Title

High-dimensional Inference for Dynamic Treatment Effects and Adaptive Split Balancing for Optimal Random Forests

Author

Ji, Weijie

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

High-dimensional Inference for Dynamic Treatment Effects and Adaptive Split Balancing for Optimal Random Forests

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

in

Mathematics with a Specialization in Statistics

by

Weijie Ji

Committee in charge:

Professor Jelena Bradic, Chair
Professor Yian Ma
Professor Dimitris Politis
Professor Danna Zhang
Professor Wenxin Zhou

2023

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

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Jelena Bradic. Her profound knowledge and unwavering passion for research have been a great source of inspiration for my academic research. I am fortunate and grateful to have had her mentorship during my PhD studies. I am also grateful for the countless hours she dedicated to reviewing drafts, providing constructive feedback, and engaging in thoughtful discussions. Her dedication to fostering a spirit of intellectual curiosity and meticulous attention to detail have left an indelible mark on my academic journey.

Next, I would like to thank Professors Yian Ma, Dimitris Politis, Danna Zhang and Wenxin Zhou for their support and serving on my thesis committee. I would also thankful to Professor Professor Yuqian Zhang. I am delighted to have had the opportunity to work with him, and I have gained valuable insights from his generous suggestions.

In addition, I extend my heartfelt thanks to my other colleagues and friends at UCSD for their companionship. They brought endless fun to my life, especially during the COVID-19 pandemic. I have truly enjoyed my time with them. Finally, I would like to express my gratitude to my parents and friends for their tremendous understanding and encouragement in the past few years.

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

Chaper 1, 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 2, in full, has been submitted for publication of the material. Zhang, Yuqian;

Ji, Weijie; Bradic, Jelena. Dynamic treatment effects: high-dimensional inference under model misspecification. The dissertation author was the primary investigator and author of this material.

Chaper 3, in full, is currently being prepared for submission for publication of the material. Bradic, Jelena; Ji, Weijie; Zhang, Yuqian. Adaptive split balancing for optimal random forests. The dissertation author was the primary investigator and author of this material.

VITA

2017	B. S. in Mathematics and Applied Mathematics, Shanghai University
2023	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

Y. Zhang, W. Ji and J. Bradic, "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`.

ABSTRACT OF THE DISSERTATION

High-dimensional Inference for Dynamic Treatment Effects and Adaptive Split Balancing for Optimal Random Forests

by

Weijie Ji

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2023

Professor Jelena Bradic, Chair

The first two chapters consider the estimation and inference of the dynamic treatment effect when the confounders are possibly high dimensional. Chapter 1 proposes a sequential doubly robust Lasso (S-DRL) estimator using ℓ_1-regularized nuisance estimates with DR-type imputations. The proposed method achieves consistency as long as at least one nuisance function is appropriately parametrized for each exposure time and treatment path. The key to achieving these results is the usage of DR representations for intermediate conditional outcome models, which offers superior inferential performance while requiring weaker as-

sumptions. We establish root-n inference based on the S-DRL estimator is guaranteed when two product-sparsity conditions are satisfied. Chapter 2 further provides root-n inference for the dynamic treatment effect even when model misspecification occurs. We provide valid inference based on a "sequential model double robust" solution as long as one of the nuisance models is correctly specified at each time spot. Chapter 3 proposes a novel construction for random forests, incorporating cyclic modification of the selection of splitting directions with the goal of achieving a faster consistency rate for the integrated mean squared error (IMSE). Setting $\alpha = 0.5$ leads to the proposed cyclic forest degenerating into cyclic median forests, obtaining a minimax optimal rate for IMSE within the Lipschitz class. We further extend our exploration to local polynomial regression within each leaf, formulating cyclic local polynomial forests as generalizations of the cyclic forests. When $\alpha = 0.5$, our cyclic local polynomial forests attain a minimax rate for IMSE, marking the first instance of achieving minimax optimal rates for random forests within the Hölder class. Furthermore, we establish minimax optimal rates for the uniform convergence rate.

Chapter 1

High-dimensional inference for dynamic treatment effects

1.1 Introduction

The complexity of a given disease or economic policy often manifests in the diversity and size of the personal characteristics pertaining to each individual or economy under consideration, causing a considerable degree of heterogeneity in observed outcomes. However, the utility of randomized control trials (RCTs), especially over time, is frequently curtailed by prohibitive costs or ethical concerns. In contrast, the accessibility of time-varying observational studies has burgeoned of late. The ubiquity of data-driven decision-making is evident in various aspects of daily life, such as the continuous monitoring of individuals' health using mobile devices and consequential medical interventions, tracking of online presence and real-time measurement of economic and social policies implemented to enhance

public health. The present study contributes novel insights to the literature by proposing a novel framework to construct confidence intervals pertaining to dynamic treatment effects amid high-dimensional observations. In a Job Corps real-data analysis, our novel framework provides more accurate estimates of the long-term impact of additional schooling over time on wages, which has important practical implications for designing effective policies aimed at increasing educational attainment and improving economic outcomes.

In light of intricate notational complexities, we exemplify our ideas and findings for two-stage trials while affirming that the same theoretical framework and methodology developed are extensible to multiple-stage trials; see, e.g., Section 1.5. Consider a two-stage series of binary treatment assignments, denoted by A_1 and A_2, and an outcome of interest, $Y \in \mathbb{R}$. Alongside this, a set of possibly high-dimensional sequential pre-treatment covariates $\mathbf{S}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{S}_2 \in \mathbb{R}^{d_2}$, possibly of different dimensions, are also observed. The potential or counterfactual outcomes, $Y(a)$, refer to the outcome that a participant would have experienced had they followed a particular treatment sequence, $a = (a_1, a_2) \in \{0, 1\}^2$, which may differ from the treatment they were observed with. Our parameter of interest is the dynamic treatment effect (DTE) between two treatment paths, a and a', which is defined as follows:

$$\theta := E[Y(a)] - E[Y(a')] = \theta_a - \theta_{a'}, \text{ with } \theta_a := E[Y(a)]. \tag{1.1}$$

Estimating the DTE is a challenging task when there are multiple exposures involved. The influence of past treatments on future confounders and treatment choices complicates the identifiability of θ [RR83]. Adjusting for confounders may not have a causal interpretation, even when all confounders are measured and the regression is correctly specified [DCDS+13].

In this context, alternative methods such as Sequential Multiple Randomized Control Trials (SMART) [HSHD$^+$16], Structural Nested Mean (SNM) [Rob97], and Marginal Structural Mean (MSM) models [MvdLRG01] have become the gold standard for addressing these challenges. This paper contributes to the field by establishing robust MSM model estimations with new effective rates.

1.1.1 The doubly robust representations

Throughout this work, we assume that any treatment-specific variable can only be affected by past treatments or past covariates; and not the future. This is sometimes called temporal ordering. We also assume a "no interference" setting and Assumption 1.1 below [Rob87, Rob00a, Mur03].

Assumption 1.1. *(a) (Sequential Ignorability)* $Y(a_1, a_2) \perp\!\!\!\perp A_1 \mid \mathbf{S}_1$ *and* $Y(a_1, a_2) \perp\!\!\!\perp A_2 \mid \mathbf{S}, A_1 = a_1$ *where* $\mathbf{S} = (\mathbf{S}_1^\top, \mathbf{S}_2^\top)^\top \in \mathbb{R}^d$ *with* $d := d_1 + d_2$. *(b) (Consistency of potential outcomes)* $Y = Y(A_1, A_2)$. *(c) (Overlap) Let* $c_0 \in (0, 1/2)$ *be a positive constant, such that* $P(c_0 \leq \pi_a(\mathbf{S}_1) \leq 1 - c_0) = 1$, *and* $P(c_0 \leq \rho_a(\mathbf{S}) \leq 1 - c_0) = 1$. *Here, the propensity scores are defined as* $\pi_a(\mathbf{s}_1) := P[A_1 = a_1 | \mathbf{S}_1 = \mathbf{s}_1]$ *and* $\rho_a(\mathbf{s}) := P[A_2 = a_2 | \mathbf{S} = \mathbf{s}, A_1 = a_1]$.

The following lemma provides a doubly robust (DR) representation of θ_a. This result is consistent with previous studies in the literature, including works by [vdLG12, ORR10, MvdLRG01, BR05]. We consider the MSM models where we adjust for confounding variables that may affect both the treatment assignment and the outcome of interest. In an MSM, the treatment assignment and the outcome of interest are modeled separately using propensity scores $\pi_a(\mathbf{s}_1)$ and $\rho_a(\mathbf{s})$ together with the first-time and second-time conditional means,

$\mu_a(\mathbf{s}_1) := E[Y(a)|\mathbf{S}_1 = \mathbf{s}_1]$ and $\nu_a(\mathbf{s}) := E[Y(a)|\mathbf{S} = \mathbf{s}, A_1 = a_1]$. Throughout this work, we use $\pi_a^*(\cdot)$ and $\rho_a^*(\cdot)$ as well as $\mu_a^*(\cdot)$ and $\nu_a^*(\cdot)$ to refer to the working models, i.e., the population-level approximations of the propensity scores and conditional means, respectively.

Lemma 1.1 (A DR representation of θ_a)**.** *Let Assumption 1.1 hold. Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified, i.e, (a) either $\mu_a^*(\cdot) = \mu_a(\cdot)$ or $\pi_a^*(\cdot) = \pi_a(\cdot)$, but not necessarily both and (b) either $\nu_a^*(\cdot) = \nu_a(\cdot)$ or $\rho_a^*(\cdot) = \rho_a(\cdot)$, but not necessarily both. Then*

$$\theta_a = E\left[\mu_a^*(\mathbf{S}_1) + \mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)} + \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right]. \tag{1.2}$$

Based on Lemma 1.1, consistent estimates of θ_a are expected as long as at least one nuisance model is correctly parametrized at each exposure time. However, this goal has not been achieved yet; see [BRR19] for an overview. The main obstacle is the estimation of interlocking nuisance functions, especially the first-time conditional mean, as it cannot be identified directly through the observable variables as $\mu_a(\mathbf{s}_1) = E[Y(a)|\mathbf{S}_1 = \mathbf{s}_1] \neq E[Y|\mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1]$. Under Assumption 1.1, existing DTE literature typically considers the following nested representation of $\mu_a(\cdot)$,

$$\mu_a(\mathbf{s}_1) = E[Y(a)|\mathbf{S}_1 = \mathbf{s}_1] = E[\nu_a(\mathbf{S})|\mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1], \tag{1.3}$$

and suggests a nested regression (NR) of the conditional means – as long as an estimate $\widehat{\nu}_a(\cdot)$ of $\nu_a(\cdot)$ is obtained, one can use $\widehat{\nu}_a(\mathbf{S}_i)$ as the imputed outcomes and perform regression to construct $\widehat{\mu}_{a,\mathrm{NR}}(\cdot)$; see, e.g., [MvdLRG01]. We formalize these properties under high-dimensional linear working models, naming the resulting DTE estimator the "dynamic treatment Lasso" (DTL) estimator. We show that the nested-regression approach faces certain limitations and fails to attain the DR property equivalent to Lemma 1.1. Among the

multiple factors contributing to this, the biggest one is arising from a peculiar model misspecification that we identified arising from the nested representation in Equation (1.3). In the event of a misspecified linear working model $\nu_a^*(\cdot)$, the corresponding $\mu_a^*(\cdot)$ will inevitably be misspecified as well, leading to $\mu_a^*(\cdot) \neq \mu_a(\cdot)$, even when $\mu_a(\cdot)$ is itself linear. Besides the linearity of $\mu_a(\cdot)$, additional conditions on $\nu_a(\cdot)$ are necessary for the correctness of the nested-regression-based linear working model, as discussed in Section 1.2.3.

This issue necessitates the use of specialized methods for which we propose a new DR representation of the first-time conditional mean function $\mu_a(\cdot)$; see (1.4) below. It provides tools to quantify the DR property of the resulting DTE estimate and to develop correction techniques that can mitigate the DR gap by achieving the estimation under model conditions equivalent to Lemma 1.1.

Theorem 1.1 (A DR representation of $\mu_a(\cdot)$)**.** *Suppose that either* $\nu_a^*(\cdot) = \nu_a(\cdot)$ *or* $\rho_a^*(\cdot) = \rho_a(\cdot)$ *holds. Let Assumption 1.1 holds. Then, for any* $\mathbf{s}_1 \in \mathbb{R}^{d_1}$*,*

$$\mu_a(\mathbf{s}_1) = E\left[\nu_a^*(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1\right]. \tag{1.4}$$

Utilizing the two DR representations (1.2) and (1.4) simultaneously, we propose a *sequential doubly robust Lasso* (S-DRL) estimator. The proposed estimator is consistent as long as either the conditional mean function is truly linear or the propensity score function is truly logistic (or both) for each exposure time. To the best of our knowledge, this is the first estimator that matches Lemma 1.1 conditions empirically. The inverse probability weighting (IPW) methods [Rob86, Rob00a, HBR01, Rob04] require all the propensity score models to be correctly parametrized. The covariate balancing methods [KS18, YS18, VB21] require all the conditional mean models to be correctly parametrized. Perhaps unexpectedly, the stan-

Table 1.1: Consistency of the S-DRL, DTL, and MR estimators in two-stage trials.

Nuisance models				Consistency		
logistic $\rho_a(\cdot)$	logistic $\pi_a(\cdot)$	linear $\mu_a(\cdot)$	linear $\nu_a(\cdot)$	S-DRL	DTL	MR
✓	✓	✓	✓	✓	✓	✓
✗	✓	✓	✓	✓	✓	✓
✓	✗	✓	✓	✓	✓	✓
✓	✓	✗	✓	✓	✓	✓
✓	✓	✓	✗	✓	✓	✓
✗	✗	✓	✓	✓	✓	✓
✗	✓	✗	✓	✓	✓	✗
✓	✗	✓	✗	✓	✗	✓
✓	✓	✗	✗	✓	✓	✓

dard low-dimensional DR methods [Rob00b, MvdLRG01, BR05, YvdL06] and the targeted maximum likelihood estimation (TMLE) [vdLG12] require either all the propensity score functions or all the conditional mean (or density) functions to be correctly parametrized. The "multiply robust" (MR) estimator of [BRR19] reaches better robustness than all of the aforementioned methods. In general T-stage trials, they allow for the first t conditional mean models and the last $T-t$ propensity score models to be correctly parametrized for any t. The DTL estimator allows the first t propensity score models and the last $T-t$ conditional mean models to be correctly parametrized. Our S-DRL estimator is strictly more robust in terms of consistency; see Table 1.1 and Remark 1.1 for further details.

The S-DRL estimator demonstrates superior estimation rates in high-dimensional contexts when compared to the DTL estimator; see Table 1.2 as well as Remark 1.5. Root-sample-size inference based on the S-DRL estimator is guaranteed when two product-sparsity

conditions are satisfied, whereas the DTL method requires three product-sparsity conditions, as demonstrated in Theorems 1.3 and 1.5. The errors in nuisance estimation at different stages have a parallel effect on the estimation; see the consistency rate in Theorem 1.6.

The estimation of the in-between outcome models is intrinsically linked to regression with imputed outcomes. We have developed a novel cone-set analysis of imputed Lasso estimates that is of independent interest to other imputed, high-dimensional regressions. Existing Lasso proof techniques provide conservative bounds only; see Section 1.4. Our results are adaptive to the imputation error and can be used to guide the selection of tuning parameters in high-dimensional regression models with imputed outcomes.

In the multi-stage exposure setting, we extend our method and develop DR representations to identify both the expected potential outcomes and conditional means, as shown in Section 1.5. While the consistency rate and asymptotic normality require intricate proofs, we anticipate they hold analogously to those in the two-stage case. It is worth noting that Theorem 1.11 provides new DR representations that are independent of any specific parametric models, allowing the sequential doubly robust (S-DR) method to be utilized with non-parametric nuisance estimates, which enhances its versatility.

1.1.2 Organization of the paper

In Section 1.2, we introduce the DR estimators of the DTE, including the proposed S-DRL estimator, the DTL estimator, and a general DR estimator. The theoretical properties of the considered DTE estimators are established in Section 1.3. In Section 1.4, we formalize the supporting theoretical discoveries, including a general theory for imputed Lasso estimation and the consistency results of the nuisance estimates. We further extend our setting

to the case of multi-stage treatments and provide general DR representations for the intermediate conditional means in Section 1.5. Section 1.6 presents numerical results, including simulation studies and an application to the National Job Corps Study. Further discussion is provided in Section 1.7.

1.1.3 Notation

For any $\alpha > 0$, let $\psi_\alpha(\cdot)$ denote the function given by $\psi_\alpha(x) := \exp(x^\alpha) - 1$, $\forall x > 0$. 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 constants $c, C > 0$. Define $g(u) = \exp(u)/\{1 + \exp(u)\}$ as the logistic function and $\phi(u) = \log(1 + \exp(u))$ as the corresponding link function throughout.

1.2 The doubly robust estimators

We observe a collection of independent and identically distributed (i.i.d.) samples $\mathcal{D} := \{W_i\}_{i=1}^N = (Y_i, \mathbf{S}_{1i}, A_{1i}, \mathbf{S}_{2i}, A_{2i})_{i=1}^N$, drawn from the same distribution as $(Y, \mathbf{S}_1, A_1, \mathbf{S}_2, A_2)$. In the following subsections, we present three DTE estimators: the new sequential doubly robust Lasso (S-DRL) estimator, the dynamic treatment Lasso (DTL) estimator, and the general DR estimator.

1.2.1 The sequential doubly robust Lasso (S-DRL) estimator

We focus on the high-dimensional scenario, and consider linear (working) models for the conditional means $\mu_a(\cdot)$ and $\nu_a(\cdot)$, along with logistic (working) models for the propensities $\pi_a(\cdot)$ and $\rho_a(\cdot)$. The population minimizer approximating $\pi_a(\mathbf{s}_1)$ is defined as $\pi_a^*(\mathbf{s}_1) = g(\mathbf{v}^\top \boldsymbol{\gamma}_a^*)$ with $\mathbf{v} = (1, \mathbf{s}_1^\top)^\top$, whereas that of approximating $\rho_a(\mathbf{s})$ is $\rho_a^*(\mathbf{s}) = g(\mathbf{u}^\top \boldsymbol{\delta}_a^*)$ with $\mathbf{u} = (1, \mathbf{s}^\top)^\top$. Here

$$\boldsymbol{\gamma}_a^* = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{d_1+1}}{\operatorname{argmin}} E\left[\phi(\mathbf{V}^\top \boldsymbol{\gamma}) - \mathbb{1}_{\{A_1 = a_1\}} \mathbf{V}^\top \boldsymbol{\gamma}\right], \ \mathbf{V} = (1, \mathbf{S}_1^\top)^\top \in \mathbb{R}^{d_1+1} \text{ and} \tag{1.5}$$

$$\boldsymbol{\delta}_a^* = \underset{\boldsymbol{\delta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} E\left[\mathbb{1}_{\{A_1 = a_1\}} \left[\phi(\mathbf{U}^\top \boldsymbol{\delta}) - \mathbb{1}_{\{A_2 = a_2\}} \mathbf{U}^\top \boldsymbol{\delta}\right]\right], \ \mathbf{U} = (1, \mathbf{S})^\top \in \mathbb{R}^{d+1}. \tag{1.6}$$

One can also consider a feature map $\varphi(\mathbf{s}_1)$ (e.g., a polynomial basis) and a working model $\pi_a^*(\mathbf{s}_1) = g(\varphi(\mathbf{s}_1)^\top \boldsymbol{\gamma}_a^*)$ with some $\boldsymbol{\gamma}_a^*$ defined correspondingly. We focus on $\varphi(\mathbf{s}_1) = \mathbf{v}$, although the results apply more broadly. The above working models can be estimated with many regularizations. Throughout this work, we focus on the ℓ_1-regularization, albeit the theoretical developments apply more broadly. With a subset of training data $\mathcal{D}_{\mathcal{J}} = \{W_i\}_{i \in \mathcal{J}} \subset \mathcal{D}$, where $\mathcal{J} \subset \{1, \ldots, N\}$, we define

$$\widehat{\boldsymbol{\gamma}}_a := \widehat{\boldsymbol{\gamma}}_a(\mathcal{D}_{\mathcal{J}}) = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{d_1+1}}{\operatorname{argmin}} \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \left[\phi(\mathbf{V}_i^\top \boldsymbol{\gamma}) - \mathbb{1}_{\{A_{1i} = a_1\}} \mathbf{V}_i^\top \boldsymbol{\gamma}\right] + \lambda_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_1, \tag{1.7}$$

$$\widehat{\boldsymbol{\delta}}_a := \widehat{\boldsymbol{\delta}}_a(\mathcal{D}_{\mathcal{J}}) = \underset{\boldsymbol{\delta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \mathbb{1}_{\{A_{1i} = a_1\}} \left[\phi(\mathbf{U}_i^\top \boldsymbol{\delta}) - \mathbb{1}_{\{A_{2i} = a_2\}} \mathbf{U}_i^\top \boldsymbol{\delta}\right] + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1, \tag{1.8}$$

with tuning parameters $\lambda_{\boldsymbol{\gamma}}, \lambda_{\boldsymbol{\delta}} > 0$. Observe that for $\widehat{\boldsymbol{\gamma}}_a$, we utilize all of the observations regardless of its treatment path, whereas for $\widehat{\boldsymbol{\delta}}_a$, only those whose treatment path matches a_1 regardless of what a_2 is. The best linear working model for the second-time conditional

mean $\nu_a(\cdot) = E[Y|\mathbf{S}, A_1 = a_1, A_2 = a_2]$ is denoted as

$$\nu_a^*(\mathbf{s}) = \mathbf{u}^\top \boldsymbol{\alpha}_a^*, \quad \boldsymbol{\alpha}_a^* := \underset{\boldsymbol{\alpha} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} E\left[\mathbb{1}_{\{A_1=a_1, A_2=a_2\}}(Y - \mathbf{U}^\top \boldsymbol{\alpha})^2\right]. \tag{1.9}$$

An estimator of (1.9) can be obtained similarly with $\lambda_{\boldsymbol{\alpha}} > 0$:

$$\widehat{\boldsymbol{\alpha}}_a := \widehat{\boldsymbol{\alpha}}_a(\mathcal{D}_{\mathcal{J}}) = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \mathbb{1}_{\{A_{1i}=a_1, A_{2i}=a_2\}}(Y_i - \mathbf{U}_i^\top \boldsymbol{\alpha})^2 + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1. \tag{1.10}$$

Algorithm 1 Sequential Double Robust Lasso (S-DRL)

Require: Observations $\mathcal{D} := \{W_i\}_{i=1}^N = (Y_i, \mathbf{S}_{1i}, A_{1i}, \mathbf{S}_{2i}, A_{2i})_{i=1}^N$, treatment path a, and control a'.

1: For any $K \geq 2$, let $\mathcal{K} = \{1, 2, \ldots, K\}$. Randomly split $\mathcal{I} = \{1, \ldots, N\}$ into K equal-sized $|\mathcal{I}_k| = n$. Define $\mathcal{I}_{-k} := \mathcal{I} \backslash \mathcal{I}_k$, and further split $\mathcal{I}_{-k}$ into two equal-sized sets $\mathcal{I}_{-k,1}$ and $\mathcal{I}_{-k,2}$.

2: Let $\mathcal{W}_{-k} := \{W_i\}_{i \in \mathcal{I}_{-k}}$ and $\mathcal{W}_{-k,j} := \{W_i\}_{i \in \mathcal{I}_{-k,j}}$ for each $j \in \{1, 2\}$.

3: **for** $k = 1, 2, ..., K$ **do**

4: **for** $c \in \{a, a'\}$ **do**

5: Using $\mathcal{W}_{-k}$, construct estimates $\widehat{\boldsymbol{\gamma}}_c$, $\widehat{\boldsymbol{\delta}}_c$, and $\widehat{\boldsymbol{\alpha}}_c$ through (1.7), (1.8), and (1.10), respectively.

6: Using $\mathcal{W}_{-k,1}$, construct estimates $\widetilde{\boldsymbol{\delta}}_c$ and $\widetilde{\boldsymbol{\alpha}}_c$ through (1.8) and (1.10), respectively.

7: For each $i \in \mathcal{I}_{-k,2}$, set $\widehat{Y}_i^{\mathrm{DR}}$ as defined in (1.14) with $\widetilde{\boldsymbol{\delta}}_c$ and $\widetilde{\boldsymbol{\alpha}}_c$ from Step 6.

8: Compute $\widehat{\boldsymbol{\beta}}_{c,1}$ through (1.15) based on the training samples $\mathcal{W}_{-k,2}$.

9: Exchange $\mathcal{W}_{-k,1}$ and $\mathcal{W}_{-k,2}$, repeat Steps 6-8 and obtain $\widehat{\boldsymbol{\beta}}_{c,2}$ analogously. Compute

$$\widehat{\boldsymbol{\beta}}_c = (\widehat{\boldsymbol{\beta}}_{c,1} + \widehat{\boldsymbol{\beta}}_{c,2})/2. \tag{1.11}$$

10: **end for**

11: Let $\hat{\eta}_c = (\widehat{\boldsymbol{\alpha}}_c, \widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\gamma}}_c, \widehat{\boldsymbol{\delta}}_c)$. Using the DR score (1.16), compute $\check{\theta}^{(k)}$ as

$$\check{\theta}^{(k)} = |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \left[\psi_a(W_i; \hat{\eta}_a) - \psi_{a'}(W_i; \hat{\eta}_{a'})\right].$$

12: **end for** **return** The S-DRL estimator and the variance estimate

$$\widehat{\theta} := K^{-1} \sum_{k \in \mathcal{K}} \check{\theta}^{(k)}, \quad \widehat{\sigma}^2 := N^{-1} \sum_{k \in \mathcal{K}, i \in \mathcal{I}_k} \left[\psi_a(W_i; \hat{\eta}_a) - \psi_{a'}(W_i; \hat{\eta}_{a'}) - \widehat{\theta}\right]^2. \tag{1.12}$$

Estimation of the first-time conditional mean $\mu_a(\cdot)$ Recall that $\boldsymbol{\delta}_a^*$ and $\boldsymbol{\alpha}_a^*$ are defined in Equations (1.6) and (1.9), respectively. We propose the following DR imputed outcome

$$Y^{\mathrm{DR}} := \mathbf{U}^\top \boldsymbol{\alpha}_a^* + \mathbb{1}_{\{A_2=a_2\}} \frac{Y - \mathbf{U}^\top \boldsymbol{\alpha}_a^*}{g(\mathbf{U}^\top \boldsymbol{\delta}_a^*)}.$$

With this in mind, we consider a linear working model for the first-time conditional mean

$$\mu_a^*(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}_a^*, \quad \boldsymbol{\beta}_a^* := \underset{\boldsymbol{\beta} \in \mathbb{R}^{d_1+1}}{\operatorname{argmin}} E\left[\mathbb{1}_{\{A_1=a_1\}} (Y^{\mathrm{DR}} - \mathbf{V}^\top \boldsymbol{\beta})^2\right]. \tag{1.13}$$

To estimate the best linear slope $\boldsymbol{\beta}_a^*$ based on a subset of training data $\mathcal{D}_{\mathcal{J}} \subset \mathcal{D}$, we consider an additional sample splitting with $\mathcal{D}_{\mathcal{J}} = \mathcal{D}_{\mathcal{J}_1} \cup \mathcal{D}_{\mathcal{J}_2}$, where $\mathcal{J}_1$ and $\mathcal{J}_2$ are disjoint subsets of $\mathcal{J}$. Using the first half of the subsamples $\mathcal{D}_{\mathcal{J}_1}$, we first obtain the second-time nuisance estimates $\widetilde{\boldsymbol{\delta}}_a := \widehat{\boldsymbol{\delta}}_a(\mathcal{D}_{\mathcal{J}_1})$ and $\widetilde{\boldsymbol{\alpha}}_a := \widehat{\boldsymbol{\alpha}}_a(\mathcal{D}_{\mathcal{J}_1})$ as (1.8) and (1.10), respectively. Then, for each $i \in \mathcal{J}_2$, we construct a DR imputed outcome

$$\widehat{Y}_i^{\mathrm{DR}} := \mathbf{U}_i^\top \widetilde{\boldsymbol{\alpha}}_a + \mathbb{1}_{\{A_{2i}=a_2\}} \frac{Y_i - \mathbf{U}_i^\top \widetilde{\boldsymbol{\alpha}}_a}{g(\mathbf{U}_i^\top \widetilde{\boldsymbol{\delta}}_a)}. \tag{1.14}$$

Based on the DR imputed outcomes $\widehat{Y}_{\mathcal{J}_2}^{\mathrm{DR}} := \{\widehat{Y}_i^{\mathrm{DR}}\}_{i \in \mathcal{J}_2}$, we propose a DR estimate:

$$\widehat{\boldsymbol{\beta}}_a := \widehat{\boldsymbol{\beta}}_a(\mathcal{D}_{\mathcal{J}_2}, \widehat{Y}_{\mathcal{J}_2}^{\mathrm{DR}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{d_1+1}}{\operatorname{argmin}} \frac{1}{|\mathcal{J}_2|} \sum_{i \in \mathcal{J}_2} \mathbb{1}_{\{A_{1i}=a_1\}} \left(\widehat{Y}_i^{\mathrm{DR}} - \mathbf{V}_i^\top \boldsymbol{\beta}\right)^2 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1, \tag{1.15}$$

where $\lambda_{\boldsymbol{\beta}} > 0$. To regain full sample size efficiency, we can always swap the samples $\mathcal{D}_{\mathcal{J}_1}$ and $\mathcal{D}_{\mathcal{J}_2}$, repeat the procedure, and average the results.

The S-DRL estimator of the DTE For each $c \in \{a, a'\}$ and for any $\eta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$, define the DR score function based on the DR representation (1.2):

$$\psi_c(W; \eta) := \mathbf{V}^\top \boldsymbol{\beta} + \mathbb{1}_{\{A_1=c_1\}} \frac{\mathbf{U}^\top \boldsymbol{\alpha} - \mathbf{V}^\top \boldsymbol{\beta}}{g(\mathbf{V}^\top \boldsymbol{\gamma})} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \mathbf{U}^\top \boldsymbol{\alpha}}{g(\mathbf{V}^\top \boldsymbol{\gamma}) g(\mathbf{U}^\top \boldsymbol{\delta}_c)}. \tag{1.16}$$

We propose the *sequential doubly robust Lasso* (S-DRL) estimator of θ:

$$\widehat{\theta} := N^{-1} \sum_{i=1}^{N} [\psi_a(W_i; \hat{\eta}_a) - \psi_{a'}(W_i; \hat{\eta}_{a'})],$$

where $\hat{\eta}_c := (\widehat{\boldsymbol{\alpha}}_c, \widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\gamma}}_c, \widehat{\boldsymbol{\delta}}_c)$ are the nuisance estimates of (1.10), (1.15), (1.7), and (1.8), respectively. A cross-fitting technique is used. The details are provided in Algorithm 1; see [CCD$^+$18, SRR19] where the cross-fitting leads to weaker sparsity restrictions than those without it, such as [Far15, Tan20].

1.2.2 The dynamic treatment Lasso (DTL) estimator

Algorithm 2 Dynamic Treatment Lasso (DTL)

Require: Observations $\{W_i\}_{i=1}^N$, number of cross-fitting subsets $K \geq 2$, treatment path a, and control a'.

1: For any $K \geq 2$, let $\mathcal{K} = \{1, 2, \ldots, K\}$. Randomly split $\mathcal{I} = \{1, \ldots, N\}$ into K equal-sized $|\mathcal{I}_k| = n$ with $\mathcal{I}_{-k} = \mathcal{I} \backslash \mathcal{I}_k$ and $\mathcal{W}_{-k} = \{W_i\}_{i \in \mathcal{I}_{-k}}$.

2: **for** $k = 1, 2, ..., K$ **do**

3: **for** $c \in \{a, a'\}$ **do**

4: Using $\mathcal{W}_{-k}$, construct estimates $\widehat{\boldsymbol{\gamma}}_c$, $\widehat{\boldsymbol{\delta}}_c$, and $\widehat{\boldsymbol{\alpha}}_c$ through (1.7), (1.8), and (1.10), respectively.

5: Compute $\widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}$ as (1.19) based on the training samples $\mathcal{W}_{-k}$ and the nuisance estimate $\widehat{\boldsymbol{\alpha}}_c$.

6: **end for**

7: Using the DR score (1.16) and let $\hat{\eta}_{c,\mathrm{NR}} = (\widehat{\boldsymbol{\alpha}}_c, \widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}, \widehat{\boldsymbol{\gamma}}_c, \widehat{\boldsymbol{\delta}}_c)$, compute $\check{\theta}^{(k)}_{\mathrm{DTL}}$ as

$$\check{\theta}^{(k)}_{\mathrm{DTL}} = |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} \left[\psi_a(W_i; \hat{\eta}_{a,\mathrm{NR}}) - \psi_{a'}(W_i; \hat{\eta}_{a',\mathrm{NR}})\right].$$

8: **end for** **return** The DTL estimator and the variance estimate

$$\widehat{\theta}_{\mathrm{DTL}} := K^{-1} \sum_{k \in \mathcal{K}} \check{\theta}^{(k)}_{\mathrm{DTL}}, \quad \widehat{\sigma}^2_{\mathrm{DTL}} := N^{-1} \sum_{k \in \mathcal{K}, i \in \mathcal{I}_k} \left[\psi_a(W_i; \hat{\eta}_{a,\mathrm{NR}}) - \psi_{a'}(W_i; \hat{\eta}_{a',\mathrm{NR}}) - \widehat{\theta}_{\mathrm{DTL}}\right]^2. \tag{1.17}$$

In this section, we formally define a dynamic treatment Lasso (DTL) estimator based on the DR score (1.2) of θ_a and the nested representation (1.3) of $\mu_a(\cdot)$. Here, ℓ_1-regularized nuisance estimates $\widehat{\boldsymbol{\gamma}}_a$, $\widehat{\boldsymbol{\delta}}_a$, and $\widehat{\boldsymbol{\alpha}}_a$ will be the same as before; see (1.7), (1.8), and (1.10)

above. Estimation of the first-time conditional mean model is different. Based on the best linear approximation $\nu_a^*(\mathbf{s}) = \mathbf{u}^\top \boldsymbol{\alpha}_a^*$, (1.9), of $\nu_a(\cdot)$, we introduce the following nested "best linear working model":

$$\mu_{a,\mathrm{NR}}^*(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}_{a,\mathrm{NR}}^*, \quad \boldsymbol{\beta}_{a,\mathrm{NR}}^* := \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}}{\operatorname{argmin}} E\left[\mathbb{1}_{\{A_1=a_1\}}(\mathbf{U}^\top \boldsymbol{\alpha}_a^* - \mathbf{V}^\top \boldsymbol{\beta})^2\right]. \tag{1.18}$$

Note that the two linear working models $\mu_{a,\mathrm{NR}}^*(\cdot)$ and $\mu_a^*(\cdot)$ are not necessarily the same; see Section 1.2.3 for detailed comparisons. We consider the following imputed Lasso estimate of $\boldsymbol{\beta}_{a,\mathrm{NR}}^*$, defined as $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} := \widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}(\mathcal{D}_{\mathcal{J}}, \widehat{\boldsymbol{\alpha}}_a)$ with

$$\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}}{\operatorname{argmin}} \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \mathbb{1}_{\{A_{1i}=a_1\}}\big(\mathbf{U}_i^\top \widehat{\boldsymbol{\alpha}}_a - \mathbf{V}_i^\top \boldsymbol{\beta}\big)^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1. \tag{1.19}$$

Now we introduce the *dynamic treatment Lasso* (DTL) estimator of θ:

$$\widehat{\theta}_{\mathrm{DTL}} := N^{-1}\sum_{i=1}^{N}[\psi_a(W_i; \hat{\eta}_{a,\mathrm{NR}}) - \psi_{a'}(W_i; \hat{\eta}_{a',\mathrm{NR}})],$$

where $\psi_c(\cdot;\cdot)$ is defined in (1.16) and $\hat{\eta}_{c,\mathrm{NR}} := (\widehat{\boldsymbol{\alpha}}_c, \widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}, \widehat{\boldsymbol{\gamma}}_c, \widehat{\boldsymbol{\delta}}_c)$ are the nuisance estimates as in (1.10), (1.19), (1.7), and (1.8), respectively; see Algorithm 2 for details.

1.2.3 Comparisons between the first-time working models $\mu_a^*(\cdot)$ and $\mu_{a,\mathbf{NR}}^*(\cdot)$

In the dynamic treatment setting, the relationship between the linear conditional mean function, $\mu_a(\cdot)$, and its corresponding approximations, $\mu_a^*(\cdot)$ and $\mu_{a,\mathrm{NR}}^*(\cdot)$, obtained via different identification strategies, is not straightforward. Specifically, a linear $\mu_a(\cdot)$ is only a necessary condition for correctly specified linear working models; it does not guarantee equality. Additional conditions are required to ensure that $\mu_a^*(\cdot) = \mu_a(\cdot)$ and $\mu_{a,\mathrm{NR}}^*(\cdot) = \mu_a(\cdot)$. In the following, we will discuss these necessary conditions in detail.

The working model $\mu_a^*(\cdot)$ The proposed S-DRL approach utilizes the following identification $\mu_a(\mathbf{s}_1) = E[Y^{\mathrm{DR}} \mid \mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1]$. This representation remains valid for a linear $\mu_a(\cdot)$ as long as either $\rho_a(\cdot)$ is truly logistic or $\nu_a(\cdot)$ is truly linear and not necessarily both. In this case, we also have the following equivalent expressions of $\boldsymbol{\beta}_a^*$:

$$\boldsymbol{\beta}_a^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}}{\operatorname{argmin}} E\left[(Y(a) - \mathbf{V}^\top\boldsymbol{\beta})^2 \mid A_1 = a_1\right] = \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_1+1}}{\operatorname{argmin}} E\left[(\mu_a(\mathbf{S}_1) - \mathbf{V}^\top\boldsymbol{\beta})^2 \mid A_1 = a_1\right].$$

That is, $\mu_a^*(\mathbf{s}_1) = \mathbf{v}^\top\boldsymbol{\beta}_a^*$ satisfies $\mu_a^*(\cdot) = \mu_a(\cdot)$. Hence, $\mu_a^*(\cdot) = \mu_a(\cdot)$ whenever (a) $\mu_a(\cdot)$ is a linear function and (b) either $\rho_a(\cdot)$ is a logistic function or $\nu_a(\cdot)$ is a linear function. It is worth noting that Condition (b) is already a prerequisite for the identification of DTE, as stated in Lemma 1.1. Consequently, there is no need to introduce any other conditions beyond those outlined in Condition (a).

The working model $\mu_{a,\mathbf{NR}}^*(\cdot)$ The DTL estimator relies on the nested-regression identification for which Conditions (a) and (b) above are insufficient for $\mu_{a,\mathrm{NR}}^*(\cdot) = \mu_a(\cdot)$; see Example 1.1 below. Additional conditions are needed. For instance, $\mu_{a,\mathrm{NR}}^*(\cdot) = \mu_a(\cdot)$ if we further assume the following:

(c) Under the treatment groups $(A_1, A_2) = (a_1, a_2)$ and $A_1 = a_1$, the best linear slopes are the same while regressing $Y(a)$ on $\mathbf{U}$, i.e., $\boldsymbol{\alpha}_a^* = \bar{\boldsymbol{\alpha}}_a^*$, with $\boldsymbol{\alpha}_a^*$ defined in (1.9) and

$$\bar{\boldsymbol{\alpha}}_a^* := \underset{\boldsymbol{\alpha}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} E\left[(Y(a) - \mathbf{U}^\top\boldsymbol{\alpha})^2 | A_1 = a_1\right].$$

One sufficient but not necessary condition for (c) is that the second-time conditional mean function $\nu_a(\cdot)$ is also linear; see further justifications in Section 1.8.1.

The misspecification errors of $\mu_a^*(\cdot)$ and $\mu_{a,\mathbf{NR}}^*(\cdot)$ Now we consider the case where $\mu_a(\cdot)$ is possibly non-linear and compare the misspecification (approximation) errors of the linear working models $\mu_a^*(\cdot)$ and $\mu_{a,\mathrm{NR}}^*(\cdot)$. As long as Condition (b) above holds, we have

$$\begin{aligned}
\mathrm{Err}_{\mathrm{NR}} &:= E\left[(\mu_a(\mathbf{S}_1) - \mathbf{V}^\top \boldsymbol{\beta}_{a,\mathrm{NR}}^*)^2 | A_1 = a_1\right] \\
&= E\left[(\mu_a(\mathbf{S}_1) - \mathbf{V}^\top \boldsymbol{\beta}_a^*)^2 | A_1 = a_1\right] + E\left[(\mathbf{V}^\top(\boldsymbol{\beta}_{a,\mathrm{NR}}^* - \boldsymbol{\beta}_a^*))^2 | A_1 = a_1\right] \\
&\geq \mathrm{Err}_{\mathrm{DR}} := E\left[(\mu_a(\mathbf{S}_1) - \mathbf{V}^\top \boldsymbol{\beta}_a^*)^2 | A_1 = a_1\right].
\end{aligned}$$

Hence, we have the following conclusions. (1) When $\nu_a(\cdot)$ is linear, $\mu_{a,\mathrm{NR}}^*(\cdot) = \mu_a^*(\cdot)$, and both of them are the best linear approximations of the true conditional mean $\mu_a(\cdot)$ among the group $A_1 = a_1$, i.e., $\mathrm{Err}_{\mathrm{NR}} = \mathrm{Err}_{\mathrm{DR}}$. (2) When $\nu_a(\cdot)$ is non-linear and $\rho_a(\cdot)$ is logistic, $\mu_{a,\mathrm{NR}}^*(\cdot) \neq \mu_a^*(\cdot)$ and $\boldsymbol{\beta}_{a,\mathrm{NR}}^* \neq \boldsymbol{\beta}_a^*$ in general. If Assumption 1.3 holds, the strict inequality above holds in that $\mathrm{Err}_{\mathrm{NR}} > \mathrm{Err}_{\mathrm{DR}}$ as long as $\boldsymbol{\beta}_{a,\mathrm{NR}}^* \neq \boldsymbol{\beta}_a^*$. That is, the nested "best linear approximation", $\mu_{a,\mathrm{NR}}^*(\cdot)$, is in general sub-optimal. If we further consider the case where $\mu_a(\cdot)$ is linear, then we have $\mathrm{Err}_{\mathrm{DR}} = 0$ and possibly $\mathrm{Err}_{\mathrm{NR}} > 0$; see also an illustration in Example 1.1 below.

Example 1.1 (A misspecified linear model when the truth is indeed linear)**.** *In what follows, we turn our attention to the nested-regression approach and offer an illustrative example that demonstrates how $\mu_{a,NR}^*(\cdot) \neq \mu_a(\cdot)$ for a linear $\mu_a(\cdot)$, a logistic $\rho_a(\cdot)$, a non-linear $\nu_a(\cdot)$, and a non-logistic $\pi_a(\cdot)$. To facilitate our discussion, we consider the case where both $\mathbf{S}_1$ and $\mathbf{S}_2$ are one-dimensional covariates with supports in $\mathbb{R}$. We assume $\mathbf{S}_1$ follows a uniform distribution on $[-1, 1]$, and consider an independent δ that follows a Bernoulli distribution*

with a probability of success 0.5. *We further assume*

$$P(A_1 = a_1 \mid \mathbf{S}_1) = \mathbf{S}_1/2 + 0.5, \quad \mathbf{S}_2 = \mathbf{S}_1(2\delta - 1),$$

$$P(A_2 = a_2 \mid \mathbf{S}, A_1 = a_1) = g(\mathbf{S}_2) = \exp(\mathbf{S}_2)/[1 + \exp(\mathbf{S}_2)], \quad Y(a) = 2\delta - 1.$$

It is not difficult to verify that $\mu_a(\cdot) \equiv 0$ *and* $\mu^*_{a,NR}(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}^*_{a,NR}$, *where* $\boldsymbol{\beta}^*_{a,NR} \approx (-0.09, 0.3)^\top$. *Notably, the nested-regression approach based on a linear working model* $\mu^*_{a,NR}(\cdot)$ *is misspecified, even though the true conditional mean is linear. Furthermore, because both the first-time conditional mean and propensity score models are misspecified, the DR representation of* θ_a *in Equation* (1.2) *is no longer valid when the nested-regression-based working model* $\mu^*_{a,NR}(\cdot)$ *is considered. However, since the second-time propensity score* $\rho_a(\cdot)$ *is truly logistic, the S-DRL approach leads to* $\mu^*_a(\cdot) = \mu_a(\cdot) \equiv 0$ *and the incidental validity of the DR representation of* θ_a *through Equation* (1.2).

1.2.4 The general DR DTE estimator

In this section, we present a general doubly robust (DR) estimator of the DTE. We assume that we have access to estimators $\widehat{\nu}_a(\cdot)$, $\widehat{\mu}_a(\cdot)$, $\widehat{\pi}_a(\cdot)$, and $\widehat{\rho}_a(\cdot)$ of $\nu_a(\cdot)$, $\mu_a(\cdot)$, $\pi_a(\cdot)$, and $\rho_a(\cdot)$, respectively. The functions $\nu_a(\mathbf{s})$, $\pi_a(\mathbf{s}_1)$, and $\rho_a(\mathbf{s})$ can be directly estimated using observable variables, while the remaining nuisance function $\mu_a(\cdot)$ can be identified using either the proposed DR representation (1.4) or the usual nested representation (1.3). We consider flexible estimation strategies for all nuisance functions, including both parametric and non-parametric methods. Using the DR representation of θ_a given by (1.2), we propose a general DR estimator of the DTE through a cross-fitting procedure. For any $K \geq 2$, randomly split $\mathcal{I} = \{1, \ldots, N\}$ into K equal-sized parts with $|\mathcal{I}_k| = n = N/K$. For the sake

of simplicity, we consider n as an integer. Based on the training samples $\mathcal{W}_{-k}$, construct $\widehat{\nu}_{c,-k}(\cdot)$, $\widehat{\mu}_{c,-k}(\cdot)$, $\widehat{\pi}_{c,-k}(\cdot)$, and $\widehat{\rho}_{c,-k}(\cdot)$ as estimates of the nuisance functions $\nu_c(\cdot)$, $\mu_c(\cdot)$, $\pi_c(\cdot)$, and $\rho_c(\cdot)$, respectively. For each $c \in \{a, a'\}$, let

$$\widehat{\psi}_{c,-k}(W) := \widehat{\mu}_{c,-k}(\mathbf{S}_1) + \mathbb{1}_{\{A_1=c_1\}} \frac{\widehat{\nu}_{c,-k}(\mathbf{S}) - \widehat{\mu}_{c,-k}(\mathbf{S}_1)}{\widehat{\pi}_{c,-k}(\mathbf{S}_1)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \widehat{\nu}_{c,-k}(\mathbf{S})}{\widehat{\pi}_{c,-k}(\mathbf{S}_1)\widehat{\rho}_{c,-k}(\mathbf{S})}. \quad (1.20)$$

The general DR DTE estimator and the corresponding variance estimate are then defined with $\widehat{\Delta}_{-k}(\cdot) = \widehat{\psi}_{a,-k}(\cdot) - \widehat{\psi}_{a',-k}(\cdot)$ and $\mathcal{K} = \{1, \cdots, K\}$ as

$$\widehat{\theta}_{\text{gen}} := \frac{1}{N} \sum_{k \in \mathcal{K}, i \in \mathcal{I}_k} \widehat{\Delta}_{-k}(W_i), \qquad \widehat{\sigma}^2_{\text{gen}} := \frac{1}{N} \sum_{k \in \mathcal{K}, i \in \mathcal{I}_k} [\widehat{\Delta}_{-k}(W_i) - \widehat{\theta}_{\text{gen}}]^2. \quad (1.21)$$

1.3 Asymptotic properties

Here we establish consistency and asymptotic normality of the S-DRL, DTL, and the general DR estimator.

1.3.1 Properties of the S-DRL estimator

We use $s_{\boldsymbol{\alpha}_a} := \|\boldsymbol{\alpha}_a^*\|_0$, $s_{\boldsymbol{\beta}_a} := \|\boldsymbol{\beta}_a^*\|_0$, $s_{\boldsymbol{\gamma}_a} := \|\boldsymbol{\gamma}_a^*\|_0$, and $s_{\boldsymbol{\delta}_a} := \|\boldsymbol{\delta}_a^*\|_0$ to denote sparsity levels of the nuisance parameters as defined in (1.9) , (1.13), (1.5) and (1.6), respectively. The number of covariates, d_1 and d, are possibly much larger than N; for simplicity, we consider $d_1 \asymp d_2 \asymp d := d_1 + d_2$.

Assumption 1.2. *Define* $\zeta_a := \mathbb{1}_{\{A_1=a_1, A_2=a_2\}}(Y(a) - \nu_a^*(\mathbf{S}))$, $\varepsilon_a := \mathbb{1}_{\{A_1=a_1\}}(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))$ *and let* $\zeta := \zeta_a + \zeta_{a'}$, $\varepsilon := \varepsilon_a + \varepsilon_{a'}$. *Suppose that there exist positive constants* $\sigma_\zeta < \infty$ *and* $\sigma_\varepsilon < \infty$, *such that* ζ *and* ε *are sub-Gaussian, with* $\|\zeta\|_{\psi_2} \le \sigma\sigma_\zeta$, $\|\varepsilon\|_{\psi_2} \le \sigma\sigma_\varepsilon$, *and*

$$\sigma^2 := E[\psi_a(W; \eta_a^*) - \psi_{a'}(W; \eta_{a'}^*) - \theta]^2, \quad (1.22)$$

where $\psi_c(\cdot;\cdot)$ *is defined in* (1.16) *and* $\eta_c^* := (\boldsymbol{\alpha}_c^*, \boldsymbol{\beta}_c^*, \boldsymbol{\gamma}_c^*, \boldsymbol{\delta}_c^*)$.

Assumption 1.3. *Let* $\mathbf{U}$ *be a sub-Gaussian vector such that* $\|\boldsymbol{x}^\top \mathbf{U}\|_{\psi_2} \leq \sigma_u \|\boldsymbol{x}\|_2$ *for* $\boldsymbol{x} \in \mathbb{R}^{d+1}$ *and* $\sigma_u > 0$. *Let* $\lambda_{\min}(E[\mathbf{U}\mathbf{U}^\top \mathbb{1}_{\{A_1=a_1\}}]) \geq \kappa_l$ *for any* $a_1 \in \{0,1\}$, *with* $\kappa_l > 0$.

Assumptions 1.2 and 1.3 are fairly general even among the high-dimensional literature. As $N \to \infty$, we allow ψ_2-norm bounds of ζ and ε to diverge or to shrink to zero. When all the nuisance models are correctly specified, under the overlap condition in Assumption 1.1, $\sigma^2 \asymp E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \geq \max\{E[\zeta^2], E[\varepsilon^2]\}$, where $\xi := \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$ denotes the centered conditional effect at the first exposure. A sufficient condition for Assumption 1.2 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 ψ_2-norms. Note that, we allow $\sigma = \sigma_N$ to be dependent on N while assuming σ_ζ and σ_ε to be constants independent of N; $\sigma \to 0$ and $\sigma \to \infty$ are both allowed as $N \to \infty$. The following Assumption 1.4 is an overlap condition for the working propensity score models, which is additionally required only when model misspecification occurs.

Assumption 1.4. *Let* $\pi_a^*(\cdot)$ *and* $\rho_a^*(\cdot)$ *be such that* $P(c_0 \leq \pi_a^*(\mathbf{S}_1) \leq 1 - c_0) = 1$, $P(c_0 \leq \rho_a^*(\mathbf{S}) \leq 1 - c_0) = 1$, *for a fixed constant* $c_0 > 0$.

The following theorem characterizes the consistency rate of the S-DRL estimator of θ.

Theorem 1.2 (Consistency of the S-DRL)**.** *Suppose that at least one of* $\mu_a^*(\cdot)$ *and* $\pi_a^*(\cdot)$ *is correctly specified, and at least one of the models* $\nu_a^*(\cdot)$ *and* $\rho_a^*(\cdot)$ *is correctly specified. Let Assumptions 1.1-1.4 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$, *and either (a)* $\|\mathbf{S}_1\|_\infty \leq C$ *almost surely, with a constant* $C > 0$, *or (b)* $s_{\boldsymbol{\delta}_a} \log^2(d) = O(N)$. *Then the*

sequential DR Lasso (S-DRL) estimator, $\widehat{\theta}$, as defined in Algorithm 1, satisfies

$$\widehat{\theta} - \theta = O_p\left(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\right),$$

as $N, d \to \infty$, with $s_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}}\}$ and $s_2' := \max\{s_{\boldsymbol{\alpha}_a}\mathbb{1}_{\{\rho_a^* \neq \rho_a\}}, s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi_a^* \neq \pi_a\}}, s_{\boldsymbol{\gamma}_a}\mathbb{1}_{\{\mu_a^* \neq \mu_a\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu_a^* \neq \nu_a\}}\}$.

We categorize bounded and unbounded covariate support and add a $\log(d)$ restriction to $s_{\boldsymbol{\delta}_a}$ for the latter. The DR imputed outcome (1.14) has an unbounded ψ_α-Orlicz norm for any $\alpha > 0$. Yet, if $\mathbf{S}_1$ has bounded support, no extra sparsity condition is required as the inverse probability weighting is stable and the DR imputation has a well-behaved tail distribution.

Remark 1.1 (Comparison with low-dimensional DR DTE estimators)**.** *[LS21] applied debiased machine learning and g-estimation techniques in the framework of SNM models. However, the "blip functions" $\gamma_1(\mathbf{s}_1, a_1)$ and $\gamma_2(\mathbf{s}, a_1, a_2)$ – which are defined as $E[Y(a_1, a_2) - Y(0, a_2)|A_1 = a_1, \mathbf{S}_1 = \mathbf{s}_1]$ and $E[Y(a_1, a_2) - Y(a_1, 0)|A_1 = a_1, A_2 = a_2, \mathbf{S} = \mathbf{s}]$, respectively – are considered low-dimensional and correctly specified for consistent estimation. MSM models with low-dimensional confounders have been studied extensively, with significant theoretical advancements made in the seminal work of [TS12]. Additionally, [Rob00b], [MvdLRG01], [BR05], and [YvdL06] explored DR DTE estimation with low-dimensional nuisances, proposing consistent and asymptotically normal DTE estimators given that either (a) all conditional mean models are correctly parametrized or (b) all propensity score models are correctly parametrized. More recently, [BRR19] proposed a multiple robust (MR) estimator that allows for an additional model misspecification scenario (c), where only the first-time*

conditional mean model and second-time propensity score model are correctly parametrized. However, all of the aforementioned approaches require low-dimensional, parametric nuisance estimates that are root-N consistent.

Our proposed method accommodates high-dimensional and possibly non-parametric nuisance estimates, which may not necessarily be root-N consistent. This approach allows for a consistent estimate of the DTE even in challenging scenarios where only the second-time conditional mean and first-time propensity score models are correctly specified (misspecification scenario (d)). This scenario is more common in practice due to the difficulty of identifying the first-time conditional mean model. When $\nu_a(\mathbf{s}) = \nu_a^*(\mathbf{s}) = \mathbf{u}^\top \boldsymbol{\alpha}_a^*$ *is linear, as per* (1.3)*, a linear* $\mu_a(\cdot)$ *would require* $E[\mathbf{U}^\top \boldsymbol{\alpha}_a^* | \mathbf{S}_1, A_1 = a_1]$ *to be linear in* $\mathbf{S}_1$ *– an unlikely scenario if any of the* $\mathbf{S}_2$ *are binary or discrete.*

When all the nuisance models are correctly specified, we further establish asymptotic normality results and the corresponding rate DR property of the S-DRL estimator.

Theorem 1.3 (Asymptotic normality of the S-DRL)**.** *Suppose that all the nuisance models* $\mu_a^*(\cdot)$*,* $\nu_a^*(\cdot)$*,* $\pi_a^*(\cdot)$*, and* $\rho_a^*(\cdot)$ *are correctly specified. Let Assumptions 1.1-1.3 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$*, and either (a)* $\|\mathbf{S}_1\|_\infty \leq C$ *almost surely, with some constant* $C > 0$*, or (b)* $s_{\boldsymbol{\delta}_a} \log^2(d) = O(N)$*. Additionally, assume the following product-rate condition:*

$$\max\{s_{\boldsymbol{\gamma}_a} s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}\} \log^2(d) = o(N). \tag{1.23}$$

Then, with σ^2 *in* (1.22) *and* $\widehat{\sigma}^2$ *in* (1.12)*, the S-DRL estimator satisfies* $\sigma^{-1}\sqrt{N}(\widehat{\theta} - \theta) \rightsquigarrow N(0,1)$ *and* $\hat{\sigma}^{-1}\sqrt{N}(\widehat{\theta} - \theta) \rightsquigarrow N(0,1)$ *as* $N, d \to \infty$*.*

Theorem 1.3, as per [BR05], indicates that the S-DRL estimator achieves the semi-parametric efficiency bound when all nuisance models are correctly specified.

Remark 1.2 (Comparison with static ATE estimators)**.** *We compare sparsity conditions in Theorem 1.3 with the literature on estimating ATE through DR for a single exposure. The ATE can be seen as a special case of the DTE where we assume that* $\mathbf{S}_1$ *and* A_1 *are completely random. This allows root-N estimation of* $\mu_a(\cdot)$ *and* $\pi_a(\cdot)$*. Consequently, Theorem 1.3 requires* $s_{\boldsymbol{\alpha}_a} + s_{\boldsymbol{\delta}_a} = o(N/\log(d))$ *and* $s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a} = o(N/\log^2(d))$*, which are less restrictive than [Far15, Tan20, DV21, DAV20, AV21] and are aligned with [CCD+18, SRR19].*

1.3.2 Properties of the DTL estimator

With slight abuse of notation, let $s_{\boldsymbol{\beta}_a} = \|\boldsymbol{\beta}^*_{a,\mathrm{NR}}\|_0$.

Theorem 1.4 (Consistency of the DTL)**.** *Suppose that at least one of* $\mu^*_{a,NR}(\cdot)$ *and* $\pi^*_a(\cdot)$ *is correctly specified, and at least one of the models* $\nu^*_a(\cdot)$ *and* $\rho^*_a(\cdot)$ *is correctly specified. Let Assumptions 1.1- 1.4 hold with* $\mu^*_a(\cdot)$ *and* $\boldsymbol{\beta}^*_a$ *replaced by* $\mu^*_{a,NR}(\cdot)$ *and* $\boldsymbol{\beta}^*_{a,NR}$*. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$*. Then the DTL estimator satisfies, as* $N, d \to \infty$*,*

$$\widehat{\theta}_{DTL} - \theta = O_p\left(\sigma\frac{s'_1\log(d)}{N} + \sigma\sqrt{\frac{s'_2\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\right), \tag{1.24}$$

with $s'_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\gamma}_a}}, \sqrt{s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a}s_{\boldsymbol{\gamma}_a}}\}$

and $s'_2 := \max\{s_{\boldsymbol{\alpha}_a}\mathbb{1}_{\{\pi^*_a\neq\pi_a \text{ or } \rho^*_a\neq\rho_a\}},\ s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi^*_a\neq\pi_a\}}, s_{\boldsymbol{\gamma}_a}\mathbb{1}_{\{\mu^*_{a,NR}\neq\mu_a\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu^*_a\neq\nu_a\}}\}$.

Theorem 1.5 (Asymptotic normality of the DTL)**.** *Suppose that all the nuisance models* $\mu^*_{a,NR}(\cdot)$*,* $\nu^*_a(\cdot)$*,* $\pi^*_a(\cdot)$*, and* $\rho^*_a(\cdot)$ *are correctly specified. Let Assumptions 1.1-1.3 hold with* $\mu^*_a(\cdot)$ *and* $\boldsymbol{\beta}^*_a$ *replaced by* $\mu^*_{a,NR}(\cdot)$ *and* $\boldsymbol{\beta}^*_{a,NR}$*. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$*.*

Additionally, assume the following product-rate condition:

$$\max\{s_{\boldsymbol{\gamma}_a} s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\gamma}_a} s_{\boldsymbol{\alpha}_a}\} \log^2(d) = o(N). \tag{1.25}$$

Then, with σ^2 *in* (1.22) *and* $\widehat{\sigma}^2_{DTL}$ *in* (1.17)*, the DTL estimator satisfies* $\sigma^{-1}\sqrt{N}(\widehat{\theta}_{DTL} - \theta) \rightsquigarrow N(0,1)$ *and* $\hat{\sigma}^{-1}_{DTL}\sqrt{N}(\widehat{\theta}_{DTL} - \theta) \rightsquigarrow N(0,1)$ *as* $N, d \to \infty$.

Remark 1.3 (Comparisons between the S-DRL and DTL estimators)**.**

Consistency. *Theorems 1.2 and 1.4 establish the consistency of two distinct estimators, S-DRL and DTL. While both estimators necessitate correct specification of at least one of* $\nu_a(\cdot)$ *and* $\rho_a(\cdot)$*, their requirements on* $\mu_a(\cdot)$ *and* $\pi_a(\cdot)$ *differ due to their distinct conditional mean models. Specifically, S-DRL mandates either* $\mu^*_a(\cdot) = \mu_a(\cdot)$ *or* $\pi^*_a(\cdot) = \pi_a(\cdot)$*, which can be guaranteed by the linearity of* $\mu_a(\cdot)$ *or the logistic form of* $\pi_a(\cdot)$*. In contrast, DTL imposes the stricter condition of either* $\mu^*_{a,NR}(\cdot) = \mu_a(\cdot)$ *or* $\pi^*_a(\cdot) = \pi_a(\cdot)$*, which may not be fulfilled even when* $\mu_a(\cdot)$ *is linear, as discussed in Section 1.2.3. For a comprehensive summary, see Table 1.1.*

Rate of estimation. *Different rate of estimation of S-DRL and DTL are presented in Table 1.2. Table 1.2 reveals a symmetrical pattern in the rates of S-DRL, whereas DTL exhibits an asymmetric behavior – the sparsity levels* $s_{\boldsymbol{\alpha}_a}$ *and* $s_{\boldsymbol{\gamma}_a}$ *appear to be more influential than* $s_{\boldsymbol{\beta}_a}$ *and* $s_{\boldsymbol{\delta}_a}$*. When either* $\rho_a(\cdot)$ *or* $\mu_a(\cdot)$ *are misspecified, there is no difference in the rates. However, when they are both correctly specified, the rate of DTL contains additional terms that involve the sparsity level* $s_{\boldsymbol{\alpha}_a}$ *(and* $s_{\boldsymbol{\gamma}_a}$ *under certain circumstances). Notably, if* $s_{\boldsymbol{\alpha}_a}$ *is relatively large, S-DRL exhibits a faster consistency rate than DTL.*

Asymptotic normality. *The S-DRL and DTL estimators are both asymptotically normal, as proven in Theorems 1.3 and 1.5. When* $\mu^*_{a,NR}(\cdot) = \mu_a(\cdot) = \mu^*_a(\cdot)$*, their asymptotic efficiency*

is the same. However, they require different sparsity conditions. The DTL estimator requires three product-sparsity conditions, specifically: (a) the first-time conditional mean $\mu_a(\cdot)$ and first-time propensity score $\pi_a(\cdot)$, (b) the second-time conditional mean $\nu_a(\cdot)$ and second-time propensity score $\rho_a(\cdot)$, and (c) the second-time conditional mean $\nu_a(\cdot)$ and first-time propensity score $\pi_a(\cdot)$, which are given in equation (1.25)*. On the other hand, the S-DRL estimator only requires two product-sparsity conditions, as defined in* (1.23)*. These correspond to (a) and (b) above, with (c) becoming irrelevant. The S-DRL estimator is known as (sequential) rate DR because it is asymptotically normal when the product-sparsity of the nuisance parameters is $o(N/\log^2(d))$ for each exposure time; see more details in Remark 1.2.*

1.3.3 Properties of the general DR estimator

In this section, we provide a new consistency result of the general DR DTE estimator. Here we consider arbitrary working models $\pi_a^*(\cdot)$, $\rho_a^*(\cdot)$, $\mu_a^*(\cdot)$, and $\nu_a^*(\cdot)$, which may not follow the logistic or linear forms as before. For each $c \in \{a, a'\}$, define the corresponding DR score function as

$$\psi_c^*(W) := \mu_c^*(\mathbf{S}_1) + \mathbb{1}_{\{A_1=c_1\}} \frac{\nu_c^*(\mathbf{S}) - \mu_c^*(\mathbf{S}_1)}{\pi_c^*(\mathbf{S}_1)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \nu_c^*(\mathbf{S})}{\pi_c^*(\mathbf{S}_1)\rho_c^*(\mathbf{S})}, \tag{1.26}$$

with

$$\sigma^2 := E\left[\psi_a^*(W) - \psi_{a'}^*(W) - \theta\right]^2. \tag{1.27}$$

Assumption 1.5. *For positive sequences $a_N = o(\sigma)$, $b_N = o(\sigma)$, $c_N = o(1)$, and $d_N = o(1)$, let $E[\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})]^2 = O_p(a_N^2)$, $E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^2 = O_p(b_N^2)$, $E[\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1)]^2 = O_p(c_N^2)$, and $E[\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S})]^2 = O_p(d_N^2)$. Moreover, for $c_0 \in (0, 1/2)$, $P(c_0 \leq \widehat{\pi}_a(\mathbf{S}_1) \leq$*

Table 1.2: Consistency rates of $\widehat{\theta}$ and $\widehat{\theta}_{\mathrm{DTL}}$ under various misspecification settings when the conditions in Theorems 1.2 and 1.4 are satisfied with $\sigma \asymp 1$. Misspecified and correctly specified models are denoted by ✗ and ✓, respectively. LHS below corresponds to the consistency rate provided to the left of that position.

Correctly specified models				Consistency rate of $\widehat{\theta}$	Consistency rate of $\widehat{\theta}_{\mathrm{DTL}}$
$\rho_a(\cdot)$	$\pi_a(\cdot)$	$\mu_a(\cdot)$	$\nu_a(\cdot)$		
✓	✓	✓	✓	$\frac{1}{\sqrt{N}} + \frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}} \log d}{N} + \frac{\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}} \log d}{N}$	LHS $+ \frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}_a}} \log d}{N}$
✗	✓	✓	✓	$\frac{\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log d}{N}}$	LHS
✓	✗	✓	✓	$\frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\beta}_a} \log d}{N}}$	$\sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\beta}_a} \log d}{N}}$
✓	✓	✗	✓	$\frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\gamma}_a} \log d}{N}}$	LHS
✓	✓	✓	✗	$\frac{\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}} \log d}{N} + \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log d}{N}}$	LHS $+ \frac{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}_a}} \log d}{N}$ *
✗	✗	✓	✓	$\sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\beta}_a} \log d}{N}}$	LHS
✗	✓	✗	✓	$\sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\gamma}_a} \log d}{N}}$	LHS
✓	✗	✓	✗	$\sqrt{\frac{s_{\boldsymbol{\beta}_a} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log d}{N}}$	LHS $+ \sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log d}{N}}$ **
✓	✓	✗	✗	$\sqrt{\frac{s_{\boldsymbol{\gamma}_a} \log d}{N}} + \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log d}{N}}$	LHS

* This consistency rate requires $\mu^*_{a,\mathrm{NR}}(\cdot) = \mu_a(\cdot)$. Without it, the DTL's rate is equal to the last row of the above table; see Section 1.2.3.

** This consistency rate requires $\mu^*_{a,\mathrm{NR}}(\cdot) = \mu_a(\cdot)$. Otherwise, the DTL is inconsistent; see Table 1.1.

$1 - c_0) = 1$ *and* $P(c_0 \leq \widehat{\rho}_a(\mathbf{S}) \leq 1 - c_0) = 1$ *with probability approaching one. For* ζ *and* ε *defined in Assumption 1.2,* $\max\{E|\zeta|^q/[E|\zeta|^2]^{q/2}, E|\varepsilon|^q/[E|\varepsilon|^2]^{q/2}, E|\xi|^q/[E|\xi|^2]^{q/2}\} \leq C$, $P(E[\zeta^2|\mathbf{S}] \leq CE[\zeta^2]) = 1$, *and* $P(E[\varepsilon^2|\mathbf{S}_1] \leq CE[\varepsilon^2]) = 1$, *for constants* $C > 0$ *and* $q > 2$.

The probability measures and corresponding expectations above are with respect to a fresh draw $\mathbf{S}$ (or $\mathbf{S}_1$). Note that Assumption 1.5 allows for $\rho_a^*(\cdot)$ to differ from $\rho_a(\cdot)$ while requiring a overlap condition consistent with the existing literature; see, e.g., [CCD$^+$18]. The max condition, satisfied by sub-Gaussian random variables, controls the tails of ζ, ε, and ξ. The last two conditions of Assumption 1.5 aim to ensure the interpretability of the results by bounding the "normalized" conditional second moments.

Theorem 1.6. *(Consistency of the general DR estimator) Suppose that at least one of* $\mu_a^*(\cdot)$ *and* $\pi_a^*(\cdot)$ *is correctly specified, and at least one of* $\nu_a^*(\cdot)$ *and* $\rho_a^*(\cdot)$ *is correctly specified. Let Assumptions 1.1, 1.4, and 1.5 hold. Additionally, let* $E[\mathbb{1}_{\{A_1=a_1\}}(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2] \leq C_\mu \sigma^2$, *for some* $C_\mu > 0$. *Then the general DR estimator,* $\widehat{\theta}_{\text{gen}}$, *satisfies* $\widehat{\theta}_{\text{gen}} - \theta = O_p(q_N)$ *as* $N, d \to \infty$, *where* $q_N = b_N c_N + a_N d_N + b_N \mathbb{1}_{\{\pi_a^* \neq \pi_a\}} + a_N \mathbb{1}_{\{\rho_a^* \neq \rho_a\}} + c_N \sigma \mathbb{1}_{\{\mu_a^* \neq \mu_a\}} + d_N \sigma \mathbb{1}_{\{\nu_a^* \neq \nu_a\}} + \sigma/\sqrt{N}$.

The aforementioned theorem yields two distinct conclusions that warrant discussion. The first pertains to the conditions that are necessary for achieving root-N consistency, while the second relates to the issue of consistency under model misspecification. If all the models are correctly specified, $\widehat{\theta}_{\text{gen}} - \theta = O_p(b_N c_N + a_N d_N + \sigma N^{-1/2})$ and root-N consistency happens as long as $b_N c_N + a_N d_N = O(N^{-1/2})$ and $\sigma = O(1)$.

If, on the other hand, at least one of the nuisance models is correctly specified at each exposure time, $\widehat{\theta}_{\text{gen}}$ is a consistent estimator as long as $\sigma = O(1)$. Model misspecification can take an asymmetric form in terms of estimation rates. Specifically, while q_N is symmetric

in the rates themselves, the dependence of b_N on a_N and/or d_N can introduce potential asymmetries. For example, when performing ℓ_1-regularized nested regression, Theorem 1.10 indicates that b_N depends additively on a_N as $b_N = b_N^* + a_N$, where $\{b_N^*\}^2 = s_{\boldsymbol{\beta}_a} \log(d)/N$ is the estimation error of $\mu_a(\cdot)$ when $\nu_a(\cdot)$ is known. As a result, the consistency rate of $\widehat{\theta}_{\text{gen}}$ includes an additional term $a_N c_N + a_N \mathbb{1}_{\{\pi_a^* \neq \pi_a\}}$, as illustrated by the DTL estimator in (1.24).

On the other hand, if we consider a new DR approach based on the DR representation (1.4) to estimate $\mu_a(\cdot)$, the corresponding b_N will depend on both a_N and d_N. For instance, when all the nuisance models are correctly specified, Theorem 1.9 indicates that ℓ_1-regularized DR estimation leads to a symmetric rate with $b_N = b_N^* + a_N d_N$, resulting in $\widehat{\theta}_{\text{gen}} - \theta = O_p(b_N^* c_N + a_N d_N + 1/\sqrt{N})$ if $\sigma \asymp 1$ and $a_N, b_N^*, c_N, d_N = o(1)$. The approach used to estimate the first-time conditional mean $\mu_a(\cdot)$ determines the persistence of the symmetry.

Theorem 1.7. *(Asymptotic normality of the general DR estimator) Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Whenever Assumptions 1.1, 1.5 hold and the rates of estimation satisfy the following product conditions*

$$b_N c_N = o(\sigma N^{-1/2}), \qquad a_N d_N = o(\sigma N^{-1/2}), \tag{1.28}$$

then the estimator $\widehat{\theta}_{\text{gen}}$ satisfies $\sigma^{-1}\sqrt{N}(\widehat{\theta}_{\text{gen}} - \theta) \rightsquigarrow N(0,1)$ and $\widehat{\sigma}_{\text{gen}}^{-1}\sqrt{N}(\widehat{\theta}_{\text{gen}} - \theta) \rightsquigarrow N(0,1)$ as $N \to \infty$ (and potentially $d \to \infty$), where σ^2 and $\widehat{\sigma}_{\text{gen}}^2$ are defined in (1.27) and (1.21), respectively.

Remark 1.4 (Rate double robustness)**.** *The topic of rate double robustness in the presence of multiple exposures has been addressed in [BHL22], but the authors require three product-rate conditions, including $a_N c_N = o(N^{-1/2})$ as stated in their Assumption 3.1, in addition*

to the two product-rates (1.28). *As a result, this does not allow for relatively large values of* a_N *and* c_N*, which is permissible in our setting. For example, consider a special case where* A_2 *is completely random and* $\mu_a(\cdot)$ *is a constant function. In conjunction with the sequential ignorability condition of Assumption 1.1, we have* $\mu_a(\mathbf{s}_1) = E[Y|\mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1, A_2 = a_2]$*, allowing* $\mu_a(\cdot)$ *to be identified directly through observable variables. In this scenario, both* $\mu_a(\cdot)$ *and* $\rho_a(\cdot)$ *can be estimated with a root-N rate. Therefore, Theorem 1.7 only requires* $a_N + c_N = o(1)$*, i.e.,* $\nu_a(\cdot)$ *and* $\pi_a(\cdot)$ *are consistently estimated. In comparison, [BHL22] additionally require* $a_N c_N = o(N^{-1/2})$*, which may not be feasible when* $\nu_a(\cdot)$ *and* $\pi_a(\cdot)$ *are only known to be Lipschitz continuous, and the covariate dimensions satisfy* $d \geq d_1 > 2$*.*

Our proof relies on a nuanced decomposition of the second-order estimation bias resulting from the estimation errors of $\widehat{\nu}_{c,-k}(\cdot)$ *and* $\widehat{\pi}_{c,-k}(\cdot)$*. Leveraging the Neyman orthogonality of the DR score* (1.26)*, we reformulate the second-order bias as the product* $E[\mathbb{1}_{\{A_1=c_1\}}(1-\mathbb{1}_{\{A_2=c_2\}}/\rho_c^*(\mathbf{S}))(\widehat{\nu}_{c,-k}(\mathbf{S})/\widehat{\pi}_{c,-k}(\mathbf{S}_1)-\nu_c^*(\mathbf{S})/\pi_c^*(\mathbf{S}_1))]$*. In our analysis, we then examine this product collectively rather than as separate terms, resulting in a cohesive flow of the arguments. We show that the population effect of this term is exactly zero whenever the model* $\rho_c^*(\cdot)$ *is correctly specified – a condition fullfilled when discussing asymptotic normality in high-dimensional regimes. We then showcase that the sample equivalent is negligible and does not contribute to the estimation error.*

1.4 Supporting theoretical discoveries

This section presents supplementary findings that, while not the primary focus of the research, may nonetheless be informative or valuable.

1.4.1 An adaptive theory for imputed Lasso with high-dimensional covariates

Let $\mathbb{S} := (Y_i^*, \mathbf{X}_i)_{i=1}^M$ be i.i.d. observations and let $(Y^*, \mathbf{X})$ be an independent copy with $Y^* \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^d$. Suppose that there exists, possibly random, $\widehat{Y}_i \in \mathbb{R}$. Note that for some, and possibly all observations, outcomes Y^* are imputed, i.e., estimated using $\hat{Y}_i$. The true population slope is defined as $\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} E[Y^* - \mathbf{X}^\top\boldsymbol{\beta}]^2$. Then its estimator is

$$\widehat{\boldsymbol{\beta}} \;:=\; \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ M^{-1}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^\top\boldsymbol{\beta}]^2 + \lambda_M\|\boldsymbol{\beta}\|_1 \right\}, \tag{1.29}$$

for $\lambda_M > 0$. The following result delineates properties of such imputed-Lasso estimator, $\widehat{\boldsymbol{\beta}}$.

Theorem 1.8 (General imputed Lasso estimators)**.** *Let $s = \|\boldsymbol{\beta}^*\|_0$ and $\varepsilon_i := Y_i^* - \mathbf{X}_i^\top\boldsymbol{\beta}^*$. Suppose that $\|\boldsymbol{a}^\top\mathbf{X}\|_{\psi_2} \le \sigma_{\mathbf{X}}\|\boldsymbol{a}\|_2$ for $\boldsymbol{a} \in \mathbb{R}^d$, $\lambda_{\min}(E[\mathbf{X}\mathbf{X}^\top]) > \lambda_{\mathbf{X}}$, and $\|\varepsilon\|_{\psi_2} \le \sigma$ with $\sigma_{\mathbf{X}}, \lambda_{\mathbf{X}} > 0$ and $\sigma = \sigma_M > 0$ potentially dependent on M. For some $\delta_M > 0$, define the event $\mathcal{E}_1 := \{M^{-1}\sum_{i=1}^M[\widehat{Y}_i - Y_i^*]^2 < \delta_M^2\}$. 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*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \max\left(\frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M, 8\kappa_1^{-1}\sqrt{s}\lambda_M\right),$$

with probability at least $1 - 2\exp(-\frac{4Mt^2}{1+2t+\sqrt{2}t}) - c_1\exp(-c_2M)$, where $\kappa_1, \kappa_2, c_1, c_2 > 0$ are 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 $\lambda_M \asymp \sigma\sqrt{\log(d)/M}$, as $M, d \to \infty$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\Big(\sigma\sqrt{\frac{s\log(d)}{M}} + \delta_M\Big). \tag{1.30}$$

The above result is of independent interests as it provides a general theory for any Lasso estimators based on imputed outcomes. It contributes to the literature in three specific

aspects: (a) The "imputation error", $\widehat{Y}_i - Y_i^*$, can be dependent on and even possibly correlated with covariates $\mathbf{X}_i$; (b) We allow every $\widehat{Y}_i$ 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; and (c) The tuning parameter λ_M is of the same order as the one chosen for the fully observed data and is independent of the imputation error or any sparsity parameter.

Compared with the existing literature, Theorem 1.8 requires weaker sparsity assumptions and provides better rates of estimation. Imputed Lasso of [ZZS19] requires $s = o(M)$, $\log(d) = o(\sqrt{M})$, and $\sqrt{s}\delta_M = o(1)$. That of [LS21] requires an ultra-sparse setup $s^2 \log(d) = o(M)$ and $s\delta_M = o(1)$. In contrast, we only require $s \log(d) = o(M)$ and $\delta_M = o(1)$. Additionally, [ZZS19] choose a tuning parameter $\lambda_M \gg \sqrt{s}\delta_M$ and provide $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{s}\delta_M + \sqrt{s/M})$ upon requiring strict conditions for assuring model selection consistency; see Theorem 2 therein. [LS21] take $\lambda_M \asymp \sqrt{\log(d)/M} + \delta_M$ and establish $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(s\sqrt{\log(d)/M} + s\delta_M)$; see Theorem 13 therein. In contrast, we allow $\lambda_M \asymp \sqrt{\log(d)/M}$. The imputation error δ_M only appears in our final estimation rate (1.30) additively, and its effect does not explode as the sparsity level grows.

Theorem 1.8 requires development of new proof techniques: the standard Lasso inequality followed by the cone-set reduction are not valid in this instance. In fact, the error, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, no longer belongs to the accustomed cone set, $\mathcal{C}(S, k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq k\|\boldsymbol{\Delta}_S\|_1\}$. Instead, we identify a new set, $\widetilde{\mathcal{C}}(S) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 16\lambda_M^{-1}\delta_M^2, \|\boldsymbol{\Delta}_S\|_1 \leq 4\lambda_M^{-1}\delta_M^2\}$, and show that the error vector belongs to the union of the above two sets. This enables us to avoid choosing a tuning parameter dependent on the imputation error, as is done in the above literature. Moreover, our results are adaptive to the imputation error in

that when there is no imputation, i.e., $\delta_M = 0$, our result reaches the standard consistency rate in the high-dimensional statistics literature, e.g., [BRT09, NRWY12, Wai19].

1.4.2 Theoretical characteristics of nuisance estimators with imputed outcomes

As a result of constraints on the length of the main file, we have included the theoretical properties of the nuisance estimates $\widehat{\boldsymbol{\alpha}}_a$, $\widehat{\boldsymbol{\gamma}}_a$, and $\widehat{\boldsymbol{\delta}}_a$ as defined by equations (1.10), (1.7), and (1.8) respectively, in the Section 1.8.1, where we show $\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 = O_p(\sigma\sqrt{s_{\boldsymbol{\alpha}_a}\log(d)/N})$, $\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2 = O_p(\sqrt{s_{\boldsymbol{\gamma}_a}\log(d_1)/N})$, and $\|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2 = O_p(\sqrt{s_{\boldsymbol{\delta}_a}\log(d)/N})$. Now we establish the properties of the first-time conditional mean model estimates, where imputation is required. We first consider the DR-imputation-based estimator $\widehat{\boldsymbol{\beta}}_a$ defined as (1.11) and the corresponding conditional mean estimate $\widehat{\mu}_a(\mathbf{s}_1) = \mathbf{v}^\top \widehat{\boldsymbol{\beta}}_a$.

Theorem 1.9. *Let Assumptions 1.1-1.4 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}\log(d),\ s_{\boldsymbol{\beta}_a}\log(d_1),\ s_{\boldsymbol{\delta}_a}\log(d)\} = o(N)$*, and either (a)* $\|\mathbf{S}_1\|_\infty \le C$ *almost surely, with some constant* $C > 0$*, or (b)* $s_{\boldsymbol{\delta}_a}\log(d_1)\log(d) = O(N)$*. Choose some* $\lambda_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\log(d)/N}$*,* $\lambda_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\log(d_1)/N}$*, and* $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log(d)/N}$*. Then for any constant* $r \ge 1$*, as* $N, d \to \infty$*, we have*

$$\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2 + \{E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^r\}^{1/r} = O_p\left(r_n\right),$$

with r_n *being determined as follows (a) whenever* $\rho_a^*(\cdot) = \rho_a(\cdot)$ *and* $\nu_a^*(\cdot) = \nu_a(\cdot)$*,* $r_n = \sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \frac{\sigma\sqrt{s_{\boldsymbol{\delta}_a}s_{\boldsymbol{\alpha}_a}}\log(d)}{N}$*, (b) whenever* $\rho_a^*(\cdot) = \rho_a(\cdot)$*,* $r_n = \sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \sigma\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}$*, (c) or whenever* $\nu_a^*(\cdot) = \nu_a(\cdot)$*,* $r_n = \sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}$*.*

Theorem 1.9 elucidates that the consistency rate of $\widehat{\boldsymbol{\beta}}_a$ is subject to the fidelity of

the second-time nuisance models $\rho_a^*(\cdot)$ and $\nu_a^*(\cdot)$. More specifically, when both models are accurately specified, the DR imputation error contributes multiplicatively to the consistency rate in Theorem 1.9(a). In contrast, when only one of $\rho_a^*(\cdot)$ and $\nu_a^*(\cdot)$ is correctly specified, the estimation error of the correctly specified model contributes additively to the rates presented in Theorem 1.9(b) and Theorem 1.9(c). It is noteworthy that these results do not rely on the correctness of the first-time conditional mean model per se.

In the following, we also provide the consistency results of the nested-regression-based estimator, $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$, defined in Equation (1.19), and the corresponding conditional mean estimate $\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{s}_1) = \mathbf{v}^\top \widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$.

Theorem 1.10. *Let Assumptions 1.1-1.3 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}\log(d),\, s_{\boldsymbol{\beta}_a}\log(d_1)\} = o(N)$. *Choose some* $\lambda_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\log(d)/N}$ *and* $\lambda_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\log(d_1)/N}$. *Then for any constant* $r \geq 1$, *as* $N, d \to \infty$, *we have with* r_n *as in Theorem 1.9(c),*

$$\|\widehat{\boldsymbol{\beta}}_{a,NR} - \boldsymbol{\beta}^*_{a,NR}\|_2 + \{E[\widehat{\mu}_{a,NR}(\mathbf{S}_1) - \mu^*_{a,NR}(\mathbf{S}_1)]^r\}^{1/r} = O_p\left(r_n\right). \tag{1.31}$$

Remark 1.5 (Comparison between $\widehat{\boldsymbol{\beta}}_a$ and $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$)**.** *The present remark compares the consistency rates of two estimators,* $\widehat{\boldsymbol{\beta}}_a$ *and* $\widehat{\boldsymbol{\beta}}_{a,NR}$, *in different scenarios. (a) In the case where* $\nu_a(\cdot)$ *is nonlinear and* $\rho_a(\cdot)$ *is logistic, estimators converge to distinct targets,* $\boldsymbol{\beta}^*_a$ *and* $\boldsymbol{\beta}^*_{a,NR}$, *respectively. Here,* $\boldsymbol{\beta}^*_a$ *represents the optimal linear slope approximating the true conditional mean function* $\mu_a(\cdot)$, *while* $\boldsymbol{\beta}^*_{a,NR}$ *is the optimal linear slope approximating the misspecified model* $\nu^*_a(\cdot)$; *see discussions in Section 1.2.3 above. When the first-time conditional mean* $\mu_a(\cdot)$ *is linear,* $\widehat{\boldsymbol{\beta}}_a$ *converges to the true linear slope, and a consistent estimate of* $\mu_a(\cdot)$ *is obtained. However,* $\widehat{\boldsymbol{\beta}}_{a,NR}$ *typically converges to some* $\boldsymbol{\beta}^*_{a,NR}$ *that differs from the true linear slope, resulting in an inconsistent estimate of* $\widehat{\mu}_{a,NR}(\cdot)$. *(b) When* $\nu_a(\cdot)$ *is linear and* $\rho_a(\cdot)$ *is*

logistic, $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}_{a,NR}^*$*. However, in this case,* $\widehat{\boldsymbol{\beta}}_a$ *exhibits a faster consistency rate than* $\widehat{\boldsymbol{\beta}}_{a,NR}$*. This can be attributed to the fact that the DR imputation error contributes to the consistency rate of* $\widehat{\boldsymbol{\beta}}_a$ *in a product form Theorem 1.9(a), while the imputation error from nested regression contributes in an additive form Theorem 1.9(c), which then dominates if* $s_{\boldsymbol{\beta}_a} = o(s_{\boldsymbol{\alpha}_a})$*. Consequently, the S-DRL estimator constructed based on* $\widehat{\boldsymbol{\beta}}_a$ *has a faster convergence rate than the DTL estimator, which is constructed based on* $\widehat{\boldsymbol{\beta}}_{a,NR}$*; see Table 1.2. The enhanced convergence rate exhibited by the S-DRL estimator implies a reduction in the requisite level of sparsity conditions necessary for the inferential guarantees. (c) In the scenario where* $\nu_a(\cdot)$ *is linear and* $\rho_a(\cdot)$ *is non-logistic, the targets* $\boldsymbol{\beta}_a^*$ *and* $\boldsymbol{\beta}_{a,NR}^*$ *are identical and* $\widehat{\boldsymbol{\beta}}_a$ *and* $\widehat{\boldsymbol{\beta}}_{a,NR}$ *have the same consistency rates, as seen in Theorem 1.9(c) and* (1.31)*.*

In general, $\mu_a(\cdot) - \mu_{a'}(\cdot)$ can be seen as a conditional average treatment effect (CATE) parameter through the well established nested representation (1.3). Outside of dynamic settings, DR approaches for CATE estimation typically rely on DR influence function representation of the conditional means. When those conditional means independently are not smooth enough, [Ken20] proposes to instead use DR imputations for the joint estimation of the difference of the conditional means. Here, the nested structure of $\mu_a(\cdot)$, where the true outcome is never observed, prevents direct influence function approaches. Instead, our approach leverages cases when $\mu_a(\cdot)$ has *sparser* structure than $\nu_a(\cdot)$.

1.5 Advancing multi-stage treatment estimation with DR methods

The objective of this section is to expand upon the methodology of sequential doubly robust estimation by considering its application in multi-stage settings. Consider $T \geq 2$ exposure times and suppose that we observe i.i.d. samples $\{W_{T,i}\}_{i=1}^{N} = (\mathbf{S}_{1i}, A_{1i}, \ldots, \mathbf{S}_{Ti}, A_{Ti}, Y)_{i=1}^{N}$. Let $W_T := (\mathbf{S}_1, A_1, \ldots, \mathbf{S}_T, A_T, Y)$ be an independent copy of $W_{T,i}$. For each $t \leq T$, let $\mathbf{S}_t \in \mathbb{R}^{d_t}$ and $A_t \in \{0,1\}$ denote the covariate vector and the treatment assignment at the t-th exposure time, respectively. Let $Y \in \mathbb{R}$ denote the observed outcome variable at the final stage. Denote $\overline{\mathbf{S}}_t := (\mathbf{S}_1, \ldots, \mathbf{S}_t)$ and $\bar{A}_t := (A_1, \ldots, A_t)$ for any $1 \leq t \leq T$. Let $Y(\overline{a}_T)$ be the counterfactual outcome corresponding to the treatment path $\overline{a}_T = (a_1, \ldots, a_T) \in \{0,1\}^T$. The DTE between any treatment paths $\overline{a}_T, \overline{a}'_T \in \{0,1\}^T$ is now defined as

$$\theta := E[Y(\overline{a}_T)] - E[Y(\overline{a}'_T)] = \theta_{\overline{a}_T} - \theta_{\overline{a}'_T}.$$

We define the conditional mean and propensity score functions as

$$\mu_t(\bar{\mathbf{s}}_t, \overline{a}_T) := E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_t = \bar{\mathbf{s}}_t, \bar{A}_{t-1} = \overline{a}_{t-1}], \quad \text{for} \quad 1 \leq t \leq T+1, \tag{1.32}$$

$$\pi_t(\bar{\mathbf{s}}_t, \overline{a}_t) := P[A_t = a_t \mid \overline{\mathbf{S}}_t = \bar{\mathbf{s}}_t, \bar{A}_{t-1} = \overline{a}_{t-1}], \quad \text{for} \quad 1 \leq t \leq T, \tag{1.33}$$

where for the sake of simplicity, we denote with $\bar{A}_0 = \overline{a}_0 = \emptyset$ and $\overline{\mathbf{S}}_{T+1} := (\mathbf{S}_1, \ldots, \mathbf{S}_T, Y)$. For each $1 \leq t \leq T$, we denote $\mu_t^*(\bar{\mathbf{s}}_t, \overline{a}_T)$ and $\pi_t^*(\bar{\mathbf{s}}_t, \overline{a}_t)$ as the working models for the conditional mean and propensity score, respectively. Additionally, with $\overline{\mathbf{S}}_0 = \bar{\mathbf{s}}_0 = \emptyset$, we set $\mu_0(\bar{\mathbf{s}}_0, \overline{a}_T) := E[Y(\overline{a}_T) | \overline{\mathbf{S}}_0 = \bar{\mathbf{s}}_0] = \theta_{\overline{a}_T}$ and $\mu_{T+1}^*(\bar{\mathbf{s}}_{T+1}, \overline{a}_T) := s_{T+1}$. Note that, under the Assumption 1.6(b) below, we have $\mu_{T+1}^*(\overline{\mathbf{S}}_{T+1}, \overline{a}_T) = \mu_{T+1}(\overline{\mathbf{S}}_{T+1}, \overline{a}_T) = Y$. To identify

$\theta_{\bar{a}_T} = E[Y(\bar{a}_T)]$ for any $\bar{a}_T \in \{0,1\}^T$, we assume a multi-stage version of Assumption 1.1 in the following; see also, e.g., [Mur03, Rob00a, Rob87].

Assumption 1.6. *(a) (Sequential Ignorability)* $Y(\bar{a}_T) \perp\!\!\!\perp A_t \mid \overline{\mathbf{S}}_t, \bar{A}_{t-1} = \bar{a}_{t-1}$ *for each* $1 \leq t \leq T$. *(b) (Consistency of potential outcomes)* $Y = Y(\bar{A}_T)$. *(c) (Overlap) Let* $c_0 \in (0, 1/2)$ *be a positive constant, such that* $P(c_0 \leq \pi_t(\bar{\mathbf{s}}_t, \bar{a}_t) \leq 1 - c_0) = 1$ *for each* $1 \leq t \leq T$.

The following proposition presents a well-known DR representation of $E[Y(\bar{a}_t)]$ under the multi-stage dynamic setting; see, e.g., [BR05, MvdLRG01].

Proposition 1.1. *Let Assumption 1.6 hold. For* $t \leq T$ *suppose that at least one of* $\mu_t^*(\cdot, \bar{a}_T)$ *and* $\pi_t^*(\cdot, \bar{a}_t)$ *is correctly specified, i.e., either* $\mu_t^*(\cdot, \bar{a}_T) = \mu_t(\cdot, \bar{a}_T)$ *or* $\pi_t^*(\cdot, \bar{a}_t) = \pi_t(\cdot, \bar{a}_t)$. *Then*

$$\theta_{\bar{a}_T} = E\left[\sum_{t=1}^{T} \frac{\mathbb{1}_{\{\bar{A}_t = \bar{a}_t\}}}{\prod_{l=1}^{t} \pi_l^*(\overline{\mathbf{S}}_l, \bar{a}_l)} (\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1}, \bar{a}_T) - \mu_t^*(\overline{\mathbf{S}}_t, \bar{a}_T)) + \mu_1^*(\mathbf{S}_1, \bar{a}_T)\right]. \tag{1.34}$$

According to Proposition 1.1, a consistent estimate of $\theta_{\bar{a}_T}$ should be achievable as long as we can consistently estimate at least one of the nuisance functions $\mu_t(\cdot, \bar{a}_T)$ or $\pi_t(\cdot, \bar{a}_t)$ for each exposure time t. In the present context, the propensity score functions of (1.33) are identifiable via observable variables. Additionally, by Assumption 1.6, $\mu_T(\bar{\mathbf{s}}_T, \bar{a}_T) = E[Y|\overline{\mathbf{S}}_T = \bar{\mathbf{s}}_T, \bar{A}_T = \bar{a}_T]$, thereby facilitating its estimation using the corresponding samples. However, the remaining conditional mean functions for stages $t \leq T-1$ cannot be identified directly. To address this challenge, we propose DR representations of these intermediate conditional means, as an alternative to the conventional nested representation of (1.35).

Theorem 1.11. *Let Assumption 1.6 hold. For* $t \leq T-1$ *and* $t+1 \leq r \leq T$, *suppose that either* $\pi_r^*(\cdot, \bar{a}_r)$ *or* $\mu_r^*(\cdot, \bar{a}_T)$ *is correctly specified, i.e., either* $\pi_r^*(\cdot, \bar{a}_r) = \pi_r(\cdot, \bar{a}_r)$ *or* $\mu_r^*(\cdot, \bar{a}_T) =$

$\mu_r(\cdot, \overline{a}_T)$. *Then* $\mu_t(\bar{\mathbf{s}}_t, \overline{a}_T) = E[\psi^*(W_T, \overline{a}_T) \mid \overline{\mathbf{S}}_t = \bar{\mathbf{s}}_t, \bar{A}_t = \overline{a}_t]$, *where*

$$\psi^*(W_T, \overline{a}_T) := \sum_{r=t+1}^{T} \frac{\prod_{l=t+1}^{r} \mathbb{1}_{\{A_l = a_l\}}}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)} (\mu_{r+1}^*(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) - \mu_r^*(\overline{\mathbf{S}}_r, \overline{a}_T)) + \mu_{t+1}^*(\overline{\mathbf{S}}_{t+1}, \overline{a}_T).$$

Theorem 1.11 can be regarded as an overarching, comprehensive, umbrella result that subsumes a range of components, particularly encompassing Proposition 1.1 as a specific case when $t = 0$. Indeed, $\theta_{\overline{a}_T} = E[Y(\overline{a}_T)|\overline{\mathbf{S}}_0 = \bar{\mathbf{s}}_0] = \mu_0(\bar{\mathbf{s}}_0, \overline{a}_T)$ is a conditional mean function at "stage zero". Theorem 1.11 indicates that $\mu_t(\cdot, \overline{a}_T)$ can be identified through a DR representation using all the conditional means and propensity scores at later stages. Therefore, $\mu_t(\cdot, \overline{a}_T)$ can be estimated sequentially backward in time based on the DR imputations.

