
\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_{\infty} = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d)\log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).
$$

(d) Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Note that

$$
\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}(\mathbf{W}_{\boldsymbol{\beta},3,i} + \mathbf{W}_{\boldsymbol{\beta},4,i}),
$$

where

$$\mathbf{W}_{\boldsymbol{\beta},3,i} := -2A_{1i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left\{1-\frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)}\right\}\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\mathbf{S}_{1i},$$
$$\mathbf{W}_{\boldsymbol{\beta},4,i} := -2A_{1i}\{\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})-\exp(-\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\}\left\{\frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)}\varepsilon_i+\zeta_i\right\}\mathbf{S}_{1i}. \tag{4.86}$$

Let $\mathbf{W}_{\boldsymbol{\beta},3}$ and $\mathbf{W}_{\boldsymbol{\beta},4}$ be independent copies of $\mathbf{W}_{\boldsymbol{\beta},3,i}$ and $\mathbf{W}_{\boldsymbol{\beta},4,i}$, respectively. Then, by the tower rule,

$$E(\mathbf{W}_{\boldsymbol{\beta},3}) = E(\mathbf{W}_{\boldsymbol{\beta},4}) = \mathbf{0} \in \mathbb{R}^{d_1}.$$

By Lemma 4.7, we have

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},3,i}\right\|_\infty^2\right) &\le M^{-1}(2e\log d_1 - e)E(\|\mathbf{W}_{\boldsymbol{\beta},3}\|_\infty^2)\\
&\le 4M^{-1}(2e\log d_1 - e)(1+c_0^{-1})^2E\left\{\exp(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\left\{\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\right\}^2\|\mathbf{S}_1\|_\infty^2\right\}\\
&\le 4M^{-1}(2e\log d_1 - e)(1+c_0^{-1})^2\left\|\exp(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\right\|_{P,3}^2\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\right\|_{P,6}^2\|\|\mathbf{S}_1\|_\infty\|_{P,6}^2\\
&\overset{(i)}{=} O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d)(\log d_1)^2}{N^2}\right),
\end{aligned}$$

where (i) holds by Lemmas 4.15, 4.17 and Lemma 4.6. Similarly, we also have

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},4,i}\right\|_\infty^2\right) &\le M^{-1}(2e\log d_1 - e)E(\|\mathbf{W}_{\boldsymbol{\beta},4}\|_\infty^2)\\
&\le 4M^{-1}(2e\log d_1 - e)E\left[\left\{\exp(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-\exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}^2\left(c_0^{-1}|\varepsilon|+|\zeta|\right)^2\|\mathbf{S}_1\|_\infty^2\right]\\
&\le 8M^{-1}(2e\log d_1 - e)\left\|\exp(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-\exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,6}^2\\
&\qquad\cdot\left(c_0^{-2}\|\varepsilon\|_{P,6}^2+\|\zeta\|_{P,6}^2\right)\|\|\mathbf{S}_1\|_\infty\|_{P,6}^2\\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d_1)^2}{N^2}\right),
\end{aligned} \tag{4.87}$$

where (i) holds by Lemmas 4.15 and Lemma 4.6. Hence, it follows that

$$E_{\mathbb{S}_{\boldsymbol{\beta}}}\left\{\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty^2\right\}$$
$$=O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d+s_{\boldsymbol{\alpha}}\log d)(\log d_1)^2}{N^2}\right).$$

By Lemma 4.4,

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty$$
$$=O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d+s_{\boldsymbol{\alpha}}\log d}\log d_1}{N}\right).$$

Together with (4.85), we have

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)\right\|_\infty$$
$$=O_p\left(\left(1+\sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d+s_{\boldsymbol{\alpha}}\log d)\log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

(e) Let $\rho(\cdot)=\rho^*(\cdot)$, $\nu(\cdot)=\nu^*(\cdot)$, and $\mu(\cdot)=\mu^*(\cdot)$. Note that

$$\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},4,i},$$

where $\mathbf{W}_{\boldsymbol{\beta},4,i}$ is defined in (4.86). By (4.87) and Lemma 4.4,

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty=O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1}\log d_1}{N}\right).$$

Together with (4.85), we have

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)\right\|_\infty=O_p\left(\left(1+\sqrt{\frac{s_{\boldsymbol{\gamma}}(\log d_1)^2}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

■

Proof of Theorem 4.4. We show the consistency rate of the nuisance estimators under correctly specified models.

(a) Let $\rho(\cdot) = \rho^*(\cdot)$. Then, by Lemma 4.3, when $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$,

$$\left\|\nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right).$$

By Lemma 4.1, we have (4.41) when $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$. In addition, by Lemma 4.15, we also have $P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1) = 1 - o(1)$. By Corollary 9.20 of [Wai19], we have

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right), \quad \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_1 = O_p\left(s_{\boldsymbol{\delta}}\sqrt{\frac{\log d}{N}}\right).$$

(b) Let $\nu(\cdot) = \nu^*(\cdot)$. Then, by Lemma 4.3, when $s_{\boldsymbol{\gamma}} = O(\frac{N}{\log d_1 \log d})$ and $s_{\boldsymbol{\delta}} = O(\frac{N}{(\log d)^2})$,

$$\left\|\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right).$$

By Lemma 4.1, we have (4.43) when $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$ and $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1$. In addition, by Lemmas 4.15 and 4.16, we also have $P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1 \cap \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1) = 1 - o(1)$. By Corollary 9.20 of [Wai19], we have

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\right), \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 = O_p\left(s_{\boldsymbol{\alpha}}\sqrt{\frac{\log d}{N}}\right).$$

(c) Let $\nu(\cdot) = \nu^*(\cdot)$ and $\mu(\mathbf{S}_1) = \mathbf{S}_1^T \boldsymbol{\beta}$. Then, by Lemma 4.3, when $s_{\boldsymbol{\gamma}} = O(\frac{N}{(\log d_1)^2})$ and $s_{\boldsymbol{\delta}} = O(\frac{N}{\log d_1 \log d})$,

$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right).$$

That is, for any $t > 0$, there exists some $\lambda_3 \asymp \sqrt{\frac{\log d_1}{N}}$, such that $\mathcal{E}_3 := \{\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\|_\infty \leq \lambda_3\}$ holds with probability at least $1 - t$. Condition on the event $\mathcal{E}_3$, and choose some $\lambda_{\boldsymbol{\beta}} > 2\lambda_3$. By the construction of $\boldsymbol{\beta}$, we have

$$\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) + \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}\|_1 \leq \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*\|_1.$$

Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ and $S = \{j \in \{1, \ldots, d_1\} : \boldsymbol{\beta}_j^* \neq 0\}$. Note that,

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)^T\boldsymbol{\Delta}.$$

Hence,

$$\begin{aligned}
&\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}\|_1 \leq -\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)^T\boldsymbol{\Delta} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*\|_1 \\
&\quad \leq \left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty \|\boldsymbol{\Delta}\|_1 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*\|_1 + |R_6| \leq \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*\|_1 + |R_6|,
\end{aligned}$$

where

$$R_6 := \left\{\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\}^T \boldsymbol{\Delta}.$$

