# THE DECAYING MISSING-AT-RANDOM FRAMEWORK: DOUBLY ROBUST CAUSAL INFERENCE WITH PARTIALLY LABELED DATA

BY YUQIAN ZHANG[1,*], ABHISHEK CHAKRABORTTY[2,*] AND JELENA BRADIC[3,*]

[1]*Institute of Statistics and Big Data, Renmin University of China* yuqianzhang@ruc.edu.cn

[2]*Department of Statistics, Texas A&M University* abhishek@stat.tamu.edu

[3]*Department of Mathematics and Halicioglu Data Science Institute, University of California San Diego* jbradic@ucsd.edu

In real-world scenarios, data collection limitations often result in partially labeled datasets, leading to difficulties in drawing reliable causal inferences. Traditional approaches in the semi-supervised (SS) and missing data literature may not adequately handle these complexities, leading to biased estimates. To address these challenges, our paper introduces a novel *decaying missing-at-random* (decaying MAR) framework. This framework tackles missing outcomes in high-dimensional settings and accounts for selection bias arising from the dependence of labeling probability on covariates. Notably, we relax the need for a positivity condition, commonly required in the missing data literature, and allow uniform decay of labeling propensity scores with sample size, accommodating faster growth of unlabeled data. Our decaying MAR framework enables easy rate double-robust (DR) estimation of average treatment effects, succeeding where other methods fail, even with correctly specified nuisance models. Additionally, it facilitates asymptotic normality under model misspecification. To achieve this, we propose adaptive new targeted bias-reducing nuisance estimators and asymmetric cross-fitting, along with a novel semi-parametric approach that fully leverages large volumes of unlabeled data. Our approach requires weak sparsity conditions. Numerical results confirm our estimators' efficacy and versatility, addressing selection bias and model misspecification.

**1. Introduction.** Semi-supervised (SS) learning's importance in estimating the average treatment effect (ATE) is increasingly recognized in a wide range of fields. Despite having a large total sample size (denoted as $N$), practical restrictions often result in missing outcomes (or labels) $Y \in \mathbb{R}$. The primary objective of this research is to explore how to effectively utilize this rich yet intricate dataset to study the causal impact of a binary treatment denoted as $T \in \{0,1\}$ on $Y$. For instance, in autonomous driving, exploiting abundant unlabeled camera footage could improve the detection of rare incidents. In the cybersecurity sector, the evaluation of extensive unlabeled data might enhance fraud detection capabilities. Wildlife conservation could benefit from using unlabeled images for population monitoring strategies. Similarly, integrative genomics studies can identify novel gene-disease associations by incorporating and analyzing large unlabeled datasets.

Under the potential outcome framework (Rubin, 1974; Imbens and Rubin, 2015), we consider potential outcomes $Y(0)$ and $Y(1)$, corresponding to treatment assignments $T=0$ and $T=1$, respectively. The observed outcome is denoted by $Y \equiv Y(T)$ and under consistency assumption $Y = TY(1) + (1-T)Y(0)$. The ATE characterizes the average causal effect of $T$ on $Y$ and is defined as

$$\mu_0 := \theta_1 - \theta_0, \text{ where } \theta_1 := \mathbb{E}\{Y(1)\} \text{ and } \theta_0 := \mathbb{E}\{Y(0)\}. \tag{1.1}$$

[*]All authors contributed equally in this work.

*Keywords and phrases:* Causal inference, decaying overlap, generalizability, high dimensional nuisances, debiasing, sparsity.

Estimating the ATE in observational studies, where the treatment $T$ and outcome $Y$ are influenced by a shared set of confounders $\mathbf{X} \in \mathbb{R}^d$, presents challenges due to confounding. The complexity deepens in semi-supervised (SS) settings. In this context, a labeled dataset $\mathcal{L} = (\mathbf{X}_i, Y_i)_{i=1}^n$ coexists with a substantial amount of unlabeled data $\mathcal{U} = (\mathbf{X}_i)_{i=n+1}^N$, where the outcome variable is missing. Classical semi-supervised (SS) approaches assume that the labeled dataset ($\mathcal{L}$) and the unlabeled dataset ($\mathcal{U}$) share the same distribution, assuming missingness completely at random (MCAR) for the outcomes (Cheng, Ananthakrishnan and Cai, 2021; Zhang and Bradic, 2022; Hou, Mukherjee and Cai, 2021; Chakrabortty, Dai and Tchetgen, 2022). This assumption enables effective utilization of both labeled and unlabeled data. However, in real-world scenarios, MCAR is often violated. The objective of this paper is to address these limitations and tackle challenges associated with selection bias, where the missingness of the observed outcome $Y$, denoted as the labeling indicator $R \in \{0,1\}$, is itself observational and possibly dependent on both $(T, \mathbf{X})$. Furthermore, in situations where the unlabeled dataset is much larger than the labeled dataset ($N \gg n$), the probability of observing $Y$ decreases as $N$ increases, which violates the positivity assumption (Crump et al., 2009).

1.1. *Decaying MAR setup.* Unlike 'traditional' SS settings, we treat $R$ as *random* here. We allow the labeled fraction $n/N > 0$ to be arbitrarily close to zero, and study the theory when both $\mathbb{P}(R=1) \to 0$ and $N \to \infty$. Note that we consider the labeling probability $p_N := \mathbb{P}(R=1)$, the marginal distribution of $R \equiv R_{(N)}$, as well as the conditional distribution of $R \mid (T, \mathbf{X})$ as (decaying) functions of $N$ and let $N \to \infty$; see Section 2 for the usefulness and necessity of this construction.

We define the true outcome regression (OR) models, the propensity score (PS) models corresponding to the treatment $T$ and the labeling indicator $R$, as well as the PS models for a 'product' indicator as follows. For $j \in \{0,1\}$ and $\mathbf{x}$ in $\mathcal{X} \subseteq \mathbb{R}^d$, the support of $\mathbf{X}$, we define:

(1.2) **OR models:** $m(j, \mathbf{x}) := \mathbb{E}\{Y(j) \mid \mathbf{X} = \mathbf{x}\}$,

(1.3) **T-PS models:** $\pi(\mathbf{x}) := \mathbb{P}(T = 1 \mid \mathbf{X} = \mathbf{x})$, $\pi(j, \mathbf{x}) := \mathbb{P}(T = j \mid R = 1, \mathbf{X} = \mathbf{x})$,

(1.4) **R-PS models:** $p_N(\mathbf{x}) := \mathbb{P}(R = 1 \mid \mathbf{X} = \mathbf{x})$, $p_N(j, \mathbf{x}) := \mathbb{P}(R = 1 \mid T = j, \mathbf{X} = \mathbf{x})$,

(1.5) **Product PS models:** $\gamma_N(j, \mathbf{x}) := \mathbb{P}(\Gamma^{(j)} = 1 \mid \mathbf{X} = \mathbf{x})$, and $\gamma_{N,j} := \mathbb{P}(\Gamma^{(j)} = 1)$,

where $\Gamma^{(1)} := TR$ and $\Gamma^{(0)} := (1-T)R$ are the *product indicators*.