For example, if we use linear working models for the conditional means and logistic models for the propensity scores, and either the true conditional mean $\mu_r(\cdot, \overline{a}_T)$ is linear or the true propensity score $\pi_r(\cdot, \overline{a}_r)$ is logistic at each later stage $r \geq t + 1$, we can get a consistent estimate of the $\mu_t(\cdot, \overline{a}_T)$ using DR imputed linear regression. By repeating this process backwards, we conclude that if either the conditional means or propensity scores are correctly parametrized at every stage t, we can estimate all nuisance functions consistently, leading to a consistent estimate of $\theta_{\overline{a}_T}$. An alternative approach to our proposed sequential doubly robust method is the nested estimator [MvdLRG01]. This approach represents all conditional means using the following equation:

$$\mu_t(\bar{\mathbf{s}}_t, \overline{a}_T) = E[\mu_{t+1}(\overline{\mathbf{S}}_{t+1}, \overline{a}_T) \mid \overline{\mathbf{S}}_t = \bar{\mathbf{s}}_t, \bar{A}_t = \overline{a}_t]. \tag{1.35}$$

However, in order to ensure the consistency of the nested estimator for $\mu_t(\cdot, \overline{a}_T)$, it is essential that all subsequent conditional mean functions exhibit true linearity. Interestingly, even the multiply robust approach presented by [BRR19] falls short in achieving the same level of

robustness as the S-DRL method. Further insights can be found in the comments following Theorem 1.1 and Table 1.1. Our method demonstrates a growing advantage as the number of exposure times increases. For instance, in the case of T exposure times, consider all the cases including correctly or incorrectly parametrized $\pi_t^*(\cdot, \overline{a}_t)$ and $\mu_t^*(\cdot, \overline{a}_T)$ for each t, our method enables 3^T out of 4^T possible cases. In contrast, the nested-regression-based and multiply robust approaches only allow for $(T+2)2^{T-1}$ cases. This conclusion is independent of the particular parametrization used – nonparametric, smooth models are permissible – and extends to the difference of means $\theta = \theta_{\overline{a}_T} - \theta_{\overline{a}'_T}$ as well.

1.6 Numerical Experiments

1.6.1 Simulation studies

We illustrate the finite sample properties of the introduced estimators in several simulated experiments; auxiliary settings are relegated to the Section 1.8.2. We consider $a = (1, 1)$ and $a' = (0, 0)$, and use $\mathbf{1}_{(q)} := (1, \ldots, 1)^\top \in \mathbb{R}^q$ as well as $\mathbf{0}_{(q)} := (0, \ldots, 0) \in \mathbb{R}^q$. Below we use $\zeta_i \sim^{\text{iid}} \text{Uniform}(-1, 1)$ and $\{\boldsymbol{\delta}_i\}_j \sim^{\text{iid}} \text{Uniform}(-1, 1)$. We decompose $\boldsymbol{\alpha}_a$ into two components as $\boldsymbol{\alpha}_a = (\boldsymbol{\alpha}_{a,1}^\top, \boldsymbol{\alpha}_{a,2}^\top)^\top$ and consider $\boldsymbol{\alpha}_{a'} = -\boldsymbol{\alpha}_a$ and $\boldsymbol{\eta}_{a'} = -\boldsymbol{\eta}_a$ unless specified differently. For each $c \in \{a, a'\}$, $\rho_c(\mathbf{S}_i) = g(\mathbf{U}_i^\top \boldsymbol{\eta}_c)$ and $A_{2i}|(\mathbf{S}_i, A_{1i} = c_1) \sim \text{Bernoulli}(\rho_c(\mathbf{S}_i))$.

M1: Correctly parametrized models Consider $\mathbf{S}_{1i} \sim^{\text{iid}} N_{d_1}(\mathbf{0}, \mathbf{I}_{d_1})$ and $A_{1i}|\mathbf{S}_{1i} \sim \text{Bernoulli}\ (\pi_a(\mathbf{S}_{1i}))$, with $\pi_a(\mathbf{S}_{1i}) = g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a)$. Let $\delta_{1i} \sim^{\text{iid}} N(0, 1)$, $\boldsymbol{\delta}_{1i} \sim^{\text{iid}} N_{d_1}(0, \mathbf{I}_{d_1})$,

and

$$\mathbf{S}_{2i} = \mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{(d_1)} + \boldsymbol{\delta}_{1i}.$$

The outcomes are $Y_i(c) = \mathbf{U}_i^\top \boldsymbol{\alpha}_c + N(0,1)$ with parameters $\boldsymbol{\alpha}_a = (-1,-1,1,-1,\mathbf{0}_{(d_1-3)}, -1,-1,1,\mathbf{0}_{(d_2-3)})^\top$, $\boldsymbol{\alpha}_{a'} = (1,1,1,-1,\mathbf{0}_{(d_1-3)},1,1,1,\mathbf{0}_{(d_2-3)})^\top$, $\boldsymbol{\gamma}_a = (0,1,1,1,\mathbf{0}_{(d_1-3)})^\top$, $\boldsymbol{\eta}_a = (0,1,1,\mathbf{0}_{(d_1-2)},1,-1,\mathbf{0}_{(d_2-2)})^\top$, and $\boldsymbol{\eta}_{a'} = (0,0.5,0,-0.5,\mathbf{0}_{(d_1-3)},0.5,0,0.5,\mathbf{0}_{(d_2-3)})^\top$.

M2: Weakly sparse $\nu_c(\cdot)$ and dense $\pi_c(\cdot)$ Let $D_i \sim^{\text{iid}} \text{Bernoulli}(0.5)$ and

$$\{\mathbf{S}_{1i}\}_j \sim D_i \cdot \text{Uniform}(-1,-0.5) + (1-D_i)\cdot \text{Uniform}(0.5,1).$$

Define $\{\overline{W}\}_j = 0.5\cdot 0.9^j$ and let $\mathbf{S}_{2i} = \mathbf{S}_{1i}^\top \overline{W}\mathbf{1}_{(d_2)} + \boldsymbol{\delta}_i$. Let $Y_i(c) = \mathbf{U}_i^\top \boldsymbol{\alpha}_c + \zeta_i$. The parameter $\boldsymbol{\alpha}_{a,1} = (-1,\mathbf{0}_{(d_1)})^\top$ and $\{\boldsymbol{\alpha}_{a,2}\}_j = -0.3\cdot 0.99^{j-1}$, $\boldsymbol{\eta}_a = (3,0.1,\mathbf{0}_{(d_1+d_2-1)})^\top$.

M3: Non-linear $\nu_c(\cdot)$ and non-logistic $\pi_c(\cdot)$ Consider $\{\mathbf{S}_{1i}\}_j \sim^{\text{iid}} \text{Uniform}(-1,1)$. Let $\pi_a(\mathbf{S}_{1i}) = \bar{g}(\mathbf{V}_i^\top \boldsymbol{\gamma}_a)$, where

$$\bar{g}(u) = (|u|/(|u|+1))\mathbb{1}_{\{u>0\}} + (1/(|u|+1))\mathbb{1}_{\{u<0\}}.$$

Define $\{\widetilde{W}(a)\}_j = 0.7\cdot 0.8^j$ and $\{\widetilde{W}(a')\}_j = 0.5\cdot 0.9^j$. Let $\mathbf{W}_{2i} = \{\widetilde{W}(A_{1i})\}^\top \mathbf{S}_{1i}\mathbf{1}_{(d_2)} + \boldsymbol{\delta}_i$ and $\{\mathbf{S}_{2i}\}_j = \sqrt{|\{\mathbf{W}_{2i}\}_j|}$. The parameter $\boldsymbol{\alpha}_a$ has the same $\boldsymbol{\alpha}_{a,1}$ as M2 and $\{\boldsymbol{\alpha}_{a,2}\}_j = -0.3\cdot 0.9^{j-1}$. The parameter $\boldsymbol{\gamma}_a = 5\cdot \mathbf{1}_{(10)}^\top$, and $\boldsymbol{\eta}_a = (2,0.1,\mathbf{0}_{(d_1-1)},0.1,\mathbf{0}_{(d_2-1)})^\top$. Let

$$Y_i(c) = \mathbf{V}_i^\top \boldsymbol{\alpha}_{c,1} + \sum_{j=1}^{d_2} \{\alpha_{c,2}\}_j \text{sgn}(\{\mathbf{W}_{2i}\}_j)\{\mathbf{S}_{2i}\}_j^2 + \zeta_i.$$

For M1, $d_1 = d_2 = 100$ and $N \in \{1000, 4000\}$; for M2, $d_2 = 500$, (N, d_1) are chosen from $(2000, 20)$, $(4000, 20)$, and $(4000, 50)$; for M3, $d_1 = 20$, (N, d_2) are chosen from $(500, 500)$, $(1000, 500)$, $(2000, 500)$, $(4000, 500)$, $(1000, 1000)$, and $(2000, 1000)$. We replicate

Table 1.3: Setting 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. Err_a: average estimation error of $\mu_a(\cdot)$; $\text{Err}_{a'}$: average estimation error of $\mu_{a'}(\cdot)$. N_1 and N_0 are the expected number of observations in groups $(1,1)$ and $(0,0)$.

Method	Bias	RMSE	Length	Coverage	ESD	ASD	Err_a	$\text{Err}_{a'}$
	$N = 1000, N_1 = 293, N_0 = 282, d_1 = 100, d_2 = 100$							
empdiff	0.734	0.734	0.274	0.004	0.234	0.070	NA	NA
oracle	0.003	0.220	1.091	0.954	0.325	0.278	0.000	0.000
IPW	0.864	0.865	1.342	0.346	0.319	0.342	NA	NA
DTL	0.124	0.189	0.876	0.894	0.264	0.223	0.141	0.216
S-DRL	0.131	0.202	0.880	0.880	0.271	0.224	0.227	0.337
S-DRL'	0.126	0.188	0.876	0.894	0.259	0.223	0.135	0.193
	$N = 4000, N_1 = 1178, N_0 = 1128, d_1 = 100, d_2 = 100$							
empdiff	0.731	0.731	0.137	0.000	0.111	0.035	NA	NA
oracle	-0.006	0.121	0.602	0.956	0.178	0.153	0.000	0.000
IPW	0.534	0.538	0.959	0.454	0.287	0.245	NA	NA
DTL	0.033	0.097	0.488	0.930	0.136	0.125	0.032	0.052
S-DRL	0.031	0.098	0.489	0.930	0.142	0.125	0.050	0.070
S-DRL'	0.028	0.098	0.489	0.932	0.138	0.125	0.033	0.044

settings 500 times. We report S-DRL as well as a version S-DRL', which has $\widehat{\boldsymbol{\beta}}_c$ constructed with $\widehat{\boldsymbol{\delta}}_c$ and $\widehat{\boldsymbol{\alpha}}_c$ build on the whole sub-sample $\mathcal{W}_{-k} = \mathcal{W}_{-k,1} \cup \mathcal{W}_{-k,2}$. We also present (a) DTL, Algorithm 2, (b) IPW with ℓ_1-regularized logistic PS, (c) an empirical difference estimator (empdiff), $\widehat{\theta}_{\text{empdiff}} := \sum_{i=1}^N A_{1i}A_{2i}Y_i / \sum_{i=1}^N A_{1i}A_{2i} - \sum_{i=1}^N (1-A_{1i})(1-A_{2i})Y_i / \sum_{i=1}^N (1-A_{1i})(1-A_{2i})$, and (d) an oracle DR estimator constructed with the true nuisances. All methods use 10-fold cross validation for selection of tuning parameters.

Tables 1.3 and 1.4 show the estimation and inference results for the DTE estimators, while Table 1.5 focuses on estimation performances, as valid inference is unlikely with misspecified models. Our summarized findings, shown in Tables 1.3-1.5, reveal that the

naive empirical difference estimator has large biases due to confounding between outcome and treatment assignments. The IPW method also performs poorly, with large biases and RMSEs, and confidence interval coverages far from the desired 95%. The DTL, S-DRL, and S-DRL' estimators behave similarly in Table 1.3 (under M1), with correctly specified nuisance models and relatively low sparsity levels. The S-DRL method's additional sample splitting in Algorithm 1 (Steps 5-7) leads to larger estimation errors in the first-time conditional mean estimates than those in the DTL and S-DRL' methods. Consequently, when $N = 1000$, the S-DRL estimator's RMSE is slightly larger than that of the DTL and S-DRL' estimators, but they have similar RMSEs when $N = 4000$. In terms of inference behaviors, the corresponding confidence interval coverages are below the desired 95% when $N = 1000$. However, increasing the total sample size to $N = 4000$ brings the coverages closer to 95%. Estimating $\nu_c(\cdot)$ under M2 is more challenging than estimating $\rho_c(\cdot)$. As a result, the DR estimates of $\mu_c(\cdot)$ in the S-DRL and S-DRL' methods have significantly smaller estimation errors compared to the nested regression used in the DTL method, as shown in Table 1.4, leading to smaller RMSEs and coverages closer to 95%. Moving on to M3, both $\nu_c(\cdot)$ and $\pi_c(\cdot)$ are misspecified. Table 1.5 shows that the estimation errors of $\mu_c(\cdot)$ with the S-DRL and S-DRL' are substantially smaller than those of the DTL. Consequently, we see an improvement of the RMSEs in the S-DRL and S-DRL' estimators.

1.6.2 Application to National Job Corps Study (NJCS)

Job Corps (JC) is the largest and most comprehensive federal job training program in the US for disadvantaged youth between 16 and 24 years old. Each year, about 50,000 participants receive vocational training and academic education at JC centers to improve

Table 1.4: Setting M2. The rest of the caption details remain the same as those in Table 1.3.

Method	Bias	RMSE	Length	Coverage	ESD	ASD	Err_a	$\text{Err}_{a'}$
	$N = 2000, N_1 = 954, N_0 = 951, d_1 = 20, d_2 = 500$							
oracle	-0.021	0.691	4.146	0.950	1.029	1.058	0.000	0.000
empdiff	-0.204	0.601	1.576	0.636	0.848	0.402	NA	NA
IPW	-0.153	2.542	12.747	0.970	3.718	3.252	NA	NA
DTL	-0.013	0.714	4.151	0.950	1.060	1.059	0.361	0.365
S-DRL	-0.023	0.686	4.144	0.952	1.008	1.057	0.139	0.136
S-DRL'	-0.027	0.689	4.145	0.952	1.019	1.057	0.126	0.122
	$N = 4000, N_1 = 1909, N_0 = 1901, d_1 = 20, d_2 = 500$							
oracle	-0.029	0.572	2.942	0.936	0.835	0.751	0.000	0.000
empdiff	-0.099	0.408	1.116	0.630	0.598	0.285	NA	NA
IPW	-0.101	2.510	11.928	0.978	3.643	3.043	NA	NA
DTL	-0.053	0.569	2.950	0.934	0.843	0.753	0.163	0.161
S-DRL	-0.026	0.554	2.942	0.940	0.838	0.750	0.040	0.040
S-DRL'	-0.029	0.560	2.942	0.940	0.843	0.751	0.040	0.042
	$N = 4000, N_1 = 1908, N_0 = 1901, d_1 = 50, d_2 = 500$							
oracle	-0.030	0.608	2.990	0.948	0.893	0.763	0.000	0.000
empdiff	-0.156	0.485	1.260	0.608	0.683	0.322	NA	NA
IPW	-0.086	2.099	10.088	0.984	3.136	2.574	NA	NA
DTL	-0.019	0.607	2.983	0.940	0.903	0.761	0.273	0.275
S-DRL	-0.013	0.565	2.988	0.948	0.854	0.762	0.082	0.083
S-DRL'	-0.012	0.574	2.987	0.948	0.863	0.762	0.080	0.081

Table 1.5: Setting M3. The rest of the caption details remain the same as those in Table 1.3.

Method	Bias	RMSE	Err_a	$\text{Err}_{a'}$	Bias	RMSE	Err_a	$\text{Err}_{a'}$
	$N = 500, d_1 = 20, d_2 = 500$				$N = 1000, d_1 = 20, d_2 = 500$			
oracle	0.013	0.119	0.000	0.000	0.007	0.098	0.000	0.000
empdiff	0.149	0.162	NA	NA	0.157	0.159	NA	NA
IPW	-0.035	0.465	NA	NA	-0.164	0.493	NA	NA
DTL	-0.003	0.273	0.151	0.196	0.012	0.199	0.076	0.077
S-DRL	0.009	0.224	0.087	0.096	0.024	0.166	0.036	0.040
S-DRL'	-0.007	0.207	0.067	0.074	0.016	0.170	0.031	0.033
	$N = 2000, d_1 = 20, d_2 = 500$				$N = 4000, d_1 = 20, d_2 = 500$			
oracle	0.005	0.062	0.000	0.000	0.003	0.047	0.000	0.000
empdiff	0.142	0.142	NA	NA	0.142	0.142	NA	NA
IPW	-0.276	0.525	NA	NA	-0.390	0.487	NA	NA
DTL	0.010	0.126	0.038	0.040	-0.001	0.101	0.021	0.021
S-DRL	0.013	0.108	0.016	0.016	0.000	0.089	0.007	0.007
S-DRL'	0.004	0.106	0.015	0.015	-0.007	0.090	0.007	0.007
	$N = 1000, d_1 = 20, d_2 = 1000$				$N = 2000, d_1 = 20, d_2 = 1000$			
oracle	0.001	0.096	0.000	0.000	0.011	0.059	0.000	0.000
empdiff	0.149	0.149	NA	NA	0.143	0.143	NA	NA
IPW	-0.221	0.469	NA	NA	-0.249	0.478	NA	NA
DTL	-0.028	0.212	0.090	0.091	0.010	0.135	0.048	0.049
S-DRL	-0.012	0.160	0.037	0.039	0.011	0.113	0.016	0.016
S-DRL'	-0.029	0.165	0.031	0.032	0.002	0.116	0.015	0.015

their job prospects. On average, a JC student spends 8 months at a local center, completing around 1,100 hours of instruction, which is roughly equivalent to one year of high school. For a more detailed description, refer to [SBM08] and [Sch01].

Numerous studies have investigated the effects of Job Corps on wages. [Lee09] highlighted sample selection issues in their analysis. [ZRM08] separated the causal effects of JC enrollment on wages from those on employment. [FFLGN12] found that longer exposure to JC training is associated with higher future earnings. [CF15] separated the effects of sample selection from noncompliance, while [HHLL20] distinguished between the causal direct and indirect effects in the presence of mediators. In addition to studying the effects of Job Corps in single-time treatment settings, researchers have also explored the dynamic treatment setting offered by Job Corps. [BHL22] investigated the effects of JC's educational and training programs and found positive impacts on fourth-year employment compared to no program participation. Meanwhile, [SXG21] analyzed the total, direct, and indirect dynamic dose response of job training on employment. Their study concluded that a few class hours in the first and second years significantly increase employment in the fourth year. In this section, we will evaluate the effects of sequential job training programs on wages using the S-DRL and DTL methods, as defined in Algorithms 1 and 2.

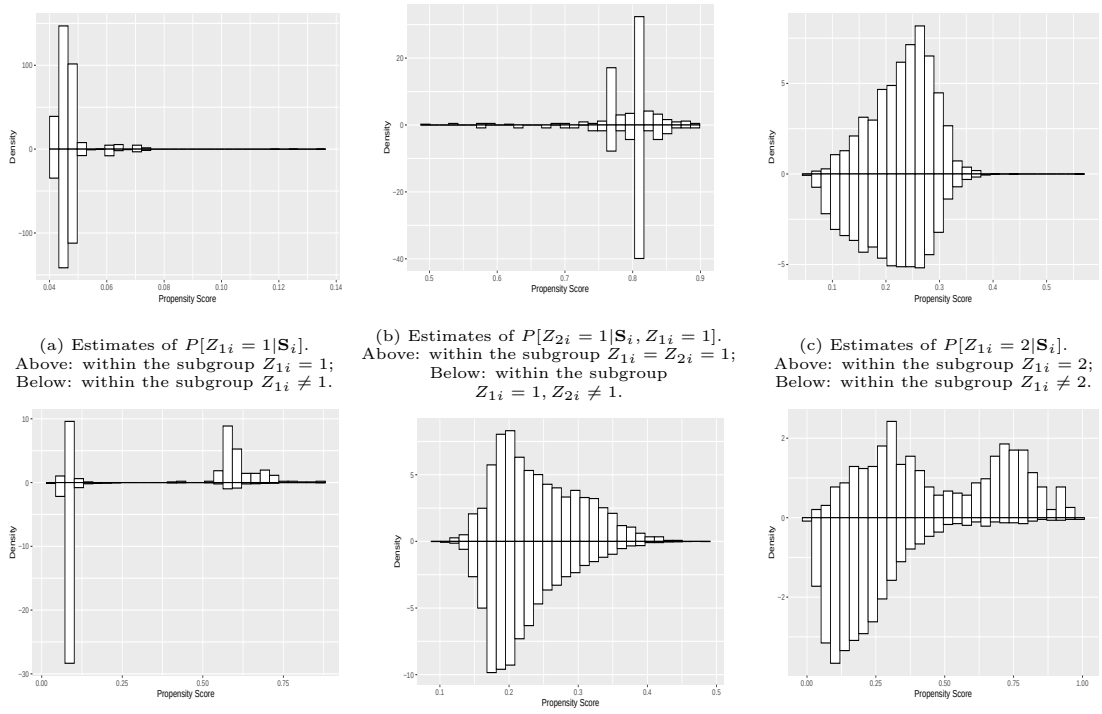
We analyze a dataset of 11,313 individuals, with 6,828 assigned to the Job Corps and 4,485 not. They are interviewed 1, 2, and 4 years post-randomization. For each year $t \in 1, 2$, $Z_t \in \{0, 1, 2, 3\}$ represents the treatment assignment in the t-th year. We assign $Z_t = 0$ for non-enrollment, $Z_t = 1$ for enrollment without program participation, $Z_t = 2$ for high-school-level education, and $Z_t = 3$ for vocational training. The baseline covariate

Table 1.6: Job Corps estimates. Here, N_a and $N_{a'}$ denote the number of observations in the treatment groups a and a', respectively. SE: the standard error; CI: 95% confidence interval; p-value: $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$.

z	z'	N_z	$N_{z'}$	Method	$\widehat{\theta}_z$	$\widehat{\theta}_{z'}$	$\widehat{\theta}$	SE	CI	p-value
$(3,3)$	$(1,1)$	568	315	DTL	6.201	5.652	0.549	0.270	[0.020, 1.078]	0.042
				S-DRL	6.208	5.641	0.567	0.273	[0.032, 1.102]	0.037
$(3,3)$	$(2,2)$	568	336	DTL	6.200	5.424	0.776	0.313	[0.163, 1.389]	0.013
				S-DRL	6.209	5.390	0.819	0.314	[0.204, 1.434]	0.009
$(2,2)$	$(1,1)$	336	315	DTL	5.410	5.639	-0.229	0.335	[-0.886, 0.428]	0.493
				S-DRL	5.371	5.626	-0.255	0.337	[-0.916, 0.406]	0.450

vector, $\mathbf{S}_1$, has 909 characteristics, while $\mathbf{S}_2$ includes 1,427 characteristics. In total, there are 2,336 covariates. The outcome is the log-transformed wage $\widetilde{Y} = \log(\text{wage} + 1) \in \mathbb{R}$. We exclude 2,610 individuals with missing treatment stages that are missing completely at random [SBRJ$^+$03] and an additional 133 with missing covariates or outcomes, resulting in a final sample of 8,570 individuals.

Table 1.6 shows estimated DTEs between treatment paths $(3,3)$ vs. $(1,1)$, $(3,3)$ vs. $(2,2)$, and $(2,2)$ vs. $(1,1)$. Both the S-DRL and DTL methods suggest that vocational training has a positive impact on achieving higher wages by showing non-zero effects between the first two paths. On the other hand, the estimates between paths $(2,2)$ and $(1,1)$ are negative, and their corresponding confidence intervals contain zero, making it impossible to determine if academic education is beneficial or detrimental. However, our analysis does suggest that individuals seeking higher-paying jobs would benefit more from vocational training compared to academic education, which only provides high school-level education without any significant vocational training. The S-DRL estimates have a slightly greater distance from zero compared to DTL's, with similar standard errors, leading to slightly smaller p-values. Figure 1.1 examines the overlap of estimated propensity scores, displaying mirror



(a) Estimates of $P[Z_{1i} = 1|\mathbf{S}_i]$. Above: within the subgroup $Z_{1i} = 1$; Below: within the subgroup $Z_{1i} \neq 1$.

(b) Estimates of $P[Z_{2i} = 1|\mathbf{S}_i, Z_{1i} = 1]$. Above: within the subgroup $Z_{1i} = Z_{2i} = 1$; Below: within the subgroup $Z_{1i} = 1, Z_{2i} \neq 1$.

(c) Estimates of $P[Z_{1i} = 2|\mathbf{S}_i]$. Above: within the subgroup $Z_{1i} = 2$; Below: within the subgroup $Z_{1i} \neq 2$.

(d) Estimates of $P[Z_{2i} = 2|\mathbf{S}_i, Z_{1i} = 2]$. Above: within the subgroup $Z_{1i} = Z_{2i} = 2$; Below: within the subgroup $Z_{1i} = 2, Z_{2i} \neq 2$.

(e) Estimates of $P[Z_{1i} = 3|\mathbf{S}_i]$. Above: within the subgroup $Z_{1i} = 3$; Below: within the subgroup $Z_{1i} \neq 3$.

(f) Estimates of $P[Z_{2i} = 3|\mathbf{S}_i, Z_{1i} = 3]$. Above: within the subgroup $Z_{1i} = Z_{2i} = 3$; Below: within the subgroup $Z_{1i} = 3, Z_{2i} \neq 3$.

Figure 1.1: Mirror histograms of propensity score overlaps.

histograms of estimated propensity scores within the treatment and control groups. The substantial overlap seen in the mirror histograms indicates that the inverse propensity score weights are relatively stable. Figure 1.1(d) displays bimodal patterns in the histograms due to a binary confounder variable that significantly influences the propensity score estimate of $P[Z_2 = 2 \mid \mathbf{S}, Z_1 = 2]$, with a close association between a participant's decision to enroll in second-year education and their attendance in the class during the final weeks of the first year.

1.7 Discussion

This paper aims to enhance the understanding of estimating causal parameters in multi-stage settings. While prior DR literature has recognized the importance of the stage-

zero DR representation for the expected potential outcome, it has overlooked the fact that all the intermediate conditional mean functions can also be identified in a DR manner. This approach leads to better theoretical guarantees and greater flexibility in modeling dynamic dependencies, which can be complex and involve multiple time exposures. Furthermore, our findings have significant practical implications beyond parametric models, especially in situations where doctors or policymakers cannot rely on randomized treatments or simplistic treatment rules. With the ability to model dynamic treatment effects using robust principles, new avenues of discovery are emerging, including optimal treatment rules and determining the best treatment times. Our approach also enables the exploration of further important issues, such as mitigating network spillover effects through robustness perspectives and enriching balancing methods with better robustness properties. The significance of our work lies in the fact that it allows researchers to estimate treatment effects in complex settings more accurately and provides a valuable tool for policymakers seeking to make informed decisions based on robust causal inference methods.

1.8 Supplementary Material

To simplify the exposition, we begin by listing some shorthand notations used throughout the supplementary document. We let $\mathbf{U} = (1, \mathbf{S})^\top \in \mathbb{R}^{d+1}$ and $\mathbf{V} = (1, \mathbf{S}_1^\top)^\top \in \mathbb{R}^{d_1+1}$. In the following it is important to follow the individuals with pre-specified treatment plan. For that purpose we introduce the following shorthand notation: $\widetilde{Y}_a = Y\mathbb{1}_{\{\boldsymbol{A}=a\}}$, $\widetilde{\mathbf{U}}_a = \mathbf{U}\mathbb{1}_{\{\boldsymbol{A}=a\}}$ where $\boldsymbol{A} = (A_1, A_2) = a$. Additionally, we use $\bar{Y}_a = Y\mathbb{1}_{\{A_1=a_1\}}$, $\bar{\mathbf{U}}_a = \mathbf{U}\mathbb{1}_{\{A_1=a_1\}}$, $\bar{\mathbf{V}}_a = \mathbf{V}\mathbb{1}_{\{A_1=a_1\}}$ to denote individuals who have taken the treatment a_1 regardless of which treatment they received at the second exposure. Where possible, we suppress the sub-index

a.

1.8.1 Further discussions on the nuisance models

Model correctness of DTL

We illustrate when will the two working outcome models $\nu_a^*(\mathbf{s}) = \mathbf{u}^\top \boldsymbol{\alpha}_a^*$ and $\mu_{a,\text{NR}}^*(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}_{a,\text{NR}}^*$, the models used for DTL, be correctly specified. If the model $\nu_a^*(\cdot)$ is misspecified, then the model $\mu_{a,\text{NR}}^*(\cdot)$ 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. Their justifications are provided in Section 1.8.1 below.

(i) If we assume that the true outcome model, $\nu_a(\cdot)$ is linear in that

$$\nu_a(\mathbf{S}) = E[Y(a)|\mathbf{S}, A_1 = a_1, A_2 = a_2] = \mathbf{U}^\top \boldsymbol{\alpha}_a \tag{1.36}$$

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

$$\mu_a(\mathbf{S}_1) = E[Y(a)|\mathbf{S}_1, A_1 = a_1] = \mathbf{V}^\top \boldsymbol{\beta}_a \tag{1.37}$$

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., $\mu_{a,\text{NR}}^*(\cdot) \neq \mu_a(\cdot)$ making $\mu_{a,\text{NR}}^*(\cdot)$ potentially misspecified.

(iii) Now, if the true outcome model (1.37) holds and in addition $\boldsymbol{\alpha}_a^*$, (3.2), is equal to $\bar{\boldsymbol{\alpha}}_a^*$, with $\bar{\boldsymbol{\alpha}}_a^*$ defined as

$$\bar{\boldsymbol{\alpha}}_a^* := \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{d+1}} E\left[(Y(a) - \mathbf{U}^\top \boldsymbol{\alpha})^2 | A_1 = a_1\right] = \left[E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]\right]^{-1} E[\bar{\mathbf{U}} Y(a)],$$

then, we have $\boldsymbol{\beta}^*_{a,\mathrm{NR}} = \boldsymbol{\beta}_a$ and $\mu^*_{a,\mathrm{NR}}(\cdot) = \mu_a(\cdot)$, i.e., $\mu^*_{a,\mathrm{NR}}(\cdot)$ is correctly specified.

(iv) Lastly, if both of the true outcome models are linear, i.e., (1.36) and (1.37) hold simultaneously, then, both $\nu^*_a(\cdot)$ and $\mu^*_{a,\mathrm{NR}}(\cdot)$ are correctly specified. Case (iv) is equivalent to requiring $E(\mathbf{S}_2^\top \boldsymbol{\alpha}_{a,2} | \boldsymbol{S}_1)$ to be linear in $\boldsymbol{S}_1$; here, $\boldsymbol{\alpha}_a = (\boldsymbol{\alpha}_{a,1}, \boldsymbol{\alpha}_{a,2})^\top$ 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 $\mu^*_{a,\mathrm{NR}}(\cdot)$ 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 $\mu^*_{a,\mathrm{NR}}(\cdot)$; however, if the true outcome model $\nu_a(\cdot)$ is also linear, then $\mu^*_{a,\mathrm{NR}}(\cdot)$ 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 $\mu^*_{a,\mathrm{NR}}(\cdot)$ 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 $S_1, S_2, Z \sim^{\mathrm{iid}} \mathrm{Uniform}(-1,1)$ with a nonlinear outcome model $\nu_a(\cdot)$, $Y(a) = S_1 + S_2^3 + Z$. Let the treatment assignments satisfy

$$\pi_a(s_1) = |s_1|, \text{ and } \rho_a(s_1, s_2) = \exp(s_1 + s_2)/\{1 + \exp(s_1 + s_2)\},$$

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 $\mu^*_{a,\mathrm{NR}}(\cdot)$. Here, $\pi^*_a(\cdot)$ and $\nu^*_a(\cdot)$ are misspecified, $\rho^*_a(\cdot)$ and $\mu^*_{a,\mathrm{NR}}(\cdot)$ are correctly specified.

Model correctness of S-DRL

Note that we consider the same working models $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ for the DTL and S-DRL estimators; only the outcome model at time two differs among these two estimators. In the following, we illustrate when will the doubly robust working model $\mu_a^*(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}_a^*$ be misspecified and when will $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}_{a,\mathrm{NR}}^*$.

(i) Let either $\nu_a^*(\cdot) = \nu_a(\cdot)$ or $\rho_a^*(\cdot) = \rho_a(\cdot)$. Then, as long as $\mu_a(\cdot)$ is linear in that (1.37) holds for some vector $\boldsymbol{\beta}_a \in \mathbb{R}^{d_1+1}$, we have $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}$ and hence $\mu_a^*(\mathbf{s}_1) = \mathbf{v}^\top \boldsymbol{\beta}_a^* = \mu_a(\mathbf{s}_1)$, i.e., the doubly robust working model $\mu_a^*(\cdot)$ is correctly specified.

(ii) If $\nu_a^*(\cdot) = \nu_a(\cdot)$, then $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}_{a,\mathrm{NR}}^*$.

The justifications of cases (i) and (ii) are also provided in Section 1.8.1 below.

Justifications

Justifications of cases (i)-(iv) in Section 1.8.1 For (i), under Assumption 1 and by the tower rule, we have

$$\begin{aligned}
\boldsymbol{\alpha}_a^* &= \left[E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\right]^{-1} E\left[\widetilde{\mathbf{U}}\widetilde{Y}\right] = \left[E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\mathbf{U}Y(a)\right] \\
&= \left[E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbf{U}E\left[Y(a)|\mathbf{U}, A_1=a_1, A_2=a_2\right] P\left[A_1=a_1, A_2=a_2|\mathbf{U}\right]\right] \\
&= \left[E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbf{U}\mathbf{U}^\top\boldsymbol{\alpha}_a E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}|\mathbf{U}\right]\right] \\
&= \left[E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\right]^{-1} E\left[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\right]\boldsymbol{\alpha}_a = \boldsymbol{\alpha}_a.
\end{aligned}$$

It follows that

$$\nu_a(\mathbf{S}) = \mathbf{U}^\top\boldsymbol{\alpha}_a = \mathbf{U}^\top\boldsymbol{\alpha}_a^* = \nu_a^*(\mathbf{S}).$$

Therefore, if the model (1.36) holds, the working model $\nu_a^*(\mathbf{S})$ is correctly specified.

For (ii), it suffices to prove a counterexample. We refer to Example 1 in Section 2.3.

For (iii), if we assume that $\bar{\boldsymbol{\alpha}}_a^* = \boldsymbol{\alpha}_a^*$, we have

$$\begin{aligned}\boldsymbol{\beta}_{a,\mathrm{NR}}^* &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}\bar{\mathbf{U}}^\top\right]\boldsymbol{\alpha}_a^* = \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}\bar{\mathbf{U}}^\top\right]\bar{\boldsymbol{\alpha}}_a^* \\ &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}\bar{\mathbf{U}}^\top\right]\left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{U}}Y(a)\right].\end{aligned}$$

By the fact that $\mathbf{U} = (\mathbf{V}^\top, \mathbf{S}_2^\top)^\top$, we can write

$$\mathbf{V} = \mathbf{Q}\mathbf{U} \text{ where } \mathbf{Q} = \begin{pmatrix} \boldsymbol{I}_{d_1+1} & \mathbf{0}_{(d_1+1)\times d_2} \end{pmatrix}, \tag{1.38}$$

and hence $\bar{\mathbf{V}} = \mathbf{Q}\bar{\mathbf{U}}$, which implies that

$$\begin{aligned}&E\left[\bar{\mathbf{V}}\bar{\mathbf{U}}^\top\right]\left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{U}}Y(a)\right] = \mathbf{Q}E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{U}}Y(a)\right] \\ &= \mathbf{Q}E\left[\bar{\mathbf{U}}Y(a)\right] = E\left[\bar{\mathbf{V}}Y(a)\right].\end{aligned}$$

Therefore,

$$\boldsymbol{\beta}_{a,\mathrm{NR}}^* = \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}Y(a)\right] = \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{V}Y(a)\right].$$

By the tower rule,

$$\begin{aligned}\boldsymbol{\beta}_{a,\mathrm{NR}}^* &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\mathbf{V}E\left[Y(a)|\mathbf{V}, A_1 = a_1\right]E\left[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{V}\right]\right] \\ &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{V}\mathbf{V}^\top\boldsymbol{\beta}_a\right] = \boldsymbol{\beta}_a.\end{aligned}$$

It follows that

$$\mu_a(\mathbf{S}_1) = \mathbf{V}^\top\boldsymbol{\beta}_a = \mathbf{V}^\top\boldsymbol{\beta}_{a,\mathrm{NR}}^* = \mu_{a,\mathrm{NR}}^*(\mathbf{S}_1).$$

Therefore, if the model (1.37) holds and $\bar{\boldsymbol{\alpha}}_a^* = \boldsymbol{\alpha}_a^*$, the working model $\mu_{a,\mathrm{NR}}^*(\mathbf{S}_1)$ is correctly specified.

Regarding (iv), based on the results in (i), we have $\boldsymbol{\alpha}_a^* = \boldsymbol{\alpha}_a$. Under Assumption 1 and (1.36), we have

$$\nu_a(\mathbf{S}) = E\left[Y(a)|\mathbf{S}, A_1 = a_1\right] = \mathbf{U}^\top \boldsymbol{\alpha}_a.$$

Hence, we also have

$$\begin{aligned}
\bar{\boldsymbol{\alpha}}_a^* &= \left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{U}}Y(a)\right] = \left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{U}Y(a)\right] \\
&= \left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbf{U}E\left[Y(a)|\mathbf{U}, A_1 = a_1\right] P\left[A_1 = a_1|\mathbf{U}\right]\right] \\
&= \left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\mathbf{U}\mathbf{U}^\top \boldsymbol{\alpha}_a E\left[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{U}\right]\right] \\
&= \left[E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top\right] \boldsymbol{\alpha}_a = \boldsymbol{\alpha}_a.
\end{aligned}$$

Therefore,

$$\boldsymbol{\alpha}_a^* = \bar{\boldsymbol{\alpha}}_a^* = \boldsymbol{\alpha}_a.$$

Together with the results in (iii), we conclude that $\mu_{a,\mathrm{NR}}^*(\cdot)$ is correctly specified.

Justifications of cases (i)-(ii) in Section 1.8.1 For (i), by the definition of $\boldsymbol{\beta}_a^*$ and the KKT condition, we have

$$\begin{aligned}
\boldsymbol{\beta}_a^* &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}Y^{\mathrm{DR}}\right] \\
&= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1\}}\mathbf{V}\left[\nu_a^*(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})}\right]\right] \\
&\overset{(i)}{=} \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[E\left[\mathbb{1}_{\{A_1=a_1\}} \mid \mathbf{S}_1\right]\mathbf{V}E\left[\nu_a^*(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right]\right] \\
&\overset{(ii)}{=} \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[E[\mathbb{1}_{\{A_1=a_1\}} \mid \mathbf{S}_1]\mathbf{V}\mu_a(\mathbf{S}_1)\right] = \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}\mu_a(\mathbf{S}_1)\right] \\
&\overset{(iii)}{=} \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\boldsymbol{\beta}_a\right] = \boldsymbol{\beta}_a,
\end{aligned}$$

where (i) holds by the tower rule; (ii) holds by Theorem 1 in Section 1.1; (iii) holds since $\bar{\mathbf{V}}\mu_a(\mathbf{S}_1) = \bar{\mathbf{V}}\mathbf{V}^\top\boldsymbol{\beta}_a = \bar{\mathbf{V}}\bar{\mathbf{V}}^\top\boldsymbol{\beta}_a$.

For (ii), we observe that

$$\begin{aligned}
\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a,\mathrm{NR}}^* &= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\bar{\mathbf{V}}Y^{\mathrm{DR}} - \bar{\mathbf{V}}\bar{\mathbf{U}}^\top\boldsymbol{\alpha}_a^*\right] \\
&= \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\mathbf{V}\frac{Y(a)-\nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})}\right] \\
&\overset{(i)}{=} \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[E[A_1=a_1 \mid \mathbf{S}]\mathbf{V}E\left[\mathbb{1}_{\{A_2=a_2\}}\frac{Y(a)-\nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}, A_1=a_1\right]\right] \\
&\overset{(ii)}{=} \left[E\left[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top\right]\right]^{-1} E\left[E[A_1=a_1 \mid \mathbf{S}]\mathbf{V}E\left[\rho_a(\mathbf{S})\frac{\nu_a(\mathbf{S})-\nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}, A_1=a_1\right]\right] = 0,
\end{aligned}$$

as long as $\nu_a^*(\cdot) = \nu_a(\cdot)$. Here, (i) holds by the tower rule; (ii) holds since $Y(a) \perp\!\!\!\perp A_2 \mid \mathbf{S}, A_1 = a_1$ under Assumption 1, $\rho_a(\mathbf{S}) = E[\mathbb{1}_{\{A_2=a_2\}} \mid \mathbf{S}, A_1 = a_1]$, and $\nu_a(\mathbf{S}) = E[Y(a) \mid \mathbf{S}, A_1 = a_1]$. Hence, $\boldsymbol{\beta}_a^* = \boldsymbol{\beta}_{a,\mathrm{NR}}^*$ as long as $\nu_a^*(\cdot) = \nu_a(\cdot)$.

1.8.2 Additional numerical experiments

In this section, we present additional simulation results under different data generating processes (DGPs) where all the nuisance functions are correctly parametrized. For each $i \le N$, generate $\mathbf{S}_{1i} \sim^{\mathrm{iid}} N_{d_1}(\mathbf{0}, \mathbf{I}_{d_1})$ and $A_{1i}|\mathbf{S}_{1i} \sim \mathrm{Bernoulli}(\pi_a(\mathbf{S}_{1i}))$, where $\pi_a(\mathbf{S}_{1i}) = g(\mathbf{V}_i^\top\boldsymbol{\gamma}_a)$. Let $\delta_{1i} \sim^{\mathrm{iid}} N(0,1)$, $\boldsymbol{\delta}_{1i} \sim^{\mathrm{iid}} N_{d_1}(0, \mathbf{I}_{d_1})$ and $\boldsymbol{\delta}_{2i} \sim^{\mathrm{iid}} N_{d_2}(0, \mathbf{I}_{d_2})$. The following models on $\mathbf{S}_{2i}|(\mathbf{S}_{1i}, A_{1i})$ are considered:

M4. (Sparse linear) $\mathbf{S}_{2i} = W_s(A_{1i})\mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$.

M5. (Dense linear) $\mathbf{S}_{2i} = W_d(A_{1i})\mathbf{S}_{1i} + A_{1i}(1+\delta_{1i})\mathbf{1}_{d_2\times 1} + \boldsymbol{\delta}_{2i}$.

M6. (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 $i \leq d_2$ and $j \leq d_1$,

$$\{W_s(a)\}_{i,j} = 0.8^{|i-j|}\mathbb{1}\{|i-j| \leq 1\}, \quad \{W_d(a)\}_{i,j} = 0.8^{|i-j|},$$

$$\{W_s(a')\}_{i,j} = 0.7^{|i-j|}\mathbb{1}\{|i-j| \leq 2\}, \quad \{W_d(a')\}_{i,j} = 0.7^{|i-j|},$$

$$\{\widetilde{W}_d(c)\}_{i,j} = \{W_d(c)\}_{i,j}\mathbb{1}\{j > 3\} \text{ for each } c \in \{a, a'\}.$$

The treatment indicators at time $t = 2$ are generated as

$$A_{2i}|(\mathbf{S}_i, A_{1i} = c_1) \sim \text{Bernoulli}(\rho_c(\mathbf{S}_i)), \text{ with}$$

$$\rho_c(\mathbf{S}_i) = g(c_1\mathbf{U}_i^\top\boldsymbol{\eta}_a + (1-c_1)\mathbf{U}_i^\top\boldsymbol{\eta}_{a'}), \text{ for each } c = (c_1, c_2) \in \{a, a'\}.$$

The outcome variables are generated as

$$Y_i = Y_i(A_{1i}, A_{2i}), \; Y_i(c) = \mathbf{U}_i^\top\boldsymbol{\alpha}_c + \zeta_i, \text{ for each } c \in \{a, a'\}, \text{ where } \zeta_i \sim^{\text{iid}} N(0,1).$$

The parameter values are chosen as $\boldsymbol{\alpha}_c = (\boldsymbol{\alpha}_{c,1}^\top, \boldsymbol{\alpha}_{c,2}^\top)^\top$, for each $c \in \{a, a'\}$, $\boldsymbol{\alpha}_{a,1} = (-1, -1, 1, -1, \mathbf{0}_{(d_1-3)})^\top$, $\boldsymbol{\alpha}_{a,2} = (-1, -1, 1, \mathbf{0}_{(d_2-3)})^\top$, $\boldsymbol{\alpha}_{a',1} = (1, 1, 1, -1, \mathbf{0}_{(d_1-3)})^\top$, $\boldsymbol{\alpha}_{a',2} = (1, 1, 1, \mathbf{0}_{(d_2-3)})^\top$, $\boldsymbol{\gamma}_a = (0, 1, 1, 1, \mathbf{0}_{(d_1-3)})^\top$, $\boldsymbol{\eta}_a = (0, 1, 1, \mathbf{0}_{(d_1-2)}, 1, -1, \mathbf{0}_{(d_2-2)})^\top$, and $\boldsymbol{\eta}_{a'} = (0, 0.5, 0, -0.5, \mathbf{0}_{(d_1-3)}, 0.5, 0, 0.5, \mathbf{0}_{(d_2-3)})^\top$, where $\mathbf{0}_q := (0, \ldots, 0) \in \mathbb{R}^q$ for any $q \geq 1$. Under the above DGPs, we have the following nuisance functions: for each $c \in \{a, a'\}$,

$$\nu_c(\mathbf{S}) = E[Y(c)|\mathbf{S}, A_1 = c_1] = \mathbf{U}^\top\boldsymbol{\alpha}_c, \tag{1.39}$$

$$\mu_c(\mathbf{S}_1) = E[Y(c)|\mathbf{S}_1, A_1 = c_1] = \mathbf{V}^\top\boldsymbol{\alpha}_{c,1} + E[\mathbf{S}_2^\top\boldsymbol{\alpha}_{c,2}|\mathbf{S}_1, A_1 = c_1] = \mathbf{V}^\top\boldsymbol{\beta}_c, \tag{1.40}$$

where $\boldsymbol{\beta}_c$ varies for different models on $\mathbf{S}_{2i}|(\mathbf{S}_{1i}, A_{1i})$ as follows:

M4. $\boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2}\boldsymbol{\alpha}_{a',2}\mathbb{1}\{c = a'\}, (W_s(c)\boldsymbol{\alpha}_{c,2})^\top)^\top$ with $\|\boldsymbol{\beta}_a\|_0 = 4$ and $\|\boldsymbol{\beta}_{a'}\|_0 = 5$.

Table 1.7: Simulation under M4. 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.

Estimator	Bias	RMSE	Length	Coverage	ESD	ASD
$N = 1000, N_1 = 279, N_0 = 312, d_1 = 100, d_2 = 50$						
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
DTL	0.063	0.218	1.121	0.934	0.326	0.286
S-DRL	0.133	0.202	0.881	0.874	0.262	0.225
S-DRL'	0.121	0.195	0.876	0.898	0.262	0.224
$N = 4000, N_1 = 1115, N_0 = 1248, d_1 = 100, d_2 = 50$						
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
DTL	0.029	0.119	0.600	0.928	0.171	0.153
S-DRL	0.031	0.122	0.601	0.926	0.169	0.153
S-DRL'	0.030	0.119	0.598	0.922	0.173	0.153

M5-6. $\boldsymbol{\beta}_c = \boldsymbol{\alpha}_{c,1} + (\sum_{j=1}^{d_2} \boldsymbol{\alpha}_{a',2} \mathbb{1}\{c = a'\}, (W_d(c)\boldsymbol{\alpha}_{c,2})^\top)^\top$ is weakly sparse in that $\|\boldsymbol{\beta}_a\|_0 = \|\boldsymbol{\beta}_{a'}\|_0 = d_1 + 1$, $\|\boldsymbol{\beta}_a\|_1 < 5.23$, and $\|\boldsymbol{\beta}_{a'}\|_1 < 7.24$.

The following choices of parameters are implemented: $N \in \{1000, 4000\}$, $d_1 = 100$, and $d_2 = d_1/2 = 50$. For each of the DGPs, we repeat the simulation for 500 times. For each replication, we consider the oracle estimator, the empirical difference estimator (empdiff), the DTL estimator, the S-DRL estimator, and the S-DRL' estimator as in Section 6.1. The results are reported in Tables 1.7-1.5.

The considered DGPs are only different on the procedure of generating $\mathbf{S}_2$ based on $\mathbf{S}_1$ and A_1. Under M4, 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 $\boldsymbol{\beta}_c$ is

a sparse vector. Under M5, we consider a dense linear dependence that the corresponding coefficient $\boldsymbol{\beta}_c$ is only weakly sparse that it's $\|\cdot\|_1$ norm is bounded. Under M6, we consider a dense quadratic dependence between $\mathbf{S}_2$ and $\mathbf{S}_1$ but the nuisance function $\mu_c(\cdot)$ is still linear - we can see that the nuisance function can be linear even when $\mathbf{S}_2$ is not linearly dependent on $\mathbf{S}_1$. Note that, although $E(\mathbf{S}_2|\mathbf{S}_1, A_1 = c_1)$ is quadratic in $\mathbf{S}_1$, $E(\mathbf{S}_2^\top \alpha_{c,2}|\mathbf{S}_1, A_1 = c_1)$ is still linear on $\mathbf{S}_1$ and hence the linear working models $\mu_c^*(\cdot) = \mu_{c,\text{NR}}^*(\cdot)$ are both correctly specified as the second-time conditional mean $\nu_c(\cdot)$ is also linear.

Table 1.8: Simulation under M5. The rest of the caption details remain the same as those in Table 1.7.

Estimator	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
DTL	0.084	0.219	1.139	0.920	0.322	0.291
S-DRL	0.097	0.225	1.140	0.922	0.323	0.291
S-DRL'	0.084	0.224	1.138	0.926	0.321	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
DTL	0.019	0.113	0.608	0.934	0.166	0.155
S-DRL	0.024	0.114	0.609	0.930	0.166	0.155
S-DRL'	0.019	0.114	0.609	0.932	0.166	0.155

We first focus on the DTL, S-DRL, and S-DRL' estimators and compare their behaviors. We can see that when the model is relatively easy (under M4), and the total sample size is relatively small ($N = 1000$), the DTL method provides better coverage but with a worse RMSE than the S-DRL and S-DRL' methods; see Table 1.7. This is because, although the DTL estimator has a smaller bias, it also has a larger ESD compared with the S-DRL and

Table 1.9: Simulation under M6. The rest of the caption details remain the same as those in Table 1.7.

Estimator	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
DTL	0.083	0.225	1.141	0.924	0.318	0.291
S-DRL	0.077	0.228	1.139	0.914	0.320	0.290
S-DRL'	0.076	0.213	1.135	0.920	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
DTL	0.019	0.114	0.610	0.936	0.166	0.156
S-DRL	0.021	0.115	0.610	0.928	0.166	0.156
S-DRL'	0.020	0.112	0.608	0.932	0.166	0.155

S-DRL' estimators. If we further increase the sample size ($N = 4000$), we can see that the coverages based on DTL, S-DRL, and S-DRL' estimators are close to each other and also overall acceptable. When the estimation of the first-time conditional mean is relatively hard as its linear parameter is only weakly sparse (under M4 and M5), we can see that the RMSEs and confidence intervals' coverages of the DTL, S-DRL, and S-DRL' methods are relatively close to each other for both $N = 1000$ and $N = 4000$; see Tables 1.8 and 1.9. In addition, we can see that the considered estimators have very similar behaviors among DGPs M5 and M6. Note that the nuisance functions, including the propensity score and conditional mean functions, are the same under M5 and M6, although the conditional densities of $\mathbf{S}_2$ given $(\mathbf{S}_1, A_1)$ are different. This observation indicates that the considered estimators' behavior mainly relies on the conditional means of the potential outcomes and treatment variables instead of the conditional densities. Lastly, we can also see that the naive empirical difference

estimator, $\widehat{\theta}_{\text{empdiff}}$, is not even consistent because of the appearance of confounders.

1.8.3 Proof of the results for the doubly robust representation

Proof of Lemma 1. By the tower rule and $Y = Y(A_1, A_2)$ under Assumption 1,

$$\begin{aligned} E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right] &= E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y(a)-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right] \\ &= E\left[E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y(a)-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \mid \mathbf{S}, A_1=a_1\right]P(A_1=a_1 \mid \mathbf{S})\right]. \end{aligned}$$

By $Y(a) \perp\!\!\!\perp A_2 \mid \mathbf{S}, A_1 = a_1$ under Assumption 1,

$$\begin{aligned} &E\left[E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y(a)-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \mid \mathbf{S}, A_1=a_1\right]P(A_1=a_1 \mid \mathbf{S})\right] \\ &= E\left[\frac{E[\mathbb{1}_{\{A_2=a_2\}} \mid \mathbf{S}, A_1=a_1](E[Y(a) \mid \mathbf{S}, A_1=a_1]-\nu_a^*(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}E[\mathbb{1}_{\{A_1=a_1\}} \mid \mathbf{S}]\right] \\ &\overset{(i)}{=} E\left[\frac{\rho_a(\mathbf{S})(\nu_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}E[\mathbb{1}_{\{A_1=a_1\}} \mid \mathbf{S}]\right] \\ &\overset{(ii)}{=} E\left[\mathbb{1}_{\{A_1=a_1\}}\frac{\rho_a(\mathbf{S})(\nu_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right], \end{aligned}$$

where (i) holds since $\rho_a(\mathbf{S}) = P[A_2 = a_2 \mid \mathbf{S}, A_1 = a_1]$ and $\nu_a(\mathbf{S}) = E[Y(a) \mid \mathbf{S}, A_1 = a_1]$; (ii) holds by the tower rule. Hence,

$$E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right] = E\left[\mathbb{1}_{\{A_1=a_1\}}\frac{\rho_a(\mathbf{S})(\nu_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right].$$

Observe that

$$\begin{aligned} &E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y-\nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}+\mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a^*(\mathbf{S})-\mu_a^*(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)}+\mu_a^*(\mathbf{S}_1)\right]-\theta_a \\ &= E\Bigg[\mathbb{1}_{\{A_1=a_1\}}\frac{\rho_a(\mathbf{S})(\nu_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}+\mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a^*(\mathbf{S})-\mu_a^*(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)} \\ &\quad +\mu_a^*(\mathbf{S}_1)-\mu_a(\mathbf{S}_1)\Bigg] \overset{(i)}{=} G_1+G_2+G_3, \end{aligned} \tag{1.41}$$

where

$$\begin{aligned} G_1 :=& E\left[\mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a^*(\mathbf{S})-\nu_a(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)}\left(1-\frac{\rho_a(\mathbf{S})}{\rho_a^*(\mathbf{S})}\right)\right], \\ G_2 :=& E\left[(\mu_a^*(\mathbf{S}_1)-\mu_a(\mathbf{S}_1))\left(1-\frac{\mathbb{1}_{\{A_1=a_1\}}}{\pi_a^*(\mathbf{S}_1)}\right)\right], \\ G_3 :=& E\left[\mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a(\mathbf{S})-\mu_a(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)}\right]. \end{aligned}$$

In the above, (i) holds by rearranging the terms after the following decomposition

$$\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1) = (\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})) + (\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1)) + (\mu_a(\mathbf{S}) - \mu_a^*(\mathbf{S}_1)).$$

By assumption, either $\nu_a^*(\cdot) = \nu_a(\cdot)$ or $\rho_a^*(\cdot) = \rho_a(\cdot)$, we have

$$G_1 = 0. \tag{1.42}$$

For G_2, by the tower rule,

$$\begin{aligned} G_2 &= E\left[E\left[(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))\left(1 - \frac{\mathbb{1}_{\{A_1=a_1\}}}{\pi_a^*(\mathbf{S}_1)}\right) \mid \mathbf{S}_1\right]\right] \\ &\overset{(i)}{=} E\left[(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))\left(1 - \frac{\pi_a(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)}\right)\right] \overset{(ii)}{=} 0, \end{aligned} \tag{1.43}$$

where (i) holds since $\pi_a(\mathbf{S}_1) = P[A_1 = a_1|\mathbf{S}_1]$; (ii) holds since $\pi_a(\mathbf{S}_1) = P[A_1 = a_1|\mathbf{S}_1]$; (ii) holds since, by assumption, either $\mu_a^*(\cdot) = \mu_a(\cdot)$ or $\pi_a^*(\cdot) = \pi_a(\cdot)$. For G_3, by the tower rule,

$$\begin{aligned} G_3 &= E\left[E\left[\mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)} \mid \mathbf{S}_1, A_1 = a_1\right] P(A_1 = a_1 \mid \mathbf{S}_1)\right] \\ &\overset{(i)}{=} E\left[\frac{\pi_a(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)}\left[E\left[\nu_a(\mathbf{S}) \mid \mathbf{S}_1, A_1 = a_1\right] - \mu_a(\mathbf{S}_1)\right]\right] \overset{(ii)}{=} 0, \end{aligned} \tag{1.44}$$

where (i) holds since $\pi_a(\mathbf{S}_1) = P[A_1 = a_1|\mathbf{S}_1]$; (ii) holds since $\mu_a(\mathbf{S}_1) = E[\nu_a(\mathbf{S}) \mid \mathbf{S}_1, A_1 = a_1]$. Combining (1.42)-(1.44) with (1.41), we have

$$\theta_a = E\left[\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} + \mathbb{1}_{\{A_1=a_1\}}\frac{\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)} + \mu_a^*(\mathbf{S}_1)\right].$$

■

Proof of Theorem 1. Observe that

$$\begin{aligned} &E\left[\nu_a^*(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right] - \mu_a(\mathbf{S}_1) \\ &\quad\overset{(i)}{=} E\left[\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y(a) - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right] \\ &\quad= E\left[(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))\left(1 - \frac{\mathbb{1}_{\{A_2=a_2\}}}{\rho_a^*(\mathbf{S})}\right) + \mathbb{1}_{\{A_2=a_2\}}\frac{Y(a) - \nu_a(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right], \end{aligned}$$

where (i) holds since $\mu_a(\mathbf{S}_1) = E[\nu_a(\mathbf{S}) \mid \mathbf{S}_1, A_1 = a_1]$ and $Y = Y(A_1, A_2)$ under Assumption 1. By the tower rule,

$$\begin{aligned} &E\left[(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))\left(1 - \frac{\mathbb{1}_{\{A_2=a_2\}}}{\rho_a^*(\mathbf{S})}\right) \mid \mathbf{S}_1, A_1 = a_1\right] \\ &\quad= E\left[E\left[(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))\left(1 - \frac{\mathbb{1}_{\{A_2=a_2\}}}{\rho_a^*(\mathbf{S})}\right) \mid \mathbf{S}, A_1 = a_1\right] \mid \mathbf{S}_1, A_1 = a_1\right] \\ &\quad\overset{(i)}{=} E\left[(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))\left(1 - \frac{\rho_a(\mathbf{S})}{\rho_a^*(\mathbf{S})}\right) \mid \mathbf{S}_1, A_1 = a_1\right] \overset{(ii)}{=} 0, \end{aligned}$$

where (i) holds since $\rho_a(\mathbf{S}) = P[A_2 = a_2 \mid \mathbf{S}, A_1 = a_1]$; (ii) holds since either $\nu_a^*(\cdot) = \nu_a(\cdot)$ or $\rho_a^*(\cdot) = \rho_a(\cdot)$ by Assumption. In addition, by the tower rule, we also have

$$
\begin{aligned}
&E\left[\mathbb{1}_{\{A_2=a_2\}} \frac{Y(a) - \nu_a(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right] \\
&\quad = E\left[E\left[\mathbb{1}_{\{A_2=a_2\}} \frac{Y(a) - \nu_a(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}, A_1 = a_1\right] \mid \mathbf{S}_1, A_1 = a_1\right] \\
&\quad \overset{(i)}{=} E\left[E\left[\mathbb{1}_{\{A_2=a_2\}} \mid \mathbf{S}, A_1 = a_1\right] \frac{E\left[Y(a) \mid \mathbf{S}, A_1 = a_1\right] - \nu_a(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1, A_1 = a_1\right] \overset{(ii)}{=} 0,
\end{aligned}
$$

where (i) holds since $Y(a) \perp\!\!\!\perp A_2 \mid \mathbf{S}, A_1 = a_1$ under Assumption 1; (ii) holds since $\nu_a(\mathbf{S}) = E\left[Y(a) \mid \mathbf{S}, A_1 = a_1\right]$. Therefore, for any $\mathbf{s}_1 \in \mathbb{R}^{d_1}$,

$$
\mu_a(\mathbf{s}_1) = E\left[\nu_a^*(\mathbf{S}) + \mathbb{1}_{\{A_2=a_2\}} \frac{Y - \nu_a^*(\mathbf{S})}{\rho_a^*(\mathbf{S})} \mid \mathbf{S}_1 = \mathbf{s}_1, A_1 = a_1\right].
$$

■

1.8.4 Convergence rates for nuisance estimators

Auxiliary Lemmas

The following lemmas will be helpful in our proofs.

Lemma 1.2 (Selection of Lemma D.1 of [CLCL19])**.** *Let $X, Y \in \mathbb{R}$ be a random variable. Then $\|cX\|_{\psi_2} = |c|\|X\|_{\psi_2}$ $\forall c \in \mathbb{R}$. If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_2} \leq \|Y\|_{\psi_2}$. Moreover, for X and Y sub-Gaussian, $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$. If X is bounded, i.e, $|X| \leq C$ a.s. for some constant C, then $\|X\|_{\psi_2} \leq (\log 2)^{-1/2} C$. If $\|X\|_{\psi_2} \leq \sigma$, then $E(|X|^m) \leq 2\sigma^m \Gamma(m/2 + 1)$ $\forall\, m \geq 1$, where $\Gamma(a) := \int_0^\infty x^{a-1} \exp(-x) dx$ $\forall a > 0$ denotes the Gamma function. Hence, $E(|X|) \leq \sigma\sqrt{\pi}$ and $E(|X|^m) \leq 2\sigma^m (m/2)^{m/2}$ $\forall\, m \geq 2$. Let $\{X_i\}_{i=1}^n$ 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 1.3. *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 $\|X\|_{\psi_2} \leq 2\sigma$.*

The following lemma provides the same type of results as used in the Assumption 3 but now for covariates at different exposure time and different treatment paths.

Lemma 1.4. *Let the overlap conditions of Assumption 1 and Assumption 3 hold. Consider the constants c_0, κ_l, σ_u defined as in Assumptions 1 and 3. Then, the following statements hold:*

(a) $0 < c_0\kappa_l \leq \lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) \leq \lambda_{\max}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) \leq 2\sigma_u^2 < \infty$ *and* $\widetilde{\mathbf{U}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^\top\widetilde{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d+1}$*;*

(b) $0 < \kappa_l \leq \lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \leq \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \leq 2\sigma_u^2 < \infty$ *and* $\bar{\mathbf{U}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^\top\bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d+1}$*;*

c) $0 < \kappa_l \leq \lambda_{\min}(E[\mathbf{U}\mathbf{U}^\top]) \leq \lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) \leq 2\sigma_u^2 < \infty$ *and* $\mathbf{U}$ *is sub-Gaussian with* $\|\boldsymbol{x}^\top\mathbf{U}\|_{\psi_2} \leq \sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$*;*

d) $0 < \kappa_l \leq \lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) \leq \lambda_{\max}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) \leq 2\sigma_u^2 < \infty$ *and* $\bar{\mathbf{V}}$ *is sub-Gaussian with* $\|\boldsymbol{x}^\top\bar{\mathbf{V}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$*;*

e) $0 < \kappa_l \leq \lambda_{\min}(E[\mathbf{V}\mathbf{V}^\top]) \leq \lambda_{\max}(E[\mathbf{V}\mathbf{V}^\top]) \leq 2\sigma_u^2 < \infty$ *and* $\mathbf{V}$ *is sub-Gaussian with* $\|\boldsymbol{x}^\top\mathbf{V}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$ *for any* $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$*.*

The second-time conditional mean model The following lemma characterizes the estimation error of the second-time conditional mean model, $\nu_a^*(\cdot)$. The corresponding conditional mean estimator is defined as $\widehat{\nu}_a(\mathbf{S}) = \mathbf{U}^\top\widehat{\boldsymbol{\alpha}}_a$.

Lemma 1.5. *Let Assumptions 1-3 hold. For any* $t > 0$*, choose* $\lambda_{\boldsymbol{\alpha}} := 32\sigma\sigma_u\sigma_\zeta(t + \sqrt{\log(d+1)/|\mathcal{J}|})$*. Let* $|\mathcal{J}| \geq \max\{\log(d+1), 100\kappa_2^2 s_{\boldsymbol{\alpha}_a}\log(d+1)\}$*. Then* $\widehat{\boldsymbol{\alpha}}_a$*, (2.6), sat-*

isfies

$$\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 \;\le\; 8\kappa_1^{-1}\lambda_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\alpha}_a}}, \quad \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}[\widetilde{\mathbf{U}}_i^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \;\le\; 32\kappa_1^{-1}\lambda_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}, \tag{1.45}$$

with probability at least $1-2\exp\left(-\frac{4|\mathcal{J}|t^2}{1+2t+\sqrt{2t}}\right)-c_1\exp(-c_2|\mathcal{J}|)$ *and some constants* $c_1, c_2, \kappa_1, \kappa_2 > 0$. *In addition, assume* $|\mathcal{J}| \asymp N$ *and* $N \gg s_{\boldsymbol{\alpha}_a}\log(d)$. *Choose some* $\lambda_{\boldsymbol{\alpha}} \asymp \sigma\sqrt{\log(d)/N}$. *Then for any constant* $r \ge 1$, *as* $N, d \to \infty$, *we have*

$$\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 = O_p\left(\sigma\sqrt{s_{\boldsymbol{\alpha}_a}\log(d)/N}\right), \tag{1.46}$$

$$\{E[\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})]^r\}^{1/r} = O_p\left(\sigma\sqrt{s_{\boldsymbol{\alpha}_a}\log(d)/N}\right). \tag{1.47}$$

In the above, the expectation of the left-hand side of (1.47) is taken respect to the distribution of a new observation's covariate vector $\mathbf{S}$.

The nested-regression-based estimator $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$ proposed in Section 2.2 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 $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$ 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 $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$, 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 (1.45) 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 1.6. *Let Assumptions of Lemma 1.5 hold. In addition, let* $|\mathcal{J}| \ge \max\{\log(d+1), (c_3+100\kappa_2^2)s_{\boldsymbol{\alpha}_a}\log(d+1)\}$, *with constant* $c_3 > 0$. *Then*

$$\frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}[\bar{\mathbf{U}}_i^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \le 288\sigma_u\kappa_1^{-2}\lambda_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a},$$

with probability at least $1-2\exp\left(-\frac{4|\mathcal{J}|t^2}{1+2t+\sqrt{2t}}\right)-c_1\exp(-c_2|\mathcal{J}|)-2\exp(-c_4|\mathcal{J}|)$ *and constants* $c_1, c_2, c_4 > 0$.

The propensity score models The following lemma provides asymptotic upper bounds on the estimation errors of the propensity score models, $\pi_a^*(\cdot)$ and $\rho_a^*(\cdot)$. The corresponding propensity score estimators are defined as $\widehat{\pi}_a(\mathbf{S}_1) = g(\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_a)$ and $\widehat{\rho}_a(\mathbf{S}) = g(\mathbf{U}^\top \widehat{\boldsymbol{\delta}}_a)$, respectively.

Lemma 1.7. *Let the overlap conditions of Assumption 1 and Assumptions 3-4 hold. Let the sample size be such that* $|\mathcal{J}| \asymp N$ *and* $N \gg \max\{s_{\boldsymbol{\gamma}_a} \log(d_1), s_{\boldsymbol{\delta}_a} \log(d)\}$. *Then, as* $N, d \to \infty$, *(a) the logistic Lasso* (2.3) *with* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log(d_1)/N}$ *satisfies*

$$\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\gamma}_a} \log(d_1)/N}\right), \tag{1.48}$$

$$E[\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1)]^2 = O_p\left(s_{\boldsymbol{\gamma}_a} \log(d_1)/N\right), \tag{1.49}$$

whereas (b) the logistic Lasso (2.4) *with* $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log(d)/N}$ *satisfies*

$$\|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\delta}_a} \log(d)/N}\right), \tag{1.50}$$

$$E[\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S})]^2 = O_p\left(s_{\boldsymbol{\delta}_a} \log(d)/N\right). \tag{1.51}$$

In the left-hand side of (1.49) *and* (1.51), *the expectations are only taken w.r.t. the distribution of the new observations* $\mathbf{S}_1$ *and* $\mathbf{S}$, *respectively. Note that Assumption 4 holds under Assumption 1 when* $\pi_a^*(\mathbf{S}_1)$ *and* $\rho_a^*(\mathbf{S})$ *are correctly specified.*

Lemma 1.8. *Let Assumptions of Lemma 1.7 hold. Define the event* $\mathcal{A} := \{\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2 \leq 1, \|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2 \leq 1\}$. *Then, as* $N, d \to \infty$, $P(\mathcal{A}) = 1 - o(1)$. *Moreover, on the event* $\mathcal{A}$, *as* $N, d \to \infty$, $\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-r}\}^{1/r}$ *and* $\{E|\widehat{\rho}_a(\mathbf{S})|^{-r}\}^{1/r}$ *are both bounded uniformly by some*

constants independent of N and for $r > 2$,

$$\left\{E\left|\widehat{\pi}_a^{-1}(\mathbf{S}_1) - \pi_a^{*-1}(\mathbf{S}_1)\right|^r\right\}^{1/r} = O_p\left(\sqrt{s_{\boldsymbol{\gamma}_a}\log(d_1)/N}\right),$$

$$\left\{E\left|\widehat{\rho}_a^{-1}(\mathbf{S}) - \rho_a^{*-1}(\mathbf{S})\right|^r\right\}^{1/r} = O_p\left(\sqrt{s_{\boldsymbol{\delta}_a}\log(d)/N}\right),$$

$$\left\{E\left|\widehat{\pi}_a^{-1}(\mathbf{S}_1)\widehat{\rho}_a^{-1}(\mathbf{S}) - \pi_a^{*-1}(\mathbf{S}_1)\rho_a^{*-1}(\mathbf{S})\right|^r\right\}^{1/r} = O_p\left(\sqrt{(s_{\boldsymbol{\gamma}_a}\log(d_1) + s_{\boldsymbol{\delta}_a}\log(d))/N}\right).$$

In the left-hand side of the equations above, the expectations are only taken w.r.t. the distribution of the new observations $\mathbf{S}_1$ or $\mathbf{S}$.

Convergence rate for the general imputed Lasso estimator

Proof of Theorem 8. By the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\frac{1}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}]^2 + \lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^\top\boldsymbol{\beta}^*]^2 + \lambda_M\|\boldsymbol{\beta}^*\|_1,$$

or, expanding and rearranging,

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 + \lambda_M\|\widehat{\boldsymbol{\beta}}\|_1 \\
&\quad\le \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - \mathbf{X}_i^\top\boldsymbol{\beta}^*]\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda_M\|\boldsymbol{\beta}^*\|_1 \\
&\quad= \frac{2}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda_M\|\boldsymbol{\beta}^*\|_1. \qquad (1.52)
\end{aligned}$$

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

$$\begin{aligned}
P\left(\max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right) &= P\left(\bigcup_{j=1}^{d}\left\{\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right\}\right) \\
&\le \sum_{j=1}^{d}P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right). \qquad (1.53)
\end{aligned}$$

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 $\|\boldsymbol{e}_j^\top \mathbf{X}\|_{\psi_2} \le \sigma_{\mathbf{X}}$ and $\|\varepsilon\|_{\psi_2} \le \sigma$, by Lemma 1.2,

$$\|\boldsymbol{e}_j^\top \mathbf{X}\varepsilon\|_{\psi_1} \le \|\boldsymbol{e}_j^\top \mathbf{X}\|_{\psi_2} \cdot \|\varepsilon\|_{\psi_2} \le \sigma\sigma_{\mathbf{X}}.$$

Note that, here we do not make any assumption on the sample gram matrix $\hat{\boldsymbol{\Sigma}} := M^{-1}\sum_{i=1}^M \mathbf{X}_i\mathbf{X}_i^\top$, e.g., $\sup_{1\le j\le d}\hat{\boldsymbol{\Sigma}}_{j,j} \le 1$ as required in [Wai19,NRWY12]. Instead, we consider $\boldsymbol{e}_j^\top \mathbf{X}\varepsilon$ as a sub-exponential random variable, and the Bernstein's inequality is applied in the following to control (1.53). Recall the definition of $\boldsymbol{\beta}^*$, we have $E[\mathbf{X}\varepsilon] = 0$. By Bernstein's inequality, for each $1 \le j \le d$,

$$P\left(\left|\frac{1}{M}\sum_{i=1}^M \mathbf{X}_{i,j}\varepsilon_i\right| \ge 2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2\right) \le 2\exp\left(-M\epsilon^2\right), \quad \text{for any } \epsilon > 0. \tag{1.54}$$

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

$$2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2 \le 4\sigma\sigma_{\mathbf{X}}\left(\sqrt{\frac{\log(d)}{M}} + t\right) = \frac{\lambda_M}{4}. \tag{1.55}$$

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 (1.54) and (1.55), we have, for each $1 \le j \le d$,

$$P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge \frac{\lambda_M}{4}\right) \le P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \ge 2\sigma\sigma_{\mathbf{X}}\epsilon + \sigma\sigma_{\mathbf{X}}\epsilon^2\right)$$
$$\le 2\exp\left(-M\epsilon^2\right) \le \frac{2}{d}\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right).$$

Together with (1.53),

$$P(\mathcal{E}_2) = P\left(\max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \le \frac{\lambda_M}{4}\right) \ge 1 - 2\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right). \tag{1.56}$$

On the event $\mathcal{E}_2$, we have

$$\left|\frac{2}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right| \le 2\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \max_{1\le j\le d}\left|\frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_{i,j}\varepsilon_i\right| \le \lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1/2. \tag{1.57}$$

As for the second term of (1.52), by the fact that $2ab \le a^2+b^2$ for any $a, b \in \mathbb{R}$, and we set $a = \sqrt{2}[\widehat{Y}_i - Y_i^*]$, $b = \mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/\sqrt{2}$, we have

$$\left|\frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right| \le \frac{2}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]^2 + \frac{1}{2M}\sum_{i=1}^{M}\left[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2$$
$$\le 2\delta_M^2 + \frac{1}{2M}\sum_{i=1}^{M}\left[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2, \tag{1.58}$$

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 (1.52) by 2, we have

$$\frac{2}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1$$
$$\le \frac{4}{M}\sum_{i=1}^{M}\varepsilon_i\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + \frac{4}{M}\sum_{i=1}^{M}[\widehat{Y}_i - Y_i^*]\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + 2\lambda_M\|\boldsymbol{\beta}^*\|_1.$$

Together with (1.57) and (1.58), on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\frac{2}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1$$
$$\le \lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 + \frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 + 4\delta_M^2 + 2\lambda_M\|\boldsymbol{\beta}^*\|_1. \tag{1.59}$$

Hence,

$$\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2+2\lambda_M\|\widehat{\boldsymbol{\beta}}\|_1\le\lambda_M\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+2\lambda_M\|\boldsymbol{\beta}^*\|_1+4\delta_M^2$$

$$=\lambda_M\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+\lambda_M\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1+2\lambda_M\|\boldsymbol{\beta}_S^*\|_1+4\delta_M^2, \tag{1.60}$$

where $S:=\{j\le d:\boldsymbol{\beta}_j^*\neq 0\}$ 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,

$$\|\widehat{\boldsymbol{\beta}}\|_1=\|\widehat{\boldsymbol{\beta}}_S\|_1+\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1\ge\|\boldsymbol{\beta}_S^*\|_1-\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1. \tag{1.61}$$

By (1.60) and (1.61), on the event $\mathcal{E}_1\cap\mathcal{E}_2$, we get that

$$\frac{1}{M}\sum_{i=1}^{M}[\mathbf{X}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2+\lambda_M\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1\le 3\lambda_M\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+4\delta_M^2. \tag{1.62}$$

By Lemma 4.5 of [ZCB21], there exist constants $\kappa_1,\kappa_2>0$, such that

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\Delta})^2\ge\kappa_1\|\boldsymbol{\Delta}\|_2\left\{\|\boldsymbol{\Delta}\|_2-\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\Delta}\|_1\right\}\quad\text{for all }\ \|\boldsymbol{\Delta}\|_2\le 1, \tag{1.63}$$

with probability at least $1-c_1\exp(-c_2M)$ and some constants $c_1,c_2>0$. Note that 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

$$\mathcal{E}_3:=\left\{\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2\ge\kappa_1\|\boldsymbol{\delta}\|_2^2-\kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1\|\boldsymbol{\delta}\|_2\right\}. \tag{1.64}$$

Let $\boldsymbol{\Delta}=\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$. Then, $\|\boldsymbol{\Delta}\|_2=1$ and hence by (1.63),

$$P(\mathcal{E}_3)\ge 1-c_1\exp(-c_2M).$$

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 (1.62),

$$\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1 + 4\lambda_M^{-1}\delta_M^2 \leq 16\lambda_M^{-1}\delta_M^2.$$

Hence,

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \leq 20\lambda_M^{-1}\delta_M^2,$$

and

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 \leq 3\lambda_M\|\boldsymbol{\delta}_S\|_1 + 4\delta_M^2 \leq 16\delta_M^2.$$

In addition, on the event $\mathcal{E}_3$,

$$\kappa_1\|\boldsymbol{\delta}\|_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1\|\boldsymbol{\delta}\|_2 \leq \frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 \leq 16\delta_M^2.$$

It follows that,

$$\begin{aligned}
\|\boldsymbol{\delta}\|_2 &\leq \frac{\kappa_1\kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1 + \sqrt{\kappa_1^2\kappa_2^2\frac{\log(d)}{M}\|\boldsymbol{\delta}\|_1^2 + 64\kappa_1\delta_M^2}}{2\kappa_1} \\
&\leq \kappa_2\sqrt{\frac{\log(d)}{M}}\|\boldsymbol{\delta}\|_1 + 4\kappa_1^{-1/2}\delta_M \leq 20\kappa_2\sqrt{\frac{\log(d)}{M}}\lambda_M^{-1}\delta_M^2 + 4\kappa_1^{-1/2}\delta_M \\
&\leq \frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M,
\end{aligned}$$

since $\lambda_M = 16\sigma\sigma_{\mathbf{X}}(\sqrt{\frac{\log(d)}{M}} + t) \geq 16\sigma\sigma_{\mathbf{X}}\sqrt{\frac{\log(d)}{M}}$.

Case 2. Case of $\|\boldsymbol{\delta}_S\|_1 \geq 4\lambda_M^{-1}\delta_M^2$. Then, by (1.62),

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 + \lambda_M\|\boldsymbol{\delta}_{S^c}\|_1 \leq \lambda_M(3\|\boldsymbol{\delta}_S\|_1 + 4\lambda_M^{-1}\delta_M^2) \leq 4\lambda_M\|\boldsymbol{\delta}_S\|_1, \tag{1.65}$$

and hence

$$\|\boldsymbol{\delta}_{S^c}\|_1 \leq 4\|\boldsymbol{\delta}_S\|_1. \tag{1.66}$$

Notice that, $\|\boldsymbol{\delta}_S\|_1 \leq \sqrt{s}\|\boldsymbol{\delta}_S\|_2$. It follows that

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \leq 5\|\boldsymbol{\delta}_S\|_1 \leq 5\sqrt{s}\|\boldsymbol{\delta}_S\|_2 \leq 5\sqrt{s}\|\boldsymbol{\delta}\|_2.$$

Hence, under the event $\mathcal{E}_3$, when $M > 100{\kappa_2}^2 s\log(d)$,

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 &\geq \kappa_1\|\boldsymbol{\delta}\|_2^2 - 5\kappa_1\kappa_2\sqrt{\frac{s\log(d)}{M}}\|\boldsymbol{\delta}\|_2^2 \\ &\geq \frac{\kappa_1}{2}\|\boldsymbol{\delta}\|_2^2 \geq \frac{\kappa_1}{2}\|\boldsymbol{\delta}_S\|_2^2 \geq \frac{\kappa_1}{2s}\|\boldsymbol{\delta}_S\|_1^2. \end{aligned} \tag{1.67}$$

Together with (1.65), we have

$$\frac{\kappa_1}{2s}\|\boldsymbol{\delta}_S\|_1^2 \leq \frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 \leq 4\lambda_M\|\boldsymbol{\delta}_S\|_1.$$

Hence, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|\boldsymbol{\delta}_S\|_1 \leq 8\kappa_1^{-1}s\lambda_M. \tag{1.68}$$

By (1.66),

$$\|\boldsymbol{\delta}\|_1 \leq \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1 \leq 5\|\boldsymbol{\delta}_S\|_1 \leq 40\kappa_1^{-1}s\lambda_M.$$

Besides, by (1.65) and (1.68),

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 \leq 4\lambda_M\|\boldsymbol{\delta}_S\|_1 \leq 32\kappa_1^{-1}s\lambda_M^2.$$

Additionally, by (1.67), when $M > 100{\kappa_2}^2 s\log(d)$,

$$\|\boldsymbol{\delta}\|_2 \leq \sqrt{\frac{2}{\kappa_1 M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2} \leq 8\kappa_1^{-1}\sqrt{s}\lambda_M.$$

To sum up, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and when $M > \max\{\log(d), 100{\kappa_2}^2 s\log(d)\}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \max\left(\frac{5\kappa_2\delta_M^2}{4\sigma\sigma_{\mathbf{X}}} + 4\kappa_1^{-1/2}\delta_M, 8\kappa_1^{-1}\sqrt{s}\lambda_M\right), \tag{1.69}$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \max\left(20\lambda_M^{-1}\delta_M^2, 40\kappa_1^{-1}s\lambda_M\right), \tag{1.70}$$

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{X}_i^\top\boldsymbol{\delta})^2 \leq \max\left(16\delta_M^2, 32\kappa_1^{-1}s\lambda_M^2\right). \tag{1.71}$$

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 $M, d \to \infty$,

$$P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - 2\exp\left(-\frac{4Mt^2}{1+2t+\sqrt{2t}}\right) - c_1\exp(-c_2 M) - o(1).$$

■

Convergence rates for nuisance estimators with imputed outcomes

DR-imputation-based estimator

Proof of Theorem 9. We first consider the DR-imputation-based estimator $\widehat{\boldsymbol{\beta}}_{a,2} = \widehat{\boldsymbol{\beta}}_a(\mathcal{D}_{\mathcal{J}_2}, \widehat{Y}^{\mathrm{DR}}_{\mathcal{J}_2})$ defined as (2.11). In this case, the expectations are taken w.r.t. the samples in $\mathcal{D}_{\mathcal{J}_2}$; with a slight abuse of notation, $\widetilde{\boldsymbol{\delta}}_a := \widehat{\boldsymbol{\delta}}_a(\mathcal{D}_{\mathcal{J}_1})$ and $\widetilde{\boldsymbol{\alpha}}_a := \widehat{\boldsymbol{\alpha}}_a(\mathcal{D}_{\mathcal{J}_1})$ are fitted using samples in $\mathcal{D}_{\mathcal{J}_1}$ and are treated as fixed or condition on. Repeat the same procedure as in (1.52), we have

$$\begin{aligned}
&\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\bar{\mathbf{V}}_i^\top \boldsymbol{\Delta}_{\boldsymbol{\beta}})^2 + \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{a,2}\|_1 \leq \frac{2}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\mathbb{1}_{\{A_{1i}=a_1\}}(\widehat{Y}_i^{\mathrm{DR}} - \mathbf{V}_i^\top\boldsymbol{\beta}_a^*)\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_a^*\|_1 \\
&= \frac{2}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\Delta_{1i}+\Delta_{2i}+\Delta_{3i}+\Delta_{4i}+\Delta_{5i}+\Delta_{6i})\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_a^*\|_1 \\
&\leq \sum_{l=1}^{3}\left\|\frac{2}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{li}\bar{\mathbf{V}}_i\right\|_\infty\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \frac{2}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\left(\sum_{l=4}^{6}\Delta_{li}\right)^2 \\
&\quad + \frac{1}{2|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}})^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_a^*\|_1,
\end{aligned}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{a,2} - \boldsymbol{\beta}_a^*$,

$$
\begin{aligned}
\Delta_{1i} :=& \bar{\mathbf{U}}_i^\top \boldsymbol{\alpha}_a^* + \frac{\widetilde{Y}_i - \widetilde{\mathbf{U}}_i^\top \boldsymbol{\alpha}_a^*}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} - \bar{\mathbf{V}}_i^\top \boldsymbol{\beta}_a^*, \\
\Delta_{2i} :=& \left\{ 1 - \frac{\mathbb{1}_{\{A_{2i}=a_2\}}}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} \right\} \bar{\mathbf{U}}_i^\top (\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*) \mathbb{1}_{\{\rho_a^*(\cdot)=\rho_a(\cdot)\}}, \\
\Delta_{3i} :=& \left\{ \frac{1}{g(\mathbf{U}_i^\top \widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} \right\} (\widetilde{Y}_i - \widetilde{\mathbf{U}}_i^\top \boldsymbol{\alpha}_a^*) \mathbb{1}_{\{\nu_a^*(\cdot)=\nu_a(\cdot)\}}, \\
\Delta_{4i} :=& - \left\{ \frac{1}{g(\mathbf{U}_i^\top \widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} \right\} \widetilde{\mathbf{U}}_i^\top (\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*), \\
\Delta_{5i} :=& \left\{ 1 - \frac{\mathbb{1}_{\{A_{2i}=a_2\}}}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} \right\} \bar{\mathbf{U}}_i^\top (\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*) \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}, \\
\Delta_{6i} :=& \left\{ \frac{1}{g(\mathbf{U}_i^\top \widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}_i^\top \boldsymbol{\delta}_a^*)} \right\} (\widetilde{Y}_i - \widetilde{\mathbf{U}}_i^\top \boldsymbol{\alpha}_a^*) \mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}},
\end{aligned}
$$

and $g(u) = \exp(u)/\{1+\exp(u)\}$ is the logistic function. Let Δ_l be an independent copy of Δ_{li} for $1 \le l \le 6$. We first show that $\Delta_l \bar{\mathbf{V}}$ are zero mean vectors for each $l \in \{1,2,3\}$. By the definition of $\boldsymbol{\beta}_a^*$, we have $E[\Delta_1 \bar{\mathbf{V}}] = \mathbf{0}$. By the tower rule, we have

$$
\begin{aligned}
&E[\Delta_2 \bar{\mathbf{V}}] \\
&= E\left[P(A_1 = a_1 \mid \mathbf{U}) E\left[1 - \frac{\mathbb{1}_{\{A_2=a_2\}}}{\rho_a^*(\mathbf{S})} \mid \mathbf{U}, A_1 = a_1 \right] \mathbf{U}^\top (\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*) \mathbf{V} \mathbb{1}_{\{\rho_a^*(\cdot)=\rho_a(\cdot)\}} \right] \\
&= E\left[P(A_1 = a_1 \mid \mathbf{U}) \left\{ 1 - \frac{\rho_a(\mathbf{S})}{\rho_a^*(\mathbf{S})} \right\} \mathbf{U}^\top (\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*) \mathbf{V} \mathbb{1}_{\{\rho_a^*(\cdot)=\rho_a(\cdot)\}} \right] = \mathbf{0}.
\end{aligned}
$$

Similarly, we also have

$$
\begin{aligned}
&E[\Delta_3 \bar{\mathbf{V}}] \\
&= E\Bigg[P(A_1 = a_1 \mid \mathbf{U}) \left\{ \frac{1}{g(\mathbf{U}^\top \widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}^\top \boldsymbol{\delta}_a^*)} \right\} \\
&\qquad \cdot E[(Y(a) - \mathbf{U}^\top \boldsymbol{\alpha}_a^*) \mathbb{1}_{\{A_2 = a_2\}} \mid \mathbf{U}, A_1 = a_1] \mathbf{V} \mathbb{1}_{\{\nu_a^*(\cdot) = \nu_a(\cdot)\}} \Bigg] \\
&= E\Bigg[P(A_1 = a_1 \mid \mathbf{U}) \left\{ \frac{1}{g(\mathbf{U}^\top \widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}^\top \boldsymbol{\delta}_a^*)} \right\} \\
&\qquad \cdot P(A_2 = a_2 \mid \mathbf{U}, A_1 = a_1)[\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S})] \mathbf{V} \mathbb{1}_{\{\nu_a^*(\cdot) = \nu_a(\cdot)\}} \Bigg] = \mathbf{0}.
\end{aligned}
$$

Let $\boldsymbol{e}_j \in \mathbb{R}^d$ be the vector whose j-th element is 1 and other elements are 0s, for each $1 \le j \le d_1 + 1$. Under Assumption 4, we have $|\Delta_1 \bar{\mathbf{V}}^\top \boldsymbol{e}_j| = |(\varepsilon_a + g^{-1}(\mathbf{U}^\top \boldsymbol{\delta}_a^*)\zeta_a) \bar{\mathbf{V}}^\top \boldsymbol{e}_j| \le (|\varepsilon_a| + c_0^{-1}|\zeta_a|)|\bar{\mathbf{V}}^\top \boldsymbol{e}_j|$ for each $1 \le j \le d_1 + 1$. By Lemma 1.2,

$$
\|\Delta_1 \bar{\mathbf{V}}^\top \boldsymbol{e}_j\|_{\psi_1} \le (\|\varepsilon_a\|_{\psi_2} + c_0^{-1}\|\zeta_a\|_{\psi_2}) \|\bar{\mathbf{V}}^\top \boldsymbol{e}_j\|_{\psi_2} \overset{(i)}{\le} \sigma(\sigma_\zeta + c_0^{-1}\sigma_\varepsilon)\sigma_u,
$$

where (i) holds by Assumptions 2 and 3. By Lemma D.4 of [CLCL19], for each $1 \le j \le d_1 + 1$ and any $t > 0$,

$$
P\left(\left| \frac{1}{|\mathcal{J}_2|} \sum_{i \in \mathcal{J}_2} \Delta_{1i} \bar{\mathbf{V}}_i^\top \boldsymbol{e}_j \right| > h(t) \right) \le 2\exp(-t - \log(d_1 + 1)).
$$

where $h(t) = \sigma(\sigma_\zeta + c_0^{-1}\sigma_\varepsilon)\sigma_u \left(2\sqrt{\frac{t + \log(d_1+1)}{|\mathcal{J}_2|}} + \frac{t + \log(d_1+1)}{|\mathcal{J}_2|} \right)$. It follows that,

$$
\begin{aligned}
P\left(\left\| \frac{1}{|\mathcal{J}_2|} \sum_{i \in \mathcal{J}_2} \Delta_{1i} \bar{\mathbf{V}}_i \right\|_\infty > h(t) \right) &\le \sum_{j=1}^{d_1+1} P\left(\left| \frac{1}{|\mathcal{J}_2|} \sum_{i \in \mathcal{J}_2} \Delta_{1i} \bar{\mathbf{V}}_i^\top \boldsymbol{e}_j \right| > h(t) \right) \\
&\le 2(d_1 + 1)\exp(-t - \log(d_1 + 1)) = 2\exp(-t).
\end{aligned}
$$

Therefore, by $|\mathcal{J}| \asymp N$, we have

$$
\left\| \frac{1}{|\mathcal{J}_2|} \sum_{i \in \mathcal{J}_2} \Delta_{1i} \bar{\mathbf{V}}_i \right\|_\infty = O_p\left(\sigma \sqrt{\frac{\log(d_1)}{N}} \right). \tag{1.72}
$$

In addition, note that $|\Delta_2 \bar{\mathbf{V}}^T \mathbf{e}_j| \leq (1 + c_0^{-1})|\mathbf{U}_i^T(\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}^*)\mathbf{V}^T\mathbf{e}_j|$ under Assumption 4. By Lemma 1.2, conditional on $\mathcal{D}_{\mathcal{J}_1}$, we have

$$\|\Delta_2 \bar{\mathbf{V}}^T \mathbf{e}_j\|_{\psi_1} \leq (1 + c_0^{-1})\|\mathbf{U}_i^T(\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}^*)\|_{\psi_2}\|\mathbf{V}^T\mathbf{e}_j\|_{\psi_2} \overset{(i)}{\leq} (1 + c_0^{-1})\|\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}^*\|_2\sigma_u^2.$$

where (i) holds by Assumption 3. By Lemma D.4 of [CLCL19] and the union bound, for any $t > 0$, we have

$$\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{2i}\bar{\mathbf{V}}_i\right\|_\infty > (1 + c_0^{-1})\|\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}^*\|_2\sigma_u^2\left(2\sqrt{\frac{t + \log(d_1 + 1)}{|\mathcal{J}_2|}} + \frac{t + \log(d_1 + 1)}{|\mathcal{J}_2|}\right)$$

with probability at most $2\exp(-t)$. Therefore, by $|\mathcal{J}| \asymp N$, we have

$$\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{2i}\bar{\mathbf{V}}_i\right\|_\infty = O_p\left(\|\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}^*\|_2\sqrt{\frac{\log(d_1)}{N}}\right) \overset{(i)}{=} o_p\left(\sigma\sqrt{\frac{\log(d_1)}{N}}\right), \tag{1.73}$$

where (i) holds by Lemma 1.5 with $s_{\boldsymbol{\alpha}_a}\log(d) = o(N)$.

Besides, by Corollary 2.3 of [DVDGVW10], we have

$$E\left[\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty^2\right] \leq \frac{(2e\log(d_1) - e)E[\|\Delta_{3i}\bar{\mathbf{V}}_i\|_\infty^2]}{|\mathcal{J}_2|}.$$

(a) If $\nu_a^*(\cdot) = \nu_a(\cdot)$ and $\|\mathbf{S}_1\|_\infty \leq C$ almost surely, we have $\|\bar{\mathbf{V}}_i\|_\infty \leq \|\mathbf{V}_i\|_\infty \leq \max\{1, C\}$, which implies

$$E\left[\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty^2\right] = O\left(\frac{E[\Delta_3^2]\log(d_1)}{N}\right),$$

since $|\mathcal{J}| \asymp N$. Under Assumption 2, by Lemma 1.2, we have $E[\zeta^8] \leq 2^9\sigma^8\sigma_\zeta^8$. Together with Lemma 1.8, we have

$$E[\Delta_3^2] \leq \sqrt{E[\Delta_3^4]} \leq \left\{E[\zeta^8]E[g^{-1}(\mathbf{U}_i^\top\widetilde{\boldsymbol{\delta}}_a) - g^{-1}(\mathbf{U}_i^\top\boldsymbol{\delta}_a^*)]^8\right\}^{1/4} = O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a}\log(d)}{N}\right). \tag{1.74}$$

Hence,

$$E\left[\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty^2\right] = O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a}\log(d)}{N}\cdot\frac{\log(d_1)}{N}\right) = o_p\left(\frac{\sigma^2\log(d)}{N}\right), \tag{1.75}$$

since $s_{\boldsymbol{\delta}_a}\log(d) = o(N)$.