Note that $\|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}_S^*\|_1 \leq \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\boldsymbol{\Delta}_S\|_1$, $\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 = \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1$, and $\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_S\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1$. Hence, we have

$$2\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_S\|_1 + 2|R_6|.$$

Observe that

$$\begin{aligned}
|R_6| &= \left|2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left\{1 - \frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})}\right\}\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\mathbf{S}_{1i}^T\boldsymbol{\Delta}\right| \\
&\leq \frac{\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_7,
\end{aligned}$$

where $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^T\boldsymbol{\Delta})^2$ and

$$R_7 := 2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left\{1 - \frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}})}\right\}^2\left\{\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2.$$

It follows that

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_S\|_1 + 2R_7. \tag{4.88}$$

Condition on $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$, where by Lemma 4.15, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$ holds with probability $1 - o(1)$. Also, condition on the event that

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d_1}{M} \|\boldsymbol{\Delta}\|_1^2, \tag{4.89}$$

which, by Lemma 4.1, holds with probability $1 - o(1)$. Since $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \geq 0$, by (4.88), we have

$$\|\boldsymbol{\Delta}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1 + 2\lambda_{\boldsymbol{\beta}}^{-1} R_7. \tag{4.90}$$

Note that $\|\boldsymbol{\Delta}_S\|_1 \leq \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}_S\|_2 \leq \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2$. Together with (4.88) and (4.89),

$$\begin{aligned}
3\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2R_7 &\geq 3\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_S\|_1 + 2R_7 \geq \delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \\
&\geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d_1}{M}\|\boldsymbol{\Delta}\|_1^2 \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d_1}{M}\left(4\|\boldsymbol{\Delta}_S\|_1 + 2\lambda_{\boldsymbol{\beta}}^{-1}R_7\right)^2 \\
&\geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - 4\kappa_2\frac{\log d_1}{M}\left(4\|\boldsymbol{\Delta}_S\|_1^2 + \lambda_{\boldsymbol{\beta}}^{-2}R_7^2\right) \\
&\geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - 4\kappa_2\frac{\log d_1}{M}\left(4s_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_2^2 + \lambda_{\boldsymbol{\beta}}^{-2}R_7^2\right) \geq \frac{\kappa_1}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{4\kappa_2\log d_1}{\lambda_{\boldsymbol{\beta}}^2 M}R_7^2,
\end{aligned}$$

when $M > 32\kappa_2 s_{\boldsymbol{\beta}}\log d_1/\kappa_1$. By Lemma 4.14,

$$\|\boldsymbol{\Delta}\|_2 \leq \frac{6\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{\frac{8\kappa_2 R_7^2\log d_1}{\kappa_1\lambda_{\boldsymbol{\beta}}^2 M} + \frac{4R_7}{\kappa_1}}. \tag{4.91}$$

Now, we upper bound the term R_7. Observe that

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\beta}}}(R_7) &= 2E\left[\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}\right\}^2\left\{\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2\right] \\
&\leq 2\left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\right\|_{P,3}\left\|1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})}\right\|_{P,6}^2\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{P,6}^2 \\
&\leq 2\left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\right\|_{P,3}\left\{1 + \left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}})\right\|_{P,6}\right\}^2\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{P,6}^2 \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\alpha}}\log d}{N}\right),
\end{aligned}$$

where (i) holds by Lemmas 4.15, 4.16, and 4.17. By Lemma 4.4,

$$R_7 = O_p\left(\frac{s_{\boldsymbol{\alpha}} \log d}{N}\right).$$

By (4.91) and $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\frac{\log d_1}{N}}$, we have

$$\|\boldsymbol{\Delta}\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}} + \frac{s_{\boldsymbol{\alpha}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d + s_{\boldsymbol{\beta}} \log d_1}{N}}\right).$$

By (4.90),

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2\lambda_{\boldsymbol{\beta}}^{-1} R_7 = O_p\left(s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}} + s_{\boldsymbol{\gamma}}\sqrt{\frac{(\log d)^2}{N \log d_1}}\right).$$

(d) Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Then, by Lemma 4.3, when $s_{\boldsymbol{\gamma}} = o(\frac{N}{(\log d_1)^2})$, $s_{\boldsymbol{\delta}} = o(\frac{N}{\log d_1 \log d})$, and $s_{\boldsymbol{\alpha}} = o(\frac{N}{\log d_1 \log d})$,

$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right).$$

That is, for any $t > 0$, there exists some $\lambda_4 \asymp \sqrt{\frac{\log d_1}{N}}$, such that

$$\mathcal{E}_4 := \{\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\|_\infty \le \lambda_4\}$$

holds with probability at least $1 - t$. Condition on the event $\mathcal{E}_4$, and choose some $\lambda_{\boldsymbol{\beta}} > 2\lambda_4$. Similarly as in part (c), we obtain

$$2\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{S^c}\|_1 \le 3\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_S\|_1 + 2|R_8|,$$

where

$$\begin{aligned}
|R_8| &= \left|\left\{\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\}^T \boldsymbol{\Delta}\right| \\
&= \left|2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} A_{2i} \exp(-\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}}) \left\{g^{-1}(\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*)\right\} \widehat{\varepsilon}_i \mathbf{S}_{1i}^T \boldsymbol{\Delta}\right| \\
&\le \frac{\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_9.
\end{aligned}$$

Here, $\widehat{\varepsilon}_i := Y_i(1,1) - \bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\alpha}}$,

$$\begin{aligned}\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) =& M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \exp(-\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}}) (\mathbf{S}_{1i}^T \boldsymbol{\Delta})^2, \\ R_9 :=& 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{2i} \exp(-\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*) \right\}^2 \widehat{\varepsilon}_i^2.\end{aligned}$$

Observe that

$$\begin{aligned}E_{\mathbb{S}_{\boldsymbol{\beta}}}(R_9) &= 2E \left[A_2 \exp(-\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \right\}^2 \widehat{\varepsilon}^2 \right] \\ &\leq 2 \left\| \exp(-\mathbf{S}_1^T \widehat{\boldsymbol{\gamma}}) \right\|_{P,3} \left\| g^{-1}(\bar{\mathbf{S}}_2^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \right\|_{P,6}^2 \|\widehat{\varepsilon}\|_{P,6}^2 \\ &\overset{(i)}{=} O_p \left(\frac{s_{\boldsymbol{\delta}} \log d}{N} \right),\end{aligned}$$

where (i) holds by Lemmas 4.15, 4.16, and 4.17. By Lemma 4.4,

$$R_9 = O_p \left(\frac{s_{\boldsymbol{\delta}} \log d}{N} \right).$$

Repeat the same procedure as in part (c), we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2 &\leq \frac{6\lambda_{\boldsymbol{\beta}} \sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{\frac{8\kappa_2 R_9^2 \log d_1}{\kappa_1 \lambda_{\boldsymbol{\beta}}^2 M} + \frac{4R_9}{\kappa_1}}, \\ \|\boldsymbol{\Delta}\|_1 &\leq 4\sqrt{s_{\boldsymbol{\beta}}} \|\boldsymbol{\Delta}\|_2 + 2\lambda_{\boldsymbol{\beta}}^{-1} R_9,\end{aligned}$$

with probability at least $1 - t - o(1)$. Hence,

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2 &= O_p \left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\beta}} \log d_1}{N}} \right), \\ \|\boldsymbol{\Delta}\|_1 &= O_p \left(s_{\boldsymbol{\delta}} \sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}} \sqrt{\frac{\log d_1}{N}} \right).\end{aligned}$$