ASSUMPTION 1 (Basic assumptions). (a) We assume the 'no unmeasured confounding' (NUC) and overlap assumptions for the treatment $T$, so that for some constant $c \in (0, 1/2)$:

$$T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid \mathbf{X} \ \rightsquigarrow \textbf{(T-NUC)} \ \text{ and } \ c < \pi(\mathbf{x}) < 1 - c \ \forall\, \mathbf{x} \in \mathcal{X} \ \rightsquigarrow \textbf{(T-overlap)}.$$

(b) We further assume missing at random (MAR) condition and a *'decaying overlap'* condition (DOC) for the labeling indicator $R \equiv R_{(N)}$ as follows:

$$R \perp\!\!\!\perp Y \mid (T, \mathbf{X}) \ \rightsquigarrow \textbf{(R-MAR)}; \ \ \textit{and} \ \text{ for any } \mathbf{x} \in \mathcal{X},\ j \in \{0,1\}, \text{ and for each fixed } N,$$

$$\{p_N(\mathbf{x}), p_N(j, \mathbf{x})\} > 0, \text{ while possibly, } \{p_N(\mathbf{x}),\ p_N(j, \mathbf{x})\} \to 0 \text{ as } N \to \infty \ \rightsquigarrow \textbf{(R-DOC)}.$$

The T-NUC assumption ('ignorability'), and the T-overlap condition are commonly assumed (Rosenbaum and Rubin, 1983; Crump et al., 2009; Imbens and Rubin, 2015). The R-MAR condition was used recently in Wei et al. (2022), but the authors didn't consider the full semi-supervised setting of $N \gg n$. Kallus and Mao (2020) discuss it but only develop theory under a relaxed case of $R \perp\!\!\!\perp T \mid \mathbf{X}$ with a troublesome assumption $\mathbb{P}(R = 1) = 0$, which leads to $n = \sum_{i=1}^{N} R_i = 0$ almost surely, i.e., a degenerate unsupervised setting. For this, the R-DOC assumption plays a critical role and is a non-standard condition. It is weaker than the traditional positivity condition, which requires the existence of a constant $c > 0$ independent of $N$ such that $p_N(\mathbf{x}) > c$ (and $p_N(j, \mathbf{x}) > c$ in causal setups) (Bang and Robins, 2005; Tsiatis, 2007). Recently, Zhang, Chakrabortty and Bradic (2023) considered a slightly different version of the R-DOC condition that only involves conditions for $p_N(\cdot)$. When both the R-MAR and R-DOC conditions are satisfied, we refer to it as the 'decaying MAR setting'.

1.2. *Our contributions.* We introduce a decaying MAR setting, redefining the non-MCAR SS setup. This transformative approach addresses a previously uncharted intersection: it tackles the often-ignored selection bias in the SS literature and challenges the traditional positivity condition in the missing data domain; see Table 1.1. By doing so, our research contributes to the literature on the 'generalizability' of randomized controlled trials (RCTs), where RCT is combined with unlabeled, external data; see Dahabreh et al. (2019), Lesko et al. (2017) and also Shi, Pan and Miao (2023) for a review. In our context, $R$ denotes whether an individual is involved in the RCT or not and lack of strict positivity condition allows our work to be impactful when external data size surpasses that of an RCT.

TABLE 1.1

*Comparison of the missing outcome settings. 'Selection bias' allows $R$ to be dependent on $(Y, \mathbf{X})$. 'Causal setup + missing $Y$' allows for the observed outcome $Y = Y(T)$ to be possibly missing with $R, T \not\equiv 1$.*

| Settings | Selection Bias | Decaying PS | Causal + missing $Y$ |
|---|---|---|---|
| Kawakita and Kanamori (2013); Azriel et al. (2022); Chakrabortty and Cai (2018); Zhang, Brown and Cai (2019); Cai and Guo (2020); Chan et al. (2020); Xue, Ma and Li (2021); Chakrabortty, Dai and Carroll (2022) | × | ✓ | × |
| Cheng, Ananthakrishnan and Cai (2021); Hou, Mukherjee and Cai (2021); Zhang and Bradic (2022); Chakrabortty, Dai and Tchetgen (2022) | × | ✓ | ✓ |
| Rubin (1976); Robins, Rotnitzky and Zhao (1994); Robins and Rotnitzky (1995); Bang and Robins (2005); Tsiatis (2007); Kang and Schafer (2007); Graham (2011); Chakrabortty et al. (2019) | ✓ | × | × |
| Dahabreh et al. (2019); Lesko et al. (2017); Kallus and Mao (2020); Wei et al. (2022) | ✓ | × | ✓ |
| Zhang, Chakrabortty and Bradic (2023) | ✓ | ✓ | × |
| The proposed (causal) decaying MAR setting | ✓ | ✓ | ✓ |

As we improve DR properties, understanding their definitions is crucial. The ***model DR*** property states that the ATE estimator is asymptotically normally distributed when either of the nuisances is correctly specified. See Tan (2020) and Dukes, Avagyan and Vansteelandt (2020), with minor modifications in Smucler, Rotnitzky and Robins (2019). The ***rate DR*** property requires both nuisances to be correctly specified with the product of their sparsities of the order of $o(N)$. The ***sparsity DR***[1] property of Bradic, Wager and Zhu (2019), needs correct model specifications with one ultra-sparse nuisance at $o(\sqrt{N})$, while the other is at $o(N)$ for PS with ultra-sparse OR, or $o(N^{3/4})$ for the OR with ultra-sparse PS.

We highlight the utility of decaying MAR framework by discussing the ease of attaining ***rate DR***; see (A.7) and Theorem A.2 with adaptive rates accounting for PS decay. Prior

[1]Note that the rate DR property and the sparsity DR property are distinct and not mutually implied. Moreover, the definitions above should be interpreted up to a logarithmic factor.

research, even those based on simpler MCAR conditions or neglecting $N \gg n$ scenarios, has been limited; see Remark 1. Decaying MAR extends to missing treatment cases as well; see Corollary 3.1. We then propose two new DR method classes, each anchored in distinct PS representations. The first is parametric, introducing the *sparsity DR* and *model DR* *bias-reduced* doubly robust decaying MAR (DR-DMAR) SS estimator, abbreviated as BRSS (Section 4.1). The second approach is semi-parametric, named the semi-parametric bias-reduced DR-DMAR SS estimator (abbreviated as SP-BRSS, see Section 4.3). This approach takes advantage of the fully observed pairs $(T_i, \mathbf{X}_i)$ and advances a broader range of *sparsity DR* and *model DR* robust techniques with a new nuisance model class. Our primary findings are Theorems 4.4, 4.5, and Corollary 4.6. Our method, subsumes existing (special) cases: Chernozhukov et al. (2018)'s rate DR method with $R \equiv 1$, Zhang and Bradic (2022) and Chakrabortty, Dai and Tchetgen (2022) DR method under a consistent PS of $n/N$, and Zhang, Chakrabortty and Bradic (2023)'s non-causal DR with $T \equiv 1$; see Remark 3. We however, exhibit better robustness and outperform these methods, even in the canonical cases, as evidenced in Figure 1 and Table 4.1.

1.3. *Organization.* Section 2 introduces the decaying MAR setting and ATE's identification. Section 3 proposes the DR-DMAR SS estimator and its theoretical properties. Sections 3.2 and 3.3 discuss decaying PS estimation and missing treatments. Section 4 defines a BRSS estimator with a parametric and a semi-parametric approaches and showcases main theoretical results. Sections 5 and 6 provide numerical results on simulated and pseudo-random datasets. Concluding discussion is in Section 7 whereas additional results, and the proofs are relegated to the Supplement.