(b) If $\nu_a^*(\cdot) = \nu_a(\cdot)$ and $s_{\boldsymbol{\delta}_a}\log(d_1)\log(d) = O(N)$, by Lemma 1.2, we have $\|\|\bar{\mathbf{V}}_i\|_\infty\|_{\psi_2}$ $\leq \sigma_u\sqrt{\log(d_1+1)+2}$, which implies that $E\|\bar{\mathbf{V}}_i\|_\infty^4 \leq 8\sigma_u^4\{\log(d_1+1)+2\}^2$ through the moment bound of Lemma 1.2. By Hölder's inequality with $|\mathcal{J}| \asymp N$,

$$E\left[\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty^2\right] = O\left(\frac{\sqrt{E[\Delta_3^4]E\|\bar{\mathbf{V}}_i\|_\infty^4}\log(d_1)}{N}\right) = O\left(\frac{\sqrt{E[\Delta_3^4]}\log^2(d_1)}{N}\right).$$

Hence, we also have

$$E\left[\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty^2\right] = O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a}\log(d)\log^2(d_1)}{N^2}\right) = O_p\left(\frac{\sigma^2\log(d_1)}{N}\right),$$

since $s_{\boldsymbol{\delta}_a}\log(d_1)\log(d) = O(N)$. Together with (1.75), we conclude that $E[\||\mathcal{J}_2|^{-1}\sum_{i\in\mathcal{J}_2}\Delta_{3i}$ $\bar{\mathbf{V}}_i\|_\infty^2] = O_p(\sigma\sqrt{\log(d_1)/N})$ when $\nu_a^*(\cdot) = \nu_a(\cdot)$ and either (a) $\|\mathbf{S}_1\|_\infty \leq C$ almost surely or (b) $s_{\boldsymbol{\delta}_a}\log(d_1)\log(d) = O(N)$. By Markov's inequality, we have

$$\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{3i}\bar{\mathbf{V}}_i\right\|_\infty = O_p\left(\sigma\sqrt{\frac{\log(d_1)}{N}}\right). \tag{1.76}$$

Together with (1.72) and (1.73),

$$\sum_{l=1}^3\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{li}\bar{\mathbf{V}}_i\right\|_\infty = O_p\left(\sigma\sqrt{\frac{\log(d_1)}{N}}\right).$$

That is, for any $t > 0$, there exists some $\lambda_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\log(d_1)/N}$ such that $\mathcal{E}_5$ occurs with probability at least $1-t$, where

$$\mathcal{E}_5 := \left\{\sum_{l=1}^3\left\|\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{li}\bar{\mathbf{V}}_i\right\|_\infty \leq \frac{\lambda_{\boldsymbol{\beta}}}{4}\right\}.$$

Condition on the event $\mathcal{E}_5$. Then, now we have

$$\begin{aligned}
&\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}})^2 + \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{a,2}\|_1 \\
&\quad\leq \frac{\lambda_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \frac{2}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\left(\sum_{l=4}^6\Delta_{li}\right)^2 + \frac{1}{2|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}})^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_a^*\|_1.
\end{aligned}$$

Since $\left(\sum_{l=4}^{6}\Delta_{li}\right)^2 \leq 3\sum_{l=4}^{6}\Delta_{li}^2$, we have

$$\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}(\bar{\mathbf{V}}_i^\top\boldsymbol{\Delta}_{\boldsymbol{\beta}})^2 + 2\lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{a,2}\|_1 \leq \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \frac{12}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\sum_{l=4}^{6}\Delta_{li}^2 + 2\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_a^*\|_1,$$

which reaches (1.59) in the proof of Theorem 1. Repeat the remaining steps therein, when $N \gg s_{\boldsymbol{\beta}_a}\log(d_1)$, with $\lambda_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\log(d_1)/N}$, we have

$$\|\widehat{\boldsymbol{\beta}}_{a,2} - \boldsymbol{\beta}_a^*\|_2 = \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \left(\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\sum_{l=4}^{6}\Delta_{li}^2\right)^{1/2}\right).$$

In the following, we further control the term $|\mathcal{J}_2|^{-1}\sum_{i\in\mathcal{J}_2}\Delta_{li}^2$ for each $l \in \{4,5,6\}$. By Lemmas 1.5 and 1.8 with the Hölder's inequality, we have

$$\begin{aligned}
E\left[\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{4i}^2\right] &= E\left[\Delta_4^2\right] \\
&\leq \left\{E\left[\left\{\frac{1}{g(\mathbf{U}^\top\widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}^\top\boldsymbol{\delta}_a^*)}\right\}\right]^4 E\left[\mathbf{U}^\top(\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\right]^4\right\}^{1/2} \\
&= O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}\log^2(d)}{N^2}\right).
\end{aligned}$$

Under Assumption 4, by Lemma 1.5 with the Hölder's inequality, we have

$$\begin{aligned}
E\left[\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{5i}^2\right] &= E\left[\left[\left\{1 - \frac{\mathbb{1}_{\{A_{2i}=a_2\}}}{g(\mathbf{U}_i^\top\boldsymbol{\delta}_a^*)}\right\}\bar{\mathbf{U}}_i^\top(\widetilde{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\right]^2 \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\right] \\
&= O_p\left(\frac{\sigma^2 s_{\boldsymbol{\alpha}_a}\log(d)}{N}\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\right).
\end{aligned}$$

Under Assumption 2, by Lemma 1.8 with the Hölder's inequality, we have

$$\begin{aligned}
E\left[\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\Delta_{6i}^2\right] &= E\left[\left[\left\{\frac{1}{g(\mathbf{U}_i^\top\widetilde{\boldsymbol{\delta}}_a)} - \frac{1}{g(\mathbf{U}_i^\top\boldsymbol{\delta}_a^*)}\right\}(\widetilde{Y}_i - \widetilde{\mathbf{U}}_i^\top\boldsymbol{\alpha}_a^*)\right]^2 \mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right] \\
&= O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a}\log(d)}{N}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right).
\end{aligned}$$

By Markov's inequality,

$$\frac{1}{|\mathcal{J}_2|}\sum_{i\in\mathcal{J}_2}\sum_{l=4}^{6}\Delta_{li}^2$$
$$= O_p\left(\frac{\sigma^2 s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}\log^2 d}{N^2} + \frac{\sigma^2 s_{\boldsymbol{\alpha}_a}\log(d)}{N}\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}} + \frac{\sigma^2 s_{\boldsymbol{\delta}_a}\log(d)}{N}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right).$$

Therefore, we have

$$\|\widehat{\boldsymbol{\beta}}_{a,2} - \boldsymbol{\beta}_a^*\|_2 = O_p(r_n),$$

with $r_n = \sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \frac{\sigma\sqrt{s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}}\log(d)}{N} + \sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\mathbb{1}_{\{\rho_a^*\neq\rho_a\}} + \sigma\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\mathbb{1}_{\{\nu_a^*\neq\nu_a\}}$. Similarly, we consider the DR-imputation-based estimator $\widehat{\boldsymbol{\beta}}_{a,1} = \widehat{\boldsymbol{\beta}}_a(\mathcal{D}_{\mathcal{J}_1}, \widehat{Y}_{\mathcal{J}_1}^{\mathrm{DR}})$ defined as (2.11). In this case, the expectations are taken w.r.t. the samples in $\mathcal{D}_{\mathcal{J}_1}$; $\widehat{\boldsymbol{\delta}}_a(\mathcal{D}_{\mathcal{J}_2})$ and $\widehat{\boldsymbol{\alpha}}_a(\mathcal{D}_{\mathcal{J}_2})$ are fitted using samples in $\mathcal{D}_{\mathcal{J}_2}$ and are treated as fixed or condition on. We also have the same consistency rate for $\widehat{\boldsymbol{\beta}}_{a,1}$. Therefore, for $\widehat{\boldsymbol{\beta}}_a = (\widehat{\boldsymbol{\beta}}_{a,1} + \widehat{\boldsymbol{\beta}}_{a,2})/2$, we have

$$\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2 = O_p(r_n).$$

By Lemma 1.4, $\|\mathbf{V}^\top(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)\|_{\psi_2} \le 2\sigma_u\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2$. By Lemma 1.2, for any $r \ge 1$,

$$\{E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^r\}^{1/r} = \left\{E[\mathbf{V}^\top(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*)]^r\right\}^{1/r} = O(\|\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a^*\|_2) = O_p(r_n).$$

■

Nested-regression-based estimator

Proof of Theorem 10. Let $\widehat{Y} = \bar{\mathbf{U}}^\top\widehat{\boldsymbol{\alpha}}_a$, $Y^* = \bar{\mathbf{U}}^\top\boldsymbol{\alpha}_a^*$, $\mathbf{X} = \bar{\mathbf{V}}$, $\mathbb{S} = (\bar{\mathbf{V}}_i)_{i\in J}$, $M = |\mathcal{J}|$, and $\delta_M^2 = 288\sigma_u\kappa_1^{-2}\lambda_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}$. For any $t > 0$, let $\lambda_{\boldsymbol{\alpha}} := 32\sigma\sigma_u\sigma_\zeta(\sqrt{\log(d+1)/|\mathcal{J}|} + t)$ and $\lambda_{\boldsymbol{\beta}} := 32\sigma\sigma_u\sigma_\varepsilon(\sqrt{\log(d_1+1)/|\mathcal{J}|} + t)$. Suppose that $|\mathcal{J}| \ge \max\{\log(d+1), (c_3 + 100\kappa_2^2)s_{\boldsymbol{\alpha}_a}\log(d+$

$1), 100\kappa_2^2 s_{\boldsymbol{\beta}_a} \log(d_1+1)\}$. Now, for the event $\mathcal{E}_1 := \{|\mathcal{J}|^{-1} \sum_{i\in\mathcal{J}} [\widehat{Y}_i - Y_i^*]^2 < \delta_M^2\}$, by Lemma 1.6, we have

$$P(\mathcal{E}_1) \geq 1 - 2\exp\left(-\frac{4|\mathcal{J}|t^2}{1+2t+\sqrt{2t}}\right) - c_1\exp(-c_2|\mathcal{J}|) - 2\exp(-c_4|\mathcal{J}|).$$

By Lemma 1.4, $\lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) \geq \kappa_l$ and $\bar{\mathbf{V}}$ is sub-Gaussian with $\|\boldsymbol{x}^\top \bar{\mathbf{V}}\|_{\psi_2} \leq 2\sigma_u \|\boldsymbol{x}\|_2$, for any $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$. Additionally, under Assumption 2, $\|\varepsilon\|_{\psi_2} \leq \sigma\sigma_\varepsilon$. Here, κ_l, σ_u, σ_ε, and σ, defined in Assumptions 2 and 3, are positive constants independent of N and d. Hence, the estimation rates of $\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}}$ in Theorem 10 follow from Theorem 8. To show the esitmation rate of $\widehat{\mu}_{a,\mathrm{NR}}(\cdot)$, by Lemma 1.2, for any $r \geq 1$,

$$\begin{aligned}\{E[\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu^*_{a,\mathrm{NR}}(\mathbf{S}_1)]^r\}^{1/r} &= \{E[\mathbf{V}^\top(\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} - \boldsymbol{\beta}^*_{a,\mathrm{NR}})]^r\}^{1/r} \\ &= O(\|\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} - \boldsymbol{\beta}^*_{a,\mathrm{NR}}\|_2) = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a}\log(d_1)}{N}} + \sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\right),\end{aligned}$$

since $\|\mathbf{V}^\top(\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} - \boldsymbol{\beta}^*_{a,\mathrm{NR}})\|_{\psi_2} \leq 2\sigma_u \|\widehat{\boldsymbol{\beta}}_{a,\mathrm{NR}} - \boldsymbol{\beta}^*_{a,\mathrm{NR}}\|_2$ by Lemma 1.4. Here, the expectation is only taken w.r.t. the distribution of the new observation $\mathbf{S}_1$. ∎

Proofs of Auxiliary Lemmas

Proof of Lemma 1.3. 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 1.4. (a) Observe that

$$\begin{aligned}
\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top \mathbb{1}_{\{A_1=a_1,A_2=a_2\}}]\boldsymbol{x} \\
&\overset{(i)}{=} \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[E[(\mathbf{U}^\top\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1,A_2=a_2\}}|\mathbf{U}, A_1=a_1]P[A_1=a_1|\mathbf{U}]] \\
&= \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^\top\boldsymbol{x})^2 \cdot P[A_2=a_2|\mathbf{U}, A_1=a_1]E[\mathbb{1}_{\{A_1=a_1\}}|\mathbf{U}]] \\
&\overset{(ii)}{=} \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^\top\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1\}} \cdot P[A_2=a_2|\mathbf{U}, A_1=a_1]], \qquad (1.77)
\end{aligned}$$

where (i) and (ii) hold by the tower rule. Under the overlap conditions of Assumption 1,

$$P(c_0 \le P[A_2=a_2|\mathbf{U}, A_1=a_1] \le 1-c_0) = 1.$$

Together with (1.77), under Assumption 3, we have

$$\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) \ge c_0 \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^\top\boldsymbol{x})^2 \mathbb{1}_{\{A_1=a_1\}}] \ge c_0\kappa_l > 0.$$

Additionally, we also have

$$\begin{aligned}
\lambda_{\max}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top \mathbb{1}_{\{A_1=a_1,A_2=a_2\}}]\boldsymbol{x} \\
&\le \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top]\boldsymbol{x} = \lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) \overset{(i)}{\le} 2\sigma_u^2,
\end{aligned}$$

where (i) holds since, by Assumption 3 and Lemma 1.2,

$$\lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) = \max_{\|\boldsymbol{x}\|_2=1} E[(\boldsymbol{x}^\top\mathbf{U})^2] \le \max_{\|\boldsymbol{x}\|_2=1} 2\sigma_u^2\|\boldsymbol{x}\|_2^2 = 2\sigma_u^2 < \infty. \qquad (1.78)$$

Besides, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^\top\widetilde{\mathbf{U}}|^{2k}] = E[|\boldsymbol{x}^\top\mathbf{U}|^{2k}\mathbb{1}_{\{A_1=a_1,A_2=a_2\}}] \le E[|\boldsymbol{x}^\top\mathbf{U}|^{2k}] \overset{(i)}{\le} 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1),$$

where (i) holds by Assumption 3 and Lemma 1.2. By Lemma 1.3, we have

$$\|\boldsymbol{x}^\top\widetilde{\mathbf{U}}\|_{\psi_2} \le 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d+1}.$$

(b) Under Assumption 3, we also have

$$\lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) = \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[(\mathbf{U}^\top\boldsymbol{x})^2\mathbb{1}_{\{A_1=a_1\}}] \geq \kappa_l > 0, \tag{1.79}$$

and by (1.78),

$$\begin{aligned}\lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top\mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &\leq \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top]\boldsymbol{x} \leq 2\sigma_u^2 < \infty.\end{aligned} \tag{1.80}$$

In addition, for any $\boldsymbol{x}\in\mathbb{R}^{d+1}$ and $k\in\mathbb{N}$,

$$E[|\boldsymbol{x}^\top\bar{\mathbf{U}}|^{2k}] = E[|\boldsymbol{x}^\top\mathbf{U}|^{2k}\mathbb{1}_{\{A_1=a_1\}}] \leq E[|\boldsymbol{x}^\top\mathbf{U}|^{2k}] \overset{(i)}{\leq} 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1), \tag{1.81}$$

where (i) holds by Assumption 3 and Lemma 1.2. By Lemma 1.3, we have

$$\|\boldsymbol{x}^\top\bar{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x}\in\mathbb{R}^{d+1}.$$

(c) Note that

$$\begin{aligned}\lambda_{\min}(E[\mathbf{U}\mathbf{U}^\top]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top]\boldsymbol{x} \\ &\geq \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top\mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \overset{(i)}{\geq} \kappa_l > 0,\end{aligned}$$

where (i) holds by (1.79). By (1.78), we know $\lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) \leq 2\sigma_u^2 < \infty$. By Assumption 3, we have

$$\|\boldsymbol{x}^\top\mathbf{U}\|_{\psi_2} \leq \sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x}\in\mathbb{R}^{d+1}.$$

(d) Recall the representation (1.38), we also have

$$\begin{aligned}\lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{V}\mathbf{V}^\top\mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &= \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{Q}\mathbf{U}\mathbf{U}^\top\mathbf{Q}^\top\mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &\overset{(i)}{\geq} \min_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top\mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \overset{(ii)}{\geq} \kappa_l > 0,\end{aligned} \tag{1.82}$$

where (i) holds since, for every $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$, $\mathbf{Q}^\top \boldsymbol{x} = (\boldsymbol{x}^\top, 0, \ldots, 0)^\top \in \mathbb{R}^{d+1}$ and hence $\|\mathbf{Q}^\top \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 = 1$; (ii) follows from (1.79). Similarly,

$$\begin{aligned}\lambda_{\max}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{V}\mathbf{V}^\top \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{Q}\mathbf{U}\mathbf{U}^\top\mathbf{Q}^\top \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} \\ &\leq \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \overset{(i)}{\leq} 2\sigma_u^2 < \infty,\end{aligned}$$

where (i) follows from (1.80). In addition, for any $k \in \mathbb{N}$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \bar{\mathbf{V}}|^{2k}] &= \sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \mathbf{Q}\bar{\mathbf{U}}|^{2k}] \\ &\overset{(i)}{\leq} \sup_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \bar{\mathbf{U}}|^{2k}] \overset{(ii)}{\leq} 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1),\end{aligned}$$

where (i) holds since, for every $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$, $\|\mathbf{Q}^\top \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 = 1$; (ii) follows from (1.81). Hence, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^\top \bar{\mathbf{V}}|^{2k}] \leq 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1).$$

By Lemma 1.3, we have $\bar{\mathbf{V}}$ is sub-Gaussian with

$$\|\boldsymbol{x}^\top \bar{\mathbf{V}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d_1+1}.$$

(e) Lastly, note that

$$\begin{aligned}\lambda_{\min}(E[\mathbf{V}\mathbf{V}^\top]) &= \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{V}\mathbf{V}^\top]\boldsymbol{x} \\ &\geq \min_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{V}\mathbf{V}^\top \mathbb{1}_{\{A_1=a_1\}}]\boldsymbol{x} = \lambda_{\min}(E[\bar{\mathbf{V}}\bar{\mathbf{V}}^\top]) \overset{(i)}{\geq} \kappa_l > 0,\end{aligned}$$

where (i) holds by (1.82). Besides,

$$\begin{aligned}\lambda_{\max}(E[\mathbf{V}\mathbf{V}^\top]) &= \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{V}\mathbf{V}^\top]\boldsymbol{x} = \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{Q}\mathbf{U}\mathbf{U}^\top\mathbf{Q}^\top]\boldsymbol{x} \\ &\leq \max_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top E[\mathbf{U}\mathbf{U}^\top]\boldsymbol{x} = \lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) \overset{(i)}{\leq} 2\sigma_u^2 < \infty,\end{aligned}$$

where (i) follows from (1.80). In addition, for any $k \in \mathbb{N}$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \mathbf{V}|^{2k}] &= \sup_{\boldsymbol{x}\in\mathbb{R}^{d_1+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \mathbf{Q}\mathbf{U}|^{2k}] \\ &\overset{(i)}{\leq} \sup_{\boldsymbol{x}\in\mathbb{R}^{d+1}:\|\boldsymbol{x}\|_2=1} E[|\boldsymbol{x}^\top \mathbf{U}|^{2k}] \overset{(ii)}{\leq} 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1),\end{aligned}$$

where (i) holds since, for every $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x} \in \mathbb{R}^{d_1+1}$, $\|\mathbf{Q}^\top \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 = 1$; (ii) follows from (1.81). Hence, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$ and $k \in \mathbb{N}$,

$$E[|\boldsymbol{x}^\top \mathbf{V}|^{2k}] \leq 2(\sigma_u\|\boldsymbol{x}\|_2)^{2k}\Gamma(k+1).$$

By Lemma 1.3, we have $\mathbf{V}$ is also sub-Gaussian with

$$\|\boldsymbol{x}^\top \mathbf{V}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2, \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^{d_1+1}.$$

■

Proof of Lemma 1.5. Now, we consider the Lasso estimator $\widehat{\boldsymbol{\alpha}}_a$ defined as (2.6), which is constructed using the outcome $\widetilde{Y}$, covariates $\widetilde{\mathbf{U}}$ and training samples $\mathcal{D}_{\mathcal{J}}$. Note that $\widehat{\boldsymbol{\alpha}}_a$ is a special case of $\widehat{\boldsymbol{\beta}}$, (4.1). Let $\widehat{Y} = Y^* = \widetilde{Y}$, $\mathbf{X} = \widetilde{\mathbf{U}}$, $\mathbb{S} = (\mathbf{X}_i)_{i\in J}$, $M = |\mathcal{J}|$, and $\delta_M = 0$. By Lemma 1.4, $\lambda_{\min}(E[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top]) \geq c_0\kappa_l$ and $\widetilde{\mathbf{U}}$ is sub-Gaussian with $\|\boldsymbol{x}^\top\widetilde{\mathbf{U}}\|_{\psi_2} \leq 2\sigma_u\|\boldsymbol{x}\|_2$, for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$. Additionally, under Assumption 2, $\|\zeta\|_{\psi_2} \leq \sigma\sigma_\zeta$. Here, c_0, κ_l, σ_u, σ_ζ, and σ, defined in Assumptions 1, 2, and 3, are positive constants independent of N and d. Hence, the estimation rates of $\widehat{\boldsymbol{\alpha}}_a$ in Lemma 1.5 follows from Theorem 8. To show the estimation rate of $\widehat{\nu}_a(\cdot)$, by Lemma 1.2, for any $r \geq 1$,

$$\begin{aligned}\{E[\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})]^r\}^{1/r} &= \{E[\mathbf{U}^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^r\}^{1/r} \\ &= O(\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2) = O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a}\log(d)}{N}}\right),\end{aligned}$$

since $\|\mathbf{U}^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\|_{\psi_2} \le \sigma_u \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2$ under Assumption 3. Here, the expectation is only taken w.r.t. the joint distribution of the new observations $\mathbf{S}$. ∎

Proof of Lemma 1.6. Let $\widehat{Y} = Y^* = \widetilde{Y}$, $\mathbf{X} = \widetilde{\mathbf{U}}$, $\mathbb{S} = (\mathbf{X}_i)_{i\in J}$, $M = |\mathcal{J}|$, and $\delta_M = 0$. Following the proof of Theorem 8, since $\delta_M = 0$, we have $\|\boldsymbol{\delta}_S\|_1 \ge 4\lambda^{-1}\delta_M^2$. That is, we are under Case 2. Hence, $\boldsymbol{\delta}$ is in the cone set as in (1.66). By Lemma 1.4, $\|\boldsymbol{a}^\top \bar{\mathbf{U}}\|_{\psi_2} \le 2\sigma_u \|\boldsymbol{a}\|_2$ for any $\boldsymbol{a} \in \mathbb{R}^{d+1}$ and $\lambda_{\min}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \ge \kappa_l$. Here, σ_u and κ_l, defined in Assumption 3, are positive constants independent of N and d. By Theorem 15 of [RZ12], with some constants $c_3, c_4 > 0$, when $|\mathcal{J}| \ge c_3 s_{\boldsymbol{\alpha}_a} \log(d+1)$,

$$\frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \left\{\bar{\mathbf{U}}_i^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)\right\}^2 \le 1.5^2 \lambda_{\max}(E[\bar{\mathbf{U}}\bar{\mathbf{U}}^\top]) \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2^2 \le 4.5\sigma_u \|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2^2,$$

with probability at least $1 - 2\exp(-c_4|\mathcal{J}|)$. In addition, by Lemma 1.5, we have

$$\|\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*\|_2 \le 8\kappa_1^{-1}\lambda_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\alpha}_a}},$$

with probability at least $1 - 2\exp(-\frac{4|\mathcal{J}|t^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2|\mathcal{J}|)$. Therefore, with probability at least $1 - 2\exp(-\frac{4|\mathcal{J}|t^2}{1+2t+\sqrt{2t}}) - c_1\exp(-c_2|\mathcal{J}|) - 2\exp(-c_4|\mathcal{J}|)$,

$$\frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} [\bar{\mathbf{U}}_i^\top(\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a^*)]^2 \le 288\sigma_u\kappa_1^{-2}\lambda_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\alpha}_a}.$$

∎

Proof of Lemma 1.7. In this Lemma, we provide estimation rates for $\widehat{\boldsymbol{\gamma}}_a$, $\widehat{\pi}_a(\cdot)$, $\widehat{\boldsymbol{\delta}}_a$, and $\widehat{\rho}_a(\cdot)$. We allow model misspecifications that $\pi_a^*(\cdot) \ne \pi_a(\cdot)$ and $\rho_a^*(\cdot) \ne \rho_a(\cdot)$. 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 (1.48) and (1.49). In part (a), the expectations are only taken w.r.t. the distribution of the new observation $\mathbf{S}_1$.

Consider the link function $\phi(u) = \log(1 + \exp(u))$, we have

$$\phi''(\mathbf{V}^\top \boldsymbol{\gamma}_a^*) = \frac{\exp(\mathbf{V}^\top \boldsymbol{\gamma}_a^*)}{(1 + \exp(\mathbf{V}^\top \boldsymbol{\gamma}_a^*))^2} = \pi_a^*(\mathbf{S}_1)(1 - \pi_a^*(\mathbf{S}_1)).$$

Under Assumption 4, we have $P(c_0^2 \leq \phi''(\mathbf{V}^\top \boldsymbol{\gamma}_a^*) \leq (1 - c_0)^2) = 1$. By Lemma 1.4,

$$\lambda_{\min}(E[\mathbf{V}\mathbf{V}^\top]) \geq \kappa_l > 0, \quad \lambda_{\max}(E[\mathbf{V}\mathbf{V}^\top]) \leq 2\sigma_u^2 < \infty, \tag{1.83}$$

and $\mathbf{V}$ is sub-Gaussian with $\|\boldsymbol{x}^\top \mathbf{V}\|_{\psi_2} \leq 2\sigma_u \|\boldsymbol{x}\|_2$ 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 $\pi_a^*(\cdot)$ not necessarily equal to $\pi_a(\cdot)$; The true $\boldsymbol{\gamma}_a$ may not exists such that $\widehat{\pi}_a(\cdot)$ 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 (1.85) 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 $\delta\ell_{\mathcal{J}}(\boldsymbol{\Delta}, \boldsymbol{\gamma}_a^*)$ defined in (1.84) can be written as

$$\delta\ell_{\mathcal{J}}(\boldsymbol{\Delta}, \boldsymbol{\gamma}_a^*) = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \left[\phi(\mathbf{V}_i^\top(\boldsymbol{\gamma}_a^* + \boldsymbol{\Delta})) - \phi(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \boldsymbol{\Delta}^\top \mathbf{V}_i \phi'(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*)\right],$$

which is function of $\mathbf{S}_{1i}$s, and A_{1i}s are not involved above. As a result, the model misspecification for $\pi_a(\mathbf{S}_1) = E(A_1|\mathbf{S}_1)$ does not affect the RSC property. In below, we consider the RSC property studied by [ZCB21]. For any $\boldsymbol{\gamma}_a, \boldsymbol{\Delta} \in \mathbb{R}^{d_1+1}$, define

$$\ell_{\mathcal{J}}(\boldsymbol{\gamma}_a) := \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \left[\phi(\mathbf{V}_i^\top \boldsymbol{\gamma}_{\boldsymbol{a}}) - \mathbb{1}_{\{A_{1i}=a_1\}} \mathbf{V}_i^\top \boldsymbol{\gamma}_a\right],$$

$$\delta\ell_{\mathcal{J}}(\boldsymbol{\Delta}, \boldsymbol{\gamma}_a^*) := \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^* + \boldsymbol{\Delta}) - \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*) - \boldsymbol{\Delta}^\top \nabla \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*). \tag{1.84}$$

By Lemma 4.5 of [ZCB21], we have the following RSC property holds:

$$\begin{aligned}\delta\ell_{\mathcal{J}}(\boldsymbol{\Delta}, \boldsymbol{\gamma}_a^*) &\geq \kappa_1 \|\boldsymbol{\Delta}\|_2 \left\{ \|\boldsymbol{\Delta}\|_2 - \kappa_2 \sqrt{\frac{\log(d_1+1)}{|\mathcal{J}|}} \|\boldsymbol{\Delta}\|_1 \right\} \\ &\geq \frac{\kappa_1}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{\kappa_1 \kappa_2^2 \log(d_1+1)}{2|\mathcal{J}|}\|\boldsymbol{\Delta}\|_1^2 \quad \text{for all } \|\boldsymbol{\Delta}\|_2 \leq 1, \end{aligned} \tag{1.85}$$

with probability at least $1 - c_1 \exp(-c_2|\mathcal{J}|)$, where $c_1, c_2, \kappa_1, \kappa_2 > 0$ are some constants.

Additionally, the gradient $\|\nabla \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*)\|_\infty$ is controlled in the following. We allow a possibly misspecified model that $\pi_a^*(\cdot) \neq \pi_a(\cdot)$. Note that, even under model misspecification, we still have (1.87) below. Hence, $\|\nabla \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*)\|_\infty$ is the maximum of zero-mean random variables. By the union bound, we have

$$\begin{aligned} P\left(\|\nabla \ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*)\|_\infty \geq \frac{\lambda_\gamma}{2} \right) &= P\left(\max_{1 \leq j \leq d_1+1} \left| \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} (g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_{1i}=a_1\}}) \mathbf{V}_{i,j} \right| \geq \frac{\lambda_\gamma}{2} \right) \\ &\leq \sum_{j=1}^{d_1+1} P\left(\left| \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} (g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_{1i}=a_1\}}) \mathbf{V}_{i,j} \right| \geq \frac{\lambda_\gamma}{2} \right), \end{aligned} \tag{1.86}$$

where $g(u) = \exp(u)/\{1+\exp(u)\}$ is the logistic function. By definition, $\boldsymbol{\gamma}_a^* = \arg\min_{\boldsymbol{\gamma}_a \in \mathbb{R}^{d_1+1}} E[\ell(\boldsymbol{\gamma}_a)]$, where for any $\boldsymbol{\gamma}_a \in \mathbb{R}^{d_1+1}$,

$$\ell(\boldsymbol{\gamma}_a) := E\left[\phi(\mathbf{V}^\top \boldsymbol{\gamma}_a) - \mathbb{1}_{\{A_1=a_1\}} \mathbf{V}^\top \boldsymbol{\gamma}_a \right].$$

By the first-order optimality condition, we know that

$$\nabla E[\ell(\gamma_a^*)] = E\left[(g(\mathbf{V}^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_1=a_1\}}) \mathbf{V} \right] = \mathbf{0} \in \mathbb{R}^{d_1+1}. \tag{1.87}$$

Additionally, since $|g(\mathbf{V}^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_1=a_1\}}| \leq 1$, by Lemma 1.2 and under Assumption 3, for any $i \in \mathcal{J}$ and $j \leq d_1 + 1$,

$$\|(g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_{1i}=a_1\}}) \mathbf{V}_{i,j}\|_{\psi_2} \leq \|\mathbf{V}_{i,j}\|_{\psi_2} \leq \sigma_u.$$

That is, $(g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_{1i}=a_1\}})\mathbf{V}_{i,j}$ is a zero-mean sub-Gaussian random variable. Hence, by Hoeffding's inequality, for each $j \le d_1 + 1$,

$$P\left(\left|\frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}(g(\mathbf{V}_i^\top \boldsymbol{\gamma}_a^*) - \mathbb{1}_{\{A_{1i}=a_1\}})\mathbf{V}_{i,j}\right| \ge \frac{\lambda_\gamma}{2}\right) \le 2\exp\left(\frac{-|\mathcal{J}|\lambda_\gamma^2}{32\sigma_u^2}\right)$$
$$\le 2\exp\big(-\log(d_1+1) - |\mathcal{J}|t^2\big) = \frac{2\exp(-|\mathcal{J}|t^2)}{d_1+1},$$

where for any $t > 0$, we set $\lambda_\gamma := 4\sqrt{2}\sigma_u(\sqrt{\frac{\log(d_1+1)}{|\mathcal{J}|}} + t)$. Together with (1.86), it follows that

$$P\left(\|\ell_{\mathcal{J}}(\boldsymbol{\gamma}_a^*)\|_\infty \le \frac{\lambda_\gamma}{2}\right) \le 1 - 2\exp(-|\mathcal{J}|t^2).$$

Together with (1.85), when $|\mathcal{J}| \ge 64\kappa_2^2 s_{\boldsymbol{\gamma}_a}\log(d_1+1)$ and $9s_{\boldsymbol{\gamma}_a}\lambda_\gamma^2 \le \kappa_1^2$, by Corollary 9.20 of [Wai19], we conclude that

$$\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2 \le \frac{3\sqrt{s_{\boldsymbol{\gamma}_a}}\lambda_\gamma}{\kappa_1}, \quad \|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_1 \le \frac{6s_{\boldsymbol{\gamma}_a}\lambda_\gamma}{\kappa_1},$$

with probability at least $1 - 2\exp(-|\mathcal{J}|t^2) - c_1\exp(-c_2|\mathcal{J}|)$. Hence, when $|\mathcal{J}| \asymp N$ and $N \gg s_{\boldsymbol{\gamma}_a}\log(d_1)$, with some $\lambda_{\mathcal{J}} \asymp \sqrt{\frac{\log(d_1)}{|\mathcal{J}|}}$,

$$\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}_a}\log(d_1)}{N}\right). \tag{1.88}$$

Now, we show the estimation rate for $\widehat{\pi}_a(\cdot)$. In the following, we will use Taylor's Theorem to control the estimation error of $\widehat{\pi}_a(\cdot)$ by the estimation error of $\widehat{\boldsymbol{\gamma}}_a$ as in (1.90). Then, we apply the estimation rate (1.88) proved above to obtain the rate for $\widehat{\pi}_a(\cdot)$.

Recall that $g(u) := \exp(u)/\{1+\exp(u)\} = \phi'(u)$ for any $u \in \mathbb{R}$. Note that, for any $u^*, \Delta \in \mathbb{R}$,

$$\frac{d(g(u^*+t\Delta) - g(u^*))^2}{dt} = 2(g(u^*+t\Delta) - g(u^*))g'(u^*+t\Delta)\Delta,$$
$$\frac{d^2(g(u^*+t\Delta) - g(u^*))^2}{dt^2} = 2(g'(u^*+t\Delta))^2\Delta^2 + 2(g(u^*+t\Delta) - g(u^*))g''(u^*+t\Delta)\Delta^2,$$

where, for any $u \in \mathbb{R}$, since $g(u) \in (0,1)$, we have

$$g'(u) = g(u)(1-g(u)) \in (0,1), \quad g''(u) = g(u)(1-g(u))(1-2g(u)) \in (-1,1). \tag{1.89}$$

Set $u^* = \mathbf{V}^\top \boldsymbol{\gamma}_a^*$ and $\Delta = \mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)$. By Taylor's Theorem, with some $\widetilde{t} \in (0,1)$,

$$\begin{aligned}
E[g(\mathbf{V}^\top\widehat{\boldsymbol{\gamma}}_a) - g(\mathbf{V}^\top\boldsymbol{\gamma}_a^*)]^2 &= E[g(u^* + 1\cdot\Delta) - g(u^*)]^2 \\
&= E[g(u^* + 0\cdot\Delta) - g(u^*)]^2 + \left.\frac{dE(g(u^*+t\Delta)-g(u^*))^2}{dt}\right|_{t=0}\cdot 1 \\
&\quad + \left.\frac{d^2E(g(u^*+t\Delta)-g(u^*))^2}{2dt^2}\right|_{t=\widetilde{t}}\cdot 1^2 \\
&= 0 + E\left[2(g(u^*+0\cdot\Delta)-g(u^*))g'(u^*+0\cdot\Delta)\Delta\right] \\
&\quad + E\left[(g'(u^*+\widetilde{t}\Delta))^2\Delta^2 + (g(u^*+\widetilde{t}\Delta)-g(u^*))g''(u^*+\widetilde{t}\Delta)\Delta^2\right] \\
&= E\left[(g'(u^*+\widetilde{t}\Delta))^2\Delta^2 + (g(u^*+\widetilde{t}\Delta)-g(u^*))g''(u^*+\widetilde{t}\Delta)\Delta^2\right] \\
&\overset{(i)}{\leq} 2E[\Delta^2] = 2E[\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)]^2,
\end{aligned}$$

where (i) holds since, by (1.89), $(g'(u^*+\widetilde{t}\Delta))^2 \leq 1$ and $(g(u^*+\widetilde{t}\Delta)-g(u^*))g''(u^*+\widetilde{t}\Delta) \leq 1$. Hence,

$$E[\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1)]^2 = E[g(\mathbf{V}^\top\widehat{\boldsymbol{\gamma}}_a) - g(\mathbf{V}^\top\boldsymbol{\gamma}_a^*)]^2 \leq 2E[\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)]^2. \tag{1.90}$$

Then, from (1.83) and (1.88), we have

$$E[\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1)]^2 \leq 2\|E[\mathbf{V}\mathbf{V}^\top]\|_2\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}_a}\log(d_1)}{N}\right). \tag{1.91}$$

(b) Now, we show (1.50) and (1.51). In part (b), the expectations are only taken w.r.t. the distribution of the new observations $\mathbf{S}$.

By Lemma 1.4, we know that the minimum and maximum eigenvalues of covariance matrix $E[\mathbf{U}\mathbf{U}^\top]$ satisfy

$$\lambda_{\min}(E[\mathbf{U}\mathbf{U}^\top]) \geq \kappa_l > 0, \quad \lambda_{\max}(E[\mathbf{U}\mathbf{U}^\top]) \leq 2\sigma_u^2 < \infty,$$

and $\mathbf{U}$ is sub-Gaussian with $\|\boldsymbol{x}^\top \mathbf{U}\|_{\psi_2} \leq \sigma_u \|\boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathbb{R}^{d+1}$. Additionally, we also have $P(c_0^2 \leq \phi''(\mathbf{U}^\top \boldsymbol{\delta}_a) \leq (1-c_0)^2) = 1$ under Assumption 4. 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

$$\begin{aligned} E[\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S})]^2 &= E[g(\mathbf{U}^\top \widehat{\boldsymbol{\delta}}_a) - g(\mathbf{U}^\top \boldsymbol{\delta}_a^*)]^2 \leq 2E[\mathbf{U}^\top(\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*)]^2 \\ &\leq 2\|E[\mathbf{U}\mathbf{U}^\top]\|_2 \|\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta}_a^*\|_2^2 = O_p\left(\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}\right). \end{aligned}$$

■

Proof of Lemma 1.8. In this proof, the expectations are only taken w.r.t. the distribution of the new observations $\mathbf{S}$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). By Lemma 1.7, we have $P(\mathcal{A}) = 1 - o(1)$. By Minkowski's inequality, we have

$$\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}} = \{E|1 + \exp(-\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_a)|^r\}^{\frac{1}{r}} \leq 1 + \{E|\exp(-\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_a)|^r\}^{\frac{1}{r}}.$$

Under Assumption 4, we know that

$$P\left(\frac{c_0}{1-c_0} \leq \exp(-\mathbf{V}^\top \boldsymbol{\gamma}_a^*) \leq \frac{1-c_0}{c_0}\right) = 1. \tag{1.92}$$

which implies that

$$\begin{aligned} \{E|\exp(-\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_a)|^r\}^{\frac{1}{r}} &= \{E|\exp(-\mathbf{V}^\top \boldsymbol{\gamma}_a^*)\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|^r\}^{\frac{1}{r}} \\ &\leq \frac{1-c_0}{c_0}\{E|\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|^r\}^{\frac{1}{r}}. \end{aligned}$$

Hence, to prove $\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly, i.e., bounded by a constant independent of N, it suffices to show $\{E|\exp(-r\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|\}^{\frac{1}{r}}$ is bounded uniformly.

Let $\mu = E[|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)|]$. By Lemma 1.4, we have

$$\|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a)\|_{\psi_2} \leq 2\sigma_u \|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a\|_2. \tag{1.93}$$

Now, condition on the event $\mathcal{A}$, we have

$$\mu \leq \sqrt{\pi_a}\sigma_u, \quad \|\mu\|_{\psi_2} \leq (\log 2)^{-1/2}\sqrt{\pi_a}\sigma_u, \tag{1.94}$$

which follows from Lemma 1.2. 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

$$\||\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)| - \mu\|_{\psi_2} \leq \|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)\|_{\psi_2} + \|\mu\|_{\psi_2} \leq [1 + (\log 2)^{-1/2}\sqrt{\pi_a}]\sigma_u,$$

and with it that the moment generating function can be bounded with

$$E[\exp\{r(|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)| - \mu)\}] \leq \exp\{2r^2[1 + (\log 2)^{-1/2}\sqrt{\pi_a}]^2\sigma_u^2\}.$$

Using (1.94), we get that

$$\begin{aligned}\{E|\exp(-r\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|\}^{\frac{1}{r}} &\leq \{E|\exp(r|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)|)|\}^{\frac{1}{r}} \\ &\leq \exp\{\sqrt{\pi_a}\sigma_u + 2r[1 + (\log 2)^{-1/2}\sqrt{\pi_a}]^2\sigma_u^2\},\end{aligned} \tag{1.95}$$

which is bounded and hence $\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly. Repeating the same procedure above, we can obtain that $\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}$ is also bounded uniformly, which will be used later on in the proof. By (1.92), we have

$$\begin{aligned}\left\{E\left|\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)}\right|^r\right\}^{\frac{1}{r}} &= \{E|\exp(-\mathbf{V}^\top\boldsymbol{\gamma}_a^*)[\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)) - 1]|^r\}^{\frac{1}{r}} \\ &\leq \frac{1-c_0}{c_0}\{E|\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)) - 1|^r\}^{\frac{1}{r}}.\end{aligned} \tag{1.96}$$

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 $t \in (0,1)$

$$\begin{aligned} |\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)) - 1| &= \exp(-t\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)| \\ &\overset{(i)}{\leq} [1 + \exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))]|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)|, \end{aligned} \tag{1.97}$$

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 (1.96) and (1.97), we have

$$\begin{aligned} \left\{ E \left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^r \right\}^{\frac{1}{r}} &\leq \frac{1-c_0}{c_0} \{E|\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)) - 1|^r\}^{\frac{1}{r}} \\ &\leq \frac{1-c_0}{c_0} \left\{ E \left| [1 + \exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))]\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*) \right|^r \right\}^{\frac{1}{r}} \\ &\overset{(i)}{\leq} \frac{1-c_0}{c_0} \left\{ E \left| \mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*) \right|^r \right\}^{\frac{1}{r}} \\ &\quad + \frac{1-c_0}{c_0} \left\{ E \left| \exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*) \right|^r \right\}^{\frac{1}{r}} \\ &\overset{(ii)}{\leq} \frac{1-c_0}{c_0} \left\{ E \left| \mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*) \right|^r \right\}^{\frac{1}{r}} \\ &\quad + \frac{1-c_0}{c_0} \left\{ E \left| \exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)) \right|^{2r} \right\}^{\frac{1}{2r}} \left\{ E \left| \mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*) \right|^{2r} \right\}^{\frac{1}{2r}}, \end{aligned}$$

where (i) holds by the Minkowski inequality; (ii) holds by Hölder's inequality.

Recall the equation (1.95), we know that $\{E|\exp(-\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*))|^{2r}\}^{\frac{1}{2r}}$ is bounded uniformly. In addition, recall the equation (1.93), by Lemma 1.2, we have

$$\{E|\mathbf{V}^\top(\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*)|^r\}^{\frac{1}{r}} = O(\|\widehat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_a^*\|_2) \overset{(i)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}_a}\log(d_1)}{N}}\right),$$

where (i) holds by Lemma 1.7. Therefore, we obtain that

$$\left\{ E \left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^r \right\}^{\frac{1}{r}} = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}_a}\log(d_1)}{N}}\right). \tag{1.98}$$

Repeating the same procedure, we obtain that $\{E|\widehat{\rho}_a(\mathbf{S})|^{-r}\}^{\frac{1}{r}}$ is bounded uniformly and

$$\left\{E\left|\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right|^r\right\}^{\frac{1}{r}}=O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right). \tag{1.99}$$

Therefore,

$$\begin{aligned}
&\left\{E\left|\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right|^r\right\}^{\frac{1}{r}}\\
&\overset{(i)}{\leq}\left\{E\left|\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)}\left(\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right)\right|^r\right\}^{\frac{1}{r}}+\left\{E\left|\frac{1}{\rho_a^*(\mathbf{S})}\left(\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)}-\frac{1}{\pi_a^*(\mathbf{S}_1)}\right)\right|^r\right\}^{\frac{1}{r}}\\
&\overset{(ii)}{\leq}\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}\left\{E\left|\frac{1}{\widehat{\rho}_a(\mathbf{S})}-\frac{1}{\rho_a^*(\mathbf{S})}\right|^{2r}\right\}^{\frac{1}{2r}}+\frac{1}{c_0}\left\{E\left|\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)}-\frac{1}{\pi_a^*(\mathbf{S}_1)}\right|^r\right\}^{\frac{1}{r}}\\
&\overset{(iii)}{=}O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}_a}\log(d_1)+s_{\boldsymbol{\delta}_a}\log(d)}{N}}\right).
\end{aligned}$$

where (i) holds by the Minkowski inequality; (ii) holds by Hölder's inequality; (iii) holds by (1.98), (1.99), and the fact that $\{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-2r}\}^{\frac{1}{2r}}$ is bounded uniformly. ∎

1.8.5 Asymptotic theory for general Dynamic Treatment Effect (DTE)

In this section, we consider general nuisance estimators and general working models. Below we introduce some shorthand notations that increase the readability of the proofs. With a slight abuse of notation, $\widehat{\nu}_c(\cdot)=\widehat{\nu}_{c,-k}(\cdot)$, $\widehat{\mu}_c(\cdot)=\widehat{\mu}_{c,-k}(\cdot)$, $\widehat{\pi}_c(\cdot)=\widehat{\pi}_{c,-k}(\cdot)$, and $\widehat{\rho}_c(\cdot)=\widehat{\rho}_{c,-k}(\cdot)$ are estimates of the nuisance functions $\nu_c(\cdot)$, $\mu_c(\cdot)$, $\pi_c(\cdot)$, and $\rho_c(\cdot)$ using the training samples $\mathcal{W}_{-k}$. We also define $\widehat{\Delta}(\cdot)=\widehat{\Delta}_{-k}(\cdot)$ and $\widehat{\psi}_c(\cdot)=\widehat{\psi}_{c,-k}(\cdot)$ for each $c\in\{a,a'\}$ and $k=1,...,K$. We suppress the dependence on k when possible. Note that we have $\widehat{\Delta}(W)=\widehat{\psi}_a(W)-\widehat{\psi}_{a'}(W)$, where for each $c\in\{a,a'\}$,

$$\widehat{\psi}_c(W):=\widehat{\mu}_c(\mathbf{S}_1)+\mathbb{1}_{\{A_1=c_1\}}\frac{\widehat{\nu}_c(\mathbf{S})-\widehat{\mu}_c(\mathbf{S}_1)}{\widehat{\pi}_c(\mathbf{S}_1)}+\mathbb{1}_{\{A_1=c_1,A_2=c_2\}}\frac{Y-\widehat{\nu}_c(\mathbf{S})}{\widehat{\pi}_c(\mathbf{S}_1)\widehat{\rho}_c(\mathbf{S})}. \tag{1.100}$$

Define $\Delta^*(W) = \psi_a^*(W) - \psi_{a'}^*(W)$, where for each $c \in \{a, a'\}$,

$$\psi_c^*(W) := \mu_c^*(\mathbf{S}_1) + \mathbb{1}_{\{A_1=c_1\}} \frac{\nu_c^*(\mathbf{S}) - \mu_c^*(\mathbf{S}_1)}{\pi_c^*(\mathbf{S}_1)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \nu_c^*(\mathbf{S})}{\pi_c^*(\mathbf{S}_1)\rho_c^*(\mathbf{S})}. \tag{1.101}$$

Define $\check{\theta}_{\text{gen}}^{(k)} = n^{-1} \sum_{i \in \mathcal{I}_k} \widehat{\Delta}(W_i)$, where $n = N/K = |\mathcal{I}_k|$. Then $\widehat{\theta}_{\text{gen}} = K^{-1} \sum_{k=1}^K \check{\theta}_{\text{gen}}^{(k)}$. For each $k = 1, ..., K$, we divide $\check{\theta}_{\text{gen}}^{(k)} - \theta$ into four terms T_1, T_2, T_3, T_4,

$$\check{\theta}_{\text{gen}}^{(k)} - \theta = n^{-1} \sum_{i \in \mathcal{I}_k} \widehat{\Delta}_{-k}(W_i) - \theta := T_1 + T_2 + T_3 + T_4, \tag{1.102}$$

where

$$T_1 := E[\Delta^*(W)] - \theta, \tag{1.103}$$

$$T_2 := T_2^{(k)} := E[\widehat{\Delta}(W) - \Delta^*(W)], \tag{1.104}$$

$$T_3 := T_3^{(k)} := \frac{1}{n} \sum_{i \in \mathcal{I}_k} \Delta^*(W_i) - E[\Delta^*(W)], \tag{1.105}$$

$$T_4 := T_4^{(k)} := \frac{1}{n} \sum_{i \in \mathcal{I}_k} [\widehat{\Delta}(W_i) - \Delta^*(W_i)] - E[\widehat{\Delta}(W) - \Delta^*(W)]. \tag{1.106}$$

Auxiliary Lemmas

Lemma 1.9. *Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumption 1 hold. Then,*

$$T_1 = 0, \tag{1.107}$$

where T_1 is defined as (1.103)*.*

Lemma 1.10. *(a) Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1, 4 and 5 hold.*

Then

$$T_2 = O_p\Bigg(b_N c_N + a_N d_N + b_N \mathbb{1}_{\{\pi_a^* \neq \pi_a\}} + a_N \mathbb{1}_{\{\rho_a^* \neq \rho_a\}} + c_N \sqrt{E[\zeta^2 + \varepsilon^2]}\mathbb{1}_{\{\mu_a^* \neq \mu_a\}} + d_N \sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^* \neq \nu_a\}}\Bigg), \tag{1.108}$$

where T_2 is defined as (1.104).

(b) Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumptions 1 and 5 hold. Then

$$T_2 = O_p\left(b_N c_N + a_N d_N\right). \tag{1.109}$$

Lemma 1.11. *(a) Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1, 4 hold. Then*

$$T_3 = O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \tag{1.110}$$

where $\xi := \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$ and T_3 is defined as (1.105).

(b) Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumption 1 hold. Then we also have (1.110).

Lemma 1.12. *(a) Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1, 4 and 5 hold. Then*

$$T_4 = O_p\left(\frac{1}{\sqrt{N}}\left[a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right]\right), \tag{1.111}$$

where T_4 is defined as (1.106).

(b) Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumptions 1 and 5 hold. Then

$$T_4 = O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right). \tag{1.112}$$

Lemma 1.13. *Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumption 1 hold.*

(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 $\sigma^2 := E(\Delta^*(W) - \theta)^2$.

(b) Let Assumption 2 hold. 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 1.14. *Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumption 1 hold. Then we have for some constants $t > 0$ and $C_t > 0$ possibly dependent with t, such that*

$$\sigma^2 := E(\Delta^*(W) - \theta)^2 = E(\Delta(W) - \theta)^2 \geq E[\zeta^2] + E[\varepsilon^2] + E[\xi^2], \tag{1.113}$$

$$E|\Delta^*(W) - \theta|^{2+t} \leq \frac{2C_t}{c_0^{4+2t}}E\left[|\zeta|^{2+t} + |\varepsilon|^{2+t} + |\xi|^{2+t}\right]. \tag{1.114}$$

Lemma 1.15. *Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumption 1 hold. Define $\widehat{\sigma}_{\text{gen}}^2 = N^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i) - \widehat{\theta}_{\text{gen}})^2$ and $\sigma^2 := E(\Delta^*(W) - \theta)^2$. If*

$$\widehat{\theta}_{\text{gen}} - \theta = O_p(\sigma/\sqrt{N}), \qquad \left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i) - \Delta^*(W_i)|^2\right]^{\frac{1}{2}} = o_p(\sigma),$$

for each $k \leq K$, *and* $[E|(\Delta^*(W) - \theta)|^{2+t}]^{\frac{2}{2+t}} < C\sigma^2$ *for some constant* C, *we have*

$$\widehat{\sigma}^2_{\text{gen}} - \sigma^2 = o_p(\sigma^2). \tag{1.115}$$

Proof of Theorem 6

Recall the representation (1.102). By Lemmas 1.9, 1.10, 1.11, and 1.12, we have

$$\begin{aligned}
T_1 &= 0, \\
T_2^{(k)} &= O_p\Bigg(b_N c_N + a_N d_N + b_N \mathbb{1}_{\{\pi_a^* \neq \pi_a\}} + a_N \mathbb{1}_{\{\rho_a^* \neq \rho_a\}} \\
&\qquad + c_N \sqrt{E[\zeta^2 + \varepsilon^2]} \mathbb{1}_{\{\mu_a^* \neq \mu_a\}} + d_N \sqrt{E[\zeta^2]} \mathbb{1}_{\{\nu_a^* \neq \nu_a\}}\Bigg), \\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \\
T_4^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right]\right).
\end{aligned}$$

Together with Lemma 1.13 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

$$\begin{aligned}
\widehat{\theta}_{\text{gen}} - \theta &= K^{-1} \sum_{k=1}^{K} (T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \\
&= O_p\Bigg(b_N c_N + a_N d_N + b_N \mathbb{1}_{\{\pi_a^* \neq \pi_a\}} + a_N \mathbb{1}_{\{\rho_a^* \neq \rho_a\}} \\
&\qquad + c_N \sigma \mathbb{1}_{\{\mu_a^* \neq \mu_a\}} + d_N \sigma \mathbb{1}_{\{\nu_a^* \neq \nu_a\}} + \frac{1}{\sqrt{N}} \sigma\Bigg).
\end{aligned}$$

Proof of Theorem 7

In this theorem, we consider correctly specified nuisance models.

Consistency Recall the representation (1.102), by Lemmas 1.9, 1.10, 1.11, and 1.12, we have

$$T_1 = 0, \tag{1.116}$$

$$T_2^{(k)} = O_p\left(b_N c_N + a_N d_N\right), \tag{1.117}$$

$$T_3^{(k)} = O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right),$$

$$T_4^{(k)} = O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right). \tag{1.118}$$

By assumption, $b_N c_N + a_N d_N = o(\sigma N^{-1/2})$. Together with Lemma 1.14, we obtain that

$$\begin{aligned}
\widehat{\theta}_{\text{gen}} - \theta &= K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \\
&= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right] + b_N c_N + a_N d_N\right) \\
&\quad + O_p\left(\frac{1}{\sqrt{N}}(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]})\right) \\
&= O_p\left(\frac{1}{\sqrt{N}}\sigma\right).
\end{aligned} \tag{1.119}$$

Asymptotic Normality Now, we demonstrate that $\sqrt{N}\sigma^{-1}(\widehat{\theta}_{\text{gen}} - \theta) \rightsquigarrow N(0,1)$. By (1.116), (1.117), and (1.118), under Assumption 5 and $b_N c_N + a_N d_N = o(\sigma N^{-1/2})$, we have

$$\sqrt{n}\sigma^{-1}(T_1 + T_2^{(k)} + T_4^{(k)}) = o_p(1)$$

for each $k \leq K$. Hence, we only need to show

$$\sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K}T_3^{(k)} = \sqrt{N}\sigma^{-1}\left(N^{-1}\sum_{i=1}^{N}\Delta^*(W_i) - \theta\right) \rightsquigarrow N(0,1),$$

where $T_3^{(k)}$ is defined as (1.105). By Lyapunov's central limit theorem, it suffices to show the following Lyapunov's condition holds: with some $t > 0$,

$$\lim_{n\to\infty}\frac{E|\Delta^*(W) - \theta|^{2+t}}{n^{\frac{t}{2}}\sigma^{2+t}} = 0. \tag{1.120}$$

Step 1 To check Lyapunov's condition, it suffices to show that for some constant $C' > 0$,

$$\frac{E|\Delta^*(W) - \theta|^{2+t}}{\sigma^{2+t}} < C'. \tag{1.121}$$

By Lemma 1.14, we have, for some constants $t > 0$ and $C_t > 0$,

$$\begin{aligned}\frac{E|\Delta^*(W) - \theta|^{2+t}}{\sigma^{2+t}} &\leq \frac{2C_t}{c_0^{4+2t}} \frac{E[|\zeta|^{2+t} + |\varepsilon|^{2+t} + |\xi|^{2+t}]}{(E[\zeta^2] + E[\varepsilon^2] + E[\xi^2])^{1+\frac{t}{2}}} \\ &\leq \frac{2C_t}{c_0^{4+2t}} \left(\frac{E[|\zeta|^{2+t}]}{(E[\zeta^2])^{1+\frac{t}{2}}} + \frac{E[|\varepsilon|^{2+t}]}{(E[\varepsilon^2])^{1+\frac{t}{2}}} + \frac{E[|\xi|^{2+t}]}{(E[\xi^2])^{1+\frac{t}{2}}} \right) \leq \frac{2CC_t}{c_0^{4+2t}},\end{aligned} \tag{1.122}$$

where the last inequality follows from Assumption 4. Taking $C' = 2CC_t/c_0^{4+2t}$, we get (1.120) and hence the Lyapunov's condition holds.

Step 2 By (1.119), we have $\widehat{\theta}_{\text{gen}} - \theta = O_p(\sigma/\sqrt{N})$. Here, we show that, for each $k \leq K$,

$$\left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} |\widehat{\Delta}(W_i) - \Delta^*(W_i)|^2 \right]^{\frac{1}{2}} = o_p(\sigma). \tag{1.123}$$

Note that

$$E\left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} |\widehat{\Delta}(W_i) - \Delta^*(W_i)|^2 \right]^{\frac{1}{2}} \overset{(i)}{\leq} \left\{ E\left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} |\widehat{\Delta}(W_i) - \Delta^*(W_i)|^2 \right] \right\}^{\frac{1}{2}} \tag{1.124}$$

$$\overset{(ii)}{=} [E|\widehat{\Delta}(W) - \Delta^*(W)|^2]^{\frac{1}{2}} \tag{1.125}$$

$$\overset{(iii)}{=} O_p\left(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]} \right),$$

where in (1.124), the expectations are taken w.r.t. the joint distribution of $(W_i)_{i \in \mathcal{I}_k}$; in (1.125), 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 the estimator of nuisance functions are independent of $\{W_i\}_{i \in \mathcal{I}_k}$ based on cross-fitting, $\{W_i\}_{i \in \mathcal{I}_k}$ are i.i.d. distributed and W is an

independent copy of them; (iii) holds by Lemma 1.12. By Markov's inequality, we have

$$\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}$$
$$=O_p\left(a_N+b_N+c_N(\sqrt{E[\zeta^2]}+\sqrt{E[\varepsilon^2]})+d_N\sqrt{E[\zeta^2]}\right)=o_p(\sigma).$$

Together with (1.119), (1.120), (1.123), and Lemma 1.15, we conclude that

$$\widehat{\sigma}^2_{\text{gen}}-\sigma^2=o_p(\sigma^2).$$

Proofs of Auxiliary Lemmas

Proof of Lemma 1.9. Recall the definition (1.103). Since $\theta=\theta_a-\theta_{a'}$ and $\Delta^*(W)=\psi^*_a(W)-\psi^*_{a'}(W)$, we have

$$T_1=(E[\psi^*_a(W)]-\theta_a)-(E[\psi^*_{a'}(W)]-\theta_{a'}).$$

By Lemma 1, we have $\theta_c=E[\psi^*_c(W)]$ for each $c\in\{a,a'\}$. Therefore, $T_1=0$. ∎

Proof of Lemma 1.10. In this proof, the expectations are taken w.r.t. the distribution of new observations $\mathbf{S}$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). We only focus on the treatment paths $a=(1,1)$ and $a'=(0,0)$. We begin by decomposing T_2, (1.104), as a sum of six terms

$$\widehat{\Delta}(W)-\Delta^*(W)=\sum_{i=1}^{6}Q_i, \tag{1.126}$$

where

$$Q_1 := \frac{A_1 A_2}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}(Y - \widehat{\nu}_a(\mathbf{S})) - \frac{A_1 A_2}{\pi^*_a(\mathbf{S}_1)\rho^*_a(\mathbf{S})}(Y - \nu^*_a(\mathbf{S})), \tag{1.127}$$

$$Q_2 := \frac{A_1}{\widehat{\pi}_a(\mathbf{S}_1)}(\widehat{\nu}_a(\mathbf{S}) - \widehat{\mu}_a(\mathbf{S}_1)) - \frac{A_1}{\pi^*_a(\mathbf{S}_1)}(\nu^*_a(\mathbf{S}) - \mu^*_a(\mathbf{S}_1)), \tag{1.128}$$

$$Q_3 := \widehat{\mu}_a(\mathbf{S}_1) - \mu^*_a(\mathbf{S}_1), \tag{1.129}$$

$$\begin{aligned} Q_4 := &-\frac{(1-A_1)(1-A_2)}{\widehat{\pi}_{a'}(\mathbf{S}_1)\widehat{\rho}_{a'}(\mathbf{S})}(Y - \widehat{\nu}_{a'}(\mathbf{S})) \\ &+ \frac{(1-A_1)(1-A_2)}{\pi^*_{a'}(\mathbf{S}_1)\rho^*_{a'}(\mathbf{S})}(Y - \nu^*_{a'}(\mathbf{S})), \end{aligned} \tag{1.130}$$

$$Q_5 := -\frac{1-A_1}{\widehat{\pi}_{a'}(\mathbf{S}_1)}(\widehat{\nu}_{a'}(\mathbf{S}) - \widehat{\mu}_{a'}(\mathbf{S}_1)) + \frac{1-A_1}{\pi^*_{a'}(\mathbf{S}_1)}(\nu^*_{a'}(\mathbf{S}) - \mu^*_{a'}(\mathbf{S}_1)), \tag{1.131}$$

$$Q_6 := -\widehat{\mu}_{a'}(\mathbf{S}_1) + \mu^*_{a'}(\mathbf{S}_1). \tag{1.132}$$

Hence, we have the following representation for T_2:

$$T_2 = E[\widehat{\Delta}(W) - \Delta^*(W)] = \sum_{i=1}^{6} E[Q_i], \tag{1.133}$$

where the expecatations are only taken w.r.t. the distribution of the new obseravtion W.

(a) We first obtain an upper bound for $E[Q_1 + Q_2 + Q_3]$. By the tower rule,

$$E[Q_1] = E\left[\frac{A_1\rho_a(\mathbf{S})}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}(\nu_a(\mathbf{S}) - \widehat{\nu}_a(\mathbf{S})) - \frac{A_1\rho_a(\mathbf{S})}{\pi^*_a(\mathbf{S}_1)\rho^*_a(\mathbf{S})}(\nu_a(\mathbf{S}) - \nu^*_a(\mathbf{S}))\right].$$

Through rearranging, we have the following representation:

$$E[Q_1 + Q_2 + Q_3] = \sum_{i=1}^{8} R_i, \tag{1.134}$$

where

$$R_1 := E\left[\frac{A_1\rho_a^*(\mathbf{S})(\widehat{\nu}_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\widehat{\pi}_a(\mathbf{S}_1)}\left(\frac{1}{\rho_a^*(\mathbf{S})}-\frac{1}{\widehat{\rho}_a(\mathbf{S})}\right)\right], \tag{1.135}$$

$$R_2 := E\left[\pi_a^*(\mathbf{S}_1)(\widehat{\mu}_a(\mathbf{S}_1)-\mu_a^*(\mathbf{S}_1))\left(\frac{1}{\pi_a^*(\mathbf{S}_1)}-\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)}\right)\right], \tag{1.136}$$

$$R_3 := E\left[\frac{A_1(\rho_a^*(\mathbf{S})-\rho_a(\mathbf{S}))(\widehat{\nu}_a(\mathbf{S})-\nu_a^*(\mathbf{S}))}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}\right], \tag{1.137}$$

$$\begin{aligned} R_4 :=& E\left[\frac{(\pi_a^*(\mathbf{S}_1)-A_1)(\widehat{\mu}_a(\mathbf{S}_1)-\mu_a^*(\mathbf{S}_1))}{\widehat{\pi}_a(\mathbf{S}_1)}\right] \\ \overset{(i)}{=}& E\left[\frac{(\pi_a^*(\mathbf{S}_1)-\pi_a(\mathbf{S}_1))(\widehat{\mu}_a(\mathbf{S}_1)-\mu_a^*(\mathbf{S}_1))}{\widehat{\pi}_a(\mathbf{S}_1)}\right], \end{aligned} \tag{1.138}$$

$$R_5 := E\left[\frac{A_1\rho_a^*(\mathbf{S})(\nu_a^*(\mathbf{S})-\nu_a(\mathbf{S}))}{\widehat{\pi}_a(\mathbf{S}_1)}\left(\frac{1}{\rho_a^*(\mathbf{S})}-\frac{1}{\widehat{\rho}_a(\mathbf{S})}\right)\right], \tag{1.139}$$

$$R_6 := E\left[A_1(\mu_a^*(\mathbf{S}_1)-\mu_a(\mathbf{S}_1))\left(\frac{1}{\pi_a^*(\mathbf{S}_1)}-\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)}\right)\right], \tag{1.140}$$

$$\begin{aligned} R_7 :=& E\left[\left(\frac{A_1}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})}-\frac{A_1}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}\right)(\rho_a^*(\mathbf{S})-\rho_a(\mathbf{S}))(\nu_a^*(\mathbf{S})-\nu_a(\mathbf{S}))\right] \\ \overset{(ii)}{=}& 0, \end{aligned} \tag{1.141}$$

$$R_8 := E\left[\frac{A_1(\widehat{\pi}_a(\mathbf{S}_1)-\pi_a^*(\mathbf{S}_1))(\mu_a(\mathbf{S}_1)-\nu_a(\mathbf{S}))}{\widehat{\pi}_a(\mathbf{S}_1)\pi_a^*(\mathbf{S}_1)}\right] \overset{(iii)}{=} 0. \tag{1.142}$$

Here, (i) holds by the tower rule; (ii) holds since either $\rho_a^*(\cdot) = \rho_a(\cdot)$ or $\mu_a^*(\cdot) = \mu_a(\cdot)$ by assumption; (iii) holds by the tower rule and the fact that $\mu_a(\mathbf{S}_1) = E[\nu_a(\mathbf{S})|\mathbf{S}_1, A_1 = a_1]$. We condition on the following event

$$\mathcal{E}_4 := \left\{P(c_0 \le \widehat{\pi}_a(\mathbf{S}_1) \le 1-c_0) = 1,\ \ P(c_0 \le \widehat{\rho}_a(\mathbf{S}) \le 1-c_0) = 1\right\}. \tag{1.143}$$

Under Assumption 5, the event $\mathcal{E}_4$ occurs with probability approaching one. In the following, we use Cauchy-Schwarz inequality to obtain an upper bound R_i ($i \in \{1,\ldots,6\}$). For R_1+R_2,

on the event $\mathcal{E}_4$, we have

$$\begin{aligned} R_1 + R_2 &\leq \frac{1}{c_0^2}[E(\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S}))^2]^{\frac{1}{2}}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} \\ &\quad + \frac{1}{c_0}[E(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\ &= O_p\left(b_N c_N + a_N d_N\right), \end{aligned} \tag{1.144}$$

under Assumption 5. For $R_3 + R_4$, on the event $\mathcal{E}_4$, we have

$$\begin{aligned} R_3 + R_4 &\leq \frac{1}{c_0^2}[E(\rho_a^*(\mathbf{S}) - \rho_a(\mathbf{S}))^2]^{\frac{1}{2}}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} \\ &\quad + \frac{1}{c_0}[E(\pi_a^*(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))^2]^{\frac{1}{2}}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\ &\leq \frac{\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}}{c_0^2}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{\mathbb{1}_{\{\pi_a^*(\cdot)\neq\pi_a(\cdot)\}}}{c_0}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}}, \end{aligned}$$

since

$$\begin{aligned} E(\rho_a^*(\mathbf{S}) - \rho_a(\mathbf{S}))^2 &= \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}E(\rho_a^*(\mathbf{S}) - \rho_a(\mathbf{S}))^2 \overset{(i)}{\leq} \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}, \\ E(\pi_a^*(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))^2 &= \mathbb{1}_{\{\pi_a^*(\cdot)\neq\pi_a(\cdot)\}}E(\pi_a^*(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))^2 \overset{(ii)}{\leq} \mathbb{1}_{\{\pi_a^*(\cdot)\neq\pi_a(\cdot)\}}, \end{aligned}$$

where (i) and (ii) hold because $\rho_a(\mathbf{S}) = E(A_2|\mathbf{S}, A_1 = a_1) \in (0,1)$, $\pi_a(\mathbf{S}_1) = E(A_1|\mathbf{S}_1) \in (0,1)$, and, under Assumption 4, $\rho_a^*(\mathbf{S}), \pi_a^*(\mathbf{S}_1) \in (0,1)$ with probability one. Hence, under Assumption 5, we have

$$R_3 + R_4 = O_p\left(b_N \mathbb{1}_{\{\pi_a^*(\cdot)\neq\pi_a(\cdot)\}} + a_N \mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}\right). \tag{1.145}$$

As for $R_5 + R_6$, similarly, we have

$$\begin{aligned} R_5 + R_6 &\leq \frac{1}{c_0^2}[E(\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S}))^2]^{\frac{1}{2}}\left[E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2]\right]^{\frac{1}{2}} \\ &\quad + \frac{1}{c_0^2}[E(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}}\left[E[A_1(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2]\right]^{\frac{1}{2}} \end{aligned} \tag{1.146}$$

Here, we need upper bound for $[E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2]]^{\frac{1}{2}}$ and $[E[A_1(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2]]^{\frac{1}{2}}$.

By definition,

$$\zeta = \zeta_a + \zeta_{a'}, \quad \varepsilon = \varepsilon_a + \varepsilon_{a'}, \quad Y = Y(a)A_1A_2 + Y(0,0)(1-A_1)(1-A_2),$$

where

$$\zeta_a = A_1A_2\left(Y(a) - \nu_a^*(\mathbf{S})\right), \quad \varepsilon_a = A_1\left(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1)\right).$$

Hence, we have

$$E[\zeta^2] \geq E[A_1A_2\zeta^2] = E[\zeta_a^2] = E[A_1A_2(Y - \nu_a^*(\mathbf{S}))^2] \tag{1.147}$$

Note that

$$\begin{aligned}
&E[A_1A_2(Y - \nu_a(\mathbf{S}))(\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))] \\
&\overset{(i)}{=} E[E[A_1A_2(Y(a) - \nu_a(\mathbf{S}))(\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))|\mathbf{S}, A_1 = a_1]P(A_1 = a_1|\mathbf{S})] \\
&\overset{(ii)}{=} E[E[A_2|\mathbf{S}, A_1 = a_1](E[Y(a)|\mathbf{S}, A_1 = a_1] - \nu_a(\mathbf{S}))(\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))P(A_1 = a_1|\mathbf{S})] \\
&\overset{(iii)}{=} 0,
\end{aligned}$$

where (i) holds by the tower rule and the fact that $A_1A_2Y = A_1A_2Y(a)$; (ii) holds under Assumption 1; (iii) holds since $\nu_a(\mathbf{S}) = E[Y(a)|\mathbf{S}, A_1 = a_1]$. Therefore,

$$\begin{aligned}
E[A_1A_2(Y - \nu_a^*(\mathbf{S}))^2] &= E[A_1A_2[(Y - \nu_a(\mathbf{S}))^2 + (\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]] \\
&\geq E[A_1A_2(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2] \overset{(i)}{=} E[A_1\rho_a(\mathbf{S})(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2] \\
&\overset{(ii)}{\geq} c_0E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2],
\end{aligned} \tag{1.148}$$

where (i) holds by the tower rule; (ii) holds under Assumption 1. Together with (1.147), we have

$$E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2] \leq \frac{1}{c_0}E[\zeta^2]. \tag{1.149}$$

Besides, note that

$$E[A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))]$$

$$= E[(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))E[(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))|\mathbf{S}_1, A_1 = a_1]P(A_1 = a_1|\mathbf{S})] = 0,$$

since $E[\nu_a(\mathbf{S})|\mathbf{S}_1, A_1 = a_1] = \mu_a(\mathbf{S}_1)$. Therefore, we have

$$E[A_1(\nu_a(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))^2] = E[A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))^2] + E[A_1(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]$$

$$\geq E[A_1(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]. \tag{1.150}$$

Additionally, observe that

$$E[A_1(\nu_a(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))^2] \leq 2E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2] + 2E[\varepsilon_a^2]$$

$$\overset{(i)}{\leq} \frac{2}{c_0}E[\zeta^2] + 2E[A_1\varepsilon^2] \leq \frac{2}{c_0}E[\zeta^2] + 2E[\varepsilon^2],$$

where (i) holds by (1.149) and the fact that $\varepsilon_a^2 = A_1\varepsilon^2$. Together with (1.150), we obtain

$$E[A_1(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2] \leq \frac{2}{c_0}E[\zeta^2] + 2E[\varepsilon^2]. \tag{1.151}$$

Therefore, under Assumption 5,

$$R_5 + R_6 \leq \frac{1}{c_0^2}[E(\widehat{\rho}_a(\mathbf{S}) - \rho_a^*(\mathbf{S}))^2]^{\frac{1}{2}}[E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2]]^{\frac{1}{2}}$$

$$+ \frac{1}{c_0^2}[E(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}}[E[A_1(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2]]^{\frac{1}{2}}$$

$$= O_p\left(c_N\sqrt{E[\zeta^2 + \varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}} + d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\right). \tag{1.152}$$

Pluging (1.141), (1.142), (1.144),(1.145), and (1.152) into (1.134), we obtain

$$E[Q_1 + Q_2 + Q_3] = O_p\Bigg(b_N c_N + a_N d_N + b_N\mathbb{1}_{\{\pi_a^*(\cdot)\neq\pi_a(\cdot)\}} + a_N\mathbb{1}_{\{\rho_a^*(\cdot)\neq\rho_a(\cdot)\}}$$

$$+ c_N\sqrt{E[\zeta^2 + \varepsilon^2]}\mathbb{1}_{\{\mu_a^*(\cdot)\neq\mu_a(\cdot)\}} + d_N\sqrt{E[\zeta^2]}\mathbb{1}_{\{\nu_a^*(\cdot)\neq\nu_a(\cdot)\}}\Bigg).$$

By repeating all the previous steps, we can obtain the same result for $E[Q_4 + Q_5 + Q_6]$. Therefore, (1.108) follows.

(b) When all the nuisance models are correct, Assumption 4 holds under Assumption 1. Hence, by part (a), we also have (1.108). Since all the nuisance models are correct, we further conclude that (1.109) holds. ∎

Proof of Lemma 1.11. (a) Recall the definition (1.105). By Chebyshev's inequality, we have for any $t > 0$,

$$P(|T_3| > t) \leq \frac{1}{t^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Delta^*(W_i)\right) \leq \frac{1}{nt^2}E[\Delta^*(W)]^2.$$

To prove (1.110), we only need to show $[E(\Delta^*(W))^2]^{\frac{1}{2}} = O(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]})$. By Minkowski inequality, we have

$$[E(\Delta^*(W))^2]^{\frac{1}{2}} \leq \sum_{i=1}^{5} T_{3,i}, \tag{1.153}$$

where

$$\begin{aligned}
T_{3,1} &:= \left[E\left(\frac{A_1A_2}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}(Y - \nu_a^*(\mathbf{S}))\right)^2\right]^{\frac{1}{2}},\\
T_{3,2} &:= \left[E\left(\frac{A_1}{\pi_a^*(\mathbf{S}_1)}(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))\right)^2\right]^{\frac{1}{2}},\\
T_{3,3} &:= \left[E\left(\frac{(1-A_1)(1-A_2)}{\pi_{a'}^*(\mathbf{S}_1)\rho_{a'}^*(\mathbf{S})}(Y - \nu_{a'}^*(\mathbf{S}))\right)^2\right]^{\frac{1}{2}},\\
T_{3,4} &:= \left[E\left(\frac{1-A_1}{\pi_{a'}^*(\mathbf{S}_1)}(\nu_{a'}^*(\mathbf{S}) - \mu_{a'}^*(\mathbf{S}_1))\right)^2\right]^{\frac{1}{2}},\\
T_{3,5} &:= \left[E\left(\mu_a^*(\mathbf{S}_1) - \mu_{a'}^*(\mathbf{S}_1) - \theta\right)^2\right]^{\frac{1}{2}}.
\end{aligned}$$

We bound each of the above terms in turn. Under Assumption 4 and recall the equation

(1.147), we have

$$T_{3,1} \leq \frac{1}{c_0^2}[E(A_1A_2(Y-\nu_a^*(\mathbf{S}))^2)]^{\frac{1}{2}} \leq \frac{1}{c_0^2}\sqrt{E[\zeta^2]}. \tag{1.154}$$

Similarly, since $E[\varepsilon^2] \geq E[A_1\varepsilon^2] = E[\varepsilon_a^2] = E[A_1(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))^2]$, we have

$$T_{3,2} \leq \frac{1}{c_0}[E(A_1(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))^2)]^{\frac{1}{2}} \leq \frac{1}{c_0}\sqrt{E[\varepsilon^2]}. \tag{1.155}$$

Repeating the same process for $T_{3,3}$ and $T_{3,4}$, we also have

$$T_{3,3} \leq \frac{1}{c_0^2}\sqrt{E[\zeta^2]}, \quad T_{3,4} \leq \frac{1}{c_0}\sqrt{E[\varepsilon^2]}. \tag{1.156}$$

Additionally,

$$\begin{aligned}\frac{2}{c_0}E[\zeta^2] + 2E[\varepsilon^2] &\overset{(i)}{\geq} E[A_1(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2] \overset{(ii)}{=} E[\pi_a(\mathbf{S}_1)(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2] \\ &\overset{(iii)}{\geq} c_0 E[(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2],\end{aligned}$$

where (i) holds by (1.151); (ii) holds by the tower rule; (iii) holds under Assumption 1. Similarly, we also have

$$\frac{2}{c_0}E[\zeta^2] + 2E[\varepsilon^2] \geq c_0 E[(\mu_{a'}^*(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1))^2].$$

By Minkowski inequality,

$$\begin{aligned}T_{3,5} &\leq [E(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E(\mu_{a'}^*(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E[\xi^2]]^{\frac{1}{2}} \\ &\leq 2\sqrt{\frac{2}{c_0^2}E[\zeta^2] + \frac{2}{c_0}E[\varepsilon^2]} + \sqrt{E[\xi^2]} \leq \frac{2\sqrt{2}}{c_0}\sqrt{E[\zeta^2]} + \frac{2\sqrt{2}}{\sqrt{c_0}}\sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}. \end{aligned} \tag{1.157}$$

Plugging (1.154)-(1.157) into (1.153), we have

$$[E(\Delta^*(W))^2]^{\frac{1}{2}} = O\left(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right).$$

(b) When all the models are correctly specified, Assumption 1 implies Assumption 4. Hence, by part (a), we also have (1.110). ∎

Proof of Lemma 1.12. In this proof, the expectations are taken w.r.t. the distribution of new observations $\mathbf{S}$ (or only $\mathbf{S}_1$ if $\mathbf{S}_2$ is not involved). Additionally, we condition on the event $\mathcal{E}_4$, defined as (1.143). Under Assumption 5, such an event occurs with probability approaching one.

(a) We first show (1.111). Recall the representation (1.126), by Minkowski inequality, we have

$$[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} \leq \sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}},$$

where Q_i ($i \in \{1, \ldots, 6\}$) are defined as(1.127)-(1.132). Then, by Chebyshev's inequality, it suffices to show

$$\sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}} = O_p\left(a_N + b_N + \sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}\right).$$

Additionally, under Assumption 4, we also have

$$P(c_0 \leq \pi_a^*(\mathbf{S}_1) \leq 1 - c_0) = 1, \;\; P(c_0 \leq \rho_a^*(\mathbf{S}) \leq 1 - c_0) = 1.$$

For the first term $[E(Q_1^2)]^{\frac{1}{2}}$, under Assumption 4 and on the event $\mathcal{E}_4$,

$$\begin{aligned}
&[E(Q_1^2)]^{\frac{1}{2}} \\
&\leq \frac{1}{c_0^4}\{E[A_1 A_2 \pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})(Y - \widehat{\nu}_a(\mathbf{S})) - A_1 A_2 \widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})(Y - \nu_a^*(\mathbf{S}))]^2\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \frac{1}{c_0^4}\{E[\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})(\nu_a^*(\mathbf{S}) + \zeta - \widehat{\nu}_a(\mathbf{S})) - \widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})\zeta]^2\}^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} \frac{1}{c_0^4}\{E[\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})]^2\}^{\frac{1}{2}} + \frac{1}{c_0^4}\{E[(\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S}) - \pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))\zeta]^2\}^{\frac{1}{2}}, \qquad (1.158)
\end{aligned}$$

where (i) holds by the fact that $|A_1| \leq 1$, $|A_2| \leq 1$ and $A_1 A_2 Y = A_1 A_2 \nu_a^*(\mathbf{S}) + A_1 A_2 \zeta$; (ii) holds from Minkowski inequality and the fact that $P(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}) \leq 1) = 1$. Since

$P(0 \leq \pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}) \leq 1) = 1$ and $P(0 \leq \widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S}) \leq 1) = 1$ under $\mathcal{E}_4$, we have

$$[E(Q_1^2)]^{\frac{1}{2}} \leq \frac{1}{c_0^4}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^4}[E(\zeta^2)]^{\frac{1}{2}} = O_p\left(b_N + \sqrt{E[\zeta^2]}\right). \tag{1.159}$$

Similarly, for the second term $[E(Q_2^2)]^{\frac{1}{2}}$, under Assumption 4 and on the event $\mathcal{E}_4$,

$$\begin{aligned}
[E(Q_2^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^2}\{E[A_1\pi_a^*(\mathbf{S}_1)(\widehat{\nu}_a(\mathbf{S}) - \widehat{\mu}_a(\mathbf{S}_1)) - A_1\widehat{\pi}_a(\mathbf{S}_1)(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\
&\overset{(i)}{\leq} \frac{1}{c_0^2}\{E[\pi_a^*(\mathbf{S}_1)(\widehat{\nu}_a(\mathbf{S}) - \widehat{\mu}_a(\mathbf{S}_1)) - \widehat{\pi}_a(\mathbf{S}_1)\varepsilon]^2\}^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} \frac{1}{c_0^2}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}} \\
&\quad + \frac{1}{c_0^2}\{E[(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a^*(\mathbf{S}_1))\varepsilon]^2\}^{\frac{1}{2}} \qquad (1.160) \\
&\overset{(iii)}{\leq} \frac{1}{c_0^2}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}\{E[\varepsilon^2]\}^{\frac{1}{2}} \\
&= O_p\left(a_N + b_N + \sqrt{E[\varepsilon^2]}\right), \qquad (1.161)
\end{aligned}$$

where (i) holds from the fact that $|A_1| \leq 1$ and $A_1\nu_a^*(\mathbf{S}) = A_1\mu_a^*(\mathbf{S}_1) + A_1\varepsilon$; (ii) holds from Minkowski inequality and $P(\pi_a^*(\mathbf{S}_1) \leq 1) = 1$; (iii) holds by the fact that $P(0 \leq \pi_a^*(\mathbf{S}_1) \leq 1) = 1$ and $P(0 \leq \widehat{\pi}_a(\mathbf{S}_1) \leq 1) = 1$ on $\mathcal{E}_4$. For the third term $[E(Q_3^2)]^{\frac{1}{2}}$, we have

$$[E(Q_3^2)]^{\frac{1}{2}} = O_p\left(b_N\right), \tag{1.162}$$

under Assumption 5. Combining (1.159), (1.161) and (1.162), 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, (1.111) follows.

(b) Now, we show (1.112). By (1.158), under Assumption 5, we have

$$\begin{aligned}[E(Q_1^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^4}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a(\mathbf{S}))^2]^{\frac{1}{2}} \\ &\quad + \frac{1}{c_0^4}\{E[\zeta^2|\mathbf{S}]\}^{\frac{1}{2}}\{E[(\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S}) - \pi_a(\mathbf{S}_1)\rho_a(\mathbf{S}))]^2\}^{\frac{1}{2}} \\ &\leq \frac{1}{c_0^4}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{\sqrt{CE[\zeta^2]}}{c_0^4}\{E[(\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S}) - \pi_a(\mathbf{S}_1)\rho_a(\mathbf{S}))]^2\}^{\frac{1}{2}}\end{aligned}$$

By Minkowski inequality and under $\mathcal{E}_4$, we have

$$\begin{aligned}&\{E[\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S}) - \pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})]^2\}^{\frac{1}{2}} \\ &\quad \leq \{E[(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))\widehat{\rho}_a(\mathbf{S})]^2\}^{\frac{1}{2}} + \{E[\pi_a(\mathbf{S}_1)(\widehat{\rho}_a(\mathbf{S}) - \rho_a(\mathbf{S}))]^2\}^{\frac{1}{2}} \\ &\quad \leq [E(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))^2]^{\frac{1}{2}} + [E(\widehat{\rho}_a(\mathbf{S}) - \rho_a(\mathbf{S}))^2]^{\frac{1}{2}} = O_p\left(c_N + d_N\right).\end{aligned}$$

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 (1.160), under Assumption 5, we have

$$\begin{aligned}[E(Q_2^2)]^{\frac{1}{2}} &\leq \frac{1}{c_0^2}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a(\mathbf{S}_1)]^2]^{\frac{1}{2}} \\ &\quad + \frac{1}{c_0^2}\{E[\varepsilon^2|\mathbf{S}_1]\}^{\frac{1}{2}}\{E[(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\ &\leq \frac{1}{c_0^2}[E(\widehat{\nu}_a(\mathbf{S}) - \nu_a(\mathbf{S}))^2]^{\frac{1}{2}} + \frac{1}{c_0^2}[E(\widehat{\mu}_a(\mathbf{S}_1) - \mu_a(\mathbf{S}_1)]^2]^{\frac{1}{2}} \\ &\quad + \frac{\sqrt{CE[\varepsilon^2]}}{c_0^2}\{E[(\widehat{\pi}_a(\mathbf{S}_1) - \pi_a(\mathbf{S}_1))]^2\}^{\frac{1}{2}} \\ &= O_p\left(a_N + b_N + c_N\sqrt{E[\varepsilon^2]}\right).\end{aligned}$$

Besides, by Assumption 5,

$$[E(Q_3^2)]^{\frac{1}{2}} = O_p\left(b_N\right).$$

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

$$
[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} = O_P\left(a_N + b_N + c_N(\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]}) + d_N\sqrt{E[\zeta^2]}\right).
$$

By Chebyshev's inequality, we conclude that (1.112) holds. ∎

Proof of Lemma 1.13. (a) We notice the following representation:

$$
\Delta^*(W) - \theta = \sum_{i=1}^{8} O_i, \tag{1.163}
$$

where

$$
O_1 := \frac{A_1 A_2 (Y - \nu_a(\mathbf{S}))}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})}, \tag{1.164}
$$

$$
O_2 := \frac{A_1}{\pi_a^*(\mathbf{S}_1)}\left(1 - \frac{A_2}{\rho_a^*(\mathbf{S})}\right)(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})), \tag{1.165}
$$

$$
O_3 := \frac{A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))}{\pi_a^*(\mathbf{S}_1)}, \tag{1.166}
$$

$$
O_4 := -\frac{(1 - A_1)(1 - A_2)(Y - \nu_{a'}(\mathbf{S}))}{\pi_{a'}^*(\mathbf{S}_1)\rho_{a'}^*(\mathbf{S})}, \tag{1.167}
$$

$$
O_5 := -\frac{1 - A_1}{\pi_{a'}^*(\mathbf{S}_1)}\left(1 - \frac{1 - A_2}{\rho_{a'}^*(\mathbf{S})}\right)(\nu_{a'}^*(\mathbf{S}) - \nu_{a'}(\mathbf{S})), \tag{1.168}
$$

$$
O_6 := -\frac{(1 - A_1)(\nu_{a'}(\mathbf{S}) - \mu_{a'}(\mathbf{S}))}{\pi_{a'}^*(\mathbf{S}_1)}, \tag{1.169}
$$

$$
\begin{aligned}
O_7 :=& \left(1 - \frac{A_1}{\pi_a^*(\mathbf{S}_1)}\right)(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1)) \\
&- \left(1 - \frac{1 - A_1}{\pi_{a'}^*(\mathbf{S}_1)}\right)(\mu_{a'}^*(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1)),
\end{aligned} \tag{1.170}
$$

$$
O_8 := \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta = \xi. \tag{1.171}
$$

In the following, we demonstrate that

$$\sigma^2 = E(\Delta^*(W) - \theta)^2 = \sum_{i=1}^{8} E[O_i^2]. \tag{1.172}$$

It suffices to show that $E[O_iO_j] = 0$ for all $i \neq j$. Firstly, since $A_1(1 - A_1) = 0$, we have

$$O_iO_j = 0, \text{ for each } i \in \{1, 2, 3\}, \text{ and } j \in \{4, 5, 6\}. \tag{1.173}$$

Step 1 We show $E[O_1O_i] = 0$ for each $i \geq 2$. By (1.173), we know that $O_1O_i = 0$ 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 $i \in \{3, 7, 8\}$,

$$E[O_1O_i] = E[O_iE[O_1|\mathbf{S}, A_1 = a_1]P(A_1 = a_1|\mathbf{S})] = 0,$$

since

$$E[O_1|\mathbf{S}, A_1 = a_1] \overset{(i)}{=} \frac{E[A_2|\mathbf{S}, A_1 = a_1]E[Y(a) - \mu_a(\mathbf{S}_1)|\mathbf{S}, A_1 = a_1]}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \overset{(ii)}{=} 0,$$

where (i) holds under Assumption 1; (ii) holds because $E[Y(a)|\mathbf{S}, A_1 = a_1] = \mu_a(\mathbf{S}_1)$. Besides, we note that

$$\begin{aligned}
E[O_1O_2] &= E\left[\frac{A_1A_2(Y - \nu_a(\mathbf{S}))(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))(\rho_a^*(\mathbf{S}) - 1)}{(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))^2}\right] \\
&\overset{(i)}{=} E\left[\frac{E[A_2(Y(a) - \nu_a(\mathbf{S}))|\mathbf{S}, A_1 = a_1](\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))(\rho_a^*(\mathbf{S}) - 1)}{(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))^2}P(A_1 = a_1|\mathbf{S})\right] \\
&\overset{(ii)}{=} E\left[\frac{\rho_a(\mathbf{S})E[Y(a) - \nu_a(\mathbf{S})|\mathbf{S}, A_1 = a_1](\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))(\rho_a^*(\mathbf{S}) - 1)}{(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))^2}P(A_1 = a_1|\mathbf{S})\right] \\
&\overset{(iii)}{=} 0,
\end{aligned}$$

where (i) holds by the tower rule; (ii) holds under Assumption 1; (iii) holds because $E[Y(a)|\mathbf{S}, A_1 = a_1] = \mu_a(\mathbf{S}_1)$.

Step 2 We show $E[O_2O_i] = 0$ for each $i \geq 3$. By (1.173), we know that $O_2O_i = 0$ 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 $i \in \{3, 7, 8\}$,

$$E[O_2O_i] = E[O_iE[O_2|\mathbf{S}, A_1 = a_1]P(A_1 = a_1|\mathbf{S})] = 0,$$

since

$$\begin{aligned} E[O_2|\mathbf{S}, A_1 = a_1] &= \frac{\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)} \left(1 - \frac{E[A_2|\mathbf{S}, A_1 = a_1]}{\rho_a^*(\mathbf{S})}\right) \\ &= \frac{\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)} \left(1 - \frac{\rho_a(\mathbf{S})}{\rho_a^*(\mathbf{S})}\right) \overset{(i)}{=} 0, \end{aligned}$$

where (i) holds because either $\nu_a^*(\cdot) = \nu_1(\cdot)$ or $\rho_a^*(\cdot) = \rho_a(\cdot)$ by assumption.

Step 3 We show $E[O_3O_i] = 0$ for each $i \geq 4$. By (1.173), we know that $O_3O_i = 0$ 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 $i \in \{7, 8\}$,

$$E[O_3O_i] = E[O_iE[O_3|\mathbf{S}_1, A_1 = a_1]P(A_1 = a_1|\mathbf{S}_1)] = 0,$$

since

$$E[O_3|\mathbf{S}_1, A_1 = a_1] = \frac{E[\nu_a(\mathbf{S})|\mathbf{S}_1, A_1 = a_1] - \mu_a(\mathbf{S}_1)}{\pi_a^*(\mathbf{S}_1)} \overset{(i)}{=} 0,$$

where (i) holds because $E[\nu_a(\mathbf{S})|\mathbf{S}_1, A_1 = a_1] = \mu_a(\mathbf{S}_1)$.

Step 4 By repeating the same procedure as in Steps 1-3, we also have $E[O_iO_j] = 0$ for each $i \in \{4, 5, 6\}$ and $j \geq i + 1$.

Step 5 We show $E[O_7O_8] = 0$. Since O_8 is a function of $\mathbf{S}_1$, we have

$$E[O_7O_8] = E[O_8E[O_7|\mathbf{S}_1]] = 0,$$

since

$$E[O_7|\mathbf{S}_1] = \left(1 - \frac{\pi_a(\mathbf{S})}{\pi_a^*(\mathbf{S}_1)}\right)(\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1)) - \left(1 - \frac{\pi_{a'}(\mathbf{S})}{\pi_{a'}^*(\mathbf{S}_1)}\right)(\mu_{a'}^*(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1))$$
$$\overset{(i)}{=} 0,$$

where (i) holds because, by assumption, 1) either $\pi_a^*(\cdot) = \pi_a(\cdot)$ or $\mu_a^*(\cdot) = \mu_a(\cdot)$, and 2) either $\pi_{a'}^*(\cdot) = \pi_{a'}(\cdot)$ or $\mu_{a'}^*(\cdot) = \mu_{a'}(\cdot)$.

Based on all Steps 1-5, we conclude that (1.172) holds. Now, note that

$$\begin{aligned}
E[O_1^2] &\geq E[A_1A_2(Y(a) - \nu_a(\mathbf{S}))^2], \\
E[O_2^2] &= E\left[\frac{A_1((\rho_a^*(\mathbf{S}))^2 - 2A_2\rho_a^*(\mathbf{S}) + A_2)}{(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))^2}(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2\right] \\
&= E\left[\frac{A_1((\rho_a^*(\mathbf{S}) - \rho_a(\mathbf{S}))^2 + \rho_a(\mathbf{S})(1 - \rho_a(\mathbf{S})))}{(\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S}))^2}(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2\right] \\
&\geq c_0^2 E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2], \\
E[O_3^2] &= E\left[\frac{A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))^2}{(\pi_a^*(\mathbf{S}_1))^2}\right] \geq E[A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))^2]
\end{aligned}$$

Hence,

$$\begin{aligned}
E[A_1A_2\zeta^2] &= E[\zeta_a^2] = E[A_1A_2(Y(a) - \nu_a^*(\mathbf{S}))^2] \\
&\overset{(i)}{=} E[A_1A_2((Y(a) - \nu_a(\mathbf{S}))^2 + (\nu_a(\mathbf{S}) - \nu_a^*(\mathbf{S}))^2)] \leq E[O_1^2] + \frac{1}{c_0^2}E[O_2^2], \qquad (1.174)
\end{aligned}$$

where (i) holds as in (1.148). Additionally,

$$\begin{aligned}
E[A_1\varepsilon^2] &= E[\varepsilon_a^2] = E[A_1(\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1))^2] \\
&\leq 3\left[E[A_1(\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S}))^2] + E[A_1(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))^2] + E[A_1(\mu_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1))^2]\right] \\
&\leq \frac{3}{c_0^2}E[O_2^2] + 3E[O_3^2] + 3C_\mu\sigma^2.
\end{aligned}$$

Repeating the process above, we also have

$$E[(1-A_1)(1-A_2)\zeta^2] \leq E[O_4^2] + \frac{1}{c_0^2}E[O_5^2], \tag{1.175}$$
$$E[(1-A_1)\varepsilon^2] \leq \frac{3}{c_0^2}E[O_5^2] + 3E[O_6^2] + 3C_\mu\sigma^2.$$

Besides, we also have

$$E[\xi^2] = E[O_8^2]. \tag{1.176}$$

Therefore, we conclude that

$$\begin{aligned}
&E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \\
&\quad = E[A_1A_2\zeta^2] + E[(1-A_1)(1-A_2)\zeta^2] + E[A_1\varepsilon^2] + E[(1-A_1)\varepsilon^2] + E[\xi^2] \\
&\quad \leq E[O_1^2 + \frac{4}{c_0^2}O_2^2 + 3O_3^2 + O_4^2 + \frac{4}{c_0^2}O_5^2 + 3O_6^2 + O_8^2] + 6C_\mu\sigma^2 \leq \left(\frac{4}{c_0^2} + 6C_\mu\right)\sigma^2,
\end{aligned}$$

since $c_0 < 1$ and (1.172) holds.