(e) Let $\rho(\cdot) = \rho^*(\cdot)$, $\nu(\cdot) = \nu^*(\cdot)$, and $\mu(\cdot) = \mu^*(\cdot)$. Then, by Lemma 4.3, when $s_{\boldsymbol{\gamma}} = o(\frac{N}{(\log d_1)^2})$, $s_{\boldsymbol{\delta}} = o(\frac{N}{\log d_1 \log d})$, and $s_{\boldsymbol{\alpha}} = o(\frac{N}{\log d_1 \log d})$,

$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right),$$
$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right),$$
$$\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right).$$

Define

$$\mathbf{a} := \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) + \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*).$$

Then, $\|\mathbf{a}\|_\infty = O_p(\sqrt{\frac{\log d_1}{N}})$. Hence, for any $t > 0$, there exists some $\lambda_5 \asymp \sqrt{\frac{\log d_1}{N}}$, such that $\mathcal{E}_5 := \{\|\mathbf{a}\|_\infty \le \lambda_5\}$ holds with probability at least $1 - t$. Condition on the event $\mathcal{E}_5$, and choose some $\lambda_{\boldsymbol{\beta}} > 2\lambda_5$. Similarly as in parts (c) and (d), we obtain

$$2\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_{S^c}\|_1 \le 3\lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_S\|_1 + 2|R_{10}|,$$

where

$$\begin{aligned}
R_{10} &= \left\{\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \mathbf{a}\right\}^T \boldsymbol{\Delta} \\
&= \left\{\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\}^T \boldsymbol{\Delta} \\
&\quad - \left\{\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\}^T \boldsymbol{\Delta} \\
&= 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} A_{2i} \exp(-\mathbf{S}_{1i}^T \widehat{\boldsymbol{\gamma}}) \left\{g^{-1}(\bar{\mathbf{S}}_{2i}^T \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^T \boldsymbol{\delta}^*)\right\} \bar{\mathbf{S}}_{2i}^T (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \mathbf{S}_{1i}^T \boldsymbol{\Delta}.
\end{aligned}$$

By Young's inequality for products,

$$|R_{10}| \le \frac{\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_{11},$$

where $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^T\boldsymbol{\Delta})^2$ and

$$R_{11} := 2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{2i}\exp(-\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})\left\{g^{-1}(\bar{\mathbf{S}}_{2i}^T\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\right\}^2\left\{\bar{\mathbf{S}}_{2i}^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2.$$

Observe that

$$\begin{aligned}
E_{\mathbb{S}_{\boldsymbol{\beta}}}(R_{11}) &= 2E\left[A_2\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\left\{g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\}^2\left\{\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2\right] \\
&\le 2\left\|\exp(-\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})\right\|_{P,3}\left\|g^{-1}(\bar{\mathbf{S}}_2^T\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,6}^2\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{P,6}^2 \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\delta}}s_{\boldsymbol{\alpha}}(\log d)^2}{N^2}\right),
\end{aligned}$$

where (i) holds by Lemmas 4.15, 4.16, and 4.17. By Lemma 4.4,

$$R_{11} = O_p\left(\frac{s_{\boldsymbol{\delta}}s_{\boldsymbol{\alpha}}(\log d)^2}{N^2}\right).$$

Repeat the same procedure as in parts (c) and (d), we have

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_2 &\le \frac{6\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{\frac{8\kappa_2 R_{11}^2\log d_1}{\kappa_1\lambda_{\boldsymbol{\beta}}^2 M} + \frac{4R_{11}}{\kappa_1}}, \\
\|\boldsymbol{\Delta}\|_1 &\le 4\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2\lambda_{\boldsymbol{\beta}}^{-1}R_{11},
\end{aligned}$$

with probability at least $1 - t - o(1)$. Hence,

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_2 &= O_p\left(\frac{\sqrt{s_{\boldsymbol{\delta}}s_{\boldsymbol{\alpha}}}\log d}{N} + \sqrt{\frac{s_{\boldsymbol{\beta}}\log d_1}{N}}\right), \\
\|\boldsymbol{\Delta}\|_1 &= O_p\left(\frac{s_{\boldsymbol{\delta}}s_{\boldsymbol{\alpha}}\log d}{N}\sqrt{\frac{(\log d)^2}{N\log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).
\end{aligned}$$

■

4.5.3 Proof of auxiliary lemmas

Proof of Lemma 4.10. We prove the lemma by considering two cases separately.

(a) If $d \le m$. Choose $S = \{1, \ldots, d\}$. Since $\mathbf{X}$ is a sub-Gaussian vector, we have

$$\sup_{\|\boldsymbol{\beta}\|_2=1} E\{(\mathbf{X}^T\boldsymbol{\beta})^2\} = O(1). \tag{4.92}$$

For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, by triangle inequality, we have

$$\begin{aligned} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2 &\le \|\boldsymbol{\Delta}\|_2^2 \sup_{\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2 \\ &\overset{(i)}{\le} \|\boldsymbol{\Delta}\|_2^2 \left[\sup_{\|\boldsymbol{\beta}\|_2=1} E\{(\mathbf{X}^T\boldsymbol{\beta})^2\} + \sup_{\|\boldsymbol{\beta}\|_2=1} \left| m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2 - E\{(\mathbf{X}^T\boldsymbol{\beta})^2\} \right| \right] \end{aligned}$$

It follows that

$$\begin{aligned} &\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2} \\ &\quad\le \left[\sup_{\|\boldsymbol{\beta}\|_2=1} E\{(\mathbf{X}^T\boldsymbol{\beta})^2\} + \sup_{\|\boldsymbol{\beta}\|_2=1} \left| m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2 - E\{(\mathbf{X}^T\boldsymbol{\beta})^2\} \right| \right] \end{aligned}$$

By Lemma 4.9 and (4.92), we have, as $m \to \infty$,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2} = O_p\left(1 + \frac{d}{m}\right) \overset{(i)}{=} O_p(1)$$

where (i) holds since $d \le m$. Hence,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} \le \sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2} = O_p(1)$$