1.4. *Notation.* Throughout this work we will use various positive constants independent of $N$ denoted as lower or capital letters $c$ and $C$. $\mathbb{P}(\cdot)$ and $\mathbb{E}(\cdot)$ indicate the joint distribution and expectation of random vector $\mathbb{Z}$. $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{E}_{\mathbf{X}}\{f(\mathbf{X})\}$ denote the marginal distribution of $\mathbf{X}$ and expectation for any function $f$, respectively. For any subset $\mathcal{A}$, $\mathbb{P}_{\mathcal{A}}$ and $\mathbb{E}_{\mathcal{A}}(\cdot)$ signify its joint distribution and respective expectation. For any $r > 0$, $\|f(\cdot)\|_{\mathbb{P},r} := \{\mathbb{E}|f(\mathbb{Z})|^r\}^{1/r}$ and for any vector $\mathbf{z} \in \mathbb{R}^d$, $\|\mathbf{z}\|_r := (\sum_{j=1}^d |\mathbf{z}_j|^r)^{1/r}$, $\|\mathbf{z}\|_0 := |\{j : \mathbf{z}_j \neq 0\}|$, and $\|\mathbf{z}\|_\infty := \max_j |\mathbf{z}_j|$. For any square matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, $\|\mathbf{\Sigma}\|_r := \sup_{\|\mathbf{v}\|\neq 0} \|\mathbf{\Sigma}\mathbf{v}\|_r/\|\mathbf{v}\|_r$. $a_N \asymp b_N$ denotes equivalent sequences. Lastly, $\mathbf{e}_j \in \mathbb{R}^d$ refers to the $j$-th column of identity matrix $\mathbb{I}_d \in \mathbb{R}^{d\times d}$.

**2. The decaying MAR setting and estimation of the ATE.** In the context of SS inference assuming MCAR, the goal is to improve the supervised approach's efficiency using unlabeled data. However, in a decaying MAR framework, this becomes invalid, and MCAR-based methods introduce bias. Hence, we face a more challenging task: addressing identification issues from scratch to achieve consistent, optimal, and efficient SS estimation within a non-standard asymptotic regime.

*Necessity and usefulness of the decaying PS.* One primary contribution of this paper is the introduction of the decaying MAR schema, which addresses the dependence of $p_N$, $p_N(\mathbf{X})$, $R \equiv R_{(N)}$, and $R \mid (T, \mathbf{X})$ on $N$ and $\mathbf{X}$ in $N \gg n$ semi-supervised (SS) contexts. This schema accounts for the asymptotic scenario where $N \to \infty$ and $\mathbb{P}(R = 1) \to 0$. Previous studies may have overlooked this aspect, leading to limitations in exploring the crucial $N \gg n$ setting and only allowing $\mathbb{P}(R = 1) = c = \lim_{N\to\infty} n/N > 0$. In conventional doubly-robust (DR) literature, estimation error control relies on $p_N^{-2}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = o(1)$ for a nuisance parameter $\boldsymbol{\alpha}^*$ and its corresponding estimator $\widehat{\boldsymbol{\alpha}}$; see, e.g., the control of $\mathcal{I}_2$ in Step 5 of the proof of Theorems 5.1 and 5.2 of Chernozhukov et al. (2018). However, this condition becomes excessively demanding when $p_N \to 0$, rendering accurate estimation of the nuisance parameters practically unattainable. For instance, as the expected labeled sample size is only $\mathbb{E}(n) = Np_N$, we have $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 = O_p((Np_N)^{-1/2})$ even in low dimensions. This leads to a very restrictive requirement on the decaying probability, $Np_N^5 \to \infty$. Through refined analysis on the decaying PS, our results only require $\mathbb{E}(n) = Np_N \to \infty$.

*Identification of the parameter(s).* Given definitions in (1.2)-(1.4) and Assumption 1, we identify multiple versions of our parameters, each with unique benefits and interpretations and involving estimable *nuisance functions* from observed data. These include conditional mean regression (Reg.), inverse probability weighting (IPW) with propensity score (PS) modeling, and augmented IPW representations that use both the conditional mean and PS. We illustrate these for $\theta_1 = \mathbb{E}\{Y(1)\}$ without loss of generality, with analogous versions for $\theta_0 = \mathbb{E}\{Y(0)\}$ by substituting $(T, Y(1))$ with $(1-T, Y(0))$. To simplify the notation, we let

$$\Gamma := \Gamma^{(1)}, \;\; \gamma_N := \gamma_{N,1}, \;\; \gamma_N(\cdot) := \gamma_N(1,\cdot), \;\text{ and }\; m(\cdot) := m(1,\cdot).$$

LEMMA 2.1 (Identification of $\theta_1$). *Let Assumption 1 hold. Then, we have the following Reg. and IPW representations (Rep.):*

$$\theta_1 \;=\; \mathbb{E}\{m(\mathbf{X})\} \;\rightsquigarrow\; (\textbf{Reg. Rep.}), \quad \theta_1 \;=\; \mathbb{E}\left\{\frac{\Gamma Y}{\gamma_N(\mathbf{X})}\right\} \;\rightsquigarrow\; (\textbf{IPW Rep.}),$$

*where $m(\cdot)$ is identifiable as $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, \Gamma = 1)$. Additionally, for any arbitrary functions $m^*(\cdot)$ and $\gamma_N^*(\cdot)$, as long as either $m^*(\cdot) = m(\cdot)$ or $\gamma_N^*(\cdot) = \gamma_N(\cdot)$ holds but not necessarily both, we have:*

(2.1)
$$\mathbb{E}\left\{\psi_{N,1}^*(\mathbf{Z})\right\} = 0 \;\rightsquigarrow\; (\textbf{DR Rep.}), \text{ with } \psi_{N,1}^*(\mathbf{Z}) = m^*(\mathbf{X}) - \theta_1 + \frac{\Gamma}{\gamma_N^*(\mathbf{X})}\{Y - m^*(\mathbf{X})\}.$$

The aforementioned representations elucidate that this can be seen as a mean estimation issue with a MAR labeling, with the *effective labeling indicator* being $\Gamma = TR$. The DR representation tolerates misspecification in $m(\cdot)$ or $\gamma_N(\cdot)$, resulting in a consistent $\theta_1$ estimator if either, but not necessarily both, are correctly estimated.

## 3. The general DR-DMAR SS ATE estimator: construction and asymptotics.

We split the samples into $\mathbb{K} \geq 2$ parts, $\mathcal{I}_1, \ldots, \mathcal{I}_{\mathbb{K}}$ of equal sizes $|\mathcal{I}_k| = M = N/\mathbb{K}$ and define $\mathcal{I}_{-k} = \mathcal{I} \setminus \mathcal{I}_k$, $\forall k \leq \mathbb{K}$. Let $\widehat{m}^{(-k)}(j,\cdot)$ and $\widehat{\gamma}_N^{(-k)}(j,\cdot)$ be estimators of $m(j,\cdot)$ and $\gamma_N(j,\cdot)$, respectively, using the samples from $\mathcal{I}_{-k}$, based on suitable (working) models with one but *not* necessarily both required to be correctly specified. The estimator $\widehat{\gamma}_N(\cdot) = \widehat{\gamma}_N(1,\cdot)$ can be obtained in multiple ways: directly modelling $\Gamma \mid \mathbf{X} \equiv TR \mid \mathbf{X}$ or modeling $T \mid \mathbf{X}$ and $R \mid (T = 1, \mathbf{X})$ or by modeling $R \mid \mathbf{X}$ and $T \mid (R = 1, \mathbf{X})$; see more discussions in Section 3.2. The estimator $\widehat{m}(\cdot) = \widehat{m}(1,\cdot)$, on the other hand, can be simply obtained via any suitable (working) regression model for $(Y \mid T = 1, R = 1, \mathbf{X}) \equiv (Y \mid TR = 1, \mathbf{X})$. For each $i \in \mathcal{I}_k$, $\widetilde{m}(j, \mathbf{X}_i) = \widehat{m}^{(-k)}(j, \mathbf{X}_i)$ and $\widetilde{\gamma}_N(j, \mathbf{X}_i) = \widehat{\gamma}_N^{(-k)}(j, \mathbf{X}_i)$. We now define our *DR decaying MAR (DR-DMAR) SS estimator* $\widehat{\theta}_{j,\text{ss}}$ of $\theta_j$ ($j \in \{0,1\}$) as

$$\widehat{\theta}_{j,\text{ss}} \;:=\; \frac{1}{N}\sum_{i=1}^{N} \widetilde{m}(j, \mathbf{X}_i) + \frac{\Gamma_i^{(j)}(Y_i - \widetilde{m}(j, \mathbf{X}_i))}{\widetilde{\gamma}_N(j, \mathbf{X}_i)}, \tag{3.1}$$

where $\Gamma_i^{(1)} := T_i R_i$ and $\Gamma_i^{(0)} := (1 - T_i)R_i$. The *DR-DMAR SS ATE estimator* is defined as:

$$\widehat{\mu}_{\text{ss}} := \widehat{\theta}_{1,\text{ss}} - \widehat{\theta}_{0,\text{ss}}. \tag{3.2}$$