(b) Now, we assume Assumption 2 holds. Same as in part (a), we also have (1.172), (1.174), (1.175), and (1.176) hold. Additionally, under Assumption 2, by Lemma 1.2, we also have

$$E[\varepsilon^2] \leq 2\sigma_\varepsilon^2\sigma^2.$$

Therefore,

$$\begin{aligned}
&E[\zeta^2] + E[\varepsilon^2] + E[\xi^2] \\
&\quad = E[A_1A_2\zeta^2] + E[(1-A_1)(1-A_2)\zeta^2] + E[\varepsilon^2] + E[\xi^2] \\
&\quad \leq E[O_1^2 + \frac{1}{c_0^2}O_2^2 + O_4^2 + \frac{1}{c_0^2}O_5^2 + O_8^2] + 2\sigma_\varepsilon^2\sigma^2 \leq \left(\frac{1}{c_0^2} + 2\sigma_\varepsilon^2\right)\sigma^2.
\end{aligned}$$

■

Proof of Lemma 1.14. We first show that (1.113) holds. By Lemma 1.13, we have

$$\Delta^*(W) - \theta = \sum_{i=1}^{8} O_i, \quad \sigma^2 = E(\Delta^*(W) - \theta)^2 = \sum_{i=1}^{8} E[O_i^2],$$

where $\{O_i\}_{i=1}^8$ are defined as (1.164)-(1.171). Since now we assume that all the models are correctly specified, we have $O_i = 0$ for $i \in \{2, 5, 7\}$ and hence

$$\Delta^*(W) - \theta = O_1 + O_3 + O_4 + O_6 + O_8, \tag{1.177}$$
$$\sigma^2 = E[O_1^2] + E[O_3^2] + E[O_4^2] + E[O_6^2] + E[O_8^2] = \sum_{i=1}^{5} V_i,$$

where

$$\begin{aligned}
V_1 :=& E\left[\left(\frac{A_1 A_2}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}(Y - \nu_a(\mathbf{S}))\right)^2\right], \\
V_2 :=& E\left[\left(\frac{A_1}{\pi_a(\mathbf{S}_1)}(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))\right)^2\right], \\
V_3 :=& E\left[\left(\frac{(1-A_1)(1-A_2)}{\pi_{a'}(\mathbf{S}_1)\rho_{a'}(\mathbf{S})}(Y - \nu_{a'}(\mathbf{S}))\right)^2\right], \\
V_4 :=& E\left[\left(\frac{1-A_1}{\pi_{a'}(\mathbf{S}_1)}(\nu_{a'}(\mathbf{S}) - \mu_{a'}(\mathbf{S}_1))\right)^2\right], \\
V_5 :=& E\left[(\mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta)^2\right].
\end{aligned}$$

We lower bound each terms above:

$$\begin{aligned}
V_1 &\overset{(i)}{=} E\left[\left(\frac{\zeta_a}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}\right)^2\right] \overset{(ii)}{=} E\left[\left(\frac{A_1 A_2}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}\zeta\right)^2\right] \overset{(iii)}{\geq} E[A_1 A_2 \zeta^2], \\
V_2 &\overset{(iv)}{=} E\left[\left(\frac{\varepsilon_a}{\pi_a(\mathbf{S}_1)}\right)^2\right] \overset{(v)}{=} E\left[\left(\frac{A_1}{\pi_a(\mathbf{S}_1)}\varepsilon\right)^2\right] \overset{(vi)}{\geq} E[A_1 \varepsilon^2],
\end{aligned}$$

where (i) and (iv) hold since $\nu_a^*(\cdot) = \nu_a(\cdot)$ and $\mu_a^*(\cdot) = \mu_a(\cdot)$; (ii) and (v) hold since $\zeta_a = A_1 A_2 \zeta$ and $\varepsilon_a = A_1 \varepsilon$; (iii) and (vi) hold since $A_1, A_2 \in \{0, 1\}$, $\pi_a(\mathbf{S}_1) \leq 1$ and $\rho_a(\mathbf{S}) \leq 1$ with probability 1 under Assumption 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, $\xi = \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$. Hence,

$$V_5 = E[\xi^2].$$

Combining all the previous results, we have

$$\begin{aligned}
\sigma^2 :&= E[\Delta^*(W) - \theta]^2 \\
&\geq E[A_1 A_2 \zeta^2 + (1 - A_1)(1 - A_2)\zeta^2] + E[A_1 \varepsilon^2 + (1 - A_1)\varepsilon^2] + E[\xi^2] \\
&= E[\zeta^2] + E[\varepsilon^2] + E[\xi^2].
\end{aligned}$$

Next, we show that (1.114) holds. Recall the representation (1.177). 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

$$\begin{aligned}
\left|\frac{\Delta^*(W) - \theta}{5}\right|^{2+t} &= \left|\frac{O_1 + O_3 + O_4 + O_6 + O_8}{5}\right|^{2+t} \\
&\leq \frac{|O_1|^{2+t} + |O_3|^{2+t} + |O_4|^{2+t} + |O_6|^{2+t} + |O_8|^{2+t}}{5}
\end{aligned}$$

Therefore,

$$\begin{aligned}
E|\Delta^*(W) - \theta|^{2+t} &\leq 5^{1+t} E[|O_1|^{2+t} + |O_3|^{2+t} + |O_4|^{2+t} + |O_6|^{2+t} + |O_8|^{2+t}] \\
&= C_t \sum_{i=1}^{5} V_i',
\end{aligned}$$

where $C_t = 5^{1+t}$ and

$$
\begin{aligned}
V_1' :=& E\left[\left|\frac{A_1 A_2}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}(Y - \nu_a(\mathbf{S}))\right|^{2+t}\right], \\
V_2' :=& E\left[\left|\frac{A_1}{\pi_a(\mathbf{S}_1)}(\nu_a(\mathbf{S}) - \mu_a(\mathbf{S}_1))\right|^{2+t}\right], \\
V_3' :=& E\left[\left|\frac{(1-A_1)(1-A_2)}{\pi_{a'}(\mathbf{S}_1)\rho_{a'}(\mathbf{S})}(Y - \nu_{a'}(\mathbf{S}))\right|^{2+t}\right], \\
V_4' :=& E\left[\left|\frac{1-A_1}{\pi_{a'}(\mathbf{S}_1)}(\nu_{a'}(\mathbf{S}) - \mu_{a'}(\mathbf{S}_1))\right|^{2+t}\right], \\
V_5' :=& E\left[|\mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta|^{2+t}\right].
\end{aligned}
$$

Now, we upper bound each of the terms above.

$$
\begin{aligned}
V_1' &\overset{(i)}{=} E\left[\left|\frac{\zeta_a}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}\right|^{2+t}\right] \overset{(ii)}{=} E\left[\left|\frac{A_1 A_2}{\pi_a(\mathbf{S}_1)\rho_a(\mathbf{S})}\zeta\right|^{2+t}\right] \overset{(iii)}{\leq} \frac{1}{c_0^{4+2t}} E[|\zeta|^{2+t}], \\
V_2' &\overset{(iv)}{=} E\left[\left|\frac{\varepsilon_a}{\pi_a(\mathbf{S}_1)}\right|^{2+t}\right] \overset{(v)}{=} E\left[\left|\frac{A_1}{\pi_a(\mathbf{S}_1)}\varepsilon\right|^{2+t}\right] \overset{(vi)}{\leq} \frac{1}{c_0^{4+2t}} E[|\varepsilon|^{2+t}],
\end{aligned}
$$

where (i) and (iv) hold since $\nu_a^*(\cdot) = \nu_a(\cdot)$ and $\mu_a^*(\cdot) = \mu_a(\cdot)$; (ii) and (v) hold since $\zeta_a = A_1 A_2 \zeta$ and $\varepsilon_a = A_1 \varepsilon$; (iii) and (vi) hold since $A_1, A_2 \in \{0, 1\}$, $\pi_a(\mathbf{S}_1), \rho_a(\mathbf{S}) \in [c_0, 1 - c_0]$ with probability 1 under Assumption 1. Similarly, we also have

$$
V_3' \leq \frac{1}{c_0^{4+2t}} E[|\zeta|^{2+t}], \quad V_4' \leq \frac{1}{c_0^{2+t}} E[|\varepsilon|^{2+t}].
$$

In addition, by definition, $\xi = \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$. Hence,

$$
V_5' = E[|\xi|^{2+t}].
$$

Therefore, we conclude that

$$
\begin{aligned}
E|\Delta^*(W) - \theta|^{2+t} &\leq C_t \left[\frac{2}{c_0^{4+2t}} E[|\zeta|^{2+t}] + \frac{2}{c_0^{2+t}} E[|\varepsilon|^{2+t}] + E[|\xi|^{2+t}]\right] \\
&\leq \frac{2C_t}{c_0^{4+2t}} E[|\zeta|^{2+t} + |\varepsilon|^{2+t} + |\xi|^{2+t}],
\end{aligned}
$$

since $0 < c_0 < 1$ and $t > 0$. ∎

Proof of Lemma 1.15. We show that for each $k = 1, ..., K$,

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\Delta^*(W_i) - \theta)^2 - \sigma^2 = o_p(\sigma^2), \tag{1.178}$$

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i) - \widehat{\theta}_{\text{gen}})^2 - \frac{1}{n}\sum_{i\in\mathcal{I}_k}(\Delta^*(W_i) - \theta)^2 = o_p(\sigma^2), \tag{1.179}$$

We first show (1.178). Let $Z_{N,i} := \sigma^{-1}(\Delta^*(W_i) - \theta)^2 - 1$, note that W_i and nuisance functions $\nu_c^*(\cdot)$, $\mu_c^*(\cdot)$, $\pi_c^*(\cdot)$, and $\rho_c^*(\cdot)$ 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 (1.178) is equivalent to

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k} Z_i = o(1).$$

By Lemma 3 of [ZB22], 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,

$$E(Z_{d,1}) = E\left[\frac{(\Delta^*(W) - \theta)^2}{\sigma^2} - 1\right] = \frac{\sigma^2}{\sigma^2} - 1 = 0.$$

In addition, by Minkowski inequality,

$$\left[E\left|\frac{(\Delta^*(W) - \theta)^2}{\sigma^2} - 1\right|^{\frac{2+t}{2}}\right]^{\frac{2}{2+t}} \leq \left[\frac{E|(\Delta^*(W) - \theta)|^{2+t}}{\sigma^{2+t}}\right]^{\frac{2}{2+t}} + 1 < C + 1.$$

It follows that

$$E|Z_{d,1}|^{\frac{2+t}{2}} = E\left|\frac{(\Delta^*(W) - \theta)^2}{\sigma^2} - 1\right|^{\frac{2+t}{2}} < (C+1)^{\frac{2+t}{2}},$$

with $(2+t)/2 > 1$. Therefore, by Lemma 3 of [ZB22], we conclude that (1.178) holds.

Next, we show (1.179). Let $a_i = \widehat{\Delta}(W_i) - \Delta^*(W_i) - (\widehat{\theta}_{\text{gen}} - \theta)$ and $b_i = \Delta^*(W_i) - \theta$.

Then, it follows from the triangle inequality that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i)-\widehat{\theta}_{\text{gen}})^2-\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\Delta^*(W_i)-\theta)^2\right| \\
&\quad\le \frac{1}{n}\sum_{i\in\mathcal{I}_k}|a_i|\cdot|a_i+2b_i| \overset{(i)}{\le} \left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}a_i^2\right]^{\frac{1}{2}}\cdot\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}(a_i+2b_i)^2\right]^{\frac{1}{2}} \\
&\quad\overset{(ii)}{\le} \left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}a_i^2\right]^{\frac{1}{2}}\cdot\left[\left(\frac{1}{n}\sum_{i\in\mathcal{I}_k}a_i^2\right)^{\frac{1}{2}}+2\left(\frac{1}{n}\sum_{i\in\mathcal{I}_k}b_i^2\right)^{\frac{1}{2}}\right],
\end{aligned}$$

where (i) follows from Cauchy-Schwarz inequality; (ii) follows from Minkowski inequality.

Recall the equation (1.178), we have

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k}b_i^2=\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\Delta^*(W_i)-\theta)^2=\sigma^2(1+o_p(1)).$$

Since, by assumption, $\widehat{\theta}_{\text{gen}}-\theta=O_p(\sigma/\sqrt{N})$ and $\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}=o_p(\sigma)$, we have

$$\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}a_i^2\right]^{\frac{1}{2}}\le\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}+|\widehat{\theta}_{\text{gen}}-\theta|=o_p(\sigma).$$

Therefore,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i)-\widehat{\theta}_{\text{gen}})^2-\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\Delta^*(W_i)-\theta)^2\right| \\
&\qquad=o_p(\sigma)\cdot[o_p(\sigma)+\sigma(1+o_p(1))]=o_p(\sigma^2).
\end{aligned}$$

Now, by (1.178) and (1.179), we have

$$\begin{aligned}
\widehat{\sigma}_{\text{gen}}-\sigma^2&=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i)-\widehat{\theta}_{\text{gen}})^2-\sigma \\
&=\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i)-\widehat{\theta}_{\text{gen}})^2-(\Delta^*(W_i)-\theta)^2+(\Delta^*(W_i)-\theta)^2-\sigma\right) \\
&=o_p(\sigma^2).
\end{aligned}$$

■

1.8.6 Asymptotic theory for Sequential Double Robust Lasso (S-DRL) estimator

In this section, we provide theoretical results for the S-DRL estimator. We consider the following nuisance estimators: $\widehat{\nu}_c(\mathbf{S}) = \mathbf{U}^\top \widehat{\boldsymbol{\alpha}}_c$, $\widehat{\mu}_c(\mathbf{S}) = \mathbf{V}^\top \widehat{\boldsymbol{\beta}}_c$, $\widehat{\pi}_c(\mathbf{S}_1) = g(\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_c)$, and $\widehat{\rho}_c(\mathbf{S}) = g(\mathbf{U}^\top \widehat{\boldsymbol{\delta}}_c)$ for each $c \in \{a, a'\}$, where $\widehat{\boldsymbol{\alpha}}_c$, $\widehat{\boldsymbol{\beta}}_c$, $\widehat{\boldsymbol{\gamma}}_c$, and $\widehat{\boldsymbol{\delta}}_c$ are defined in Section 2.1. Then $\widehat{\psi}_c(\cdot)$, defined as (1.100), can be written as

$$\widehat{\psi}_c(W) = \mathbf{V}^\top \widehat{\boldsymbol{\beta}}_c + \mathbb{1}_{\{A_1=c_1\}} \frac{\mathbf{U}^\top \widehat{\boldsymbol{\alpha}}_c - \mathbf{V}^\top \widehat{\boldsymbol{\beta}}_c}{g(\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_c)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \mathbf{U}^\top \widehat{\boldsymbol{\alpha}}_c}{g(\mathbf{V}^\top \widehat{\boldsymbol{\gamma}}_c) g(\mathbf{U}^\top \widehat{\boldsymbol{\delta}}_c)}.$$

We consider the following target nuisance functions: $\nu_c^*(\mathbf{S}) = \mathbf{U}^\top \boldsymbol{\alpha}_c^*$, $\mu_c^*(\mathbf{S}) = \mathbf{V}^\top \boldsymbol{\beta}_c^*$, $\pi_c^*(\mathbf{S}_1) = g(\mathbf{V}^\top \boldsymbol{\gamma}_c^*)$, and $\rho_c^*(\mathbf{S}) = g(\mathbf{U}^\top \boldsymbol{\delta}_c^*)$ for each $c \in \{a, a'\}$, where $\boldsymbol{\alpha}_c^*$, $\boldsymbol{\beta}_c^*$, $\boldsymbol{\gamma}_c^*$, and $\boldsymbol{\delta}_c^*$ are defined in Section 2.1. Then $\psi_c^*(\cdot)$, defined as (1.101), can be written as

$$\psi_c^*(W) = \mathbf{V}^\top \boldsymbol{\beta}_c^* + \mathbb{1}_{\{A_1=c_1\}} \frac{\mathbf{U}^\top \boldsymbol{\alpha}_c^* - \mathbf{V}^\top \boldsymbol{\beta}_c^*}{g(\mathbf{V}^\top \boldsymbol{\gamma}_c^*)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \mathbf{U}^\top \boldsymbol{\alpha}_c^*}{g(\mathbf{V}^\top \boldsymbol{\gamma}_c^*) g(\mathbf{U}^\top \boldsymbol{\delta}_c^*)}.$$

Auxiliary Lemmas

Lemma 1.16. *(a) Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1-4 hold. Assume that $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$, and either (a) $\|\mathbf{S}_1\|_\infty \leq C$ almost surely, with some constant $C > 0$, or (b) $s_{\boldsymbol{\delta}_a} \log^2(d) = O(N)$. Then,*

$$T_2 = O_p\left(\sigma \frac{s_1 \log(d)}{N} + \sigma \sqrt{\frac{s_2 \log(d)}{N}}\right), \tag{1.180}$$

where T_2 is defined as (1.104) *and*

$$s_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}}\},$$

$$s_2 := \max\left\{s_{\boldsymbol{\alpha}_a} \mathbb{1}_{\{\rho_a^* \neq \rho_a\}}, s_{\boldsymbol{\beta}_a} \mathbb{1}_{\{\pi_a^* \neq \pi_a\}}, s_{\boldsymbol{\gamma}_a} \mathbb{1}_{\{\mu_a^* \neq \mu_a\}}, s_{\boldsymbol{\delta}_a} \mathbb{1}_{\{\nu_a^* \neq \nu_a\}}\right\}.$$

b) Suppose that all the nuisance models $\mu_a^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumptions 1-3 hold. Assume that $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$*, and either (a)* $\|\mathbf{S}_1\|_\infty \leq C$ *almost surely, with some constant* $C > 0$*, or (b)* $s_{\boldsymbol{\delta}_a}\log^2(d) = O(N)$*. Then, Then,*

$$T_2 = O_p\left(\sigma\frac{s_1\log(d)}{N}\right). \tag{1.181}$$

Lemma 1.17. *Suppose that at least one of $\mu_a^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions of Lemma 1.16 (a) hold. Then,*

$$[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right), \tag{1.182}$$

$$T_4 = O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right), \tag{1.183}$$

where T_4 is defined as (1.106)*.*

Proof of Theorem 2

Let $\xi := \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$. Recall the representation (1.102). By Lemmas 1.9, 1.16, 1.11, and 1.17, we have

$$\begin{aligned}
T_1 &= 0,\\
T_2^{(k)} &= O_p\left(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}}\right),\\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right),\\
T_4^{(k)} &= O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).
\end{aligned}$$

for each $k \leq K$. Therefore, by Lemma 1.13 with $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$, we obtain

$$\begin{aligned}\widehat{\theta} - \theta &= K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \\ &= O_p\left(\sigma\frac{s_1\log(d)}{N} + \sigma\sqrt{\frac{s_2\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\right),\end{aligned}$$

with $s_1 := \max\{\sqrt{s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a}s_{\boldsymbol{\gamma}_a}}\}$ and

$$s_2 := \max\left\{s_{\boldsymbol{\alpha}_a}\mathbb{1}_{\{\rho_a^*\neq\rho_a\}}, s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi_a^*\neq\pi_a\}}, s_{\boldsymbol{\gamma}_a}\mathbb{1}_{\{\mu_a^*\neq\mu_a\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu_a^*\neq\nu_a\}}\right\}.$$

Proof of Theorem 3

In this theorem, we consider the setting where all the nuisance models are correctly specified. Note that, Assumption 4 holds under Assumption 1 when all the nuisance models are correct.

Consistency Let $\xi := \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$. Recall the representation (1.102). By Lemmas 1.9, 1.16, 1.11, and 1.17, we have

$$\begin{aligned}T_1 &= 0, \\ T_2^{(k)} &= O_p\left(\sigma\frac{s_1\log(d)}{N}\right), \\ T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \\ T_4^{(k)} &= O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).\end{aligned}$$

for each $k \leq K$. By Assumption, $s_1\log(d) = o(\sqrt{N})$ and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$. Together with Lemma 1.14 , we obtain that

$$\widehat{\theta} - \theta = K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) = O_p\left(\frac{1}{\sqrt{N}}\sigma\right). \tag{1.184}$$

Asymptotic Normality By Lemmas 1.9, 1.16 and 1.17 with $s_1 \log(d) = o(\sqrt{N})$ and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$, 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

$$\sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K} T_3^{(k)} = \sqrt{N}\sigma^{-1}\left(N^{-1}\sum_{i=1}^{N}\Delta^*(W_i) - \theta\right) \rightsquigarrow N(0,1),$$

where $T_3^{(k)}$ is defined as (1.105). Here, $\Delta_{N,i} := \Delta^*(W_i)$ is possibly dependent with N since both W_i and nuisance parameters $(\boldsymbol{\alpha}_c^*, \boldsymbol{\beta}_c^*, \boldsymbol{\gamma}_c^*, \boldsymbol{\delta}_c^*)$ 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, $\{\Delta_{N,i}\}_{N,i}$ 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:

$$\lim_{n\to\infty} \frac{E|\Delta^*(W) - \theta|^{2+t}}{n^{\frac{t}{2}}\sigma^{2+t}} = 0. \tag{1.185}$$

Step 1 In order to check Lyapunov's condition, we show that for some constant C',

$$\frac{E|\Delta^*(W) - \theta|^{2+t}}{\sigma^{2+t}} < C'. \tag{1.186}$$

By Lemma 1.14, we have, for some constants $t > 0$ and $C_t > 0$,

$$\frac{E|\Delta^*(W) - \theta|^{2+t}}{\sigma^{2+t}} \leq \frac{2C_t}{c_0^{4+2t}}\left(\frac{E[|\zeta|^{2+t}}{\sigma^{2+t}} + \frac{E[|\varepsilon|^{2+t}]}{\sigma^{2+t}} + \frac{E|\xi|^{2+t}}{[E|\xi|^2]^{1+\frac{t}{2}}}\right).$$

Let $\boldsymbol{e}_1 = (1, \mathbf{0}_{1\times d_1})^\top$, then we write $\xi = \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta = \mathbf{V}^\top(\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta)$. By Lemma 1.4, under Assumption 3, we have

$$\|\xi\|_{\psi_2} = \|(\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta)^\top\mathbf{V}\|_{\psi_2} \le 2\sigma_u\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2.$$

It follows from Lemma 1.2 that

$$E[|\xi|^{2+t}] \le 2^{3+t}\sigma_u^{2+t}\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2^{2+t}\Gamma(2+t/2). \tag{1.187}$$

Similarly, by Assumption 2, we have

$$E[|\zeta|^{2+t}] \le 2^{3+t}\sigma^{2+t}\sigma_\zeta^{2+t}\Gamma(2+t/2), \tag{1.188}$$

$$E[|\varepsilon|^{2+t}] \le 2^{3+t}\sigma^{2+t}\sigma_\varepsilon^{2+t}\Gamma(2+t/2). \tag{1.189}$$

By Lemma 1.4, under Assumption 3, we also have

$$\begin{aligned} E[|\xi|^2] &= E[|\mathbf{V}^\top(\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta)|^2] \ge \|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2^2 \cdot \lambda_{\min}(E[\mathbf{V}\mathbf{V}^\top]) \\ &\ge \kappa_l\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2^2. \end{aligned} \tag{1.190}$$

Using (1.187) and (1.190), we get that

$$\frac{E|\xi|^{2+t}}{[E|\xi|^2]^{1+\frac{t}{2}}} \le \frac{2^{3+t}\sigma_u^{2+t}\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{a'}^* - \boldsymbol{e}_1\theta\|_2^{2+t}} = \frac{2^{3+t}\sigma_u^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}}. \tag{1.191}$$

Using (1.188), (1.189) and (1.191), then we obtain that

$$\begin{aligned} &\frac{E|\Delta^*(W) - \theta|^{2+t}}{\sigma^{2+t}} \\ &\le \frac{2C_t}{c_0^{4+2t}}\left(2^{3+t}\sigma_\zeta^{2+t}\Gamma(2+t/2) + 2^{3+t}\sigma_\varepsilon^{2+t}\Gamma(2+t/2) + \frac{2^{3+t}\sigma_u^{2+t}\Gamma(2+t/2)}{\kappa_l^{1+t/2}}\right) =: C'. \end{aligned}$$

That is, (1.186) holds and hence the Lyapunov's condition is satisfied.

Step 2 Now, we show the consistency of the asymptotic variance's estimator. In this step, the expectations are taken w.r.t. the joint distribution of $(W_i)_{i\in\mathcal{I}_k}$. By (1.184), we have $\widehat{\theta}-\theta=O_p(\sigma/\sqrt{N})$. Then, we show, for each $k\leq K$,

$$\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}=o_p(\sigma). \tag{1.192}$$

It follows from Jensen's inequality that

$$\begin{aligned}E\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}&\leq\left\{E\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]\right\}^{\frac{1}{2}}\\&=[E|\widehat{\Delta}(W)-\Delta^*(W)|^2]^{\frac{1}{2}}\overset{(i)}{=}O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a},s_{\boldsymbol{\beta}_a},s_{\boldsymbol{\gamma}_a},s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right),\end{aligned}$$

where (i) follows from (1.182) in Lemma 1.17 with correctly specified nuisance models. By Markov's inequality with $\max\{s_{\boldsymbol{\alpha}_a},s_{\boldsymbol{\beta}_a},s_{\boldsymbol{\gamma}_a},s_{\boldsymbol{\delta}_a}\}\log(d)=o(N)$, we have

$$\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i)-\Delta^*(W_i)|^2\right]^{\frac{1}{2}}=O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a},s_{\boldsymbol{\beta}_a},s_{\boldsymbol{\gamma}_a},s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right)=o_p(\sigma).$$

Therefore, using (1.184), (1.186) and (1.192), we get $\widehat{\sigma}^2-\sigma^2=o_p(\sigma^2)$ by Lemma 1.15.

Proofs of Auxiliary Lemmas

Proof of Lemma 1.16. In this proof, the expectations are taken w.r.t. the distribution of a new observation W. Recall the representation (1.133) that $T_2=E[\widehat{\Delta}(W)-\Delta^*(W)]=\sum_{i=1}^6E[Q_i]$. Here, we first upper bound $E[Q_1+Q_2+Q_3]$. Same as in the proof of Lemma 1.10, we also have (1.134) holds, with R_is defined in (1.135)-(1.142). Same as in (1.141) and (1.142), we have $R_7=R_8=0$. Now, we obtain an upper bound for R_i ($i\in\{1,\ldots,6\}$). For

$R_1 + R_2$, since $|A_1| \leq 1$, $|\pi_a^*(\mathbf{S}_1)| \leq 1$ and $|\rho_a^*(\mathbf{S})| \leq 1$, we have

$$\begin{aligned} R_1 + R_2 &\overset{(i)}{\leq} \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \left\{ E \left| \frac{1}{\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\rho_a^*(\mathbf{S})} \right|^3 \right\}^{\frac{1}{3}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^3\}^{\frac{1}{3}} \\ &\quad + \left\{ E \left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} \{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \\ &\overset{(ii)}{=} O_p\bigg(\frac{\sigma\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}} \log(d)}{N} + \frac{\sigma\sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}} \log(d)}{N} \\ &\quad + \frac{\sigma\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}_a}} \log(d)}{N} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} + \frac{\sigma\sqrt{s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\gamma}_a}} \log(d)}{N} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \bigg), \end{aligned} \tag{1.193}$$

where (i) holds by Hölder's inequality; (ii) follows from Lemmas 1.5, 1.8 and Theorem 9 with $s_{\boldsymbol{\gamma}_a} \log(d) = o(N)$ and $d_1 \asymp d$. Similarly, for $R_3 + R_4$, since $|A_1| \leq 1$, $|\rho_a^*(\mathbf{S}) - \rho_a(\mathbf{S})| \leq 1$, $|\pi_a^*(\mathbf{S}_1) - \pi_a(\mathbf{S}_1)| \leq 1$,

$$\begin{aligned} R_3 + R_4 \leq & \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\rho}_a(\mathbf{S})|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^3\}^{\frac{1}{3}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \\ & + \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-2}\}^{\frac{1}{2}} \{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot)\}} \\ \overset{(i)}{=} & O_p\bigg(\sigma\sqrt{\frac{s_{\boldsymbol{\beta}_a} \log(d)}{N}} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot)\}} + \sigma\sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log(d)}{N}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \\ & + \frac{\sigma\sqrt{s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}} \log(d)}{N} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot)\}} + \sigma\sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot), \nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \bigg), \end{aligned} \tag{1.194}$$

where (i) holds by Lemmas 1.5, 1.8 and Theorem 9 with $d_1 \asymp d$. For $R_5 + R_6$, since $|\rho_a^*(\mathbf{S})| \leq 1$,

$$\begin{aligned} R_5 + R_6 &\leq \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}} \left\{ E \left| \frac{1}{\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\rho_a^*(\mathbf{S})} \right|^4 \right\}^{\frac{1}{4}} \{E[A_1|\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})|^2]\}^{\frac{1}{2}} \\ &\quad + \left\{ E \left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} E[A_1|\{\mu_a^*(\mathbf{S}_1) - \mu_a(\mathbf{S}_1)|^2]\}^{\frac{1}{2}} \\ &\overset{(i)}{=} O_p\left(\sigma\sqrt{\frac{s_{\boldsymbol{\gamma}_a} \log(d)}{N}} \mathbb{1}_{\{\mu_a^*(\cdot) \neq \mu_a(\cdot)\}} + \sigma\sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \right). \end{aligned} \tag{1.195}$$

where (i) follows from Lemma 1.8, (1.149), (1.151), and Lemma 1.13. Combining (1.193)-(1.195) with $s_{\boldsymbol{\gamma}_a} \log(d) = o(N)$, we have

$$E[Q_1 + Q_2 + Q_3] = O_p\left(\sigma \frac{s_1 \log(d)}{N} + \sigma\sqrt{\frac{s_2 \log(d)}{N}}\right).$$

Analogously to $E[Q_1 + Q_2 + Q_3]$, we have the same result for $E[Q_4 + Q_5 + Q_6]$. Theorefore, (1.180) follows.

(b) When all the models are correctly specified, we have $s_2 = 0$. Hence, by part (a), (1.181) holds. ∎

Proof of Lemma 1.17. In this proof, the expectations are taken w.r.t. a new observation W, unless stated otherwise. We first show that (1.182) holds. Recall the representation (1.126), by Minkowski inequality, we have

$$[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} \leq \sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}},$$

where Q_i ($i \in \{1, \ldots, 6\}$) are defined as(1.127)-(1.132). In the following, we show that

$$\sum_{i=1}^{6} [E(Q_i^2)]^{\frac{1}{2}} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d)}{N}}\right).$$

By Minkowski's inequality,

$$
\begin{aligned}
[E(Q_1^2)]^{\frac{1}{2}} \leq & \left\{ E\left[\frac{A_1 A_2}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} (\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})) \right]^2 \right\}^{\frac{1}{2}} \\
& + \left\{ E\left[\left(\frac{A_1 A_2}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} - \frac{A_1 A_2}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \right) (Y - \nu_a^*(\mathbf{S})) \right]^2 \right\}^{\frac{1}{2}} \\
\overset{(i)}{\leq} & \left\{ E\left[\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} (\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})) \right]^2 \right\}^{\frac{1}{2}} \\
& + \left\{ E\left[\left(\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \right) \zeta \right]^2 \right\}^{\frac{1}{2}} \\
\overset{(ii)}{\leq} & \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-6}\}^{\frac{1}{6}} \{E|\widehat{\rho}_a(\mathbf{S})|^{-6}\}^{\frac{1}{6}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^6\}^{\frac{1}{6}} \\
& + \{E|\zeta|^4\}^{\frac{1}{4}} \left\{ E\left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\pi_a^*(\mathbf{S}_1)\rho_a^*(\mathbf{S})} \right|^4 \right\}^{\frac{1}{4}} \\
\overset{(iii)}{=} & O_p\left(\sigma \sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d)}{N}} \right),
\end{aligned} \tag{1.196}
$$

where (i) holds by the fact that $|A_1| \leq 1$, $|A_2| \leq 1$ and $A_1 A_2 \zeta = \zeta_a = A_1 A_2 (Y - \nu_a^*(\mathbf{S}))$; (ii) holds by Hölder's inequality; (iii) follows from Lemmas 1.5, 1.8, and under Assumption 2, by Lemma 1.2,

$$
E[|\zeta|^4] \leq 8\sigma^4 \sigma_\zeta^4, \quad E[|\varepsilon|^4] \leq 8\sigma^4 \sigma_\varepsilon^4. \tag{1.197}
$$

Then, similarly as above, we obtain

$$
\begin{aligned}
[E(Q_2^2)]^{\frac{1}{2}} &\leq \left\{ E\left[\frac{A_1}{\widehat{\pi}_a(\mathbf{S}_1)} (\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})) \right]^2 \right\}^{\frac{1}{2}} + \left\{ E\left[\frac{A_1}{\widehat{\pi}_a(\mathbf{S}_1)} (\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)) \right]^2 \right\}^{\frac{1}{2}} \\
&\quad + \left\{ E\left[\left(\frac{A_1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{A_1}{\pi_a^*(\mathbf{S}_1)} \right) (\nu_a^*(\mathbf{S}) - \mu_a^*(\mathbf{S}_1)) \right]^2 \right\}^{\frac{1}{2}} \\
&\leq \left\{ E\left[\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} (\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})) \right]^2 \right\}^{\frac{1}{2}} + \left\{ E\left[\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} (\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)) \right]^2 \right\}^{\frac{1}{2}} \\
&\quad + \left\{ E\left[\left(\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right) \varepsilon \right]^2 \right\}^{\frac{1}{2}} \\
&\leq \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^4\}^{\frac{1}{4}} + \left\{ E\left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^4 \right\}^{\frac{1}{4}} \{E|\varepsilon|^4\}^{\frac{1}{4}} \\
&\quad + \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}} \{E|\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)|^4\}^{\frac{1}{4}} \\
&\overset{(i)}{=} O_p\left(\sigma \sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}\} \log(d)}{N}} + \sigma \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \right), \qquad (1.198)
\end{aligned}
$$

where (i) follows from Lemmas 1.5, 1.8, (1.197), and Theorem 9 with $s_{\boldsymbol{\delta}_a} \log(d) = o(N)$ and $d_1 \asymp d$. By Theorem 9 , we also have

$$
\begin{aligned}
[E(Q_3^2)]^{\frac{1}{2}} &= \left\{ E[\widehat{\mu}_a(\mathbf{S}_1) - \mu_a^*(\mathbf{S}_1)]^2 \right\}^{1/2} \\
&= O_p\left(\sigma \sqrt{\frac{s_{\boldsymbol{\beta}_a} \log(d_1)}{N}} + \frac{\sigma \sqrt{s_{\boldsymbol{\delta}_a} s_{\boldsymbol{\alpha}_a}} \log(d)}{N} \right. \\
&\quad \left. + \sigma \sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log(d)}{N}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} + \sigma \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \right). \qquad (1.199)
\end{aligned}
$$

Combining (1.196)-(1.199), we have

$$
[E(Q_1^2)]^{\frac{1}{2}} + [E(Q_2^2)]^{\frac{1}{2}} + [E(Q_3^2)]^{\frac{1}{2}} = O_p\left(\sigma \sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d)}{N}} \right).
$$

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, (1.182) holds. Now, we show (1.183). Recall the definition (1.106), we have

$T_4 := n^{-1}\sum_{i\in\mathcal{I}_k}[\widehat{\Delta}(W_i) - \Delta^*(W_i)] - E[\widehat{\Delta}(W) - \Delta^*(W)]$. By Chebyshev's inequality, we have for any $t > 0$,

$$P(|T_4| > t) \leq \frac{1}{t^2}\mathrm{Var}\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}(\widehat{\Delta}(W_i) - \Delta^*(W_i))\right] \tag{1.200}$$
$$\leq \frac{1}{nt^2}E[\widehat{\Delta}(W) - \Delta^*(W)]^2.$$

In the righ-hand side of (1.200), the variance is taken over the joint distribution of $(W_i)_{i\in\mathcal{I}_k}$. Note that, based on the sample-splitting, the nuisance estimates are independent of $(W_i)_{i\in\mathcal{I}_k}$. Together with (1.182), we conclude that (1.183) holds. ∎

1.8.7 Asymptotic theory for Dynamic Treatment Lasso (DTL) estimator

In this section, we provide theoretical results for the DTL estimator. The ℓ_1-regularized nuisance estimates $\widehat{\boldsymbol{\alpha}}_c$, $\widehat{\boldsymbol{\gamma}}_c$, $\widehat{\boldsymbol{\delta}}_c$ and the target nuisance estimates $\boldsymbol{\alpha}_c^*$, $\boldsymbol{\gamma}_c^*$, $\boldsymbol{\delta}_c^*$ are the same as in Section 1.8.6. For the first-time conditional mean function, we consider the nested-regression-based estimator $\widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}$ defined in Section 2.2. With a slight abuse of notation, we consider set the general nuisance estimates as $\widehat{\nu}_c(\mathbf{S}) = \mathbf{U}^\top\widehat{\boldsymbol{\alpha}}_c$, $\widehat{\mu}_c(\mathbf{S}) = \widehat{\mu}_{c,\mathrm{NR}}(\mathbf{S}) = \mathbf{V}^\top\widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}$, $\widehat{\pi}_c(\mathbf{S}_1) = g(\mathbf{V}^\top\widehat{\boldsymbol{\gamma}}_c)$, and $\widehat{\rho}_c(\mathbf{S}) = g(\mathbf{U}^\top\widehat{\boldsymbol{\delta}}_c)$ for each $c \in \{a, a'\}$. Then $\widehat{\psi}_c(\cdot)$, defined as (1.100), can be written as

$$\widehat{\psi}_c(W) = \mathbf{V}^\top\widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}} + \mathbb{1}_{\{A_1=c_1\}}\frac{\mathbf{U}^\top\widehat{\boldsymbol{\alpha}}_c - \mathbf{V}^\top\widehat{\boldsymbol{\beta}}_{c,\mathrm{NR}}}{g(\mathbf{V}^\top\widehat{\boldsymbol{\gamma}}_c)} + \mathbb{1}_{\{A_1=c_1,A_2=c_2\}}\frac{Y - \mathbf{U}^\top\widehat{\boldsymbol{\alpha}}_c}{g(\mathbf{V}^\top\widehat{\boldsymbol{\gamma}}_c)g(\mathbf{U}^\top\widehat{\boldsymbol{\delta}}_c)}.$$

With a slight abuse of notations, we set the general working models as $\nu_c^*(\mathbf{S}) = \mathbf{U}^\top\boldsymbol{\alpha}_c^*$, $\mu_c^*(\mathbf{S}) = \mu_{c,\mathrm{NR}}^*(\mathbf{S}) = \mathbf{V}^\top\boldsymbol{\beta}_{c,\mathrm{NR}}^*$, $\pi_c^*(\mathbf{S}_1) = g(\mathbf{V}^\top\boldsymbol{\gamma}_c^*)$, and $\rho_c^*(\mathbf{S}) = g(\mathbf{U}^\top\boldsymbol{\delta}_c^*)$ for each $c \in \{a, a'\}$.

Then $\psi_c^*(\cdot)$, defined as (1.101), can be written as

$$\psi_c^*(W) = \mathbf{V}^\top \boldsymbol{\beta}_{c,\mathrm{NR}}^* + \mathbb{1}_{\{A_1=c_1\}} \frac{\mathbf{U}^\top \boldsymbol{\alpha}_c^* - \mathbf{V}^\top \boldsymbol{\beta}_{c,\mathrm{NR}}^*}{g(\mathbf{V}^\top \boldsymbol{\gamma}_c^*)} + \mathbb{1}_{\{A_1=c_1, A_2=c_2\}} \frac{Y - \mathbf{U}^\top \boldsymbol{\alpha}_c^*}{g(\mathbf{V}^\top \boldsymbol{\gamma}_c^*) g(\mathbf{U}^\top \boldsymbol{\delta}_c^*)}.$$

Auxiliary Lemmas

Lemma 1.18. *(a) Suppose that at least one of $\mu_{a,NR}^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1-4 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$*. Then,*

$$T_2 = O_p\left(\sigma \frac{s_1' \log(d)}{N} + \sigma \sqrt{\frac{s_2' \log(d)}{N}}\right), \tag{1.201}$$

where T_2 is defined as (1.104) *and*

$$s_1' := \max\{\sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\gamma}_a}}, \sqrt{s_{\boldsymbol{\alpha}_a} s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a} s_{\boldsymbol{\gamma}_a}}\},$$

$$s_2' := \max\left\{ s_{\boldsymbol{\alpha}_a} \mathbb{1}_{\{\pi_a^* \neq \pi_a \text{ or } \rho_a^* \neq \rho_a\}}, s_{\boldsymbol{\beta}_a} \mathbb{1}_{\{\pi_a^* \neq \pi_a\}}, s_{\boldsymbol{\gamma}_a} \mathbb{1}_{\{\mu_{a,NR}^* \neq \mu_a\}}, s_{\boldsymbol{\delta}_a} \mathbb{1}_{\{\nu_a^* \neq \nu_a\}} \right\}.$$

(b) Suppose that all the nuisance models $\mu_{a,NR}^*(\cdot)$, $\nu_a^*(\cdot)$, $\pi_a^*(\cdot)$, and $\rho_a^*(\cdot)$ are correctly specified. Let Assumptions 1-3 hold. Assume that $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$ *Then,*

$$T_2 = O_p\left(\sigma \frac{s_1' \log(d)}{N}\right). \tag{1.202}$$

Lemma 1.19. *Suppose that at least one of $\mu_{a,NR}^*(\cdot)$ and $\pi_a^*(\cdot)$ is correctly specified, and at least one of the models $\nu_a^*(\cdot)$ and $\rho_a^*(\cdot)$ is correctly specified. Let Assumptions 1-4 hold. Assume that* $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d) = o(N)$*. Then,*

$$[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} = O_p\left(\sigma \sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d)}{N}}\right), \tag{1.203}$$

$$T_4 = O_p\left(\sigma \frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\} \log(d)}}{N}\right), \tag{1.204}$$

where T_4 is defined as (1.106)*.*

Proof of Theorem 4

Let $\xi = \mu_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu_{a',\mathrm{NR}}(\mathbf{S}_1) - \theta$. Recall the representation (1.102). By Lemmas 1.9, 1.18, 1.11, and 1.19, we have

$$\begin{aligned}
T_1 &= 0, \\
T_2^{(k)} &= O_p\left(\sigma\frac{s_1'\log(d)}{N} + \sigma\sqrt{\frac{s_2'\log(d)}{N}}\right), \\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \\
T_4^{(k)} &= O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).
\end{aligned}$$

for each $k \leq K$. Therefore, by Lemma 1.13 with $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$, we obtain that

$$\begin{aligned}
\widehat{\theta}_{\mathrm{DTL}} - \theta &= K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \\
&= O_p\left(\sigma\frac{s_1'\log(d)}{N} + \sigma\sqrt{\frac{s_2'\log(d)}{N}} + \frac{1}{\sqrt{N}}\sigma\right),
\end{aligned}$$

where

$$\begin{aligned}
s_1' &:= \max\{\sqrt{s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\gamma}_a}}, \sqrt{s_{\boldsymbol{\alpha}_a}s_{\boldsymbol{\delta}_a}}, \sqrt{s_{\boldsymbol{\beta}_a}s_{\boldsymbol{\gamma}_a}}\}, \\
s_2' &:= \max\left\{s_{\boldsymbol{\alpha}_a}\mathbb{1}_{\{\pi_a^*\neq\pi_a \text{ or } \rho_a^*\neq\rho_a\}}, s_{\boldsymbol{\beta}_a}\mathbb{1}_{\{\pi_a^*\neq\pi_a\}}, s_{\boldsymbol{\gamma}_a}\mathbb{1}_{\{\mu_{a,\mathrm{NR}}^*\neq\mu_a\}}, s_{\boldsymbol{\delta}_a}\mathbb{1}_{\{\nu_a^*\neq\nu_a\}}\right\}.
\end{aligned}$$

Proof of Theorem 5

In this theorem, we consider the setting where all the nuisance models are correctly specified. Note that, Assumption 4 holds under Assumption 1 when all the nuisance models are correct.

Consistency Let $\xi = \mu_a(\mathbf{S}_1) - \mu_{a'}(\mathbf{S}_1) - \theta$. Recall the representation (1.102), by Lemmas 1.9, 1.18, 1.11, and 1.19 in that order we have

$$\begin{aligned}
T_1 &= 0, \\
T_2^{(k)} &= O_p\left(\sigma \frac{s_1' \log(d)}{N}\right), \\
T_3^{(k)} &= O_p\left(\frac{1}{\sqrt{N}}\left[\sqrt{E[\zeta^2]} + \sqrt{E[\varepsilon^2]} + \sqrt{E[\xi^2]}\right]\right), \\
T_4^{(k)} &= O_p\left(\sigma \frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).
\end{aligned}$$

for each $k \leq K$. By Assumption, $s_1' \log(d) = o(\sqrt{N})$ and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$. Together with Lemma 1.14 , we obtain that

$$\widehat{\theta}_{\text{DTL}} - \theta = K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) = O_p\left(\frac{1}{\sqrt{N}}\sigma\right). \tag{1.205}$$

Asymptotic Normality By Lemmas 1.9, 1.18, and 1.19 with $s_1' \log(d) = o(\sqrt{N})$ and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$, we have

$$\sqrt{n}\sigma^{-1}(T_1 + T_2^{(k)} + T_4^{(k)}) = o_p(1)$$

for each $k \leq K$. In addition, repeating Section 1.8.6 in the proof of Theorem 3, we also have

$$\sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K} T_3^{(k)} \rightsquigarrow N(0,1),$$

which implies,

$$\sqrt{N}\sigma^{-1}(\widehat{\theta}_{\text{DTL}} - \theta) = \sqrt{N}\sigma^{-1}K^{-1}\sum_{k=1}^{K}(T_1 + T_2^{(k)} + T_3^{(k)} + T_4^{(k)}) \rightsquigarrow N(0,1).$$

By (1.203) and $\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d) = o(N)$, we have

$$\left[\frac{1}{n}\sum_{i\in\mathcal{I}_k}|\widehat{\Delta}(W_i) - \Delta^*(W_i)|^2\right]^{\frac{1}{2}} = o_p(\sigma).$$

Together with (1.205) and (1.186), we have $\widehat{\sigma}^2_{\text{DTL}} - \sigma^2 = o_p(\sigma^2)$ by Lemma 1.15.

Proofs of Auxiliary Lemmas

Proof of Lemma 1.18. In this proof, the expectations are taken w.r.t. the distribution of a new observation W. We repeat the proof of Lemma 1.16, except here we consider the nested-regression-based estimate $\widehat{\mu}_c(\cdot) = \widehat{\mu}_{c,\mathrm{NR}}(\cdot)$ and the corresponding target $\mu_c^*(\cdot) = \mu_{c,\mathrm{NR}}^*(\cdot)$. Note that

$$T_2 = E[\widehat{\Delta}(W) - \Delta^*(W)] = \sum_{i=1}^{6} E[Q_i], \tag{1.206}$$

with $E[Q_1 + Q_2 + Q_3] = \sum_{i=1}^{6} R_i$, where R_is are defined in (1.135)-(1.140). For $R_1 + R_2$, we have

$$\begin{aligned} R_1 + R_2 &\overset{(i)}{\leq} \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \left\{ E\left| \frac{1}{\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\rho_a^*(\mathbf{S})} \right|^3 \right\}^{\frac{1}{3}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^3\}^{\frac{1}{3}} \\ &\quad + \left\{ E\left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} \{E|\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \\ &\overset{(ii)}{=} O_p\left(\sigma \frac{s_1' \log(d)}{N} \right), \end{aligned} \tag{1.207}$$

where (i) holds by Hölder's inequality; (ii) follows from Lemmas 1.5, 1.8 and Theorem 10 with $d_1 \asymp d$. Similarly, for $R_3 + R_4$,

$$\begin{aligned} R_3 + R_4 \leq & \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\rho}_a(\mathbf{S})|^{-3}\}^{\frac{1}{3}} \{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^3\}^{\frac{1}{3}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \\ & + \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-2}\}^{\frac{1}{2}} \{E|\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}^*(\mathbf{S}_1)|^2\}^{\frac{1}{2}} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot)\}} \\ = & O_p\left(\sigma \sqrt{\frac{(s_{\boldsymbol{\alpha}_a} + s_{\boldsymbol{\beta}_a}) \log(d)}{N}} \mathbb{1}_{\{\pi_a^*(\cdot) \neq \pi_a(\cdot)\}} + \sigma \sqrt{\frac{s_{\boldsymbol{\alpha}_a} \log(d)}{N}} \mathbb{1}_{\{\rho_a^*(\cdot) \neq \rho_a(\cdot)\}} \right). \end{aligned} \tag{1.208}$$

Repeating the similar procedure of (1.151), we have

$$E[A_1(\mu_{a,\mathrm{NR}}^*(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}(\mathbf{S}_1))^2] \leq \frac{2}{c_0} E[\zeta^2] + 2E[\varepsilon^2]. \tag{1.209}$$

For $R_5 + R_6$,

$$R_5 + R_6 \leq \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}} \left\{ E \left| \frac{1}{\widehat{\rho}_a(\mathbf{S})} - \frac{1}{\rho_a^*(\mathbf{S})} \right|^4 \right\}^{\frac{1}{4}} \{E[A_1|\nu_a^*(\mathbf{S}) - \nu_a(\mathbf{S})|^2]\}^{\frac{1}{2}}$$
$$+ \left\{ E \left| \frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)} \right|^2 \right\}^{\frac{1}{2}} \{E[A_1|\mu_{a,\mathrm{NR}}^*(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}(\mathbf{S}_1)|^2]\}^{\frac{1}{2}}$$
$$\overset{(i)}{=} O_p \left(\sigma \sqrt{\frac{s_{\boldsymbol{\gamma}_a} \log(d)}{N}} \mathbb{1}_{\{\mu_{a,\mathrm{NR}}^*(\cdot) \neq \mu_{a,\mathrm{NR}}(\cdot)\}} + \sigma \sqrt{\frac{s_{\boldsymbol{\delta}_a} \log(d)}{N}} \mathbb{1}_{\{\nu_a^*(\cdot) \neq \nu_a(\cdot)\}} \right), \quad (1.210)$$

where (i) follows from Lemma 1.8, (1.149), (1.209), and Lemma 1.13. Combining (1.207)-(1.210), we have

$$E[Q_1 + Q_2 + Q_3] = \sum_{i=1}^{6} R_i = O_p \left(\sigma \frac{s_1' \log(d)}{N} + \sigma \sqrt{\frac{s_2' \log(d)}{N}} \right).$$

Note that $E[Q_4 + Q_5 + Q_6]$ can be controlled similarly as $E[Q_1 + Q_2 + Q_3]$. By (1.206), we have (1.201) holds.

(b) When all the models are correctly specified, we have $s_2' = 0$. Hence, by part (a), (1.202) holds. ∎

Proof of Lemma 1.19. In this proof, the expectations are taken w.r.t. a new observation W, unless stated otherwise. We repeat the proof of Lemma 1.17, except here we consider the nested-regression-based estimate $\widehat{\mu}_c(\cdot) = \widehat{\mu}_{c,\mathrm{NR}}(\cdot)$ and the corresponding target $\mu_c^*(\cdot) = \mu_{c,\mathrm{NR}}^*(\cdot)$. Note that the estimation error of $\mu_c^*(\cdot)$ only appears in steps (1.198) and (1.199) when controlling the terms $[E(Q_2^2)]^{1/2}$ and $[E(Q_3^2)]^{1/2}$. By Lemmas 1.5, 1.8, (1.197), and

Theorem 10 with with $d_1 \asymp d$, we have

$$
\begin{aligned}
&[E(Q_2^2)]^{\frac{1}{2}} \\
&\quad \leq \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}}\{E|\widehat{\nu}_a(\mathbf{S}) - \nu_a^*(\mathbf{S})|^4\}^{\frac{1}{4}} + \left\{E\left|\frac{1}{\widehat{\pi}_a(\mathbf{S}_1)} - \frac{1}{\pi_a^*(\mathbf{S}_1)}\right|^4\right\}^{\frac{1}{4}}\{E|\varepsilon|^4\}^{\frac{1}{4}} \\
&\qquad + \{E|\widehat{\pi}_a(\mathbf{S}_1)|^{-4}\}^{\frac{1}{4}}\{E|\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}^*(\mathbf{S}_1)|^4\}^{\frac{1}{4}} \\
&\quad = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}\}\log(d)}{N}}\right).
\end{aligned}
$$

By Theorem 10, we also have

$$
[E(Q_3^2)]^{\frac{1}{2}} = \left\{E[\widehat{\mu}_{a,\mathrm{NR}}(\mathbf{S}_1) - \mu_{a,\mathrm{NR}}^*(\mathbf{S}_1)]^2\right\}^{1/2} = O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}\}\log(d)}{N}}\right).
$$

Repeating the remaining steps of the proof of Lemma 1.17, we have

$$
\begin{aligned}
[E(\widehat{\Delta}(W) - \Delta^*(W))^2]^{\frac{1}{2}} &= O_p\left(\sigma\sqrt{\frac{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}{N}}\right), \\
T_4 &= O_p\left(\sigma\frac{\sqrt{\max\{s_{\boldsymbol{\alpha}_a}, s_{\boldsymbol{\beta}_a}, s_{\boldsymbol{\gamma}_a}, s_{\boldsymbol{\delta}_a}\}\log(d)}}{N}\right).
\end{aligned}
$$

■

1.8.8 Proof of the results for multi-stage treatment estimation with DR methods

Proof of Theorem 11. By construction, we have $\mu_{T+1}^*(\overline{\mathbf{S}}_{T+1}, \overline{a}_T) = Y$, and it follows that

$$
\begin{aligned}
&E\left[\sum_{r=t+1}^{T}\frac{\prod_{l=t+1}^{r}\mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r}\pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(\mu_{r+1}^*(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) - \mu_r^*(\overline{\mathbf{S}}_r, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right] \\
&\qquad := H_T + \sum_{r=t+1}^{T-1}(H_{r,1} + H_{r,2} + H_{r,3}), \qquad (1.211)
\end{aligned}
$$

where

$$H_T = E\left[\frac{\prod_{l=t+1}^{T} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{T} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(Y - \mu_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right],$$

and for any $r \in \{t+1, \ldots, T\}$,

$$H_{r,1} = E\left[\frac{\prod_{l=t+1}^{r} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(\mu_{r+1}^*(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) - \mu_{r+1}(\overline{\mathbf{S}}_{r+1}, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right],$$
$$H_{r,2} = E\left[\frac{\prod_{l=t+1}^{r} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(\mu_{r+1}(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) - \mu_r(\overline{\mathbf{S}}_r, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right],$$
$$H_{r,3} = E\left[\frac{\prod_{l=t+1}^{r} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(\mu_r(\overline{\mathbf{S}}_r, \overline{a}_T) - \mu_r^*(\overline{\mathbf{S}}_r, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right].$$

Define $\widetilde{A}_r = (A_{t+1}, A_{t+2}, \ldots, A_r)$ and $\widetilde{a}_r = (a_{t+1}, a_{t+2}, \ldots, a_r)$ for $t+1 \le r \le T$. For H_T, by the tower rule with $Y(\bar{A}_T) = Y$ under Assumption 6, we have

$$\begin{aligned}
H_T &= E\Bigg[E\Bigg[\frac{\prod_{l=t+1}^{T} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{T} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(Y(\overline{a}_T) - \mu_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)) \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}\Bigg] \\
&\qquad \cdot P(\widetilde{A}_{T-1} = \widetilde{a}_{T-1} \mid \overline{\mathbf{S}}_T, \bar{A}_t = \overline{a}_t) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\Bigg] \\
&\overset{(i)}{=} E\Bigg[\frac{E[\mathbb{1}_{\{A_T=a_T\}} \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}]}{\prod_{l=t+1}^{T} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}] - \mu_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)) \\
&\qquad \cdot E(\mathbb{1}_{\{\widetilde{A}_{T-1}=\widetilde{a}_{T-1}\}} \mid \overline{\mathbf{S}}_T, \bar{A}_t = \overline{a}_t) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\Bigg] \\
&\overset{(ii)}{=} E\left[\frac{\prod_{l=t+1}^{T-1} \mathbb{1}_{\{A_l=a_l\}} \pi_T(\overline{\mathbf{S}}_T, \overline{a}_T)}{\prod_{l=t+1}^{T-1} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l) \pi_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)}(\mu_T(\overline{\mathbf{S}}_T, \overline{a}_T) - \mu_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right] \\
&\overset{(iii)}{=} E\left[\frac{\prod_{l=t+1}^{T-1} \mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{T-1} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)}(\mu_T(\overline{\mathbf{S}}_T, \overline{a}_T) - \mu_T^*(\overline{\mathbf{S}}_T, \overline{a}_T)) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t\right] = -H_{T-1,1}, \quad (1.212)
\end{aligned}$$

where (i) holds since $Y(\overline{a}_T) \perp\!\!\!\perp A_T \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}$ under the Assumption 6; (ii) holds since $\pi_T(\overline{\mathbf{S}}_T, \overline{a}_T) = P[A_T = a_T \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}]$ and $\mu_T(\overline{\mathbf{S}}_T, \overline{a}_T) = E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_T, \bar{A}_{T-1} = \overline{a}_{T-1}]$ under the Assumption 6; (iii) holds since either $\pi_T^*(\cdot, \overline{a}_T) = \pi_T(\cdot, \overline{a}_T)$ or $\mu_T^*(\cdot, \overline{a}_T) = \mu_T(\cdot, \overline{a}_T)$

by assumption. For $H_{r,2}$, by the tower rule, we have

$$\begin{aligned} H_{r,2} = E\Bigg[E\Bigg[\frac{\prod_{l=t+1}^{r} \mathbb{1}_{\{A_l = a_l\}}}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_l)} (\mu_{r+1}(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) - \mu_r(\overline{\mathbf{S}}_r, \overline{a}_T)) \mid \overline{\mathbf{S}}_r, \bar{A}_r = \overline{a}_r \Bigg] \\ \cdot P(\widetilde{A}_r = \widetilde{a}_r \mid \overline{\mathbf{S}}_r, \bar{A}_t = \overline{a}_t) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t \Bigg] \\ = E\Bigg[\frac{E[\mu_{r+1}(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) \mid \overline{\mathbf{S}}_r, \bar{A}_r = \overline{a}_r] - \mu_r(\overline{\mathbf{S}}_r, \overline{a}_T)}{\prod_{l=t+1}^{r} \pi_l^*(\overline{\mathbf{S}}_l, \overline{a}_{l-1})} \\ \cdot P(\widetilde{A}_r = \widetilde{a}_r \mid \overline{\mathbf{S}}_r, \bar{A}_t = \overline{a}_t) \mid \overline{\mathbf{S}}_t, \bar{A}_t = \overline{a}_t \Bigg]. \end{aligned} \tag{1.213}$$

For any $r \in \{1, \ldots, T\}$, we have

$$\begin{aligned} \mu_r(\overline{\mathbf{S}}_r, \overline{a}_T) &= E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_r, \bar{A}_{r-1} = \overline{a}_{r-1}] \overset{(i)}{=} E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_r, \bar{A}_r = \overline{a}_r] \\ &\overset{(ii)}{=} E[E[Y(\overline{a}_T) \mid \overline{\mathbf{S}}_{r+1}, \bar{A}_r = \overline{a}_r] \mid \overline{\mathbf{S}}_r, \bar{A}_r = \overline{a}_r] \\ &= E[\mu_{r+1}(\overline{\mathbf{S}}_{r+1}, \overline{a}_T) \mid \overline{\mathbf{S}}_r, \bar{A}_r = \overline{a}_r], \end{aligned} \tag{1.214}$$

where (i) holds since $Y(\overline{a}_T) \perp\!\!\!\perp A_r \mid \overline{\mathbf{S}}_r, \bar{A}_{r-1} = \overline{a}_{r-1}$ under the Assumption 6; (ii) holds by the tower rule. Together with (1.213), we conclude that

$$H_{r,2} = 0. \tag{1.215}$$

For $H_{r,3}$, by the tower rule, we have

$$
\begin{aligned}
H_{r,3} &= E\Bigg[E\Bigg[\frac{\prod_{l=t+1}^{r}\mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)}(\mu_r(\overline{\mathbf{S}}_r,\overline{a}_T)-\mu_r^*(\overline{\mathbf{S}}_r,\overline{a}_T)) \mid \overline{\mathbf{S}}_r,\bar{A}_{r-1}=\overline{a}_{r-1}\Bigg] \\
&\qquad \cdot P(\widetilde{A}_{r-1}=\widetilde{a}_{r-1} \mid \overline{\mathbf{S}}_r,\bar{A}_t=\overline{a}_t) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\Bigg] \\
&= E\Bigg[\frac{E[\mathbb{1}_{\{A_r=a_r\}} \mid \overline{\mathbf{S}}_r,\bar{A}_{r-1}=\overline{a}_{r-1}]}{\prod_{l=t+1}^{r}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)}(\mu_r(\overline{\mathbf{S}}_r,\overline{a}_T)-\mu_r^*(\overline{\mathbf{S}}_r,\overline{a}_T)) \\
&\qquad \cdot E(\mathbb{1}_{\{\widetilde{A}_{r-1}=\widetilde{a}_{r-1}\}} \mid \overline{\mathbf{S}}_r,\bar{A}_t=\overline{a}_t) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\Bigg] \\
&\overset{(i)}{=} E\left[\frac{\prod_{l=t+1}^{r-1}\mathbb{1}_{\{A_l=a_l\}}\pi_r(\overline{\mathbf{S}}_r,\overline{a}_r)}{\prod_{l=t+1}^{r-1}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)\pi_r^*(\overline{\mathbf{S}}_r,\overline{a}_r)}(\mu_r(\overline{\mathbf{S}}_r,\overline{a}_T)-\mu_r^*(\overline{\mathbf{S}}_r,\overline{a}_T)) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\right] \\
&\overset{(ii)}{=} E\left[\frac{\prod_{l=t+1}^{r-1}\mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r-1}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)}(\mu_r(\overline{\mathbf{S}}_r,\overline{a}_T)-\mu_r^*(\overline{\mathbf{S}}_r,\overline{a}_T)) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\right]=-H_{r-1,1}, \qquad (1.216)
\end{aligned}
$$

where (i) holds by the tower rule and that $\pi_r(\overline{\mathbf{S}}_r,\overline{a}_{r-1})=P[A_r=a_r \mid \overline{\mathbf{S}}_r,\bar{A}_{r-1}=\overline{a}_{r-1}]$; (ii) holds since either $\pi_r^*(\cdot,\overline{a}_T)=\pi_r(\cdot,\overline{a}_T)$ or $\mu_r^*(\cdot,\overline{a}_T)=\mu_r(\cdot,\overline{a}_T)$ by assumption. Combining (1.212)-(1.216) with (1.211), we have

$$
\begin{aligned}
&E\Bigg[\sum_{r=t+1}^{T}\frac{\prod_{l=t+1}^{r}\mathbb{1}_{\{A_l=a_l\}}}{\prod_{l=t+1}^{r}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)}(\mu_{r+1}^*(\overline{\mathbf{S}}_{r+1},\overline{a}_T)-\mu_r^*(\overline{\mathbf{S}}_r,\overline{a}_T)) \\
&\qquad +\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t=\bar{\mathbf{s}}_t,\bar{A}_t=\overline{a}_t\Bigg] \\
&= H_T+\sum_{r=t+1}^{T-1}(H_{r,1}+H_{r,2}+H_{r,3})+E[\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t] \\
&= -H_{T-1,1}+\sum_{r=t+1}^{T-1}(H_{r,1}-H_{r-1,1})+E[\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t] \\
&= -H_{t,1}+E[\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t] \\
&\overset{(i)}{=} E\left[\mu_{t+1}(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\right] \overset{(ii)}{=} \mu_t(\overline{\mathbf{S}}_t,\overline{a}_T), \text{ for any } t\in\{1,\ldots,T\},
\end{aligned}
$$

where (i) holds since $H_{t,1}=E\left[\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T)-\mu_{t+1}(\overline{\mathbf{S}}_{t+1},\overline{a}_T) \mid \overline{\mathbf{S}}_t,\bar{A}_t=\overline{a}_t\right]$; (ii) holds since (1.214) holds for any $r\in\{1,\ldots,T\}$. ∎

Proof of Proposition 1. By Theorem 11 with $t = 0$, we have

$$E\left[\sum_{t=1}^{T} \frac{\mathbb{1}_{\{\bar{A}_t=\bar{a}_t\}}}{\prod_{l=1}^{t}\pi_l^*(\overline{\mathbf{S}}_l,\overline{a}_l)}(\mu_{t+1}^*(\overline{\mathbf{S}}_{t+1},\overline{a}_T)-\mu_t^*(\overline{\mathbf{S}}_t,\overline{a}_T))+\mu_1^*(\mathbf{S}_1,\overline{a}_T)\right]$$

$$=\mu_0(\bar{\mathbf{s}}_0,\overline{a}_T)=E[Y(\overline{a}_T)|\overline{\mathbf{S}}_0=\bar{\mathbf{s}}_0]=\theta_{\overline{a}_T}.$$

■

1.9 Acknowledgement

Chaper 1, 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 2

Dynamic treatment effects: high-dimensional inference under model misspecification

2.1 Introduction

Statistical inference and estimation of causal relationships have a long tradition. 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 more ranging from biomedical studies and public health to political science. This work considers statistical inference of causal effects for dynamic and observational data with possibly high-dimensional confounding. In dynamic treatment settings, model misspecification is more likely to occur in practice due to the many possible

dependencies; previous treatments can arbitrarily affect future treatments and/or outcomes. High-dimensional confounding is a real possibility as multiple covariates collected over time quickly outgrow the treatment-specific sample size. The double robust inference that allows model misspecification has been a long-standing open problem; even in low-dimensional settings, only a few separate advances have been successfully made. We hope to bring to the literature a distinct double-robust solution.

We consider the dynamic setting with binary treatments at two exposure times, although our results extend to finite exposure times, and collect independent and identically distributed samples $\mathbb{S} = \{\mathbf{W}_i\}_{i=1}^N$, $\mathbf{W}_i = (Y_i, A_{1i}, A_{2i}, \mathbf{S}_{1i}, \mathbf{S}_{2i})$, with $\mathbf{W}$ being an independent copy of $\mathbf{W}_i$. Here, $Y \in \mathbb{R}$ denotes the observed outcome at the final or last treatment stage. The causal setting of interest is framed through potential outcomes $Y(a_1, a_2)$ with $a_1, a_2 \in \{0, 1\}$ denoting treatment at first and second exposure time. We assume the consistency of potential outcomes with $Y = Y(A_1, A_2)$ and A_1, A_2 denoting the observed binary treatment assignments at the first and the second exposure time, respectively. At each exposure, we also observe covariates $\mathbf{S}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{S}_2 \in \mathbb{R}^{d_2}$. We let the first coordinate of $\mathbf{S}_1$ be an intercept term. Covariate history up to time two is denoted with $\bar{\mathbf{S}}_2 := (\mathbf{S}_1^\top, \mathbf{S}_2^\top)^\top \in \mathbb{R}^d$, where the dimension $d := d_1 + d_2$, is potentially much larger than N. The dynamic treatment effect (DTE) is defined as DTE $:= \theta_{a_1,a_2} - \theta_{a_1',a_2'}$, where $\theta_{a_1,a_2} := \mathbb{E}\{Y(a_1, a_2)\}$ and (a_1, a_2) and (a_1', a_2') denote the treatment and control paths. Without loss of generality, we focus on the inference of the counterfactual mean $\theta_{1,1}$. To identify the parameter of interest, we consider the marginal structural mean (MSM) models [MvdLRG01], where $\pi(\mathbf{s}_1) := \mathbb{P}(A_1 = 1 \mid \mathbf{S}_1 = \mathbf{s}_1)$ and $\rho(\bar{\mathbf{s}}_2) := \mathbb{P}(A_2 = 1 \mid \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, A_1 = 1)$ denote the

true treatment assignment probabilities, i.e., propensity scores (PS) at both exposure times, and $\mu(\mathbf{s}_1) := \mathbb{E}\{Y(1,1) \mid \mathbf{S}_1 = \mathbf{s}_1\}$ and $\nu(\bar{\mathbf{s}}_2) := \mathbb{E}\{Y(1,1) \mid \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, A_1 = 1\}$ denote the true outcome regressions (OR).

As naive average of the outcomes is biased due to confounding effects, a common approach to consider is that of the inverse propensity weighting (IPW) [Rob86, Rob00a, HBR01, Rob04]. Under standard identification conditions (see Assumption 2.1 below), the IPW representation

$$\theta_{1,1} = \mathbb{E}\{\psi_{\text{IPW}}(\mathbf{W}; \pi^*, \rho^*)\} \text{ holds when } \pi^* = \pi \text{ and } \rho^* = \rho, \tag{2.1}$$

where with a little abuse in the notation, $\psi_{\text{IPW}}(\mathbf{W}; \pi^*, \rho^*) := A_1 A_2 Y / \{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)\}$, π^* and ρ^* denote the working models for the PS at the first and second exposure times, respectively. Here, we refer to working models as the population-level approximations of the true nuisance functions. However, IPW requires correctly specified PS models, with model misspecification leading to inconsistent estimators. In the following, we first propose a new, model-robust IPW representation that is unbiased under model misspecification.

Lemma 2.1 (Model-robust IPW)**.** *Let Assumption 2.1 holds. Suppose that either the true OR or PS model are linear or logistic, respectively, at each exposure time, i.e., 1) either* $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^0)$ *or* $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top \boldsymbol{\beta}^0$ *with some* $\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$ *and 2) either* $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^0)$ *or* $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$ *with some* $\boldsymbol{\delta}^0, \boldsymbol{\alpha}^0 \in \mathbb{R}^d$*. Then we have*

$$\theta_{1,1} = \mathbb{E}\{\psi_{\text{IPW}}(\mathbf{W}; \pi^*, \rho^*)\} \text{ where } \pi^*(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^*) \text{ and } \rho^*(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^*), \tag{2.2}$$

with $\boldsymbol{\gamma}^* \in \mathbb{R}^{d_1}$ *and* $\boldsymbol{\delta}^* \in \mathbb{R}^d$ *being the solutions of*

$$\mathbb{E}\left[\left\{1 - \frac{A_1}{\pi^*(\mathbf{S}_1)}\right\}\mathbf{S}_1\right] = \mathbf{0}, \quad \text{and} \quad \mathbb{E}\left[\frac{A_1}{\pi^*(\mathbf{S}_1)}\left\{1 - \frac{A_2}{\rho^*(\bar{\mathbf{S}}_2)}\right\}\bar{\mathbf{S}}_2\right] = \mathbf{0}. \tag{2.3}$$

Unlike the standard IPW literature [Rob86, Rob00a, HBR01, Rob04, BAWM18], new representations (2.3) allow for misspecified PS models even within the IPW framework, as long as the corresponding OR models are linear. We expect that root-N inferential results follow easily when the working PS models are root-N estimable under the robustness setting of Lemma 2.1, for example, when the covariates are low-dimensional. However, in high dimensions or modern non-parametric settings, the nuisances' estimation errors are typically non-ignorable in general, and hence we cannot guarantee root-N inference for $\theta_{1,1}$. Therefore, we consider a well-known double-robust representation of $\theta_{1,1}$ and propose novel estimators of the outcome models, together with new PS estimators based on (2.3). We use the doubly robust (DR) score [NBW21, TYWK$^+$19, vdLG11, ORR10, MvdLRG01],

$$\psi(\mathbf{W};\boldsymbol{\eta}^*) := \mu^*(\mathbf{S}_1) + \frac{A_1\{\nu^*(\bar{\mathbf{S}}_2) - \mu^*(\mathbf{S}_1)\}}{\pi^*(\mathbf{S}_1)} + \frac{A_1A_2\{Y - \nu^*(\bar{\mathbf{S}}_2)\}}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}, \tag{2.4}$$

for which we show DR properties whenever at least one of the nuisance models is correctly specified at each exposure, i.e., whenever Assumption 2.2 holds, $\theta_{1,1} = \mathbb{E}\{\psi(\mathbf{W};\boldsymbol{\eta}^*)\}$. Here, $\boldsymbol{\eta}^* = ({\boldsymbol{\gamma}^*}^\top, {\boldsymbol{\delta}^*}^\top, {\boldsymbol{\alpha}^*}^\top, {\boldsymbol{\beta}^*}^\top)^\top$, where $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ are defined such that (2.3) holds. We introduce working OR models $\mu^*(\mathbf{s}_1) = \mathbf{s}_1^\top\boldsymbol{\beta}^*$ and $\nu^*(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top\boldsymbol{\alpha}^*$ with newly proposed $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ below. Let $\widehat{\boldsymbol{\eta}}$ be an estimator of $\boldsymbol{\eta}^*$. Note that, $\psi(\mathbf{W};\boldsymbol{\eta}^*) = W_1(\boldsymbol{\eta}^*) + W_2(\boldsymbol{\eta}^*) + \psi_{\text{IPW}}(\mathbf{W};\pi^*,\rho^*)$, where $W_1(\boldsymbol{\eta}^*) = \{1 - A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\}\mathbf{S}_1^\top\boldsymbol{\beta}^*$ and $W_2(\boldsymbol{\eta}^*) = A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}^*$. The first two terms, $W_1(\boldsymbol{\eta}^*)$ and $W_2(\boldsymbol{\eta}^*)$ can be viewed as bias correction terms in the presence of model misspecification, and we propose moment-targeting estimators that will remove it asymptotically. To set ideas, let us consider a special case where $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$, $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, and we only focus on the estimation error coming from

$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$. Then the biases of the IPW and DR score are

$$\mathbb{E}\{\psi_{\mathrm{IPW}}(\mathbf{W}; \pi^*, \widehat{\rho}) - \psi_{\mathrm{IPW}}(\mathbf{W}; \pi^*, \rho^*)\} = \mathbb{E}\left[\frac{A_1 A_2}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\left\{\frac{1}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})} - \frac{1}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\} Y\right], \quad (2.5)$$

$$\mathbb{E}\{\psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\} = \mathbb{E}\left[\frac{A_1 A_2}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\left\{\frac{1}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})} - \frac{1}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\} \big(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*\big)\right], \quad (2.6)$$

where $\widetilde{\boldsymbol{\eta}} = (\boldsymbol{\gamma}^{*\top}, \widehat{\boldsymbol{\delta}}^\top, \boldsymbol{\alpha}^{*\top}, \boldsymbol{\beta}^{*\top})^\top$. Consequently, the bias (2.6) is potentially much smaller than (2.5); when $\nu(\bar{\mathbf{s}}_2) = \nu^*(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^*$, we have $\mathbb{E}\{\psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\} = 0$. However, when model misspecification occurs and $\nu(\cdot) \neq \nu^*(\cdot)$, $\mathbb{E}\{\psi(\mathbf{W}; \widetilde{\boldsymbol{\eta}}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\} \neq 0$ in general. In the following, we design the nuisance parameters $\boldsymbol{\eta}^*$ such that the bias, i.e. (2.6), is asymptotically negligible even under such model misspecification. Whenever $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$ is small enough, we can approximate (2.6) as

$$\mathbb{E}\Big[\frac{A_1 A_2}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)} \exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\big(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*\big)\bar{\mathbf{S}}_2^\top\Big](\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*),$$

and mitigate the bias effects by constructing $\boldsymbol{\alpha}^*$, i.e., ν^*, as the solution of

$$\mathbb{E}\Big[\frac{A_1 A_2}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)} \exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\big(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*\big)\bar{\mathbf{S}}_2\Big] = \mathbb{E}\Big[\frac{A_1 A_2\{1 - \rho^*(\bar{\mathbf{S}}_2)\}}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\big\{Y - \nu^*(\bar{\mathbf{S}}_2)\big\}\bar{\mathbf{S}}_2\Big] = \mathbf{0}. \quad (2.7)$$

Similar approximation of (2.6), whenever $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ is small enough, leads to a construction of new moment defining $\boldsymbol{\beta}^*$, i.e., μ^* as the solution of

$$\mathbb{E}\left[A_1\Big\{1 - \frac{1}{\pi^*(\mathbf{S}_1)}\Big\}\Big\{\frac{A_2(Y - \nu^*(\bar{\mathbf{S}}_2))}{\rho^*(\bar{\mathbf{S}}_2)} + \nu^*(\bar{\mathbf{S}}_2) - \mu^*(\mathbf{S}_1)\Big\}\mathbf{S}_1\right] = \mathbf{0}. \quad (2.8)$$

We name the parameters satisfying (2.3), (2.7), and (2.8) as the *moment-targeted nuisance parameters*.

Double-robust literature, unlike IPW or covariate balancing approaches [KS18, YS18] which require correctly specified PS and OR models, respectively, enables certain forgiveness: only one and not both of them need to be correctly specified. However, with dynamic

treatment regimes, the equivalent of this forgiveness has not been achieved yet. With the presence of high-dimensional covariates, DR estimators of the DTE have been recently studied by [BHL22, BJZ21, LS21]. However, statistical inference for the DTE under model misspecification has not been considered. We provide a new estimator establishing statistical inference for the DTE allowing model misspecification in high dimensions. Specifically, we allow for the following four settings:

The OR models at the first and second exposure are correctly specified; (2.9)

The PS models at the first and second exposure are correctly specified; (2.10)

The first OR model and the second PS model are correctly specified; (2.11)

The second OR model and the first PS model are correctly specified. (2.12)

Therefore, we require at least one of the models to be correctly specified at each exposure. This property is named *sequentially model doubly robust* (SMDR); see Theorem 2.1. Even in low dimensions, SMDR is the most generous conditions up to date: [Rob00b, MvdLRG01, BR05, YvdL06] establish confidence intervals when either (2.9) or (2.10) holds, but (2.11) and (2.12) are not allowed. [BRR19], arguably the best result up to date, proposed a "multiple robust" estimator (also in low dimensions), which allows for (2.9), (2.10), or (2.11) but does not allow for (2.12). This work, therefore, provides a solution to a long-standing open problem of interest.

The average treatment effect (ATE) estimation problem is closely related to the DTE – it can be seen as a special, degenerate DTE estimation problem. The ATE has a long tradition [Rub74] and has attracted a significant amount of attention with the advent of high-dimensional models [Far15, AIW18, CCD+18, SRR19, BWZ19, Tan20]. Statistical inference

for the ATE under model misspecification has been studied recently by [SRR19, Tan20, DV20, DAV20, AV21]. They propose "model doubly robust" estimators, which are shown to be asymptotically normal as long as either the OR model or the PS model is correctly specified, a condition that we successfully naturally mimic with the proposed SMDR; for the first time in the dynamic setting. To achieve model DR, authors therein, similarly to [BWZ19], discuss the construction of novel and not off-the-shelf ready nuisance estimates. Our moment-targeting estimates are motivated by the above work. However, identification of the DTE is non-trivially distinct from the problem of identifying an ATE, leading to distinct functionals from those considered in the prior work. In turn, the estimation problem here is differentiated and the estimators developed have sequential estimation structure. Dynamic treatment effects rely on a key identification condition, regarding outcome models in particular, which ensures that DTE quantifies the effect of a sequence of treatments on the outcome of interest. Authors in previous work could not explicitly design a DR estimator that meets this condition and have therefore resorted to particular relaxations of DR notions for dynamic settings.

The manuscript is organized as follows. In Section 2.2, we introduce the doubly robust representation for the counterfactual mean and further motivate the working nuisance models, which are constructed to achieve inference under model misspecification. In Section 2.3, we propose a sequential model doubly robust estimator for the DTE based on the moment-targeted nuisance estimators. Our main theoretical results, which demonstrate the inference results under possible model misspecification, are provided in Theorems 2.1 and 2.2. The theoretical results for the nuisance estimators are further provided in Section 2.4.

In Section 2.5, we illustrate the finite sample performance of the proposed estimator under simulations and semi-synthetic experiments. Additional justifications and the proofs of the main results are provided in the Supplementary Material.

We use the following notation throughout. Let $\mathbb{P}(\cdot)$ and $\mathbb{E}(\cdot)$ denote the probability measure and expectation characterizing the joint distribution of the underlying random vector $\mathbb{W} := (\{Y(a_1,a_2)\}_{a_1,a_2\in\{0,1\}}, A_1, A_2, \mathbf{S}_1, \mathbf{S}_2)$ (independent of the observed samples), respectively. For any $\alpha > 0$, let $\psi_\alpha(x) := \exp(x^\alpha) - 1$, $\forall x > 0$. The ψ_α-Orlicz norm $\|\cdot\|_{\psi_\alpha}$ of a random variable $X \in \mathbb{R}$ is defined as $\|X\|_{\psi_\alpha} := \inf\{c > 0 : \mathbb{E}[\psi_\alpha(|X|/c)] \leq 1\}$. Two special cases are given by $\psi_2(x) = \exp(x^2) - 1$ and $\psi_1(x) = \exp(x) - 1$. We use $a_N \asymp b_N$ to denote $cb_N \leq a_N \leq Cb_N$ for all $N \geq 1$ and constants $c, C > 0$. For any $\tilde{\mathbb{S}} \subseteq \mathbb{S} = (\mathbf{Z}_i)_{i=1}^N$, define $\mathbb{P}_{\tilde{\mathbb{S}}}$ as the joint distribution of $\tilde{\mathbb{S}}$ and $\mathbb{E}_{\tilde{\mathbb{S}}}(f) = \int f d\mathbb{P}_{\tilde{\mathbb{S}}}$. 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 $\|\mathbf{z}\|_\infty := \max_j |\mathbf{z}_j|$. We denote the logistic function with $g(u) = \exp(u)/(1+\exp(u))$, for all $u \in \mathbb{R}$. A d dimensional vector of all ones and zeros are denoted with $\mathbf{1}_{(d)}$ and $\mathbf{0}_{(d)}$, respectively.

2.2 Moment-targeted nuisance estimators

To identify the counterfactual mean θ_{a_1,a_2}, we assume the standard sequential ignorability, consistency, and overlap conditions; see, e.g., [IR15, LM10, Mur03, Rob00a, Rob87].

Assumption 2.1 (Basic assumptions)**.** *(a) Sequential ignorability:* $Y(a_1,a_2) \perp\!\!\!\perp A_1 \mid \mathbf{S}_1$, $Y(a_1,a_2) \perp\!\!\!\perp A_2 \mid (\mathbf{S}_1, \mathbf{S}_2, A_1 = a_1)$. *(b) Consistency:* $Y = Y(A_1, A_2)$. *(c) Overlap: let* $\mathbb{P}(c_0 < \pi(\mathbf{S}_1) < 1 - c_0) = 1$, $\mathbb{P}(c_0 < \rho(\bar{\mathbf{S}}_2) < 1 - c_0) = 1$ *with some constant* $c_0 \in (0, 1)$. *Additionally, let* $\pi^*(\cdot)$ *and* $\rho^*(\cdot)$ *be some functions satisfying* $\mathbb{P}(c_0 < \pi^*(\mathbf{S}_1) < 1 - c_0) =$

1, $\mathbb{P}(c_0 < \rho^*(\bar{\mathbf{S}}_2) < 1 - c_0) = 1$.

In this paper, we consider linear and logistic working models for the OR and PS models as defined above in (2.4). The model correctness conditions, named sequential model double robustness (SMDR), are introduced below.

Assumption 2.2 (Sequential model double robustness)**.** *Let (a) either $\pi = \pi^*$ or $\mu = \mu^*$ holds, but not necessarily both; and (b) either $\rho = \rho^*$ or $\nu = \nu^*$, but not necessarily both.*

To reduce the bias under model misspecification, we construct the moment-targeted nuisance parameters, $\boldsymbol{\eta}^*$, as the solution of $\mathbb{E}\{\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ even when models are possibly misspecified as in Assumption 2.2. Note that this is more ambitious task than the Neyman orthogonality, which ensures $\mathbb{E}\{\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \mathbf{0}$ but requires correctly specified models. With that in mind, we introduce new estimators of $\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$, in that order; nuisances are intertwined and require sequential estimation. Justifications of the claims below are provided in equations (2.43)-(2.46) of the Supplementary Material.

The first is $\boldsymbol{\gamma}^*$ and $\pi^*(\mathbf{S}_1) = g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)$, introduced to balance the model misspecification in PS models through equation (2.3) (the left-hand side). This, in turn, leads to a loss function ℓ_1 and a targeted parameter of interest

$$\boldsymbol{\gamma}^* := \text{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^{d_1}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\}, \quad \text{for} \quad \ell_1(\mathbf{W};\boldsymbol{\gamma}) := (1-A_1)\mathbf{S}_1^\top\boldsymbol{\gamma} + A_1\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}). \quad (2.13)$$

Additionally, we observe that $\mathbb{E}\{\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\} = \boldsymbol{\nabla}_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma}^*)\}$ with the later being zero at $\boldsymbol{\gamma}^*$ under the SMDR setting. The next parameter to be defined is $\boldsymbol{\delta}^*$, for which $\rho^*(\bar{\mathbf{S}}_2) = g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)$, and which needs to satisfy the targeted moment condition (2.3) (the

right-hand side). We show that this is then equivalent to the population M-estimator

$$\boldsymbol{\delta}^* := \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^d} \mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\}, \text{ for } \ell_2(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta}) := \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma})}\Big\{(1-A_2)\bar{\mathbf{S}}_2^\top\boldsymbol{\delta} + A_2\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})\Big\}, \tag{2.14}$$

where $\boldsymbol{\delta}^* = \boldsymbol{\delta}^*(\boldsymbol{\gamma}^*)$. Moreover, we show $\mathbb{E}\left\{\nabla_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} = \nabla_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\}$, therefore enabling vanishing the effect of the first order bias under SMDR. For the remainder two OR models, we design moment conditions such that (2.7) and (2.8) hold. To do so, we observe that the corresponding moments are

$$\boldsymbol{\alpha}^* := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \mathbb{E}\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha})\}, \quad \text{and} \quad \boldsymbol{\beta}^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} \mathbb{E}\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta})\}, \tag{2.15}$$

with the new loss functions ℓ_3 and ℓ_4:

$$\ell_3(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}) := \frac{A_1A_2\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})}{g(\mathbf{S}_1^\top\boldsymbol{\gamma})}\left(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}\right)^2, \tag{2.16}$$

$$\ell_4(\mathbf{W};\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}) := A_1\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma})\left\{\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha} + \frac{A_2(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha})}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})} - \mathbf{S}_1^\top\boldsymbol{\beta}\right\}^2. \tag{2.17}$$

For convenience, we suppress the intrinsic notation of $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)$ with $\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*$ (and $\boldsymbol{\alpha}^*$) defined in (2.13) and (2.14). The above losses enable us to reduce the estimation bias under SMDR assumptions by ensuring $\mathbb{E}\left\{\nabla_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} = \nabla_{\boldsymbol{\beta}}\mathbb{E}\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\}/2$ and $\mathbb{E}\left\{\nabla_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} = \nabla_{\boldsymbol{\alpha}}\mathbb{E}\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\}/2$. The loss functions (2.13), (2.14), (2.16), and (2.17) are named as *moment-targeting loss functions*, which ensure the SMDR property of the resulting estimator and therefore solving an open problem of double-robustness in dynamic exposure times. Moment-targeting estimators in non-dynamic settings [SRR19, Tan20, AV21, BWZ19] match up only with (2.13) above; whereas, other losses, even in low-dimensional settings, are entirely new and are the first to achieve SMDR.

The uniqueness of the moment-targeted nuisance parameters is discussed in Section 2.7.1 of the Supplementary Material. We next discuss their identification. The nested models are hard to interpret, especially the OR model at the first exposure, $\mu(\cdot)$. This is not a caviat of our representation but of a dynamic nature of the problem and is widely recognized; see, e.g., [BRR19]. (a) We say $\pi^*(\cdot)$ is correctly specified when $\pi^*(\cdot) = \pi(\cdot)$, which occurs if and only if (iff) there exists some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$, such that $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^0)$ holds. Additionally, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^0$. (b) We say $\rho^*(\cdot)$ is correctly specified when $\rho^*(\cdot) = \rho(\cdot)$, which occurs iff there exists some $\boldsymbol{\delta}^0 \in \mathbb{R}^d$, such that $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^0)$ holds. Additionally, $\boldsymbol{\delta}^* = \boldsymbol{\delta}^0$. (c) We say $\nu^*(\cdot)$ is correctly specified when $\nu^*(\cdot) = \nu(\cdot)$, which occurs iff there exists some $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$, such that $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$ holds. Additionally, $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0$. (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 $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top \boldsymbol{\beta}^0$ and, furthermore, either case (b) or (c) holds. Additionally, $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$. Note that, $\boldsymbol{\delta}^*$, (2.14), is a function of $\boldsymbol{\gamma}^*$. However, (b) specifies that the correctness of $\rho^*(\cdot)$ does not depend on $\boldsymbol{\gamma}^*$. Analogous result for $\pi^*(\cdot)$, $\rho^*(\cdot)$ and $\nu^*(\cdot)$ can be found in (a)-(c) therefore establishing that their correctness has no effect on each other. However, this is not the case for the OR model at the first exposure time, $\mu(\cdot)$. Namely, if $\mu(\cdot)$ is linear with $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top \boldsymbol{\beta}^0$ for some $\boldsymbol{\beta}^0$, this does *not* imply that $\mu^*(\cdot)$ is correctly specified, as $\boldsymbol{\beta}^*$ in (2.15) may not be equal to $\boldsymbol{\beta}^0$. From (d), we see that $\mu^*(\cdot)$ is correctly specified if additionally either $\rho^*(\cdot)$ or $\nu^*(\cdot)$ is (or both are) correctly specified, a condition that always assumed through SMDR settings of Assumption 2.2. Further details and justifications can be found in Section 2.7.2 of the Supplementary Material. Based on the moment-targeted nuisance estimators, and new loss functions, ℓ_1, ℓ_2, ℓ_3, and ℓ_4 as defined in (2.13), (2.14), (2.16), and (2.17), we propose a

sequential model doubly robust estimator for the counterfactual mean, $\theta_{1,1} = \mathbb{E}\{Y(1,1)\}$ in Algorithm 3. Note that the estimators therein are sequential in that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}}(\widehat{\boldsymbol{\gamma}})$, $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}})$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}})$. Whenever possible, we avoid arduous exposition for readability.

2.3 Sequential model doubly robust inference

In this section, we choose the tuning parameters as $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1/N}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log d/N}$, $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/N}$, $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log d_1/N}$. 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.

Assumption 2.3 (Sparsity)**.** *Let* $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\beta}} = o(N/\log d_1)$, $s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = o(N/\log d)$, *and* $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$.

The sparsity conditions of the type $s = o\,(N/\log d)$ are very common in the high-dimensional statistics literature and guarantee estimation consistency. If we further assume that $\|\bar{\mathbf{S}}_2\|_\infty < C$, as in, e.g., [BWZ19, Tan20, SRR19], then the condition $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$ is no longer required. The following assumption imposes some standard moment conditions.

Assumption 2.4 (Sub-Gaussianity)**.** *Let* $\bar{\mathbf{S}}_2$ *be a sub-Gaussian random vector with* $\|\mathbf{v}^\top \bar{\mathbf{S}}_2\|_{\psi_2} \le \sigma_{\mathbf{S}} \|\mathbf{v}\|_2$ *for all* $\mathbf{v} \in \mathbb{R}^d$. *Let* $\varepsilon := Y(1,1) - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*$ *and* $\zeta := \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - \mathbf{S}_1^\top \boldsymbol{\beta}^*$ *be sub-Gaussian with* $\|\varepsilon\|_{\psi_2} \le \sigma_\varepsilon$ *and* $\|\zeta\|_{\psi_2} \le \sigma_\zeta$. *In addition, let* $Var\{Y(1,1)\} > c_Y$ *and the smallest eigenvalue of* $\mathbb{E}(A_1 \bar{\mathbf{S}}_2 \bar{\mathbf{S}}_2^\top)$ *is bounded bellow by* $c_{\min}$. *Here,* $\sigma_{\mathbf{S}}, \sigma_\varepsilon, \sigma_\zeta, c_Y, c_{\min}$ *are some positive constants.*

Algorithm 3 The sequential model doubly robust (SMDR) counterfactual mean estimator

Require: Observations $\mathbb{S} = (\mathbf{W}_i)_{i=1}^N$ and the treatment path $(a_1, a_2) = (1, 1)$.

1: Let $\mathcal{I} = \{1, 2, \ldots, N\} = \cup_{k=1}^{\mathbb{K}} \mathcal{I}_k$ with equal sized splits $n = N/\mathbb{K}$ and $\mathbb{K} \geq 2$.

2: **for** $k = 1, 2, ..., \mathbb{K}$ **do**

3: $\quad \mathcal{I}_{-k} \leftarrow \mathcal{I} \setminus \mathcal{I}_k$

4: $\quad \mathcal{I}_{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\delta}}, \mathcal{I}_{\boldsymbol{\alpha}}, \mathcal{I}_{\boldsymbol{\beta}} \leftarrow$ size M disjoint partition of $\mathcal{I}_{-k}$ with $M = N(\mathbb{K}-1)/(4\mathbb{K})$.

5: $\quad$ Propensity at the first exposure

$$\widehat{\boldsymbol{\gamma}}_{-k} \leftarrow \widehat{\boldsymbol{\gamma}} \ := \ \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{d_1}} \Big\{ M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\gamma}}} \ell_1(\mathbf{W}_i; \boldsymbol{\gamma}) + \lambda_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_1 \Big\}, \tag{2.18}$$

6: $\quad$ Propensity at the last/second exposure

$$\widehat{\boldsymbol{\delta}}_{-k} \leftarrow \widehat{\boldsymbol{\delta}} \ := \ \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^{d}} \Big\{ M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} \ell_2(\mathbf{W}_i; \widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}) + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1 \Big\}, \tag{2.19}$$

7: $\quad$ Outcome at the last exposure

$$\widehat{\boldsymbol{\alpha}}_{-k} \leftarrow \widehat{\boldsymbol{\alpha}} \ := \ \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{d}} \Big\{ M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} \ell_3(\mathbf{W}_i; \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}) + \lambda_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \Big\}, \tag{2.20}$$

8: $\quad$ Outcome at the first exposure

$$\widehat{\boldsymbol{\beta}}_{-k} \leftarrow \widehat{\boldsymbol{\beta}} \ := \ \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d_1}} \Big\{ M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} \ell_4(\mathbf{W}_i; \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \Big\}, \tag{2.21}$$

9: **end for**

10: **return** SMDR estimator is

$$\widehat{\theta}_{1,1} = N^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}), \tag{2.22}$$

where $\widehat{\boldsymbol{\eta}}_{-k} = (\widehat{\boldsymbol{\gamma}}_{-k}^\top, \widehat{\boldsymbol{\delta}}_{-k}^\top, \widehat{\boldsymbol{\alpha}}_{-k}^\top, \widehat{\boldsymbol{\beta}}_{-k}^\top)^\top$ and $\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k})$ is defined through (2.4) as

$$\Big\{1 - \frac{A_{1i}}{g(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}_{-k})}\Big\} \mathbf{S}_{1i}^\top \widehat{\boldsymbol{\beta}}_{-k} + \frac{A_{1i}}{g(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}_{-k})} \Big\{1 - \frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}_{-k})}\Big\} \bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\alpha}}_{-k} + \frac{A_{1i} A_{2i} Y_i}{g(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}_{-k}) g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}_{-k})}$$

Theorem 2.1 (Inference under model misspecification)**.** *Let Assumptions 2.1-2.4 hold. Let the following product sparsity conditions hold*

$$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} = o\left(\frac{N}{(\log d_1)^2}\right), \quad s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right). \tag{2.23}$$

We assume the following additional conditions if model misspecification occurs:

$$\text{if } \rho(\cdot) \neq \rho^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{\log d_1 \log d}\right); \tag{2.24}$$

$$\text{if } \nu(\cdot) \neq \nu^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} = o\left(\frac{N}{\log d_1 \log d}\right), \ s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right); \tag{2.25}$$

$$\text{if } \mu(\cdot) \neq \mu^*(\cdot), \text{ further let } s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d_1}\right), \ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{\log d_1 \log d}\right). \tag{2.26}$$

Then, as $N, d_1, d_2 \to \infty$, in distribution,

$$\sigma^{-1} N^{-1/2} (\widehat{\theta}_{1,1} - \theta_{1,1}) \ \to \ \mathcal{N}(0,1), \ \text{ where } \ \sigma^2 \ := \ \mathbb{E}\left\{\psi(\mathbf{W}; \boldsymbol{\eta}^*) - \theta_{1,1}\right\}^2. \tag{2.27}$$

In addition, define

$$\widehat{\sigma}^2 \ := \ N^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \left\{\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \widehat{\theta}_{1,1}\right\}^2. \tag{2.28}$$

Then, as $N, d_1, d_2 \to \infty$, $\widehat{\sigma}^2 = \sigma^2\{1 + o_p(1)\}$.

Remark 2.1 (Sequential model double robustness)**.** *In Theorem 2.1, we demonstrate the "sequential model double robustness" (SMDR) property of our proposed estimator: root-N inference is provided as long as at least one nuisance model is correctly specified at each exposure; see Assumption 2.2. To the best of our knowledge, this is the first result that establishes DR property in its full generality. In high-dimensional dynamic settings, no inferential guarantees exist up to date that allow model misspecification. Among the low-dimensional literature, the recent work of [BRR19] provides the best results so far on model*

robustness. The estimator therein is asymptotically normal when either (2.9)*,* (2.10) *or* (2.11) *holds. However, our Assumption 2.2 allows an additional case* (2.12) *therefore filling in the important gap.*

Remark 2.2 (Required sparsity conditions under model misspecification)**.** *Here we discuss the sparsity conditions required in Theorem 2.1 for root-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 2.3 and* (2.23)*. 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 constructed based on. Recall that we construct the nuisance estimators sequentially in the order: $\widehat{\boldsymbol{\gamma}}$ then $\widehat{\boldsymbol{\delta}}$ followed by $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$. For instance, when $\mu^*(\cdot)$ is misspecified, as shown in* (2.26)*, we need a product condition between 1) $s_{\boldsymbol{\gamma}}$ and 2) $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}}$. Moreover, whenever OR model at the exposure time t is misspecified, based on the pattern we discussed above, we always require an ultra-sparse PS parameter at that exposure. More details are listed in Table 2.1.*

In addition, consider the degenerate case with one exposure time. 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 [Tan20, AV21], where both $s_{\boldsymbol{\gamma}} = o(\sqrt{N}/\log d_1)$ and $s_{\boldsymbol{\beta}} = o(\sqrt{N}/\log d_1)$ are required. [BWZ19] imposed different conditions with 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* $s_{\boldsymbol{\gamma}} = o(\sqrt{N}/\log d_1)$.

Table 2.1: 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
$\pi^*(\cdot)$	$\rho^*(\cdot)$	$\nu^*(\cdot)$	$\mu^*(\cdot)$	
✓	✓	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✓	✗	$s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✓	$s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✓	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✓	✗	✗	$s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✓	✗	✗	✓	$s_{\boldsymbol{\gamma}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✓	✗	✓	$s_{\boldsymbol{\delta}} = o\left(\frac{\sqrt{N}}{\log d}\right),\ s_{\boldsymbol{\gamma}} s_{\boldsymbol{\delta}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$
✗	✗	✓	✓	$s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} + s_{\boldsymbol{\gamma}} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o\left(\frac{N}{(\log d)^2}\right)$

Whenever all the nuisance models are correctly specified, we have the following result.