(b) If $m > d$. Choose any set $S \subseteq \{1, \ldots, d\}$ such that $s := |S| \asymp m$. For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, define $\widetilde{\boldsymbol{\Delta}} = (\widetilde{\boldsymbol{\Delta}}_S^T, \widetilde{\boldsymbol{\Delta}}_{S^c}^T)^T \in \mathbb{R}^d$ such that

$$\widetilde{\boldsymbol{\Delta}}_S = s^{-1}\|\boldsymbol{\Delta}\|_1(1, \ldots, 1)^T \in \mathbb{R}^s, \quad \widetilde{\boldsymbol{\Delta}}_{S^c} = \boldsymbol{\Delta}_{S^c} \in \mathbb{R}^{d-s}.$$

Then,

$$\|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 = \|\boldsymbol{\Delta}_{S^c}\|_1 \le \|\boldsymbol{\Delta}\|_1 = \|\widetilde{\boldsymbol{\Delta}}_S\|_1.$$

Hence, $\widetilde{\boldsymbol{\Delta}} \in \mathbb{C}(S,3) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \le 3\|\boldsymbol{\Delta}_S\|_1\}$. In addition, since $(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta})_{S^c} = \mathbf{0} \in \mathbb{R}^{d-s}$, we also have $\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta} \in \mathbb{C}(S,3)$. Therefore, by the fact that $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\begin{aligned} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2 &\le 2m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\widetilde{\boldsymbol{\Delta}})^2 + 2m^{-1}\sum_{i=1}^m \left\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta})\right\}^2 \\ &\le 2\left(\|\widetilde{\boldsymbol{\Delta}}\|_2^2 + \|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_2^2\right) \sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2. \end{aligned}$$

Now, we observe that

$$\begin{aligned} \|\widetilde{\boldsymbol{\Delta}}\|_2^2 &= \|\widetilde{\boldsymbol{\Delta}}_S\|_2^2 + \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_2^2 = s^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}_{S^c}\|_2^2, \\ \|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_2^2 &= \|\widetilde{\boldsymbol{\Delta}}_S - \boldsymbol{\Delta}_S\|_2^2 \le 2\|\widetilde{\boldsymbol{\Delta}}_S\|_2^2 + 2\|\boldsymbol{\Delta}_S\|_2^2 = 2s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 2\|\boldsymbol{\Delta}_S\|_2^2. \end{aligned}$$

Hence, we have

$$m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2 \le 2\left(3s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 2\|\boldsymbol{\Delta}\|_2^2\right) \sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d,$$

since $\|\widetilde{\boldsymbol{\Delta}}\|_2^2 + \|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_2^2 \le 3s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 2\|\boldsymbol{\Delta}\|_2^2$. It follows that

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{6s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 4\|\boldsymbol{\Delta}\|_2^2} \le \sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\beta})^2,$$

By Lemma 4.9 and (4.92), as $m \to \infty$,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{s^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p\left(1 + \sqrt{\frac{s}{m}}\right).$$

Besides, note that $s \asymp m$ and hence $1 + \sqrt{s/m} = O(1)$. It follows that

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^T\boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

■

Proof of Lemma 4.11. Note that

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}) :=& \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta} - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1 \\ =& \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta} + R_1(\boldsymbol{\Delta}) - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1, \qquad (4.93)\end{aligned}$$

where

$$\begin{aligned}R_1(\boldsymbol{\Delta}) :=& \left\{\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\}^T\boldsymbol{\Delta} \\ =& M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}\left\{1 - A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\right\}\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta}.\end{aligned}$$

Let $\lambda_{\boldsymbol{\delta}} > 2\sigma_{\boldsymbol{\delta}}\sqrt{\frac{t+\log d}{M}}$ with some $t > 0$. By Lemma 4.2, we have $P_{\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1) \geq 1 - 2\exp(-t)$. On the event $\mathcal{A}_1$, we have $|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta}| \leq \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1/2$. Note that $\|\boldsymbol{\delta}^*\|_1 = \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}}\|_1 \leq \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}} + \boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1$, $\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}^c}\|_1$, and $\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 = \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}} + \boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}^c}\|_1$. Recall (4.93), it follows that

$$2\mathcal{F}(\boldsymbol{\Delta}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}^c}\|_1 - 3\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta})|.$$

Hence,

$$2\mathcal{F}(\boldsymbol{\Delta}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1 - 4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta})|. \qquad (4.94)$$

Under the overlap condition in Assumption 4.1 and since $|A_1| \leq 1$,

$$\begin{aligned}|R_1(\boldsymbol{\Delta})| \leq& (1 + c_0^{-1})M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta} \\ \overset{(i)}{\leq}& (1 + c_0^{-1})\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}^2}\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2},\end{aligned}$$

where (i) holds by the Cauchy–Schwarz inequality. It follows that

$$\begin{aligned}&\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{|R_1(\boldsymbol{\Delta})|}{\|\boldsymbol{\Delta}\|_1/\sqrt{N} + \|\boldsymbol{\Delta}\|_2} \\ &\leq (1 + c_0^{-1})\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}^2}\sqrt{\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2}{N^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2}},\end{aligned}$$

since $(\|\boldsymbol{\Delta}\|_1/\sqrt{N}+\|\boldsymbol{\Delta}\|_2)^2 > N^{-1}\|\boldsymbol{\Delta}\|_1^2+\|\boldsymbol{\Delta}\|_2^2$. Note that

$$E_{\mathbb{S}_{\boldsymbol{\delta}}}\left[M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})-g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}^2\right] = E\left[\left\{g^{-1}(\mathbf{S}_1^T\widehat{\boldsymbol{\gamma}})-g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\}^2\right]$$
$$\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}\right),$$

where (i) holds by Lemma 4.15. By Lemma 4.4,

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}})-g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}\right).$$

Besides, by Lemma 4.10, we also have

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta})^2}{N^{-1}\|\boldsymbol{\Delta}\|_1^2+\|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

Hence,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{|R_1(\boldsymbol{\Delta})|}{\|\boldsymbol{\Delta}\|_1/\sqrt{N}+\|\boldsymbol{\Delta}\|_2} = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right).$$

That is, with any $t>0$, there exists some constant $c>0$, such that $P_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_2)\geq 1-t$. Hence,

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1\cap\mathcal{A}_2)\geq 1-t-2\exp(-t).$$

Now, condition on $\mathcal{A}_1\cap\mathcal{A}_2$, we have

$$2\mathcal{F}(\boldsymbol{\Delta})\geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta})+\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1-4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1-2c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\left(\frac{\|\boldsymbol{\Delta}\|_1}{\sqrt{N}}+\|\boldsymbol{\Delta}\|_2\right).$$

With some $\lambda_{\boldsymbol{\delta}}\asymp\sqrt{\frac{\log d}{N}}$, since $d_1\leq d$ and $s_{\boldsymbol{\gamma}}=o(N)$, we have $\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N^2}}=o(\lambda_{\boldsymbol{\delta}})$. Hence, with some $N_0>0$, when $N>N_0$, we have $4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N^2}}\leq\lambda_{\boldsymbol{\delta}}$. It follows that

$$4\mathcal{F}(\boldsymbol{\Delta})\geq 4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta})+\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1-8\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1-4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\|\boldsymbol{\Delta}\|_2.$$