3.1. *Asymptotic properties.* Let $\gamma_N^*(j,\cdot) = \gamma_N(j,\cdot)$ and $m^*(j,\cdot) = m(j,\cdot)$. The Supplement's Theorems A.2-A.3 provide a full description of the main and supporting results; see Supplement. Here, we emphasize conclusions and their significance. Assume $\mathbb{E}\{\gamma_N^{-1}(j,\mathbf{X})\} < \infty$ and define

$$a_{N,j} := [\mathbb{E}\{\gamma_N^{-1}(j,\mathbf{X})\}]^{-1} \text{ and } a_N := \min(a_{N,1}, a_{N,0}).$$

For each $j \in \{0,1\}$, define the DR score

$$\psi_{N,j}^{opt}(\mathbf{Z}) := m(j,\mathbf{X}) - \theta_j + \frac{\Gamma^{(j)}}{\gamma_N(j,\mathbf{X})}\{Y(j) - m(j,\mathbf{X})\}.$$

Under the listed full conditions in the Supplement and with the product rate condition satisfied for each $j \in \{0,1\}$,

$$\mathbb{E}_{\mathbf{X}}\{\widehat{m}(j,\mathbf{X}) - m(j,\mathbf{X})\}^2 \mathbb{E}_{\mathbf{X}}\left\{1 - \frac{\gamma_N(j,\mathbf{X})}{\widehat{\gamma}_N(j,\mathbf{X})}\right\}^2 := O_p(c_{N,j}^2 d_{N,j}^2) = o_p(1/(Na_N)), \tag{3.3}$$

the DR-DMAR SS estimator is ***rate DR*** in that

$$\sqrt{Na_N}(\widehat{\mu}_{\mathrm{ss}} - \mu_0) = O_p(1) \text{ and } \sqrt{N}\left(\Sigma_N^{opt}\right)^{-1/2}(\widehat{\mu}_{\mathrm{ss}} - \mu_0) \xrightarrow{d} \mathcal{N}(0,1), \tag{3.4}$$

where $\Sigma_N^{opt} := \mathrm{Var}\{\psi_{N,1}^{opt}(\mathbf{Z}) - \psi_{N,0}^{opt}(\mathbf{Z})\} \asymp a_N^{-1}$. Here, $Na_N$ denotes the *'effective sample size'* within a decaying MAR context, while $a_N$ signifies the deceleration factor resulting from the decaying PS, giving rise to atypical convergence rates. A simpler non-causal problem was studied in Zhang, Chakrabortty and Bradic (2023). In causal scenarios, Kallus and Mao (2020) proposed a rate DR theory for the ATE, setting semi-parametric efficiency bounds for possible missing outcomes and observable surrogate variables. However, their modeling is founded in a critically flawed $\mathbb{P}(R = 1) = 0$ (see discussion in Section 1.1) and their Theorems 4.1 and 4.2 rely on $R \perp\!\!\!\perp T \mid \mathbf{X}$ which we remove. Yet our result achieves their semi-parametric efficiency bound.

REMARK 1 (Comparing with Special Cases of Decaying MAR Studies). In the *supervised causal setting*, $Y$ is always observable ($R \equiv 1$). Our findings in (A.7) (and Theorem A.2) are in line with but are distinct from those of double machine learning (Chernozhukov et al., 2018); directly adopting their result would lead to sub-optimal convergence rates. For accuracy, we control the non-standard ratio $\Gamma_i^{(j)}/\widetilde{\gamma}_N(j,\mathbf{X}_i)$ of (3.1) where most $\Gamma_i^{(j)}$ are zeros, and $\widetilde{\gamma}_N(j,\mathbf{x})$ may decay uniformly over $\mathbf{x} \in \mathcal{X}$. A low-dimensional DR solution of Wei et al. (2022) uses strict positivity condition ($p_N > c > 0$) which forbids a truly SS setting where $N \gg n$. The *regular SS setup* includes MCAR-missing outcomes only (Cheng, Ananthakrishnan and Cai, 2021; Zhang, Brown and Cai, 2019; Hou, Mukherjee and Cai, 2021). The ATE estimators of Zhang and Bradic (2022); Chakrabortty, Dai and Tchetgen (2022) are special case of ours, but exhibit bias when missingness is not completely at random.

3.2. *Estimation of the PS.* Estimating PS in the decaying MAR setting is challenging due to extreme imbalance: the proportion of the labeled group relative to the full data size becomes exceptionally small. However, the decaying MAR provides three representations of the $\gamma_N(j,\cdot)$ function, enhancing flexibility in model formulation, robustness, and theoretical prerequisites. For simplicity, we consider $a_{N,1} \asymp a_{N,0} \asymp a_N \to 0$. The PS function $\gamma_N(\cdot) = \gamma_N(1,\cdot)$ can be represented as

$$\gamma_N(\cdot) = \mathbb{P}(\Gamma = 1 \mid \cdot) \quad \text{or} \quad \gamma_N(\cdot) = \pi(1,\cdot)p_N(\cdot) \quad \text{or} \quad \gamma_N(\cdot) = \pi(\cdot)p_N(1,\cdot).$$

We will explore each case individually, assuming a linear OR model with a slope of sparsity $s_{\boldsymbol{\alpha},1}$, and a Lasso estimate resulting in $c_{N,1} \asymp \sqrt{s_{\alpha,1}\log d/(Na_N)}$ of (3.3).

*Model 1: a logistic $\gamma_N(\cdot)$.* This setting is especially suitable when the treatment and labeling indicators, $T$ and $R$, are affected by the same set of covariates. Here $\gamma_N(\mathbf{x})$ is modeled directly as a logistic function with a *diverging* offset:

$$\gamma_N(\mathbf{x}) = \frac{\gamma_N \exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\gamma_N \exp(\mathbf{x}^T\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}+\log\gamma_N)}{1+\exp(\mathbf{x}^T\boldsymbol{\beta}+\log\gamma_N)}, \ \forall \mathbf{x}\in\mathcal{X}, \tag{3.5}$$

where $\boldsymbol{\beta}\in\mathbb{R}^d$ and the decaying nature of the PS function is captured by the diverging 'offset' term $\log\gamma_N$ where $\gamma_N := \mathbb{E}(\Gamma)\to 0$ as $N$ grows. Zhang, Chakrabortty and Bradic (2023) considered the offset-based logistic model to capture $p_N(\mathbf{X}) = \mathbb{P}(R=1 \mid \mathbf{X})$ in non-causal contexts. Under conditions of Theorem 4.2 and Lemma 4.3 therein, an offset estimator above satisfies Assumptions 8, 9, and 10 with $a_N \asymp \gamma_N$, $d_{N,1}\asymp\sqrt{s_{\beta,1}\log d/(Na_N)}$ (see (3.3)), and $s_{\beta,1} := \|\boldsymbol{\beta}\|_0$. implying the 'product-rate' condition: $s_{\boldsymbol{\alpha},1}s_{\boldsymbol{\beta},1} = o(Na_N/\log^2 d)$.