Theorem 2.2 (Inference under correctly specified models)**.** *Suppose all the nuisance models are correctly specified. Let Assumptions 2.1, 2.3 and 2.4 hold, as well as the product sparsity* (2.23)*. Then, as* $N, d_1, d_2 \to \infty$*,*

$$\sigma^{-1} N^{-1/2} (\widehat{\theta}_{1,1} - \theta_{1,1}) \ \to\ \mathcal{N}(0,1)$$

in distribution, where σ^2 *is defined in* (2.27)*. With* $\widehat{\sigma}^2$ *as in* (2.28)*, we also have* $\widehat{\sigma}^2 = \sigma^2\{1 + o_p(1)\}$.

Remark 2.3 (Sequential rate double robustness)**.** *As shown in Theorem 2.2, root-N inference requires product sparsity conditions between the nuisance parameters' sparsity levels at*

each exposure, i.e., (2.23)*; we name such a property as "sequential rate double robustness". We need the same product sparsity conditions as the Sequential Double Robust Lasso estimator proposed by [BJZ21]; such conditions are weaker than [BHL22] where an additional product sparsity condition* $s_{\boldsymbol{\gamma}} s_{\boldsymbol{\alpha}} = o(N/(\log d_1 \log d))$ *is imposed. With one time exposure, our conditions coincide with the "rate double robustness" of [CCD*$^+$*18] and [SRR19] and is weaker than, e.g., the sparsity conditions in [Far15, Tan20, AV21].*

2.4 Theoretical results for the nuisance estimators

We develop theoretical properties of the proposed moment-targeted nuisance estimators, $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\delta}}$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\boldsymbol{\beta}}$, defined in (2.18)-(2.21). In Section 2.4.1, we demonstrate the consistency of the nuisance estimators allowing all the models to be misspecified. In Section 2.4.2, we provide faster consistency rates for the nuisance estimators assuming some models being correctly specified.

2.4.1 Results for misspecified models

We first 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 previously constructed nuisance estimators, i.e., they are all inter-dependent. We carefully control the errors originated from the previous steps' estimation.

Theorem 2.3. *Let Assumptions 2.1 and 2.4 hold. Then, as* $N, d_1, d_2 \to \infty$*, the following holds:* ***(a)*** *If* $s_{\boldsymbol{\gamma}} = o(N/\log d_1)$ *and* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1/N}$*, then* $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\gamma}} \log d_1/N}\right)$.

(b) *In addition to (a), if* $s_{\boldsymbol{\delta}} = o(N/\log d)$ *and* $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log d/N}$*, then* $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d)/N}\right)$. ***(c)*** *In addition to (a) and (b), if* $s_{\boldsymbol{\alpha}} = o(N/\log d)$ *and* $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/N}$*, then* $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = O_p\left(\sqrt{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d)/N}\right)$. ***(d)*** *In addition to (a), (b), and (c), if* $s_{\boldsymbol{\beta}} = o(N/\log d_1)$ *and* $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log d_1/N}$*, then* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d + s_{\boldsymbol{\beta}} \log d_1/N}\right)$.

To establish the convergence rates in Theorem 2.3, we show the restricted strong convexity (RSC) conditions in Lemma 2.17 and control the loss functions' gradients in Lemma 2.18; see the Supplementary Material. Among the results in Theorem 2.3, part (b) is the most challenging to show. Notice that $\widehat{\boldsymbol{\delta}}$ is constructed based on a first-stage estimate $\widehat{\boldsymbol{\gamma}}$. Due to the occurrence of the imputation error $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$, the estimation error $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$ no longer belongs to the usual cone set $\mathbb{C}(S, k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq k\|\boldsymbol{\Delta}_S\|_1\}$. A similar problem has been recently studied by [BJZ21], where their Theorem 8 provides consistency rates of imputed Lasso estimates. The problem we consider here is even more technically challenging in that the loss function (2.14) is non-quadratic with respect to $\boldsymbol{\delta}$. We consider a cone set $\widetilde{C}(s, k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \leq k\sqrt{s}\|\boldsymbol{\Delta}\|_2\}$ that is "larger" than the usual $\mathbb{C}(S, k)$ and also different from the cone set studied by [BJZ21]. We show that $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^* \in \widetilde{C}(s, k)$ with high probability and some $k, s > 0$; see details in Lemma 2.10. Together with some empirical process results as in Lemma 2.8, we control the imputation error's effect and finally reach the consistency rates introduced above; see Lemma 2.9 and the proof of Theorem 2.3. Although we focus on a specific loss function (2.14), the results of part (b) in fact apply more broadly to other smooth and convex loss functions. As for parts (c) and (d), the corresponding loss functions are (weighted) least squares. By controlling all the imputation errors from multiple

stages, we show the consistency rates of the nested estimators. Since the nuisance estimators $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$ are constructed sequentially, and the later estimators depend on all the previous ones, the estimation errors of the nuisance parameters are cumulative, i.e., the consistency rate depends on the sparsity levels of all the nuisance parameters up to the current one.

2.4.2 Results for correctly specified models

If we have additional information that some of the nuisance models are correctly specified, we are able to achieve better consistency results than Theorem 2.3.

Theorem 2.4. *As $N, d_1, d_2 \to \infty$, the following holds:* ***(a)*** *Let $\rho(\cdot) = \rho^*(\cdot)$. Let the assumptions in part (b) of Theorem 2.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$. Then $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\delta}} \log d/N}\right)$.* ***(b)*** *Let $\nu(\cdot) = \nu^*(\cdot)$. Let the assumptions in part (c) of Theorem 2.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$ and $s_{\boldsymbol{\delta}} = O(N/(\log d)^2)$. Then $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\alpha}} \log d/N}\right)$.* ***(c)*** *Let $\nu(\cdot) = \nu^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 2.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$ and $s_{\boldsymbol{\delta}} = O(N/(\log d)^2)$. Then $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{(s_{\boldsymbol{\alpha}} \log d + s_{\boldsymbol{\beta}} \log d_1)/N}\right)$.* ***(d)*** *Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 2.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$. Then $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{(s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\beta}} \log d_1)/N}\right)$.* ***(e)*** *Let $\rho(\cdot) = \rho^*(\cdot)$, $\nu(\cdot) = \nu^*(\cdot)$, and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (d) of Theorem 2.3 hold. Additionally, let $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$ and $s_{\boldsymbol{\delta}} = O(N/(\log d)^2)$. Then $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}} \log d/N + \sqrt{s_{\boldsymbol{\beta}} \log d_1/N}\right)$. Further, if $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o(N/(\log d)^2)$, then $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\sqrt{s_{\boldsymbol{\beta}} \log d_1/N}\right)$.*

The new convergence rates in Theorem 2.4 are established through Lemmas 2.17 and

2.19 of the Supplementary Material. Assuming certain nuisance models being correct, unlike Theorem 2.3 and Lemma 2.18, we can control the gradients involving the *estimated* nuisance parameters and control the imputation errors from the previous steps' nuisances in a more efficient way; see more details in Lemma 2.19. As a result, we obtain faster convergence rates than Theorem 2.3 given additional model correctness information. For the "first" nuisance estimator $\widehat{\boldsymbol{\gamma}}$, as shown in case (a) of Theorem 2.3, we have $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{s_{\boldsymbol{\gamma}} \log d_1 / N})$ regardless of the correctness of $\pi^*(\cdot)$. When $\rho^*(\cdot)$ is correctly specified, the convergence rate of $\widehat{\boldsymbol{\delta}}$ depends only on $s_{\boldsymbol{\delta}}$; see part (a) of Theorem 2.4 and part (b) of Theorem 2.3 where $\rho^*(\cdot)$ is possibly misspecified. When $\nu^*(\cdot)$ is correctly specified, the convergence rate of $\widehat{\boldsymbol{\alpha}}$ depends only on $s_{\boldsymbol{\alpha}}$; see parts (b) and (c) of Theorems 2.4 and Theorem 2.3, respectively. 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, $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} = o(N/(\log d)^2)$, is assumed as in (2.23) of Theorem 2.1, the product sparsity $s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}$ can also be omitted.

Remark 2.4 (Bounded covariates)**.** *If we further assume that* $\|\bar{\mathbf{S}}_2\|_\infty < C < \infty$*, then the following conditions can be omitted:* $s_{\boldsymbol{\gamma}} = O((N/\log d_1 \log d))$ *in case (a);* $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$ *and* $s_{\boldsymbol{\delta}} = O(N/(\log d)^2)$ *in cases (b), (c), and (e);* $s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}} = O(N/(\log d_1 \log d))$ *in case (d).*

2.5 Numerical Experiments

2.5.1 Simulation studies

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 $a = (a_1, a_2) = (1, 1)$ and $a' = (a'_1, a'_2) = (0, 0)$. We describe the considered data generating processes below. The outcome variables are generated as $Y_i = A_{1i}A_{2i}Y_i(1,1) + (1 - A_{1i})(1 - A_{2i})Y_i(0,0)$.

Setting (a): Non-linear $\mu(\cdot)$ and non-logistic $\rho(\cdot)$ Generate covariates at the first exposure: for each $i \leq N$, $\mathbf{S}_{1i} \sim^{\text{iid}} N_{d_1}(\mathbf{0}, \mathbf{I}_{d_1})$. The treatment indicators of the first exposure are generated as $A_{1i}|\mathbf{S}_{1i} \sim \text{Bernoulli}(g(\mathbf{S}_{1i}^\top \boldsymbol{\gamma}))$. Covariates at the second exposure satisfy $\mathbf{S}_{2i} = 0.5Q(A_{1i})(\mathbf{S}_{1i}^2 - 1) + Q(A_{1i})\mathbf{S}_{1i} + A_i(1 + \delta_{1i})\mathbf{1}_{(d_2)} + \boldsymbol{\delta}_{1i}$, where $\mathbf{S}_{1i}^2 \in \mathbb{R}^{d_1}$ is the coordinate-wise square of $\mathbf{S}_{1i}$, $\boldsymbol{\delta}_{1i} \sim^{\text{iid}} N_{d_2}(0, \mathbf{I}_{d_2})$, and a matrix Q is defined with $\{Q(1)\}_{i,j} = 0.8^{|i-j|}\mathbb{1}\{|i - j| \leq 1\}$ and $\{Q(0)\}_{i,j} = 0.7^{|i-j|}\mathbb{1}\{|i - j| \leq 2\}$ for $i \leq d_2$ and $j \leq d_1$. The treatment indicators at the second exposure are generated as $A_{2i}|(\bar{\mathbf{S}}_{2i}, A_{1i}) \sim \text{Bernoulli}(A_{1i}\tilde{g}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}) + (1 - A_{1i})\tilde{g}(-\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}))$, where $\tilde{g}(u) := (|u + 1| + 0.1)/(|u + 1| + 1)$. Lastly, $Y_i(1,1) = \bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha} + 1 + \epsilon_i$, $Y_i(0,0) = -\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha} - 1 + \epsilon_i$ and $\epsilon_i \sim^{\text{iid}} N(0, 1)$. We consider $\boldsymbol{\alpha} = (1, \mathbf{0}_{(d_1-1)}, 0.5, 0.5, 0.5, 0.5, \mathbf{0}_{(d_2-4)})^\top$, $\boldsymbol{\gamma} = (1, 1, \mathbf{0}_{(d_1-2)})^\top$ and $\boldsymbol{\delta} = (1, \mathbf{0}_{(d_1-1)}, 0.5, 0.5, 0.5, 0.5, \mathbf{0}_{(d_2-4)})^\top$.

Setting (b): Non-linear $\mu(\cdot)$ and non-linear $\nu(\cdot)$ At the first exposure, generate covariates from a centered Beta distribution, i.e., $\mathbf{S}_{1ij} \sim^{\text{iid}} \text{Beta}(1, 2) - 1/3$ for each $i \leq N$ and $j \leq d_1$; generate $A_{1i}|\mathbf{S}_{1i} \sim \text{Bernoulli}(g(\mathbf{S}_{1i}^\top \boldsymbol{\gamma}))$. At the second exposure, generate $\mathbf{S}_{2i} =$

$W(A_{1i})\mathbf{S}_{1i}+A_{2i}\mathbf{1}_{(d_2)}+\boldsymbol{\delta}_i$ and $A_{2i}|(\bar{\mathbf{S}}_{2i}, A_{1i}) \sim \text{Bernoulli}(A_{1i}g(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta})+(1-A_{1i})g(-\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}))$, where $\boldsymbol{\delta}_{ij} \sim^{\text{iid}} \text{Beta}(1,4)-1/5$, $\{W(1)\}_{i,j} = 0.2^{|i-j|}\mathbb{1}\{|i-j| \le 1\}$ and $\{W(0)\}_{i,j} = 0.2^{|i-j|}\mathbb{1}\{|i-j| \le 2\}+0.1\mathbb{1}\{|i-j| = 2\}$ for each $i \le d_2$ and $j \le d_1$. Here, $Y_i(1,1) = \bar{\mathbf{S}}_{2i}^\top\boldsymbol{\alpha}-1+2r_i+\epsilon_i$, $Y_i(0,0) = -\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\alpha}+1-2r_i+\epsilon_i$ and $\epsilon_i \sim^{\text{iid}} N(0,1)$. Here, we consider non-linear signals with r_i as the standardized version of $\mathbf{S}_{1i1}\mathbf{S}_{1i2}\mathbb{1}\{\mathbf{S}_{1i2} > 0.3\}+\mathbf{S}_{1i1}\mathbf{S}_{1i3}\mathbb{1}\{\mathbf{S}_{1i1} > 0.3\}+\mathbf{S}_{1i2}\mathbf{S}_{1i3}\mathbb{1}\{\mathbf{S}_{1i1} > 0.3\}$. The parameters are $\boldsymbol{\alpha} = (-1,0,0,1/18,\mathbf{0}_{(d_1-4)},-1,-1,-1,\mathbf{0}_{(d_2-3)})^\top$, $\boldsymbol{\gamma} = (1,1,\mathbf{0}_{(d_1-2)})^\top$ and $\boldsymbol{\delta} = (-2,-2,\mathbf{0}_{(d_1+d_2-2)})^\top$.

For each of the settings, we consider the following choices of the dimensions: $(d_1,d_2) \in \{(10,10),(100,50)\}$, with sample sizes N varying from 4000 to 16000. The experiments are repeated 200 times. The proposed SMDR estimator is denoted as SMDR1; see Algorithm 3 with $\mathbb{K} = 5$. We also report a slightly different version, SMDR2, which constructs all the nuisances on the whole sub-sample of $\mathcal{I}_{-k}$ in Steps 4-7 of Algorithm 3. We also report the standard IPW estimator, where PS and OR models are estimated using ℓ_1-regularized logistic and Lasso estimators and no cross-fitting is performed. In addition, we consider the Sequential Doubly Robust Lasso (S-DRL) estimator [BJZ21] and two version of the Dynamic Treatment Lasso estimator [BJZ21, BHL22] (also with $\mathbb{K} = 5$), denoted as DTL2 and DTL1; DTL2's nuisances use samples in $\mathcal{I}_{-k}$, whereas, DTL1's use different sub-samples in $\mathcal{I}_{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\delta}}, \mathcal{I}_{\boldsymbol{\alpha}}, \mathcal{I}_{\boldsymbol{\beta}}$. Here, DTL1 and SMDR1 share the same type of sample splitting; DTL2 and SMDR2 share the same type of sample splitting. The tuning parameters are all chosen through 5-fold cross-validations. In addition, 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 DR estimator, $\widehat{\theta}_{\text{oracle}}$, which uses DR score with correct nuisance functions.

The results are reported in Tables 2.2-2.5.

Table 2.2: Simulation under Setting (a) with $d_1 = d_2 = 10$. 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. All the reported values (except Coverage) are based on robust (median-type) estimates. N_1 and N_0 denote the expected numbers of observations in the treatment groups $(1, 1)$ and $(0, 0)$, respectively.

Method	Bias	RMSE	Length	Coverage	Bias	RMSE	Length	Coverage
	$N = 4000, N_1 = 1368, N_0 = 678$				$N = 8000, N_1 = 2736, N_0 = 1355$			
oracle	-0.002	0.087	0.534	0.940	-0.005	0.070	0.387	0.955
empdiff	-0.393	0.393	0.155	0.015	-0.402	0.402	0.110	0.000
IPW	0.977	0.977	0.706	0.055	0.959	0.959	0.530	0.010
DTL1	0.181	0.193	0.591	0.760	0.143	0.144	0.430	0.735
DTL2	0.108	0.137	0.620	0.890	0.071	0.097	0.449	0.900
S-DRL	0.102	0.133	0.621	0.900	0.071	0.099	0.450	0.905
SMDR1	0.049	0.127	0.558	0.900	0.018	0.078	0.401	0.945
SMDR2	0.004	0.098	0.553	0.935	0.004	0.064	0.400	0.935

Due to the confounding factors, the naive empirical difference estimator $\widehat{\theta}_{\text{empdiff}}$ is not consistent with large biases and poor coverage; see Tables 2.2-2.5. The IPW estimator also has very large biases and provides bad coverage results under Setting (a), where the PS model at the second exposure is misspecified. In Setting (b) where both PS models are correctly specified, surprisingly, the IPW estimator provides acceptable coverages although there is no theoretical guarantees from existing work in high dimensions. However, whenever $d_1 = 100$ and $d_2 = 50$, the RMSEs of IPW are comparable with SMDR1 and SMDR2 when $N = 12000$, and worse than SMDR1 and SMDR2 when $N = 16000$; see Table 2.5.

The DTL1 estimator has relatively poor performance overall, especially when $d_1 = 100$ and $d_2 = 50$: bias is often close to RMSE and coverages are far below the desired 95%. The bad performance mainly results from two reasons: 1) The DTL2 estimators are only

Table 2.3: Simulation under Setting (a) with $d_1 = 100, d_2 = 50$. The rest of the caption details remain the same as those in Table 2.2.

Method	Bias	RMSE	Length	Coverage	Bias	RMSE	Length	Coverage
	$N = 12000, N_1 = 4103, N_0 = 2033$				$N = 16000, N_1 = 5471, N_0 = 2710$			
oracle	-0.002	0.053	0.317	0.945	0.007	0.056	0.276	0.955
empdiff	-0.401	0.401	0.090	0.000	-0.397	0.397	0.078	0.000
IPW	0.946	0.946	0.418	0.000	0.957	0.957	0.369	0.000
DTL1	0.243	0.243	0.329	0.260	0.212	0.212	0.293	0.235
DTL2	0.137	0.141	0.355	0.670	0.122	0.122	0.314	0.655
S-DRL	0.143	0.143	0.356	0.650	0.123	0.123	0.313	0.670
SMDR1	0.053	0.075	0.311	0.890	0.048	0.069	0.269	0.920
SMDR2	0.020	0.058	0.319	0.935	0.013	0.053	0.277	0.925

shown to be consistent when model misspecification occurs [BJZ21], they are not necessarily $\sqrt{N}$-consistent nor asymptotically normal; 2) The sample splitting method of DTL1 is not efficient in finite samples – only 1/5 of the samples are used to obtain each nuisance estimator when $\mathbb{K} = 5$. DTL2 is constructed using a more efficient sample splitting. It provides smaller biases than DTL1, however fails to reach satisfactory coverage guarantees in high dimensions; see Tables 2.3 and 2.5. The S-DRL estimator is constructed similarly as DTL2, except with a different doubly robust estimation strategy for the first OR model. The S-DRL method provides RMSEs similar to (see Tables 2.2-2.4) or smaller than (see Tables 2.5) the DTL2 estimator. It also provides relatively satisfactory coverages when $d_1 = d_2 = 10$ (see Tables 2.2 and 2.4), but in high dimensions, the coverages are far below the desired 95% (see Tables 2.3 and 2.5).

The proposed SMDR1 estimator outperforms DTL1, DTL2, S-DRL, and IPW in the sense of estimation – smaller biases as well as RMSEs are observed in all considered settings; see Tables 2.2-2.5. As for the inference results, SMDR1 outperforms DTL1, DTL2, and

Table 2.4: Simulation under Setting (b) with $d_1 = 10 = d_2 = 10$. The rest of the caption details remain the same as those in Table 2.2.

Method	Bias	RMSE	Length	Coverage	Bias	RMSE	Length	Coverage
	$N = 4000, N_1 = 1088, N_0 = 909$				$N = 8000, N_1 = 2176, N_0 = 1817$			
oracle	-0.005	0.056	0.332	0.950	0.003	0.042	0.235	0.955
empdiff	-0.095	0.107	0.119	0.290	-0.100	0.101	0.085	0.210
IPW	0.027	0.076	0.469	0.975	0.014	0.049	0.337	0.980
DTL1	-0.096	0.104	0.436	0.845	-0.074	0.078	0.308	0.850
DTL2	-0.057	0.081	0.434	0.925	-0.047	0.064	0.312	0.935
S-DRL	-0.053	0.081	0.431	0.940	-0.039	0.063	0.309	0.935
SMDR1	-0.033	0.073	0.404	0.915	-0.017	0.050	0.285	0.945
SMDR2	-0.005	0.071	0.397	0.960	-0.004	0.048	0.281	0.950

Table 2.5: Simulation under Setting (b) with $d_1 = 100, d_2 = 50$. The rest of the caption details remain the same as those in Table 2.2.

Method	Bias	RMSE	Length	Coverage	Bias	RMSE	Length	Coverage
	$N = 12000, N_1 = 3264, N_0 = 2726$				$N = 16000, N_1 = 4352, N_0 = 3634$			
oracle	-0.005	0.034	0.192	0.950	0.002	0.030	0.166	0.960
empdiff	-0.103	0.103	0.069	0.135	-0.094	0.094	0.060	0.125
IPW	0.034	0.048	0.274	0.945	0.033	0.045	0.237	0.965
DTL1	-0.123	0.123	0.237	0.475	-0.100	0.100	0.210	0.540
DTL2	-0.072	0.073	0.246	0.775	-0.060	0.062	0.217	0.880
S-DRL	-0.060	0.069	0.246	0.790	-0.048	0.051	0.215	0.815
SMDR1	-0.035	0.051	0.227	0.890	-0.018	0.037	0.198	0.950
SMDR2	-0.013	0.047	0.228	0.940	0.000	0.034	0.199	0.935

S-DRL in Setting (a) overall, whereas coverages of DTL2, S-DRL, and SMDR1 are close to each other when $N = 4000$ and $d_1 = d_2 = 10$; however, note that DTL2 and S-DRL uses more samples than SMDR1 in nuisances' estimation. In high-dimensional settings, coverage of SMDR1 outperforms DTL1, DTL2, and S-DRL. Overall, we can see that SMDR1 provides coverages close to 95% for large enough N ($N = 8000$ when $d_1 = d_2 = 10$ and $N = 16000$

when $d_1 = 100$, $d_2 = 50$) – such a result coincides with our theory. Here, although the total sample size $N = 16000$ looks large compared with the dimension $d = d_1 + d_2$ when $d_1 = 100$ and $d_2 = 50$, the "effective sample size" in nuisance parameters estimation is only, e.g., $N_0/5 = 542$ for the treatment path $(0, 0)$ under Setting (a). Lastly, as the SMDR2 estimator uses samples more efficiently, it outperforms SMDR1 in finite samples. However, from the theoretical perspective, we believe that the SMDR2 estimator may require more stringent sparsity conditions compared with SMDR1.

2.5.2 A semi-synthetic analysis based on the National Job Corps Study (NJCS)

In this section, we compare the estimation and inference performance of the DTE estimators through semi-synthetic experiments. We consider a dataset from the National Job Corps Study, which is the largest and most comprehensive job training program in the US established in 1964, and serves approximately 50,000 disadvantaged youths aged 16-24 each year by providing vocational training and academic education. A detailed description of the original design and main effects can be found in [SBM08] and [Sch01].

We consider a dataset of 11,313 individuals, with 6,828 assigned to the Job Corps and 4,485 not. Treatments, denoted as $Z_{ti} \in \{0, 1, 2, 3\}$ $(t \in 1, 2)$, are assigned to the ith individual in the first and second years after the initial randomization, where $Z_{ti} = 0$ represents non-enrollment, $Z_{ti} = 1$ enrollment without program participation, $Z_{ti} = 2$ high-school-level education, and $Z_{ti} = 3$ vocational training. The baseline covariate vector, $\mathbf{S}_{1i}$, has 909 characteristics, while $\mathbf{S}_{2i}$ includes 1,427 characteristics that are evaluated before the second-year treatment assignment. We exclude 2,610 individuals whose treatment stages are

missing completely at random [SBRJ+03], resulting in a final sample of 8,703 individuals. We also exclude the binary characteristics, if the 0/1 groups are extremely unbalanced in that the minority group's size is less than 10 within the 8703 individuals, resulting in the final $\mathbf{S}_{1i}$ with 891 characteristics and $\mathbf{S}_{2i}$ with 1350 characteristics. After standardizing the covariates, we generate the potential outcomes $\widetilde{Y}_i(z)$ corresponding to treatment paths $z = (z_1, z_2) \in \{0, 1, 2, 3\}^2$ based on Settings (a) and (b) below. The observed outcome is generated as $Y_i = Y_i(Z_{i1}, Z_{i2}) = \sum_{z\in\{0,1,2,3\}^2} \mathbb{1}_{\{(Z_{1i},Z_{2i})=z\}} Y_i(z)$.

We consider estimation of the DTE, $\theta = E\{\widetilde{Y}_i(z)\} - E\{\widetilde{Y}_i(z')\}$, focusing on the treatment path $z = (z_1, z_2) = (3, 3)$ and the control path $z' = (z'_1, z'_2) = (1, 1)$. To estimate the expected potential outcome $E\{\widetilde{Y}_i(z)\}$, we set $A_{1i} = \mathbb{1}_{\{Z_{1i}=z_1\}}$, $A_{2i} = \mathbb{1}_{\{Z_{2i}=z_2\}}$, and $Y_i(1, 1) = \widetilde{Y}_i(z)$. Then $E\{\widetilde{Y}_i(z)\} = E\{Y_i(1, 1)\}$ can be estimated using Algorithm 3. The control arm $E\{\widetilde{Y}_i(z')\} = E\{Y_i(0, 0)\}$ can be estimated analogously, and the final DTE estimator is constructed as the difference of the obtained estimates. Let $\epsilon_i \sim^{\text{iid}} N(0, 1)$. The potential outcomes $Y_i(1, 1) = \widetilde{Y}_i(z)$ and $Y_i(0, 0) = \widetilde{Y}_i(z')$ are generated as below.

Setting (a): Linear $\nu(\cdot)$. $Y_i(1, 1) = \bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha} + \epsilon_i$ and $Y_i(0, 0) = -\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha} + \epsilon_i$, where $\boldsymbol{\alpha} = 0.5 \cdot (\alpha_0, \mathbf{1}_{(8)}, \mathbf{0}_{(d_1-8)}, \mathbf{1}_{(4)}, \mathbf{0}_{(d_2-4)})^\top$ with α_0 varying from 0.15 to 0.3.

Setting (b): Non-linear $\nu(\cdot)$ $Y_i(1, 1) = \mathbf{S}_{1i}^\top \boldsymbol{\alpha}_1 + (\mathbf{S}_{2i}^2 - 1)^\top \boldsymbol{\alpha}_2 + \epsilon_i$ and $Y_i(0, 0) = -\mathbf{S}_{1i}^\top \boldsymbol{\alpha}_1 - (\mathbf{S}_{2i}^2 - 1)^\top \boldsymbol{\alpha}_2 + \epsilon_i$, where $\mathbf{S}_{2i}^2$ is the coordinate-wise square of $\mathbf{S}_{2i}$, $\boldsymbol{\alpha}_1 = 0.5 \cdot (\alpha_0, \mathbf{1}_{(8)}, \mathbf{0}_{(d_1-8)})^\top$, $\boldsymbol{\alpha}_2 = 0.05 \cdot (\mathbf{1}_{(4)}, \mathbf{0}_{(d_2-4)})^\top$, and α_0 varies from 0.25 to 0.5.

For each setting, we implement the DTL2, S-DRL, and the proposed SMDR1 estimators (see Section 2.5.1). The results are reported in Table 2.6, where the biases are calculated

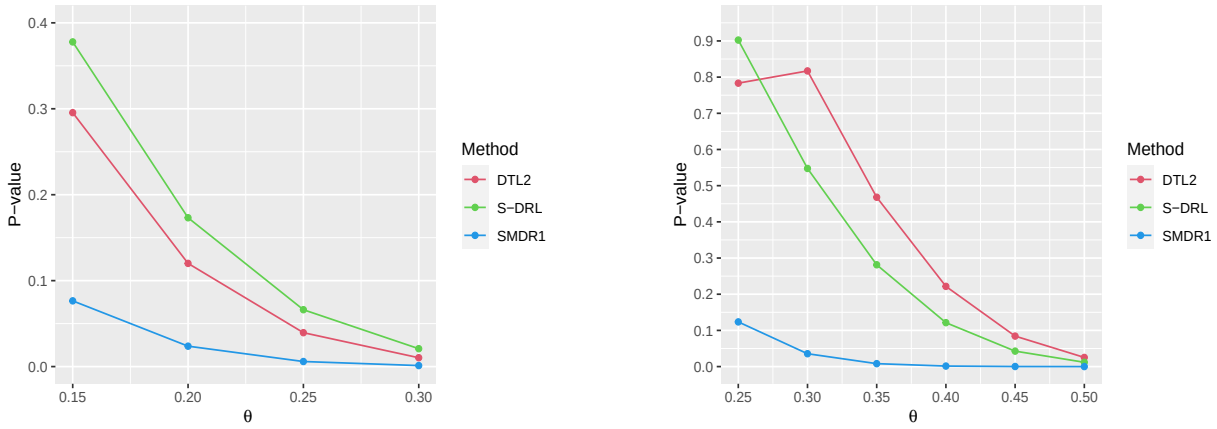
based on the oracle difference-in-mean estimates $\widehat{\theta}_O := N^{-1}\sum_{i=1}^{N}\{Y_i(1,1) - Y_i(0,0)\} = \alpha_0$. Recall that under our simulated outcome setting, we get to see all potential outcomes: two per individual. Under both Settings (a) and (b), the proposed SMDR1 method provides smaller absolute biases than the DTL2 and S-DRL estimators. In addition, under Setting (a), where the OR model at the second exposure is truly linear, all the constructed confidence intervals contain the oracle estimate $\widehat{\theta}_O$. However, when the potential outcome is generated through a quadratic function (under Setting (b)), the oracle estimate $\widehat{\theta}_O$ does not lie in the confidence intervals based on the DTL2 and S-DRL methods; on the other hand, the proposed SMDR1 method leads to confidence intervals containing the oracle estimate. Moreover, considering the hypothesis testing problem with the null $H_0 : \theta = 0$ and the alternative $H_1 : \theta \neq 0$, the reported p-values decay as $\widehat{\theta}_O = \alpha_0$ grows; see Figure 2.1. When α_0 is large enough, all the methods return p-values smaller than 0.05; however, different methods require different signal levels to detect the causal effect and reject the null successfully. Under Setting (a), the proposed SMDR1 method is able to detect the causal effect with a significance level of 95% when $\alpha_0 = 0.2$; however, under the same signal level, both the DTL2 and S-DRL methods fail to reject the null as the corresponding p-values are larger than 0.05. Similarly, under Setting (b) with $\widehat{\theta}_O = \alpha_0 = 0.3$, the proposed SMDR1 method is able to detect the causal effect, whereas the p-value based on the DTL2 and S-DRL methods are both very large. Therefore, we observed a significantly better power in the SMDR1 method than both DTL2 and S-DRL.

Table 2.6: Semi-synthetic analysis. Bias: empirical bias; SE: the standard error; CI: the 95% confidence interval; p-value: the p-value of $H_0 : \theta = 0$ v.s. $H_1 : \theta \neq 0$.

Method	$\widehat{\theta}_O$	$\widehat{\theta}$	Bias	SE	CI	p-value	$\widehat{\theta}_O$	$\widehat{\theta}$	Bias	SE	CI	p-value
	Setting (a)											
DTL2		0.103	-0.047	0.099	[-0.090, 0.297]	0.295		0.153	-0.047	0.099	[-0.040, 0.347]	0.120
S-DRL	0.15	0.092	-0.058	0.104	[-0.101, 0.308]	0.378	0.20	0.142	-0.058	0.104	[-0.051, 0.358]	0.173
SMDR1		0.180	0.030	0.102	[-0.019, 0.380]	0.076		0.230	0.030	0.102	[0.031, 0.430]	0.024
DTL2		0.203	-0.047	0.099	[0.010, 0.397]	0.039		0.253	-0.047	0.099	[0.060, 0.447]	0.010
S-DRL	0.25	0.192	-0.058	0.104	[-0.001, 0.408]	0.066	0.30	0.241	-0.059	0.104	[0.049, 0.458]	0.021
SMDR1		0.280	0.030	0.102	[0.080, 0.480]	0.005		0.330	0.030	0.102	[0.131, 0.530]	0.001
	Setting (b)											
DTL2		-0.027	-0.277	0.100	[-0.223,0.168]	0.783		0.023	-0.277	0.100	[-0.172,0.218]	0.817
S-DRL	0.25	0.013	-0.237	0.105	[-0.232, 0.178]	0.902	0.30	0.063	-0.237	0.105	[-0.182,0.228]	0.548
SMDR1		0.141	-0.109	0.091	[-0.038, 0.320]	0.123		0.192	-0.108	0.091	[0.013, 0.371]	0.035
DTL2		0.072	-0.278	0.100	[-0.123,0.268]	0.468		0.122	-0.278	0.100	[-0.074,0.317]	0.222
S-DRL	0.35	0.113	-0.237	0.105	[-0.133,0.277]	0.281	0.40	0.162	-0.238	0.105	[-0.083,0.327]	0.121
SMDR1		0.242	-0.108	0.091	[0.063, 0.421]	0.008		0.291	-0.109	0.091	[0.111, 0.470]	0.001
DTL2		0.172	-0.278	0.100	[-0.023,0.367]	0.084		0.223	-0.277	0.100	[0.028,0.418]	0.025
S-DRL	0.45	0.212	-0.238	0.105	[-0.033,0.377]	0.043	0.50	0.263	-0.237	0.105	[0.018,0.428]	0.012
SMDR1		0.340	-0.110	0.091	[0.161, 0.519]	0.000		0.406	-0.094	0.092	[0.225, 0.586]	0.000

2.6 Discussion

This paper develops new techniques to establish statistical inference for treatment effects under dynamic, high-dimensional, and possibly misspecified settings. Based on a set of newly proposed loss functions for the nuisance models, we establish root-N inference results as long as at least one nuisance model is correctly specified at each exposure. To the best of our knowledge, this result is more robust than the existing literature, even those focused on low-dimensional cases. Our results indicate the importance of nuisance parameters' estimation – naive, off-the-shelf estimators cannot reach desired robustness level even if coupled with double-robust loss functions. We allow non-parametric regression methods based on linear/logistic forms with basis functions, e.g., B-splines. However, due to the challenge of estimating the nested outcome regression models, the treatment effect's estimation based on other non-parametric estimators, such as random forests and boosting, still needs to be studied.



Setting (a) Setting (b)

Figure 2.1: The p-value of the estimators as θ varies.

In the context of dynamic treatment regimes, a related but different DR property have been explored. However, existing results additionally require correctly specified contrast models at all the later stages of the estimation – such a condition is very restrictive when multiple time exposures are involved and is unnecessary in our work. How should we make decisions if the contrast models cannot be correctly specified at later stages? Our results indicate that the nested OR model $\mu(\cdot)$ can be estimated consistently under such a circumstance (see Theorem 2.3) and hence optimizing $\widehat{\mu}(\cdot)$ seems to be a feasible but conservative option.

2.7 Supplementary Material

We begin by introducing some additional notation used throughout the document. Constants $c, C > 0$, independent of N and d, may change from one line to the other. For any $r > 0$, let $\|f(\cdot)\|_{r,\mathbb{P}} := \{\mathbb{E}|f(\mathbf{Z})|^r\}^{1/r}$. Denote $\mathbf{e}_j$ as the vector whose j-th element is 1 and other elements are 0s. For any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succ \mathbf{B}$ denotes that $\mathbf{A} - \mathbf{B}$ is positive definite and $\mathbf{A} \succeq \mathbf{B}$ denotes that $\mathbf{A} - \mathbf{B}$ is positive semidefinite; denote

$\lambda_{\min}(\mathbf{A})$ as the smallest eigenvalue of $\mathbf{A}$. For the sake of simplicity, we denote $Y_{1,1} := Y(1,1)$ in this supplementary document. Let $\mathbb{S}_{\boldsymbol{\gamma}} = (\mathbf{W}_i)_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}$, $\mathbb{S}_{\boldsymbol{\delta}} = (\mathbf{W}_i)_{i\in\mathcal{I}_{\boldsymbol{\delta}}}$, $\mathbb{S}_{\boldsymbol{\alpha}} = (\mathbf{W}_i)_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}$, and $\mathbb{S}_{\boldsymbol{\beta}} = (\mathbf{W}_i)_{i\in\mathcal{I}_{\boldsymbol{\beta}}}$ be the subsets of $\mathbb{S}$ corresponding to the index sets $\mathcal{I}_{\boldsymbol{\gamma}}$, $\mathcal{I}_{\boldsymbol{\delta}}$, $\mathcal{I}_{\boldsymbol{\alpha}}$, and $\mathcal{I}_{\boldsymbol{\beta}}$, respectively.

2.7.1 Uniqueness of moment-targeted parameters

In this section, we discuss the uniqueness of moment-targeted nuisance parameters $\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ defined in Section 2.2. Note that the parameters are defined through optimization problems, (2.13)-(2.15). We first consider the Hessian matrices of the objective functions:

$$\begin{aligned}
\boldsymbol{\nabla}^2_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\} &= \mathbb{E}\left[A_1\{g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}) - 1\}\mathbf{S}_1\mathbf{S}_1^\top\right], \\
\boldsymbol{\nabla}^2_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\} &= \mathbb{E}\left[A_1A_2g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}) - 1\}\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top\right], \\
\boldsymbol{\nabla}^2_{\boldsymbol{\alpha}}\mathbb{E}\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha})\} &= 2\mathbb{E}\left[A_1A_2g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*) - 1\}\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top\right], \\
\boldsymbol{\nabla}^2_{\boldsymbol{\beta}}\mathbb{E}\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta})\} &= 2\mathbb{E}\left[A_1\{g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*) - 1\}\mathbf{S}_1\mathbf{S}_1^\top\right],
\end{aligned}$$

where $g(u) = \exp(u)/\{1+\exp(u)\}$ is the logistic function. In the following, we prove by contradiction

$$\boldsymbol{\nabla}^2_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\} \succ \mathbf{0}, \ \ \forall\boldsymbol{\gamma}\in\mathbb{R}^{d_1}. \tag{2.29}$$

Assume there exists some $\mathbf{a}, \boldsymbol{\gamma}\in\mathbb{R}^{d_1}$ such that $\mathbf{a}^\top\boldsymbol{\nabla}^2_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\}\mathbf{a} = 0$ and $\|\mathbf{a}\|_2 = 1$. Then $A_1\{g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma})-1\}(\mathbf{S}_1^\top\mathbf{a})^2 = 0$ almost surely. Since $g(u)\in(0,1)$ for all $u\in\mathbb{R}$, $g^{-1}(\mathbf{s}_1^\top\boldsymbol{\gamma})-1 > 0$ for all $\mathbf{s}_1\in\mathbb{R}^{d_1}$ and hence $A_1(\mathbf{S}_1^\top\mathbf{a})^2 = 0$ almost surely. It follows that $\mathbb{E}\{A_1(\mathbf{S}_1^\top\mathbf{a})^2\} = 0$ and hence $\lambda_{\min}(\mathbb{E}(A_1\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top)) \le \lambda_{\min}(\mathbb{E}(A_1\mathbf{S}_1\mathbf{S}_1^\top)) = 0$, which conflicts Assumption 2.4. Therefore,

(2.29) holds, the solution $\boldsymbol{\gamma}^*$ is unique and can be equivalently defined as the solution of the first-order optimality condition $\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma})\} = \mathbf{0}$.

In addition, we also note that

$$\begin{aligned}\boldsymbol{\nabla}^2_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\} &= \mathbb{E}\left(\mathbb{E}\left[A_2 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}) - 1\}\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right)\\ &= \mathbb{E}\left[A_1\rho(\bar{\mathbf{S}}_2) g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}) - 1\}\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top\right]\\ &\succeq c_0\mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}) - 1\}\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top\right]\end{aligned}$$

under Assumption 2.1. In the following, we also prove by contradiction to show that

$$\boldsymbol{\nabla}^2_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\} \succ \mathbf{0}, \ \ \forall\boldsymbol{\delta}\in\mathbb{R}^d. \tag{2.30}$$

Assume there exists some $\mathbf{a}, \boldsymbol{\delta} \in \mathbb{R}^d$ such that $\mathbf{a}^\top\boldsymbol{\nabla}^2_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\}\mathbf{a} = 0$ and $\|\mathbf{a}\|_2 = 1$. Then $A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}) - 1\}(\bar{\mathbf{S}}_2^\top\mathbf{a})^2 = 0$ almost surely. Since $g(u) \in (0,1)$ for all $u \in \mathbb{R}$, $g^{-1}(\mathbf{s}_1^\top\boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{s}}_2^\top\boldsymbol{\delta}) - 1\} > 0$ for all $\bar{\mathbf{s}}_2 \in \mathbb{R}^d$ and hence $A_1(\bar{\mathbf{S}}_2^\top\mathbf{a})^2$ almost surely. It follows that $\mathbb{E}\{A_1(\bar{\mathbf{S}}_2^\top\mathbf{a})^2\} = 0$ and hence $\lambda_{\min}(\mathbb{E}(A_1\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top)) = 0$, which conflicts Assumption 2.4. Therefore, (2.30) holds, the solution $\boldsymbol{\delta}^*$ is unique and can be equivalently defined as the solution of $\boldsymbol{\nabla}_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta})\} = \mathbf{0}$.

Besides, we note that

$$\boldsymbol{\nabla}^2_{\boldsymbol{\alpha}}\mathbb{E}\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha})\} = 2\boldsymbol{\nabla}^2_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\} \succ \mathbf{0}, \ \ \forall\boldsymbol{\alpha}\in\mathbb{R}^d, \tag{2.31}$$

$$\boldsymbol{\nabla}^2_{\boldsymbol{\beta}}\mathbb{E}\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta})\} = 2\boldsymbol{\nabla}^2_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma}^*)\} \succ \mathbf{0}, \ \ \forall\boldsymbol{\beta}\in\mathbb{R}^{d_1}. \tag{2.32}$$

Therefore, the solutions $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ are also unique.

2.7.2 Justifications for Section 2.2

In this section, we provide justifications for identification in Section 2.2. Before we provide detailed justifications, we first introduce equivalent expressions for the OR functions $\mu(\cdot)$ and $\nu(\cdot)$: under Assumption 2.1,

$$\begin{aligned}\mu(\mathbf{s}_1) :=& \mathbb{E}(Y_{1,1} \mid \mathbf{S}_1 = \mathbf{s}_1) \overset{(i)}{=} \mathbb{E}(Y_{1,1} \mid \mathbf{S}_1 = \mathbf{s}_1, A_1 = 1) \\ \overset{(ii)}{=}& \mathbb{E}\{\nu(\bar{\mathbf{S}}_2) \mid \mathbf{S}_1 = \mathbf{s}_1, A_1 = 1\}, \\ \nu(\bar{\mathbf{s}}_2) :=& \mathbb{E}(Y_{1,1} \mid \bar{\mathbf{S}}_2 = \mathbf{s}_2, A_1 = 1) \overset{(iii)}{=} \mathbb{E}(Y_{1,1} \mid \bar{\mathbf{S}}_2 = \mathbf{s}_2, A_1 = 1, A_2 = 1).\end{aligned} \tag{2.33}$$

Here, (i) holds since $Y_{1,1} \perp\!\!\!\perp A_1 \mid \mathbf{S}_1$; (ii) holds by the tower rule; (iii) holds since $Y_{1,1} \perp\!\!\!\perp A_2 \mid \mathbf{S}_1, A_1 = 1$.

Now, the justifications of the cases (a)-(d) for identification are provided below:

(a.1) Assume $\pi^*(\cdot) = \pi(\cdot)$. Then $\pi(\mathbf{s}_1) = \pi^*(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^*)$ and hence $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^0)$ with $\boldsymbol{\gamma}^0 = \boldsymbol{\gamma}^*$.

(a.2) Assume there exists some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$ such that $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^0)$. By the construction of $\boldsymbol{\gamma}^*$ and note that the Hessian matrix satisfies (2.29), $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ is the unique solution of

$$\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \mathbb{E}\{\ell_1(\mathbf{W}; \boldsymbol{\gamma})\} = \mathbb{E}\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma})\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

Meanwhile, we also have

$$\begin{aligned}\mathbb{E}\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^0)\}\mathbf{S}_1\right] &= \mathbb{E}\left(\mathbb{E}\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^0)\}\mathbf{S}_1 \mid \mathbf{S}_1\right]\right) \\ &= \mathbb{E}\left[\{1 - \pi(\mathbf{S}_1) g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^0)\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.\end{aligned}$$

By the uniqueness of $\boldsymbol{\gamma}^*$, we conclude that $\boldsymbol{\gamma}^0 = \boldsymbol{\gamma}^*$ and hence $\pi(\cdot) = \pi^*(\cdot)$.

(b.1) Assume $\rho^*(\cdot) = \rho(\cdot)$. Then $\rho(\bar{\mathbf{s}}_2) = \rho^*(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^*)$ and hence $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^0)$ with $\boldsymbol{\delta}^0 = \boldsymbol{\delta}^*$.

(b.2) Assume there exists some $\boldsymbol{\delta}^0 \in \mathbb{R}^d$ such that $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top \boldsymbol{\delta}^0)$. By the construction of $\boldsymbol{\delta}^*$ and note that the Hessian matrix satisfies (2.30), $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ is the unique solution of

$$\boldsymbol{\nabla}_{\boldsymbol{\delta}} \mathbb{E}\{\ell_2(\mathbf{W}; \boldsymbol{\gamma}^*, \boldsymbol{\delta})\} = \mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta})\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d.$$

Meanwhile, we also have

$$\begin{aligned}
&\mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^0)\}\bar{\mathbf{S}}_2\right] \\
&\quad = \mathbb{E}\left(\mathbb{E}\left[g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^0)\}\bar{\mathbf{S}}_2 \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right) \\
&\quad = \mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{1 - \rho(\bar{\mathbf{S}}_2) g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^0)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d.
\end{aligned}$$

By the uniqueness of $\boldsymbol{\delta}^*$, we conclude that $\boldsymbol{\delta}^0 = \boldsymbol{\delta}^*$ and hence $\rho(\cdot) = \rho^*(\cdot)$.

(c.1) Assume $\nu^*(\cdot) = \nu(\cdot)$. Then $\nu(\bar{\mathbf{s}}_2) = \nu^*(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^*$ and hence $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$ with $\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}^*$.

(c.2) Assume there exists some $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$ such that $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$. By the construction of $\boldsymbol{\alpha}^*$ and note that the Hessian matrix satisfies (2.31), $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ is the unique solution of

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\alpha}} \mathbb{E}\{\ell_3(\mathbf{W}; \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha})\} &= 2\mathbb{E}\left[A_1 A_2 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - 1\}(\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha} - Y)\bar{\mathbf{S}}_2\right] \\
&= \mathbf{0} \in \mathbb{R}^d.
\end{aligned}$$

Meanwhile, we also have

$$
\begin{aligned}
&\mathbb{E}\left[A_1 A_2 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - 1\}(\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^0 - Y)\bar{\mathbf{S}}_2\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[A_2 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - 1\}\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^0 - Y_{1,1}\}\bar{\mathbf{S}}_2 \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left[A_2\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^0 - Y_{1,1}\} \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - 1\}\bar{\mathbf{S}}_2 \mid \bar{\mathbf{S}}_2\right]\right) \\
&\overset{(i)}{=} \mathbb{E}\left[A_1 \rho(\bar{\mathbf{S}}_2) g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - 1\}\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^0 - \nu(\bar{\mathbf{S}}_2)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d,
\end{aligned}
$$

where (i) holds since $Y_{1,1} \perp\!\!\!\perp A_2 \mid (\bar{\mathbf{S}}_2, A_1 = 1)$ under Assumption 2.1. By the uniqueness of $\boldsymbol{\alpha}^*$, we conclude that $\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}^*$ and hence $\nu(\cdot) = \nu^*(\cdot)$.

(d) Assume there exists some $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$ such that $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top \boldsymbol{\beta}^0$ and either $\rho^*(\cdot) = \rho(\cdot)$ or $\nu^*(\cdot) = \nu(\cdot)$ holds. By the construction of $\boldsymbol{\beta}^*$ and note that the Hessian matrix satisfies (2.32), $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ is the unique solution of

$$
\begin{aligned}
&\boldsymbol{\nabla}_{\boldsymbol{\beta}} \mathbb{E}\{\ell_4(\mathbf{W}; \boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta})\} \\
&\quad = 2\mathbb{E}\left[A_1\{g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) - 1\}\left\{\mathbf{S}_1^\top \boldsymbol{\beta} - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)\right\}\mathbf{S}_1\right] \\
&\quad = \mathbf{0} \in \mathbb{R}^{d_1}.
\end{aligned}
$$

Meanwhile, note that

$$
\begin{aligned}
&\mathbb{E}\left[\exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\mathbf{S}_1^\top \boldsymbol{\beta}^0 - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)\right\}\mathbf{S}_1 \mid \bar{\mathbf{S}}_2, A_1 = 1\right] \\
&= \mathbb{E}\left(\exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left[\mu(\mathbf{S}_1) - \nu^*(\bar{\mathbf{S}}_2) - \frac{A_2\{Y_{1,1} - \nu^*(\bar{\mathbf{S}}_2)\}}{\rho^*(\bar{\mathbf{S}}_2)}\right]\mathbf{S}_1 \mid \bar{\mathbf{S}}_2, A_1 = 1\right) \\
&\overset{(i)}{=} \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left[\mu(\mathbf{S}_1) - \nu^*(\bar{\mathbf{S}}_2) - \frac{\rho(\bar{\mathbf{S}}_2)\{\nu(\bar{\mathbf{S}}_2) - \nu^*(\bar{\mathbf{S}}_2)\}}{\rho^*(\bar{\mathbf{S}}_2)}\right]\mathbf{S}_1 \\
&= \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left[\mu(\mathbf{S}_1) - \nu(\bar{\mathbf{S}}_2) + \left\{1 - \frac{\rho(\bar{\mathbf{S}}_2)}{\rho^*(\bar{\mathbf{S}}_2)}\right\}\left\{\nu(\bar{\mathbf{S}}_2) - \nu^*(\bar{\mathbf{S}}_2)\right\}\right]\mathbf{S}_1 \\
&\overset{(ii)}{=} \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \nu(\bar{\mathbf{S}}_2)\}\mathbf{S}_1,
\end{aligned}
$$

where (i) holds since $Y_{1,1} \perp\!\!\!\perp A_2 \mid (\bar{\mathbf{S}}_2, A_1 = 1)$ under Assumption 2.1; (ii) follows since either $\rho^*(\cdot) = \rho(\cdot)$ or $\nu^*(\cdot) = \nu(\cdot)$ holds. Hence, by the tower rule,

$$\begin{aligned}
&\mathbb{E}\left[A_1\{g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) - 1\}\left\{\mathbf{S}_1^\top \boldsymbol{\beta}^0 - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)\right\}\mathbf{S}_1\right] \\
&\quad = \mathbb{E}\left[\exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \nu(\bar{\mathbf{S}}_2)\}\mathbf{S}_1 \mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right] \\
&\quad = \mathbb{E}\left[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \nu(\bar{\mathbf{S}}_2)\}\mathbf{S}_1\right] \\
&\quad = \mathbb{E}\left(\mathbb{E}\left[\exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \nu(\bar{\mathbf{S}}_2)\}\mathbf{S}_1 \mid \mathbf{S}_1, A_1 = 1\right]\mathbb{E}(A_1 \mid \mathbf{S}_1)\right) \\
&\quad \overset{(i)}{=} \mathbf{0} \in \mathbb{R}^{d_1},
\end{aligned}$$

where (i) follows from (2.33). By the uniqueness of $\boldsymbol{\beta}^*$, we conclude that $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^*$ and hence $\mu(\cdot) = \mu^*(\cdot)$.

2.7.3 Auxiliary lemmas

The following Lemmas will be useful in the proofs.

Lemma 2.2 (Lemma A.1 of [ZCB21]). *Let $(X_N)_{N\geq 1}$ and $(Y_N)_{N\geq 1}$ be sequences of random variables in $\mathbb{R}$. If $\mathbb{E}(|X_N|^r|Y_N) = O_p(1)$ for any $r \geq 1$, then $X_N = O_p(1)$.*

Lemma 2.3 (Lemma S.4 of [BJZ21]). *Let Assumptions 2.1 and 2.4 hold. Then the smallest eigenvalues of $\mathbb{E}(A_1\mathbf{S}_1\mathbf{S}_1^\top)$ and $\mathbb{E}(A_1A_2\bar{\mathbf{S}}_2\bar{\mathbf{S}}_2^\top)$ are both lower bounded by some constant $c'_{\min} > 0$. Additionally, $\|\mathbf{v}^\top \mathbf{S}_1\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$, $\|A_1\mathbf{v}^\top \mathbf{S}_1\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^{d_1}$ and $\|A_1A_2\mathbf{v}^\top \bar{\mathbf{S}}_2\|_{\psi_2} \leq \sigma'_{\mathbf{S}}\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^d$, with some constant $\sigma'_{\mathbf{S}} > 0$.*

Lemma 2.4 (Lemma D.1 (ii), (iv) and (vi) of [CLCL19]). *Let $X, Y \in \mathbb{R}$ be a random variable. If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_2} \leq \|Y\|_{\psi_2}$. Then $\|cX\|_{\psi_2} = |c|\|X\|_{\psi_2}$ $\forall c \in \mathbb{R}$. If $\|X\|_{\psi_2} \leq \sigma$, then $E(X) \leq \sigma\sqrt{\pi}$ and $E(|X|^m) \leq 2\sigma^m(m/2)^{m/2}$ for all $m \geq 2$. Let $\{X_i\}_{i=1}^n$ be*

random variables (possibly dependent) with $\max_{1\le i\le n}\|X_i\|_{\psi_2}\le\sigma$, *then* $\|\max_{1\le i\le n}|X_i|\|_{\psi_2}\le\sigma(\log n+2)^{1/2}$.

Lemma 2.5 (Corollary 2.3 of [DVDGVW10])**.** *Let* $\{X_i\}_{i=1}^n$ *be identically distributed, then*

$$\mathbb{E}\left[\left\|n^{-1}\sum_{i=1}^n X_i\right\|_\infty^2\right]\le n^{-1}(2e\log d-e)\mathbb{E}\left[\|X_i\|_\infty^2\right].$$

Lemma 2.6. *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.,* $\|\mathbf{a}^\top\mathbf{U}\|_{\psi_2}\le\sigma_{\mathbf{U}}\|\mathbf{a}\|_2$ *for all* $\mathbf{a}\in\mathbb{R}^d$ *with some constant* $\sigma_{\mathbf{U}}>0$. *Additionally, suppose the smallest eigenvalue of* $\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ *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* $\mathbb{E}\{|\mathbf{U}^\top\boldsymbol{\eta}|^c\}<C$ *with some constants* $c,C>0$, *there exists constants* $\kappa_1,\kappa_2,c_1,c_2>0$, *such that*

$$\begin{aligned}&\mathbb{P}_{\mathbb{S}'}\left(M^{-1}\sum_{i\in\mathcal{J}}\phi(\mathbf{U}_i^\top(\boldsymbol{\eta}+v\boldsymbol{\Delta}))(\mathbf{U}_i^\top\boldsymbol{\Delta})^2\ge\kappa_1\|\boldsymbol{\Delta}\|_2^2-\kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2,\ \ \forall\|\boldsymbol{\Delta}\|_2\le1\right)\\&\qquad\ge1-c_1\exp(-c_2M).\end{aligned}\tag{2.34}$$

Lemma 2.6 follows directly by repeating the proof of Lemma 4.3 of [ZCB21]; see also for other slightly different versions in Proposition 2 of [NRWY10] and Theorem 9.36 and Example 9.17 of [Wai19].

Lemma 2.7. *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* $S\subseteq\{1,\dots,d_1\}$ *and* $s=|S|$. *Then, as* $m,d\to\infty$,

$$\sup_{\boldsymbol{\Delta}\in\{\boldsymbol{\Delta}_{S^c}=\mathbf{0},\|\boldsymbol{\Delta}\|_2=1\}}\left|m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2-\mathbb{E}\{(\mathbf{X}^\top\boldsymbol{\Delta})^2\}\right|=O_p\left(\sqrt{\frac{s}{m}}\right).$$

If we further assume that $S\subset\{1,\dots,d\}$. *Then, as* $m,d\to\infty$,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\Delta}\|_2=1}\left|m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2-\mathbb{E}\{(\mathbf{X}^\top\boldsymbol{\Delta})^2\}\right|=O_p\left(\sqrt{\frac{s}{m}}\right),$$

where $\mathbb{C}(S,3) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\Delta}_S\|_1\}$.

Lemma 2.7 is an analog of Lemmas 15 and 16 of [BWZ19]. It can be shown by repeating the proof of [BWZ19], with replacing $\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^\top$ therein by $\mathbb{E}(\mathbf{X}\mathbf{X}^\top)$.

Lemma 2.8. *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* $m, d \to \infty$*,*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

We introduce notation needed for the next set of results. For any $s, k > 0$, define $\widetilde{C}(s,k) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \leq k\sqrt{s}\|\boldsymbol{\Delta}\|_2\}$ 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

$$\mathcal{F}(\boldsymbol{\Delta}) := \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta} - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1,$$

where $\bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta})$ and $\delta\bar{\ell}_2(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\Delta})$ are defined in (2.56) and (2.59), respectively. 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 with a different cone set $\widetilde{C}(s,k)$ (and $\widetilde{K}(s,k,1)$) defined above.

Lemma 2.9. *Let Assumptions 2.1 and 2.4 hold,* $s_{\boldsymbol{\gamma}} = o(N/\log d_1)$*,* $s_{\boldsymbol{\delta}} = o(N/\log d)$*, and consider some* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1/N}$*. For any* $0 < t < \kappa_1^2 M/(16^2\sigma_{\boldsymbol{\delta}}^2 s_{\boldsymbol{\delta}})$*, let* $\lambda_{\boldsymbol{\delta}} = 2\sigma_{\boldsymbol{\delta}}\sqrt{(t+\log d)/M}$*. Define*

$$\mathcal{A}_1 := \{\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\|_\infty \leq \lambda_{\boldsymbol{\delta}}/2\}, \tag{2.35}$$

$$\mathcal{A}_2 := \left\{|R_1(\boldsymbol{\Delta})| \leq c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\left(\frac{\|\boldsymbol{\Delta}\|_1}{\sqrt{N}} + \|\boldsymbol{\Delta}\|_2\right),\ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right\}, \tag{2.36}$$

$$\mathcal{A}_3 := \left\{\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2,\ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_2 \leq 1\right\}, \tag{2.37}$$

where $R_1(\boldsymbol{\Delta}) := \left\{\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\}^\top \boldsymbol{\Delta}$ *and* $c > 0$ *is some constant. Let* $\bar{s}_{\boldsymbol{\delta}} := s_{\boldsymbol{\gamma}} \log d_1 / \log d + s_{\boldsymbol{\delta}}$. *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* $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - t - 2\exp(-t)$.

As the nuisance estimator $\widehat{\boldsymbol{\delta}}$ is constructed based off of $\widehat{\boldsymbol{\gamma}}$, using the standard proof techniques developed for Lasso-type estimators, we would need to take care of the term $\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}}$ as in (2.101), where $\boldsymbol{\Delta}_{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$ and the gradient $\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)$ is evaluated at the target value $\boldsymbol{\delta}^*$ and the estimated value $\widehat{\boldsymbol{\gamma}}$. Under a possible model misspecification, where $\rho(\cdot) \neq \rho^*(\cdot)$, unlike part (a) of Lemma 2.19 below, we cannot provide a small enough upper bound for the infinity norm $\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\|_\infty$ as $\mathbb{E}_{\mathbb{S}_{\boldsymbol{\delta}}}\{\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\}$ might be different from zero. Hence, instead of directly using the inequality $|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}}| \leq \|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\|_\infty \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1$, here we consider $|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}}| = |\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}} + R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}})| \leq \|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\|_\infty \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 + |R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}})|$. In Lemma 2.9, we can show that $\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\|_\infty = O_p(\sqrt{\log d / N})$ and the term $R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}})$ can be controlled through the imputation error originated from the estimation of $\boldsymbol{\gamma}^*$ as well as the ℓ_1- and ℓ_2-norms of $\boldsymbol{\Delta}_{\boldsymbol{\delta}}$. Together with the restricted strong convexity (RSC) condition developed in (2.37), we further show that the error term $\boldsymbol{\Delta}_{\boldsymbol{\delta}}$ belongs to a cone set $\widetilde{C}(s, k)$ in the following lemma.

Lemma 2.10. *Let Assumptions 2.1 and 2.4 hold. Let* $s_{\boldsymbol{\gamma}} = o(N / \log d_1)$ *and consider some* $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1 / N}$. *Define* $\boldsymbol{\Delta}_{\boldsymbol{\delta}} := \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*$. *For any* $t > 0$, *let* $\lambda_{\boldsymbol{\delta}} = 2\sigma_{\boldsymbol{\delta}}\sqrt{(t + \log d)/M}$. *Events* $\mathcal{A}_1$ *and* $\mathcal{A}_2$ *are defined in* (2.35) *and* (2.36). *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 $\bar{s}_{\boldsymbol{\delta}} := s_{\boldsymbol{\gamma}} \log d_1 / \log d + s_{\boldsymbol{\delta}}$.

Lemma 2.11. *Let the assumptions in Lemma 2.9 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 2.12. *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 2.13. *Let the assumptions in part (a) of Theorem 2.3 hold. Let $r > 0$ be any positive constant. Then, as $N, d_1, d_2 \to \infty$,*

$$\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} &= \left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} \\ &= O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right). \end{aligned} \tag{2.38}$$

Define

$$\mathcal{E}_1 := \left\{ \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1 \text{ and } \left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},12} \leq C,\ \forall \boldsymbol{\gamma} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\} \right\}. \tag{2.39}$$

Then, as $N, d_1, d_2 \to \infty$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1 - o(1).$$

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

$$\left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} \leq C, \quad \left\|\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} \leq C, \quad \left\|\exp(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} \leq C',$$

with some constant $C' > 0$.

Lemma 2.14. *Let $r > 0$ be any positive constant.*

(a) Let the assumptions in part (b) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} &= \left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} \\ &= O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right). \end{aligned} \tag{2.40}$$

(b) Let the assumptions in part (a) of Theorem 2.4 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right),$$

and

$$\begin{aligned}\left\|\exp(-\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} &= \left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} \\ &= O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right). \end{aligned} \tag{2.41}$$

Let either (a) or (b) holds. Let $C > 0$ be some constant, define

$$\mathcal{E}_2 := \left\{\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1 \text{ and } \left\|g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})\right\|_{\mathbb{P},6} \le C,\ \forall \boldsymbol{\delta} \in \left\{w\boldsymbol{\delta}^* + (1-w)\widehat{\boldsymbol{\delta}} : w \in [0,1]\right\}\right\}. \tag{2.42}$$

Then, as $N, d_1, d_2 \to \infty$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{E}_2) = 1 - o(1).$$

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

$$\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})\right\|_{\mathbb{P},r'} \le C, \quad \left\|\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})\right\|_{\mathbb{P},r'} \le C, \quad \left\|\exp(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})\right\|_{\mathbb{P},r'} \le C',$$

with some constant $C' > 0$.

Lemma 2.15. *Let $r > 0$ be any positive constant.*

(a) Let the assumptions in part (c) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}}\right).$$

(b) Let the assumptions in part (b) of Theorem 2.4 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\alpha}} \log d}{N}}\right).$$

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 $\widetilde{\varepsilon} := Y_{1,1} - \bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\alpha}}$. Then, for any constant $r > 0$, $\|\widetilde{\varepsilon}\|_{\mathbb{P},r} = O_p(1)$.

Lemma 2.16. *Let $r > 0$ be any positive constant.*

(a) Let the assumptions in part (d) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right) = O_p\left(\sqrt{\frac{(s_{\boldsymbol{\gamma}} + s_{\boldsymbol{\beta}}) \log d_1 + (s_{\boldsymbol{\delta}} + s_{\boldsymbol{\alpha}}) \log d}{N}}\right).$$

(b) Let the assumptions in part (c) or part (d) or part (e) of Theorem 2.4 hold. Then, as $N, d_1, d_2 \to \infty$,

$$\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right)$$

Let either (a) or (b) holds, and let either (a) or (b) of 2.15 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 $\widetilde{\zeta} := \bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\alpha}} - \mathbf{S}_1^\top \widetilde{\boldsymbol{\beta}}$. Then, for any constant $r > 0$, $\|\widetilde{\zeta}\|_{\mathbb{P},r} = O_p(1)$.

2.7.4 Proofs of the main results

Proofs of the results in Section 2.1

Proof of Lemma 2.1. We first notice that

$$\mathbb{E}\left\{\frac{A_1A_2Y}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} \overset{(i)}{=} \mathbb{E}\left\{\frac{A_1A_2Y_{1,1}}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} \overset{(ii)}{=} \mathbb{E}\left[\mathbb{E}\left\{\frac{A_2Y_{1,1}}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)} \mid \bar{\mathbf{S}}_2, A_1 = 1\right\}\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right]$$
$$\overset{(iii)}{=} \mathbb{E}\left\{\frac{\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right\} \overset{(iv)}{=} \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\},$$

where (i) holds since $A_1A_2Y = A_1A_2Y_{1,1}$ under Assumption 2.1; (ii) and (iv) hold by the tower rule; (iii) holds since $A_2 \perp\!\!\!\perp Y_{1,1} \mid \bar{\mathbf{S}}_2, A_1 = 1$ under Assumption 2.1. We consider the following four cases:

(a) Both PS models are logistic, i.e., $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top\boldsymbol{\gamma}^0)$ and $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top\boldsymbol{\delta}^0)$ with some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$ and $\boldsymbol{\delta}^0 \in \mathbb{R}^d$. Then

$$\mathbb{E}\left\{\frac{A_1A_2Y}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} = \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} \overset{(i)}{=} \mathbb{E}\left\{\frac{A_1\nu(\bar{\mathbf{S}}_2)}{\pi(\mathbf{S}_1)}\right\}$$
$$\overset{(ii)}{=} \mathbb{E}\left[\mathbb{E}\left\{\frac{\nu(\bar{\mathbf{S}}_2)}{\pi(\mathbf{S}_1)} \mid \mathbf{S}_1, A_1 = 1\right\}\mathbb{E}(A_1 \mid \mathbf{S}_1)\right] \overset{(iii)}{=} \mathbb{E}\left\{\mu(\mathbf{S}_1)\right\} \overset{(iv)}{=} \theta_{1,1},$$

where (i) holds by $\pi^*(\cdot) = \pi(\cdot)$ and $\rho^*(\cdot) = \rho(\cdot)$ since (a.2) and (b.2); (ii) and (iv) hold by the tower rule; (iii) holds by (2.33).

(b) The OR model at time 1 is linear and the PS model at time 2 is logistic, i.e., $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top\boldsymbol{\beta}^0$ and $\rho(\bar{\mathbf{s}}_2) = g(\bar{\mathbf{s}}_2^\top\boldsymbol{\delta}^0)$ with some $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$ and $\boldsymbol{\delta}^0 \in \mathbb{R}^{d_1}$. Then

$$\mathbb{E}\left\{\frac{A_1A_2Y}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} = \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} \overset{(i)}{=} \mathbb{E}\left\{\frac{A_1\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)}\right\}$$
$$\overset{(ii)}{=} \mathbb{E}\left[\mathbb{E}\left\{\frac{\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)} \mid \mathbf{S}_1, A_1 = 1\right\}\mathbb{E}(A_1 \mid \mathbf{S}_1)\right] \overset{(iii)}{=} \mathbb{E}\left\{\frac{A_1\mu(\mathbf{S}_1)}{\pi^*(\mathbf{S}_1)}\right\}$$
$$= \mathbb{E}\left\{\frac{A_1\mathbf{S}_1^\top}{\pi^*(\mathbf{S}_1)}\right\}\boldsymbol{\beta}^0 \overset{(iv)}{=} \mathbb{E}\left(\mathbf{S}_1^\top\boldsymbol{\beta}^0\right) = \mathbb{E}\{\mu(\mathbf{S}_1)\} \overset{(v)}{=} \theta_{1,1},$$

where (i) holds by $\rho^*(\cdot) = \rho(\cdot)$ since (b.2); (ii) and (v) hold by the tower rule; (iii) holds by (2.33); (iv) holds since $\mathbb{E}[\{1 - A_1/\pi^*(\mathbf{S}_1)\}\mathbf{S}_1] = \mathbf{0} \in \mathbb{R}^{d_1}$.

(c) The PS model at time 1 is logistic and the OR model at time 2 is linear, i.e., $\pi(\mathbf{s}_1) = g(\mathbf{s}_1^\top \boldsymbol{\gamma}^0)$ and $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$ with some $\boldsymbol{\gamma}^0 \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$. Then

$$\begin{aligned}\mathbb{E}\left\{\frac{A_1A_2Y}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} &= \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} = \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\bar{\mathbf{S}}_2^\top}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\}\boldsymbol{\alpha}^0 \\ &\overset{(i)}{=} \mathbb{E}\left\{\frac{A_1\bar{\mathbf{S}}_2^\top}{\pi^*(\mathbf{S}_1)}\right\}\boldsymbol{\alpha}^0 \overset{(ii)}{=} \mathbb{E}\left\{\frac{A_1\nu(\bar{\mathbf{S}}_2)}{\pi(\mathbf{S}_1)}\right\} \overset{(iii)}{=} \theta_{1,1},\end{aligned}$$

where (i) holds since $\mathbb{E}\left[A_1/\pi^*(\mathbf{S}_1)\{1 - A_2/\rho^*(\bar{\mathbf{S}}_2)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d$ and by the tower rule, $\mathbb{E}\left[A_1/\pi^*(\mathbf{S}_1)\{1 - \rho(\bar{\mathbf{S}}_2)/\rho^*(\bar{\mathbf{S}}_2)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d$; (ii) holds by $\pi^*(\cdot) = \pi(\cdot)$ since (a.2); (iii) holds following the steps (ii)-(iv) of part (a).

(d) Both OR models are linear, i.e., $\mu(\mathbf{s}_1) = \mathbf{s}_1^\top \boldsymbol{\beta}^0$ and $\nu(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2^\top \boldsymbol{\alpha}^0$ with some $\boldsymbol{\beta}^0 \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha}^0 \in \mathbb{R}^d$. Then

$$\begin{aligned}\mathbb{E}\left\{\frac{A_1A_2Y}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} &= \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\} = \mathbb{E}\left\{\frac{A_1\rho(\bar{\mathbf{S}}_2)\bar{\mathbf{S}}_2^\top}{\pi^*(\mathbf{S}_1)\rho^*(\bar{\mathbf{S}}_2)}\right\}\boldsymbol{\alpha}^0 \\ &\overset{(i)}{=} \mathbb{E}\left\{\frac{A_1\bar{\mathbf{S}}_2^\top}{\pi^*(\mathbf{S}_1)}\right\}\boldsymbol{\alpha}^0 = \mathbb{E}\left\{\frac{A_1\nu(\bar{\mathbf{S}}_2)}{\pi^*(\mathbf{S}_1)}\right\} \overset{(ii)}{=} \theta_{1,1},\end{aligned}$$

where (i) holds following the step (i) of part (c); (ii) holds following the steps (ii)-(iv) of part(b). ∎

Proofs of the results in Section 2.3

Proof of Theorem 2.1. Recall the definition of the score function, (2.4). Observe that

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}) =& - A_1 \exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}) \left\{ \frac{A_2(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha})}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})} + \bar{\mathbf{S}}_2^\top\boldsymbol{\alpha} - \mathbf{S}_1^\top\boldsymbol{\beta} \right\} \mathbf{S}_1,\\
\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}) =& - \frac{A_1 A_2 \exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha})}{g(\mathbf{S}_1^\top\boldsymbol{\gamma})}\bar{\mathbf{S}}_2,\\
\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}) =& \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma})} \left\{ 1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta})} \right\} \bar{\mathbf{S}}_2,\\
\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}) =& \left\{ 1 - \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma})} \right\} \mathbf{S}_1.
\end{aligned}$$

By the constructions in (2.13)-(2.15), we have

$$\begin{aligned}
\mathbb{E}\left\{\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -\mathbb{E}\left[A_1 \exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}^*) \left\{ \frac{A_2(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)} + \bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}^* - \mathbf{S}_1^\top\boldsymbol{\beta}^* \right\} \mathbf{S}_1 \right]\\
&= \frac{1}{2}\boldsymbol{\nabla}_{\boldsymbol{\beta}}\mathbb{E}\{\ell_4(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\} = \mathbf{0} \in \mathbb{R}^{d_1}, \qquad (2.43)\\
\mathbb{E}\left\{\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= -\mathbb{E}\left[\frac{A_1 A_2 \exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)(Y-\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}^*)}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\bar{\mathbf{S}}_2 \right]\\
&= \frac{1}{2}\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\mathbb{E}\{\ell_3(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\} = \mathbf{0} \in \mathbb{R}^{d}, \qquad (2.44)\\
\mathbb{E}\left\{\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= \mathbb{E}\left[\frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)} \left\{ 1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)} \right\} \bar{\mathbf{S}}_2 \right] = \boldsymbol{\nabla}_{\boldsymbol{\delta}}\mathbb{E}\{\ell_2(\mathbf{W};\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\} = \mathbf{0} \in \mathbb{R}^{d}, \qquad (2.45)\\
\mathbb{E}\left\{\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\boldsymbol{\eta}^*)\right\} &= \mathbb{E}\left[\left\{ 1 - \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)} \right\} \mathbf{S}_1 \right] = \boldsymbol{\nabla}_{\boldsymbol{\gamma}}\mathbb{E}\{\ell_1(\mathbf{W};\boldsymbol{\gamma}^*)\} = \mathbf{0} \in \mathbb{R}^{d_1}, \qquad (2.46)
\end{aligned}$$

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

$$\Delta_{k,1} = n^{-1} \sum_{i \in \mathcal{I}_k} \{\psi(\mathbf{W}_i; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}_i; \boldsymbol{\eta}^*)\} - \mathbb{E}\{\psi(\mathbf{W}; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\},$$

$$\Delta_{k,2} = \mathbb{E}\{\psi(\mathbf{W}; \widehat{\boldsymbol{\eta}}_{-k}) - \psi(\mathbf{W}; \boldsymbol{\eta}^*)\}.$$

Step 1. We demonstrate that

$$\mathbb{E}\{\psi(\mathbf{W}; \boldsymbol{\eta}^*)\} - \theta_{1,1} = 0. \tag{2.47}$$

Here, (2.47) can be shown under the Assumption 2.2:

$$\begin{aligned}
&\mathbb{E}\{\psi(\mathbf{W}; \boldsymbol{\eta}^*)\} - \theta_{1,1} \\
&= \mathbb{E}\left[\left\{1 - \frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - Y_{1,1}\}\right] + \mathbb{E}\left[\frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - Y_{1,1}\}\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\left\{1 - \frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \mu(\mathbf{S}_1)\}\right] + \mathbb{E}\left[\frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\left\{1 - \frac{\rho(\bar{\mathbf{S}}_2)}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - \nu(\bar{\mathbf{S}}_2)\}\right] \\
&\overset{(ii)}{=} 0,
\end{aligned}$$

where (i) holds by the tower rule, (ii) holds under the Assumption 2.2.

Step 2. We demonstrate that, for each $k \leq \mathbb{K}$ and any $\theta \in \mathbb{R}$, as $N, d_1, d_2 \to \infty$,

$$\Delta_{k,2} = o_p(N^{-1/2}). \tag{2.48}$$

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

$$\Delta_{k,3} = \mathbb{E}\left[\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\left\{1 - \frac{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})}\right\}\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right],$$

$$\Delta_{k,4} = \mathbb{E}\left[\left\{1 - \frac{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\right\}\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right],$$

$$\Delta_{k,5} = \mathbb{E}\left[\frac{1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\left\{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) - A_1\right\}\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right],$$

$$\Delta_{k,6} = \mathbb{E}\left[\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})}\left\{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - A_2\right\}\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right],$$

$$\Delta_{k,7} = \mathbb{E}\left[\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\left\{\frac{A_2}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})} - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\}(Y_{1,1} - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)\right],$$

$$\Delta_{k,8} = \mathbb{E}\left[\left\{\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})} - \frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}(Y_{1,1} - \mathbf{S}_1^\top \boldsymbol{\beta}^*)\right],$$

$$\Delta_{k,9} = \mathbb{E}\left[\left\{\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})} - \frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\left\{\frac{A_2}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)} - 1\right\}(Y_{1,1} - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)\right].$$

By the tower rule, we have

$$\begin{aligned}
\Delta_{k,5} &= \mathbb{E}\left(\mathbb{E}\left[\frac{\pi^*(\mathbf{S}_1) - A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*) \mid \mathbf{S}_1\right]\right) \\
&= \mathbb{E}\left[\frac{\pi^*(\mathbf{S}_1) - \pi(\mathbf{S}_1)}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right] = 0, \quad \text{when } \pi(\cdot) = \pi^*(\cdot); \\
\Delta_{k,6} &= \mathbb{E}\left(\mathbb{E}\left[\frac{\rho^*(\bar{\mathbf{S}}_2) - A_2}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})}\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*) \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right) \\
&= \mathbb{E}\left[\frac{A_1\{\rho^*(\bar{\mathbf{S}}_2) - \rho(\bar{\mathbf{S}}_2)\}}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})}\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right] = 0, \quad \text{when } \rho(\cdot) = \rho^*(\cdot); \\
\Delta_{k,7} &= \mathbb{E}\left(\mathbb{E}\left[\frac{A_2\{Y_{1,1} - \nu^*(\bar{\mathbf{S}}_2)\}}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\left\{\frac{1}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})} - \frac{1}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\} \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right) \\
&\overset{(i)}{=} \mathbb{E}\left[\frac{A_1\rho(\bar{\mathbf{S}}_2)\{\nu(\bar{\mathbf{S}}_2) - \nu^*(\bar{\mathbf{S}}_2)\}}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})}\left\{\frac{1}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})} - \frac{1}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\}\right] = 0, \quad \text{when } \nu(\cdot) = \nu^*(\cdot); \\
\Delta_{k,8} &= \mathbb{E}\left(\mathbb{E}\left[\left\{\frac{A_1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})} - \frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\{Y_{1,1} - \mu^*(\mathbf{S}_1)\} \mid \mathbf{S}_1\right]\right) \\
&\overset{(ii)}{=} \mathbb{E}\left(\mathbb{E}\left[\left\{\frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})} - \frac{\pi(\mathbf{S}_1)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\{\mu(\mathbf{S}_1) - \mu^*(\mathbf{S}_1)\} \mid \mathbf{S}_1\right]\right) = 0, \quad \text{when } \mu(\cdot) = \mu^*(\cdot).
\end{aligned}$$

Here, (i) holds since $Y_{1,1} \perp\!\!\!\perp A_2 \mid (\bar{\mathbf{S}}_2, A_1 = 1)$ under Assumption 2.1; (ii) holds since $Y_{1,1} \perp\!\!\!\perp A_1 \mid \mathbf{S}_1$ under Assumption 2.1. Note that, the expectation $\mathbb{E}(\cdot)$ corresponds to the joint distribution of the underlying random vector $\mathbb{W} := (\{Y(a_1, a_2)\}_{a_1,a_2\in\{0,1\}}, A_1, A_2, \mathbf{S}_1, \mathbf{S}_2)$, which is independent of the observed samples $(\mathbf{W}_i)_{i=1}^N$ and hence also independent of the nuisance estimators $\widehat{\boldsymbol{\eta}}_{-k}$. Such an expectation will be used throughout the document unless otherwise stated.

Additionally, we also have

$$\begin{aligned}\Delta_{k,9} &= \mathbb{E}\left(\mathbb{E}\left[\left\{\frac{Y_{1,1}-\nu^*(\bar{\mathbf{S}}_2)}{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})} - \frac{Y_{1,1}-\nu^*(\bar{\mathbf{S}}_2)}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\right\}\frac{A_2-\rho^*(\bar{\mathbf{S}}_2)}{\rho^*(\bar{\mathbf{S}}_2)} \mid \bar{\mathbf{S}}_2, A_1 = 1\right]\mathbb{E}(A_1 \mid \bar{\mathbf{S}}_2)\right)\\ &\overset{(i)}{=} \mathbb{E}\left[\left\{\frac{A_1}{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})} - \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\right\}\frac{\rho(\bar{\mathbf{S}}_2)-\rho^*(\bar{\mathbf{S}}_2)}{\rho^*(\bar{\mathbf{S}}_2)}\left\{\nu(\bar{\mathbf{S}}_2)-\nu^*(\bar{\mathbf{S}}_2)\right\}\right] \overset{(ii)}{=} 0,\end{aligned}$$

where (i) holds since $Y_{1,1} \perp\!\!\!\perp A_2 \mid (\bar{\mathbf{S}}_2, A_1 = 1)$ under Assumption 2.1; (ii) holds since either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$ under Assumption 2.2. Therefore,

$$\Delta_{k,2} = \Delta_{k,3} + \Delta_{k,4} + \Delta_{k,5}\mathbb{1}_{\pi\neq\pi^*} + \Delta_{k,6}\mathbb{1}_{\rho\neq\rho^*} + \Delta_{k,7}\mathbb{1}_{\nu\neq\nu^*} + \Delta_{k,8}\mathbb{1}_{\mu\neq\mu^*},$$

Now, condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ are defined as (2.39) and (2.42), respectively. By Lemmas 2.13 and 2.14, $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with probability $1 - o(1)$. By Hölder's inequality,

$$\begin{aligned}|\Delta_{k,3}| &\leq \left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})\right\|_{\mathbb{P},4}\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k})\right\|_{\mathbb{P},4}\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},4}\\ &\quad\cdot\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},4}\\ &\overset{(i)}{=} O_p\left(\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2\right),\end{aligned}$$

where (i) follows from Lemmas 2.13, 2.14 and 2.15. Similarly,

$$
\begin{aligned}
|\Delta_{k,4}| &\leq \left\|g^{-1}(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})\right\|_{\mathbb{P},4} \left\|g^{-1}(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},4} \left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},2} \\
&\overset{(i)}{=} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right),
\end{aligned}
$$

where (i) follows from Lemmas 2.13 and 2.16. In addition,

$$
\begin{aligned}
|\Delta_{k,5}| &\overset{(i)}{=} \left|\mathbb{E}\left[\left\{\frac{1}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k})} - \frac{1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\left\{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) - A\right\}\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right]\right| \\
&\overset{(ii)}{=} \left\|g^{-1}(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k}) - g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},2} \left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},2} \\
&\overset{(iii)}{=} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right),
\end{aligned}
$$

where (i) follows from (2.46); (ii) holds by Hölder's inequality and the fact that $|g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) - A| \leq 1$; (iii) follows from Lemmas 2.13 and 2.16. Besides,

$$
\begin{aligned}
|\Delta_{k,6}| &\overset{(i)}{=} \left|\mathbb{E}\left(\left[\frac{A_1\left\{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - A_2\right\}}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k}) g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k})} - \frac{A_1\left\{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - A_2\right\}}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right]\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right)\right| \\
&\overset{(ii)}{\leq} \left\|g^{-1}(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k}) g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\right\|_{\mathbb{P},2} \left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},2} \\
&\overset{(iii)}{=} O_p\left(\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\right)\|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2\right),
\end{aligned}
$$

where (i) follows from (2.45); (ii) holds by Hölder's inequality and the fact that $|A_1\{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) - A_2\}| \leq 1$; (iii) follows from Lemma 2.15 and the fact that

$$
\begin{aligned}
&\left\|g^{-1}(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}_{-k}) g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}_{-k}) - g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\right\|_{\mathbb{P},2} \\
&\quad = O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2\right). \qquad (2.49)
\end{aligned}
$$

We verify (2.49) below:

$$
\begin{aligned}
&\left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k})-g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},2}\\
&\overset{(i)}{\leq}\left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})\left\{g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k})-g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\}\right\|_{\mathbb{P},2}\\
&\qquad+\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\left\{g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})-g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\}\right\|_{\mathbb{P},2}\\
&\overset{(ii)}{\leq}\left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})\right\|_{\mathbb{P},4}\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k})-g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},4}\\
&\qquad+\left\|g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\right\|_{\mathbb{P},4}\left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})-g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},4}\\
&=O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*\|_2\right),
\end{aligned}
$$

where (i) holds by Minkowski inequality; (ii) holds by (generalized) Hölder's inequality; (iii) follows from Lemmas 2.13 and 2.14. As for the term $\Delta_{k,7}$, 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|\mathbb{E}\left[\frac{A_2}{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}_{-k})}\left\{\frac{A_1}{g(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}}_{-k})}-\frac{A_1}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)}\right\}\varepsilon\right]\right|\\
&\overset{(ii)}{\leq}\left|\mathbb{E}\left\{\frac{A_1A_2}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^\top\right\}(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*)\right|\\
&\quad+\left|\mathbb{E}\left\{A_1A_2\exp(-\mathbf{S}_1^\top\widetilde{\boldsymbol{\gamma}}_1)\exp(-\bar{\mathbf{S}}_2^\top\widetilde{\boldsymbol{\delta}})\varepsilon\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*)\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*)\right\}\right|\\
&\quad+\left|\mathbb{E}\left[\frac{A_1A_2}{g(\mathbf{S}_1^\top\widetilde{\boldsymbol{\gamma}}_1)}\exp(-\bar{\mathbf{S}}_2^\top\widetilde{\boldsymbol{\delta}})\varepsilon\left\{\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*)\right\}^2\right]\right|\\
&\overset{(iii)}{\leq}\left\|\exp(-\mathbf{S}_1^\top\widetilde{\boldsymbol{\gamma}}_1)\right\|_{\mathbb{P},4}\left\|\exp(-\bar{\mathbf{S}}_2^\top\widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},4}\|\varepsilon\|_{\mathbb{P},4}\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*)\right\|_{\mathbb{P},8}\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},8}\\
&\quad+\left\|g^{-1}(\mathbf{S}_1^\top\widetilde{\boldsymbol{\gamma}}_1)\right\|_{\mathbb{P},4}\left\|\exp(-\bar{\mathbf{S}}_2^\top\widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},4}\|\varepsilon\|_{\mathbb{P},4}\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^*)\right\|_{\mathbb{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 (2.44) and Hölder's inequality; (iv) holds by Lemmas 2.13 and 2.14. Similarly, by

Taylor's theorem, with some $\widetilde{\boldsymbol{\gamma}}_2$ lies between $\boldsymbol{\gamma}^*$ and $\widehat{\boldsymbol{\gamma}}_{-k}$, we have

$$
\begin{aligned}
|\Delta_{k,8}| \leq & \left|\mathbb{E}\left\{A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)(Y_{1,1} - \mathbf{S}_1^\top \boldsymbol{\beta}^*)\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}\right| \\
& + \left|\mathbb{E}\left\{A_1 \exp(-\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}}_2)(Y_{1,1} - \mathbf{S}_1^\top \boldsymbol{\beta}^*)\left\{\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}^2\right\}\right| \\
\overset{(i)}{=} & \left|\mathbb{E}\left[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\frac{A_2(Y - \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)} + \bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - \mathbf{S}_1^\top \boldsymbol{\beta}^*\right\}\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right]\right| \\
& + \left|\mathbb{E}\left[A_1 \exp(-\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}}_2)(\varepsilon + \zeta)\left\{\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\}^2\right]\right| \\
\overset{(ii)}{\leq} & 0 + \left\|\exp(-\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}}_2)\right\|_{\mathbb{P},4} \|\varepsilon + \zeta\|_{\mathbb{P},4} \left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},4}^2 \\
\overset{(iii)}{=} & O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2^2\right),
\end{aligned}
$$

where (i) holds since either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$; (ii) holds by (2.43) and Hölder's inequality; (iii) holds by Lemma 2.13. To sum up, we have

$$
\begin{aligned}
\Delta_{k,2} = & O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 \|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2\right) \\
& + \mathbb{1}_{\rho \neq \rho^*} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2\right) \\
& + \mathbb{1}_{\nu \neq \nu^*} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2^2\right) \\
& + \mathbb{1}_{\mu \neq \mu^*} O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2^2\right). \qquad (2.50)
\end{aligned}
$$

Define

$$
r_{\boldsymbol{\gamma}} := \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}, \quad r_{\boldsymbol{\delta}} := \sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}, \quad r_{\boldsymbol{\alpha}} := \sqrt{\frac{s_{\gamma} \log d}{N}}, \quad r_{\boldsymbol{\beta}} := \sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}}.
$$

By Theorems 2.3 and 2.4,

$$
\begin{aligned}
\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 &= O_p\left(r_{\boldsymbol{\gamma}}\right), \\
\|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 &= \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 \mathbb{1}_{\rho = \rho^*} + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 \mathbb{1}_{\rho \neq \rho^*} = O_p\left(r_{\boldsymbol{\delta}} + r_{\boldsymbol{\gamma}} \mathbb{1}_{\rho \neq \rho^*}\right), \\
\|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2 &= \|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2 \mathbb{1}_{\nu = \nu^*} + \|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2 \mathbb{1}_{\nu \neq \nu^*} = O_p\left(r_{\boldsymbol{\alpha}} + (r_{\boldsymbol{\gamma}} + r_{\boldsymbol{\delta}}) \mathbb{1}_{\nu \neq \nu^*}\right).
\end{aligned}
$$

Additionally, note that either $\rho(\cdot) = \rho^*(\cdot)$ or $\nu(\cdot) = \nu^*(\cdot)$ (or both) holds. By Theorems 2.3 and 2.4, 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,

$$\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_{\boldsymbol{\alpha}} + (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_{\boldsymbol{\alpha}} + (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_{\boldsymbol{\alpha}} + r_{\boldsymbol{\gamma}} r_{\boldsymbol{\alpha}}\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_{\boldsymbol{\alpha}}) + \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_{\boldsymbol{\alpha}}) + \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 $\nu(\cdot)=\nu^*(\cdot)$ holds. Note that,

$$\begin{aligned}
r_{\boldsymbol{\gamma}}r_{\boldsymbol{\beta}}+r_{\boldsymbol{\delta}}r_{\boldsymbol{\alpha}} &= \frac{\sqrt{s_{\boldsymbol{\gamma}}s_{\boldsymbol{\beta}}}\log d_1}{N}+\frac{\sqrt{s_{\boldsymbol{\delta}}s_{\boldsymbol{\alpha}}}\log d}{N}\\
&= o(N^{-1/2}), \qquad \text{when (2.23) is assumed;}\\
\mathbb{1}_{\rho\neq\rho^*}r_{\boldsymbol{\gamma}}r_{\boldsymbol{\alpha}} &= \mathbb{1}_{\rho\neq\rho^*}\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}s_{\boldsymbol{\beta}}}\log d_1}{N}+\frac{\sqrt{s_{\boldsymbol{\gamma}}s_{\boldsymbol{\alpha}}\log d_1\log d}}{N}\right)\\
&= o(N^{-1/2}), \qquad \text{when (2.24) is assumed;}\\
\mathbb{1}_{\nu\neq\nu^*}(r_{\boldsymbol{\gamma}}r_{\boldsymbol{\delta}}+r_{\boldsymbol{\delta}}^2) &= \mathbb{1}_{\nu\neq\nu^*}\frac{\sqrt{s_{\boldsymbol{\delta}}\log d(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d)}}{N}\\
&= o(N^{-1/2}), \qquad \text{when (2.25) is assumed;}\\
\mathbb{1}_{\mu\neq\mu^*}r_{\boldsymbol{\gamma}}(r_{\boldsymbol{\gamma}}+r_{\boldsymbol{\delta}}+r_{\boldsymbol{\alpha}}) &= \mathbb{1}_{\mu\neq\mu^*}\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1\{s_{\boldsymbol{\gamma}}\log d_1+(s_{\boldsymbol{\delta}}+s_{\boldsymbol{\alpha}})\log d\}}}{N}\\
&= o(N^{-1/2}), \qquad \text{when (2.26) is assumed.}
\end{aligned}$$

Hence, we conclude that

$$\Delta_{k,2}=o_p(N^{-1/2}).$$

Step 3. We demonstrate that, for each $k\leq\mathbb{K}$ and any $\theta\in\mathbb{R}$, as $N,d_1,d_2\to\infty$,

$$\Delta_{k,1}=o_p(N^{-1/2}).$$

By construction, we have $\mathbb{E}_{\mathbb{S}_k}(\Delta_{k,1})=0$, where $\mathbb{E}_{\mathbb{S}_k}(\cdot)$ denotes the expectation corresponding to the joint distribution of the sub-sample $\mathbb{S}_k$. In addition, by Taylor's theorem, with some $\widetilde{\boldsymbol{\eta}}=(\widetilde{\boldsymbol{\gamma}}^\top,\widetilde{\boldsymbol{\delta}}^\top,\widetilde{\boldsymbol{\alpha}}^\top,\widetilde{\boldsymbol{\beta}}^\top)^\top$ lies between $\boldsymbol{\eta}^*$ and $\widehat{\boldsymbol{\eta}}_{-k}$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_k}(\Delta_{k,1}^2)=n^{-1}\mathbb{E}\left[\{\psi(\mathbf{W};\widehat{\boldsymbol{\eta}}_{-k})-\psi(\mathbf{W};\boldsymbol{\eta}^*)\}^2\right]\\
&=2n^{-1}\mathbb{E}\left[\{\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})-\psi(\mathbf{W};\boldsymbol{\eta}^*)\}\,\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\eta}}_{-k}-\boldsymbol{\eta}^*)\right]\\
&\overset{(i)}{\leq} 2n^{-1}\left\{\left\|\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})-\mathbf{S}_1^\top\boldsymbol{\beta}^*\right\|_{\mathbb{P},2}+\left\|\psi(\mathbf{W};\boldsymbol{\eta}^*)-\mathbf{S}_1^\top\boldsymbol{\beta}^*\right\|_{\mathbb{P},2}\right\}\left\|\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\eta}}_{-k}-\boldsymbol{\eta}^*)\right\|_{\mathbb{P},2},
\end{aligned}$$

where (i) holds by Hölder's inequality and Minkowski inequality. Note that, with probability 1,

$$\left\|\psi(\mathbf{W};\boldsymbol{\eta}^*) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\right\|_{\mathbb{P},2} \le c^{-1}\left\|\zeta\right\|_{\mathbb{P},2} + c^{-2}\left\|\varepsilon\right\|_{\mathbb{P},2} = O(1).$$

Define

$$\widetilde{\varepsilon} := Y_{1,1} - \bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\alpha}}, \quad \widetilde{\zeta} := \bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\alpha}} - \mathbf{S}_1^\top \widetilde{\boldsymbol{\beta}}.$$

Condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$. By Hölder's inequality and Lemmas 2.15 and 2.16, we also have

$$\begin{aligned}
&\left\|\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}}) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\right\|_{\mathbb{P},2} \\
&\quad\le \left\|g^{-1}(\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},4} \left\|\widetilde{\zeta}\right\|_{\mathbb{P},4} + \left\|g^{-1}(\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},6} \left\|g^{-1}(\bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},6} \left\|\widetilde{\varepsilon}\right\|_{\mathbb{P},6} + \left\|\mathbf{S}_1^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},2} \\
&\quad= O_p\left(1 + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right) = O_p\left(1 + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2\right) = O_p(1).
\end{aligned}$$

In addition, by Minkowski inequality,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\eta}}_{-k} - \boldsymbol{\eta}^*)\right\|_{\mathbb{P},2} \\
&\quad\le \left\|\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},2} + \left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*)\right\|_{\mathbb{P},2} \\
&\qquad + \left\|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},2} + \left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*)\right\|_{\mathbb{P},2}.
\end{aligned}$$

By Hölder's inequality, Minkowski inequality, and Lemma 2.13,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},2} \\
&\quad\le \left\|\exp(-\mathbf{S}_1^\top \widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},6} \left\{\left\|g^{-1}(\bar{\mathbf{S}}_2^\top \widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},6} \left\|\widehat{\varepsilon}\right\|_{\mathbb{P},6} + \left\|\widehat{\zeta}\right\|_{\mathbb{P},3}\right\} \left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},6} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2\right).
\end{aligned}$$

Similarly, for the second term, using Lemmas 2.14 and 2.15,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^{\top}(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^{*})\right\|_{\mathbb{P},2} \\
&\quad\leq \left\|g^{-1}(\mathbf{S}_1^{\top}\widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},6}\left\|\exp(-\bar{\mathbf{S}}_2^{\top}\widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},6}\|\widetilde{\varepsilon}\|_{\mathbb{P},12}\left\|\bar{\mathbf{S}}_2^{\top}(\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^{*})\right\|_{\mathbb{P},12} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^{*}\|_2\right).
\end{aligned}$$

For the third term, using Lemma 2.15,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^{\top}(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^{*})\right\|_{\mathbb{P},2} \\
&\quad\leq \left\|g^{-1}(\mathbf{S}_1^{\top}\widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},6}\left\{1+\left\|g^{-1}(\bar{\mathbf{S}}_2^{\top}\widetilde{\boldsymbol{\delta}})\right\|_{\mathbb{P},6}\right\}\left\|\bar{\mathbf{S}}_2^{\top}(\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^{*})\right\|_{\mathbb{P},6} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^{*}\|_2\right).
\end{aligned}$$

Lastly, using Lemma 2.16,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^{\top}(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^{*})\right\|_{\mathbb{P},2} \\
&\quad\leq \left\{1+\left\|g^{-1}(\mathbf{S}_1^{\top}\widetilde{\boldsymbol{\gamma}})\right\|_{\mathbb{P},4}\right\}\left\|\mathbf{S}_1^{\top}(\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^{*})\right\|_{\mathbb{P},4} = O_p\left(\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^{*}\|_2\right).
\end{aligned}$$

To sum up, we have

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\eta}}\psi(\mathbf{W};\widetilde{\boldsymbol{\eta}})^{\top}(\widehat{\boldsymbol{\eta}}_{-k}-\boldsymbol{\eta}^{*})\right\|_{\mathbb{P},2} \\
&\quad= O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^{*}\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^{*}\|_2+\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^{*}\|_2+\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^{*}\|_2\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_k}(\Delta_{k,1}^2) &= n^{-1}\mathbb{E}\left[\{\psi(\mathbf{W};\widehat{\boldsymbol{\eta}}_{-k})-\psi(\mathbf{W};\boldsymbol{\eta}^{*})\}^2\right] \\
&= N^{-1}O_p\left(\|\widehat{\boldsymbol{\gamma}}_{-k}-\boldsymbol{\gamma}^{*}\|_2+\|\widehat{\boldsymbol{\delta}}_{-k}-\boldsymbol{\delta}^{*}\|_2+\|\widehat{\boldsymbol{\alpha}}_{-k}-\boldsymbol{\alpha}^{*}\|_2+\|\widehat{\boldsymbol{\beta}}_{-k}-\boldsymbol{\beta}^{*}\|_2\right). \quad (2.51)
\end{aligned}$$

By Lemma 2.2,

$$\Delta_{k,1} = O_p\left(N^{-1/2}\sqrt{\|\widehat{\boldsymbol{\gamma}}_{-k} - \boldsymbol{\gamma}^*\|_2 + \|\widehat{\boldsymbol{\delta}}_{-k} - \boldsymbol{\delta}^*\|_2 + \|\widehat{\boldsymbol{\alpha}}_{-k} - \boldsymbol{\alpha}^*\|_2 + \|\widehat{\boldsymbol{\beta}}_{-k} - \boldsymbol{\beta}^*\|_2}\right)$$
$$= o_p(N^{-1/2}).$$

Step 4. We show that, as $N, d_1, d_2 \to \infty$,

$$\sigma^{-1}N^{-1/2}\sum_{i=1}^{N}\psi(\mathbf{W}_i;\boldsymbol{\eta}^*) - \theta_{1,1} \to \mathcal{N}(0,1). \tag{2.52}$$

By Lyapunov's central limit theorem, it suffices to show that, for some $t > 2$,

$$\sigma^{-t}\mathbb{E}\left\{|\psi(\mathbf{W};\boldsymbol{\eta}^*) - \theta_{1,1}|^t\right\} < C, \tag{2.53}$$

with some constant $C > 0$. Note that

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\{Y_{1,1} - \theta_{1,1}\}\right]^2 + \mathbb{E}\left(\left[\left\{1 - \frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^\top\boldsymbol{\beta}^* - Y_{1,1}\}\right]^2\right) \\
&\quad + \mathbb{E}\left(\left[\frac{A_1}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^\top\boldsymbol{\alpha}^* - Y_{1,1}\}\right]^2\right) \\
&\geq \mathbb{E}\left[\{Y_{1,1} - \theta_{1,1}\}^2\right] \overset{(i)}{\geq} \mathbb{E}\left[\{Y_{1,1} - \theta_{1,1}\}^2\right]/2 + \mathbb{E}[\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2]/2 \\
&\overset{(ii)}{\geq} c_Y/2 + c_0(1-c_0)^{-1}\mathbb{E}[A_1\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2]/2,
\end{aligned}$$

where (i) holds since $\mathbb{E}\left[\{Y_{1,1} - \theta_{1,1}\}^2\right] = \mathbb{E}\left[\{Y_{1,1} - \mu(\mathbf{S}_1)\}^2\right] + \mathbb{E}\left[\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2\right]$; (ii) holds since $\exp(\mathbf{S}_1^\top\boldsymbol{\gamma}^*) > c_0(1-c_0)^{-1}$ under Assumption 2.1, $\mathbb{E}[\{Y_{1,1} - \theta_{1,1}\}]^2 \geq c_Y$ under Assumption 2.4, and $A_1 \leq 1$. Based on the construction of $\boldsymbol{\beta}^*$ as in (2.15), and since either $\rho^*(\cdot)$ or $\nu^*(\cdot)$ is correctly specified, we have

$$\begin{aligned}
\boldsymbol{\beta}^* &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} \mathbb{E}\left[A_1\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\left\{\nu(\bar{\mathbf{S}}_2) - \mathbf{S}_1^\top\boldsymbol{\beta}\right\}^2\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_1}} \mathbb{E}\left[A_1\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\left\{\mu(\mathbf{S}_1) - \mathbf{S}_1^\top\boldsymbol{\beta}\right\}^2\right],
\end{aligned}$$

which implies that

$$\mathbb{E}\left[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\mu(\mathbf{S}_1) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}.$$

Under Assumptions 2.1 and 2.4, it follows that

$$\begin{aligned}
&\mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \theta_{1,1}\}^2] \\
&\quad \overset{(i)}{=} \mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\}^2] + \mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \theta_{1,1}\}^2] \\
&\quad \geq \mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \theta_{1,1}\}^2] = \mathbb{E}[\pi(\mathbf{S}_1) \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \theta_{1,1}\}^2] \\
&\quad \geq c_0(c_0^{-1} - 1)\mathbb{E}[\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \theta_{1,1}\}^2] \geq (1 - c_0)c_{\min}\|\boldsymbol{\beta}^*\|_2^2,
\end{aligned}$$

where (i) holds since

$$\begin{aligned}
&\mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\}\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - \theta_{1,1}\}] \\
&\quad = \mathbb{E}[A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\{\mu(\mathbf{S}_1) - \mathbf{S}_1^\top \boldsymbol{\beta}^*\}\mathbf{S}_1^\top\{\boldsymbol{\beta}^* - \theta_{1,1}\mathbf{e}_1\}] = 0.
\end{aligned}$$

Therefore, we have

$$\sigma^2 \geq c_Y/2 + c_0 c_{\min}\|\boldsymbol{\beta}^*\|_2^2/2.$$

Additionally, for any $r > 0$, by Minkowski inequality,

$$\begin{aligned}
&\|\psi(\mathbf{W}; \boldsymbol{\eta}^*) - \theta_{1,1}\|_{\mathbb{P},r} \\
&\quad \leq \|Y_{1,1} - \theta_{1,1}\|_{\mathbb{P},r} + \left\|\left\{1 - \frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\right\}\{\mathbf{S}_1^\top \boldsymbol{\beta}^* - Y_{1,1}\}\right\|_{\mathbb{P},r} \\
&\qquad + \left\|\frac{A_1}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)}\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}\right\}\{\bar{\mathbf{S}}_2^\top \boldsymbol{\alpha}^* - Y_{1,1}\}\right\|_{\mathbb{P},r} \\
&\quad \leq \|\mathbf{S}_1^\top \boldsymbol{\beta}^*\|_{\mathbb{P},r} + \|\varepsilon\|_{\mathbb{P},r} + \|\zeta\|_{\mathbb{P},r} + |\theta_{1,1}| + (1 + c_0^{-1})\|\varepsilon + \zeta\|_{\mathbb{P},r} \\
&\qquad + c_0^{-1}(1 + c_0^{-1})\|\varepsilon\|_{\mathbb{P},r} \\
&\quad \overset{(i)}{=} O\left(\|\boldsymbol{\beta}^*\|_2 + 1\right),
\end{aligned}$$

where (i) holds by $|\theta_{1,1}| = |\mathbb{E}(\mathbf{S}_1^\top \boldsymbol{\beta}^*)| \leq \|\mathbf{S}_1^\top \boldsymbol{\beta}^*\|_{\mathbb{P},1}$. Therefore,

$$\sigma \asymp \|\boldsymbol{\beta}^*\|_2 + 1, \tag{2.54}$$

and

$$\begin{aligned}\sigma^{-t}\mathbb{E}\left\{|\psi(\mathbf{W};\boldsymbol{\eta}^*) - \theta_{1,1}|^t\right\} &= \left\{\frac{\|\psi(\mathbf{W};\boldsymbol{\eta}^*) - \theta_{1,1}\|_{\mathbb{P},t}}{\sigma}\right\}^t \\ &= O\left(\frac{\|\boldsymbol{\beta}^*\|_2 + 1}{c_Y/2 + c_{\min}\|\boldsymbol{\beta}^*\|_2^2/2}\right) = O(1),\end{aligned}$$

and (2.53) follows.