Note that $\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 \leq \sqrt{s_{\boldsymbol{\delta}}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_2 \leq \sqrt{s_{\boldsymbol{\delta}}}\|\boldsymbol{\Delta}\|_2$. Hence,

$$4\mathcal{F}(\boldsymbol{\Delta}) \geq 4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1 - \left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} + 4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right)\|\boldsymbol{\Delta}\|_2. \tag{4.95}$$

For any $\boldsymbol{\Delta} \in \widetilde{K}(\bar{s}_{\boldsymbol{\delta}}, k_0, 1)$, we have

$$4\mathcal{F}(\boldsymbol{\Delta}) \geq 4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1 - \left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} + 4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right),$$

on the event $\mathcal{A}_1 \cap \mathcal{A}_2$ and when $N > N_0$. Here, on the event $\mathcal{A}_3$, we have

$$\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2 \overset{(i)}{\geq} \kappa_1 - \kappa_2 k_0^2\frac{\bar{s}_{\boldsymbol{\delta}}\log d}{M},$$

where (i) holds since $\boldsymbol{\Delta} \in \widetilde{K}(\bar{s}_{\boldsymbol{\delta}}, k_0, 1)$. Therefore, condition on the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$\mathcal{F}(\boldsymbol{\Delta}) \geq \kappa_1 - \kappa_2 k_0^2\frac{\bar{s}_{\boldsymbol{\delta}}\log d}{M} - 2\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} - \frac{c}{2}\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}} \geq \kappa_1/2,$$

when $N > N_1$ with some constant $N_1 > 0$, since $\frac{\bar{s}_{\boldsymbol{\delta}}\log d}{M} = \frac{s_{\boldsymbol{\gamma}}\log d_1}{M} + \frac{s_{\boldsymbol{\delta}}\log d}{M} = o(1)$, $\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} \asymp \sqrt{\frac{s_{\boldsymbol{\delta}}\log d}{N}} = o(1)$, and $\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}} = o(1)$. ∎

Proof of Lemma 4.12. Based on the construction of $\widehat{\boldsymbol{\delta}}$, we have

$$\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\widehat{\boldsymbol{\delta}}\|_1 \leq \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1.$$

By definition (4.38), we have $\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) = \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}) - \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta}_{\boldsymbol{\delta}}$. It follows that

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) &= \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\widehat{\boldsymbol{\delta}}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta}_{\boldsymbol{\delta}} - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1 \\
&= \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^T\boldsymbol{\Delta}_{\boldsymbol{\delta}} + R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1 \leq 0,
\end{aligned}$$

where

$$\begin{aligned}
R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) &= \left\{\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\}^T\boldsymbol{\Delta}_{\boldsymbol{\delta}} \\
&= M^{-1}\sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i}\left\{g^{-1}(\mathbf{S}_{1i}^T\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^T\boldsymbol{\gamma}^*)\right\}\left\{1 - A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^T\boldsymbol{\delta}^*)\right\}\bar{\mathbf{S}}_{2i}^T\boldsymbol{\Delta}_{\boldsymbol{\delta}}.
\end{aligned}$$

Repeat the same procedure in the proof of Lemma 4.12 for obtaining (4.94) and (4.95). Then, condition on $\mathcal{A}_1$, we have

$$0 \geq 2\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 - 4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta},S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}})|. \tag{4.96}$$

Condition on $\mathcal{A}_1 \cap \mathcal{A}_2$, we further have

$$0 \geq 4\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) \geq 4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 - \left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} + 4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right)\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2.$$

Hence,

$$4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \leq \left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}} + 4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right)\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2.$$

Recall (4.64), we have $\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) \geq 0$. Therefore, with some $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\frac{\log d}{N}}$, there exists some constant $k_0 > 0$, such that

$$\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 \leq k_0\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{\log d} + s_{\boldsymbol{\delta}}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 = k_0\sqrt{\bar{s}_{\boldsymbol{\delta}}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2,$$

on $\mathcal{A}_1 \cap \mathcal{A}_2$ and when $N > N_0$. ∎

Proof of Lemma 4.13. We prove by contradiction. Suppose that $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 > 1$. Let $\widetilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta}_{\boldsymbol{\delta}}/\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2$. Then, $\|\widetilde{\boldsymbol{\Delta}}\|_2 = 1$. When $\boldsymbol{\Delta}_{\boldsymbol{\delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$, we have

$$\|\widetilde{\boldsymbol{\Delta}}\|_1 = \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1/\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq k_0\sqrt{\bar{s}_{\boldsymbol{\delta}}} = k_0\sqrt{\bar{s}_{\boldsymbol{\delta}}}\|\widetilde{\boldsymbol{\Delta}}\|_2.$$

That is, $\widetilde{\boldsymbol{\Delta}} \in \widetilde{C}(\bar{s}_{\boldsymbol{\delta}}, k_0)$, and hence $\widetilde{\boldsymbol{\Delta}} \in \widetilde{K}(\bar{s}_{\boldsymbol{\delta}}, k_0, 1)$. Let $u = \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2^{-1}$. Then, $0 < u < 1$. Note that $\mathcal{F}(\cdot)$ is a convex function. Hence, when $N > N_1$,

$$\mathcal{F}(\widetilde{\boldsymbol{\Delta}}) = \mathcal{F}(u\boldsymbol{\Delta}_{\boldsymbol{\delta}} + (1-u)\mathbf{0}) \leq u\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) + (1-u)\mathcal{F}(\mathbf{0}) \overset{(i)}{=} u\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) \overset{(ii)}{\leq} 0,$$

where (i) holds since $\mathcal{F}(\mathbf{0}) = 0$ by construction of $\mathcal{F}(\cdot)$; (ii) holds by the construction of $\widehat{\boldsymbol{\delta}}$. However, by Lemma 4.11, $\mathcal{F}(\widetilde{\boldsymbol{\Delta}}) > 0$. Thus, we conclude that $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$. ∎

Proof of Lemma 4.14.

$$x \leq \frac{b+\sqrt{b^2+4ac}}{2a} \leq \frac{b+\sqrt{b^2}+\sqrt{4ac}}{2a} = \frac{b}{a} + \sqrt{\frac{c}{a}}.$$

■

Proof of Lemma 4.15. Let $\mathcal{X}$ the support of $\mathbf{S}_1$. Under Assumption 4.1, for all $\mathbf{S}_1 \in \mathcal{X}$, there exists some constant $c > 0$ such that

$$\exp(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \leq c, \quad \exp(-\mathbf{S}_1^T \boldsymbol{\gamma}^*) < g^{-1}(\mathbf{S}_1^T \boldsymbol{\gamma}^*) \leq c.$$