*Model 2: a logistic $p_N(1,\cdot)$ and a non-parametric $\pi(\cdot)$.* Model 2 is preferable when distinct covariates influence $T$ and $R$, especially if only a fraction affects $R$. We propose a semi-parametric approach, modeling $\gamma_N(\mathbf{x}) = \pi(\mathbf{x})p_N(1,\mathbf{x})$. Here, $\pi(\mathbf{X})$ is non-parametrically representing $P(T=1|\mathbf{X})$, and $p_N(1,\mathbf{X}) = P(R=1|T=1,\mathbf{X})$ is modeled as an offset-based logistic function parameterized by $\boldsymbol{\beta}_{p,1}\in\mathbb{R}^d$:

$$p_N(1,\mathbf{x}) = \frac{p_{N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{p,1})}{1+p_{N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{p,1})}, \ \forall \mathbf{x}\in\mathcal{X}, \tag{3.6}$$

where $p_{N,1} := \mathbb{E}\{p_N(1,\mathbf{X})\}\asymp\gamma_N$ decays to zero with growing $N$. In this semi-parametric approach, we exploit the entire dataset, including both labeled and unlabeled samples denoted as $N$, to estimate $\pi(\cdot)$. In contrast, the estimation error of $p_N(1,\cdot)$ is controlled by the rate $Na_N$, considerably smaller than $N$ as per the decaying MAR with $a_N\to 0$ as $N$ grows. For example, an optimally tuned random forest yields an estimation error $O_p(N^{-1/(11.1\alpha_1\vee 16.1)})$ for some $\alpha_1>0$ (Chi et al., 2022). With the estimation error of $p_N(1,\cdot)$ being $O_p(\sqrt{s_{p,1}\log d/(Na_N)})$, with $s_{p,1} = \|\boldsymbol{\beta}_{p,1}\|_0$, whenever $a_N \ll s_{p,1}N^{2/(11.1\alpha_1\vee 16.1)-1}\log d$, the non-parametric error of $\pi(\cdot)$ is of smaller order than $p_N(1,\cdot)$, resulting in a 'product-rate' condition of $s_{\boldsymbol{\alpha},1}s_{p,1} = o(Na_N/\log^2 d)$ arises. In Section 4.3 below, we will revisit and refine the approach here and provide ATE inference even under model misspecification along with weaker sparsity conditions.

*Model 3: a logistic $p_N(\cdot)$ and a logistic $\pi(1,\cdot)$.* In the data integration context, we generalize causal inference from randomized trials to a broader population containing both clinical trial data, where outcome $Y$ is observed, and non-randomized data with missing $Y$. Randomization indicator $R$ is confounded by $\mathbf{X}$, and treatment $T$ is often assigned after $R$. We propose alternative identification assumptions: $R \perp\!\!\!\perp \{Y(0),Y(1)\}\mid\mathbf{X}$ and $T\perp\!\!\!\perp\{Y(0),Y(1)\}\mid(R,\mathbf{X})$, aligned with Dahabreh et al. (2019). Our framework accommodates diminishing trial participation, allowing $p_N(\mathbf{x})\to 0$ as $N\to\infty$. Lemma 2.1 and Theorem A.2 continue to apply, ensuring the validity of the DR-DMAR SS estimator, including its asymptotic normality. Here, we model $\gamma_N(\cdot) = p_N(\cdot)\pi(1,\cdot)$, where $p_N(\mathbf{x}) = \mathbb{P}(R=1\mid\mathbf{X}=\mathbf{x})$ and $\pi(1,\mathbf{x}) = \mathbb{P}(T=1\mid R=1,\mathbf{X}=\mathbf{x})$ follow offset-based and standard logistic models, respectively,

$$p_N(\mathbf{x}) = \frac{p_N\exp(\mathbf{x}^T\boldsymbol{\beta}_p)}{1+p_N\exp(\mathbf{x}^T\boldsymbol{\beta}_p)}, \qquad \pi(1,\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_{\pi,1})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta}_{\pi,1})}, \ \forall\mathbf{x}\in\mathcal{X},$$

with $\boldsymbol{\beta}_p,\boldsymbol{\beta}_{\pi,1}\in\mathbb{R}^d$ and $s_p := \|\boldsymbol{\beta}_p\|_0$ and $s_{\pi,1} := \|\boldsymbol{\beta}_{\pi,1}\|_0$. The decaying nature is captured in $p_N := E(R)\to 0$ as $N\to\infty$. The product estimate $\widehat{\gamma}_N(\cdot) = \widehat{p}_N(\cdot)\widehat{\pi}(1,\cdot)$ yields

$$\mathbb{E}_{\mathbf{X}}\left\{1-\frac{\gamma_N(\mathbf{X})}{\widehat{\gamma}_N(\mathbf{X})}\right\}^2 = \mathbb{E}_{\mathbf{X}}\left[1-\frac{p_N(\mathbf{X})}{\widehat{p}_N(\mathbf{X})}+1-\frac{\pi(1,\mathbf{X})}{\widehat{\pi}(1,\mathbf{X})}-\left\{1-\frac{p_N(\mathbf{X})}{\widehat{p}_N(\mathbf{X})}\right\}\left\{1-\frac{\pi(1,\mathbf{X})}{\widehat{\pi}(1,\mathbf{X})}\right\}\right]^2$$

$$= O_p\left(\|\widehat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p\|_2^2 + \|\widehat{\boldsymbol{\beta}}_{\pi,1} - \boldsymbol{\beta}_{\pi,1}\|_2^2\right) = O_p\left(\frac{(s_p + s_{\pi,1})\log d}{Na_N}\right),$$

if $\mathbf{X}$ is sub-Gaussian, where one can leverage existing results to obtain $\|\widehat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p\|_2 = O_p(\sqrt{s_p \log d/(Na_N)})$ and $\|\widehat{\boldsymbol{\beta}}_{\pi,1} - \boldsymbol{\beta}_{\pi,1}\|_2 = O_p(\sqrt{s_{\pi,1}\log d/(Na_N)})$. Then, $d_{N,1} \asymp \sqrt{(s_p + s_{\pi,1})\log d/(Na_N)}$ and the product-rate condition is $s_{\boldsymbol{\alpha},1}(s_{\pi,1}+s_p) = o(Na_N/\log^2 d)$.

3.3. *Missing treatment settings.* Our methods and findings extend to cases with severe missingness in the treatment $T$ across four different setups. In setting a (Missing outcome), we observe $R, T, RY, \mathbf{X}$. In setting b (Missing treatment), $R, RT, Y, \mathbf{X}$, where $R$ indicates $T$ observation. In setting c (Simultaneously missing treatment and outcome), $R, RT, RY, \mathbf{X}$, with $R$ signifying observation of both $(T, Y)$. In Setting d (Non-simultaneously missing treatment and outcome), $R_T, R_Y, R_T T, R_Y Y, \mathbf{X}$, where $R_T$ and $R_Y$ signify the observation of $T$ and $Y$, and $R := R_T R_Y$ is their simultaneous observation indicator. A uniform approach for identifying causal effects across these settings is provided in the following corollary.