Step 5. Finally, we prove that, as $N, d_1, d_2 \to \infty$,

$$\widehat{\sigma}^2 = \sigma^2\{1 + o_p(1)\}. \tag{2.55}$$

Note that

$$\mathbb{E}_{\mathbb{S}}\left[\left\{N^{-1}\sum_{i=1}^N \psi(\mathbf{W}_i;\boldsymbol{\eta}^*) - \theta_{1,1}\right\}^2\right] = N^{-1}\sigma^2 \asymp \frac{\|\boldsymbol{\beta}^*\|_2^2 + 1}{N},$$

where $\mathbb{E}_{\mathbb{S}}(\cdot)$ denotes the expectation corresponding to the whole observed samples $\mathbb{S} = (\mathbf{W}_i)_{i=1}^N$. By Lemma 2.2,

$$N^{-1}\sum_{i=1}^N \psi(\mathbf{W}_i;\boldsymbol{\eta}^*) - \theta_{1,1} = O_p\left(\frac{\|\boldsymbol{\beta}^*\|_2 + 1}{\sqrt{N}}\right).$$

By (2.51), (2.53), (2.54), and Lemma A.4 of [ZCB21], we have (2.55) holds. ∎

Proof of Theorem 2.2. Theorem 2.2 follows directly from Theorem 2.1 as a special case that all the nuisance models are correctly specified. ∎

Proofs of the results in Section 2.4.1

To begin with, we first demonstrate some useful lemmas before we obtain the asymptotic results for the moment-targeted nuisance estimators. For any $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\delta}, \boldsymbol{\alpha} \in \mathbb{R}^d$,

define

$$\bar{\ell}_1(\boldsymbol{\gamma}) := M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}\ell_1(\mathbf{W}_i;\boldsymbol{\gamma}), \quad \bar{\ell}_2(\boldsymbol{\gamma},\boldsymbol{\delta}) := M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\ell_2(\mathbf{W}_i;\boldsymbol{\gamma},\boldsymbol{\delta}), \tag{2.56}$$

$$\bar{\ell}_3(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}) := M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}\ell_3(\mathbf{W}_i;\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}), \quad \bar{\ell}_4(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}) := M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\ell_4(\mathbf{W}_i;\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}),$$

where the loss functions are defined as (2.13), (2.14), (2.16), and (2.17). For any $\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\Delta} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha},\boldsymbol{\delta} \in \mathbb{R}^d$, define

$$\delta\bar{\ell}_1(\boldsymbol{\gamma},\boldsymbol{\Delta}) := \bar{\ell}_1(\boldsymbol{\gamma}+\boldsymbol{\Delta}) - \bar{\ell}_1(\boldsymbol{\gamma}) - \nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma})^\top\boldsymbol{\Delta}, \tag{2.57}$$

$$\delta\bar{\ell}_4(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Delta}) := \bar{\ell}_4(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}+\boldsymbol{\Delta}) - \bar{\ell}_4(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\beta})^\top\boldsymbol{\Delta}. \tag{2.58}$$

Similarly, for any $\boldsymbol{\gamma} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\Delta} \in \mathbb{R}^d$, define

$$\delta\bar{\ell}_2(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\Delta}) := \bar{\ell}_2(\boldsymbol{\gamma},\boldsymbol{\delta}+\boldsymbol{\Delta}) - \bar{\ell}_2(\boldsymbol{\gamma},\boldsymbol{\delta}) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma},\boldsymbol{\delta})^\top\boldsymbol{\Delta}, \tag{2.59}$$

$$\delta\bar{\ell}_3(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha},\boldsymbol{\Delta}) := \bar{\ell}_3(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \bar{\ell}_3(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}) - \nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha})^\top\boldsymbol{\Delta}. \tag{2.60}$$

We demonstrate the following restricted strong convexity (RSC) conditions. Note that, the nuisance estimators are constructed based on different samples, and the probability measures in (2.61)-(2.64) are different.

Lemma 2.17. *Let Assumptions 2.1 and 2.4 hold. Define* $f_{M,d_1}(\boldsymbol{\Delta}) := \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\|\boldsymbol{\Delta}\|_1^2 \log d_1/M$ *for any* $\boldsymbol{\Delta} \in \mathbb{R}^{d_1}$ *and* $f_{M,d}(\boldsymbol{\Delta}) := \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\|\boldsymbol{\Delta}\|_1^2 \log d/M$ *for any* $\boldsymbol{\Delta} \in \mathbb{R}^d$. *Then, with some constants* $\kappa_1,\kappa_2,c_1,c_2 > 0$ *and note that* $M \asymp N$, *we have*

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\delta\bar{\ell}_1(\boldsymbol{\gamma}^*,\boldsymbol{\Delta}) \geq f_{M,d_1}(\boldsymbol{\Delta}),\ \ \forall\|\boldsymbol{\Delta}\|_2 \leq 1\right) \geq 1 - c_1\exp(-c_2M). \tag{2.61}$$

Further, let $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$. *Then*

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta}) \geq f_{M,d}(\boldsymbol{\Delta}),\ \ \forall\|\boldsymbol{\Delta}\|_2 \leq 1\right) \geq 1 - c_1\exp(-c_2M), \tag{2.62}$$

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*,\boldsymbol{\Delta}) \geq f_{M,d_1}(\boldsymbol{\Delta}),\ \ \forall\boldsymbol{\Delta} \in \mathbb{R}^{d_1}\right) \geq 1 - c_1\exp(-c_2M). \tag{2.63}$$

In (2.62)*, we only consider the randomness in* $\mathbb{S}_{\boldsymbol{\delta}}$*, and* $\widehat{\boldsymbol{\gamma}}$ *is treated as fixed (or conditional on). Similarly, in* (2.64)*,* $\widehat{\boldsymbol{\gamma}}$*,* $\widehat{\boldsymbol{\delta}}$*, and* $\widehat{\boldsymbol{\alpha}}$ *are all treated as fixed.*

Moreover, let $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1$*. Then*

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \geq f_{M,d}(\boldsymbol{\Delta}),\ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right) \geq 1 - c_1 \exp(-c_2 M), \tag{2.64}$$

where $\widehat{\boldsymbol{\gamma}}$ *and* $\widehat{\boldsymbol{\delta}}$ *are treated as fixed.*

Proof of Lemma 2.17. We show that, with high probability, the RSC property holds for each of the loss functions. By Taylor's theorem, with some $v_1, v_2 \in (0,1)$,

$$\delta\bar{\ell}_1(\boldsymbol{\gamma}^*, \boldsymbol{\Delta}) = (2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\gamma}}} A_{1i} \exp\{-\mathbf{S}_{1i}^\top(\boldsymbol{\gamma}^* + v_1\boldsymbol{\Delta})\}(\mathbf{S}_{1i}^\top \boldsymbol{\Delta})^2,$$

$$\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) = (2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i} g^{-1}(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \exp\{-\bar{\mathbf{S}}_{2i}^\top(\boldsymbol{\delta}^* + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2, \tag{2.65}$$

$$\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i} g^{-1}(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \exp(-\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}})(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2, \tag{2.66}$$

$$\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^\top \boldsymbol{\Delta})^2. \tag{2.67}$$

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 2.1, $|\mathbf{U}^\top \boldsymbol{\eta}| \leq |\mathbf{S}_1^\top \boldsymbol{\gamma}^*| < C$ with some constant $C > 0$. By Lemmas 2.3 and 2.6, we have (2.61) holds. Note that, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}$, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}$, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}$, and $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}$ are the probability measures corresponding to disjoint (and independent) sub-samples $\mathbb{S}_{\boldsymbol{\gamma}}$, $\mathbb{S}_{\boldsymbol{\delta}}$, $\mathbb{S}_{\boldsymbol{\alpha}}$, and $\mathbb{S}_{\boldsymbol{\beta}}$, respectively.

Part 2. Now we treat $\widehat{\boldsymbol{\gamma}}$ as fixed (or conditional on) and suppose that $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$. Note that $g^{-1}(u) = 1 + \exp(-u)$ and $\mathbf{S} = (\mathbf{S}_1^\top, \mathbf{S}_2^\top)^\top$. Hence,

$$\begin{aligned}\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) = {} & (2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i} \exp\{-\bar{\mathbf{S}}_{2i}^\top(\boldsymbol{\delta}^* + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2 \\ & + (2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i}A_{2i} \exp\{-\bar{\mathbf{S}}_{2i}^\top(\boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}} + v_2\boldsymbol{\Delta})\}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2,\end{aligned}$$

where $\check{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}^\top, 0, \ldots, 0)^\top \in \mathbb{R}^d$. Let $\mathbf{U} = A_1 A_2 \mathbf{S}$, $\mathbb{S}' = (A_{1i} A_{2i} \bar{\mathbf{S}}_{2i})_{i \in \mathcal{I}_{\boldsymbol{\delta}}}$, $\phi(u) = \exp(-u)$, $v = v_2$, and $\boldsymbol{\eta} = \boldsymbol{\delta}^*$. Note that, under Assumption 2.1, we have $|\mathbf{U}^\top \boldsymbol{\eta}| \leq |\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*| < C$ with some constant $C > 0$. By Lemmas 2.3 and 2.6, we have

$$(2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i} A_{2i} \exp\{-\bar{\mathbf{S}}_{2i}^\top (\boldsymbol{\delta}^* + v_2 \boldsymbol{\Delta})\} (\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2$$
$$\geq \kappa_1' \|\boldsymbol{\Delta}\|_2^2 - \kappa_2' \frac{\log d}{M} \|\boldsymbol{\Delta}\|_1^2, \quad \forall \|\boldsymbol{\Delta}\|_2 \leq 1, \tag{2.68}$$

with probability $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}$ at least $1 - c_1' \exp(-c_2' M)$ and some constants $\kappa_1', \kappa_2', c_1', c_2' > 0$.

Similarly, let $\mathbf{U} = A_1 A_2 \mathbf{S}$, $\mathbb{S}' = (A_{1i} A_{2i} \bar{\mathbf{S}}_{2i})_{i \in \mathcal{I}_{\boldsymbol{\delta}}}$, $\phi(u) = \exp(-u)$, $v = v_2$, and $\boldsymbol{\eta} = \boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}}$. On the event $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$, under Assumptions 2.1 and 2.4, we have $\mathbb{E}\{|\mathbf{U}^\top \boldsymbol{\eta}|\} \leq \mathbb{E}(|\mathbf{S}_1^\top \boldsymbol{\gamma}^*|) + \mathbb{E}(|\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*|) + \mathbb{E}\{|\mathbf{S}_1^\top (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)|\} < C$ with some constant $C > 0$. By Lemmas 2.3 and 2.6, we have

$$(2M)^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i} A_{2i} \exp\{-\bar{\mathbf{S}}_{2i}^\top (\boldsymbol{\delta}^* + \check{\boldsymbol{\gamma}} + v_2 \boldsymbol{\Delta})\} (\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2$$
$$\geq \kappa_1' \|\boldsymbol{\Delta}\|_2^2 - \kappa_2' \frac{\log d}{M} \|\boldsymbol{\Delta}\|_1^2, \quad \forall \|\boldsymbol{\Delta}\|_2 \leq 1, \tag{2.69}$$

with probability $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}$ at least $1 - c_1' \exp(-c_2' M)$. Hence, (2.62) follows from (2.68) and (2.69).

Part 3. We treat both $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\delta}}$ as fixed (or conditional on) and suppose that $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$, $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1$. Note that

$$\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} A_{1i} A_{2i} \exp(-\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}) (\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2$$
$$+ M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} A_{1i} A_{2i} \exp\{-\bar{\mathbf{S}}_{2i}^\top (\widehat{\boldsymbol{\delta}} + \check{\boldsymbol{\gamma}})\} (\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta})^2. \tag{2.70}$$

Let $\mathbf{U} = A_1 A_2 \mathbf{S}$, $\mathbb{S}' = (A_{1i} A_{2i} \bar{\mathbf{S}}_{2i})_{i \in \mathcal{I}_{\boldsymbol{\alpha}}}$, $\phi(u) = \exp(-u)$, $v = 0$, and $\boldsymbol{\eta} = \widehat{\boldsymbol{\delta}}$. Here, $\mathbb{E}(|\mathbf{U}^\top \boldsymbol{\eta}|) \leq \mathbb{E}(|\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*|) + \mathbb{E}\{|\bar{\mathbf{S}}_2^\top (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)|\} < C$ with some constant $C > 0$. By Lemmas 2.3 and 2.6, we

have

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp(-\bar{\mathbf{S}}_{2i}^{\top}\widehat{\boldsymbol{\delta}})(\bar{\mathbf{S}}_{2i}^{\top}\boldsymbol{\Delta})^2 \geq \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2\leq 1, \tag{2.71}$$

with probability $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}$ 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 $\mathbb{E}(|\mathbf{U}^{\top}\boldsymbol{\eta}|)\leq\mathbb{E}(|\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*|)+\mathbb{E}(|\bar{\mathbf{S}}_2^{\top}\boldsymbol{\delta}^*|)+\mathbb{E}\{|\mathbf{S}_1^{\top}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)|\}+\mathbb{E}\{|\bar{\mathbf{S}}_2^{\top}(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}^*)|\}<C$ with some constant $C>0$. By Lemmas 2.3 and 2.6, we have

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\exp\{-\bar{\mathbf{S}}_{2i}^{\top}(\widehat{\boldsymbol{\delta}}+\check{\boldsymbol{\gamma}})\}(\bar{\mathbf{S}}_{2i}^{\top}\boldsymbol{\Delta})^2 \geq \kappa_1'\|\boldsymbol{\Delta}\|_2^2 - \kappa_2'\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \quad \forall\|\boldsymbol{\Delta}\|_2\leq 1, \tag{2.72}$$

with probability $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}$ 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 (2.71) and (2.72) 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 (2.71) and (2.72) hold if we multiply the LHS and RHS by a factor $\|\boldsymbol{\Delta}'\|_2^2$. Therefore, (2.64) holds by combining the lower bounds with (2.70) .

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, $\mathbb{E}\{|\mathbf{U}^{\top}\boldsymbol{\eta}|\}\leq\mathbb{E}(|\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*|)+\mathbb{E}\{|\mathbf{S}_1^{\top}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)|\}<C$ with some constant $C>0$. Then (2.63) holds by Lemmas 2.3 and 2.6. 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. ∎

Additionally, we upper bound the gradients of the loss functions evaluated at the target population parameter values. Note that the nuisance parameters $\boldsymbol{\gamma}^*$, $\boldsymbol{\delta}^*$, $\boldsymbol{\alpha}^*$, and $\boldsymbol{\beta}^*$ are defined as the minimizers of the corresponding loss functions, (2.13), (2.14), (2.16), and (2.17). By the KKT condition, the gradients of the loss functions' expectations are zero

vectors; see also (2.73), (2.74), (2.75), and (2.76). Therefore, the gradients of the loss functions' empirical averages $\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)$, $\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)$, $\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)$, and $\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ are averages of i.i.d. random vectors with zero means even under model misspecification. Hence, we can use the union bound techniques to control the infinite norms by the usual rates $O_p(\sqrt{\log d/M})$ or $O_p(\sqrt{\log d_1/M})$.

Lemma 2.18. *Let Assumption 2.4 holds. Let $\sigma_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\delta}}, \sigma_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\beta}} > 0$ be some constants and note that $M \asymp N$. Then, for any $t > 0$,*

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left\|\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)\right\|_\infty \le \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t+\log d}{M}}\right) \ge 1 - 2\exp(-t).$$

Further, let the Assumption 2.1 holds. Then, for any $t > 0$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty \le \sigma_{\boldsymbol{\delta}}\sqrt{\frac{t+\log d}{M}}\right) \ge 1 - 2\exp(-t),$$

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\|_\infty \le \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}} + \frac{t+\log d}{M}\right)\right) \ge 1 - 2\exp(-t),$$

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty \le \sigma_{\boldsymbol{\beta}}\left(2\sqrt{\frac{t+\log d_1}{M}} + \frac{t+\log d_1}{M}\right)\right) \ge 1 - 2\exp(-t).$$

Proof of Lemma 2.18. Now, we control the gradients of the loss functions.

Part 1. Note that

$$\nabla_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\gamma}}}\{1 - A_{1i}g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\}\mathbf{S}_{1i}.$$

By the construction of $\boldsymbol{\gamma}^*$, we have

$$\mathbb{E}\left[\{1 - A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}. \tag{2.73}$$

Also, for each $1 \le j \le d_1$, $|\{1 - A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\}\mathbf{S}_1^\top\mathbf{e}_j| \le (1 + c_0^{-1})|\mathbf{S}_1^\top\mathbf{e}_j|$ and hence, by Lemma 2.4,

$$\|\{1 - A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\}\mathbf{S}_1^\top\mathbf{e}_j\|_{\psi_2} \le (1 + c_0^{-1})\|\mathbf{S}_1^\top\mathbf{e}_j\|_{\psi_2} \le (1 + c_0^{-1})\sigma_{\mathbf{S}}.$$

Let $\sigma_{\boldsymbol{\gamma}} := \sqrt{8}(1 + c_0^{-1})\sigma_{\mathbf{S}}$. By Lemma D.2 of [CLCL19], for each $1 \le j \le d_1$ and any $t > 0$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) \le 2\exp(-t - \log d_1).$$

It follows that,

$$\begin{aligned}\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left\|\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)\right\|_\infty > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) &\le \sum_{j=1}^{d_1}\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\gamma}}\bar{\ell}_1(\boldsymbol{\gamma}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\gamma}}\sqrt{\frac{t + \log d_1}{M}}\right) \\ &\le 2d_1\exp(-t - \log d_1) = 2\exp(-t).\end{aligned}$$

Part 2. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) = M^{-1}\sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i}g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\{1 - A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_{2i}.$$

By the construction of $\boldsymbol{\delta}^*$, we have

$$\mathbb{E}\left[A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_2\right] = \mathbf{0} \in \mathbb{R}^d. \tag{2.74}$$

Under Assumption 2.1, we have $|A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_2^\top\mathbf{e}_j| \le c_0^{-1}(1 + c_0^{-1})|\bar{\mathbf{S}}_2^\top\mathbf{e}_j|$ for each $1 \le j \le d$. By Lemma D.1 (i) and (ii),

$$\|A_1 g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_2^\top\mathbf{e}_j\|_{\psi_2} \le (c_0^{-2} + c_0^{-1})\|\bar{\mathbf{S}}_2^\top\mathbf{e}_j\|_{\psi_2} \le (c_0^{-2} + c_0^{-1})\sigma_{\mathbf{S}}.$$

Let $\sigma_{\boldsymbol{\delta}} := \sqrt{8}(c_0^{-2} + c_0^{-1})\sigma_{\mathbf{S}}$. By Lemma D.2 of [CLCL19], for each $1 \le j \le d$ and any $t > 0$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\delta}}\sqrt{\frac{t + \log d}{M}}\right) \le 2\exp(-t - \log d).$$

It follows that,

$$\begin{aligned}\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty > \sigma_{\boldsymbol{\delta}}\sqrt{\frac{t + \log d}{M}}\right) &\le \sum_{j=1}^{d}\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\delta}}\sqrt{\frac{t + \log d}{M}}\right) \\ &\le 2d\exp(-t - \log d) = 2\exp(-t).\end{aligned}$$

Part 3. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*) = -2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\varepsilon_i\bar{\mathbf{S}}_{2i}.$$

By the construction of $\boldsymbol{\alpha}^*$, we have

$$\mathbb{E}\left\{-2A_1A_2g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2\right\} = \mathbf{0}\in\mathbb{R}^d. \tag{2.75}$$

Under Assumption 2.1, we have $|-2A_1A_2g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^\top\mathbf{e}_j| \le 2c_0^{-1}(c_0^{-1}-1)|\varepsilon\bar{\mathbf{S}}_2^\top\mathbf{e}_j|$ for each $1\le j\le d$. By Lemma D.1 (i), (ii), and (v),

$$\begin{aligned}&\|-2A_1A_2g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)\varepsilon\bar{\mathbf{S}}_2^\top\mathbf{e}_j\|_{\psi_1}\\ &\quad\le 2c_0^{-1}(c_0^{-1}-1)\|\varepsilon\|_{\psi_2}\|\bar{\mathbf{S}}_2^\top\mathbf{e}_j\|_{\psi_2} \le 2c_0^{-1}(c_0^{-1}-1)\sigma_\varepsilon\sigma_{\mathbf{S}}.\end{aligned}$$

Let $\sigma_{\boldsymbol{\alpha}} := 2c_0^{-1}(c_0^{-1}-1)\sigma_\varepsilon\sigma_{\mathbf{S}}$. By Lemma 2.4 and Lemma D.4 of [CLCL19], for each $j\le d$ and any $t>0$,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)^\top\mathbf{e}_j\right| > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right) \le 2\exp(-t-\log d).$$

It follows that,

$$\begin{aligned}&\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left\|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)\right\|_\infty > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right)\\ &\quad\le \sum_{j=1}^d \mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}\left(\left|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*)^\top\mathbf{e}_j\right| > \sigma_{\boldsymbol{\alpha}}\left(2\sqrt{\frac{t+\log d}{M}}+\frac{t+\log d}{M}\right)\right)\\ &\quad\le 2d\exp(-t-\log d) = 2\exp(-t).\end{aligned}$$

Part 4. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = -2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\left\{\zeta_i + A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\varepsilon_i\right\}\mathbf{S}_{1i}.$$

By the construction of $\boldsymbol{\beta}^*$, we have

$$\mathbb{E}\left[-2A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\zeta + A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\varepsilon\right\}\mathbf{S}_1\right] = \mathbf{0} \in \mathbb{R}^{d_1}. \tag{2.76}$$

Under Assumption 2.1, we have $|-2A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\zeta + A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\varepsilon\right\}\mathbf{S}_1^\top \mathbf{e}_j| \leq 2(c_0^{-1} - 1)(|\zeta| + c_0^{-1}|\varepsilon|)|\mathbf{S}_1^\top \mathbf{e}_j|$ for each $1 \leq j \leq d$. By Lemma D.1 (i), (ii), and (v),

$$\begin{aligned}
&\|-2A_1 \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left\{\zeta + A_2 g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)\varepsilon\right\}\mathbf{S}_1^\top \mathbf{e}_j\|_{\psi_1} \\
&\quad \leq 2(c_0^{-1} - 1)(\|\zeta\|_{\psi_2} + c_0^{-1}\|\varepsilon\|_{\psi_2})\|\mathbf{S}_1^\top \mathbf{e}_j\|_{\psi_2} \leq 2(c_0^{-1} - 1)(\sigma_\zeta + c_0^{-1}\sigma_\varepsilon)\sigma_{\mathbf{S}}.
\end{aligned}$$

Let $\sigma_{\boldsymbol{\beta}} := 2(c_0^{-1} - 1)(\sigma_\zeta + c_0^{-1}\sigma_\varepsilon)\sigma_{\mathbf{S}}$. By Lemma 2.4 and Lemma D.4 of [CLCL19], for each $1 \leq j \leq d_1$ and any $t > 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\beta}}\left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M}\right)\right) \\
&\quad \leq 2\exp(-t - \log d_1).
\end{aligned}$$

It follows that,

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty > \sigma_{\boldsymbol{\beta}}\left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M}\right)\right) \\
&\quad \leq \sum_{j=1}^{d_1} \mathbb{P}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left|\nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^\top \mathbf{e}_j\right| > \sigma_{\boldsymbol{\beta}}\left(2\sqrt{\frac{t + \log d_1}{M}} + \frac{t + \log d_1}{M}\right)\right) \\
&\quad \leq 2d_1 \exp(-t - \log d_1) = 2\exp(-t).
\end{aligned}$$

■

Proof of Theorem 2.3. We proof the consistency rates of the nuisance parameter estimators when the models are possibly misspecified.

(a) By Lemmas 2.17 and 2.18, 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 2.9, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - t - 2\exp(-t)$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined in (2.35) and (2.36), respectively. By Lemma 2.10, conditional 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 $\bar{s}_{\boldsymbol{\delta}} = \sqrt{s_{\boldsymbol{\gamma}} \log d_1 / \log d + s_{\boldsymbol{\delta}}}$ and $k_0 > 0$ is a constant. Additionally, by Lemma 2.11, we also have $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$. By (a), we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1\}) = 1 - o(1)$. Then, by (2.62) in Lemma 2.17, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_3) \geq 1 - o(1) - c_1 \exp(-c_2 M) = 1 - o(1)$, where $\mathcal{A}_3$ is defined in (2.37). 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 2.10; (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 $\lambda_{\boldsymbol{\delta}} \|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 / 4 \geq 0$; (iv) holds for large enough N, since $\bar{s}_{\boldsymbol{\delta}} \log d / M = s_{\boldsymbol{\gamma}} \log d_1 / M + s_{\boldsymbol{\delta}} \log d / M = o(1)$. Therefore, conditional on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$
\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq \frac{4\lambda_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}}}{\kappa_1} + \frac{2c}{\kappa_1} \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}} = O\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),
$$

with some $\lambda_{\boldsymbol{\delta}} = 2\sigma_{\boldsymbol{\delta}} \sqrt{(t + \log d)/M} \asymp \sqrt{\log d / N}$. 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 $t > 0$, let $\lambda_{\boldsymbol{\alpha}} = 2\sigma_{\boldsymbol{\alpha}}\{2\sqrt{(t+\log d)/M} + (t+\log d)/M\}$. Choose some $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1/N}$ and $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log d/N}$. Define

$$\mathcal{A}_4 := \{\|\nabla_{\boldsymbol{\alpha}} \bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\|_\infty \le \lambda_{\boldsymbol{\alpha}}/2\}, \tag{2.77}$$

$$\mathcal{A}_5 := \left\{\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \ge \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d}{M}\|\boldsymbol{\Delta}\|_1^2, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right\}. \tag{2.78}$$

By Lemma 2.18, we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_4) \ge 1 - 2\exp(-t)$. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$. Similar to the proof of Lemma 2.10 for obtaining (2.102), we have, on the event $\mathcal{A}_4$,

$$2\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}\|_1 \le 4\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_1 + 2|R_2|. \tag{2.79}$$

where

$$\begin{aligned} R_2 &= \left\{\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*) - \nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\}^\top \boldsymbol{\Delta} \\ &= 2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}} A_{1i}A_{2i}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})\exp(-\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*))\right\}\varepsilon_i\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\Delta}. \end{aligned}$$

By the fact that $2ab \le a^2/2 + 2b^2$,

$$|R_2| \le \frac{1}{2}\delta\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) + 2R_3,$$

where

$$R_3 = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}\left(\frac{\exp(-\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)}{g(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)}\right)^2\frac{g(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}})}\varepsilon_i^2.$$

By (a) and (b) of Theorem 2.3, we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1\}) = 1 - o(1)$. Note that

$$\begin{aligned} \mathbb{E}_{\mathbb{S}_{\boldsymbol{\alpha}}}[R_3] &= \mathbb{E}\left[\left(\frac{\exp(-\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\right)^2\frac{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})}\varepsilon^2\right] \\ &\le \left\|\frac{\exp(-\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top\boldsymbol{\delta}^*)}{g(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)}\right\|_{\mathbb{P},6}^2\left\|\frac{g(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})}{\exp(-\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}})}\right\|_{\mathbb{P},3}\|\varepsilon\|_{\mathbb{P},6}^2 \\ &\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d}{N}\right). \end{aligned}$$

where (i) holds by Lemma 2.4, as well as the fact that

$$\begin{aligned}
&\left\| \frac{\exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)} \right\|_{\mathbb{P},6} \\
&\quad \le \left\| g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) \left\{ \exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) - \exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \right\} \right\|_{\mathbb{P},6} \\
&\qquad + \left\| \exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) \left\{ g^{-1}(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) - g^{-1}(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*) \right\} \right\|_{\mathbb{P},6} \\
&\quad = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}} \right)
\end{aligned} \tag{2.80}$$

using Minkowski inequality, (generalized) Hölder's inequality, and Lemmas 2.13 and 2.14. Hence,

$$R_3 = O_p\left(\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N} \right). \tag{2.81}$$

Because of the inequality (2.79), 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 the equation (2.66). We have $\delta \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\Delta}) \ge 0$. Then

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 + \frac{2R_3}{\lambda_{\boldsymbol{\alpha}}} \tag{2.82}$$

Then, by Lemma 2.17, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}} \cup \mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_5) \ge 1 - o(1) - c_1 \exp(-c_2 M) = 1 - o(1)$, where $\mathcal{A}_5$ is defined in (2.78). Now, conditional 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 (2.82) and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$; (iv) holds for large enough N, since $s_{\boldsymbol{\alpha}} \log d/M = o(1)$. Hence, on the event $\mathcal{A}_4 \cap \mathcal{A}_5$, for large enough N,

$$\kappa_1 \|\boldsymbol{\Delta}\|_2^2 - 8\lambda_{\boldsymbol{\alpha}} \sqrt{s_{\boldsymbol{\alpha}}} \|\boldsymbol{\Delta}\|_2 - 16\kappa_2 R_3^2 \frac{\log d}{M\lambda_{\boldsymbol{\alpha}}^2} - 4R_3 \leq 0.$$

Choose some $\lambda_{\boldsymbol{\alpha}} = 2\sigma_{\boldsymbol{\alpha}}\{2\sqrt{(t+\log d)/M} + (t+\log d)/M\} \asymp \sqrt{\log d/N}$. It follows from Lemma 2.12 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 $\lambda_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\alpha}}} \asymp \sqrt{s_{\boldsymbol{\alpha}} \log d/N}$ and (2.81). Recall the inequality (2.82). 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 $t > 0$, let $\lambda_{\boldsymbol{\beta}} = 2\sigma_{\boldsymbol{\beta}}\{2\sqrt{(t+\log d_1)/M} + (t+\log d_1)/M\}$. Choose some $\lambda_{\boldsymbol{\gamma}} \asymp \sqrt{\log d_1/N}$, $\lambda_{\boldsymbol{\delta}} \asymp \sqrt{\log d/N}$, and $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/N}$. Define

$$\mathcal{A}_6 := \left\{ \left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty \leq \lambda_{\boldsymbol{\beta}}/2 \right\}, \tag{2.83}$$

$$\mathcal{A}_7 := \left\{ \delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d_1}{M} \|\boldsymbol{\Delta}\|_1^2, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^{d_1} \right\}. \tag{2.84}$$

By Lemma 2.18, we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\alpha}}}(\mathcal{A}_6) \geq 1 - 2\exp(-t)$. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Similar to the proof of Lemma 2.10 for obtaining (2.102), we have, on the event $\mathcal{A}_6$,

$$2\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}\|_1 \leq 4\lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_{S_{\boldsymbol{\beta}}}\|_1 + 2|R_4|. \tag{2.85}$$

where

$$R_4 = \left\{ \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\}^\top \boldsymbol{\Delta}$$
$$= 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \Bigg\{ \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\alpha}} - \mathbf{S}_{1i}^\top \boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\alpha}})}{g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}})} \right)$$
$$- \exp(-\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*) \left(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha}^* - \mathbf{S}_{1i}^\top \boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*)} \right) \Bigg\} \mathbf{S}_{1i}^\top \boldsymbol{\Delta}.$$

By the fact that $2ab \le a^2/2 + 2b^2$,

$$|R_4| \le \frac{1}{2} \delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + 2R_5,$$

where

$$R_5 = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} \frac{1}{\exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}})} \Bigg\{ \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\alpha}} - \mathbf{S}_{1i}^\top \boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\alpha}})}{g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}})} \right)$$
$$- \exp(-\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*) \left(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha}^* - \mathbf{S}_{1i}^\top \boldsymbol{\beta}^* + \frac{A_{2i}(Y_i - \bar{\mathbf{S}}_{2i}^\top \boldsymbol{\alpha}^*)}{g(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*)} \right) \Bigg\}^2$$

Note that

$$\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}[R_5] = \mathbb{E}\left[\frac{1}{\exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})} (Q_1 + Q_2 + Q_3)^2 \right]$$

where

$$Q_1 = \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) \left(1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})} \right) \bar{\mathbf{S}}_2^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*),$$

$$Q_2 = \left\{ \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*) \right\} \zeta,$$

$$Q_3 = B \left\{ \frac{\exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\gamma}})}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\delta}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^\top \boldsymbol{\delta}^*)} \right\} \varepsilon.$$

By (a) and (b) of Theorem 2.3, we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \le 1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \le 1\}) = 1 - o(1)$.

Then by Hölder's inequality,

$$\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}[R_5] \le \left\| \frac{1}{\exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})} \right\|_{\mathbb{P},2} \left(\|Q_1\|_{\mathbb{P},4} + \|Q_2\|_{\mathbb{P},4} + \|Q_3\|_{\mathbb{P},4} \right)^2$$

and

$$\begin{aligned}
\|Q_1\|_{\mathbb{P},4} &\le \left\|\exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})\right\|_{\mathbb{P},12} \left\|\left(1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})}\right)\right\|_{\mathbb{P},12} \left\|\bar{\mathbf{S}}_2^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},12} \\
&\overset{(i)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}}\right), \\
\|Q_2\|_{\mathbb{P},4} &\le \left\|\left\{\exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\right\}\right\|_{\mathbb{P},8} \|\zeta\|_{\mathbb{P},8} \overset{(ii)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right), \\
\|Q_3\|_{\mathbb{P},4} &\le \left\|\frac{\exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\gamma}})}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\delta}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\gamma}^*)}{g(\mathbf{S}_1^\top \boldsymbol{\delta}^*)}\right\|_{\mathbb{P},8} \|\varepsilon\|_{\mathbb{P},8} \overset{(iii)}{=} O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right),
\end{aligned}$$

where (i) and (ii) hold by Lemmas 2.13, 2.14, 2.15 and Lemma 2.4; (iii) holds analogously as in (2.80). Hence,

$$R_5 = O_p\left(\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d}{N}\right) \tag{2.86}$$

Recall the inequality (2.85). 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 the equation (2.67). We have $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \ge 0$. Then

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\beta}}} \|\boldsymbol{\Delta}\|_2 + \frac{2R_3}{\lambda_{\boldsymbol{\beta}}} \tag{2.87}$$

Then, by Lemma 2.17, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}} \cup \mathbb{S}_{\boldsymbol{\beta}}}(\mathcal{A}_7) \ge 1 - o(1) - c_1 \exp(-c_2 M) = 1 - o(1)$, where $\mathcal{A}_7$ is defined in (2.84). The remaining parts of the proof can be shown analogously as (c) of

Theorem 2.3. Now, conditional on $\mathcal{A}_6 \cap \mathcal{A}_7$, for large enough N,

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2 &\leq \frac{8\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{16R_3^2\frac{\kappa_2 \log d_1}{\kappa_1 M\lambda_{\boldsymbol{\beta}}^2} + \frac{4R_3}{\kappa_1}} \\ &= O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d + s_{\boldsymbol{\beta}}\log d_1}{N}}\right).\end{aligned}$$

with some $\lambda_{\boldsymbol{\beta}} = 2\sigma_{\boldsymbol{\beta}}\{2\sqrt{(t+\log d_1)/M} + (t+\log d_1)/M\} \asymp \sqrt{\log d_1/M}$. Recall the inequality (2.87). We have

$$\|\boldsymbol{\Delta}\|_1 = O_p\left(s_{\boldsymbol{\gamma}}\sqrt{\frac{\log d_1}{N}} + s_{\boldsymbol{\delta}}\sqrt{\frac{(\log d)^2}{N\log d_1}} + s_{\boldsymbol{\alpha}}\sqrt{\frac{(\log d)^2}{N\log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

■

Proofs of the results in Section 2.4.2

Assuming correctly specified models, we control the gradients in Lemma 2.19 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 2.18, we can upper bound the gradients involving the estimated nuisance parameters. For instance, in part (a) of Lemma 2.19 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 2.19. ***(a)*** *Let $\rho(\cdot) = \rho^*(\cdot)$. Let the assumptions in part (a) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,*

$$\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 \log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

(b) *Let $\nu(\cdot) = \nu^*(\cdot)$. Let the assumptions in part (b) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,*

$$\left\|\nabla_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d)\log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

(c) *Let $\nu(\cdot) = \nu^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (c) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,*

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d) \log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

(d) *Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (c) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,*

$$\begin{aligned}&\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\|_\infty \\ &\quad = O_p\left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d + s_{\boldsymbol{\alpha}} \log d) \log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).\end{aligned}$$

(e) *Let $\rho(\cdot) = \rho^*(\cdot)$, $\nu(\cdot) = \nu^*(\cdot)$, and $\mu(\cdot) = \mu^*(\cdot)$. Let the assumptions in part (c) of Theorem 2.3 hold. Then, as $N, d_1, d_2 \to \infty$,*

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\left(1 + \sqrt{\frac{s_{\boldsymbol{\gamma}} (\log d_1)^2}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

Proof of Lemma 2.19. By Lemma 2.18, we have

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right), \tag{2.88}$$

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}}\right), \tag{2.89}$$

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right). \tag{2.90}$$

(a) Let $\rho(\cdot) = \rho^*(\cdot)$. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) = M^{-1}\sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} \mathbf{W}_{\boldsymbol{\delta},i},$$

where

$$\mathbf{W}_{\boldsymbol{\delta},i} := A_{1i}\{g^{-1}(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*)\}\{1 - A_{2i} g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*)\}\bar{\mathbf{S}}_{2i}.$$

Let $\mathbf{W}_{\boldsymbol{\delta}}$ be an indepenent copy of $\mathbf{W}_{\boldsymbol{\delta},i}$. Then, by the tower rule,

$$\mathbb{E}(\mathbf{W}_{\boldsymbol{\delta}}) = \mathbf{0} \in \mathbb{R}^d.$$

By Lemma 2.5, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\delta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\mathbf{W}_{\boldsymbol{\delta},i}\right\|_\infty^2\right) &\le M^{-1}(2e\log d - e)\mathbb{E}(\|\mathbf{W}_{\boldsymbol{\delta}}\|_\infty^2) \\
&\overset{(i)}{\le} (1+c_0^{-1})M^{-1}(2e\log d - e)\mathbb{E}\left\{\left|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right|^2\|\bar{\mathbf{S}}_2\|_\infty^2\right\} \\
&\overset{(ii)}{\le} (1+c_0^{-1})M^{-1}(2e\log d - e)\left\|g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},4}^2\left\|\|\bar{\mathbf{S}}_2\|_\infty\right\|_{\mathbb{P},4}^2 \\
&\overset{(iii)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d)^2}{N^2}\right),
\end{aligned}$$

where (i) holds since $|A_1\{1 - A_2 g^{-1}(\bar{\mathbf{S}}_2\boldsymbol{\delta}^*)\}| \le 1 + c_0^{-1}$ almost surely under Assumption 2.1; (ii) holds by Hölder's inequality; (iii) holds by Lemma 2.13 and Lemma 2.4. By Lemma 2.2,

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*)\right\|_\infty = \left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\mathbf{W}_{\boldsymbol{\delta},i}\right\|_\infty = O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1}\log d}{N}\right).$$

Together with (2.88), we have

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*)\right\|_\infty = O_p\left(\left(1+\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1\log d}{N}}\right)\sqrt{\frac{\log d}{N}}\right).$$

The remaining parts of the proof can be shown analogously as in (a).

(b) Let $\nu(\cdot) = \nu^*(\cdot)$. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*) - \boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_3(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*) = M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\alpha}}}\mathbf{W}_{\boldsymbol{\alpha},i},$$

where

$$\mathbf{W}_{\boldsymbol{\alpha},i} := -2A_{1i}A_{2i}\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})\exp(-\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\exp(-\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\}\varepsilon_i\bar{\mathbf{S}}_{2i}.$$

Let $\mathbf{W}_{\boldsymbol{\alpha}}$ be an indepenent copy of $\mathbf{W}_{\boldsymbol{\alpha},i}$. Then, by the tower rule,

$$\mathbb{E}(\mathbf{W}_{\boldsymbol{\alpha}}) = \mathbf{0} \in \mathbb{R}^d.$$

By Lemma 2.5, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\alpha}}} \left(\left\| M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\alpha}}} \mathbf{W}_{\boldsymbol{\alpha},i} \right\|_\infty^2 \right) &\leq M^{-1}(2e \log d - e) \mathbb{E}(\|\mathbf{W}_{\boldsymbol{\alpha}}\|_\infty^2) \\
&\overset{(i)}{\leq} 2M^{-1}(2e \log d - e) \mathbb{E} \left\{ \left| \frac{\exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)} \right|^2 \varepsilon^2 \|\bar{\mathbf{S}}_2\|_\infty^2 \right\} \\
&\overset{(ii)}{\leq} 2M^{-1}(2e \log d - e) \left\| \frac{\exp(-\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}})}{g(\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}})} - \frac{\exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*)}{g(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)} \right\|_{\mathbb{P},6}^2 \left\| \|\varepsilon\|_{\mathbb{P},6}^2 \|\|\bar{\mathbf{S}}_2\|_\infty \right\|_{\mathbb{P},6}^2 \\
&\overset{(iii)}{=} O_p \left(\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d)(\log d)^2}{N^2} \right),
\end{aligned}$$

where (i) holds since $|A_1 A_2| \leq 1$; (ii) holds by Hölder's inequality; (iii) holds by Lemma 2.4 and (2.80). By Lemma 2.2,

$$\left\| \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*) - \boldsymbol{\nabla}_{\boldsymbol{\delta}} \bar{\ell}_3(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*) \right\|_\infty = O_p \left(\frac{\sqrt{s_{\boldsymbol{\gamma}} \log d_1} \log d}{N} \right).$$

Together with (2.89), we have

$$\left\| \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*) \right\|_\infty = O_p \left(\left(1 + \sqrt{\frac{(s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d) \log d}{N}} \right) \sqrt{\frac{\log d}{N}} \right).$$

(c) Let $\nu(\cdot) = \nu^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} (\mathbf{W}_{\boldsymbol{\beta},1,i} + \mathbf{W}_{\boldsymbol{\beta},2,i}),$$

where

$$\begin{aligned}
\mathbf{W}_{\boldsymbol{\beta},1,i} &:= -2A_{1i} \{ \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) - \exp(-\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*) \} \zeta_i \mathbf{S}_{1i}, \\
\mathbf{W}_{\boldsymbol{\beta},2,i} &:= -2A_{1i} A_{2i} \{ \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) g^{-1}(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}) - \exp(-\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*) g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*) \} \varepsilon_i \mathbf{S}_{1i}.
\end{aligned}$$

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,

$$\mathbb{E}(\mathbf{W}_{\boldsymbol{\beta},1}) = \mathbb{E}(\mathbf{W}_{\boldsymbol{\beta},2}) = \mathbf{0} \in \mathbb{R}^{d_1}.$$

By Lemma 2.5, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},1,i}\right\|_{\infty}^{2}\right) &\leq M^{-1}(2e\log d_1 - e)\mathbb{E}(\|\mathbf{W}_{\boldsymbol{\beta},1}\|_{\infty}^{2}) \\
&\overset{(i)}{\leq} 4M^{-1}(2e\log d_1 - e)\mathbb{E}\left\{\left|\exp(-\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}}) - \exp(\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)\right|^2\zeta^2\|\mathbf{S}_1\|_{\infty}^{2}\right\} \\
&\overset{(ii)}{\leq}{}^{-1} (2e\log d_1 - e)\left\|\exp(-\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}}) - \exp(\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},6}^{2}\left\|\zeta\|_{\mathbb{P},6}^{2}\right\|\|\mathbf{S}_1\|_{\infty}\|_{\mathbb{P},6}^{2} \\
&\overset{(iii)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d_1)^2}{N^2}\right),
\end{aligned}$$

where (i) holds since $|A_1| \leq 1$; (ii) holds by Hölder's inequality; (iii) holds by Lemma 2.13 and Lemma 2.4. Similarly, by Lemma 2.5, we also have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},2,i}\right\|_{\infty}^{2}\right) &\leq M^{-1}(2e\log d_1 - e)\mathbb{E}(\|\mathbf{W}_{\boldsymbol{\beta},2}\|_{\infty}^{2}) \\
&\overset{(i)}{\leq} 4M^{-1}(2e\log d_1 - e)\mathbb{E}\left\{\left|\frac{\exp(-\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^{\top}\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^{\top}\boldsymbol{\delta}^*)}\right|^2\varepsilon^2\|\mathbf{S}_1\|_{\infty}^{2}\right\} \\
&\overset{(ii)}{\leq} 4M^{-1}(2e\log d_1 - e)\left\|\frac{\exp(-\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^{\top}\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^{\top}\boldsymbol{\delta}^*)}\right\|_{\mathbb{P},6}^{2}\left\|\varepsilon\|_{\mathbb{P},6}^{2}\right\|\|\mathbf{S}_1\|_{\infty}\|_{\mathbb{P},6}^{2} \\
&\overset{(iii)}{=} O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d)(\log d_1)^2}{N^2}\right),
\end{aligned}$$

where (i) holds since $|A_1A_2| \leq 1$; (ii) holds by Hölder's inequality; (iii) holds by Lemma 2.4 and, analogously as in (2.80),

$$\left\|\frac{\exp(-\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}})}{g(\bar{\mathbf{S}}_2^{\top}\widehat{\boldsymbol{\delta}})} - \frac{\exp(-\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)}{g(\bar{\mathbf{S}}_2^{\top}\boldsymbol{\delta}^*)}\right\|_{\mathbb{P},6} = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d}{N}}\right).$$

Hence, it follows that

$$\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left\{\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty^2\right\}$$
$$=O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d)(\log d_1)^2}{N^2}\right).$$

By Lemma 2.2,

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty=O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d}\log d_1}{N}\right).$$

Together with (2.90), we have

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty=O_p\left(\left(1+\sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d)\log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).$$

(d) Let $\rho(\cdot)=\rho^*(\cdot)$ and $\mu(\cdot)=\mu^*(\cdot)$. Note that

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)-\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}(\mathbf{W}_{\boldsymbol{\beta},3,i}+\mathbf{W}_{\boldsymbol{\beta},4,i}),$$

where

$$\begin{aligned}\mathbf{W}_{\boldsymbol{\beta},3,i}&:=-2A_{1i}\exp(-\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})\left\{1-\frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)}\right\}\bar{\mathbf{S}}_{2i}^\top(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\mathbf{S}_{1i},\\ \mathbf{W}_{\boldsymbol{\beta},4,i}&:=-2A_{1i}\{\exp(-\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})-\exp(-\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\}\left\{\frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)}\varepsilon_i+\zeta_i\right\}\mathbf{S}_{1i}.\end{aligned}\tag{2.91}$$

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,

$$\mathbb{E}(\mathbf{W}_{\boldsymbol{\beta},3})=\mathbb{E}(\mathbf{W}_{\boldsymbol{\beta},4})=\mathbf{0}\in\mathbb{R}^{d_1}.$$

By Lemma 2.5, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},3,i}\right\|_{\infty}^{2}\right) &\leq M^{-1}(2e\log d_1 - e)\mathbb{E}(\|\mathbf{W}_{\boldsymbol{\beta},3}\|_{\infty}^{2}) \\
&\overset{(i)}{\leq} (2e\log d_1 - e)(1+c_0^{-1})^2\mathbb{E}\left\{\exp(\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}})\left\{\bar{\mathbf{S}}_2^{\top}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\right\}^2\|\mathbf{S}_1\|_{\infty}^{2}\right\} \\
&\overset{(ii)}{\leq} (2e\log d_1 - e)(1+c_0^{-1})^2\left\|\exp(\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}})\right\|_{\mathbb{P},3}^{2}\left\|\bar{\mathbf{S}}_2^{\top}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\right\|_{\mathbb{P},6}^{2}\left\|\|\mathbf{S}_1\|_{\infty}\right\|_{\mathbb{P},6}^{2} \\
&\overset{(iii)}{=} O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d)(\log d_1)^2}{N^2}\right),
\end{aligned}$$

where (i) holds since $|A_1\{1 - A_2/g(\bar{\mathbf{S}}_2^{\top}\boldsymbol{\delta}^*)\}| \leq (1 + c_0^{-1})$ under Assumption 2.1; (ii) holds by Hölder's inequality; (iii) holds by Lemmas 2.13, 2.15, and Lemma 2.4. Similarly, we also have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left(\left\|M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta},4,i}\right\|_{\infty}^{2}\right) &\leq M^{-1}(2e\log d_1 - e)\mathbb{E}(\|\mathbf{W}_{\boldsymbol{\beta},4}\|_{\infty}^{2}) \\
&\leq 4M^{-1}(2e\log d_1 - e)\mathbb{E}\left[\left\{\exp(\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}}) - \exp(\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)\right\}^2\left(c_0^{-1}|\varepsilon| + |\zeta|\right)^2\|\mathbf{S}_1\|_{\infty}^{2}\right] \\
&\leq 8M^{-1}(2e\log d_1 - e)\left\|\exp(\mathbf{S}_1^{\top}\widehat{\boldsymbol{\gamma}}) - \exp(\mathbf{S}_1^{\top}\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},6}^{2}\left(c_0^{-2}\|\varepsilon\|_{\mathbb{P},6}^{2} + \|\zeta\|_{\mathbb{P},6}^{2}\right)\left\|\|\mathbf{S}_1\|_{\infty}\right\|_{\mathbb{P},6}^{2} \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1(\log d_1)^2}{N^2}\right), \qquad (2.92)
\end{aligned}$$

where (i) holds by Lemmas 2.13 and Lemma 2.4. Hence, it follows that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}\left\{\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_{\infty}^{2}\right\} \\
&\qquad = O_p\left(\frac{(s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d)(\log d_1)^2}{N^2}\right).
\end{aligned}$$

By Lemma 2.2,

$$\begin{aligned}
&\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_{\infty} \\
&\qquad = O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1 + s_{\boldsymbol{\delta}}\log d + s_{\boldsymbol{\alpha}}\log d}\log d_1}{N}\right).
\end{aligned}$$

Together with (2.90), we have

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta}^*)\right\|_\infty\\&\quad=O_p\left(\left(1+\sqrt{\frac{(s_{\boldsymbol{\gamma}}\log d_1+s_{\boldsymbol{\delta}}\log d+s_{\boldsymbol{\alpha}}\log d)\log d_1}{N}}\right)\sqrt{\frac{\log d_1}{N}}\right).\end{aligned}$$

(e) Let $\rho(\cdot)=\rho^*(\cdot)$, $\nu(\cdot)=\nu^*(\cdot)$, and $\mu(\cdot)=\mu^*(\cdot)$. 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 (2.91). By (2.92) and Lemma 2.2,

$$\left\|\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)-\nabla_{\boldsymbol{\beta}}\bar{\ell}_4(\boldsymbol{\gamma}^*,\boldsymbol{\delta}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)\right\|_\infty=O_p\left(\frac{\sqrt{s_{\boldsymbol{\gamma}}\log d_1}\log d_1}{N}\right).$$

Together with (2.90), 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 2.4. We show the consistency rate of the nuisance estimators under correctly specified models.

(a) Let $\rho(\cdot)=\rho^*(\cdot)$. Then, by Lemma 2.19, when $s_{\boldsymbol{\gamma}}=O(N/(\log d_1\log d))$,

$$\left\|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*)\right\|_\infty=O_p\left(\sqrt{\frac{\log d}{N}}\right).$$

By Lemma 2.17, we have (2.62) when $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\le 1$. In addition, by Lemma 2.13, we also have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\le 1)=1-o(1)$. 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 2.19, when $s_{\boldsymbol{\gamma}} = O(N/(\log d_1 \log d))$ and $s_{\boldsymbol{\delta}} = O(N/(\log d)^2)$,

$$\left\| \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \bar{\ell}_3(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*) \right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N}} \right).$$

By Lemma 2.17, we have (2.64) when $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$ and $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1$. In addition, by Lemmas 2.13 and 2.14, we also have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}} \cup \mathbb{S}_{\boldsymbol{\delta}}}(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1 \cap \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 \leq 1) = 1 - o(1)$. 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 $\mu(\mathbf{S}_1) = \mathbf{S}_1^\top \boldsymbol{\beta}$. Then, by Lemma 2.19, when $s_{\boldsymbol{\gamma}} = O(N/(\log d_1)^2)$ and $s_{\boldsymbol{\delta}} = O(N/(\log d_1 \log d))$,

$$\left\| \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}} \right).$$

That is, for any $t > 0$, there exists some $\lambda_3 \asymp \sqrt{\log d_1 / N}$ such that

$$\mathcal{E}_3 := \{ \|\boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\|_\infty \leq \lambda_3 \}$$

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,

$$\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)^\top \boldsymbol{\Delta}.$$

Hence,

$$\begin{aligned}
&\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\widehat{\boldsymbol{\beta}}\|_1 \leq -\boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)^\top \boldsymbol{\Delta} + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}^*\|_1 \\
&\leq \left\| \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty \|\boldsymbol{\Delta}\|_1 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}^*\|_1 + |R_6| \leq \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}\|_1 / 2 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}^*\|_1 + |R_6|,
\end{aligned}$$

where

$$R_6 := \left\{ \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\}^\top \boldsymbol{\Delta}.$$

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

$$\begin{aligned} |R_6| &= \left| 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left\{ 1 - \frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}})} \right\} \bar{\mathbf{S}}_{2i}^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \mathbf{S}_{1i}^\top \boldsymbol{\Delta} \right| \\ &\leq \frac{\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_7, \end{aligned}$$

where $\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) (\mathbf{S}_{1i}^\top \boldsymbol{\Delta})^2$ and

$$R_7 := 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left\{ 1 - \frac{A_{2i}}{g(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}})} \right\}^2 \left\{ \bar{\mathbf{S}}_{2i}^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \right\}^2.$$

It follows that

$$\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_S\|_1 + 2R_7. \tag{2.93}$$

Condition on $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$, where by Lemma 2.13, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1$ holds with probability $1 - o(1)$. Also, condition on the event that

$$\delta \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \frac{\log d_1}{M} \|\boldsymbol{\Delta}\|_1^2, \tag{2.94}$$

which, by Lemma 2.17, 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 (2.93), we have

$$\|\boldsymbol{\Delta}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1 + 2\lambda_{\boldsymbol{\beta}}^{-1} R_7. \tag{2.95}$$

Note that $\|\boldsymbol{\Delta}_S\|_1 \le \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}_S\|_2 \le \sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2$. Hence,

$$
\begin{aligned}
3\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2R_7 &\ge 3\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_S\|_1 + 2R_7 \overset{(i)}{\ge} \delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) \\
&\overset{(ii)}{\ge} \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d_1}{M}\|\boldsymbol{\Delta}\|_1^2 \overset{(iii)}{\ge} \kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log d_1}{M}\left(4\|\boldsymbol{\Delta}_S\|_1 + 2\lambda_{\boldsymbol{\beta}}^{-1}R_7\right)^2 \\
&\ge \kappa_1\|\boldsymbol{\Delta}\|_2^2 - 4\kappa_2\frac{\log d_1}{M}\left(4\|\boldsymbol{\Delta}_S\|_1^2 + \lambda_{\boldsymbol{\beta}}^{-2}R_7^2\right) \\
&\ge \kappa_1\|\boldsymbol{\Delta}\|_2^2 - 4\kappa_2\frac{\log d_1}{M}\left(4s_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_2^2 + \lambda_{\boldsymbol{\beta}}^{-2}R_7^2\right) \ge \frac{\kappa_1}{2}\|\boldsymbol{\Delta}\|_2^2 - \frac{4\kappa_2\log d_1}{\lambda_{\boldsymbol{\beta}}^2 M}R_7^2,
\end{aligned}
$$

when $M > 32\kappa_2 s_{\boldsymbol{\beta}}\log d_1/\kappa_1$. Here, (i) follows from (2.93) and the fact that $\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{S^c}\|_1 \ge 0$; (ii) holds under the event that (2.94) occurs; (iii) follows from (2.95). By Lemma 2.12,

$$
\|\boldsymbol{\Delta}\|_2 \le \frac{6\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{\frac{8\kappa_2 R_7^2\log d_1}{\kappa_1\lambda_{\boldsymbol{\beta}}^2 M} + \frac{4R_7}{\kappa_1}}. \tag{2.96}
$$

Now, we upper bound the term R_7. Observe that

$$
\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}(R_7) &= 2\mathbb{E}\left[\exp(-\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})\left\{1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})}\right\}^2\left\{\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2\right] \\
&\overset{(i)}{\le} 2\left\|\exp(-\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})\right\|_{\mathbb{P},3}\left\|1 - \frac{A_2}{g(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})}\right\|_{\mathbb{P},6}^2\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},6}^2 \\
&\overset{(ii)}{\le} 2\left\|\exp(-\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}})\right\|_{\mathbb{P},3}\left\{1 + \left\|g^{-1}(\bar{\mathbf{S}}_2^\top\widehat{\boldsymbol{\delta}})\right\|_{\mathbb{P},6}\right\}^2\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\|_{\mathbb{P},6}^2 \\
&\overset{(iii)}{=} O_p\left(\frac{s_{\boldsymbol{\alpha}}\log d}{N}\right),
\end{aligned}
$$

where (i) holds by Hölder's inequality; (ii) holds by Minkowski inequality; (iii) holds by Lemmas 2.13, 2.14, and 2.15. By Lemma 2.2,

$$
R_7 = O_p\left(\frac{s_{\boldsymbol{\alpha}}\log d}{N}\right).
$$

By (2.96) and since $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log d_1/N}$, 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 (2.95),

$$\|\boldsymbol{\Delta}\|_1 \le 4\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2\lambda_{\boldsymbol{\beta}}^{-1}R_7 = O_p\left(s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}} + s_\gamma\sqrt{\frac{(\log d)^2}{N\log d_1}}\right).$$

(d) Let $\rho(\cdot) = \rho^*(\cdot)$ and $\mu(\cdot) = \mu^*(\cdot)$. Then, by Lemma 2.19, when $s_{\boldsymbol{\gamma}} = o(N/(\log d_1)^2)$, $s_{\boldsymbol{\delta}} = o(N/(\log d_1 \log d))$, and $s_{\boldsymbol{\alpha}} = o(N/(\log d_1 \log d))$,

$$\left\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\|_\infty = O_p\left(\sqrt{\frac{\log d_1}{N}}\right).$$

That is, for any $t > 0$, there exists some $\lambda_4 \asymp \sqrt{\log d_1/N}$ such that

$$\mathcal{E}_4 := \{\|\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\|_\infty \le \lambda_4\}$$

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

$$\begin{aligned}
|R_8| &= \left|\left\{\boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}}\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*)\right\}^\top \boldsymbol{\Delta}\right| \\
&= \left|2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}A_{2i}\exp(-\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})\left\{g^{-1}(\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\right\}\widehat{\varepsilon}_i\mathbf{S}_{1i}^\top\boldsymbol{\Delta}\right| \\
&\le \frac{\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_9.
\end{aligned}$$

Here, $\widehat{\varepsilon}_i := Y_i(1,1) - \bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\alpha}}$,

$$\begin{aligned}
\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) =& M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{1i}\exp(-\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^\top\boldsymbol{\Delta})^2, \\
R_9 :=& 2M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\beta}}} A_{2i}\exp(-\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}})\left\{g^{-1}(\bar{\mathbf{S}}_{2i}^\top\widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\delta}^*)\right\}^2\widehat{\varepsilon}_i^2.
\end{aligned}$$

Observe that

$$\begin{aligned}\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}(R_9) &= 2\mathbb{E}\left[A_2 \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \right\}^2 \widehat{\varepsilon}^2 \right] \\ &\leq 2 \left\| \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) \right\|_{\mathbb{P},3} \left\| g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \right\|_{\mathbb{P},6}^2 \|\widehat{\varepsilon}\|_{\mathbb{P},6}^2 \\ &\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\delta}} \log d}{N} \right),\end{aligned}$$

where (i) holds by Lemmas 2.13, 2.14, and 2.15. By Lemma 2.2,

$$R_9 = O_p\left(\frac{s_{\boldsymbol{\delta}} \log d}{N} \right).$$

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 2.19, when $s_{\boldsymbol{\gamma}} = o(N/(\log d_1)^2)$, $s_{\boldsymbol{\delta}} = o(N/(\log d_1 \log d))$, and $s_{\boldsymbol{\alpha}} = o(N/(\log d_1 \log d))$,

$$\begin{aligned}\left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty &= O_p\left(\sqrt{\frac{\log d_1}{N}} \right), \\ \left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) \right\|_\infty &= O_p\left(\sqrt{\frac{\log d_1}{N}} \right), \\ \left\| \nabla_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\|_\infty &= O_p\left(\sqrt{\frac{\log d_1}{N}} \right).\end{aligned}$$

Define

$$\mathbf{a} := \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) + \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*).$$

Then $\|\mathbf{a}\|_\infty = O_p(\sqrt{\frac{\log d_1}{N}})$. Hence, for any $t > 0$, there exists some $\lambda_5 \asymp \sqrt{\log d_1/N}$ such that $\mathcal{E}_5 := \{\|\mathbf{a}\|_\infty \leq \lambda_5\}$ 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

$$2\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}_S\|_1 + 2|R_{10}|,$$

where

$$\begin{aligned}
R_{10} &= \left\{ \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \mathbf{a} \right\}^\top \boldsymbol{\Delta} \\
&= \left\{ \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\}^\top \boldsymbol{\Delta} \\
&\quad - \left\{ \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*) - \boldsymbol{\nabla}_{\boldsymbol{\beta}} \bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \right\}^\top \boldsymbol{\Delta} \\
&= 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} A_{2i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*) \right\} \bar{\mathbf{S}}_{2i}^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \mathbf{S}_{1i}^\top \boldsymbol{\Delta}.
\end{aligned}$$

By Young's inequality for products,

$$|R_{10}| \leq \frac{\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta})}{2} + R_{11},$$

where $\delta\bar{\ell}_4(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^*, \boldsymbol{\Delta}) = M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{1i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}})(\mathbf{S}_{1i}^\top \boldsymbol{\Delta})^2$ and

$$R_{11} := 2M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\beta}}} A_{2i} \exp(-\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_{2i}^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*) \right\}^2 \left\{ \bar{\mathbf{S}}_{2i}^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \right\}^2.$$

Observe that

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\beta}}}(R_{11}) &= 2\mathbb{E}\left[A_2 \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) \left\{ g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \right\}^2 \left\{ \bar{\mathbf{S}}_2^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \right\}^2 \right] \\
&\leq 2 \left\| \exp(-\mathbf{S}_1^\top \widehat{\boldsymbol{\gamma}}) \right\|_{\mathbb{P},3} \left\| g^{-1}(\bar{\mathbf{S}}_2^\top \widehat{\boldsymbol{\delta}}) - g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \right\|_{\mathbb{P},6}^2 \left\| \bar{\mathbf{S}}_2^\top (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \right\|_{\mathbb{P},6}^2 \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} (\log d)^2}{N^2} \right),
\end{aligned}$$

where (i) holds by Lemmas 2.13, 2.14, and 2.15. By Lemma 2.2,

$$R_{11} = O_p\left(\frac{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} (\log d)^2}{N^2}\right).$$

Repeat the same procedure as in parts (c) and (d), we have

$$\|\boldsymbol{\Delta}\|_2 \leq \frac{6\lambda_{\boldsymbol{\beta}}\sqrt{s_{\boldsymbol{\beta}}}}{\kappa_1} + \sqrt{\frac{8\kappa_2 R_{11}^2 \log d_1}{\kappa_1 \lambda_{\boldsymbol{\beta}}^2 M} + \frac{4R_{11}}{\kappa_1}},$$

$$\|\boldsymbol{\Delta}\|_1 \leq 4\sqrt{s_{\boldsymbol{\beta}}}\|\boldsymbol{\Delta}\|_2 + 2\lambda_{\boldsymbol{\beta}}^{-1} R_{11},$$

with probability at least $1 - t - o(1)$. Hence,

$$\|\boldsymbol{\Delta}\|_2 = O_p\left(\frac{\sqrt{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}}}\log d}{N} + \sqrt{\frac{s_{\boldsymbol{\beta}} \log d_1}{N}}\right),$$

$$\|\boldsymbol{\Delta}\|_1 = O_p\left(\frac{s_{\boldsymbol{\delta}} s_{\boldsymbol{\alpha}} \log d}{N}\sqrt{\frac{(\log d)^2}{N \log d_1}} + s_{\boldsymbol{\beta}}\sqrt{\frac{\log d_1}{N}}\right).$$

■

2.7.5 Proofs of the auxiliary Lemmas

Proof of Lemma 2.8. We prove the lemma by considering two cases separately.

(a) If $d \leq m$. Choose $S = \{1, \ldots, d\}$. Since $\mathbf{X}$ is a sub-Gaussian vector, we have

$$\sup_{\|\boldsymbol{\beta}\|_2 = 1} \mathbb{E}\{(\mathbf{X}^\top \boldsymbol{\beta})^2\} = O(1). \tag{2.97}$$

For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, by triangle inequality, we have

$$\begin{aligned} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2 &\leq \|\boldsymbol{\Delta}\|_2^2 \sup_{\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\beta})^2 \\ &\leq \|\boldsymbol{\Delta}\|_2^2 \left[\sup_{\|\boldsymbol{\beta}\|_2=1} \mathbb{E}\{(\mathbf{X}^\top \boldsymbol{\beta})^2\} + \sup_{\|\boldsymbol{\beta}\|_2=1} \left| m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\beta})^2 - \mathbb{E}\{(\mathbf{X}^\top \boldsymbol{\beta})^2\} \right| \right] \end{aligned}$$

It follows that

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2}$$
$$\leq \left[\sup_{\|\boldsymbol{\beta}\|_2=1}\mathbb{E}\{(\mathbf{X}^\top\boldsymbol{\beta})^2\}+\sup_{\|\boldsymbol{\beta}\|_2=1}\left|m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\beta})^2-\mathbb{E}\{(\mathbf{X}^\top\boldsymbol{\beta})^2\}\right|\right]$$

By Lemma 2.7 and (2.97), we have, as $m, d\to\infty$,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2} = O_p\left(1+\sqrt{\frac{d}{m}}\right)\overset{(i)}{=}O_p(1)$$

where (i) holds since $d\leq m$. Hence,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2+\|\boldsymbol{\Delta}\|_2^2} \leq \sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_2^2} = O_p(1)$$

(b) If $d>m$. Choose any set $S\subseteq\{1,\ldots,d\}$ such that $s:=|S|\asymp m$. For any $\boldsymbol{\Delta}\in\mathbb{R}^d$, define $\widetilde{\boldsymbol{\Delta}}=(\widetilde{\boldsymbol{\Delta}}_S^\top,\widetilde{\boldsymbol{\Delta}}_{S^c}^\top)^\top\in\mathbb{R}^d$ such that

$$\widetilde{\boldsymbol{\Delta}}_S = s^{-1}\|\boldsymbol{\Delta}\|_1(1,\ldots,1)^\top\in\mathbb{R}^s,\quad \widetilde{\boldsymbol{\Delta}}_{S^c}=\boldsymbol{\Delta}_{S^c}\in\mathbb{R}^{d-s}.$$

Then

$$\|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1=\|\boldsymbol{\Delta}_{S^c}\|_1\leq\|\boldsymbol{\Delta}\|_1=\|\widetilde{\boldsymbol{\Delta}}_S\|_1.$$

Hence, $\widetilde{\boldsymbol{\Delta}}\in\mathbb{C}(S,3):=\{\boldsymbol{\Delta}\in\mathbb{R}^d:\|\boldsymbol{\Delta}_{S^c}\|_1\leq 3\|\boldsymbol{\Delta}_S\|_1\}$. 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 $(a+b)^2\leq 2a^2+2b^2$, we have

$$m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\Delta})^2\leq 2m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\widetilde{\boldsymbol{\Delta}})^2+2m^{-1}\sum_{i=1}^m\left\{\mathbf{X}_i^\top(\widetilde{\boldsymbol{\Delta}}-\boldsymbol{\Delta})\right\}^2$$
$$\leq 2\left(\|\widetilde{\boldsymbol{\Delta}}\|_2^2+\|\widetilde{\boldsymbol{\Delta}}-\boldsymbol{\Delta}\|_2^2\right)\sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m(\mathbf{X}_i^\top\boldsymbol{\beta})^2.$$

Now, we observe that

$$\|\widetilde{\boldsymbol{\Delta}}\|_2^2 = \|\widetilde{\boldsymbol{\Delta}}_S\|_2^2 + \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_2^2 = s^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}_{S^c}\|_2^2,$$

$$\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_2^2 = \|\widetilde{\boldsymbol{\Delta}}_S - \boldsymbol{\Delta}_S\|_2^2 \le 2\|\widetilde{\boldsymbol{\Delta}}_S\|_2^2 + 2\|\boldsymbol{\Delta}_S\|_2^2 = 2s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 2\|\boldsymbol{\Delta}_S\|_2^2.$$

Hence, we have

$$m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2 \le 2\left(3s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 2\|\boldsymbol{\Delta}\|_2^2\right) \sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\beta})^2, \ \ \forall \boldsymbol{\Delta}\in\mathbb{R}^d,$$

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

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2}{6s^{-1}\|\boldsymbol{\Delta}\|_1^2 + 4\|\boldsymbol{\Delta}\|_2^2} \le \sup_{\boldsymbol{\beta}\in\mathbb{C}(S,3)\cap\|\boldsymbol{\beta}\|_2=1} m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\beta})^2,$$

By Lemma 2.7 and (2.97), as $m, d \to \infty$,

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2}{s^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p\left(1 + \sqrt{\frac{s}{m}}\right).$$

Besides, note that $s \asymp m$ and hence $1 + \sqrt{s/m} = O(1)$. It follows that

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{m^{-1}\sum_{i=1}^m (\mathbf{X}_i^\top \boldsymbol{\Delta})^2}{m^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

■

Proof of Lemma 2.9. Note that

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}) :=& \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta} - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1 \\ =& \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta} + R_1(\boldsymbol{\Delta}) - \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^*\|_1, \qquad (2.98)\end{aligned}$$

where

$$\begin{aligned}R_1(\boldsymbol{\Delta}) :=& \left\{\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)\right\}^\top \boldsymbol{\Delta} \\ =& M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}\left\{g^{-1}(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*)\right\}\left\{1 - A_{2i}g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*)\right\}\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta}.\end{aligned}$$

Let $\lambda_{\boldsymbol{\delta}} = 2\sigma_{\boldsymbol{\delta}}\sqrt{(t+\log d)/M}$ with some $t > 0$. By Lemma 2.18, we have $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1) \geq 1 - 2\exp(-t)$. On the event $\mathcal{A}_1$, we have $|\nabla_{\boldsymbol{\delta}}\bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}| \leq \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1/2$. Note that $\|\boldsymbol{\delta}^*\|_1 = \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}}\|_1 \leq \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}} + \boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1$, $\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S^c_{\boldsymbol{\delta}}}\|_1$, and $\|\boldsymbol{\delta}^* + \boldsymbol{\Delta}\|_1 = \|\boldsymbol{\delta}^*_{S_{\boldsymbol{\delta}}} + \boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 + \|\boldsymbol{\Delta}_{S^c_{\boldsymbol{\delta}}}\|_1$. Recall the equation (2.98). It follows that

$$2\mathcal{F}(\boldsymbol{\Delta}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S^c_{\boldsymbol{\delta}}}\|_1 - 3\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta})|.$$

Hence,

$$2\mathcal{F}(\boldsymbol{\Delta}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1 - 4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta})|. \tag{2.99}$$

Under the overlap condition in Assumption 2.1 and since $|A_1| \leq 1$,

$$\begin{aligned}
|R_1(\boldsymbol{\Delta})| &\leq (1 + c_0^{-1})M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}} A_{1i}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\right\}\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\Delta} \\
&\overset{(i)}{\leq} (1 + c_0^{-1})\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\right\}^2}\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\Delta})^2},
\end{aligned}$$

where (i) holds by the Cauchy–Schwarz inequality. It follows that

$$\begin{aligned}
&\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{|R_1(\boldsymbol{\Delta})|}{\|\boldsymbol{\Delta}\|_1/\sqrt{N} + \|\boldsymbol{\Delta}\|_2} \\
&\leq (1 + c_0^{-1})\sqrt{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\right\}^2}\sqrt{\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}}\frac{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\Delta})^2}{N^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2}},
\end{aligned}$$

since $(\|\boldsymbol{\Delta}\|_1/\sqrt{N} + \|\boldsymbol{\Delta}\|_2)^2 > N^{-1}\|\boldsymbol{\Delta}\|_1^2 + \|\boldsymbol{\Delta}\|_2^2$. Note that

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\boldsymbol{\delta}}}\left[M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\right\}^2\right] &= \mathbb{E}\left[\left\{g^{-1}(\mathbf{S}_1^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\}^2\right] \\
&\overset{(i)}{=} O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}\right),
\end{aligned}$$

where (i) holds by Lemma 2.13. By Lemma 2.2,

$$M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}\left\{g^{-1}(\mathbf{S}_{1i}^\top\widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top\boldsymbol{\gamma}^*)\right\}^2 = O_p\left(\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}\right).$$

Besides, by Lemma 2.8, we also have

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d/\{\mathbf{0}\}} \frac{M^{-1}\sum_{i\in\mathcal{I}_{\boldsymbol{\delta}}}(\bar{\mathbf{S}}_{2i}^\top\boldsymbol{\Delta})^2}{N^{-1}\|\boldsymbol{\Delta}\|_1^2+\|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

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 $t>0$, when N is large enough, there exists some constant $c>0$ such that $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_2)\geq 1-t$. Hence,

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}\cup\mathbb{S}_{\boldsymbol{\delta}}}(\mathcal{A}_1\cap\mathcal{A}_2)\geq 1-t-2\exp(-t).$$

Recall the definitions (2.35) and (2.36). Now, conditional on $\mathcal{A}_1\cap\mathcal{A}_2$, we have

$$2\mathcal{F}(\boldsymbol{\Delta})\geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta})+\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1-4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{S_{\boldsymbol{\delta}}}\|_1-2c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\left(\frac{\|\boldsymbol{\Delta}\|_1}{\sqrt{N}}+\|\boldsymbol{\Delta}\|_2\right).$$

Since $\lambda_{\boldsymbol{\delta}}=2\sigma_{\boldsymbol{\delta}}\sqrt{(t+\log d)/M}\geq 2\sigma_{\boldsymbol{\delta}}\sqrt{\log d_1/M}$, $M\asymp N$, and $s_{\boldsymbol{\gamma}}=o(N)$, we have $\sqrt{s_{\boldsymbol{\gamma}}\log d_1/N^2}=o(\lambda_{\boldsymbol{\delta}})$. Hence, with some $N_0>0$, when $N>N_0$, we have $4c\sqrt{s_{\boldsymbol{\gamma}}\log d_1/N^2}\leq\lambda_{\boldsymbol{\delta}}$. 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,

$$4\mathcal{F}(\boldsymbol{\Delta})\geq 4\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}},\boldsymbol{\delta}^*,\boldsymbol{\Delta})+\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}\|_1-\left(8\lambda_{\boldsymbol{\delta}}\sqrt{s_{\boldsymbol{\delta}}}+4c\sqrt{\frac{s_{\boldsymbol{\gamma}}\log d_1}{N}}\right)\|\boldsymbol{\Delta}\|_2. \tag{2.100}$$

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, conditional on the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, when $N > N_1$ with some constant $N_1 > 0$,

$$\mathcal{F}(\boldsymbol{\Delta}) \geq \kappa_1 - \kappa_2 k_0^2 \frac{\bar{s}_{\boldsymbol{\delta}} \log d}{M} - 2\lambda_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}} - \frac{c}{2} \sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}} \geq \kappa_1/2,$$

since as $N \to \infty$, we have $\bar{s}_{\boldsymbol{\delta}} \log d/M = s_{\boldsymbol{\gamma}} \log d_1/M + s_{\boldsymbol{\delta}} \log d/M = o(1)$, $\sqrt{s_{\boldsymbol{\gamma}} \log d_1/N} = o(1)$, and $2\lambda_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}} = 4\sigma_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}}(t + \log d)/M} \leq 4\sigma_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}} t/M} + 4\sigma_{\boldsymbol{\delta}} \sqrt{s_{\boldsymbol{\delta}} \log d/M} \leq \kappa_1/4 + o(1)$ when $t < \kappa_1^2 M/(16^2 \sigma_{\boldsymbol{\delta}}^2 s_{\boldsymbol{\delta}})$. ∎

Proof of Lemma 2.10. 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 (2.59), we have $\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) = \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}) - \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}}$. It follows that

$$\begin{aligned} \mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) &= \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}} \|\widehat{\boldsymbol{\delta}}\|_1 + \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}} - \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}^*\|_1 \qquad (2.101) \\ &= \delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}^* + \boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 + \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*)^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}} + R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) - \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}^*\|_1 \leq 0, \end{aligned}$$

where

$$\begin{aligned} R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) &= \left\{ \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*) - \nabla_{\boldsymbol{\delta}} \bar{\ell}_2(\boldsymbol{\gamma}^*, \boldsymbol{\delta}^*) \right\}^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}} \\ &= M^{-1} \sum_{i \in \mathcal{I}_{\boldsymbol{\delta}}} A_{1i} \left\{ g^{-1}(\mathbf{S}_{1i}^\top \widehat{\boldsymbol{\gamma}}) - g^{-1}(\mathbf{S}_{1i}^\top \boldsymbol{\gamma}^*) \right\} \left\{ 1 - A_{2i} g^{-1}(\bar{\mathbf{S}}_{2i}^\top \boldsymbol{\delta}^*) \right\} \bar{\mathbf{S}}_{2i}^\top \boldsymbol{\Delta}_{\boldsymbol{\delta}}. \end{aligned}$$

Repeat the same procedure in the proof of Lemma 2.9 for obtaining (2.99) and (2.100). Then, conditional on $\mathcal{A}_1$, we have

$$0 \geq 2\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\delta}}) \geq 2\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) + \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_1 - 4\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\Delta}_{\boldsymbol{\delta},S_{\boldsymbol{\delta}}}\|_1 - 2|R_1(\boldsymbol{\Delta}_{\boldsymbol{\delta}})|. \tag{2.102}$$

Conditional 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 the equation (2.65). We have $\delta\bar{\ell}_2(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\delta}^*, \boldsymbol{\Delta}_{\boldsymbol{\delta}}) \geq 0$. Since $\lambda_{\boldsymbol{\delta}} = 2\sigma_{\boldsymbol{\delta}}\sqrt{(t+\log d)/M} \geq 2\sigma_{\boldsymbol{\delta}}\sqrt{\log d/M}$ and $N \asymp M$, 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 $N > N_0$ with some $N_0 > 0$. ∎

Proof of Lemma 2.11. 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 2.9, $\mathcal{F}(\widetilde{\boldsymbol{\Delta}}) > 0$. Thus, we conclude that $\|\boldsymbol{\Delta}_{\boldsymbol{\delta}}\|_2 \leq 1$. ∎

Proof of Lemma 2.12.

$$x \leq \frac{b+\sqrt{b^2+4ac}}{2a} \leq \frac{b+\sqrt{b^2}+\sqrt{4ac}}{2a} = \frac{b}{a} + \sqrt{\frac{c}{a}}.$$

■

Proof of Lemma 2.13. Let $\mathcal{X}$ the support of $\mathbf{S}_1$. Under Assumption 2.1, for all $\mathbf{S}_1 \in \mathcal{X}$, there exists some constant $c > 0$ such that

$$\exp(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) \leq c, \quad \exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*) < g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*) \leq c.$$

By Theorem 2.3,

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right).$$

Since $\mathbf{S}_1$ is a sub-Gaussian random vector under Assumption 2.4, by Theorem 2.6 of [Wai19],

$$\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1}{N}}\right).$$

Additionally, note that $s_{\boldsymbol{\gamma}} = o(N/\log d_1)$. It follows that

$$\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq 1) = 1 - o(1).$$

For any $\boldsymbol{\gamma} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\}$, we have

$$\begin{aligned}\left\|g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}) - g^{-1}(\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} &= \left\|\exp(-\mathbf{S}_1^\top \boldsymbol{\gamma}^*)\left[\exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right\} - 1\right]\right\|_{\mathbb{P},r} \\ &\leq c\left\|\exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right\} - 1\right\|_{\mathbb{P},r} \end{aligned} \tag{2.103}$$

By Taylor's Thorem, for any $\mathbf{S}_1 \in \mathcal{X}$, with some $v \in (0,1)$,

$$\begin{aligned}\left|\exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right\} - 1\right| &= \exp\left\{-v\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right\}\left|\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right| \\ &\leq \left[1 + \exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right\}\right]\left|\mathbf{S}_1^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right|. \end{aligned} \tag{2.104}$$

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

$$\begin{aligned}
\left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}) - g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} &\overset{(i)}{\le} c\left\|\left[1+\exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\right]\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} \\
&\overset{(ii)}{\le} c\left\|\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} + c\left\|\exp\left\{-\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\}\right\|_{\mathbb{P},2r}\left\|\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},2r} \\
&\overset{(iii)}{=} O\left(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\right), \qquad (2.105)
\end{aligned}$$

where (i) holds by (2.103) and (2.104); (ii) holds by Minkowski inequality and Hölder's inequality; (iii) holds by Theorem 2.6 of [Wai19] under Assumption 2.4. It follows that,

$$\left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r} \le \left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r} + O\left(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\right) \le C,$$

with some constant $C > 0$, since $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 \le 1$. Therefore, we conclude that $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1-o(1)$. Moreover, by the fact that $\exp(-u) = g^{-1}(u) - 1 < g^{-1}(u)$ and $\|X\|_{\mathbb{P},r'} \le \|X\|_{\mathbb{P},12}$ for any $X \in \mathbb{R}$ and $1 \le r' \le 12$, we have

$$\left\|g^{-1}(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} \le C, \quad \left\|\exp(-\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} \le C.$$

Moreoever, we have (2.38), since $\widehat{\boldsymbol{\gamma}} \in \{w\boldsymbol{\gamma}^* + (1-w)\widehat{\boldsymbol{\gamma}} : w \in [0,1]\}$, $\mathbb{P}_{\mathbb{S}_{\boldsymbol{\gamma}}}(\mathcal{E}_1) = 1-o(1)$, and (2.105) holds. Besides, note that

$$\begin{aligned}
\left\|\exp(\mathbf{S}_1^\top\boldsymbol{\gamma}) - \exp(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r'} &\le c\left\|\exp\{\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\} - 1\right\|_{\mathbb{P},r'} \\
&\le c\left[\left\|\exp\{\mathbf{S}_1^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\}\right\|_{\mathbb{P},r'} + 1\right] = O(1),
\end{aligned}$$

since $\mathbf{S}_1$ is sub-Gaussian and $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le 1$. Therefore,

$$\begin{aligned}
\left\|\exp(\mathbf{S}_1^\top\boldsymbol{\gamma})\right\|_{\mathbb{P},r'} &\le \left\|\exp(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r'} + \left\|\exp(\mathbf{S}_1^\top\boldsymbol{\gamma}) - \exp(\mathbf{S}_1^\top\boldsymbol{\gamma}^*)\right\|_{\mathbb{P},r'} \\
&\le c + O(1) = O(1).
\end{aligned}$$

■

Proof of Lemma 2.14. Let $\mathcal{S}$ the support of $\bar{\mathbf{S}}_2$. Under Assumption 2.1, there exists some constant $c > 0$, such that, for all $\bar{\mathbf{S}}_2 \in \mathcal{S}$,

$$\exp(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \leq c, \quad \exp(-\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) < g^{-1}(\bar{\mathbf{S}}_2^\top \boldsymbol{\delta}^*) \leq c.$$

(a) By Theorem 2.3,

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right) = o_p(1).$$

By Assumption 2.4 and Theorem 2.6 of [Wai19],

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\gamma}} \log d_1 + s_{\boldsymbol{\delta}} \log d}{N}}\right).$$

(b) By Theorem 2.4,

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2 = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right) = o_p(1).$$

Similarly, by Assumption 2.4 and Theorem 2.6 of [Wai19],

$$\left\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*)\right\|_{\mathbb{P},r} = O\left(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2\right) = O_p\left(\sqrt{\frac{s_{\boldsymbol{\delta}} \log d}{N}}\right).$$

The remaining proof is an analog of the proof of Lemma 2.13. ∎

Proof of Lemma 2.15. The upper bounds for $\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{\mathbb{P},r}$ follow directly from Theorems 2.3, 2.4, Theorem 2.6 of [Wai19], and the sub-Gaussianity of $\bar{\mathbf{S}}_2$ assumed in Assumption 2.4. Let either (a) or (b) holds. Then we have $\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{\mathbb{P},r} = o_p(1)$. Note that, $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = (1 - v_1)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$. Therefore,

$$\begin{aligned}\|\widetilde{\varepsilon}\|_{\mathbb{P},r} &\leq \|\varepsilon\|_{\mathbb{P},r} + \|\bar{\mathbf{S}}_2^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{\mathbb{P},r} = \|\varepsilon\|_{\mathbb{P},r} + (1 - v_1)\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_{\mathbb{P},r} \\ &= O(1) + o_p(1) = O_p(1).\end{aligned}$$

∎

Proof of Lemma 2.16. The upper bounds for $\left\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_{\mathbb{P},r}$ follow directly by Theorems 2.3, 2.4, Theorem 2.6 of [Wai19], and the sub-Gaussianity of $\mathbf{S}_1$ assumed in Assumption 2.4. Let either (a) or (b) of Lemma 2.16 holds, and let either (a) or (b) of 2.15 holds. Then we have $\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_{\mathbb{P},r} = o_p(1)$ and $\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\|_{\mathbb{P},r} = o_p(1)$. 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,

$$\begin{aligned}
\|\widetilde{\zeta}\|_{\mathbb{P},r} &\le \|\zeta\|_{\mathbb{P},r} + \|\mathbf{S}_1^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_{\mathbb{P},r} + \|\bar{\mathbf{S}}_2^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\|_{\mathbb{P},r} \\
&= \|\zeta\|_{\mathbb{P},r} + (1-v_1)\|\mathbf{S}_1^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_{\mathbb{P},r} + (1-v_2)\|\bar{\mathbf{S}}_2^\top(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\|_{\mathbb{P},r} \\
&= O(1) + o_p(1) = O_p(1).
\end{aligned}$$

■

2.8 Acknowledgement

Chaper 2, in full, has been submitted for publication of the material. Zhang, Yuqian; Ji, Weijie; Bradic, Jelena. Dynamic treatment effects: high-dimensional inference under model misspecification. The dissertation author was the primary investigator and author of this material.

Chapter 3

Adaptive split balancing for optimal random forests

The random forests method, introduced by [BFSO84, Bre01], currently stands as one of the most popular approaches for tackling classification and regression problems, exhibiting significant empirical success across a diverse range of real-world applications. Extensions of random forests to address other statistical challenges have also been extensively investigated, encompassing quantile estimation [MR06], survival analysis [IKBL08, IK10], and feature selection or importance evaluation [GPB11, MH14, LWSG13, LWB$^+$19, BWLY22]. However, despite its widespread use, the theoretical analysis of this method remains incomplete.

Let $\mathbb{S}_N := (Y_i, \mathbf{X}_i)_{i=1}^N$ be independent and identically distributed (i.i.d.) samples, and denote $(Y, \mathbf{X})$ as an independent copy of $(Y_i, \mathbf{X}_i)$. Here, $Y \in \mathbb{R}$ is the response variable and $\mathbf{X} \in [0,1]^d$ denotes the covariate vector. Consider the estimation of the conditional mean function $m(\mathbf{x}) := \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ for any $\mathbf{x} \in [0,1]^d$. In this paper, we mainly focus on the

integrated mean squared error (IMSE) $\mathbb{E}_{\mathbf{x}}[\widehat{m}(\mathbf{x}) - m(\mathbf{x})]^2$, where $\widehat{m}(\cdot)$ denotes the random forest constructed based on $\mathbb{S}_N$, and the expectation above is only taken with respect to the new observation $\mathbf{x}$.