By Theorem 4.3,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right).$$

Since $\mathbf{S}_1$ is a sub-Gaussian random vector under Assumption 4.4, by Theorem 2.6 of [Wai19],

$$\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right).$$

Additionally, note that $s_{\boldsymbol{\gamma}} = o(\frac{N}{\log d_1})$. It follows that

$$P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1) = 1 - o(1).$$

For any $\boldsymbol{\gamma} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\}$, we have

$$\begin{aligned}\left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r} &= \left\|\exp(-\mathbf{S}_1^T\boldsymbol{\gamma}^*)\left[\exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\} - 1\right]\right\|_{P,r} \\ &\leq c\left\|\exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\} - 1\right\|_{P,r}\end{aligned}$$

By Taylor's Thorem, for any $\mathbf{S}_1 \in \mathcal{X}$, with some $v \in (0,1)$,

$$\begin{aligned}\left|\exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\} - 1\right| &= \exp\left\{-v\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\left|\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right| \\ &\leq \left[1 + \exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\right]\left|\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right|.\end{aligned}$$

Condition on the event $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 \le 1$. Note that $\boldsymbol{\gamma}-\boldsymbol{\gamma}^* = (1-w)(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)$ and $1-w \in [0,1]$, we have

$$\begin{aligned}
&\left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}) - g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r} \\
&\quad\le c\left\|\left[1+\exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\right]\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{P,r} \\
&\quad\le c\left\|\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{P,r} \\
&\qquad + c\left\|\exp\left\{-\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\right\|_{P,2r}\left\|\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{P,2r} \\
&\quad= O\left(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right), \qquad (4.97)
\end{aligned}$$

using Assumption 4.4 and Theorem 2.6 of [Wai19]. It follows that,

$$\left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r} \le \left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r} + O\left(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\right) \le C,$$

with some constant $C > 0$, since $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 \le 1$. Therefore, we conclude that $P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1-o(1)$. Moreover, by the fact that $\exp(-u) = g^{-1}(u) - 1 < g^{-1}(u)$ and $\|X\|_{P,r'} \le \|X\|_{P,12}$ for any $X \in \mathbb{R}$ and $1 \le r' \le 12$, we have

$$\left\|g^{-1}(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} \le C, \quad \left\|\exp(-\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} \le C.$$

Moreoever, we have (4.48), since $\widehat{\boldsymbol{\gamma}} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\}$, $P_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1-o(1)$, and (4.97) holds. Besides, note that

$$\begin{aligned}
\left\|\exp(\mathbf{S}_1^T\boldsymbol{\gamma}) - \exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r'} &\le c\left\|\exp\{\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\} - 1\right\|_{P,r'} \\
&\le c\left[\left\|\exp\{\mathbf{S}_1^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\}\right\|_{P,r'} + 1\right] = O(1),
\end{aligned}$$

since $\mathbf{S}_1$ is sub-Gaussian and $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le 1$. Therefore,

$$\begin{aligned}
\left\|\exp(\mathbf{S}_1^T\boldsymbol{\gamma})\right\|_{P,r'} &\le \left\|\exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r'} + \left\|\exp(\mathbf{S}_1^T\boldsymbol{\gamma}) - \exp(\mathbf{S}_1^T\boldsymbol{\gamma}^*)\right\|_{P,r'} \\
&\le c + O(1) = O(1).
\end{aligned}$$

■

Proof of Lemma 4.16. Let $\mathcal{S}$ the support of $\bar{\mathbf{S}}_2$. Under Assumption 4.1, there exists some constant $c > 0$, such that, for all $\bar{\mathbf{S}}_2 \in \mathcal{S}$,

$$\exp(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \leq c, \quad \exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) < g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*) \leq c.$$

(a) By Theorem 4.3,

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right) = o_p(1).$$

By Assumption 4.4 and Theorem 2.6 of [Wai19],

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right).$$

(b) By Theorem 4.4,

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right) = o_p(1).$$

Similarly, by Assumption 4.4 and Theorem 2.6 of [Wai19],

$$\left\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{P,r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right).$$

The remaining proof is an analog of the proof of Lemma 4.15. Let either (a) or (b) holds. Then, we have $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = o_p(1)$. Hence,

$$P_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1) = 1 - o(1).$$

For any $\boldsymbol{\delta} \in \{w\boldsymbol{\delta}^* + (1-w)\widehat{\boldsymbol{\delta}} : w \in [0,1]\}$, we have

$$\begin{aligned}\left\|g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}) - g^{-1}(\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)\right\|_{P,r} &= \left\|\exp(-\bar{\mathbf{S}}_2^T \boldsymbol{\delta}^*)\left[\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta} - \boldsymbol{\delta}^*)\right\} - 1\right]\right\|_{P,r} \\ &\leq c\left\|\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta} - \boldsymbol{\delta}^*)\right\} - 1\right\|_{P,r}\end{aligned}$$

By Taylor's Thorem, for any $\bar{\mathbf{S}}_2 \in \mathcal{S}$, with some $v \in (0,1)$,

$$\begin{aligned}\left|\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\}-1\right| &= \exp\left\{-v\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\}\left|\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right| \\ &\leq \left[1+\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\}\right]\left|\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right|.\end{aligned}$$

Condition on the event $\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*\|_2 \leq 1$. Note that $\boldsymbol{\delta}-\boldsymbol{\delta}^* = (1-w)(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*)$ and $1-w \in [0,1]$, we have

$$\begin{aligned}&\left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})-g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} \\ &\quad\leq c\left\|\left[1+\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\}\right]\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\|_{P,r} \\ &\quad\leq c\left\|\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\|_{P,r} \\ &\qquad + c\left\|\exp\left\{-\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\}\right\|_{P,2r}\left\|\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\right\|_{P,2r} \\ &\quad= O\left(\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*\|_2\right) = O(1),\end{aligned} \tag{4.98}$$

using Assumption 4.4 and Theorem 2.6 of [Wai19]. It follows that,

$$\left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r} \leq \left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r} + O(1) \leq C,$$

with some constant $C > 0$. Therefore, we conclude that $P_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{E}_2) = 1-o(1)$. Moreover, by the fact that $\exp(-u) = g^{-1}(u) - 1 < g^{-1}(u)$ and $\|X\|_{P,r'} \leq \|X\|_{P,12}$ for any $X \in \mathbb{R}$ and $1 \leq r' \leq 12$, we have

$$\left\|g^{-1}(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} \leq C, \quad \left\|\exp(-\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} \leq C.$$