COROLLARY 3.1. *Let Assumption 1 hold with $R := R_T R_Y$. Define $\Gamma := TR = TR_T R_Y$, $\gamma_N(\mathbf{x}) := \mathbb{P}(\Gamma = 1 \mid \mathbf{X} = \mathbf{x})$, and $m(\mathbf{x}) := \mathbb{E}\{Y(1) \mid \mathbf{X}\}$. Then, $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, \Gamma = 1)$. Additionally, for any arbitrary functions $m^*(\cdot)$ and $\gamma_N^*(\cdot)$, we have (2.1) holds as long as either $m^*(\cdot) = m(\cdot)$ or $\gamma_N^*(\cdot) = \gamma_N(\cdot)$.*

Under Setting d (and consequently, under Settings b and c), both $\Gamma = TR_T R_Y$ and $\Gamma Y = TR_T Y R_Y$ are observable. This allows us to estimate the nuisance functions $m(\mathbf{X}) = \mathbb{E}(Y \mid \mathbf{X}, \Gamma = 1) = \mathbb{E}(R_Y Y \mid \mathbf{X}, \Gamma = 1)$ and $\gamma_N(\mathbf{X}) = \mathbb{E}(\Gamma \mid \mathbf{X})$. As a result, the DR-DMAR SS estimator (3.2) remains valid, Theorem A.2 still holds with the newly defined $R$ and $\Gamma$, allowing the asymptotic results (A.7) to remain applicable; check Section B in the Supplement for estimating nuisance functions without complete $T_i$ data. Corollary 3.1 assumes an R-MAR condition, where $R_T R_Y \perp\!\!\!\perp Y \mid (T, \mathbf{X})$. Unlike previous studies that rely on restrictive monotone conditions (Manski, 1997; Manski and Pepper, 2000; Molinari, 2010; Mebane and Poast, 2013), our results avoid such assumptions between treatment and potential outcomes. Another MAR condition considered by Zhang et al. (2016), $R \perp\!\!\!\perp T \mid (\mathbf{X}, Y)$, is generally inappropriate as $Y$ is usually evaluated after the treatment assignment.

**4. Refined DR estimators.** We introduce two DR estimates and their theoretical properties when $d$ is large compared with the 'effective sample size', $Na_N$. We first propose the bias-reduced DR-DMAR SS estimator based on parametric nuisance models in Section 4.1. Section 4.2 explains the rationale behind the construction of the proposed estimator. Then, making full use of the unlabeled samples, we propose a new semi-parametric approach in Section 4.3. Theoretical properties of the proposed nuisance and ATE estimators are provided in Sections 4.4 and 4.5, respectively.

4.1. *DR-DMAR SS (BRSS) ATE estimator.* We improve the estimator from Section 3 by defining refined nuisance estimates to minimize bias. While employing the DR representation (2.1), we introduce an asymmetric cross-fitting. Model 1 from Section 3.2 is used. We define the following *working* models for $m(\cdot)$ and $\gamma_N(\cdot)$ with 'targets':

$$m^*(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\alpha}^* \text{ and } \gamma_N^*(\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\beta}^* + \log\gamma_N) = \frac{\gamma_N \exp(\mathbf{x}^T\boldsymbol{\beta}^*)}{1 + \gamma_N\exp(\mathbf{x}^T\boldsymbol{\beta}^*)}, \tag{4.1}$$

where $g(u) := \exp(u)/\{1 + \exp(u)\}$ is the logistic function, $\gamma_N := \mathbb{E}(\Gamma)$. In the above, $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^* \in \mathbb{R}^d$ are the *targeted bias-reducing nuisance parameters* defined as follows:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\{l(\Gamma, \mathbf{X}, \boldsymbol{\beta}, \gamma_N)\} \quad\text{and}\quad \boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \mathbb{E}\{h(\Gamma, \mathbf{X}, Y, \boldsymbol{\alpha}, \boldsymbol{\beta}^*, \gamma_N)\},$$

with the loss functions $l$ and $h$ defined as

$$l(\Gamma,\mathbf{X},\boldsymbol{\beta},\gamma_N)=(1-\Gamma)\mathbf{X}^T\boldsymbol{\beta}+\frac{\Gamma}{\gamma_N}\exp(-\mathbf{X}^T\boldsymbol{\beta})\quad\text{and}$$

$$h(\Gamma,\mathbf{X},Y,\boldsymbol{\alpha},\boldsymbol{\beta},\gamma_N)=\frac{\Gamma}{\gamma_N}\exp(-\mathbf{X}^T\boldsymbol{\beta})(Y-\mathbf{X}^T\boldsymbol{\alpha})^2.$$

Let $s_{\boldsymbol{\beta}}:=\|\boldsymbol{\beta}^*\|_0$ and $s_{\boldsymbol{\alpha}}:=\|\boldsymbol{\alpha}^*\|_0$ be the corresponding *sparsity* levels, and assume $s_{\boldsymbol{\beta}},s_{\boldsymbol{\alpha}}\geq 1$ for the sake of simplicity. Note that $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ always exist, and also equal the corresponding 'true' model parameters when the working models in (4.1) are correctly specified for $m(\cdot)$ or $\gamma_N(\cdot)$. In what follows, we only require either $m^*(\cdot)=m(\cdot)$ or $\gamma_N^*(\cdot)=\gamma_N(\cdot)$, but *not* necessarily both.

We define *targeted bias-reducing nuisance estimators* as

$$\widehat{\gamma}_N^{(k)} := M^{-1}\sum_{i\in\mathcal{I}_k}\Gamma_i, \tag{4.2}$$

$$\widehat{\boldsymbol{\beta}}^{(k)} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} M^{-1}\sum_{i\in\mathcal{I}_k} l(\Gamma_i,\mathbf{X}_i,\boldsymbol{\beta},\widehat{\gamma}_N^{(k)})+\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1, \tag{4.3}$$

$$\widehat{\boldsymbol{\alpha}}^{(k)} := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} M^{-1}\sum_{i\in\mathcal{I}_k} h(\Gamma_i,\mathbf{X}_i,Y_i,\boldsymbol{\alpha},\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\gamma}_N^{(k)})+\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1, \tag{4.4}$$

where $\lambda_{\boldsymbol{\alpha}},\lambda_{\boldsymbol{\beta}}\geq 0$ denote the respective tuning parameters for the $\ell_1$-regularizations above. To build robust inference for the ATE, Tan (2020); Smucler, Rotnitzky and Robins (2019); Avagyan and Vansteelandt (2021) examined nuisance estimates akin to (4.3) and (4.4) under degenerate supervised settings *only*. We adapt these to the decaying MAR context with our proposed bias-reducing estimators, incorporating an apriori chosen estimate $\widehat{\gamma}_N^{(k)}$ to counterbalance the impact of diminishing PSs. In (4.3), we aggregate two sums: $\sum_{i\in\mathcal{I}_{-k}:\Gamma_i=0}\mathbf{X}_i^T\boldsymbol{\beta}$ and $\sum_{i\in\mathcal{I}_{-k}:\Gamma_i=1}\exp(-\mathbf{X}_i^T\boldsymbol{\beta})/\widehat{\gamma}_N^{(k)}$. The set where $\Gamma_i=0$ predominates over the set where $\Gamma_i=1$. To compensate for this imbalance, we amplify the latter group by a large factor of $1/\widehat{\gamma}_N^{(k)}$, ensuring a balanced influence from both groups.