The consistency of Breiman's original algorithm has been demonstrated by [SBV15]; nevertheless, they did not provide a specific consistency rate. Recently, [CVFL22] established the consistency rate of the original algorithm under a "sufficient impurity decrease" (SID) condition. Their findings indicate that Breiman's original algorithm maintains consistency even in scenarios where the regression function is discontinuous. However, the established consistency rate is slow for smooth functions, as illustrated in Table 3.1.

Due to the theoretical challenges associated with analyzing Breiman's original random forests, [Bia12] investigated a simplified version named "centered random forests." In centered random forests, splitting directions are chosen randomly, and splitting points are selected as the midpoint of the parent nodes. Centered forests fall under the category of "purely random forests" [MGS20, OT21, Bia12, BDL08, AG14, Klu21], where the trees grow independently from all the samples. Among these studies, [MGS20] achieved the minimax rate for the Hölder class $\mathcal{H}^{0,\beta}$ ($\beta \in (0, 1]$), and [OT21] further attained the minimax rate for the class $\mathcal{H}^{1,\beta}$. However, all purely random forests methods grow trees independently of the observed samples, limiting the utilization of information from the data during the tree-growing process. Recently, [GXZ22] analyzed the usage of random forests for classification problems and established nearly optimal rates for the classification error when the conditional probability function is Lipschitz continuous. They considered a modified version of centered forests employing an "early stopping" technique to prevent overfitting; however,

their proposed method is applicable only when the outcome variable is discrete.

A slightly more sophisticated variant, termed "median forests", has been investigated by [Klu21, DS18]. In median forests, the splitting directions are also randomly chosen. However, in contrast to centered forests, the sample medians are selected as the splitting points rather than the center points. The splitting rules of such methods depend on the covariates but are independent of the outcomes. Notably, the achieved consistency rates are relatively slow, with minimax optimal rates only attained when $d = 1$; refer to Table 3.1.

In a recent line of work, [ATW19, WW15, WA18, FTAW20] explored another variant known as "honest forests", which differs from the original algorithm in three aspects: (a) similarly to centered and median forests, the splitting directions are randomly chosen; (b) the splitting point is determined such that child nodes contain at least a fraction of $\alpha \leq 0.5$ of the samples in the parent nodes; and (c) the forest is "honest" in that two independent sub-sampless are chosen for each tree – only the outcomes from one sub-samples, along with all the covariates, are used for splitting, while the outcomes from the other sub-samples are used solely for local averaging. Unlike other variants, their proposed methods allow the splits to depend on both the covariates and outcomes (when $\alpha < 0.5$). The splitting points can be determined through minimizing the empirical mean squared error within each node, as long as the α-fraction constraint is satisfied. In fact, the α-fraction constraint is crucial for achieving a fast convergence rate. As discussed by [Ish15, BFSO84, CKT22], without an such a constraint, the splits tend to concentrate along the endpoints of the parent node, resulting in slow convergence rates for the single trees, as certain leaves only contain a very small number of samples, making local averaging inaccurate. By imposing the α-

fraction constraint, the child nodes are ensured to contain a certain fraction of the parents, stabilizing the local averaging procedure. However, the consistency rates established for honest forests were also relatively slow; see Table 3.1. We make a further modification to the above approach. Instead of picking a direction uniformly, as in the above aspect (a), we choose the splitting directions in a cyclic way; see details in Algorithm 4.

In the following, we present the rationale behind the cyclic splitting procedure. Let us initially consider a straightforward case with $\alpha = 0.5$, where the proposed method degenerates into splitting at the median each time. In such a degenerate example, the only distinction between the proposed algorithm and standard median forests is how we choose the splitting directions – randomly or cyclically. When the splitting directions are chosen randomly, there is a non-negligible probability that some directions are over-selected while others are barely selected. Consequently, the terminal leaves tend to be too wide in certain directions and too narrow in others. It should be noted that the appearance of this long and narrow leaf structure is entirely determined by auxiliary randomness and has nothing to do with the data – indeed, the splittings are even independent of outcome variables when $\alpha = 0.5$. To avoid the adverse effects of such unnecessary (and even harmful) randomness, we consider a cyclic splitting procedure. When the directions are chosen cyclically, the lengths of leaves in different directions tend to be balanced. This results in better control for the leaves' diameter (see Lemma 3.1), ensuring better control for the algorithm's bias. As shown in Theorem 3.1, the cyclic method leads to a minimax optimal rate for Lipschitz functions and is faster than existing median forests (as long as $d > 1$), where directions are randomly chosen [Klu21, DS18]. Our results indicate that the sub-optimal rates in the existing litera-

ture originated from the inappropriate method of selecting splitting directions. Indeed, when $d = 1$, the cyclic and random methods are the same as there is no need to choose the splitting directions. As a result, the minimax rates have been achieved by [Klu21, DS18]. When $d > 1$, the existing results are sub-optimal, and a simple but essential cyclic modification of the splitting directions' selection leads to a minimax optimal rate.

When $\alpha < 0.5$, we permit CART-like splitting criteria, and the splitting points are allowed to depend on both the covariates and outcomes. In this case, it is possible for the terminal leaves to be relatively wide in certain directions and narrow in others. However, unlike the centered and median forests with random splitting directions [Klu21, DS18], the appearance of the long and narrow leaf structure depends on the data. This distinguishes it from other methods with data-independent splitting rules (e.g., [MGS20, OT21, GXZ22]), allowing us to leverage information from the data during the tree-growing process. This enhancement in empirical performance is particularly notable when different covariates have distinct local effects on the outcome. The tuning parameter α controls the desired balance in the lengths of the leaves. Any constant $\alpha > 0$ prevents making splits near any endpoints of the parent node. In other words, although we allow a certain gap between the lengths of different sides of the leaves, such a gap cannot be too extreme, and the lengths still need to be relatively balanced among different directions.

As an extension of the proposed method, we further consider local polynomial regression within each leaf and provide faster convergence rates when higher-order smoothness conditions are satisfied. The proposed method differs from the local linear forests studied by [FTAW20] in two aspects: (a) the splitting directions are chosen cyclically instead of

randomly, and (b) we allow more general q-th order local polynomial forests with any $q \geq 1$. The proposed method ensures a faster convergence rate even when $q = 1$ – this corresponds to the local linear forests; however, the cyclic splitting procedure leads to a better result.

Other methods that can exploit the higher-order smoothness of regression functions include [CMDY23, CKU23], where data-independent splitting rules are considered. [CMDY23] proposed an extrapolated random tree method. If the regression function belongs to the Hölder class $\mathcal{H}^{q,\beta}$ with some $q \in \mathbb{N}$ and $\beta \in (0, 1]$, their method has a nearly optimal in-sample excess risk; however, they did not provide any upper bounds for the out-of-sample errors. Additionally, [CKU23] proposed a debiased technique based on Mondrian forests and established minimax optimal rates in the point-wise mean squared error $\mathbb{E}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2$ for any interior point $\mathbf{x}$. However, as pointed out in their Section 5.3, their debiasing procedure is only designed to handle interior bias and does not provide any correction for boundary bias. Indeed, since their splitting locations are chosen uniformly from a leaf's side, there is a non-negligible probability that the terminal leaves nearby the boundary only contain a small number of samples. In contrast, the α-constraint imposed in our method avoids making splits near the boundaries. When we set $\alpha = 0.5$, our proposed method yields a minimax rate for the IMSE – marking the first instance of achieving minimax optimal rates for random forests when the Hölder smoothness condition holds with $q > 1$. Furthermore, we establish minimax optimal rates for the uniform convergence rate, as detailed in Section 3.3; these findings represent a novel contribution to the literature for any $q \in \mathbb{N}$.

Of particular significance, when we choose any tuning parameter $\alpha < 0.5$, in contrast

to [CMDY23, CKU23], we harness information from the data, encompassing both $\mathbf{X}_i$ and Y_i, to enhance the empirical performance of the forests. Notably, data-dependent splitting emerges as a crucial factor for random forests to excel in practical applications, surpassing the performance of traditional kernel methods.

Moreover, we employ the proposed methods to estimate the average treatment effect (ATE) using the double machine learning (DML) technique [CCD$^+$17]. The DML method involves the estimation of three nuisance functions: two outcome regression functions and one propensity score function. We leverage the proposed forests to estimate these nuisance functions and scrutinize the performance of forests-based ATE estimator. Although random forests have found extensive applications in causal inference problems, to the best of our knowledge, we are the first to provide theoretical underpinnings for their utilization in ATE estimation and inference; refer to the inherent challenges outlined in Remark 3.1. In contrast to [WA18], which focuses on the estimation of conditional average treatment effect (CATE), we concentrate on the population-level parameter ATE.

3.0.1 Notation

We use the following notations throughout. For any vector $\mathbf{x} \in \mathbb{R}^d$, let $||\mathbf{x}||$ denote the Euclidean norm. Let multi-index $\boldsymbol{\alpha} := (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_d)$ denote a d-tuple of nonnegative integers, and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) \in \mathbb{R}^d$ denote a d-dimensional random variable. We define the following multi-index notations: $|\boldsymbol{\alpha}| := \sum_{i=1}^d \boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}! := \Pi_{i=1}^d \boldsymbol{\alpha}_i!$, and $\mathbf{x}^{\boldsymbol{\alpha}} := \Pi_{i=1}^d \mathbf{x}_i^{\boldsymbol{\alpha}_i}$. Moreover, we denote the partial derivative as $D^{\boldsymbol{\alpha}} f := \frac{\partial^{|\boldsymbol{\alpha}|} f(\mathbf{x})}{\partial \mathbf{x}_1^{\boldsymbol{\alpha}_1} \partial \mathbf{x}_2^{\boldsymbol{\alpha}_2} \cdots \partial \mathbf{x}_d^{\boldsymbol{\alpha}_d}}$. For any $a \in \mathbb{R}$, $\lfloor a \rfloor$ denotes the largest integer no larger than a.

Table 3.1: Comparison of random forests' consistency rates. The reported rates correspond to the integrated mean squared error (IMSE), except that [CKU23] only provided the point-wise mean squared error at interior points, [FTAW20] established normal results at a given $\mathbf{x}$ and established upper bounds for the asymptotic variance, and [CMDY23] considered the in-sample excess risk.

Methods	Consistency rate	Functional class	Algorithm	Splitting criterion
[Bia12]	$N^{-\frac{3/4}{q\log(2)+3/4}}$	q-sparse $\mathcal{H}^{0,1}$	Centered forests	Data-independent
[Gen12]	$N^{-2/3}, d=1$	$\mathcal{H}^{0,1}$	Purely uniform RFs	Data-independent
[AG14]	$N^{\frac{-2\log(1-1/(2d))}{2\log(1-1/(2d))-\log(2)}}$	$\mathcal{H}^{1,1}$, $d\leq 3$	Centered forests	Data-independent
	$N^{\delta-2\log(\frac{2d-1}{2d})}$, $\delta>0$	$\mathcal{H}^{1,1}$, $d\geq 4$		
[MGS20]	$N^{-\frac{2}{d+2}}$	$\mathcal{H}^{0,\beta}$, $\beta\in(0,1]$	Mondrian forests	Data-independent
	$N^{-\frac{2(1+\beta)}{d+2(1+\beta)}}$	$\mathcal{H}^{1,\beta}$, $\beta\in(0,1/2]$		
	$N^{-\frac{3}{d+3}}$	$\mathcal{H}^{1,\beta}$, $\beta\in(1/2,1]$		
[OT21]	$N^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$	$\mathcal{H}^{q,\beta}$, $q\in\{0,1\}$, $\beta\in(0,1]$	Tessellation forests	Data-independent
[CKU23]	$N^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$	$\mathcal{H}^{q,\beta}$, $q\in\mathbb{N}$, $\beta\in(0,1]$	Debiased Mondrian forests	Data-independent
[CMDY23]	$(N/\log(N))^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$	$\mathcal{H}^{q,\beta}$, $q\in\mathbb{N}$, $\beta\in(0,1]$	Extrapolated random tree	Data-independent
[Klu21]	$(N(\log(N))^{(d-1)/2})^{-r}$, $r=\frac{2\log(1-1/(2d))}{2\log(1-1/(2d))-\log(2)}$	$\mathcal{H}^{0,\beta}$, $\beta=1$	Centered forests	Data-independent
	$N^{-\frac{2\log(1-1/(2d))}{2\log(1-1/(2d))-\log(2)}}$	$\mathcal{H}^{0,\beta}$, $\beta=1$	Median forests	Depends on $\mathbf{X}_i$
[DS18]	$N^{-\frac{\log(1-3/(4d))}{\log(1-3/(4d))-\log(2)}}$	$\mathcal{H}^{0,1}$	Median forests	Depends on $\mathbf{X}_i$
[WW15]	$N^{-\frac{\log(1-3/(4q))}{\log(1-3/(4q))-\log(2)}}$	q-sparse $\mathcal{H}^{0,1}$	Median forests	Depends on $\mathbf{X}_i$
[SBV15]	Only $o_p(1)$	Additive model with continuous components	Breiman's original forests	Depends on both $\mathbf{X}_i$ and Y_i
[KT22]	$O_p(1/\log(N))$	Additive model with bounded total variation	Breiman's original forests	Depends on both $\mathbf{X}_i$ and Y_i
[CVFL22]	$N^{-\frac{c}{\alpha_1\alpha_2}}+N^{-\eta}$, $\alpha_2>1$, $c\in(0,1/4)$, $\eta\in(0,1/8)$	SID(α_1), $\alpha_1\geq 1$	Breiman's original forests	Depends on both $\mathbf{X}_i$ and Y_i
[FTAW20]	$N^{\delta-\left(1+\frac{d\log(\alpha)}{1.3\pi\log(1-\alpha)}\right)^{-1}}$, $\alpha\leq 0.2, \pi\leq 1/d$, $\delta>0$	$\mathcal{H}^{1,1}$	Local linear forests	Depends on both $\mathbf{X}_i$ and Y_i
Ours	$N^{\frac{-2\log(1-\alpha)}{d\log(\alpha)+2\log(1-\alpha)}}$, $\alpha\in(0,0.5]$	$\mathcal{H}^{0,1}$	Cyclic forests	Depends on $\mathbf{X}_i$, also depends on Y_i when $\alpha<0.5$
	$N^{\frac{-2(q+\beta)\log(1-\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}}$, $\alpha\in(0,0.5]$	$\mathcal{H}^{q,\beta}$, $q\in\mathbb{N}$, $\beta\in(0,1]$	Cyclic q-th order local polynomial forests	

3.1 Cyclic Forest

Consider the regression model

$$Y = m(\mathbf{X}) + \varepsilon, \tag{3.1}$$

where $m(\mathbf{x}) := \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ is the true conditional mean and $\varepsilon := Y - m(\mathbf{X})$ is the noise variable. We aim to estimate the function $m(\cdot)$ using i.i.d. samples $\mathbb{S}_N := (Y_i, \mathbf{X}_i)_{i=1}^N$.

The regression tree models the function $m(\cdot)$ by recursively partitioning the feature space $[0,1]^d$ into non-overlapping rectangles, generally called leaves. For any given point $\mathbf{x} \in [0,1]^d$, a regression tree estimates $m(\mathbf{x})$ using the average of responses for those samples in the same leaf as $\mathbf{x}$:

$$\mathrm{T}(\mathbf{x}, \xi) = \sum_{i \in \mathcal{I}} \frac{\mathbb{1}_{\{\mathbf{X}_i \in L(\mathbf{x},\xi)\}}}{\#\{l : \mathbf{X}_l \in L(\mathbf{x}, \xi)\}} Y_i, \tag{3.2}$$

where $\xi \in \Xi$ denotes all the auxiliary randomness in the tree-growing process and is independent of the samples, $\mathcal{I} \subseteq \{1, \ldots, N\}$ is the indices of training samples used for local averaging and possibly depends on ξ, $L(\mathbf{x}, \xi)$ represents the terminal leaf containing the point $\mathbf{x} \in [0,1]^d$, and $\#\{l : \mathbf{X}_l \in L(\mathbf{x}, \xi)\} = \sum_{l=1}^N \mathbb{1}_{\{\mathbf{X}_l \in L(\mathbf{x},\xi)\}}$ is the number of samples in this leaf.

To mitigate the impact from the auxiliary randomness, random forests consider ensembles of regression trees, where the forests' predictions are the average of all the tree predictions. Let $\{\mathrm{T}(\mathbf{x}, \xi_j), j = 1, \ldots, B\}$ denote the collection of regression trees in a forest, where B is the number of trees and $\xi_1, \ldots, \xi_B \in \Xi$ are i.i.d. auxiliary variables. For any $B \geq 1$, random forests estimate the conditional mean as

$$\widehat{m}(\mathbf{x}) := B^{-1} \sum_{j=1}^{B} \mathrm{T}(\mathbf{x}, \xi_j) = \mathbb{E}_\xi[\mathrm{T}(\mathbf{x}, \xi)],$$

where for any function $f(\cdot)$, $\mathbb{E}_\xi[f(\mathbf{x})] = B^{-1}\sum_{j=1}^{B} f(\xi_j)$ denotes the empirical average over the auxiliary variables, and we omit the dependence of such an expectation on B for the sake of notation simplicity.

Random forests can also be represented as a weighted average of the outcomes:

$$\widehat{m}(\mathbf{x}) = \mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi) Y_i\right], \text{ where } \omega_i(\mathbf{x},\xi) := \frac{\mathbb{1}_{\{\mathbf{X}_i \in L(\mathbf{x},\xi)\}}}{\#\{l : \mathbf{X}_l \in L(\mathbf{x},\xi)\}}. \tag{3.3}$$

To study the estimation behavior of random forests, we consider the following decomposition of the integrated mean squared error (IMSE):

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2 \le 2R_1 + 2R_2,$$

where $R_1 := \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\varepsilon_i\right]\right]^2$ is the estimation error originating from the random noise ε_i, and $R_2 := \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)(m(\mathbf{X}_i) - m(\mathbf{x}))\right]\right]^2$ can be viewed as the approximation error of the tree models. Let k be the minimum leaf size. Standard techniques lead to $R_1 = O_p(1/k)$ for the estimation error, as shown in (3.29) of the Supplement, and similar results can also be found in [Klu21,DS18,Bia12]. The control of the remaining approximation error is the key to reaching an optimal overall estimation error.

In this section, we restrict our attention to the class of Lipschitz continuous functions; see Assumption 3.1 below. The more general Hölder smooth functions will be further studied in Section 3.2.

Assumption 3.1 (Lipschitz continuous)**.** *Assume that $m(\cdot)$ satisfies $|m(\mathbf{x}) - m(\mathbf{x}')| \le L_0\|\mathbf{x} - \mathbf{x}'\|$ for all $\mathbf{x}, \mathbf{x}' \in [0,1]^d$ with some constant $L_0 > 0$.*

For any leaf $L \subseteq [0,1]^d$, denote $\mathrm{diam}(L) := \sup_{\mathbf{x},\mathbf{x}'\in L}\|\mathbf{x} - \mathbf{x}'\|$ as its diameter. Under

the Lipschitz condition, the approximation error can be controlled by the leaves' diameters:

$$R_2 \leq L_0^2 \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\xi} \left[\operatorname{diam}^2(L(\mathbf{x}, \xi)) \right] \right]. \tag{3.4}$$

Therefore, it suffices to obtain an upper bound for the diameters. When the splitting directions are chosen randomly, [Klu21] showed that both center and median forests lead to an upper bound $(k/N)^{2\log_2\left(\frac{2d}{2d-1}\right)}$ for the right-hand side of (3.4); a lower bound with the same rate has also been established for center forests. By choosing an optimal k that balances the estimation and approximation error, their methods lead to an overall IMSE with the rate $N^{\frac{-2\log(1-1/(2d))}{2\log(1-1/(2d))-\log(2)}}$ – this is *not* minimax optimal for Lipschitz functions. The sub-optimality stems from the excessive dependence of the forests on auxiliary randomness, rendering a substantial portion of the splits redundant and inefficient. This, in turn, leads to a relatively large approximation error. Indeed, as discussed in Section 2.1, choosing splitting directions randomly leads to a non-negligible fraction of terminal leaves becoming long and narrow. However, as leaves' diameters mainly depend on the longest side, these long and narrow leaves lead to large diameters and hence result in a relatively large approximation error for the forests method.

3.1.1 A cyclic approach

In order to reduce the large approximation error caused by auxiliary randomness, we propose a simple yet crucial modification to the existing methods. Instead of choosing splitting directions randomly, we adopt a more controlled and less random approach. Each time a leaf is split, we only randomly select a direction from one of the sides that has been split the least times. In other words, the splitting directions are chosen in a cyclic fashion –

Selected Direction	Probabilities of making splits on a direction											
	1	2	$\cdots$	i_1	$\cdots$	i_2	$\cdots$	i_{d-1}	$\cdots$	i_d	$\cdots$	d
i_1	$\frac{1}{d}$	$\frac{1}{d}$		$\frac{1}{d}$		$\frac{1}{d}$		$\frac{1}{d}$		$\frac{1}{d}$		$\frac{1}{d}$
i_2	$\frac{1}{d-1}$	$\frac{1}{d-1}$		0		$\frac{1}{d-1}$		$\frac{1}{d-1}$		$\frac{1}{d-1}$		$\frac{1}{d-1}$
	$\frac{1}{d-2}$	$\frac{1}{d-2}$		0		0		$\frac{1}{d-2}$		$\frac{1}{d-2}$		$\frac{1}{d-2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$
i_{d-1}	0	0		0		0		$\frac{1}{2}$		$\frac{1}{2}$		0
i_d	0	0		0		0		0		1		0
	0	0		0		0		0		0		0

Figure 3.1: Probabilities of making splits on each direction within a round.

we have to split once in each direction before proceeding to the next round of splitting; see illustrations in Figure 3.1. This approach helps reduce the impact of auxiliary randomness and enables more efficient splitting.

To further enhance the practical performance of forests, we permit data-dependent splitting rules contingent on both $\mathbf{X}_i$ and Y_i. This flexibility proves particularly valuable when the local smoothness level varies in different directions and locations. We consider a sample splitting procedure: for each tree, partition the samples into two sets, denoted by $\mathcal{I}$ and $\mathcal{J}$. The outcomes $(Y_i)_{i\in\mathcal{I}}$ are exclusively used for local averaging and are independent of the leaves. This structure is commonly referred to as "honest", as initially proposed by [AI16]. Additionally, we also impose constraints on the child node fraction and terminal leaf size, as observed in [MR06,WW15,WA18,FTAW20]. Specifically, with tuning parameters $\alpha \in (0, 0.5]$ and $k \in \mathbb{N}$, we require the following conditions to hold for the $\mathcal{I}$ sample: (a)

each child node contains at least an α-fraction of observations within the parent node, and (b) the number of observations within terminal leaves is between k and $2k-1$. The splitting locations are then determined to minimize the empirical mean squared error within each parent node, selecting from the set of points satisfying the above conditions. For more details on the proposed method, refer to Algorithm 4.

3.1.2 Theoretical results

For the sake of simplicity, we consider uniformly distributed $\mathbf{X}$ with support $[0,1]^d$. We first demonstrate the advantage of the cyclic splitting rule in the following lemma.

Lemma 3.1. *For any $r \geq 1$ and $\alpha \in (0, 0.5]$, the leaves' diameters of cyclic tree satisfy*

$$\sup_{\mathbf{x}\in[0,1]^d, \xi\in\Xi} \mathbb{E}_{\mathbb{S}_{\mathcal{I}}} \left[\mathrm{diam}^r(L(\mathbf{x},\xi))\right] < d^{r/2} \exp(r^2) \left(\frac{\lfloor wN \rfloor}{2k-1}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}}. \tag{3.5}$$

By Lemma 3.1, the proposed forests' approximation error (3.4) can be upper bounded by $O_p((k/N)^{\frac{2\log(1-\alpha)}{d\log(\alpha)}})$. When $\alpha = 0.5$, the algorithm degenerates into a cyclic version of median forests and results in an optimal rate $(k/N)^{\frac{2}{d}}$. For standard median forests where splitting directions are chosen randomly, Lemma 1 of [WA18] showed that at a given $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, $\mathbb{P}(\mathrm{diam}^2(L(\mathbf{x},\xi)) \geq C(k/N)^{\frac{1.98(1-\delta)}{d}}) = O((k/N)^{\frac{\delta^2}{2d\log 2}})$ for any $\delta > 0$ and some $C > 0$. Their result implies that for any given leaf, the diameter has a nearly optimal rate $\mathrm{diam}(L(\mathbf{x},\xi)) = O_p((k/N)^{\frac{1.98(1-\delta)}{d}})$. However, the corresponding tail probability is not small enough; in other words, there is a non-negligible probability that the leaf's diameter is large. As a result, integrating over both $\mathbf{x}$ and ξ (or taking the supremum as in our Lemma 3.1) leads to a slower rate. Indeed, as shown in [Klu21], the integrated square diameter (3.4) of both standard center and median forests are of the order $(k/N)^{2\log_2\left(\frac{2d}{2d-1}\right)}$ – this is strictly

Algorithm 4 Cyclic Forest

Require: Observations $\mathbb{S}_N = (\mathbf{X}_i, Y_i)_{i=1}^N$, with parameters $B \geq 1$, $\alpha \in (0, 0.5]$, $w \in (0, 1]$, and $k \leq \lfloor wN \rfloor$.

1: **for** $b = 1, \ldots, B$ **do**

2: Divide $\mathbb{S}_N$ into two disjoint sets $\mathbb{S}_{\mathcal{I}}^{(b)}$ and $\mathbb{S}_{\mathcal{J}}^{(b)}$, indexed by $\mathcal{I}^{(b)}$ and $\mathcal{J}^{(b)}$, with sizes $\#\mathcal{I}^{(b)} = \lfloor wN \rfloor$ and $\#\mathcal{J}^{(b)} = N - \lfloor wN \rfloor$.

3: **repeat**

4: For each current node $L \subseteq [0, 1]^d$, randomly select a direction j along which the node has been split the least number of times.

5: Partition the node along the j-th direction by minimizing the empirical mean squared error using samples $\mathbb{S}_{\mathcal{J}}^{(b)}$. That is, find the splitting point that minimizes

$$\sum_{i \in \mathcal{J}^{(b)}} (Y_i - \overline{Y}_1)^2 \mathbb{1}\{\mathbf{X}_i \in L_1\} + \sum_{i \in \mathcal{J}^{(b)}} (Y_i - \overline{Y}_2)^2 \mathbb{1}\{\mathbf{X}_i \in L_2\},$$

where L_1 and L_2 are the resulting child nodes, $\overline{Y}_1$ and $\overline{Y}_2$ are the average responses within the nodes L_1 and L_2, respectively. The splitting points are chosen such that $\#\{i \in \mathbb{S}_{\mathcal{I}}^{(b)} : \mathbf{X}_i \in L_j\} \geq \alpha \#\{i \in \mathbb{S}_{\mathcal{I}}^{(b)} : \mathbf{X}_i \in L\}$ for each $j \in \{1, 2\}$.

6: **until** The number of samples $\mathbb{S}_{\mathcal{I}}^{(b)}$ contained within each current node is between k and $2k - 1$.

7: The b-th cyclic tree estimates $m(\mathbf{x})$ using observations of $\mathbb{S}_{\mathcal{I}}^{(b)}$ within the terminal leaf containing $\mathbf{x}$ as in (3.2).

8: **end for**

9: **return** The cyclic forest is the average of B cyclic trees.

slower than our cyclic version as long as $k/N \to 0$ and $d > 1$. When $d = 1$, the rates are the same as there is no need to choose a splitting direction under such a degenerate situation. Clearly, choosing splitting directions randomly is sub-optimal as over-reliance on auxiliary randomness in the tree-growing process brings in a large approximation error.

To characterize the forests' overall IMSE, we further assume the following standard condition for the noise variable.

Assumption 3.2. *Assume that* $\mathbb{E}[\varepsilon^2 \mid \mathbf{X}] \leq M$ *almost surely with some constant* $M > 0$.

Theorem 3.1. *Let Assumptions 3.1 and 3.2 hold. Suppose that* $w \in (0, 1]$ *and* $\alpha \in (0, 0.5]$ *are both constants. Choose any* $B \in \mathbb{N}$ *and* $k \leq \lfloor wN \rfloor$. *Then, as* $N \to \infty$,

$$\mathbb{E}_{\mathbf{x}} \left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2 = O_p\left(\frac{1}{k} + \left(\frac{k}{N}\right)^{\frac{2\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.6}$$

Moreover, let $k \asymp N^{\frac{2\log(1-\alpha)}{d\log(\alpha)+2\log(1-\alpha)}}$, *we have*

$$\mathbb{E}_{\mathbf{x}} \left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2 = O_p\left(N^{-\frac{2\log(1-\alpha)}{d\log(\alpha)+2\log(1-\alpha)}}\right). \tag{3.7}$$

The results established in Theorem 3.1 are applicable for any $\alpha \in (0, 0.5]$. As long as $\alpha < 0.5$, the splitting locations are not restricted to the medians. Instead, we can leverage information from both $\mathbf{X}_i$ and Y_i to further enhance the empirical performance of the method. It is important to note that any regression tree can be regarded as a weighted average of response variables, using the weights defined in (3.3). Unlike standard kernel methods, the response variables affect the weights by influencing the leaves. This data-dependent learning aspect is shared with neural networks, which are also widely popular today. Neural networks learn representations in a data-dependent way, while random forests learn weights in a data-

dependent manner. This similarity might partially explain why both random forests and neural networks perform well in various complex applications.

Furthermore, our results provide additional insights into median forests. As indicated by (3.7), cyclic median forests (with $\alpha = 0.5$) achieve a minimax optimal rate of $N^{-\frac{2}{d+2}}$. The reason existing results [Klu21, DS18, Bia12] fall short of reaching the minimax optimal rate is mainly attributed to the inappropriate splitting rule considered in the current literature. We argue that random forests should not be excessively random, as an over-reliance on auxiliary randomness leads to poor estimation efficiency. It is noteworthy that the above results hold for any $B \geq 1$. With careful control of the impact of auxiliary randomness through the cyclic procedure, each individual cyclic median tree achieves minimax optimality for the Lipschitz class. Although averaging over multiple trees does not result in a faster convergence rate, we believe it is still worthwhile to do so to enhance finite-sample performance in practical applications.

3.2 Cyclic Local Polynomial Forest

In this section, we extend our focus to Hölder smooth functions and introduce cyclic forests capable of exploiting higher-order smoothness levels.

Assumption 3.3 (Hölder smooth)**.** *Assume that $m(\cdot) \in \mathcal{H}^{q,\beta}$ with some $q \in \mathbb{N}$ and $\beta \in (0, 1]$. The Hölder class $\mathcal{H}^{q,\beta}$ contains all functions $f : [0,1]^d \to \mathbb{R}$ that are q times continuously differentiable, with (a) $|D^{\boldsymbol{\alpha}} f(\mathbf{x})| \leq L_0$ for all $\mathbf{x} \in [0,1]^d$ and multi-index $\boldsymbol{\alpha}$ satisfying $|\boldsymbol{\alpha}| \leq q$, and (b) $|D^{\boldsymbol{\alpha}} f(\mathbf{x}) - D^{\boldsymbol{\alpha}} f(\mathbf{x}')| \leq L_0 \|\mathbf{x} - \mathbf{x}'\|^{\beta}$ for all $\mathbf{x}, \mathbf{x}' \in [0,1]^d$ and $\boldsymbol{\alpha}$ satisfying $|\boldsymbol{\alpha}| = q$, where $L_0 > 0$ is a constant.*

3.2.1 Cyclic local polynomial forest

To capture the higher-order smoothness of the conditional mean function $m(\cdot)$, we propose to fit a local polynomial regression within the leaves. We first introduce the polynomial basis with order $q \in \mathbb{N}$. For any $\mathbf{x} \in [0,1]^d$ and $j \in \{0, 1, \ldots, q\}$, let $\mathbf{g}_j(\mathbf{x}) := (\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathcal{A}_j} \in \mathbb{R}^{d^j}$, where $\mathcal{A}_j := \{\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d) \in \mathbb{N}^d : |\boldsymbol{\alpha}| = j\}$. For instance, $\mathbf{g}_0(\mathbf{x}) = 1$, $\mathbf{g}_1(\mathbf{x}) = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)^\top$, and $\mathbf{g}_2(\mathbf{x}) = (\mathbf{x}_1^2, \mathbf{x}_1\mathbf{x}_2, \ldots, \mathbf{x}_d^2)^\top$. Denote $\mathbf{G}(\mathbf{x}) := (\mathbf{g}_0(\mathbf{x}), \mathbf{g}_1(\mathbf{x})^\top, \ldots, \mathbf{g}_q(\mathbf{x})^\top)^\top \in \mathbb{R}^{\bar{d}}$ as the q-th order polynomial basis, where $\bar{d} := \sum_{j=0}^{q} d^j$.

For any $\xi \in \Xi$, define the weights $\omega_i(\mathbf{x}, \xi)$ as in (3.3), where we postpone the detailed tree-growing process for later. Using the training samples indexed by $\mathcal{I}$, consider the weighted polynomial regression:

$$\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi) := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{\bar{d}}} \sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi)(Y_i - \mathbf{G}(\mathbf{X}_i)^\top \boldsymbol{\beta})^2. \tag{3.8}$$

The cyclic q-th order local polynomial forest is proposed as

$$\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) := \mathbb{E}_\xi[\mathbf{G}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi)]. \tag{3.9}$$

Now, let us delve into the tree-growing process. The cyclic q-th order local polynomial forests are formulated as generalizations of the cyclic forests introduced in Algorithm 4. We incorporate a sample splitting mechanism to ensure the "honesty" of the forests and adopt a cyclic approach for selecting splitting directions, as elaborated in Section 3.1.1. In contrast to Algorithm 4, where local averages serve as tree predictions, our approach involves performing polynomial regressions within the terminal leaves. Consequently, our objective is to construct leaves that optimize the behavior of the final polynomial regressions.

For any current node $L \subseteq [0,1]^d$, the ideal approach is to find the optimal splitting

point that minimizes

$$\sum_{i\in\mathcal{J}^{(b)}}(Y_i - G(\mathbf{X}_i)^\top\widehat{\boldsymbol{\beta}}_1)^2\mathbb{1}\{\mathbf{X}_i\in L_1\} + \sum_{i\in\mathcal{J}^{(b)}}(Y_i - G(\mathbf{X}_i)^\top\widehat{\boldsymbol{\beta}}_2)^2\mathbb{1}\{\mathbf{X}_i\in L_2\}, \tag{3.10}$$

where L_1 and L_2 are the resulting child nodes, and $\widehat{\boldsymbol{\beta}}_j$ is the least squares estimate (using the polynomial basis) within the node L_j for each $j \in \{1,2\}$. However, this procedure requires calculating the least squares estimates for each candidate splitting point, making it computationally intractable. Drawing inspiration from [FTAW20], we minimize

$$\sum_{i\in\mathcal{J}^{(b)}}(\widehat{Y}_i - \widetilde{Y}_1)^2\mathbb{1}\{\mathbf{X}_i\in L_1\} + \sum_{i\in\mathcal{J}^{(b)}}(\widehat{Y}_i - \widetilde{Y}_1)^2\mathbb{1}\{\mathbf{X}_i\in L_2\}, \tag{3.11}$$

where $\widehat{Y}_i := Y_i - G(\mathbf{X}_i)^\top\widehat{\boldsymbol{\beta}}$ with $\widehat{\boldsymbol{\beta}}$ denoting the least squares estimate within the parent node L, and $\widetilde{Y}_j$ is the average of $\widehat{Y}_i$ within the node L_j for each $j \in \{1,2\}$. As discussed in Section 3.1.1, we also require that both child nodes contain at least an α-fraction of samples from the parent node. Additional specifics are outlined in Algorithm 5. It is noteworthy that Algorithm 4 is a special case of Algorithm 5 when $q = 0$.

To minimize (3.11), we only need to obtain the least squares estimate once, and the same $\widehat{\boldsymbol{\beta}}$ is used to calculate the error for each candidate splitting point within the node. Note that (3.11) can be viewed as an approximation of (3.10), where we substitute $\widehat{\boldsymbol{\beta}}_j$ with $\widetilde{\boldsymbol{\beta}}_j = (\widetilde{\boldsymbol{\beta}}_{1j}, \widehat{\boldsymbol{\beta}}_{-1}^\top)^\top$, and $\widetilde{\boldsymbol{\beta}}_{1j} = \arg\min_{\beta\in\mathbb{R}}\sum_{i\in\mathcal{J}^{(b)}}(Y_i^\top - \mathbf{G}(\mathbf{X}_i)^\top(\beta, \widehat{\boldsymbol{\beta}}_{-1}^\top)^\top)^2\mathbb{1}\{\mathbf{X}_i\in L_j\}$. In essence, we replace the slope coefficients in the child nodes with those in the parent node and find the least squares solution only for the intercept term.

3.2.2 Theoretical results

The following theorem characterizes the convergence rate of the proposed cyclic q-th order local polynomial forests under Hölder smooth conditions.

Algorithm 5 Cyclic q-th order Local Polynomial Forest

Require: Observations $\mathbb{S}_N = (\mathbf{X}_i, Y_i)_{i=1}^N$, with parameters $B \geq 1$, $\alpha \in (0, 0.5]$, $w \in (0, 1]$, $k \leq \lfloor wN \rfloor$, and $q \in \mathbb{N}$.

1: Calculate the polynomial basis $G(\mathbf{X}_i) \in \mathbb{R}^{\bar{d}}$ for each $i \leq N$.

2: **for** $b = 1, \ldots, B$ **do**

3: Divide $\mathbb{S}_N$ into two disjoint sets $\mathbb{S}_{\mathcal{I}}^{(b)}$ and $\mathbb{S}_{\mathcal{J}}^{(b)}$, indexed by $\mathcal{I}^{(b)}$ and $\mathcal{J}^{(b)}$, with sizes $\#\mathcal{I}^{(b)} = \lfloor wN \rfloor$ and $\#\mathcal{J}^{(b)} = N - \lfloor wN \rfloor$.

4: **repeat**

5: For each current node L, randomly select a direction j along which the node has been split the least number of times.

6: Partition the node along the j-th direction by minimizing (3.11) using samples $\mathbb{S}_{\mathcal{J}}^{(b)}$, ensuring that $\#\{i \in \mathbb{S}_{\mathcal{I}}^{(b)} : \mathbf{X}_i \in L_j\} \geq \alpha \#\{i \in \mathbb{S}_{\mathcal{I}}^{(b)} : \mathbf{X}_i \in L\}$ for each $j \in \{1, 2\}$.

7: **until** The number of samples $\mathbb{S}_{\mathcal{I}}^{(b)}$ contained within each current node is between k and $2k - 1$.

8: The b-th cyclic polynomial tree estimates $m(\mathbf{x})$ using observations of $\mathbb{S}_{\mathcal{I}}^{(b)}$ as $T_{\mathrm{CLPF}}(\mathbf{x}, \xi_b) := G(\mathbf{x})^\top \widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi_b)$, where $\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi_b)$ is defined as (3.8).

9: **end for**

10: **return** The cyclic q-th order local polynomial forest is the average of B cyclic polynomial trees.

Theorem 3.2. *Let Assumptions 3.2 and 3.3 hold. Suppose that $w \in (0, 1]$ and $\alpha \in (0, 0.5]$ are both constants. Choose any $B \in \mathbb{N}$ and $k \leq \lfloor wN \rfloor$ satisfying $k \gg \log(N)$. Then, as*

$N \to \infty$,

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})\right]^2 = O_p\left(\frac{1}{k} + \left(\frac{k}{N}\right)^{\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.12}$$

Moreover, let $k \asymp N^{\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}}$*, we have*

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})\right]^2 = O_p\left(N^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}}\right). \tag{3.13}$$

When $\alpha = 0.5$, (3.13) leads to the minimax optimal rate $N^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$ for the Hölder class $\mathcal{H}^{q,\beta}$. To the best of our knowledge, this is the first random forest method shown to achieve an IMSE reaching minimax optimality when $q > 1$; see Table 3.1.

In the following, we compare with existing results that have exploited Hölder smooth functions with $q \geq 1$. For $q = 1$, [OT21] proposed Tessellation forests and demonstrated that the corresponding IMSE reaches the minimax rate $N^{-\frac{2(1+\beta)}{d+2(1+\beta)}}$. For arbitrary $q \in \mathbb{N}$, [CMDY23] provided a nearly optimal rate of $(N/\log(N))^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$; however, they only obtained results for the *in-sample* excess risk, lacking theoretical guarantees for prediction performance on new observations. Additionally, [CKU23] obtained a minimax rate of $N^{-\frac{2(q+\beta)}{d+2(q+\beta)}}$ for point-wise mean squared error at *interior* points; however, their results do not lead to an optimal rate for the IMSE, as their debiased method is only valid for interior points.

It is worth mentioning that all the aforementioned works grow the trees completely independent of the samples. In contrast, we allow supervised splitting rules to further improve the forests' practical performance, as long as $\alpha < 0.5$ after appropriate tuning. Only [BTYW16, FTAW20] considered data-dependent splitting rules and studied local linear forests under the special case $q = 1$. However, [BTYW16] only demonstrated the consistency of their method, without providing any explicit rate of convergence. [FTAW20]

provided asymptotic normal results at a given $\mathbf{x} \in [0,1]^d$. However, their asymptotic variance is relatively large since the splitting directions are chosen randomly. In addition, their results rely on a technical condition that $\kappa_N := (1 - \mathbf{d}^\top \mathbf{S}^{-1}\mathbf{d})^{-1} = O(1)$, where $\mathbf{d} := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{G}_{-1}(\mathbf{X}_i - \mathbf{x})$, $\mathbf{S} := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{G}_{-1}(\mathbf{X}_i - \mathbf{x})\mathbf{G}_{-1}(\mathbf{X}_i - \mathbf{x})^\top$, and $\mathbf{G}(\mathbf{x}) = (1, \mathbf{G}_{-1}(\mathbf{x})^\top)^\top$ for any $\mathbf{x} \in [0,1]^d$. However, it is unclear when such a condition holds. Instead of forcing an upper bound for the random quantity κ_N by assumption, we prove that this quantity is bounded above with high probability; see Lemma 3.7.

3.3 Uniform results

In this section, we extend our analysis to encompass uniform-type results for the estimation error of the forests. As the cyclic forest introduced in Algorithm 4 constitutes a specific instance of the more general cyclic q-th order local polynomial forest outlined in Algorithm 5, we focus on presenting results for the latter.

To commence, we establish a uniform bound on the leaves' diameter as follows.

Lemma 3.2. *Suppose that $r \geq 1$, $w \in (0,1]$, and $\alpha \in (0, 0.5]$ are constants. Choose any $k \leq \lfloor wN \rfloor$ satisfying $k \gg \log^2(N)$. Then, as $N \to \infty$,*

$$\sup_{\mathbf{x}\in[0,1]^d, \xi\in\Xi} \{\mathrm{diam}^r(L(\mathbf{x},\xi))\} \leq C \left(\frac{N}{k}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}}, \tag{3.14}$$

with probability at least $1 - \log(\lfloor wN \rfloor/k)/(\sqrt{k}\log((1-\alpha)^{-1}))$ and some constant $C > 0$.

The requirement $k \gg \log^2(N)$ is slightly more stringent than the one assumed in Theorem 3.12, where we need $k \gg \log(N)$. Under this slightly stronger restriction on the minimum leaf size, we establish a uniform upper bound for the diameters, which holds with

high probability. It is important to note that we cannot directly apply Markov's inequality based on Lemma 3.1 to obtain uniform-type results, as the expectation $\mathbb{E}_{\mathbb{S}_N}(\cdot)$ is taken before the supremum, not after.

Subsequently, we present a uniform upper bound for the estimation error of the forests.

Theorem 3.3. *Let Assumption 3.3 hold. Suppose that $|Y| \leq M$. Let $M > 0$, $w \in (0, 1]$, and $\alpha \in (0, 0.5]$ be constants. Choose any $B \in \mathbb{N}$ and $k \leq \lfloor wN \rfloor$ satisfying $k \gg \log^2(N)$. Then, as $N \to \infty$,*

$$\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| = O_p\left(\sqrt{\frac{\log(N)}{k}} + \left(\frac{k}{N}\right)^{\frac{(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.15}$$

Moreover, let $k \asymp N^{\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}} (\log(N))^{\frac{d\log(\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}}$, we have

$$\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| = O_p\left(\left(\frac{\log(N)}{N}\right)^{\frac{(q+\beta)\log(1-\alpha)}{d\log(\alpha)+2(q+\beta)\log(1-\alpha)}}\right). \tag{3.16}$$

Comparing with the results in Theorem 3.2, the rates in (3.15)-(3.16) consist of additional logarithm terms. This is due to the cost of seeking uniform bounds. When $\alpha = 0.5$, an optimally tuned k leads to the rate $(\log(N)/N)^{\frac{q+\beta}{d+2(q+\beta)}}$, which is minimax optimal for sup-norms; see, e.g., [Sto82]. To the best of our knowledge, we are the first to establish minimax optimal uniform bounds for forests over the Hölder class $\mathcal{H}^{q,\beta}$ for any $q \in \mathbb{N}$.

3.4 Application to ATE estimation in causal inference

In this section, we apply the proposed forests to estimate the average treatment effect (ATE) in the context of causal inference. Let us consider i.i.d. samples $(\mathbf{W}_i)_{i=1}^N := (Y_i, \mathbf{X}_i, A_i)_{i=1}^N$, and denote $\mathbf{W} = (Y, \mathbf{X}, A)$ as its independent copy. Here, $Y \in \mathbb{R}$ denotes

the outcome of interest, $A \in \{0,1\}$ is a binary treatment variable, and $\mathbf{X} \in \mathbb{R}^d$ represents a vector of covariates uniformly distributed in $[0,1]^d$. We operate within the potential outcome framework and assume the existence of potential outcomes $Y(1)$ and $Y(0)$, where $Y(a)$ represents the outcome that would be observed if an individual receives treatment $a \in \{0,1\}$. The ATE is defined as $\theta := \mathbb{E}[Y(1) - Y(0)]$, representing the average effect of the treatment A on the outcome Y.

In order to identify causal effects, we make the following standard assumptions:

Assumption 3.4. *(a) Unconfoundedness:* $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid \mathbf{X}$. *(b) Consistency:* $Y = Y(A)$. *(c) Overlap:* $\mathbb{P}(c_0 < \pi^*(\mathbf{X}) < 1 - c_0) = 1$, *where* $c_0 \in (0, 1/2)$ *is a constant and the propensity score (PS) function is defined as* $\pi^*(\mathbf{x}) := \mathbb{P}(A = 1 \mid \mathbf{X} = \mathbf{x})$ *for any* $\mathbf{x} \in [0,1]^d$.

Define the true outcome regression function $\mu_a^*(\mathbf{x}) := \mathbb{E}[Y(a) \mid \mathbf{X} = \mathbf{x}]$ for $a \in \{0,1\}$ and consider the doubly robust score function: for any $\eta = (\mu_1, \mu_0, \pi)$,

$$\psi(\mathbf{W}; \eta) := \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \frac{A(Y - \mu_1(\mathbf{X}))}{\pi(\mathbf{X})} - \frac{(1-A)(Y - \mu_0(\mathbf{X}))}{1 - \pi(\mathbf{X})}. \tag{3.17}$$

As the ATE parameter can be represented as $\theta = \mathbb{E}[\psi(\mathbf{W}; \eta^*)]$, it can be estimated as the empirical average of the score functions as long as we plug in appropriate estimates of the nuisance functions $\eta^* = (\mu_1^*, \mu_0^*, \pi^*)$. In the following, we consider the double machine learning (DML) method of [CCD+17] and apply the proposed forests in Section 3.2 to estimate the nuisance functions.

For any fixed integer $K \geq 2$, split the samples into K equal-sized parts, indexed by $(\mathcal{I}_k)_{k=1}^K$. For the sake of simplicity, we assume $n := \#\mathcal{I}_k = N/K \in \mathbb{N}$. For each $k \leq K$, denote $\mathcal{I}_{-k} = \mathcal{I} \setminus \mathcal{I}_k$. Under Assumption 3.17, we can identify the outcome regression

function as $\mu_a^*(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, A = a)$ for each $a \in \{0, 1\}$. Hence, we construct $\widehat{\mu}_a^{-k}(\cdot)$ using Algorithm 5, based on samples $(Y_i, \mathbf{X}_i)_{i \in \{i \in \mathcal{I}_{-k} : A_i = a\}}$. Additionally, we also construct $\widehat{\pi}^{-k}(\cdot)$ using Algorithm 5, based on samples $(A_i, \mathbf{X}_i)_{i \in \mathcal{I}_{-k}}$. For the sake of simplicity, we denote $\mu_2(\cdot) := \pi(\cdot)$. The number of trees B and the orders of polynomial forests are chosen in advance, where we use q_j to denote the polynomial orders considered in the estimation of $\mu_j(\cdot)$ for each $j \in \{0, 1, 2\}$. Further denote $h_j := (\alpha_j, w_j, k_j)$ as the hyperparameters for estimating $\mu_j(\cdot)$. To appropriately select h_j, we further split the samples indexed by $\mathcal{I}_{-k}$ into training and validation sets. We train the forests based on the training samples and use the validation set to find the optimal tuning parameters that offer the smallest mean squared error. Note that the number of trees B is not a tuning parameter and is selected in advance – it essentially controls the computation error and should be large enough as long as the computing power allows. After obtaining the nuisance estimates $\widehat{\eta}^{-k} := (\widehat{\mu}_1^{-k}, \widehat{\mu}_0^{-k}, \widehat{\pi}^{-k})$ for each $k \leq K$, we define the ATE estimator as

$$\widehat{\theta} := N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \psi(\mathbf{W}_i; \widehat{\eta}^{-k}).$$

In the following, we first show that the proposed forests provide stable PS estimates that are away from zero and one with high probability.

Lemma 3.3. *Let Assumptions 3.4(c) hold and $\pi^* \in \mathcal{H}^{q_2, \beta_2}$, where $q_2 \in \mathbb{N}$ and $\beta_2 \in (0, 1]$. Let $M > 0$, $w_2 \in (0, 1]$, and $\alpha_2 \in (0, 0.5]$ be constants. Choose any $B \geq 1$ and $k_2 \gg \log^2(N)$. Then, as $N \to \infty$,*

$$\mathbb{P}_{\mathbf{X}}(c_1 < \widehat{\pi}^{-k}(\mathbf{X}) \leq 1 - c_1) = 1, \ \ \textit{for each} \ \ k \leq K, \tag{3.18}$$

with probability approaching one and some constant $c_1 \in (0, 1/2)$. Note that the left-hand-side of (3.18) *is a random quantity as the probability is only taken with respect to a new*

observation $\mathbf{X}$.

Lemma 3.3 demonstrates the stability of the inverse PS estimates, a requirement often assumed in the context of non-parametric nuisance estimates, as discussed in [CCD$^+$17]. The above results suggest that, under the assumption of overlap, there is typically no necessity to employ any form of trimming or truncation techniques on the estimated propensities, provided the chosen tuning parameters α and k are not too small.

Now, we are ready to introduce theoretical properties of the ATE estimator.

Theorem 3.4. *Let Assumption 3.4 hold,* $|Y| \leq M$, *and* $\mathbb{E}[\mathbb{1}_{\{A=a\}}(Y(a)-\mu_a^*)]^2 \geq C_0$ *for each* $a \in \{0,1\}$, *with some positive constants* M *and* C_0. *Suppose that* $\mu_0^* \in \mathcal{H}^{q_0,\beta_0}$, $\mu_1^* \in \mathcal{H}^{q_1,\beta_1}$, *and* $\pi^* \in \mathcal{H}^{q_2,\beta_2}$, *where* $q_j \in \mathbb{N}$ *and* $\beta_j \in (0,1]$ *for each* $j \in \{0,1,2\}$. *Let* $w_j \in (0,1]$ *and* $\alpha_j \in (0,0.5]$ *be constants. Choose any* $B \geq 1$ *and* $k_j \asymp N^{\frac{2(q_j+\beta_j)\log(1-\alpha_j)}{d\log(\alpha_j)+2(q_j+\beta_j)\log(1-\alpha_j)}}$. *Moreover, let* $d^2 \leq \frac{4(q_a+\beta_a)(q_2+\beta_2)\log(1-\alpha_a)\log(1-\alpha_2)}{\log(\alpha_a)\log(\alpha_2)}$ *for each* $a \in \{0,1\}$. *Then, as* $N \to \infty$, $\sigma^{-1}\sqrt{N}(\widehat{\theta}-\theta) \rightsquigarrow N(0,1)$ *and* $\widehat{\sigma}^{-1}\sqrt{N}(\widehat{\theta}-\theta) \rightsquigarrow N(0,1)$, *where* $\widehat{\sigma}^2 := N^{-1}\sum_{k=1}^{N}\sum_{i\in\mathcal{I}_k}[\psi(W_i;\hat{\eta}^{-k})-\widehat{\theta}]^2$.

Remark 3.1 (Technical challenges of forests-based ATE estimation)**.** *It is worth emphasizing that the following aspects are the main challenges in our analysis:*

(a) Establish convergence rates for the integrated mean squared error (IMSE) of the nuisance estimates. As the ATE is a parameter defined through integration over the entire population, we require convergence results for the IMSE; point-wise mean squared error results are insufficient. This distinguishes our work from [WA18], which focused on the estimation and inference for the conditional average treatment effect (CATE).

(b) Develop sufficiently fast convergence rates through higher-order smoothness. The asymptotic normality of the DML method [CCD$^+$17] requires a product-rate condition for

the nuisance estimation errors. If we only utilize the Lipschitz continuity of the nuisance functions, root-N inference is ensured only when $d = 1$. In other words, we need to establish methods that can exploit the higher-order smoothness of nuisance functions as long as $d > 1$. As shown in Theorem 3.4, the higher the smoothness levels are, the larger dimension d we allow for.

(c) Construct stable propensity score (PS) estimates. The construction of the DML ATE estimator involves the inversion of PS estimates. Using an early stopping technique that ensures a sufficiently large minimum leaf size, we guarantee that each terminal leaf contains a non-negligible fraction of samples in both treatment groups, as long as the overlap condition holds for the true PS function. Therefore, the early stopping technique stabilizes the PS estimates, and consequently, the ATE estimate.

3.5 Numerical Experiments

In this section, we assess the numerical performance of the proposed methods through simulation studies. We focus on the estimation of conditional mean function $m(x) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$. Generate i.i.d. covariates $\mathbf{X}_i \sim \text{Uniform}[0,1]^d$ and noise $\varepsilon_i \sim N(0,1)$ for each $i \leq N$. Consider the following outcome regression models:

(a) $Y_i = 10\sin(\pi \mathbf{X}_{i1}\mathbf{X}_{i2}) + 20(\mathbf{X}_{i3} - 5)^2 + 10\mathbf{X}_{i4} + 5\mathbf{X}_{i5} + \varepsilon_i$,

(b) $Y_i = 20\exp((\sum_{j=1}^{s} \mathbf{X}_{ij} - 0.5s)/\sqrt{s}) + \varepsilon_i$.

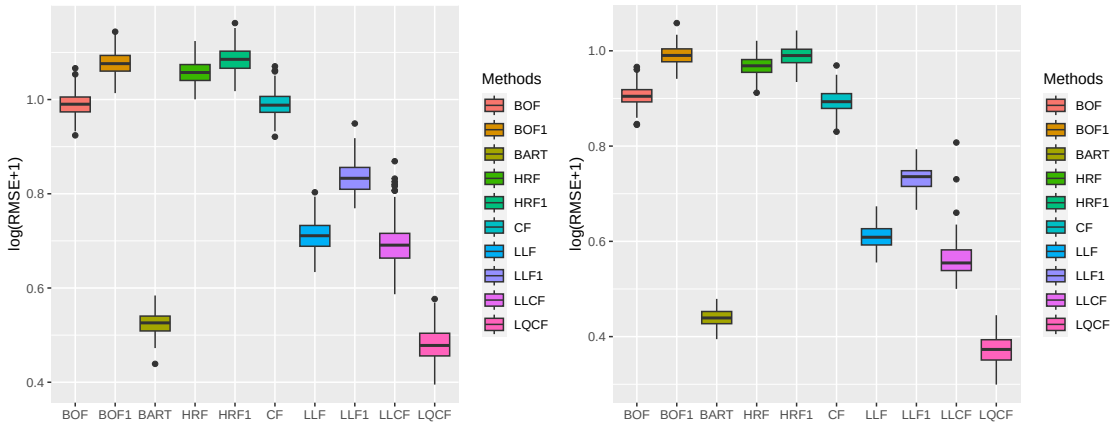
In setting (a), we employ the well-known Friedman function proposed by [Fri91], a commonly used benchmark for assessing non-parametric regression methods [ZL12, HZ21,

LH21]. We set the covariates' dimension to $d = 5$ and consider sample sizes $N \in \{500, 1000\}$. Additionally, we investigate the performance of the forests under various sparsity levels, maintaining $d = 10$, $N = 1000$, and choosing $s \in \{2, 6, 10\}$.

We implement the proposed cyclic forest (CF), local linear cyclic forest (LLCF), and local quadratic cyclic forest (LQCF), where LLCF and LQCF correspond to the cyclic q-th order local polynomial forests with $q = 1$ and $q = 2$, respectively. We choose $B = 200$ and use 80% of samples for training purposes, reserving the remaining 20% for validation to find the optimal tuning parameters (α, k). For the sake of simplicity, we fix $w = 0.5$.

We also consider Breiman's original forest (BOF), honest random forest (HRF), local linear forest (LLF), and Bayesian additive regression trees (BART). BOF is implemented using the R package `ranger` [WWPW19], HRF and LLF are implemented using the R package `grf` [TAF+23], and BART is implemented by the `BART` package [SSM19]. HRF and LLF methods involve an additional tuning parameter mtry, denoting the number of directions tried for each split. For comparison purposes, we also consider modified versions with fixed mtry = 1. This corresponds to the case where splitting directions are randomly chosen and is the only case that has been thoroughly studied theoretically [WA18, FTAW20]. We denote the modified versions of HRF and LLF as HRF1 and LLF1, respectively. The only difference between HRF1 and CF is that CF considers a cyclic splitting rule; a parallel difference exists between LLF1 and LLCF. Additionally, we introduce a modified version of BOF, denoted as BOF1.

We evaluate the considered methods' root mean square error (RMSE) within 1000 test points and repeat the procedure 200 times. Figures 3.2 and 3.3 depict boxplots comparing



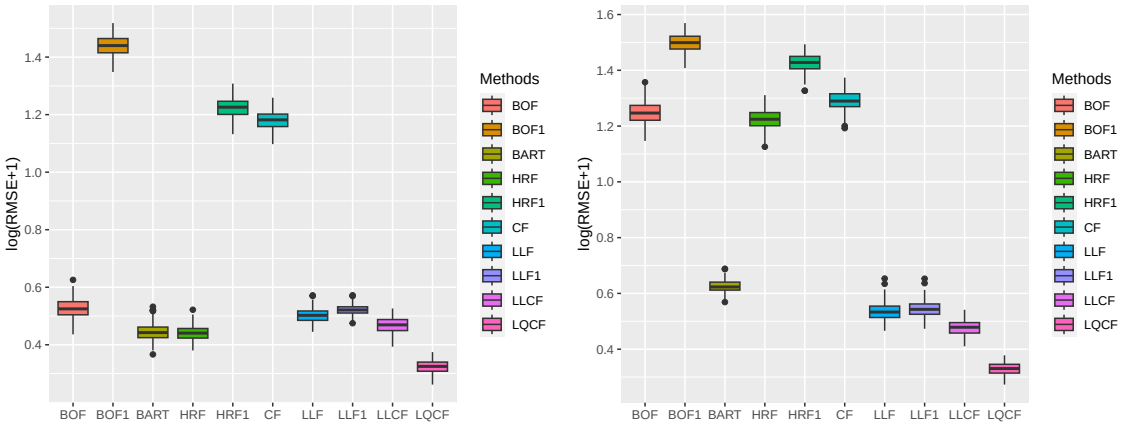
(a) Sample size $N = 500$

(b) Sample size $N = 1000$

Figure 3.2: Boxplots of $\log(\text{RMSE} + 1)$ under Setting (a) with a varying sample size.

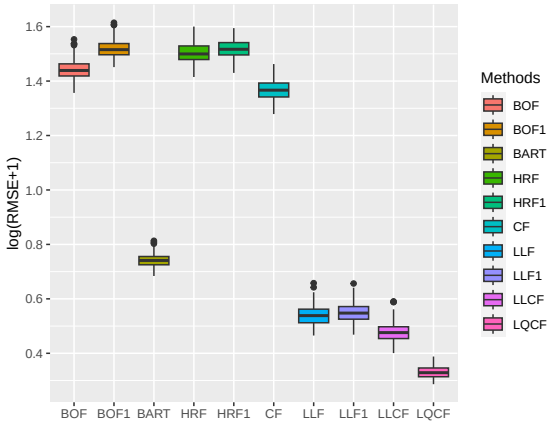
the log-transformed RMSE, $\log(\text{RMSE} + 1)$, of all the considered methods across various settings introduced above. From the boxplot figures, it is evident that LQCF always exhibits the best performance across all settings. Additionally, both our methods, LLCF and CF, consistently outperform LLF1 and HRF1, respectively, highlighting the distinct advantages offered by our cyclic method in contrast to random direction selection.

In dense scenarios, the proposed CF method demonstrates superior performance compared to all the existing local averaging methods, including BOF, BOF1, HRF, and HRF1; refer to Figure 3.2 and Figure 3.3(c). Additionally, LLCF also outperforms both the existing local linear methods LCF and LCF1. Under sparse scenarios, BOF and HRF demonstrate superior performance compared to BOF1, HRF1, and CF, as seen in Figure 3.3(a)-(b). This is because their splitting directions are chosen in a data-dependent fashion, which is more suitable when redundant covariates are included. However, we can still see that both the proposed LLCF and LQCF methods outperform all the existing ones in Figure 3.3(b). Even when the sparsity level is small in Figure 3.3(a), the LQCF remains to provide the smallest



(a) Sparsity level $s = 2$

(b) Sparsity level $s = 6$



(c) Sparsity level $s = 10$

Figure 3.3: Boxplots of $\log(\text{RMSE} + 1)$ under Setting (b) with a varying sparsity level.

RMSE.

3.6 Supplement

Notation We denote rectangles $L \in [0,1]^d$ by $R = \bigotimes_{j=1}^{d}[a_j, b_j]$, where $0 \le a_j < b_j \le 1$ for all $j = 1, \ldots, d$, writing the Lebesgue measure of L as $\lambda(L) = \prod_{j=1}^{d}(b_j - a_j)$. The indicator function of a subset A of a set X is a function $\mathbb{1}_A$ defined as $\mathbb{1}_A = 1$ if $x \in A$, and $\mathbb{1}_A = 0$ if

$x \notin A$. For any rectangle $L \in [0,1]^d$, we denote $\mu(L) := \mathbb{E}[\mathbb{1}_{\{\mathbf{X} \in L\}}]$ as the expected fraction of training examples falling within L. Denote $\#L := \sum_{i \in \mathcal{I}} \mathbb{1}_{\{\mathbf{X}_i \in L\}}$ as the number of training samples $\mathbf{X}_i$ falling within L. For any $n \times n$ matrix $\mathbf{A}$, let $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of the matrix $\mathbf{A}$, respectively. A d-dimensional vector of all ones is denoted with $\mathbf{1}_d$. A tree grown by recursive partitioning is called (α, k)-regular for some $\alpha \in (0, 0.5]$ and $k \in \mathbb{N}$ if the following conditions to hold for the $\mathcal{I}$ sample: (a) each child node contains at least an α-fraction of observations within the parent node, and (b) the number of observations within terminal leaves is between k and $2k-1$.

3.6.1 Proof of the results for Cyclic Forest

Proof of Lemma 3.1. For any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, let $c(\mathbf{x}, \xi)$ be the number of splits leading to the leaf $L(\mathbf{x}, \xi)$, and let $c_j(\mathbf{x}, \xi)$ be the number of such splits along the j-th coordinate. Let $t = \min_{1 \le j \le d} c_j(\mathbf{x}, \xi)$. By the cyclic splitting rule, we know that the number of splits along different coordinates differs by at most one. That is, $c_j(\mathbf{x}, \xi) \in \{t, t+1\}$ for all $1 \le j \le d$. Since $c(\mathbf{x}, \xi) = \sum_{j=1}^{d} c_j(\mathbf{x}, \xi)$, $c(\mathbf{x}, \xi)$ can be written as $c(\mathbf{x}, \xi) = td + l$, with some $0 \le l \le d-1$. Let $n_0 \ge n_1 \ge \cdots \ge n_{td+l}$ be the number of points in the successive nodes containing $\mathbf{x}$, where $n_0 = \lfloor wN \rfloor$. By (α, k)-regular, we know that $\alpha n_{i-1} \le n_i \le (1-\alpha) n_{i-1}$ for each $1 \le i \le td + l$. Hence, for any $1 \le i \le td + l$,

$$\alpha^i n_0 \le n_i \le (1-\alpha)^i n_0, \quad \text{and} \tag{3.19}$$

$$\alpha^{td+l-i} n_i \le n_{td+l} \le (1-\alpha)^{td+l-i} n_i. \tag{3.20}$$

For any $1 \le j \le d$ and closed set $L \subset [0,1]^d$, let $\mathrm{diam}_j(L)$ be the length of the longest segment parallel to the j-th axis that is a subset of L. For any $1 \le i \le t+1$ and $1 \le j \le d$, let $k_{i,j}$ be

the number of splits after the j-th coordinate has been split for i times, then by the cyclic splitting rule, we have $(i-1)d+1 \le k_{i,j} \le id$ for each $1 \le i \le t$. Let $L_{k_{i,j}}(\mathbf{x}, \xi)$ be the node containing $\mathbf{x}$ after $k_{i,j}$ splits, then the node $L_{k_{i,j}}(\mathbf{x}, \xi)$ contains $n_{k_{i,j}}$ samples of $\mathbb{S}_{\mathcal{I}}$. Denote $\bar{n}_j = (n_{k_{1,j}-1}, n_{k_{1,j}}, \ldots, n_{k_{t,j}-1}, n_{k_{t,j}})$. Then, for the i-th time splitting a node along the j-th coordinate, conditional on $\bar{n}_j$ and $L_{k_{i,j}-1}(\mathbf{x}, \xi)$, $\mathrm{diam}_j(L_{k_{i,j}}(\mathbf{x}, \xi))/\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x}, \xi))$ is at most the $(n_{k_{i,j}}+1)$-th order statistic of $n_{k_{i,j}-1}$ i.i.d. uniform random variables with support $[0,1]$. Note that for any i.i.d. uniform random variables $U_1, \ldots, U_{n_0} \sim \mathrm{Uniform}[0,1]$, the i-th order statistic follows a beta distribution $U_{(i)} \sim \mathrm{Beta}(i, n_0 - i + 1)$. Hence, we have

$$\frac{\mathrm{diam}_j(L_{k_{i,j}}(\mathbf{x}, \xi))}{\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x}, \xi))} \le B_{i,j},$$

with some $B_{i,j} \mid \bar{n}_j \sim \mathrm{Beta}(n_{k_{i,j}}+1, n_{k_{i,j}-1} - n_{k_{i,j}})$ and $(B_{i,j})_{i=1}^t$ are independent conditional on $\bar{n}_j$. Additionally, note that $\mathrm{diam}_j(L_{k_{1,j}-1}(\mathbf{x}, \xi)) = 1$. Therefore, for any $\mathbf{x} \in [0,1]^d$, $\xi \in \Xi$, and $1 \le j \le d$, we have $\mathrm{diam}_j(L(\mathbf{x}, \xi)) \le \mathrm{diam}_j(L_{k_{t,j}}(\mathbf{x}, \xi))$ and

$$\mathrm{diam}_j(L_{k_{t,j}}(\mathbf{x}, \xi)) \le \prod_{i=1}^{t} B_{i,j}.$$

By the conditional independency, we have

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}_j^r(L(\mathbf{x}, \xi)) \mid \bar{n}_j\right] \le \prod_{i=1}^{t} \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[B_{i,j}^r \mid \bar{n}_j\right].$$

For any $1 \le j \le d$ and $1 \le i \le t$, note that

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[B_{i,j}^r \mid \bar{n}_j\right] &= \frac{(n_{k_{i,j}}+1)\cdots(n_{k_{i,j}}+r)}{(n_{k_{i,j}-1}+1)\cdots(n_{k_{i,j}-1}+r)} \\
&\overset{(i)}{\le} (1-\alpha)^r \left(1 + \frac{\alpha/(1-\alpha)}{n_{k_{i,j}-1}+1}\right)\left(1 + \frac{2\alpha/(1-\alpha)}{n_{k_{i,j}-1}+2}\right)\cdots\left(1 + \frac{r\alpha/(1-\alpha)}{n_{k_{i,j}-1}+r}\right),
\end{aligned}$$

where (i) holds as $n_{k_{i,j}} \leq (1-\alpha)n_{k_{i,j}-1}$ by (α, k)-regular . Since $k_{i,j} \leq id$ for any $1 \leq j \leq d$ and $1 \leq i \leq t$, we have

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[B_{i,j}^r \mid \bar{n}_j\right] \leq (1-\alpha)^r \left(1 + \frac{\alpha/(1-\alpha)}{n_{id-1}+1}\right)\left(1 + \frac{2\alpha/(1-\alpha)}{n_{id-1}+2}\right)\cdots\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right).$$

It follows that, for any $\mathbf{x} \in [0,1]^d$, $\xi \in \Xi$ and $1 \leq j \leq d$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}_j^r(L(\mathbf{x},\xi)) \mid \bar{n}_j\right] \\
&\quad\leq (1-\alpha)^{tr} \prod_{i=1}^{t} \left(1 + \frac{\alpha/(1-\alpha)}{n_{id-1}+1}\right)\left(1 + \frac{2\alpha/(1-\alpha)}{n_{id-1}+2}\right)\cdots\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right) \\
&\quad\leq (1-\alpha)^{tr} \prod_{i=1}^{t} \left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right)^r. \qquad (3.21)
\end{aligned}$$

Since $\log(1+x) < x$ for all $x > 0$, we have

$$\log\left(\prod_{i=1}^{t}\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right)^r\right) = r\sum_{i=1}^{t}\log\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right) < r^2\sum_{i=1}^{t}\frac{\alpha/(1-\alpha)}{n_{id-1}+r}.$$

By $r \geq 1$ and (3.20), we have for each $i \leq t$,

$$\frac{1}{n_{id-1}+r} \leq \frac{1}{n_{id-1}} \leq (1-\alpha)^{l+1}\frac{(1-\alpha)^{(t-i)d}}{n_{td+l}},$$

which implies that

$$\log\left(\prod_{i=1}^{t}\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right)^r\right) < \frac{r^2\alpha(1-\alpha)^l}{n_{td+l}}\sum_{i=1}^{t}(1-\alpha)^{(t-i)d} = \frac{r^2\alpha(1-\alpha)^l}{n_{td+l}}\frac{1-(1-\alpha)^{td}}{1-(1-\alpha)^d}.$$

By $t, d > 0$ and $\alpha \in (0, 0.5]$, we have

$$\log\left(\prod_{i=1}^{t}\left(1 + \frac{r\alpha/(1-\alpha)}{n_{id-1}+r}\right)^r\right) < \frac{r^2(1-\alpha)^l}{n_{td+l}} \overset{(i)}{\leq} \frac{r^2(1-\alpha)^l}{k} \overset{(ii)}{\leq} r^2,$$

where (i) holds by (α, k)-regular; (ii) holds since $k \geq 1$, $\alpha \in (0, 0.5)$, and $l \geq 0$. Together with (3.21), for any $\mathbf{x} \in [0,1]^d$, $\xi \in \Xi$, and $1 \leq j \leq d$,

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}_j^r(L(\mathbf{x},\xi))\right] < (1-\alpha)^{tr}\exp(r^2). \qquad (3.22)$$

By (α, k)-regular and (3.19), we have $n_{td+l} \in [k, 2k-1]$ for some $k \in \mathbb{N}$ and $\alpha^{td+l}\lfloor wN \rfloor \leq n_{td+l} \leq (1-\alpha)^{td+l}\lfloor wN \rfloor$. Hence, we have $k \leq (1-\alpha)^{td+l}\lfloor wN \rfloor$ and $\alpha^{td+l}\lfloor wN \rfloor \leq 2k-1$, which implies that

$$\frac{\log((2k-1)/\lfloor wN \rfloor)}{d\log(\alpha)} - \frac{l}{d} \leq t \leq \frac{\log(k/\lfloor wN \rfloor)}{d\log(1-\alpha)} - \frac{l}{d}. \tag{3.23}$$

It follows that

$$(1-\alpha)^{tr} \leq (1-\alpha)^{\frac{lr}{d}} \left(\frac{\lfloor wN \rfloor}{2k-1}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}} \overset{(i)}{\leq} \left(\frac{\lfloor wN \rfloor}{2k-1}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}},$$

where (i) holds since $0 \leq l \leq d-1$. By (3.22), for any $\mathbf{x} \in [0,1]^d$, $\xi \in \Xi$, and $1 \leq j \leq d$,

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}_j^r(L(\mathbf{x},\xi))\right] < \left(\frac{\lfloor wN \rfloor}{2k-1}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}} \exp(r^2). \tag{3.24}$$

By the finite form of Jensen's inequality, we have for any $r \geq 1$,

$$\left(\frac{\sum_{j=1}^d 1 \cdot \mathrm{diam}_j^2(L(\mathbf{x},\xi))}{d}\right)^{r/2} \leq \frac{\sum_{j=1}^d 1 \cdot \mathrm{diam}_j^r(L(\mathbf{x},\xi))}{d},$$

which implies that

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^r(L(\mathbf{x},\xi))\right] = \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\sum_{j=1}^d \mathrm{diam}_j^2(L(\mathbf{x},\xi))\right]^{r/2} \leq d^{(r-2)/2}\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\sum_{j=1}^d \mathrm{diam}_j^r(L(\mathbf{x},\xi))\right].$$

By (3.24), for any $r \geq 1$, $\mathbf{x} \in [0,1]^d$, and $\xi \in \Xi$,

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^r(L(\mathbf{x},\xi))\right] < d^{r/2}\exp(r^2)\left(\frac{\lfloor wN \rfloor}{2k-1}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}}.$$

■

Proof of Theorem 3.1. By Jensen's inequality and the fact that $(a-b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2 &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)(Y_i - m(\mathbf{x}))\right]\right]^2 \\ &\leq \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)(Y_i - m(\mathbf{x}))\right]^2\right] \leq 2\mathbb{E}_{\mathbf{x}}\left[T_1(\mathbf{x})\right] + 2\mathbb{E}_{\mathbf{x}}\left[T_2(\mathbf{x})\right],\end{aligned} \tag{3.25}$$

where for any $\mathbf{x} \in [0,1]^d$,

$$T_1(\mathbf{x}) := \mathbb{E}_\xi \left[\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \varepsilon_i \right]^2, \tag{3.26}$$

$$T_2(\mathbf{x}) := \mathbb{E}_\xi \left[\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \left(m(\mathbf{X}_i) - m(\mathbf{x}) \right) \right]^2, \tag{3.27}$$

with $\varepsilon_i = Y_i - m(\mathbf{X}_i)$. By Fubini's theorem,

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\mathbb{E}_\mathbf{x} \left[T_1(\mathbf{x}) \right] \right] = \mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \varepsilon_i \right]^2 \right] \right].$$

Note that

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \varepsilon_i \right]^2 = \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \varepsilon_i^2 \right] + \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i,j \in \mathcal{I}, i \neq j} \omega_i(\mathbf{x}, \xi) \omega_j(\mathbf{x}, \xi) \varepsilon_i \varepsilon_j \right].$$

For any $i, j \in \mathcal{I}$ with $i \neq j$,

$$\begin{aligned} \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \omega_j(\mathbf{x}, \xi) \varepsilon_i \varepsilon_j \right] &\overset{(i)}{=} \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \omega_j(\mathbf{x}, \xi) \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\varepsilon_i \varepsilon_j \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l \in \mathcal{J}} \right] \right] \\ &\overset{(ii)}{=} \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \omega_j(\mathbf{x}, \xi) \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\varepsilon_i \mid \mathbf{X}_i \right] \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\varepsilon_j \mid \mathbf{X}_j \right] \right] \overset{(iii)}{=} 0, \end{aligned}$$

where (i) holds by the tower rule and "honesty" of the forests; (ii) holds by the independence of the samples; (iii) holds since $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$. Therefore, we have

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\mathbb{E}_\mathbf{x} \left[T_1(\mathbf{x}) \right] \right] = \mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \varepsilon_i^2 \right] \right] \right].$$

By the tower rule,

$$\begin{aligned} &\mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \varepsilon_i^2 \right] \right] \right] \\ &\quad \overset{(i)}{=} \mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\varepsilon_i^2 \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l \in \mathcal{J}} \right] \right] \right] \right] \\ &\quad \overset{(ii)}{=} \mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\varepsilon_i^2 \mid \mathbf{X}_i \right] \right] \right] \right] \\ &\quad \overset{(iii)}{\leq} M \mathbb{E}_\xi \left[\mathbb{E}_\mathbf{x} \left[\mathbb{E}_{\mathbb{S}_\mathcal{I}} \left[\sum_{i \in \mathcal{I}} \left[\omega_i(\mathbf{x}, \xi) \right]^2 \right] \right] \right], \end{aligned}$$

where (i) holds by "honesty" of the forests; (ii) holds by the independence of the samples; (iii) holds by Assumption 3.2. Therefore, we have

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_\mathbf{x}\left[T_1(\mathbf{x})\right]\right] \leq M\mathbb{E}_\xi\left[\mathbb{E}_\mathbf{x}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\left[\omega_i(\mathbf{x},\xi)\right]^2\right]\right]\right].$$

Since $\left(\mathbb{1}_{\{\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\right)^2 = \mathbb{1}_{\{\mathbf{X}_i\in L(\mathbf{x},\xi)\}}$, we have

$$\omega_i^2(\mathbf{x},\xi) = \frac{\omega_i(\mathbf{x},\xi)}{\#\left\{l : \mathbf{X}_l \in L(\mathbf{x},\xi)\right\}} \overset{(i)}{\leq} \frac{\omega_i(\mathbf{x},\xi)}{k}, \tag{3.28}$$

where (i) holds by (α,k)-regular. By $\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)=1$, we have $\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_\mathbf{x}\left[T_1(\mathbf{x})\right]\right] \leq M/k$. By Markov's inequality, as $N\to\infty$, we have

$$\mathbb{E}_\mathbf{x}\left[T_1(\mathbf{x})\right] = O_p\left(\frac{1}{k}\right). \tag{3.29}$$

Additionally, note that $\mathbb{E}_\mathbf{x}\left[T_2(\mathbf{x})\right] = \mathbb{E}_\mathbf{x}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)\right]^2\right]$. By Cauchy-Schwarz inequality and the fact that $\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)=1$,

$$\begin{aligned}\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)\right]^2 &\leq \left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\right]\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)^2\right]\\ &= \sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)^2.\end{aligned}$$

Then, we have

$$\mathbb{E}_\mathbf{x}\left[T_2(\mathbf{x})\right] \leq \mathbb{E}_\mathbf{x}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)^2\right]\right].$$

By the Lipschitz continuity of $m(\cdot)$, we have

$$\begin{aligned}\mathbb{E}_\mathbf{x}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(m(\mathbf{X}_i)-m(\mathbf{x})\right)^2\right]\right] &\leq \mathbb{E}_\mathbf{x}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\left(L_0\|\mathbf{X}_i-\mathbf{x}\|\right)^2\right]\right]\\ &\leq L_0^2\mathbb{E}_\mathbf{x}\left[\mathbb{E}_\xi\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right],\end{aligned}$$

where L_0 is the Lipschitz constant. Then, we have

$$\mathbb{E}_{\mathbf{x}}\left[T_2(\mathbf{x})\right] \leq L_0^2 \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right] \overset{(i)}{=} L_0^2 \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right],$$

where (i) holds by $\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi) = 1$. By Fubini's theorem,

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[T_2(\mathbf{x})\right]\right] \leq L_0^2 \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right]\right] = L_0^2 \mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right]\right].$$

By Lemma 3.1,

$$\begin{aligned}\mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^2(L(\mathbf{x},\xi))\right]\right]\right] &\leq \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^2(L(\mathbf{x},\xi))\right] \\ &< d\exp(4)\left(\frac{\lfloor wN\rfloor}{2k-1}\right)^{-\frac{2\log(1-\alpha)}{d\log(\alpha)}}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[T_2(\mathbf{x})\right]\right] < L_0^2 d\exp(4)\left(\frac{\lfloor wN\rfloor}{2k-1}\right)^{-\frac{2\log(1-\alpha)}{d\log(\alpha)}}.$$

By Markov's inequality, as $N\to\infty$, we have

$$\mathbb{E}_{\mathbf{x}}\left[T_2(\mathbf{x})\right] = O_p\left(\left(\frac{N}{k}\right)^{-\frac{2\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.30}$$

Combining (3.25), (3.29), and (3.30), we conclude that (3.6) holds.

■

3.6.2 Proofs of the results for Cyclic Polynomial Forest

Proof of Theorem 3.2. Recall the definition of $\widehat{m}_{\mathrm{CLPF}}(\mathbf{x})$, (3.9),

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})\right]^2 &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\mathbf{G}(\mathbf{x})^{\top}\left(\widehat{\boldsymbol{\beta}}(\mathbf{x},\xi) - \boldsymbol{\beta}\right)\right]\right]^2 \\ &\overset{(i)}{\leq} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\xi}\left[\mathbf{G}(\mathbf{x})^{\top}\left(\widehat{\boldsymbol{\beta}}(\mathbf{x},\xi) - \boldsymbol{\beta}\right)\right]^2\right] \overset{(ii)}{=} \mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{x}}\left[\mathbf{G}(\mathbf{x})^{\top}\left(\widehat{\boldsymbol{\beta}}(\mathbf{x},\xi) - \boldsymbol{\beta}\right)\right]^2\right],\end{aligned} \tag{3.31}$$

where (i) holds by Jensen's inequality and (ii) holds by Fubini's theorem. In the following, we condition on the event $\mathcal{B} \cap \mathcal{C}$ defined as (3.56) and (3.58). By Lemmas 3.7 and 3.8, we know that $\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top$, $\mathbf{S}_L$, $\mathbf{S}$, $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^\top$ and $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top$ are all positive-definite, with $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{B} \cap \mathcal{C}) = 1 - o(1)$. Recall the definition of $\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi)$, (3.8),

$$\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi) = \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top \right)^{-1} \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) Y_i \right).$$

Let $\boldsymbol{\alpha} := (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_d)$ be the multi-index, where each $\boldsymbol{\alpha}_i$ is a nonnegative integer. Define $r_i = Y_i - \mathbf{G}(\mathbf{X}_i)^\top \boldsymbol{\beta} - \varepsilon_i$ with $\mathbf{G}(\mathbf{X}_i)^\top \boldsymbol{\beta} = \sum_{|\boldsymbol{\alpha}|=0}^{q} D^{\boldsymbol{\alpha}} m(\mathbf{x}) (\mathbf{X} - \mathbf{x})^{\boldsymbol{\alpha}} / \boldsymbol{\alpha}!$. Then, we have

$$\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi) - \boldsymbol{\beta} = \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top \right)^{-1} \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) (\varepsilon_i + r_i) \right).$$

Note that there exists some $\bar{d} \times \bar{d}$ lower triangular matrix $\mathbf{T}$ with 1 on main diagonal such that

$$\mathbf{G}(\mathbf{X}_i - \mathbf{x}) = \mathbf{T} \mathbf{G}(\mathbf{X}_i), \tag{3.32}$$

which implies

$$\begin{aligned} &\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi) - \boldsymbol{\beta} \\ &\quad = \mathbf{T}^\top \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i - \mathbf{x}) \mathbf{G}(\mathbf{X}_i - \mathbf{x})^\top \right)^{-1} \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i - \mathbf{x}) (\varepsilon_i + r_i) \right). \end{aligned}$$

To simplify the exposition, we let $\boldsymbol{\Delta}_i := \mathbf{G}(\mathbf{X}_i - \mathbf{x})$. By $\mathbf{T}\mathbf{G}(\mathbf{x}) = \mathbf{G}(\mathbf{0}) = \mathbf{e}_1$,

$$\mathbf{G}(\mathbf{x})^\top \left(\widehat{\boldsymbol{\beta}}(\mathbf{x}, \xi) - \boldsymbol{\beta} \right) = \mathbf{e}_1^\top \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^\top \right)^{-1} \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i (\varepsilon_i + r_i) \right). \tag{3.33}$$

By (3.31), we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}} \left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x}) \right]^2 \\ &\quad \leq \mathbb{E}_\xi \left[\mathbb{E}_{\mathbf{x}} \left[\mathbf{e}_1^\top \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^\top \right)^{-1} \left(\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i (\varepsilon_i + r_i) \right) \right]^2 \right]. \end{aligned}$$

Define $\mathbf{U}_i := \left(\mathbf{g}_1(\mathbf{X}_i - \mathbf{x})^\top, \mathbf{g}_2(\mathbf{X}_i - \mathbf{x})^\top, \ldots, \mathbf{g}_q(\mathbf{X}_i - \mathbf{x})^\top\right)^\top = (Z_{i1}, \ldots, Z_{id}, Z_{i1}^2, Z_{i1}Z_{i2}, \ldots, Z_{id}^2, \ldots, Z_{id}^q)^\top \in \mathbb{R}^{\bar{d}}$ with $Z_{ij} := \mathbf{X}_{ij} - \mathbf{x}_j$ for any $i \in \mathcal{I}$ and $j \leq d$, and $\bar{d} = \sum_{i=1}^q d^i$. Since $\boldsymbol{\Delta}_i = (1, \mathbf{U}_i^\top)^\top$, we have

$$\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top = \begin{pmatrix} 1 & \mathbf{d}^\top \\ \mathbf{d} & \mathbf{S} \end{pmatrix},$$

where $\mathbf{d} := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i$ and $\mathbf{S} := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i\mathbf{U}_i^\top$. By Schur decomposition,

$$\mathbf{e}_1^\top \left(\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top\right)^{-1} = \begin{pmatrix} (1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1} & (1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1}\mathbf{d}^\top\mathbf{S}^{-1} \end{pmatrix}, \tag{3.34}$$

Since $\boldsymbol{\Delta}_i = (1, \mathbf{U}_i^\top)^\top$, we also have

$$\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\boldsymbol{\Delta}_i(\varepsilon_i + r_i) = \begin{pmatrix} \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)(\varepsilon_i + r_i) & \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i^\top(\varepsilon_i + r_i) \end{pmatrix}^\top. \tag{3.35}$$

It follows that

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq \mathbb{E}_\xi\Bigg[\mathbb{E}_{\mathbf{x}}\Bigg[&(1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1}\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)(\varepsilon_i + r_i) \\ &+ (1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1}\mathbf{d}^\top\mathbf{S}^{-1}\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i(\varepsilon_i + r_i)\Bigg]^2\Bigg]. \end{aligned} \tag{3.36}$$

Define $\mathbf{U}_i^L := \left(Z_{i1}^L, \ldots, Z_{id}^L, (Z_{i1}^L)^2, Z_{i1}^L Z_{i2}^L, \ldots, (Z_{id}^L)^2, \ldots, (Z_{i1}^L)^q\right)^\top \in \mathbb{R}^{\bar{d}}$ with $Z_{ij}^L := (\mathbf{X}_{ij} - \mathbf{x}_j)/\mathrm{diam}_j(L(\mathbf{x}, \xi))$ for any $i \in \mathcal{I}$ and $j \leq d$. Define a $\bar{d} \times \bar{d}$ diagonal matrix $\mathbf{D}_L := \mathrm{diag}\,(\mathrm{diam}_1(L(\mathbf{x}, \xi)), \ldots, \mathrm{diam}_d(L(\mathbf{x}, \xi)), \mathrm{diam}_1^2(L(\mathbf{x}, \xi)), \mathrm{diam}_1(L(\mathbf{x}, \xi))\mathrm{diam}_2(L(\mathbf{x}, \xi)), \ldots, \mathrm{diam}_1^2(L(\mathbf{x}, \xi)), \ldots, \mathrm{diam}_d^q(L(\mathbf{x}, \xi)))$. Then,

$$\mathbf{U}_i = \mathbf{D}_L\mathbf{U}_i^L, \;\; \mathbf{d} = \mathbf{D}_L\mathbf{d}_L, \;\; \text{and} \;\; \mathbf{S} = \mathbf{D}_L\mathbf{S}_L\mathbf{D}_L, \;\; \text{where} \tag{3.37}$$

$$\mathbf{d}_L := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i^L \;\; \text{and} \;\; \mathbf{S}_L := \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top. \tag{3.38}$$

Plugging (3.37) into (3.36), we have

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x})-m(\mathbf{x})\right]^2 \leq \mathbb{E}_{\xi}\Bigg[\mathbb{E}_{\mathbf{x}}\Bigg[(1-\mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\mathbf{d}_L)^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)(\varepsilon_i+r_i) + (1-\mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\mathbf{d}_L)^{-1}\mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\varepsilon_i+r_i)\Bigg]^2\Bigg]. \tag{3.39}$$

Let $\mathbf{c}_L := \mathbf{S}_L - \mathbf{d}_L\mathbf{d}_L^{\top}$. On the event $\mathcal{B}$, the matrix $\mathbf{c}_L$ is invertible. Since $\mathbf{d}_L\left(\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L+1\right) = \left(\mathbf{d}_L\mathbf{d}_L^{\top}+\mathbf{c}_L\right)\mathbf{c}_L^{-1}\mathbf{d}_L$ and $\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L \geq 0$, we have $\mathbf{d}_L = \left(\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L+1\right)^{-1}\left(\mathbf{d}_L\mathbf{d}_L^{\top}+\mathbf{c}_L\right)\mathbf{c}_L^{-1}\mathbf{d}_L$. It follows that

$$\begin{aligned}\mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\mathbf{d}_L &= \mathbf{d}_L^{\top}\left(\mathbf{d}_L\mathbf{d}_L^{\top}+\mathbf{c}_L\right)^{-1}\mathbf{d}_L \\ &= \mathbf{d}_L^{\top}\left(\mathbf{d}_L\mathbf{d}_L^{\top}+\mathbf{c}_L\right)^{-1}\left(\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L+1\right)^{-1}\left(\mathbf{d}_L\mathbf{d}_L^{\top}+\mathbf{c}_L\right)\mathbf{c}_L^{-1}\mathbf{d}_L \\ &= \frac{\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L}{\mathbf{d}_L^{\top}\mathbf{c}_L^{-1}\mathbf{d}_L+1} \leq 1.\end{aligned} \tag{3.40}$$

Then, we have $\left(1-\mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\mathbf{d}_L\right)^{-1} = 1+\mathbf{d}_L^{\top}\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^{\top}\right)^{-1}\mathbf{d}_L$. Therefore,

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x})-m(\mathbf{x})\right]^2 \leq \mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{4}\Delta_i(\mathbf{x},\xi)\right]^2 \sup_{\mathbf{x}\in[0,1]^d}\left\{1+\mathbf{d}_L^{\top}\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^{\top}\right)^{-1}\mathbf{d}_L\right\}\right],$$

where for any $\mathbf{x}\in[0,1]^d$ and $\xi\in\Xi$,

$$\Delta_1(\mathbf{x},\xi) := \mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L r_i, \quad \Delta_2(\mathbf{x},\xi) := \sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)r_i, \tag{3.41}$$

$$\Delta_3(\mathbf{x},\xi) := \mathbf{d}_L^{\top}\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L \varepsilon_i, \quad \Delta_4(\mathbf{x},\xi) := \sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\varepsilon_i. \tag{3.42}$$

By the finite form of Jensen's inequality, we have

$$\left[\frac{1}{4}\sum_{i=1}^{4}\Delta_i(\mathbf{x},\xi)\right]^2 \leq \frac{1}{4}\sum_{i=1}^{4}\left[\Delta_i(\mathbf{x},\xi)\right]^2,$$

which implies that

$$\mathbb{E}_{\mathbf{x}}\left[\widehat{m}_{\mathrm{CLPF}}(\mathbf{x})-m(\mathbf{x})\right]^2 \leq \mathbb{E}_{\xi}\left[4\sum_{i=1}^{4}\mathbb{E}_{\mathbf{x}}\left[\Delta_i(\mathbf{x},\xi)\right]^2 \sup_{\mathbf{x}\in[0,1]^d}\left\{1+\mathbf{d}_L^{\top}\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^{\top}\right)^{-1}\mathbf{d}_L\right\}\right]. \tag{3.43}$$

By Lemma 3.7, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi}\left\{1+\mathbf{d}_L^\top\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^\top\right)^{-1}\mathbf{d}_L\right\}=O_p(1). \tag{3.44}$$

Since $m\in\mathcal{H}^{q,\beta}$, by the Taylor's theorem, we have

$$m(\mathbf{X}_i)=P_{q-1}(\mathbf{X}_i)+R_{q-1}(\mathbf{X}_i),\ \text{ where }\ P_{q-1}(\mathbf{X}_i):=\sum_{|\boldsymbol{\alpha}|=0}^{q-1}\frac{D^{\boldsymbol{\alpha}}m(\mathbf{x})}{\boldsymbol{\alpha}!}(\mathbf{X}_i-\mathbf{x})^{\boldsymbol{\alpha}}$$

and $R_{q-1}(\mathbf{X}_i):=\sum_{|\boldsymbol{\alpha}|=q}D^{\boldsymbol{\alpha}}m(\boldsymbol{\xi})(\mathbf{X}_i-\mathbf{x})^{\boldsymbol{\alpha}}/\boldsymbol{\alpha}!$ for some $\boldsymbol{\xi}_i$ between $\mathbf{x}$ and $\mathbf{X}_i$. By definition, $r_i=m(\mathbf{X}_i)-\mathbf{G}(\mathbf{X}_i)^\top\boldsymbol{\beta}=R_{q-1}(\mathbf{X}_i)-(\mathbf{G}(\mathbf{X}_i)^\top\boldsymbol{\beta}-P_{q-1}(\mathbf{X}_i))=\sum_{|\boldsymbol{\alpha}|=q}(D^{\boldsymbol{\alpha}}m(\boldsymbol{\xi}_i)-D^{\boldsymbol{\alpha}}m(\mathbf{x}))(\mathbf{X}_i-\mathbf{x})^{\boldsymbol{\alpha}}/\boldsymbol{\alpha}!$ since $\mathbf{G}(\mathbf{X}_i)^\top\boldsymbol{\beta}=\sum_{|\boldsymbol{\alpha}|=0}^{q}D^{\boldsymbol{\alpha}}m(\mathbf{x})(\mathbf{X}-\mathbf{x})^{\boldsymbol{\alpha}}/\boldsymbol{\alpha}!$. By Assumption 3.3, we have

$$r_i\le\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\|\boldsymbol{\xi}_i-\mathbf{x}\|^\beta\|\mathbf{X}_i-\mathbf{x}\|^q\le\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}. \tag{3.45}$$

It follows that, for any $\xi\in\Xi$,

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_1(\mathbf{x},\xi)\right]^2\le\left[\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\right]^2\mathbb{E}_{\mathbf{x}}\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right]^2.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}&\left(\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right)^2\\&\quad\le\left(\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\mathbf{S}_L^{-1}\mathbf{d}_L\right)\left(\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\right)\\&\quad\overset{(i)}{=}\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{d}_L\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\overset{(ii)}{\le}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)},\end{aligned} \tag{3.46}$$

where (i) holds by the fact that $\mathbf{S}_L=\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top$; (ii) holds by (3.40). Then, we have

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_1(\mathbf{x},\xi)\right]^2\le\left[\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\right]^2\mathbb{E}_{\mathbf{x}}\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\right].$$

By construction, we have

$$\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\leq\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\overset{(i)}{=}\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi)),\quad(3.47)$$

where (i) holds by $\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)=1$. By Lemma 3.1, for any $\xi\in\Xi$,

$$\begin{aligned}\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\right]\right]&\leq\sup_{\mathbf{x}\in[0,1]^d}\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\right]\\&\leq d^{q+\beta}\exp\left(4(q+\beta)^2\right)\left(\frac{\lfloor wN\rfloor}{2k-1}\right)^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\end{aligned}$$

By Markov's inequality, as $N\to\infty$, we have

$$\mathbb{E}_{\mathbf{x}}\left[\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\right]=O_p\left(\left(\frac{N}{k}\right)^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right).\quad(3.48)$$

Therefore, for any $\xi\in\Xi$,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[\Delta_1(\mathbf{x},\xi)\right]^2&\leq\left[\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\right]^2\mathbb{E}_{\mathbf{x}}\left[\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\right]\\&=O_p\left(\left(\frac{N}{k}\right)^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right).\end{aligned}\quad(3.49)$$