Moreoever, we have (4.50) and (4.51), since $\widehat{\boldsymbol{\delta}} \in \{w\boldsymbol{\delta}^* + (1-w)\widehat{\boldsymbol{\delta}} : w \in [0,1]\}$, $P_{\mathbb{S}_{\boldsymbol{\gamma}}\cap\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{E}_2) = 1-o(1)$, and (4.98) holds. Besides, note that

$$\begin{aligned}\left\|\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})-\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r'} &\leq c\left\|\exp\{\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\}-1\right\|_{P,r'} \\ &\leq c\left[\left\|\exp\{\bar{\mathbf{S}}_2^T(\boldsymbol{\delta}-\boldsymbol{\delta}^*)\}\right\|_{P,r'}+1\right] = O(1),\end{aligned}$$

since $\bar{\mathbf{S}}_2$ is sub-Gaussian and $\|\boldsymbol{\delta} - \boldsymbol{\delta}^*\|_2 \leq 1$. Therefore,

$$\begin{aligned}\left\|\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta})\right\|_{P,r'} &\leq \left\|\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r'} + \left\|\exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}) - \exp(\bar{\mathbf{S}}_2^T\boldsymbol{\delta}^*)\right\|_{P,r'} \\ &\leq c + O(1) = O(1).\end{aligned}$$

■

Proof of Lemma 4.17. The upper bounds for $\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r}$ follow directly by Theorems 4.3, 4.4, Theorem 2.6 of [Wai19], and the sub-Gaussianity of $\bar{\mathbf{S}}_2$ assumed in Assumption 4.4. Let either (a) or (b) holds. Then, we have $\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} = o_p(1)$. Note that, $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = (1 - v_1)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$. Therefore,

$$\begin{aligned}\|\widetilde{\varepsilon}\|_{P,r} &\leq \|\varepsilon\|_{P,r} + \|\bar{\mathbf{S}}_2^T(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} = \|\varepsilon\|_{P,r} + (1 - v_1)\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} \\ &= O(1) + o_p(1) = O_p(1).\end{aligned}$$

■

Proof of Lemma 4.18. The upper bounds for $\left\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{P,r}$ follow directly by Theorems 4.3, 4.4, Theorem 2.6 of [Wai19], and the sub-Gaussianity of $\mathbf{S}_1$ assumed in Assumption 4.4. Let either (a) or (b) of Lemma 4.18 holds, and let either (a) or (b) of 4.17 holds. Then, we have $\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{P,r} = o_p(1)$ and $\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} = o_p(1)$. Note that, $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = (1 - v_1)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$ and $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (1 - v_2)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Therefore,

$$\begin{aligned}\|\widetilde{\zeta}\|_{P,r} &\leq \|\zeta\|_{P,r} + \|\mathbf{S}_1^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{P,r} + \|\bar{\mathbf{S}}_2^T(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} \\ &= \|\zeta\|_{P,r} + (1 - v_1)\|\mathbf{S}_1^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_{P,r} + (1 - v_2)\|\bar{\mathbf{S}}_2^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{P,r} \\ &= O(1) + o_p(1) = O_p(1).\end{aligned}$$

■

4.6 Acknowledgement

Chaper 4, in full, is currently being prepared for submission for publication of the material. Zhang, Yuqian; Bradic, Jelena; Ji, Weijie. Dynamic treatment effects: high-dimensional inference under model misspecification. The dissertation author was the primary investigator and author of this material.

Bibliography

[ABS+21] David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

[Acc74] Frank William Accomando. *Optimal Asymptotic Test of a Composite Statistical Hypothesis*. PhD thesis, University of Maine, 1974.

[AIW18] 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.

[AK05] Alan Agresti and Bernhard Klingenberg. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):691–706, 2005.

[AV21] Vahe Avagyan and Stijn Vansteelandt. High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics & Epidemiology*, pages 1–18, 2021.

[BAWM18] Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.

[BCH14] 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.

[BCK15] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[BHL20] Hugo Bodory, Martin Huber, and Lukáš Lafférs. Evaluating (weighted) dynamic treatment effects by double machine learning. *arXiv preprint arXiv:2012.00370*, 2020.

[BJZ21] Jelena Bradic, Weijie Ji, and Yuqian Zhang. High-dimensional inference for dynamic treatment effects. *arXiv preprint arXiv:2110.04924*, 2021.

[BMZ04] Doron Blatt, Susan A Murphy, and Ji Zhu. A-learning for approximate planning. *Ann Arbor*, 1001:48109–2122, 2004.

[BR05] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[BRR19] Lucia Babino, Andrea Rotnitzky, and James Robins. Multiple robust estimation of marginal structural mean models for unconstrained outcomes. *Biometrics*, 75(1):90–99, 2019.

[BRS21] Iavor Bojinov, Ashesh Rambachan, and Neil Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196, 2021.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[BS19] Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019.

[BSB+06] John D Baxter, Jonathan M Schapiro, Charles AB Boucher, Veronika M Kohlbrenner, David B Hall, Joseph R Scherer, and Douglas L Mayers. Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. *Journal of Virology*, 80(21):10794–10801, 2006.

[BVDBS+15] Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope-adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103, 2015.

[BWZ19] Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

[CAC18] David Cheng, Ashwin Ananthakrishnan, and Tianxi Cai. Efficient and robust semi-supervised estimation of average treatment effects in electronic medical records data. *arXiv preprint arXiv:1804.00195*, 2018.

[CC18] Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

[CCD+17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[CG20] T Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):391–419, 2020.

[CHIM09] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[CLCL19] Abhishek Chakrabortty, Jiarui Lu, T Tony Cai, and Hongzhe Li. High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*, 2019.

[CM14] Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

[CMS10] Bibhas Chakraborty, Susan Murphy, and Victor Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*, 19(3):317–343, 2010.

[CRL+10] Lauren E Cain, James M Robins, Emilie Lanoy, Roger Logan, Dominique Costagliola, and Miguel A Hernán. When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2), 2010.

[CSZ09] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[CZW21] Yuan Chen, Donglin Zeng, and Yuanjia Wang. Learning individualized treatment rules for multiple-domain latent outcomes. *Journal of the American Statistical Association*, 116(533):269–282, 2021.

[DAV20] Oliver Dukes, Vahe Avagyan, and Stijn Vansteelandt. Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics*, 76(4):1190–1200, 2020.

[DCDS+13] Rhian M Daniel, SN Cousens, BL De Stavola, Michael G Kenward, and JAC Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618, 2013.

[DV20] Oliver Dukes and Stijn Vansteelandt. Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 2020.

[DVDGVW10] Lutz Dümbgen, Sara A Van De Geer, Mark C Veraar, and Jon A Wellner. Nemirovski's inequalities revisited. *The American Mathematical Monthly*, 117(2):138–160, 2010.

[EACR+16] Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of ℓ_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

[EHvdL20] Ashkan Ertefaie, Nima S Hejazi, and Mark J van der Laan. Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. *arXiv preprint arXiv:2005.11303*, 2020.

[Far15] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[FLM21] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[FWXY20] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[GB05] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005.

[GC18] Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):579–594, 2018.

[Gei75] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

[GLTC20] Jessica Gronsbell, Molei Liu, Lu Tian, and Tianxi Cai. Efficient estimation and evaluation of prediction rules in semi-supervised settings under stratified sampling. *arXiv preprint arXiv:2010.09443*, 2020.

[Gra11] Bryan S Graham. Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2):437–452, 2011.