With a special asymmetric cross-fitting, we propose the *bias-reduced DR-DMAR SS counterfactual mean estimator* $\widehat{\theta}_{1,\mathrm{BRSS}}$ of $\theta_1$ as: $\widehat{\theta}_{1,\mathrm{BRSS}}:=(\widehat{\theta}_{1,\mathrm{BRSS}}^{(1)}+\widehat{\theta}_{1,\mathrm{BRSS}}^{(2)})/2$, where $k\neq k'\in\{1,2\}$,

$$\widehat{\theta}_{1,\mathrm{BRSS}}^{(k)} := M^{-1}\sum_{i\in\mathcal{I}_k}\left\{\mathbf{X}_i^T\widehat{\boldsymbol{\alpha}}^{(k')}+\frac{\Gamma_i(Y_i-\mathbf{X}_i^T\widehat{\boldsymbol{\alpha}}^{(k')})}{\widehat{\gamma}_N^{(k)}(\mathbf{X}_i)}\right\}, \tag{4.5}$$

with $\widehat{\gamma}_N^{(k)}(\mathbf{X}_i)=g(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}^{(k)}+\log\widehat{\gamma}_N^{(k)})$. Analogously, we propose a bias-reduced estimator $\widehat{\theta}_{0,\mathrm{BRSS}}$ of $\theta_0$, and the *bias-reduced DR-DMAR SS (BRSS) ATE estimator* as:

$$\widehat{\mu}_{\mathrm{BRSS}} := \widehat{\theta}_{1,\mathrm{BRSS}}-\widehat{\theta}_{0,\mathrm{BRSS}}. \tag{4.6}$$

4.2. *The asymmetric cross-fitting.* In the following, we introduce the rationale behind the asymmetric cross-fitting strategy focusing on the simplest case where both the OR and PS models are correctly specified. For $i\in\mathcal{I}_1$, the PS function $\gamma_N(\mathbf{X}_i)$ is estimated using a non-cross-fitted $\widehat{\boldsymbol{\beta}}^{(1)}\not\perp\!\!\!\perp\mathbf{Z}_i$, whereas the OR function $m(\mathbf{X}_i)$ is estimated using a cross-fitted $\widehat{\boldsymbol{\alpha}}^{(2)}\perp\!\!\!\perp\mathbf{Z}_i$. W.l.o.g., we consider $\widehat{\theta}_{1,\mathrm{BRSS}}^{(1)} = M^{-1}\sum_{i\in\mathcal{I}_1}\psi_{N,1}^*(\mathbf{Z})+\Delta_1+\Delta_2+\Delta_3$, where $\psi_{N,1}^*(\mathbf{Z})\equiv\psi_{N,1}(\mathbf{Z};\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*):=m^*(\mathbf{X})-\theta_1+\Gamma\{Y-m^*(\mathbf{X})\}/\gamma_N^*(\mathbf{X})$, and

$$\Delta_1 := -M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widehat{\gamma}_N^{(1)}(\mathbf{X}_i)}-\frac{\Gamma_i}{\gamma_N^*(\mathbf{X}_i)}\right\}\mathbf{X}_i^T(\widehat{\boldsymbol{\alpha}}^{(2)}-\boldsymbol{\alpha}^*),$$

$$\Delta_2 := M^{-1} \sum_{i \in \mathcal{I}_1} \left\{ 1 - \frac{\Gamma_i}{\gamma_N^*(\mathbf{X}_i)} \right\} \mathbf{X}_i^T (\widehat{\boldsymbol{\alpha}}^{(2)} - \boldsymbol{\alpha}^*),$$

$$\Delta_3 := M^{-1} \sum_{i \in \mathcal{I}_1} \left\{ \frac{\Gamma_i}{\widehat{\gamma}_N^{(1)}(\mathbf{X}_i)} - \frac{\Gamma_i}{\gamma_N^*(\mathbf{X}_i)} \right\} (Y_i - \mathbf{X}_i^T \boldsymbol{\alpha}^*).$$

The 'drift term' $\Delta_1$ in the influence function $\psi_{N,1}(\mathbf{Z}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ represents estimation error linked to the product of two nuisance parameters. A condition involving sparsity in this product is sufficient to minimize this error. To demonstrate the doubly robust (DR) property, it's crucial to manage the expectations of the remaining terms, $\Delta_2$ and $\Delta_3$. Our asymmetric cross-fitting approach controls the former term, sometimes merging it with $\Delta_1$, depending on active sparsity conditions. In contrast, we use in-sample control through the Karush-Kuhn-Tucker (KKT) condition to handle the latter term. The asymmetry of our approach is that the propensity score (PS) estimator does not require cross-fitting. This leads to less stringent sparsity conditions for valid inference, as demonstrated in Theorems 4.4 and 4.5 below.

Our proposed approach innovatively combines the strengths of existing strategies to enhance robustness for ATE inference. We draw on the methodology of Farrell (2015); Tan (2020); Avagyan and Vansteelandt (2021) for non-cross-fitted PS estimates and the cross-fitted OR estimate approach of Chernozhukov et al. (2018); Smucler, Rotnitzky and Robins (2019). As opposed to previous works, our method relaxes the ultra-sparsity requirements on both $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ by introducing asymmetric cross-fitting. This novel combination provides superior robustness under degenerate supervised settings and requires weaker sparsity conditions. To our knowledge, only Bradic, Wager and Zhu (2019) employs cross-fitting akin to ours. Our method proves easier to implement, provides ATE inference even with a misspecified nuisance model – unlike their requirement for both models to be accurate – and under correctly specified models, demands weaker sparsity conditions for valid inference. Detailed comparisons can be found in Table 4.1 and Remark 3.

4.3. *A semi-parametric bias-reduced DR-DMAR SS (SP-BRSS) estimator.* Here, we introduce a semi-parametric model. Unlike the BRSS method, which combines missingness patterns of $R$ and $T$ as $\Gamma = R \cdot T$, our approach separates them, concentrating on $R$ directly and allowing non-parametric modeling of $T$. This transition offers numerous benefits. First, it enriches the complexity and robustness of the model, providing a clearer understanding of the involved components. Secondly, fully utilizing all $N$ of the $(T_i, \mathbf{X}_i)$ pairs enables accurate estimation of $\pi(\mathbf{x}) = \mathbb{P}(T = 1 \mid \mathbf{X} = \mathbf{x})$, thereby improving overall estimation accuracy. Third, non-parametric treatment modeling for $T$ allows greater flexibility, capturing complex treatment effects with precision. Simultaneously, focusing on $R$ enhances method robustness by addressing the unique challenge of missingness in the outcome variable, ultimately yielding more reliable causal inference, particularly in complex dependency scenarios.

The semi-parametric approach posit the following working models,

$$\widetilde{m}^*(\mathbf{x}) = \mathbf{x}^T \widetilde{\boldsymbol{\alpha}}^*, \;\; \widetilde{\gamma}_N^*(\mathbf{x}) = \pi^*(\mathbf{x}) p_{N,1}^*(\mathbf{x}), \;\; \text{and} \;\; p_{N,1}^*(\mathbf{x}) = g(\mathbf{x}^T \boldsymbol{\beta}_{p,1}^* + \log p_{N,1}).$$

where $\pi^*(\cdot)$ is (possibly) a non-parametric model of $\pi(\cdot)$ and $p_{N,1}^*(\cdot)$ is an offset-based logistic model of $p_{N,1}(\cdot) := p_N(1, \cdot)$. The new (target) nuisance parameters $\widetilde{\boldsymbol{\alpha}}^*, \boldsymbol{\beta}_{p,1}^* \in \mathbb{R}^d$ use novel-reparametrization together with the loss functions of Section 4.1 and are defined as

$$\boldsymbol{\beta}_{p,1}^* = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathbb{E} \left\{ l\Big( \frac{\Gamma}{\pi^*(\mathbf{X})}, \mathbf{X}, \boldsymbol{\beta}, p_{N,1} \Big) \right\} \quad \text{and}$$

$$\widetilde{\boldsymbol{\alpha}}^* = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \mathbb{E} \Big\{ h\Big( \frac{\Gamma}{\pi^*(\mathbf{X})}, \mathbf{X}, Y, \boldsymbol{\alpha}, \boldsymbol{\beta}_{p,1}^*, p_{N,1} \Big) \Big\}.$$

Let $s_{p,1} := \|\boldsymbol{\beta}^*_{p,1}\|_0$ and $s_{\widetilde{\boldsymbol{\alpha}}} := \|\widetilde{\boldsymbol{\alpha}}^*\|_0$ be the corresponding sparsity levels, and assume $s_{p,1}, s_{\widetilde{\boldsymbol{\alpha}}} \geq 1$ for the sake of simplicity..