In addition, by (3.45), for any $\xi\in\Xi$,

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_2(\mathbf{x},\xi)\right]^2\leq\left[\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\right]^2\mathbb{E}_{\mathbf{x}}\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right]^2.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}\left[\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right]^2&\leq\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\\&\overset{(i)}{=}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{2(q+\beta)}\overset{(ii)}{\leq}\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi)),\end{aligned}\quad(3.50)$$

where (i) holds by $\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)=1$; (ii) holds by (3.47). Therefore, we have

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_2(\mathbf{x},\xi)\right]^2\leq\left[\sum_{|\boldsymbol{\alpha}|=q}\frac{L_0}{\boldsymbol{\alpha}!}\right]^2\mathbb{E}_{\mathbf{x}}\left[\mathrm{diam}^{2(q+\beta)}(L(\mathbf{x},\xi))\right].$$

Together with (3.48), for any $\xi \in \Xi$, we have

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_2(\mathbf{x},\xi)\right]^2 = O_p\left(\left(\frac{N}{k}\right)^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.51}$$

As for the term $\Delta_3(\mathbf{x},\xi)$, for any $\xi \in \Xi$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] \\
&\quad = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top \mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\varepsilon_i^2\mathbf{S}_L^{-1}\mathbf{d}_L\right]\right] \\
&\qquad + \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\sum_{i,j\in\mathcal{I},i\neq j}\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\mathbf{U}_{L,j}^\top\mathbf{S}_L^{-1}\mathbf{d}_L\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x})\varepsilon_i\varepsilon_j\right]\right].
\end{aligned}$$

By the tower rule, for any $i,j \in \mathcal{I}$ with $i \neq j$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\mathbf{U}_{L,j}^\top\mathbf{S}_L^{-1}\mathbf{d}_L\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x})\varepsilon_i\varepsilon_j\right] \\
&\quad \overset{(i)}{=} \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\mathbf{U}_{L,j}^\top\mathbf{S}_L^{-1}\mathbf{d}_L\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x})\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\varepsilon_i\varepsilon_j \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}\right]\right] \\
&\quad \overset{(ii)}{=} \mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{U}_i^L\mathbf{U}_{L,j}^\top\mathbf{S}_L^{-1}\mathbf{d}_L\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x})\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\varepsilon_i \mid \mathbf{X}_i\right]\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\varepsilon_j \mid \mathbf{X}_j\right]\right] \overset{(iii)}{=} 0,
\end{aligned}$$

where (i) holds by "honesty" of the forests; (ii) holds by the independency of the samples; (iii) holds since $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$. Therefore, we have

$$\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\varepsilon_i^2\mathbf{S}_L^{-1}\mathbf{d}_L\right]\right].$$

By the tower rule, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] \\
&\quad \overset{(i)}{=} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\varepsilon_i^2\mathbf{S}_L^{-1}\mathbf{d}_L \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}\right]\right]\right] \\
&\quad \overset{(ii)}{=} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\mathbf{S}_L^{-1}\mathbf{d}_L\mathbb{E}[\varepsilon_i^2 \mid \mathbf{X}_i]\right]\right]\right] \\
&\quad \overset{(iii)}{\leq} M\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_{\mathcal{I}}}\left[\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top\mathbf{S}_L^{-1}\mathbf{d}_L\right]\right],
\end{aligned}$$

where (i) holds by "honesty" of the forests; (ii) holds by the independency of the samples; (iii) holds by Assumption 3.2. By (3.28), we have

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] \leq \frac{M}{k}\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbf{d}_L^\top \mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top \mathbf{S}_L^{-1}\mathbf{d}_L\right]\right]$$

Since $\mathbf{S}_L = \sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top$ and (3.40) holds, we have $\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top \mathbf{S}_L^{-1}\mathbf{d}_L = \mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{d}_L \leq 1$, and hence

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2\right] \leq \frac{M}{k}.$$

By Markov's inequality, for any $\xi \in \Xi$, we have

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_3(\mathbf{x},\xi)\right]^2 = O_p\left(\frac{1}{k}\right). \tag{3.52}$$

Lastly, for the term $\Delta_4(\mathbf{x},\xi)$, with any $\xi \in \Xi$,

$$\begin{aligned}\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_4(\mathbf{x},\xi)\right]^2\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\Delta_4(\mathbf{x},\xi)\right]^2\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\varepsilon_i^2\right]\right] + \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i,j\in\mathcal{I},i\neq j}\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x})\varepsilon_i\varepsilon_j\right]\right].\end{aligned}$$

Using the tower rule, we also have

$$\begin{aligned}\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x},\xi)\varepsilon_i\varepsilon_j\right] &\overset{(i)}{=} \mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x},\xi)\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\varepsilon_i\varepsilon_j \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}\right]\right] \\ &\overset{(ii)}{=} \mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\omega_i(\mathbf{x},\xi)\omega_j(\mathbf{x},\xi)\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\varepsilon_i \mid \mathbf{X}_i\right]\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\varepsilon_j \mid \mathbf{X}_j\right]\right] \overset{(iii)}{=} 0,\end{aligned}$$

where (i) holds by "honesty" of the forests; (ii) holds by the independency of the samples; (iii) holds since $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_4(\mathbf{x},\xi)\right]^2\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\varepsilon_i^2\right]\right] \\
&\overset{(i)}{=} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\varepsilon_i^2 \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}\right]\right]\right] \\
&\overset{(iii)}{=} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\varepsilon_i^2 \mid \mathbf{X}_i\right]\right]\right] \\
&\overset{(iv)}{\leq} M\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\sum_{i\in\mathcal{I}}\omega_i^2(\mathbf{x},\xi)\right]\right],
\end{aligned}$$

where (i) holds by the tower rule and "honesty" of the forests; (ii) holds by the independency of the samples; (iii) holds by Assumption 3.2. By (3.28) and $\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi) = 1$, we have

$$\mathbb{E}_{\mathbb{S}_\mathcal{I}}\left[\mathbb{E}_{\mathbf{x}}\left[\Delta_4(\mathbf{x},\xi)\right]^2\right] \leq \frac{M}{k}.$$

By Markov's inequality, for any $\xi \in \Xi$, we have

$$\mathbb{E}_{\mathbf{x}}\left[\Delta_4(\mathbf{x},\xi)\right]^2 = O_p\left(\frac{1}{k}\right). \tag{3.53}$$

Combining (3.49), (3.51), (3.52) and (3.53) with (3.44), we have

$$\begin{aligned}
&\mathbb{E}_\xi\left[4\sum_{i=1}^4 \mathbb{E}_{\mathbf{x}}\left[\Delta_i(\mathbf{x},\xi)\right]^2 \sup_{\mathbf{x}\in[0,1]^d}\left\{1 + \mathbf{d}_L^\top\left(\mathbf{S}_L - \mathbf{d}_L\mathbf{d}_L^\top\right)^{-1}\mathbf{d}_L\right\}\right] \\
&\quad = O_p\left(\frac{1}{k} + \left(\frac{N}{k}\right)^{-\frac{2(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right).
\end{aligned}$$

Together with (3.43), we conclude that (3.12) holds. ∎

3.6.3 Auxiliary Lemmas

Lemma 3.4 (Theorem 7 of [WW15]). *Let $\mathcal{D} = \{1, 2, \ldots, d\}$ and $\omega, \epsilon \in (0, 1)$. Then, there exists a set of rectangles $\mathcal{R}_{\mathcal{D},\omega,\epsilon}$ such that the following properties hold. Any rectangle L of*

volume $\lambda(L) \geq \omega$ can be well approximated by elements in $\mathcal{R}_{\mathcal{D},\omega,\epsilon}$ from both above and below in terms of Lebesgue measure. Specifically, there exist rectangles $R_-, R_+ \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ such that

$$R_- \subseteq L \subseteq R_+ \ \ and \ \ \exp\{-\epsilon\}\lambda(R_+) \leq \lambda(L) \leq \exp\{\epsilon\}\lambda(R_-).$$

Moreover, as $\omega, \epsilon \to 0$, the set $\mathcal{R}_{\mathcal{D},\omega,\epsilon}$ has cardinality bounded by

$$\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} = \frac{1}{\omega}\left(\frac{8d^2}{\epsilon^2}\left(1 + \log_2\left\lfloor\frac{1}{\omega}\right\rfloor\right)\right)^d \cdot (1 + O(\epsilon)).$$

Lemma 3.5 (Theorem 10 of [WW15])**.** *Suppose that $w \in (0, 1]$, and $\alpha \in (0, 0.5]$ are constants. Choose any $k \leq n = \lfloor wN \rfloor$ satisfying $k \gg \log(N)$. Let $\mathcal{L}$ be the collection of all possible leaves of partitions satisfying (α, k)-regular. Let $\mathcal{R}_{\mathcal{D},\omega,\epsilon}$ be as defined in Lemma 3.4, with ω and ϵ choosing as*

$$\omega = \frac{k}{2n} \ \ and \ \ \epsilon = \frac{1}{\sqrt{k}}. \tag{3.54}$$

Then, there exists an $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$, the following statement holds with probability at at least $1 - n^{-1/2}$: for each leaf $L \in \mathcal{L}$, we can select a rectangle $\bar{R} \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ such that $\bar{R} \subseteq L$, $\lambda(L) \leq \exp\{\epsilon\}\lambda(\bar{R})$, and

$$\#L - \#\bar{R} \leq 3\epsilon\#L + 2\sqrt{3\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})\#L} + O\left(\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})\right).$$

Lemma 3.6 (Lemma 12 of [WW15])**.** *Fix a sequence $\delta(n) > 0$, and define the event*

$$\mathcal{A} := \left\{\sup\left\{\frac{|\#R - n\mu(R)|}{\sqrt{n\mu(R)}} : R \in \mathcal{R}, \mu(R) \geq \mu_{\min}\right\} \leq \sqrt{3\log\left(\frac{\#\mathcal{R}}{\delta}\right)}\right\}$$

for any set of rectangles $\mathcal{R}$ and threshold $\mu_{\min}$, where $\#R := \#\{i : \mathbf{X}_i \in R\}$ and $\#\mathcal{R}$ is the number of rectangles of the set $\mathcal{R}$. Then, for any sequence of problems indexed by n with

$$\lim_{n\to\infty}\frac{\log(\#\mathcal{R})}{n\mu_{\min}} = 0 \ \ and \ \ \lim_{n\to\infty}\frac{\delta^{-1}}{\#\mathcal{R}} = 0, \tag{3.55}$$

there is a threshold $n_0 \in \mathbb{N}$ *such that, for all* $n \geq n_0$*, we have* $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$*. Note that, above,* $\mathcal{A}$*,* $\mathcal{R}$*,* $\mu_{\min}$ *and* δ *are all implicitly changing with* n*.*

Lemma 3.7. *Suppose that* $w \in (0, 1]$*, and* $\alpha \in (0, 0.5]$ *are constants. Choose any* $k \leq n = \lfloor wN \rfloor$ *satisfying* $k \gg \log(N)$*. Then, there exists a positive constant* $\Lambda_0 > 0$ *such that the event*

$$\mathcal{B} := \left\{ \inf_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \Lambda_{\min} \left(\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top \right) \geq \Lambda_0 \right\} \tag{3.56}$$

satisfies $\lim_{N \to \infty} \mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{B}) = 1$*, where* $\mathbf{d}_L$ *and* $\mathbf{S}_L$ *are defined as in* (3.38)*. In addition, on the event* $\mathcal{B}$*, the matrices* $\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top$ *and* $\mathbf{S}_L$ *are both positive-definite, and we also have*

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \mathbf{d}_L^\top (\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top)^{-1} \mathbf{d}_L \leq \frac{\bar{d}}{\Lambda_0}. \tag{3.57}$$

Lemma 3.8. *Let the assumptions in Lemma 3.7 hold. Define the event*

$$\mathcal{C} := \left\{ \mathrm{diam}_j(L(\mathbf{x}, \xi)) \neq 0, \ \text{for all } 1 \leq j \leq d, \ \mathbf{x} \in [0,1]^d, \ \xi \in \Xi \right\}. \tag{3.58}$$

Then, we have $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{C}) = 1$*. Moreover, on the event* $\mathcal{B} \cap \mathcal{C}$*, we have* $\mathbf{S}$*,* $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^\top$ *and* $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top$ *are both positive-definite, where* $\boldsymbol{\Delta}_i = \mathbf{G}(\mathbf{X}_i - \mathbf{x})$*.*

3.6.4 Proofs of the auxiliary Lemmas

Proof of Lemma 3.7. Choose ω and ϵ as in (3.54). By Lemma 3.4, there exists some $\widetilde{\mathcal{R}}_{\mathcal{D},\omega,\epsilon}$ satisfying the approximation property as in Lemma 3.4 with $\log(\#\widetilde{\mathcal{R}}_{\mathcal{D},\omega,\epsilon}) = O(\log(N))$. Therefore, we can choose some $\mathcal{R}_{\mathcal{D},\omega,\epsilon} \supseteq \widetilde{\mathcal{R}}_{\mathcal{D},\omega,\epsilon}$ satisfying $\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}) = O(\log(N))$ and $\sqrt{n} = o(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})$. Condition on the event $\mathcal{A}$ defined in Lemma 3.6, with $\mathcal{R} = \mathcal{R}_{\mathcal{D},\omega,\epsilon}$, $\mu_{min} = \omega$, and $\delta = 1/\sqrt{n}$. By $k \gg \log(N)$, as $N \to \infty$, we have

$$\frac{\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k} = O\left(\frac{\log(N)}{k} \right) = o(1) \ \text{ and } \ \frac{\sqrt{n}}{\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}} = o(1). \tag{3.59}$$

Thus, the condition (3.55) is satisfied. By Lemma 3.6, there exists $n_0 \in \mathbb{N}$ such that

$$\mathbb{P}_{\mathbb{S}_{\mathcal{I}}}(\mathcal{A}) \geq 1 - \frac{1}{\sqrt{n}} \text{ for any } n \geq n_0. \tag{3.60}$$

Condition on the event $\mathcal{A}$ above. For any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, under (α, k)-regular, and by Corollary 14 of [WW15], we have

$$\mu(L(\mathbf{x}, \xi)) \geq \omega. \tag{3.61}$$

By Lemmas 3.4 and 3.5, we can choose some $\bar{R} := \bar{R}(\mathbf{x}, \xi) \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ as an inner approximation of $L(\mathbf{x}, \xi)$ satisfying $\bar{R} \subseteq L(\mathbf{x}, \xi)$,

$$\lambda(L(\mathbf{x}, \xi)) \overset{(i)}{=} \mu(L(\mathbf{x}, \xi)) \leq \exp\{\epsilon\}\lambda(\bar{R}) \overset{(i)}{=} \exp\{\epsilon\}\mu(\bar{R}), \text{ and} \tag{3.62}$$

$$\frac{\#L - \#\bar{R}}{\#L} \leq \frac{3}{\sqrt{k}} + 2\sqrt{\frac{3\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{\#L}} + O\left(\frac{\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{\#L}\right), \tag{3.63}$$

where we denote $\#L := \#L(\mathbf{x}, \xi)$ for the sake of simplicity and (i) holds since $\mathbf{X}_i \sim$ Uniform$[0,1]^d$. Define $\omega_i^L := \omega_i(\mathbf{x}, \xi) = \mathbb{1}_{\{\mathbf{X}_i \in L(\mathbf{x},\xi)\}}/\#L$ and $\omega_i^{\bar{R}} := \mathbb{1}_{\left\{\mathbf{X}_i \in \bar{R}\right\}}/\#\bar{R}$. Note that

$$\mathbf{S}_L - \mathbf{d}_L\mathbf{d}_L^\top = \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i^L(\mathbf{U}_i^L)^\top - \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)\mathbf{U}_i^L \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x}, \xi)(\mathbf{U}_i^L)^\top = \sum_{i=1}^4 \mathbf{Q}_i,$$

where

$$\begin{aligned}
\mathbf{Q}_1 &:= \sum_{i\in\mathcal{I}} \omega_i^L \mathbf{U}_i^L(\mathbf{U}_i^L)^\top - \sum_{i\in\mathcal{I}} \omega_i^L \mathbf{U}_i^L \sum_{i\in\mathcal{I}} \omega_i^L(\mathbf{U}_i^L)^\top \\
&\quad - \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}} \mathbf{U}_i^L(\mathbf{U}_i^L)^\top + \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}} \mathbf{U}_i^L \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}}(\mathbf{U}_i^L)^\top, \\
\mathbf{Q}_2 &:= \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}} \mathbf{U}_i^L(\mathbf{U}_i^L)^\top - \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}} \mathbf{U}_i^L \sum_{i\in\mathcal{I}} \omega_i^{\bar{R}}(\mathbf{U}_i^L)^\top - \mathrm{Var}\left(\mathbf{U} \mid \mathbf{X} \in \bar{R}\right), \\
\mathbf{Q}_3 &:= \mathrm{Var}\left(\mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right) - \mathrm{Var}\left(\mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x}, \xi)\right), \\
\mathbf{Q}_4 &:= \mathrm{Var}\left(\mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x}, \xi)\right),
\end{aligned}$$

where $\mathbf{U}$ and $\mathbf{U}^L$ are independent copies of $\mathbf{U}_i$ and $\mathbf{U}^L$, respectively. By the triangle inequality,

$$\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi}\Lambda_{\min}\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^\top\right)=\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\sum_{i=1}^4\mathbf{a}^\top\mathbf{Q}_i\mathbf{a}$$
$$\geq\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\mathbf{a}^\top\mathbf{Q}_4\mathbf{a}-\sum_{i=1}^3\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_i\mathbf{a}\right|.$$

In the following, we show that there exists some constant $\Lambda_0>0$ such that

$$\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi}\Lambda_{\min}\left(\mathbf{S}_L-\mathbf{d}_L\mathbf{d}_L^\top\right)\geq\Lambda_0,$$

with probability approaching one as $N\to\infty$.

Step 1. We first demonstrate that on the event $\mathcal{A}$, as $N\to\infty$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_1\mathbf{a}\right|=o(1). \tag{3.64}$$

By the triangle inequality, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_1\mathbf{a}\right|\leq\sum_{j=1}^2\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_{1,j}\mathbf{a}\right|, \tag{3.65}$$

where for any $\mathbf{x}\in[0,1]^d$ and $\xi\in\Xi$,

$$\mathbf{Q}_{1,1}:=\mathbf{Q}_{1,1}(\mathbf{x},\xi)=\sum_{i\in\mathcal{I}}\omega_i^L\mathbf{U}_i^L(\mathbf{U}_i^L)^\top-\sum_{i\in\mathcal{I}}\omega_i^{\bar{R}}\mathbf{U}_i^L(\mathbf{U}_i^L)^\top,$$
$$\mathbf{Q}_{1,2}:=\mathbf{Q}_{1,1}(\mathbf{x},\xi)=\sum_{i\in\mathcal{I}}\omega_i^L\mathbf{U}_i^L\sum_{i\in\mathcal{I}}\omega_i^L(\mathbf{U}_i^L)^\top-\sum_{i\in\mathcal{I}}\omega_i^{\bar{R}}\mathbf{U}_i^L\sum_{i\in\mathcal{I}}\omega_i^{\bar{R}}(\mathbf{U}_i^L)^\top.$$

Note that $\bar{R}\subseteq L(\mathbf{x},\xi)$, for any $\mathbf{a}\in\mathbb{R}^{\bar{d}}$, we have

$$\left|\mathbf{a}^\top\mathbf{Q}_{1,1}\mathbf{a}\right|\leq\left|\frac{1}{\#L}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\left(\mathbf{a}^\top\mathbf{U}_i^L\right)^2-\frac{1}{\#\bar{R}}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\left(\mathbf{a}^\top\mathbf{U}_i^L\right)^2\right|$$
$$+\left|\frac{1}{\#L}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\setminus\bar{R}\right\}}\left(\mathbf{a}^\top\mathbf{U}_i^L\right)^2\right|$$
$$\leq\frac{2(\#L-\#\bar{R})}{\#L}\sup_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\left(\mathbf{a}^\top\mathbf{U}_i^L\right)^2.$$

For any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, if $\omega_i(\mathbf{x},\xi) \neq 0$, i.e., $\mathbf{X}_i \in L(\mathbf{x},\xi)$, we have $(\mathbf{X}_{ij} - \mathbf{x}_j)/\operatorname{diam}_j(L(\mathbf{x},\xi)) \in [-1,1]$. By the construction of $\mathbf{U}_i^L$,

$$\|\mathbf{U}_i^L\|_2 \leq \sqrt{\bar{d}}\|\mathbf{U}_i^L\|_\infty \leq \sqrt{\bar{d}}, \ \ \forall i \in \{i \in \mathcal{I} : \omega_i(\mathbf{x},\xi) \neq 0\}, \tag{3.66}$$

where $\bar{d} = \sum_{i=1}^q d^i$. Hence, it follows that

$$\sup_{i \in \{i \in \mathcal{I}: \mathbf{X}_i \in L(\mathbf{x},\xi)\}, \|\mathbf{a}\|_2 = 1} \left\{ \left(\mathbf{a}^\top \mathbf{U}_i^L\right)^2 \right\} = \sup_{i \in \{i \in \mathcal{I}: \mathbf{X}_i \in L(\mathbf{x},\xi)\}} \left\|\mathbf{U}_i^L\right\|_2^2 \leq \bar{d}. \tag{3.67}$$

Therefore,

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi, \|\mathbf{a}\|_2 = 1} \left|\mathbf{a}^\top \mathbf{Q}_{1,1} \mathbf{a}\right| \leq 2\bar{d} \sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \left\{ \frac{\#L - \#\bar{R}}{\#L} \right\}. \tag{3.68}$$

By (3.59) and (3.63), as $N \to \infty$, we have $(\#L - \#\bar{R})/\#L = o(1)$ for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, which implies that

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \left\{ \frac{\#L - \#\bar{R}}{\#L} \right\} = o(1). \tag{3.69}$$

By (3.68), we have

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi, \|\mathbf{a}\|_2 = 1} \left|\mathbf{a}^\top \mathbf{Q}_{1,1} \mathbf{a}\right| = o(1). \tag{3.70}$$

In addition, note that for any $\mathbf{a} \in \mathbb{R}^{\bar{d}}$,

$$\begin{aligned} \left|\mathbf{a}^\top \mathbf{Q}_{1,2} \mathbf{a}\right| = & \left| \frac{1}{\#L} \sum_{i \in \{i \in \mathcal{I}: \mathbf{X}_i \in L(\mathbf{x},\xi)\}} \mathbf{a}^\top \mathbf{U}_i^L + \frac{1}{\#\bar{R}} \sum_{i \in \left\{i \in \mathcal{I}: \mathbf{X}_i \in \bar{R}\right\}} \mathbf{a}^\top \mathbf{U}_i^L \right| \\ & \cdot \left| \frac{1}{\#L} \sum_{i \in \{i \in \mathcal{I}: \mathbf{X}_i \in L(\mathbf{x},\xi)\}} \mathbf{a}^\top \mathbf{U}_i^L - \frac{1}{\#\bar{R}} \sum_{i \in \left\{i \in \mathcal{I}: \mathbf{X}_i \in \bar{R}\right\}} \mathbf{a}^\top \mathbf{U}_i^L \right|. \end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}
&\left|\frac{1}{\#L}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\mathbf{a}^\top\mathbf{U}_i^L-\frac{1}{\#\bar{R}}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\mathbf{a}^\top\mathbf{U}_i^L\right| \\
&\quad\le\left|\frac{1}{\#L}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\mathbf{a}^\top\mathbf{U}_i^L-\frac{1}{\#\bar{R}}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\mathbf{a}^\top\mathbf{U}_i^L\right|+\left|\frac{1}{\#L}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\setminus\bar{R}\right\}}\mathbf{a}^\top\mathbf{U}_i^L\right| \\
&\quad\le\frac{2(\#L-\#\bar{R})}{\#L}\sup_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\left|\mathbf{a}^\top\mathbf{U}_i^L\right|.
\end{aligned}$$

Besides, we also have

$$\left|\frac{1}{\#L}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\mathbf{a}^\top\mathbf{U}_i^L+\frac{1}{\#\bar{R}}\sum_{i\in\left\{i\in\mathcal{I}:\mathbf{X}_i\in\bar{R}\right\}}\mathbf{a}^\top\mathbf{U}_i^L\right|\le 2\sup_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\left|\mathbf{a}^\top\mathbf{U}_i^L\right|.$$

Therefore,

$$\left|\mathbf{a}^\top\mathbf{Q}_{1,2}\mathbf{a}\right|\le\frac{4(\#L-\#\bar{R})}{\#L}\sup_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in L(\mathbf{x},\xi)\}}\left|\mathbf{a}^\top\mathbf{U}_i^L\right|^2.$$

By (3.67) and (3.69), we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_{1,2}\mathbf{a}\right|=o(1). \tag{3.71}$$

Combining (3.70) and (3.71) with (3.65), we conclude that (3.64) holds.

Step 2. We now demonstrate that on the event $\mathcal{A}$, as $N\to\infty$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_3\mathbf{a}\right|=o(1). \tag{3.72}$$

By the triangle inequality, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_3\mathbf{a}\right|\le\sum_{j=1}^{2}\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_{3,j}\mathbf{a}\right|, \tag{3.73}$$

where for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$,

$$\begin{aligned}
\mathbf{Q}_{3,1} := \mathbf{Q}_{3,1}(\mathbf{x},\xi) =& \mathbb{E}\left(\mathbf{U}^L(\mathbf{U}^L)^\top \mid \mathbf{X} \in \bar{R}\right) - \mathbb{E}\left(\mathbf{U}^L(\mathbf{U}^L)^\top \mid \mathbf{X} \in L(\mathbf{x},\xi)\right), \\
\mathbf{Q}_{3,2} := \mathbf{Q}_{3,1}(\mathbf{x},\xi) =& \mathbb{E}\left(\mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right) \mathbb{E}\left((\mathbf{U}^L)^\top \mid \mathbf{X} \in \bar{R}\right) - \\
& \mathbb{E}\left(\mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) \mathbb{E}\left((\mathbf{U}^L)^\top \mid \mathbf{X} \in L(\mathbf{x},\xi)\right).
\end{aligned}$$

Let $\mu_L = \mathbb{E}\left[\mathbb{1}_{\{\mathbf{X} \in L(\mathbf{x},\xi)\}}\right]$ and $\mu_{\bar{R}} = \mathbb{E}\left[\mathbb{1}_{\{\mathbf{X} \in \bar{R}\}}\right]$. By the triangle inequality,

$$\begin{aligned}
\left|\mathbf{a}^\top \mathbf{Q}_{3,1} \mathbf{a}\right| \le& \left|\frac{1}{\mu_L}\mathbb{E}\left[\left(\mathbf{a}^\top \mathbf{U}^L\right)^2 \mathbb{1}_{\{\mathbf{X} \in \bar{R}\}}\right] - \frac{1}{\mu_{\bar{R}}}\mathbb{E}\left[\left(\mathbf{a}^\top \mathbf{U}^L\right)^2 \mathbb{1}_{\{\mathbf{X} \in \bar{R}\}}\right]\right| \\
& + \left|\frac{1}{\mu_L}\mathbb{E}\left[\left(\mathbf{a}^\top \mathbf{U}^L\right)^2 \mathbb{1}_{\{\mathbf{X} \in L(\mathbf{x},\xi) \setminus \bar{R}\}}\right]\right| \\
\le& \frac{2(\mu_L - \mu_{\bar{R}})}{\mu_L} \sup_{\mathbf{X} \in [0,1]^d} \left(\mathbf{a}^\top \mathbf{U}^L\right)^2 \mathbb{1}_{\{\mathbf{X} \in L(\mathbf{x},\xi)\}}.
\end{aligned}$$

Similarly as in (3.67), we also have

$$\sup_{\mathbf{X} \in [0,1]^d, \|\mathbf{a}\|_2 = 1} \left(\mathbf{a}^\top \mathbf{U}^L\right)^2 \mathbb{1}_{\{\mathbf{X} \in L(\mathbf{x},\xi)\}} \le \bar{d}.$$

Hence,

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi, \|\mathbf{a}\|_2 = 1} \left|\mathbf{a}^\top \mathbf{Q}_{3,1} \mathbf{a}\right| \le 2\bar{d} \sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \left\{\frac{\mu_L - \mu_{\bar{R}}}{\mu_L}\right\}. \tag{3.74}$$

By (3.62) since $\mathbf{X}_i \sim \text{Uniform}[0,1]^d$, we have $\mu_{\bar{R}} \ge \exp\{-\epsilon\}\mu_L$ for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, which implies that

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \left\{\frac{\mu_L - \mu_{\bar{R}}}{\mu_L}\right\} \le 1 - \exp\{-\epsilon\} = 1 - \exp\{-1/\sqrt{k}\} = o(1) \tag{3.75}$$

as $N \to \infty$. Together with (3.74), we have

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi, \|\mathbf{a}\|_2 = 1} \left|\mathbf{a}^\top \mathbf{Q}_{3,1} \mathbf{a}\right| = o(1). \tag{3.76}$$

In addition, for any $\mathbf{a} \in \mathbb{R}^{\bar{d}}$,

$$\begin{aligned}\left|\mathbf{a}^\top \mathbf{Q}_{3,2}\mathbf{a}\right| = &\left|\mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) + \mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right)\right| \\ &\cdot \left|\mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) - \mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right)\right|.\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}&\left|\mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) - \mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right)\right| \\ &\quad \le \left|\frac{1}{\mu_L}\mathbb{E}\left[\mathbf{a}^\top \mathbf{U}^L \mathbb{1}_{\left\{\mathbf{X}\in\bar{R}\right\}}\right] - \frac{1}{\mu_{\bar{R}}}\mathbb{E}\left[\mathbf{a}^\top \mathbf{U}^L \mathbb{1}_{\left\{\mathbf{X}\in\bar{R}\right\}}\right]\right| \\ &\qquad + \left|\frac{1}{\mu_L}\mathbb{E}\left[\mathbf{a}^\top \mathbf{U}^L \mathbb{1}_{\left\{\mathbf{X}\in L(\mathbf{x},\xi)\setminus\bar{R}\right\}}\right]\right| \\ &\quad \le \frac{2(\mu_L - \mu_{\bar{R}})}{\mu_L} \sup_{\mathbf{X}\in[0,1]^d} \left|\mathbf{a}^\top \mathbf{U}^L\right| \mathbb{1}_{\{\mathbf{X}\in L(\mathbf{x},\xi)\}}.\end{aligned}$$

Besides, we also have

$$\left|\mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) + \mathbb{E}\left(\mathbf{a}^\top \mathbf{U}^L \mid \mathbf{X} \in \bar{R}\right)\right| \le 2 \sup_{\mathbf{X}\in[0,1]^d} \left|\mathbf{a}^\top \mathbf{U}^L\right| \mathbb{1}_{\{\mathbf{X}\in L(\mathbf{x},\xi)\}}.$$

Therefore,

$$\left|\mathbf{a}^\top \mathbf{Q}_{3,2}\mathbf{a}\right| \le \frac{4(\mu_L - \mu_{\bar{R}})}{\mu_L} \sup_{\{i:\mathbf{X}_i\in L(\mathbf{x},\xi)\}} \left|\mathbf{a}^\top \mathbf{U}_i^L\right|^2.$$

By (3.67) and (3.75), we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \left|\mathbf{a}^\top \mathbf{Q}_{3,2}\mathbf{a}\right| = o(1). \tag{3.77}$$

Combining (3.76) and (3.77) with (3.73), we conclude that (3.72) holds.

Step 3. We next demonstrate that condition on the event $\mathcal{A}$, as $N \to \infty$, with probability at least $1 - 2/\sqrt{n}$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \left|\mathbf{a}^\top \mathbf{Q}_2\mathbf{a}\right| = o(1). \tag{3.78}$$

For any $i \in \mathcal{I}$ and $j \leq d$, define $\mathbf{U}_i^{\bar{R}} := \left(Z_{i1}^{\bar{R}}, \ldots, Z_{id}^{\bar{R}}, (Z_{i1}^{\bar{R}})^2, Z_{i1}^{\bar{R}} Z_{i2}^{\bar{R}}, \ldots, (Z_{id}^{\bar{R}})^2, \ldots, (Z_{id}^{\bar{R}})^q\right)^\top$ with $Z_{ij}^{\bar{R}} := (\mathbf{X}_{ij} - \mathbf{x}_j)/\mathrm{diam}_j(\bar{R})$. Then, we have

$$\mathbf{U}_i^L = \mathbf{D}_{\bar{R}} \mathbf{U}_i^{\bar{R}}, \tag{3.79}$$

where $\mathbf{D}_{\bar{R}} := \mathrm{diag}\left(\frac{\mathrm{diam}_1(\bar{R})}{\mathrm{diam}_1(L(\mathbf{x},\xi))}, \ \ldots, \ \frac{\mathrm{diam}_d(\bar{R})}{\mathrm{diam}_d(L(\mathbf{x},\xi))}, \ \frac{\mathrm{diam}_1^2(\bar{R})}{\mathrm{diam}_1^2(L(\mathbf{x},\xi))}, \ \frac{\mathrm{diam}_1(\bar{R})\mathrm{diam}_2(\bar{R})}{\mathrm{diam}_1(L(\mathbf{x},\xi))\mathrm{diam}_2(L(\mathbf{x},\xi))}, \ \ldots, \ \frac{\mathrm{diam}_2^2(\bar{R})}{\mathrm{diam}_1^2(L(\mathbf{x},\xi))}, \ldots, \frac{\mathrm{diam}_d^q(\bar{R})}{\mathrm{diam}_d^q(L(\mathbf{x},\xi))} \right)$. Let $\bar{R}_j = [\bar{r}_j^-, \bar{r}_j^+] \subseteq [0,1]$ be the interval of the j-axis of the rectangle $\bar{R}$ for each $1 \leq j \leq d$. Define $\bar{V}_{ij}^{\bar{R}} := (\mathbf{X}_{ij} - \bar{r}_j^-)/\mathrm{diam}_j(\bar{R})$ and $\bar{c}_{\bar{R},j} = (\bar{r}_j^- - x_j)/\mathrm{diam}_j(\bar{R})$ for any $1 \leq j \leq d$. Then, the $\bar{d}$-dimensional vector $\mathbf{U}_i^{\bar{R}}$ can be represented as $\mathbf{U}_i^{\bar{R}} = (\bar{V}_{i1}^{\bar{R}} + \bar{c}_{\bar{R},1}, \ldots, \bar{V}_{id}^{\bar{R}} + \bar{c}_{\bar{R},d}, (\bar{V}_{i1}^{\bar{R}} + \bar{c}_{\bar{R},1})^2, (\bar{V}_{i1}^{\bar{R}} + \bar{c}_{\bar{R},1})(\bar{V}_{i2}^{\bar{R}} + \bar{c}_{\bar{R},2}), \ldots, (\bar{V}_{id}^{\bar{R}} + \bar{c}_{\bar{R},d})^2, \ldots, (\bar{V}_{id}^{\bar{R}} + \bar{c}_{\bar{R},d})^q)^\top$. Note that there exists some $\bar{d} \times \bar{d}$ lower triangular matrix $\mathbf{P}_{\bar{R}}$ with 1 on main diagonal such that

$$\mathbf{U}_i^{\bar{R}} = \mathbf{P}_{\bar{R}} \mathbf{V}_i^{\bar{R}} + \mathbf{C}_{\bar{R}}, \tag{3.80}$$

where $\mathbf{V}_i^{\bar{R}} := (V_{i1}^{\bar{R}}, \ldots, V_{id}^{\bar{R}}, (V_{i1}^{\bar{R}})^2, V_{i1}^{\bar{R}} V_{i2}^{\bar{R}}, \ldots, (V_{id}^{\bar{R}})^2, \ldots, (V_{id}^{\bar{R}})^q)^\top$ and $\mathbf{C}_{\bar{R}} := (c_{\bar{R},1}, \ldots, c_{\bar{R},d}, c_{\bar{R},1}^2, c_{\bar{R},1} c_{\bar{R},2}, \ldots, c_{\bar{R},d}^2, \ldots, c_{\bar{R},d}^q)^\top$. By (3.79) and (3.80), we have $\mathbf{U}_i^L = \mathbf{D}_{\bar{R}} \mathbf{P}_{\bar{R}} \mathbf{V}_i^{\bar{R}} + \mathbf{D}_{\bar{R}} \mathbf{C}_{\bar{R}}$. Let $\mathbf{V}^{\bar{R}}$ be an independent copy of $\mathbf{V}_i^{\bar{R}}$. Then, we can express $\mathbf{Q}_2$ as

$$\mathbf{Q}_2 = \mathbf{D}_{\bar{R}} \mathbf{P}_{\bar{R}} \mathbf{Q}_{\bar{R}} \mathbf{P}_{\bar{R}}^\top \mathbf{D}_{\bar{R}},$$

where $\mathbf{Q}_{\bar{R}} := \sum_{i \in \mathcal{I}} \omega_i^{\bar{R}} \mathbf{V}_i^{\bar{R}} (\mathbf{V}_i^{\bar{R}})^\top - \sum_{i \in \mathcal{I}} \omega_i^{\bar{R}} \mathbf{V}_i^{\bar{R}} \sum_{i \in \mathcal{I}} \omega_i^{\bar{R}} (\mathbf{V}_i^{\bar{R}})^\top - \mathrm{Var}\left(\mathbf{V}^{\bar{R}} \mid \mathbf{X} \in \bar{R}\right)$. By the sub-multiplicative property of matrix norm, we have

$$\|\mathbf{Q}_2\|_2 \leq \|\mathbf{D}_{\bar{R}}\|_2^2 \|\mathbf{P}_{\bar{R}}\|_2^2 \|\mathbf{Q}_{\bar{R}}\|_2.$$

Since $\mathbf{D}_{\bar{R}}$ is a diagonal matrix and its largest eigenvalue is smaller than 1, we have $\|\mathbf{D}_{\bar{R}}\|_2^2 \leq 1$. In addition, since the eigenvalues of an lower triangular matrix are the diagonal entries of

the matrix, we also have $\|\mathbf{P}_{\bar{R}}\|_2^2 = 1$. Therefore, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \left|\mathbf{a}^\top \mathbf{Q}_2 \mathbf{a}\right| = \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \|\mathbf{Q}_2\|_2 \leq \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \|\mathbf{Q}_{\bar{R}}\|_2.$$

For any $R \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$, let $R_j = [r_j^-, r_j^+] \subseteq [0,1]$ be the interval of the j-axis of the rectangle R for each $1 \leq j \leq d$. For any $i \in \mathcal{I}$, let $\omega_i^R := \mathbb{1}_{\{\mathbf{X}_i \in R\}}/\#R$ and $\mathbf{V}_i^R := (V_1^R, \ldots, V_{id}^R, (V_{i1}^R)^2, V_{i1}^R V_{i2}^R, \ldots, (V_{id}^R)^2, \ldots, (V_{id}^R)^q)^\top$ with $V_{ij}^R = (\mathbf{X}_{ij} - r_j^-)/\mathrm{diam}_j(R)$. Define $\mathbf{Q}_R := \sum_{i\in\mathcal{I}} \omega_i^R \mathbf{V}_i^R (\mathbf{V}_i^R)^\top - \sum_{i\in\mathcal{I}} \omega_i^R \mathbf{V}_i^R \sum_{i\in\mathcal{I}} \omega_i^R (\mathbf{V}_i^R)^\top - \mathrm{Var}\left(\mathbf{V}^R \mid \mathbf{X} \in R\right)$, where $\mathbf{V}^R$ is an independent copy of $\mathbf{V}_i^R$. Recall that on the event $\mathcal{A}$, (3.69) holds. Hence, there exists $n_1 \in \mathbb{N}$ such that $\#\bar{R} \geq k/2$ whenever $n \geq n_1$. In addition, note that $\bar{R} \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ for all $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$. Therefore, when $n \geq n_1$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \|\mathbf{Q}_{\bar{R}}\|_2 \leq \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_R\|_2 .$$

Let $\mathbf{m}_R := \mathbb{E}\left[\mathbf{V}^R \mid \mathbf{X} \in R\right]$. By the triangle inequality, we have

$$\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_R\|_2 \leq \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_{R,1}\|_2 + \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_{R,2}\|_2 ,$$

where

$$\mathbf{Q}_{R,1} := \sum_{i\in\mathcal{I}} \omega_i^R \left(\mathbf{V}_i^R - \mathbf{m}_R\right)\left(\mathbf{V}_i^R - \mathbf{m}_R\right)^\top - \mathrm{Var}\left(\mathbf{V}^R \mid \mathbf{X} \in R\right),$$
$$\mathbf{Q}_{R,2} := \left(\sum_{i\in\mathcal{I}} \omega_i^R \mathbf{V}_i^R - \mathbf{m}_R\right)\left(\sum_{i\in\mathcal{I}} \omega_i^R \mathbf{V}_i^R - \mathbf{m}_R\right)^\top .$$

Therefore, on the event $\mathcal{A}$ and provided that $n \geq n_1$, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \left|\mathbf{a}^\top \mathbf{Q}_2 \mathbf{a}\right| \leq \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_{R,1}\|_2 + \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} \|\mathbf{Q}_{R,2}\|_2 . \tag{3.81}$$

For all $i \in \{i \in \mathcal{I} : \mathbf{X}_i \in R\}$ and $j \leq d$, we have $\mathbf{V}_{ij}^R \in [0,1]$, where $\mathbf{V}_{ij}^R$ denotes the j-th coordinate of $\mathbf{V}_i^R$. Note that $(\mathbf{V}_i^R)_{i\in\mathcal{I}:\mathbf{X}_i\in R}$ are i.i.d. random vectors condition on the

indicators $\left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}$. As shown in Example 2.4 of [Wai19], condition on $\left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}$, $(\mathbf{V}_{ij}^R)_{i\in\mathcal{I}:\mathbf{X}_i\in R, j\le d}$ are sub-Gaussian with parameter at most $\sigma = 1$. Note that $\mathbb{E}[\mathbf{Q}_{R,1} \mid \left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}] = \mathbf{0}$. By Theorem 6.5 of [Wai19], for all $\zeta_1 \ge 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\|\mathbf{Q}_{R,1}\|_2 \ge C_1\left(\sqrt{\frac{\bar{d}}{\#R}} + \frac{\bar{d}}{\#R}\right) + \zeta_1 \mid \left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}, \mathcal{A}\right)\\
&\quad\le C_2 \exp\left\{-C_3 \#R \min\{\zeta_1, \zeta_1^2\}\right\},
\end{aligned}$$

where C_1, C_2, and C_3 are some positive constants. By the union bound, we have

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\|\mathbf{Q}_{R,1}\|_2 \ge C_1\left(\sqrt{\frac{2\bar{d}}{k}} + \frac{2\bar{d}}{k}\right) + \zeta_1 \mid \left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}, \mathcal{A}\right)\\
&\quad\le C_2 \#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \exp\left\{-\frac{C_3}{2} k \min\{\zeta_1, \zeta_1^2\}\right\}.
\end{aligned}$$

By the tower rule, we further obtain that

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\|\mathbf{Q}_{R,1}\|_2 \ge C_1\left(\sqrt{\frac{2\bar{d}}{k}} + \frac{2\bar{d}}{k}\right) + \zeta_1 \mid \mathcal{A}\right)\\
&\quad\le C_2 \#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \exp\left\{-\frac{C_3}{2} k \min\{\zeta_1, \zeta_1^2\}\right\}.
\end{aligned}$$

Let $\zeta_1 = \sqrt{\frac{4\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{C_3 k}}$. By (3.59), there exists $n_2 \in \mathbb{N}$ such that $\zeta_1^2 \le \zeta_1$ and $\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \ge \max\{C_2, \bar{d}/2\}\sqrt{n}$ whenever $n \ge n_2$. Thus, provided that $n \ge \max\{n_1, n_2\}$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\|\mathbf{Q}_{R,1}\|_2 \ge C_1\left(\sqrt{\frac{2\bar{d}}{k}} + \frac{2\bar{d}}{k}\right) + \sqrt{\frac{4\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{C_3 k}} \mid \mathcal{A}\right)\\
&\quad\le C_2/\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \le 1/\sqrt{n}. \qquad (3.82)
\end{aligned}$$

Additionally, note that

$$\begin{aligned}
\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\|\mathbf{Q}_{R,2}\|_2 &= \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\left\|\sum_{i\in\mathcal{I}}\omega_i^R\mathbf{V}_i^R - \mathbf{m}_R\right\|_2^2\\
&= \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon}, \#R\ge k/2}\sum_{j=1}^{\bar{d}}\left(\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}}\mathbf{V}_{ij}^R - \mathbf{m}_{R,j}\right)^2, \qquad (3.83)
\end{aligned}$$

where $\mathbf{m}_{R,j}$ is the j-th coordinate of $\mathbf{m}_R$. Since $\mathbf{V}_{ij}^R \in [0,1]$ for all $i \in \{i \in \mathcal{I} : \mathbf{X}_i \in R\}$ and $j \leq d$, by Theorem 2 of [Hoe94], for any $j \leq \bar{d}$ and $\zeta_2 > 0$,

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\left|\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}}\mathbf{V}_{ij}^R - \mathbf{m}_{R,j}\right| \geq \zeta_2 \mid \left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}, \mathcal{A}\right) \leq 2\exp\left\{-2\#R\zeta_2^2\right\}.$$

By the union bound, for all $\zeta_2 \geq 0$,

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2}\sum_{j=1}^{\bar{d}}\left(\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}}\mathbf{V}_{ij}^R - \mathbf{m}_{R,j}\right)^2 \geq \bar{d}\zeta_2^2 \mid \left\{\mathbb{1}_{\{\mathbf{X}_i\in R\}}\right\}_{i\in\mathcal{I}}, \mathcal{A}\right)$$
$$\leq 2\bar{d}\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}\exp\left\{-k\zeta_2^2\right\}.$$

Together with (3.83) and using the tower rule, for all $\zeta_2 \geq 0$,

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2}\|\mathbf{Q}_{R,2}\|_2 \geq \bar{d}\zeta_2^2 \mid \mathcal{A}\right) \leq 2\bar{d}\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}\exp\left\{-k\zeta_2^2\right\}.$$

Let $\zeta_2 = \sqrt{\frac{2\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k}}$. Then, we have

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2}\|\mathbf{Q}_{R,2}\|_2 \geq \frac{2\bar{d}\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k} \mid \mathcal{A}\right) \leq 2\bar{d}/\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \leq 1/\sqrt{n}, \tag{3.84}$$

since $\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \geq 2\bar{d}\sqrt{n}$ whenever $n \geq n_2$. Combining (3.82) and (3.84) with (3.81), provided that $n \geq \max\{n_1, n_2\}$, we have

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\Bigg(\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_2\mathbf{a}\right| \geq C_1\left(\sqrt{\frac{2\bar{d}}{k}} + \frac{2\bar{d}}{k}\right) + \sqrt{\frac{4\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{C_3 k}}$$
$$+ \frac{2\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k} \mid \mathcal{A}\Bigg) \leq 2/\sqrt{n}.$$

Therefore, condition on the event $\mathcal{A}$, as $N \to \infty$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1}\left|\mathbf{a}^\top\mathbf{Q}_2\mathbf{a}\right| = O\left(\sqrt{\frac{1}{k}} + \frac{1}{k} + \sqrt{\frac{\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k}} + \frac{\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{k}\right) = o(1),$$

with probability at least $1 - 2/\sqrt{n}$.

Step 4. We demonstrate that there exists some constant $\Lambda_0 > 0$ such that

$$\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbf{Q}_4 \mathbf{a} \geq 2\Lambda_0. \tag{3.85}$$

Let $L_j(\mathbf{x},\xi) = [a_j, b_j] \subseteq [0,1]$ be the interval of the j-axis of the leaf $L(\mathbf{x},\xi)$ for each $j \leq d$. Define $V_j^L = (\mathbf{X}_j - a_j)/\mathrm{diam}_j(L(\mathbf{x},\xi))$ and $c_{L,j} = (a_j - \mathbf{x}_j)/\mathrm{diam}_j(L(\mathbf{x},\xi))$ for any $j \leq d$. The $\bar{d}$-dimensional vector $\mathbf{U}^L$ can be represented as $\mathbf{U}^L = (V_1^L + c_{L,1}, \ldots, V_d^L + c_{L,d}, (V_1^L + c_{L,1})^2, (V_1^L + c_{L,1})(V_2^L + c_{L,2}), \ldots, (V_d^L + c_{L,d})^2, \ldots, (V_d^L + c_{L,d})^q)^\top$. Then, there exists some $\bar{d} \times \bar{d}$ lower triangular matrix $\mathbf{P}_L$ with 1 on main diagonal such that

$$\mathbf{U}^L = \mathbf{P}_L \mathbf{V}^L + \mathbf{C}_L, \tag{3.86}$$

where $\mathbf{V}^L := (V_{L,1}, \ldots, V_{L,d}, V_{L,1}^2, V_{L,1}V_{L,2}, \ldots, V_{L,d}^2, \ldots, V_{L,d}^q)^\top$ and $\mathbf{C}_L := (c_{L,1}, \ldots, c_{L,d}, c_{L,1}^2, c_{L,1}c_{L,2}, \ldots, c_{L,d}^2, \ldots, c_{L,d}^q)^\top$. Here, $\mathbb{P}_L$ and $\mathbf{C}_L$ are both deterministic given $L(\mathbf{x},\xi)$. Plugging $\mathbf{U}^L = \mathbf{P}_L \mathbf{V}^L + \mathbf{C}_L$ into $\mathbf{Q}_4$, we have

$$\mathbf{Q}_4 = \mathrm{Var}\left(\mathbf{P}_L \mathbf{V}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) = \mathbf{P}_L \mathrm{Var}\left(\mathbf{V}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) \mathbf{P}_L^\top.$$

By the sub-multiplicative property of matrix norm, we have

$$\|\mathbf{Q}_4^{-1}\|_2 \leq \|\mathbf{P}_L^{-1}\|_2^2 / \Lambda_{\min}(\mathrm{Var}\left(\mathbf{V}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right))$$

Since $\mathbf{P}_L$ is a lower triangular matrix with 1 on main diagonal, we know that $\mathbf{P}_L^{-1}$ is an upper triangular matrix with 1 on main diagonal, and it follows that $\left\|\mathbf{P}_L^{-1}\right\|_2 = 1$. Therefore,

$$\inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbf{Q}_4 \mathbf{a} \qquad \geq \inf_{\mathbf{x}\in[0,1]^d,\xi\in\Xi,\|\mathbf{a}\|=1} \mathrm{Var}\left(\mathbf{a}^\top \mathbf{V}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right).$$

Since the coordinates of $\mathbf{X}$ are i.i.d. uniformly distributed, we know that $(V_j^L)_{j=1}^d$ are also i.i.d. uniformly distributed given $\mathbf{X} \in L(\mathbf{x},\xi)$. Let $(\widetilde{V}_j)_{j=1}^d$ be a sequence of i.i.d. uniform random variables with support $[0,1]$, and denote $\widetilde{\mathbf{V}} := (\widetilde{V}_1, \ldots, \widetilde{V}_d, \widetilde{V}_1^2, \widetilde{V}_1\widetilde{V}_2, \ldots, \widetilde{V}_d^2, \ldots, \widetilde{V}_d^q)^\top$.

Then, for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, we have

$$\inf_{\|\mathbf{a}\|=1} \mathrm{Var}\left(\mathbf{a}^\top \mathbf{V}^L \mid \mathbf{X} \in L(\mathbf{x},\xi)\right) = \inf_{\|\mathbf{a}\|_2=1} \mathrm{Var}\left(\mathbf{a}^\top \widetilde{\mathbf{V}}\right).$$

Let $\Lambda_0 = \inf_{\|\mathbf{a}\|_2=1} \mathrm{Var}\left(\mathbf{a}^\top \widetilde{\mathbf{V}}\right)/2$. Note that the quantity Λ_0 is deterministic given the dimension $\bar{d}$ and hence is independent of the sample size N. Suppose that $\mathrm{Var}\left(\mathbf{a}^\top \widetilde{\mathbf{V}}\right) = 0$ with some $\mathbf{a} \neq \mathbf{0}$. Then, we have $\mathbb{P}(\mathbf{a}^\top \widetilde{\mathbf{V}} = a_0) = 1$ with some constant $a_0 \in \mathbb{R}$. However, note that $\mathbf{a}^\top \widetilde{\mathbf{V}} - a_0$ is a q-th polynomial function of $(\widetilde{V}_1, \widetilde{V}_2, \ldots, \widetilde{V}_d)$. As shown in Section 2.6.5 of [Fed14], $\mathbb{P}(\mathbf{a}^\top \widetilde{\mathbf{V}} = a_0) = 1$ occurs only if $a_0 = 0$ and $\mathbf{a} = \mathbf{0}$; this contradicts with $\mathbf{a} \neq \mathbf{0}$. Therefore, we conclude that $\Lambda_0 > 0$ and $\inf_{\mathbf{x}\in[0,1]^d, \xi\in\Xi, \|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbf{Q}_4 \mathbf{a} \geq 2\Lambda_0$.

Combining the results of Steps 1-4 and note that (3.60) holds, we conclude that

$$\lim_{N\to\infty} \mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{B}) = 1.$$

Lastly, condition on the event $\mathcal{B}$, (3.56). Then, $\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top$ and $\mathbf{S}_L$ are both positive-definite. In addition, we also have

$$\begin{aligned}
\sup_{\mathbf{x}\in[0,1]^d, \xi\in\Xi} \mathbf{d}_L^\top (\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top)^{-1} \mathbf{d}_L &\leq \frac{1}{\Lambda_0} \sup_{\mathbf{x}\in[0,1]^d, \xi\in\Xi} \left\| \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi) \mathbf{U}_i^L \right\|_2^2 \\
&\overset{(i)}{\leq} \frac{1}{\Lambda_0} \sup_{\mathbf{x}\in[0,1]^d, \xi\in\Xi} \left(\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi) \left\| \mathbf{U}_i^L \right\|_2 \right)^2 \overset{(ii)}{\leq} \frac{\bar{d}}{\Lambda_0},
\end{aligned}$$

where (i) holds by the triangle inequality; (ii) holds by (3.66) and $\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi) = 1$. ∎

Proof of Lemma 3.8. Condition on the event $\mathcal{B}$, (3.56). Then, the matrix $\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top$ is positive-definite, which implies that $\mathbf{S}_L$ is also positive-definite. Recall that $\mathbf{D}_L := \mathrm{diag}\,(\mathrm{diam}_1(L(\mathbf{x},\xi)), \ldots, \mathrm{diam}_d(L(\mathbf{x},\xi)), \mathrm{diam}_1^2(L(\mathbf{x},\xi)), \mathrm{diam}_1(L(\mathbf{x},\xi))\mathrm{diam}_2(L(\mathbf{x},\xi)), \ldots, \mathrm{diam}_1^2(L(\mathbf{x},\xi)), \ldots, \mathrm{diam}_d^q(L(\mathbf{x},\xi)))$. On the event $\mathcal{C}$, (3.58), the diagonal matrix $\mathbf{D}_L$ is invertible. By

(3.37), we have $\mathbf{S} = \mathbf{D}_L \mathbf{S}_L \mathbf{D}_L$ and $\mathbf{S} - \mathbf{d}\mathbf{d}^\top = \mathbf{D}_L(\mathbf{S}_L - \mathbf{d}_L\mathbf{d}_L^\top)\mathbf{D}_L$. Hence, on the event $\mathcal{B} \cap \mathcal{C}$, we have $\mathbf{S}$ and $\mathbf{S} - \mathbf{d}\mathbf{d}^\top$ are both positive-definite. In addition, note that

$$\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{\Delta}_i \mathbf{\Delta}_i^\top = \begin{pmatrix} 1 & \mathbf{d}^\top \\ \mathbf{d} & \mathbf{S} \end{pmatrix}.$$

For any $(a, \mathbf{b}) \in \mathbb{R}^{\bar{d}+1} \setminus \{\mathbf{0}\}$, we have

$$\begin{pmatrix} a & \mathbf{b}^\top \end{pmatrix} \begin{pmatrix} 1 & \mathbf{d}^\top \\ \mathbf{d} & \mathbf{S} \end{pmatrix} \begin{pmatrix} a \\ \mathbf{b} \end{pmatrix} = (a + \mathbf{d}^\top \mathbf{b})^2 + \mathbf{b}^\top (\mathbf{S} - \mathbf{d}\mathbf{d}^\top)\mathbf{b} > 0,$$

since $(a + \mathbf{d}^\top \mathbf{b})^2 = \mathbf{b}^\top (\mathbf{S} - \mathbf{d}\mathbf{d}^\top)\mathbf{b} = 0$ only when $\mathbf{b} = \mathbf{0}$ and $a = -d^\top \mathbf{b} = 0$. Hence, $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{\Delta}_i \mathbf{\Delta}_i^\top$ is positive-definite on the event $\mathcal{B} \cap \mathcal{C}$. Recall that the lower triangular matrix $\mathbf{T}$ is invertible. By (3.32), we have $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top = \mathbf{T}^{-1} \sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{\Delta}_i \mathbf{\Delta}_i^\top (\mathbf{T}^{-1})^\top$. Hence, $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \mathbf{G}(\mathbf{X}_i) \mathbf{G}(\mathbf{X}_i)^\top$ is also positive-definite on the event $\mathcal{B} \cap \mathcal{C}$.

In the following, we further show that $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{C}) = 1$. Let $\mathbf{X}_{ij}$ be the j-th coordinate of the vector $\mathbf{X}_i$ and $c \in [0, 1]$ be some constant. Then, we have

$$\begin{aligned}
\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{C}^c) &= \mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\exists\, j \le d,\ \mathbf{x} \in [0,1]^d,\ \xi \in \Xi,\ \text{s.t. } \operatorname{diam}_j(L(\mathbf{x}, \xi)) = 0\right) \\
&= \mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\exists\, j \le d,\ \mathbf{x} \in [0,1]^d,\ \xi \in \Xi,\ \text{s.t. } \mathbf{X}_{ij} = \mathbf{X}_{i'j}\ \forall i, i' \in \{i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi)\}\right) \\
&\overset{(i)}{\le} \mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\exists\, j \le d, i, i' \in \{i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi)\} \text{ s.t. } i \ne i' \text{ and } \mathbf{X}_{ij} = \mathbf{X}_{i'j}\right) \\
&\overset{(ii)}{\le} \sum_{1 \le j \le d, i, i' \in \mathcal{I}, i \ne i'} \mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathbf{X}_{ij} = \mathbf{X}_{i'j}) \overset{(iii)}{=} 0,
\end{aligned}$$

where (i) holds since the minimum leaf size $\#\{l : \mathbf{X}_l \in L(\mathbf{x}, \xi)\} \ge k \ge 2$; (ii) holds by the union bound; (iii) holds since $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathbf{X}_{ij} = \mathbf{X}_{i'j}) = 0$ as $\mathbf{X}_{ij}$ and $\mathbf{X}_{i'j}$ are independent uniform random variables for any $i \ne i'$ and $j \le d$. Therefore, we conclude that $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{C}) = 1$ holds.

■

3.6.5 Proofs of the uniform results

Proof of Lemma 3.2. In this proof, we use the same notation as Lemma 3.1. Note that for any $1 \leq j \leq d$, we have

$$\mathrm{diam}_j(L(\mathbf{x},\xi)) \leq \mathrm{diam}_j(L_{k_{t,j}}(\mathbf{x},\xi)) \overset{(i)}{=} \prod_{i=1}^{t} \frac{\mathrm{diam}_j(L_{k_{i,j}}(\mathbf{x},\xi))}{\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x},\xi))}, \tag{3.87}$$

where (i) holds by $\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x},\xi)) = \mathrm{diam}_j(L_{k_{i-1,j}}(\mathbf{x},\xi))$ for any $2 \leq i \leq t$ and $\mathrm{diam}_j(L_{k_{1,j}}(\mathbf{x},\xi)) = 1$. Choose $\omega_i = (1-\alpha)/2$, $\epsilon_i = 1/\sqrt{(1-\alpha)n_{k_{i,j}-1}}$ and $\delta_i = 1/\sqrt{n_{k_{i,j}-1}}$ thoughout this proof. Define the event

$$\mathcal{A}_{ij} := \left\{ \sup \left\{ \frac{|\#R - n_{k_{i,j}-1}\widetilde{\mu}(R)|}{\sqrt{n_{k_{i,j}-1}\widetilde{\mu}(R)}} : R \in \mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}, \widetilde{\mu}(R) \geq \omega_i \right\} \leq \sqrt{3\log\left(\frac{\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}}{\delta_i}\right)} \right\}, \tag{3.88}$$

for any $1 \leq i \leq t$ and $1 \leq j \leq d$, where $\#R := \#\{i : \mathbf{X}_i \in R\}$, $\widetilde{\mu}(R) := \mathbb{E}\Big[\mathbb{1}_{\{\mathbf{X}\in R\}} \mid \left\{\mathbb{1}_{\left\{\mathbf{X}_l \in L_{k_{i,j}-1}(\mathbf{x},\xi)\right\}}\right\}_{l\in\mathcal{I}}\Big]$ and $\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}$ is the number of rectangles of the set $\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}$. Note that as $n_{k_{i,j}-1} \to \infty$, we have

$$\frac{\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{n_{k_{i,j}-1}\omega_i} = O\left(\frac{\log(n_{k_{i,j}-1})}{n_{k_{i,j}-1}}\right) = o(1) \text{ and } \frac{\sqrt{n_{k_{i,j}-1}}}{\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}} = o(1).$$

By Lemma 3.6, for $n_{k_{i,j}-1}$ is large enough, we have

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\mathcal{A}_i \mid \left\{\mathbb{1}_{\left\{\mathbf{X}_l \in L_{k_{i,j}-1}(\mathbf{x},\xi)\right\}}\right\}_{l\in\mathcal{I}}\right) \geq 1 - \frac{1}{\sqrt{n_{k_{i,j}-1}}} \geq 1 - \frac{1}{\sqrt{k}}.$$

By the tower rule, we have $\mathbb{P}(\mathcal{A}_{ij}) \geq 1 - 1/\sqrt{k}$. By the union bound and (3.23), we have

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\cup_{j=1}^{d} \cup_{i=1}^{t} \mathcal{A}_{ij}^c\right) \leq \frac{td}{\sqrt{k}} \leq \frac{\log(\lfloor wN \rfloor/k)}{\sqrt{k}\log\left((1-\alpha)^{-1}\right)}.$$

Then, we have

$$\mathbb{P}_{\mathbb{S}_{\mathcal{I}}}\left(\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i\right)=1-\mathbb{P}_{\mathbb{S}_{\mathcal{I}}}\left(\cup_{j=1}^{d}\cup_{i=1}^{t}\mathcal{A}_{ij}^{c}\right)\geq 1-\frac{\log(\lfloor wN\rfloor)/k)}{\sqrt{k}\log\left((1-\alpha)^{-1}\right)}.$$

By $k\gg\log^2(N)$, we have

$$\mathbb{P}_{\mathbb{S}_{\mathcal{I}}}\left(\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i\right)=1-o(1). \tag{3.89}$$

Given the event $\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$, under (α,k)-regular, and by Corollary 14 of [WW15], we have $\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))\geq\omega_i$ for any $\mathbf{x}\in[0,1]^d$, $\xi\in\Xi$ and $1\leq i\leq t$. By Lemma 3.4, we can choose some $\widetilde{R}_i:=\widetilde{R}_i(\mathbf{x},\xi)\in\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}$ as an inner approximation of $L_{k_{i,j}}(\mathbf{x},\xi)$ satisfying $\widetilde{R}_i\subseteq L_{k_{i,j}}(\mathbf{x},\xi)$ with $\#\widetilde{R}_i\leq n_{k_{i,j}}$ and $\exp\{-\epsilon_i\}\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))\leq\widetilde{\mu}(\widetilde{R}_i)\leq\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))$, where $\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))=\mathrm{diam}_j(L_{k_{i,j}}(\mathbf{x},\xi))/\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x},\xi))$. Conditional on the event $\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$, we have for any $\mathbf{x}\in[0,1]^d$, $\xi\in\Xi$ and $1\leq i\leq t$,

$$\#\widetilde{R}_i\geq n_{k_{i,j}-1}\mu(\widetilde{R}_i)-\sqrt{3n_{k_{i,j}-1}\mu(\widetilde{R}_i)\log\left(\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}/\delta_i\right)},$$

which implies

$$n_{k_{i,j}}\geq\exp\{-\epsilon_i\}n_{k_{i,j}-1}\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))-\sqrt{3n_{k_{i,j}-1}\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))\log\left(\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}/\delta_i\right)}.$$

Hence, for any $\mathbf{x}\in[0,1]^d$, $\xi\in\Xi$ and $1\leq i\leq t$,

$$\begin{aligned}\sqrt{\frac{\mathrm{diam}_j(L_{k_{i,j}}(\mathbf{x},\xi))}{\mathrm{diam}_j(L_{k_{i,j}-1}(\mathbf{x},\xi))}}&=\sqrt{\widetilde{\mu}(L_{k_{i,j}}(\mathbf{x},\xi))}\leq 2\exp\{\epsilon_i\}\left(\sqrt{\frac{3}{n_{k_{i,j}-1}}\log\left(\frac{\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}}{\delta_i}\right)}\right.\\&\left.+\sqrt{\frac{3}{n_{k_{i,j}-1}}\log\left(\frac{\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}}{\delta_i}\right)+4\exp\{-\epsilon_i\}\frac{n_{k_{i,j}}}{n_{k_{i,j}-1}}}\right)\\&\leq 4\exp\{\epsilon_i\}\sqrt{\frac{3}{n_{k_{i,j}-1}}\log\left(\frac{\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}}{\delta_i}\right)+4\exp\{-\epsilon_i\}\frac{n_{k_{i,j}}}{n_{k_{i,j}-1}}}.\end{aligned}$$

By (3.87), conditional on the event $\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$, by (3.61), we have for any $\mathbf{x}\in[0,1]^d$ and $\xi\in\Xi$,

$$\mathrm{diam}_j(L(\mathbf{x},\xi))\leq 16\left\{\prod_{i=1}^{t}\exp\{2\epsilon_i\}\left[\frac{3}{n_{k_{i,j}-1}}\log\left(\frac{\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i}}{\delta_i}\right)+4\exp\{-\epsilon_i\}\frac{n_{k_{i,j}}}{n_{k_{i,j}-1}}\right]\right\}$$

By $\log(\#\mathcal{R}_{\mathcal{D},\omega_i,\epsilon_i})=O(\log(n_{k_{i,j}-1}))$, as $N\to\infty$, we have

$$\mathrm{diam}_j(L(\mathbf{x},\xi))=O\left(\prod_{i=1}^{t}\left[\frac{\log(n_{k_{i,j}-1})}{n_{k_{i,j}-1}}+\frac{n_{k_{i,j}}}{n_{k_{i,j}-1}}\right]\right)\overset{(i)}{=}O\left((1-\alpha)^t\left[1+\frac{\log(\lfloor wN\rfloor)}{(1-\alpha)k}\right]^t\right),$$

where (i) holds by $k\leq n_{k_{i,j}-1}\leq\lfloor wN\rfloor$ and $n_{k_{i,j}}\leq(1-\alpha)n_{k_{i,j}-1}$. By (3.23) and $k\gg\log^2(N)$, we have

$$\lim_{N\to\infty}\left[1+\frac{\log(\lfloor wN\rfloor)}{(1-\alpha)k}\right]^t=\lim_{N\to\infty}\exp\left\{\frac{t\log(\lfloor wN\rfloor)}{(1-\alpha)k}\right\}=1.$$

By (3.23) with $l\leq d$ and $\alpha\in(0,0.5]$, we have

$$(1-\alpha)^t\leq(1-\alpha)^{\frac{\log((2k-1)/\lfloor wN\rfloor)}{d\log(\alpha)}-\frac{l}{d}}\leq 2\left(\frac{\lfloor wN\rfloor}{2k-1}\right)^{-\frac{\log(1-\alpha)}{d\log(\alpha)}}.$$

Conditional on the event $\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$, as $N\to\infty$, we have for any $\mathbf{x}\in[0,1]^d$ and $\xi\in\Xi$,

$$\mathrm{diam}_j(L(\mathbf{x},\xi))=O\left(\left(\frac{N}{k}\right)^{-\frac{\log(1-\alpha)}{d\log(\alpha)}}\right).$$

Since $\mathrm{diam}^r(L(\mathbf{x},\xi))=\left[\sum_{j=1}^{d}\mathrm{diam}_j^2(L(\mathbf{x},\xi))\right]^{r/2}$ for any $r\geq 1$, conditional on the event $\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$, as $N\to\infty$, we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi}\{\mathrm{diam}^r(L(\mathbf{x},\xi))\}=O\left(\left(\frac{N}{k}\right)^{-\frac{r\log(1-\alpha)}{d\log(\alpha)}}\right).$$

■

Proof of Lemma 3.3. In this proof, we condition on the event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap\bar{\mathcal{A}}$, where $\bar{\mathcal{A}}:=\cap_{j=1}^{d}\cap_{i=1}^{t}\mathcal{A}_i$. Let $n=\lfloor wN\rfloor$. The event $\mathcal{A}$ is defined in Lemma 3.6, with $\mathcal{R}=\mathcal{R}_{\mathcal{D},\omega,\epsilon}$,

$\mu_{min} = \omega$, and $\delta = 1/\sqrt{n}$. The events $\mathcal{B}$, $\mathcal{C}$ and $\cap_{j=1}^{d} \cap_{i=1}^{t} \mathcal{A}_i$ are defined as (3.56), (3.58) and (3.88). By Lemmas 3.7 and 3.8, we know that $\mathbf{S}_L - \mathbf{d}_L \mathbf{d}_L^\top$, $\mathbf{S}_L$, $\mathbf{S}$, $\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top$ and $\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{G}(\mathbf{X}_i)\mathbf{G}(\mathbf{X}_i)^\top$ are all positive-definite, with $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{B}\cap\mathcal{C}) = 1 - o(1)$. Together with Lemma 3.6 and (3.89), we have $\mathbb{P}_{\mathbb{S}_\mathcal{I}}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap\bar{\mathcal{A}}) = 1 - o(1)$. Recall the definition of $\widehat{m}_{\mathrm{CLPF}}(\mathbf{x})$, (3.9); we have

$$\begin{aligned}\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| &= \sup_{\mathbf{x}\in[0,1]^d} \left|\mathbb{E}_\xi\left[\mathbf{G}(\mathbf{x})^\top\left(\widehat{\boldsymbol{\beta}}(\mathbf{x},\xi) - \boldsymbol{\beta}\right)\right]\right| \\ &\le \sup_{\mathbf{x}\in[0,1]^d}\left[\mathbb{E}_\xi\left|\mathbf{G}(\mathbf{x})^\top\left(\widehat{\boldsymbol{\beta}}(\mathbf{x},\xi) - \boldsymbol{\beta}\right)\right|\right].\end{aligned}$$

By (3.33), we have

$$\begin{aligned}&\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| \\ &\quad\le \sup_{\mathbf{x}\in[0,1]^d}\left[\mathbb{E}_\xi\left|\mathbf{e}_1^\top\left(\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top\right)^{-1}\left(\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\boldsymbol{\Delta}_i(\varepsilon_i + r_i)\right)\right|\right] \\ &\quad\overset{(i)}{\le} \sup_{\mathbf{x}\in[0,1]^d}\left[\mathbb{E}_\xi\left|(1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)(\varepsilon_i + r_i)\right.\right. \\ &\qquad\left.\left. + (1 - \mathbf{d}^\top\mathbf{S}^{-1}\mathbf{d})^{-1}\mathbf{d}^\top\mathbf{S}^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i(\varepsilon_i + r_i)\right|\right]. \qquad (3.90)\end{aligned}$$

where (i) hold by (3.34) and (3.35). Plugging (3.37) into (3.90), we have

$$\begin{aligned}\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| \le \sup_{\mathbf{x}\in[0,1]^d}\Bigg[\mathbb{E}_\xi\Bigg|&(1 - \mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{d}_L)^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)(\varepsilon_i + r_i) \\ &+ (1 - \mathbf{d}_L^\top\mathbf{S}_L^{-1}\mathbf{d}_L)^{-1}\mathbf{d}_L^\top\mathbf{S}_L^{-1}\sum_{i\in\mathcal{I}}\omega_i(\mathbf{x},\xi)\mathbf{U}_i^L(\varepsilon_i + r_i)\Bigg|\Bigg].\end{aligned}$$

By (3.40) and the triangle inequality,

$$\sup_{\mathbf{x}\in[0,1]^d} |\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})| \le \sum_{i=1}^{4}\left[\sup_{\mathbf{x}\in[0,1]^d}\mathbb{E}_\xi|\Delta_i(\mathbf{x},\xi)|\right]\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi}\left\{1 + \mathbf{d}_L^\top\left(\mathbf{S}_L - \mathbf{d}_L\mathbf{d}_L^\top\right)^{-1}\mathbf{d}_L\right\}, \qquad (3.91)$$

where $\Delta_i(\mathbf{x},\xi)$ ($i \in \{1,2,3,4\}$) are defined as (3.41)-(3.42). Since $m \in \mathcal{H}^{q,\beta}$, by (3.45), we have

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_1(\mathbf{x},\xi)\right| \leq \left[\sum_{|\boldsymbol{\alpha}|=q} \frac{L_0}{\boldsymbol{\alpha}!}\right] \sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{d}_L^\top \mathbf{S}_L^{-1}\mathbf{U}_i^L \|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right|,$$

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_2(\mathbf{x},\xi)\right| \leq \left[\sum_{|\boldsymbol{\alpha}|=q} \frac{L_0}{\boldsymbol{\alpha}!}\right] \sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right|.$$

By (3.46), (3.47) and (3.50), we have

$$\left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{d}_L^\top \mathbf{S}_L^{-1}\mathbf{U}_i^L \|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right| \leq \operatorname{diam}^{(q+\beta)}(L(\mathbf{x},\xi)),$$

$$\left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\|\mathbf{X}_i-\mathbf{x}\|^{q+\beta}\right| \leq \operatorname{diam}^{(q+\beta)}(L(\mathbf{x},\xi)).$$

Then, we have

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_1(\mathbf{x},\xi)\right| \leq \left[\sum_{|\boldsymbol{\alpha}|=q} \frac{L_0}{\boldsymbol{\alpha}!}\right] \sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left[\operatorname{diam}^{(q+\beta)}(L(\mathbf{x},\xi))\right],$$

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_2(\mathbf{x},\xi)\right| \leq \left[\sum_{|\boldsymbol{\alpha}|=q} \frac{L_0}{\boldsymbol{\alpha}!}\right] \sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left[\operatorname{diam}^{(q+\beta)}(L(\mathbf{x},\xi))\right].$$

Hence, by Lemma 3.2 with $r = q+\beta$, conditional on the event $\bar{\mathcal{A}}$, as $N \to \infty$, we have

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_1(\mathbf{x},\xi)\right| = O_p\left(\left(\frac{N}{k}\right)^{-\frac{(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right), \tag{3.92}$$

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_2(\mathbf{x},\xi)\right| = O_p\left(\left(\frac{N}{k}\right)^{-\frac{(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right). \tag{3.93}$$

Conditional on the event $\mathcal{A}$ above, we follow the proof of Lemma 3.7 to choose some $\bar{R} := \bar{R}(\mathbf{x},\xi) \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ as an inner approximation of $L(\mathbf{x},\xi)$ satisfying $\bar{R} \subseteq L(\mathbf{x},\xi)$ with (3.62) and (3.63). Recall the definition $\omega_i^L := \omega_i(\mathbf{x},\xi) = \mathbb{1}_{\{\mathbf{X}_i \in L(\mathbf{x},\xi)\}}/\#L$ and $\omega_i^{\bar{R}} := \mathbb{1}_{\left\{\mathbf{X}_i\in\bar{R}\right\}}/\#\bar{R}$, where $\#L := \#L(\mathbf{x},\xi)$ and $\#\bar{R} := \#\bar{R}(\mathbf{x},\xi)$. By the triangle inequality,

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_3(\mathbf{x},\xi)\right| \leq \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \left|\mathbf{d}_L^\top \mathbf{S}_L^{-1} \sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\mathbf{U}_i^L \varepsilon_i\right| \leq \sum_{j=1}^{2} \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} |\Delta_{3,j}| \tag{3.94}$$

where for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$,

$$\Delta_{3,1} := \Delta_{3,1}(\mathbf{x}, \xi) = \sum_{i \in \mathcal{I}} \omega_i^L \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i - \sum_{i \in \mathcal{I}} \omega_i^{\bar{R}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i,$$

$$\Delta_{3,2} := \Delta_{3,2}(\mathbf{x}, \xi) = \sum_{i \in \mathcal{I}} \omega_i^{\bar{R}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i.$$

Note that

$$\begin{aligned}
|\Delta_{3,1}| \leq & \left| \frac{1}{\#L} \sum_{i \in \left\{ i \in \mathcal{I} : \mathbf{X}_i \in \bar{R} \right\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i - \frac{1}{\#\bar{R}} \sum_{i \in \left\{ i \in \mathcal{I} : \mathbf{X}_i \in \bar{R} \right\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i \right| \\
& + \left| \frac{1}{\#L} \sum_{i \in \left\{ i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi) \setminus \bar{R} \right\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i \right| \leq \frac{2(\#L - \#\bar{R})}{\#L} \sup_{i \in \{ i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi) \}} \left| \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i \right| \\
& \overset{(i)}{\leq} 4M \frac{(\#L - \#\bar{R})}{\#L} \sup_{i \in \{ i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi) \}} \left| \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \right|, \qquad (3.95)
\end{aligned}$$

where (i) holds by $|\varepsilon| = |Y - \mathbb{E}[Y \mid \mathbf{X}]| \leq 2M$ since $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$ and $Y \in [-M, M]$. By the triangle inequality, we have

$$\|\mathbf{d}_L\|_2 \leq \sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) \left\| \mathbf{U}_i^L \right\|_2 \overset{(i)}{\leq} \sqrt{\bar{d}},$$

where (i) holds by (3.66) and $\sum_{i \in \mathcal{I}} \omega_i(\mathbf{x}, \xi) = 1$. By Cauchy-Schwarz inequality and the sub-multiplicative property of matrix norm, we have for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$,

$$\sup_{i \in \{ i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi) \}} \left| \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \right| \leq \sup_{i \in \{ i \in \mathcal{I} : \mathbf{X}_i \in L(\mathbf{x}, \xi) \}} \|\mathbf{d}_L\|_2 \|\mathbf{S}_L\|_2 \|\mathbf{U}_i^L\|_2 \overset{(i)}{\leq} \frac{\bar{d}}{\Lambda_0}. \qquad (3.96)$$

where (i) holds by Lemma 3.7 and (3.66). By (3.59) and (3.63), conditional on the event $\mathcal{A}$, as $N \to \infty$, we have $(\#L - \#\bar{R})/\#L = O(\sqrt{\log(N)/k})$ for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$, which implies that

$$\sup_{\mathbf{x} \in [0,1]^d, \xi \in \Xi} \left\{ \frac{\#L - \#\bar{R}}{\#L} \right\} = O\left(\sqrt{\frac{\log(N)}{k}} \right). \qquad (3.97)$$

Combining (3.96) and (3.97) with (3.95), conditional on the event $\mathcal{A}$, as $N \to \infty$ we have

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} |\Delta_{3,1}| = O\left(\sqrt{\frac{\log(N)}{k}}\right). \tag{3.98}$$

By $k \gg \log(N)$ and (3.97), conditional on the event $\mathcal{A}$, as $N \to \infty$, we have $(\#L - \#\bar{R})/\#L = o(1)$ for any $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$. Hence, there exists $n_1 \in \mathbb{N}$ such that $\#\bar{R} \geq k/2$ whenever $n \geq n_1$. Note that $\bar{R} \in \mathcal{R}_{\mathcal{D},\omega,\epsilon}$ for all $\mathbf{x} \in [0,1]^d$ and $\xi \in \Xi$. Therefore, when $n \geq n_1$,

$$\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} |\Delta_{3,2}| \leq \sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2} |\Delta_{3,2}|. \tag{3.99}$$

By the tower rule and $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$, we have $\mathbb{E}\left[\sum_{i\in\mathcal{I}} \omega_i^{\bar{R}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}\right] = 0$. Since $\varepsilon_i \in [-2M, 2M]$ for all $i \in \{i \in \mathcal{I} : \mathbf{X}_i \in R\}$ and (3.96), by Theorem 2 of [Hoe94], for any $\zeta > 0$,

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\left|\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i\right| \geq \zeta \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}, \mathcal{A}\right) \leq 2\exp\left\{-\frac{\Lambda_0^2 \#R\zeta^2}{8M^2\bar{d}^2}\right\}.$$

By the union bound, for all $\zeta \geq 0$,

$$\begin{aligned}&\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2}\left|\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i\right| \geq \zeta \mid \{\mathbf{X}_l\}_{l=1}^N, \{Y_l\}_{l\in\mathcal{J}}, \mathcal{A}\right)\\ &\quad\leq 2\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}\exp\left\{-\frac{\Lambda_0^2 k\zeta^2}{16M^2\bar{d}^2}\right\}.\end{aligned}$$

By the tower rule, for all $\zeta \geq 0$,

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{R\in\mathcal{R}_{\mathcal{D},\omega,\epsilon},\#R\geq k/2}\left|\frac{1}{\#R}\sum_{i\in\{i\in\mathcal{I}:\mathbf{X}_i\in R\}} \mathbf{d}_L^\top \mathbf{S}_L^{-1} \mathbf{U}_i^L \varepsilon_i\right| \geq \zeta \mid \mathcal{A}\right) \leq 2\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}\exp\left\{-\frac{\Lambda_0^2 k\zeta^2}{16M^2\bar{d}^2}\right\}.$$

Let $\zeta = \sqrt{\frac{32M^2\bar{d}^2\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{\Lambda_0^2 k}}$. By (3.59), there exists $n_2 \in \mathbb{N}$ such that $\#\mathcal{R}_{\mathcal{D},\omega,\epsilon} \geq 2\sqrt{n}$ whenever $n \geq n_2$. Together with (3.99), provided that $n \geq \max\{n_1, n_2\}$, we have

$$\mathbb{P}_{\mathbb{S}_\mathcal{I}}\left(\sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} |\Delta_{3,2}| \geq \sqrt{\frac{32M^2\bar{d}^2\log(\#\mathcal{R}_{\mathcal{D},\omega,\epsilon})}{\Lambda_0^2 k}} \mid \mathcal{A}\right) \leq \frac{2}{\#\mathcal{R}_{\mathcal{D},\omega,\epsilon}} \leq \frac{1}{\sqrt{n}}. \tag{3.100}$$

Combining (3.98) and (3.100) with (3.94), conditional on the event $\mathcal{A}$, as $N \to \infty$,

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_3(\mathbf{x},\xi)\right| = O_p\left(\sqrt{\frac{\log(N)}{k}}\right). \tag{3.101}$$

Note that

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_4(\mathbf{x},\xi)\right| = \sup_{\mathbf{x}\in[0,1]^d} \left[\mathbb{E}_\xi \left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\varepsilon_i\right|\right] \leq \sup_{\mathbf{x}\in[0,1]^d,\xi\in\Xi} \left|\sum_{i\in\mathcal{I}} \omega_i(\mathbf{x},\xi)\varepsilon_i\right|.$$

Repeating the same procedure as (3.101) except replacing $\mathbf{d}_L^\top \mathbf{S}_L^{-1}\mathbf{U}_i^L\varepsilon_i$ with ε_i, conditional on the event $\mathcal{A}$, as $N \to \infty$, we have

$$\sup_{\mathbf{x}\in[0,1]^d} \mathbb{E}_\xi \left|\Delta_4(\mathbf{x},\xi)\right| = O_p\left(\sqrt{\frac{\log(N)}{k}}\right). \tag{3.102}$$

Combining (3.92), (3.93), (3.101), (3.102), (3.44) with (3.91), we have

$$\sup_{\mathbf{x}\in[0,1]^d} \left|\widehat{m}_{\mathrm{CLPF}}(\mathbf{x}) - m(\mathbf{x})\right| = O_p\left(\sqrt{\frac{\log(N)}{k}} + \left(\frac{N}{k}\right)^{-\frac{(q+\beta)\log(1-\alpha)}{d\log(\alpha)}}\right).$$

■

Proof of Theorem 3.4. For this proof, it sufficient to check the conditions of Assumptions 2.1 from Theorem 2.1 of [CCD+17]. Let $V := A - \pi^*(\mathbf{X})$ and $U := U_1 + U_0$ with $U_a := \mathbb{1}_{\{A=a\}}(Y(a) - \mu_a^*(\mathbf{X}))$ for $a \in \{0,1\}$. By the definition of $\pi^*(\mathbf{X})$ and $\mu_a^*(\mathbf{X})$, we have $\mathbb{E}[V \mid \mathbf{X}] = 0$ and $\mathbb{E}[U_a \mid \mathbf{X}, A = a] = 0$ for $a \in \{0,1\}$. By the law of total probability, we have $\mathbb{E}[U \mid \mathbf{X}, A] = \mathbb{E}[U_1 \mid \mathbf{X}, A = 1]\mathbb{P}(A = 1 \mid \mathbf{X}) + \mathbb{E}[U_0 \mid \mathbf{X}, A = 0]\mathbb{P}(A = 0 \mid \mathbf{X}) = 0$. Hence, the condition (i) of Assumptions 2.1 is satisfied. Let $r > 4$ be any fixed positive constant. Since $|Y| \leq M$, we have $|\mu_a^*(\mathbf{X})| \leq M$, which implies $\{\mathbb{E}[\mu_a^*(\mathbf{X})]^r\}^{1/r} \leq M$. By $|Y| \leq M$, we also get $\{\mathbb{E}[Y]^r\}^{1/r} \leq M$. By the triangle inequality and $\mathbb{1}_{\{A=a\}} \leq 1$, we have $|U| \leq 2|U_a| \leq 2|Y - \mu_a^*(\mathbf{X})| \leq 2|Y| + 2|\mu_a^*(\mathbf{X})| \leq 4M$, which implies $\mathbb{P}(\mathbb{E}[U^2 \mid \mathbf{X}] \leq 4M) = 1$.

Since $\mathbb{E}[\mathbb{1}_{\{A=a\}}(Y(a)-\mu_a^*)]^2 \geq C_0$, we have $\{\mathbb{E}[U]^2\}^{1/2} = \{\mathbb{E}[U_0]^2 + \mathbb{E}[U_1]^2\}^{1/2} \geq \sqrt{2C_0}$. By overlap condition under Assumption 3.4, we have $\mathbb{P}(|A - \pi^*(\mathbf{X})| \geq c_0) = 1$, which implies $\{\mathbb{E}[V^2]\}^{1/2} \geq c_0$. Hence, the condition (ii) of Assumptions 2.1 is satisfied. By Theorem 3.2, we have

$$\{\mathbb{E}_{\mathbf{X}}\left[\widehat{\mu}_a^{-k}(\mathbf{X}) - \mu_a^*(\mathbf{X})\right]^2\}^{1/2} = O_p\left(N^{-\frac{(q_a+\beta_a)\log\left((1-\alpha_a)^{-1}\right)}{d\log\left(\alpha_a^{-1}\right)+2(q_a+\beta_a)\log\left((1-\alpha_a)^{-1}\right)}}\right) = o_p(1),$$

$$\{\mathbb{E}_{\mathbf{X}}\left[\widehat{\pi}^{-k}(\mathbf{X}) - \pi^*(\mathbf{X})\right]^2\}^{1/2} = O_p\left(N^{-\frac{(q_2+\beta_2)\log\left((1-\alpha_2)^{-1}\right)}{d\log\left(\alpha_2^{-1}\right)+2(q_2+\beta_2)\log\left((1-\alpha_2)^{-1}\right)}}\right) = o_p(1).$$

By $d \leq 2\sqrt{\frac{(q_a+\beta_a)(q_2+\beta_2)\log((1-\alpha_a)^{-1})\log((1-\alpha_2)^{-1})}{\log\left(\alpha_a^{-1}\right)\log\left(\alpha_2^{-1}\right)}}$ for $a = \{0,1\}$, we have

$$\frac{(q_a+\beta_a)\log\left((1-\alpha_a)^{-1}\right)}{d\log\left(\alpha_a^{-1}\right)+2(q_a+\beta_a)\log\left((1-\alpha_a)^{-1}\right)} + \frac{(q_2+\beta_2)\log\left((1-\alpha_2)^{-1}\right)}{d\log\left(\alpha_2^{-1}\right)+2(q_2+\beta_2)\log\left((1-\alpha_2)^{-1}\right)} \geq \frac{1}{2},$$

which implies

$$\{\mathbb{E}_{\mathbf{X}}\left[\widehat{\mu}_a^{-k}(\mathbf{X}) - \mu_a^*(\mathbf{X})\right]^2\}^{1/2}\{\mathbb{E}_{\mathbf{X}}\left[\widehat{\pi}^{-k}(\mathbf{X}) - \pi^*(\mathbf{X})\right]^2\}^{1/2} = o_p(N^{-1/2}).$$

By Lemma 3.3, we have $\mathbb{P}(c_1 < \widehat{\pi}^{-k}(\mathbf{X}) \leq 1 - c_1) = 1$ with probability approaching one and some constant $c_1 \in (0, 1/2)$. Hence, the condition (iii) of Assumptions 2.1 is satisfied. ∎

Proof of Lemma 3.3. By Theorem 3.3, as $N \to \infty$, we have $\sup_{\mathbf{x}\in[0,1]^{d_1}} |\widehat{\pi}^{-k}(\mathbf{x}) - \pi^*(\mathbf{x})| = o_p(1)$. Then, we have

$$\lim_{N\to\infty} \mathbb{P}_{\mathbb{S}_n}\left(\sup_{\mathbf{x}\in[0,1]^{d_1}} |\widehat{\pi}^{-k}(\mathbf{x}) - \pi^*(\mathbf{x})| \leq c_0/2\right) = 1,$$

which implies

$$\lim_{N\to\infty} \mathbb{P}_{\mathbb{S}_n}\left[\mathbb{P}_{\mathbf{X}}\left(\pi^*(\mathbf{S}_1) - c_0/2 < \widehat{\pi}^{-k}(\mathbf{X}) \leq \pi^*(\mathbf{X}) + c_0/2\right) = 1\right] = 1,$$

By the overlap condition $\mathbb{P}_{\mathbf{X}}(c_0 < \pi^*(\mathbf{X}) < 1 - c_0) = 1$, we have

$$\lim_{N\to\infty} \mathbb{P}_{\mathbb{S}_n}\left(\mathbb{P}_{\mathbf{X}}(c_0/2 < \widehat{\pi}^{-k}(\mathbf{X}) \leq 1 - c_0/2) = 1\right) = 1.$$

Let $c_1 = c_0/2$. Then, we have

$$\lim_{N\to\infty} \mathbb{P}_{\mathbb{S}_n}\left(\mathbb{P}_{\mathbf{X}}(c_1 < \widehat{\pi}^{-k}(\mathbf{X}) \leq 1 - c_1) = 1\right) = 1.$$

■

3.7 Acknowledgement

Chaper 3, in full, is currently being prepared for submission for publication of the material. Bradic, Jelena; Ji, Weijie; Zhang, Yuqian. Adaptive split balancing for optimal random forests. The dissertation author was the primary investigator and author of this material.

Bibliography

[AG14] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

[AI16] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[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.

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

[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.

[BDL08] Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

[BFSO84] Leo Breiman, Jerome Friedman, Charles J Stone, and RA Olshen. *Classification and Regression Trees*. CRC Press, 1984.

[BHL22] Hugo Bodory, Martin Huber, and Lukáš Lafférs. Evaluating (weighted) dynamic treatment effects by double machine learning. *The Econometrics Journal*, 25(3):628–648, 2022.

[Bia12] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[BJZ21] Jelena Bradic, Weijie Ji, and Yuqian Zhang. High-dimensional inference for dynamic treatment effects. *arXiv preprint arXiv:2110.04924*, 2021.

[BR05] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[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.

[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.

[BTYW16] Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. Supervised neighborhoods for distributed nonparametric regression. In *Artificial Intelligence and Statistics*, pages 1450–1459. PMLR, 2016.

[BWLY22] Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences*, 119(22):e2118636119, 2022.

[BWZ19] Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

[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.

[CF15] Xuan Chen and Carlos A Flores. Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of job corps. *Journal of Business & Economic Statistics*, 33(4):523–540, 2015.

[CKT22] Mattias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.10805*, 2022.

[CKU23] Matias D Cattaneo, Jason M Klusowski, and William G Underwood. Inference with mondrian random forests. *arXiv preprint arXiv:2310.09702*, 2023.

[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.

[CMDY23] Yuchao Cai, Yuheng Ma, Yiwei Dong, and Hanfang Yang. Extrapolated random tree for regression. 2023.

[CVFL22] Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.

[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.

[DS18] Roxane Duroux and Erwan Scornet. Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128, 2018.

[DV20] Oliver Dukes and Stijn Vansteelandt. Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 2020.

[DV21] Oliver Dukes and Stijn Vansteelandt. Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 108(2):321–334, 2021.

[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.

[Far15] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[Fed14] Herbert Federer. *Geometric measure theory*. Springer, 2014.

[FFLGN12] Carlos A Flores, Alfonso Flores-Lagunes, Arturo Gonzalez, and Todd C Neumann. Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Review of Economics and Statistics*, 94(1):153–171, 2012.

[Fri91] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[FTAW20] Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517, 2020.

[Gen12] Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

[GPB11] Benjamin A Goldstein, Eric C Polley, and Farren BS Briggs. Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, 10(1), 2011.

[GXZ22] Wei Gao, Fan Xu, and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022.

[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.

[HHLL20] Martin Huber, Yu-Chin Hsu, Ying-Ying Lee, and Layal Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840, 2020.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[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.

[HZ21] Torsten Hothorn and Achim Zeileis. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 30(4):1181–1196, 2021.

[IK10] Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064, 2010.

[IKBL08] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

[IR15] 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.

[Ish15] Hemant Ishwaran. The effect of splitting on random forests. *Machine learning*, 99:75–118, 2015.

[Ken20] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[Klu21] Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.

[KS18] Nathan Kallus and Michele Santacatterina. Optimal balancing of time-dependent confounders for marginal structural models. *arXiv preprint arXiv:1806.01083*, 2018.

[KT22] J Klusowski and P Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 2022.

[Lee09] David S Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102, 2009.

[LH21] Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation. *The Journal of Machine Learning Research*, 22(1):386–426, 2021.

[LM10] Michael Lechner and Ruth Miquel. Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics*, 39:111–137, 2010.

[LS21] Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects. *Advances in Neural Information Processing Systems*, 34:22695–22707, 2021.

[LWB+19] Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased mdi feature importance measure for random forests. *Advances in Neural Information Processing Systems*, 32, 2019.

[LWSG13] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, 26, 2013.

[MGS20] Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for mondrian trees and forests. *The Annals of Statistics*, 48(4):2253–2276, 2020.

[MH14] Lucas Mentch and Giles Hooker. Ensemble trees and clts: Statistical inference for supervised learning. *stat*, 1050:25, 2014.

[MR06] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

[Mur03] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[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.

[OT21] Eliza O'Reilly and Ngoc Mai Tran. Minimax rates for high-dimensional random tessellation forests. *arXiv preprint arXiv:2109.10541*, 2021.

[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.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[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.

[SBM08] Peter Z Schochet, John Burghardt, and Sheena McConnell. Does job corps work? impact findings from the national job corps study. *American Economic Review*, 98(5):1864–1886, 2008.

[SBRJ+03] P Schochet, J Bellotti, C Ruo-Jiao, S Glazerman, A Grady, M Gritz, S McConnell, T Johnson, and J Burghardt. National job corps study: data documentation and public use files. *vols. I-IV). Washington, DC: Mathematica Policy Research, Inc*, 2003.

[SBV15] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

[Sch01] Peter Z Schochet. *National Job Corps Study: The impacts of Job Corps on participants' employment and related outcomes*. US Department of Labor, Employment and Training Administration, Office of Policy and Research, 2001.

[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.

[SSM19] Rodney Sparapani, Charles Spanbauer, and Robert McCulloch. The bart r package. *Accessed on Aug*, 21:2019, 2019.

[Sto82] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[SXG21] Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for multistage causal inference: Mediation analysis and dynamic treatment effects. *arXiv preprint arXiv:2111.03950*, 2021.

[TAF+23] Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. Package 'grf', 2023.

[Tan20] 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.

[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.

[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.

[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.

[vdLG12] Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1), 2012.

[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.

[WW15] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[WWPW19] Marvin N Wright, Stefan Wager, Philipp Probst, and Maintainer Marvin N Wright. Package 'ranger'. *Version 0.11*, 2, 2019.

[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.

[ZB22] Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

[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.

[ZL12] Guoyi Zhang and Yan Lu. Bias-corrected random forests in regression. *Journal of Applied Statistics*, 39(1):151–160, 2012.

[ZRM08] Junni L Zhang, Donald B Rubin, and Fabrizia Mealli. Evaluating the effects of job training programs on wages through principal stratification. In *Modelling and Evaluating Treatment Effects in Econometrics*. Emerald Group Publishing Limited, 2008.

[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.