[HBH05] Ruth Hummel, Senin Banga, and Thomas P Hettmansperger. Better confidence intervals for the variance in a random sample. Technical report, Citeseer, 2005.

[HBR01] Miguel A Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

[HLL20] Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):32–47, 2020.

[Hol88] Paul W Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988.

[HP11] Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*, 2011.

[HR10] Miguel A Hernán and James M Robins. Causal inference, 2010.

[HSHD+16] Miguel A Hernán, Brian C Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75, 2016.

[Imb04] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[IR15a] Kosuke Imai and Marc Ratkovic. Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, 110(511):1013–1023, 2015.

[IR15b] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[JYF10] Marshall M Joffe, Wei Peter Yang, and Harold I Feldman. Selective ignorability assumptions in causal inference. *The International Journal of Biostatistics*, 6(2), 2010.

[KC18] Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond subgaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

[KK13] Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine learning*, 91(2):189–209, 2013.

[KM20] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

[KS07] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

[KS18] Nathan Kallus and Michele Santacatterina. Optimal balancing of time-dependent confounders for marginal structural models. *arXiv preprint arXiv:1806.01083*, 2018.

[KSBY19] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[KT10] Shakeeb Khan and Elie Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

[LLK+20] Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020. PMID: 32952236.

[LLL+19] Ning Liu, Ying Liu, Brent Logan, Zhiyuan Xu, Jian Tang, and Yanzhi Wang. Learning the dynamic treatment regimes from medical registry data through deep q-network. *Scientific Reports*, 9(1):1–10, 2019.

[LLQ+14] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225, 2014.

[LLS14] Eric B Laber, Kristin A Linn, and Leonard A Stefanski. Interactive model building for q-learning. *Biometrika*, 101(4):831–847, 2014.

[LLS17] Kristin A Linn, Eric B Laber, and Leonard A Stefanski. Interactive q-learning for quantiles. *Journal of the American Statistical Association*, 112(518):638–649, 2017.

[LM05] Michael Lechner and Ruth Miquel. Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *University of St. Gallen Economics Discussion Paper*, (2005-17), 2005.

[LS20] Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects via g-estimation. *arXiv preprint arXiv:2002.07285*, 2020.

[LZC20] Molei Liu, Yi Zhang, and Tianxi Cai. Doubly robust covariate shift regression with semi-nonparametric nuisance models. *arXiv preprint arXiv:2010.02521*, 2020.

[MC18] Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *Journal of Machine Learning Research*, 19(1):3074–3100, 2018.

[MR10] Erica EM Moodie and Thomas S Richardson. Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146, 2010.

[Mur03] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[Mur05] Susan A Murphy. A generalization error for q-learning. 2005.

[MvdLRG01] Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[NBW21] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

[NRWY10] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *arXiv preprint arXiv:1010.2731*, 2010.

[NRWY12] 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.

[ORR10] Liliana Orellana, Andrea Rotnitzky, and James M Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The International Journal of Biostatistics*, 6(2), 2010.

[Owe07] Art B Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(Apr):761–773, 2007.

[Qin98] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[RGK+03] Soo-Yon Rhee, Matthew J Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela, and Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31(1):298–303, 2003.

[RLSR12] Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

[Rob86] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[Rob87] James M Robins. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987.

[Rob97] James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

[Rob00a] James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

[Rob00b] James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[Rob04] James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

[Rot17] Christoph Rothe. Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660, 2017.

[RR95] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[RRZ94] 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.

[RS19] Ashesh Rambachan and Neil Shephard. A nonparametric dynamic causal model for macroeconometrics. *Available at SSRN 3345325*, 2019.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[RWG19] Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019.

[RWY10] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[RZ12] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 10.1–10.24, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings.

[SC17] Vira Semenova and Victor Chernozhukov. Estimation and inference about conditional average treatment effect and other structural functions. *arXiv preprint arXiv:1702.06240*, 2017.

[SFSL18] Chengchun Shi, Alin Fan, Rui Song, and Wenbin Lu. High-dimensional a-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, 46(3):925, 2018.

[SNDS90] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[SRR99] 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.

[SRR19] Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

[Sto74] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[SWZK15] Rui Song, Weiwei Wang, Donglin Zeng, and Michael R Kosorok. Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901, 2015.

[SZF20] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[Tan20a] Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837, 2020.

[Tan20b] Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020.

[Tib97] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[TS12] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *The Annals of Statistics*, 40(3):1816, 2012.

[Tsi07] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[TYWK+19] Linh Tran, Constantin Yiannoutsos, Kara Wools-Kaloustian, Abraham Siika, Mark Van Der Laan, and Maya Petersen. Double robust efficient estimators of longitudinal treatment effects: comparative performance in simulations and a case study. *The International Journal of Biostatistics*, 15(2), 2019.

[VB21] Davide Viviano and Jelena Bradic. Dynamic covariate balancing: estimating treatment effects over time. *arXiv preprint arXiv:2103.01280*, 2021.

[VdGBRD14] 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.

[vdGL13] Sara van de Geer and Johannes Lederer. The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1-2):225–250, 2013.

[vdLG11] Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of an intervention specific mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 290, 2011.

[vdLPJ05] Mark J van der Laan, Maya L Petersen, and Marshall M Joffe. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics*, 1(1), 2005.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[VG03] Stijn Vansteelandt and Els Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):817–835, 2003.

[VZ18] Giancarlo Visconti and José R Zubizarreta. Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4:217–249, 2018.

[WA18] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Wan20] HaiYing Wang. Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR, 2020.

[Wat89] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

[WD92] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[WL08] Larry Wasserman and John D Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.

[WW15] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[YD17] Shu Yang and Peng Ding. Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *arXiv preprint arXiv:1704.00666*, 2017.

[YS18] Sean Yiu and Li Su. Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika*, 105(3):709–722, 2018.

[YvdL06] Zhuo Yu and Mark van der Laan. Double robust estimation in longitudinal marginal structural models. *Journal of Statistical Planning and Inference*, 136(3):1061–1089, 2006.

[YZ10] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

[ZB18] Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

[ZB21] Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 2021. asab042.

[ZBC19] Anru Zhang, Lawrence D Brown, and T Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

[ZCB21] Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Double robust semi-supervised inference for the mean: Selection bias under mar labeling with decaying overlap. *arXiv preprint arXiv:2104.06667*, 2021.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[Zhu05] Xiaojin Zhu. Semi-supervised learning literature survey. *World*, 10:10, 2005.

[ZTD+12] Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

[ZW18] Xiang Zhou and Geoffrey T Wodtke. Residual balancing weights for marginal structural models: with application to analyses of time-varying treatments and causal mediation. *arXiv preprint arXiv:1807.10869*, 2018.

[ZW20] Xiang Zhou and Geoffrey T Wodtke. Residual balancing: a method of constructing weights for marginal structural models. *Political Analysis*, 28(4):487–506, 2020.

[ZZ14] 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.

[ZZS19] Wensheng Zhu, Donglin Zeng, and Rui Song. Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417, 2019.