*The semi-parametric bias-reduced SS ATE estimator.* Let $\widehat{\pi}^{(1)}(\cdot)$ and $\widehat{\pi}^{(2)}(\cdot)$ be (non-parametric) estimates of $\pi(\cdot)$, using samples $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. For $k \in \{1,2\}$, let

$$\begin{aligned}
\widehat{p}^{(k)}_{N,1} &= \sum_{i\in\mathcal{I}_k}\Gamma_i / \sum_{i\in\mathcal{I}_k} T_i,\\
\widehat{\boldsymbol{\beta}}^{(k)}_{p,1} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} M^{-1}\sum_{i\in\mathcal{I}_k} l\Big(\frac{\Gamma_i}{\widehat{\pi}^{(k)}(\mathbf{X}_i)}, \mathbf{X}_i, \boldsymbol{\beta}, \widehat{p}^{(k)}_{N,1}\Big) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1, \quad \text{and}\\
\widetilde{\boldsymbol{\alpha}}^{(k)} &= \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} M^{-1}\sum_{i\in\mathcal{I}_k} h\Big(\frac{\Gamma_i}{\widehat{\pi}^{(k)}(\mathbf{X}_i)}, \mathbf{X}_i, Y_i, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}^{(k)}_{p,1}, \widehat{p}^{(k)}_{N,1}\Big) + \lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1,
\end{aligned}$$

where with a slight abuse of notations, $\lambda_{\boldsymbol{\alpha}}, \lambda_{\boldsymbol{\beta}} \geq 0$ denote the tuning parameters. Then, the *semi-parametric bias-reduced SS counterfactual mean estimator* $\widehat{\theta}_{1,\text{SP-BRSS}}$ of $\theta_1$ is $\widehat{\theta}_{1,\text{SP-BRSS}} := (\widehat{\theta}^{(1)}_{1,\text{SP-BRSS}} + \widehat{\theta}^{(2)}_{1,\text{SP-BRSS}})/2$, where for any $k \neq k' \in \{1,2\}$,

$$\widehat{\theta}^{(k)}_{1,\text{SP-BRSS}} := M^{-1}\sum_{i\in\mathcal{I}_k}\left\{\mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(k')} + \frac{\Gamma_i(Y_i - \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(k')})}{\widehat{p}^{(k)}_{N,1}(\mathbf{X}_i)\widehat{\pi}^{(k)}(\mathbf{X}_i)}\right\},$$

with $\widehat{p}^{(k)}_{N,1}(\mathbf{X}_i) = g(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}^{(k)}_{p,1} + \log\widehat{p}^{(k)}_{N,1})$. Analogously, we define $\widehat{\theta}_{0,\text{SP-BRSS}}$ and the ATE estimator $\widehat{\mu}_{\text{SP-BRSS}} := \widehat{\theta}_{1,\text{SP-BRSS}} - \widehat{\theta}_{0,\text{SP-BRSS}}$. The variance estimator is also considered

(4.7)
$$\widehat{\Sigma}_{1,\text{SP-BRSS}} := N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\left\{\mathbf{X}_i^T\overline{\boldsymbol{\alpha}} + \frac{\Gamma_i(Y_i - \mathbf{X}_i^T\overline{\boldsymbol{\alpha}})}{\widehat{\pi}^{(k)}(\mathbf{X}_i)g(\mathbf{X}_i^T\bar{\boldsymbol{\beta}} + \log\bar{p}_{N,1})} - \widehat{\theta}_{1,\text{SP-BRSS}}\right\}^2, \text{ with}$$

$$\overline{\boldsymbol{\alpha}} := \frac{\widetilde{\boldsymbol{\alpha}}^{(1)} + \widetilde{\boldsymbol{\alpha}}^{(2)}}{2}, \ \bar{\boldsymbol{\beta}}_{p,1} := \frac{\widehat{\boldsymbol{\beta}}^{(1)}_{p,1} + \widehat{\boldsymbol{\beta}}^{(2)}_{p,1}}{2}, \text{ and } \bar{p}_{N,1} := \frac{\widehat{p}^{(1)}_{N,1} + \widehat{p}^{(2)}_{N,1}}{2}.$$

4.4. *Theoretical properties of the targeted bias-reducing nuisance estimators.* Theorems 4.1 - 4.3 discuss adaptive nuisances and should be of independent interest – owing to the non-parametric initial step, the high dimensionality, and the decaying PS setting.

ASSUMPTION 2 (Sub-Gaussian covariates). Assume that $\mathbf{X}$ is a sub-Gaussian random vector with $\|\mathbf{X}^T\mathbf{v}\|_{\psi_2} \leq \sigma\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^d$ and $\|\mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^*\|_{\psi_2} \leq \sigma$, for some constant $\sigma > 0$. Additionally, for some constant $\kappa_l > 0$, assume that $\inf_{\|\mathbf{v}\|_2=1}\mathbb{E}\{\Gamma(\mathbf{X}^T\mathbf{v})^2\} \geq \gamma_N\kappa_l$.

ASSUMPTION 3 (Moment condition on the PS). There exists some $q > 1$, such that $\mathbb{E}\{\gamma_N^q(\mathbf{X})\} \leq C\gamma_N^q$, for some constant $C > 0$.

ASSUMPTION 4 (Non-parametric estimation of $\pi(\cdot)$). With some constant $c > 0$, let: (a) $c < \pi^*(\mathbf{x}) < 1 - c$ for all $\mathbf{x} \in \mathcal{X}$ and (b) the events $\mathcal{E}_\pi := \{c < \widehat{\pi}^{(k)}(\mathbf{x}) < 1 - c, \forall k \in \{1,2\}, \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{E}_\zeta := \{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_k}\Gamma_i\{\widehat{\pi}^{(k)}(\mathbf{X}_i) - \pi^*(\mathbf{X}_i)\}^2 \leq \zeta_N^2, \forall k \in \{1,2\}\}$ occur with probability approaching one as $N \to \infty$ with some $\zeta_N = o(1/\sqrt{\log N})$.

ASSUMPTION 5 (Bounded covariates and coefficients). Let: (a) $\|\mathbf{x}\|_\infty \leq C_{\mathbf{X}}$ for all $\mathbf{x} \in \mathcal{X}$ and (b) $\|\boldsymbol{\beta}^*_{p,1}\|_1 \leq C_{\boldsymbol{\beta}}$, for some constants $C_{\mathbf{X}}, C_{\boldsymbol{\beta}} > 0$.

If Assumption 2 holds and $p^*_{N,1}(\cdot) = p_{N,1}(\cdot)$ then Assumption 3 holds for any constant $q > 1$, although we only need it for some $q > 1$. Conditions similar to Assumption 4 (a) appear in Tan (2020); Ning, Sida and Imai (2020) as 'overlap conditions'. Moreover, Assumption 4 (b) can be simplified to an in sample mean squared error of $\widehat{\pi}^{(k)}$ as long as Assumption 3 holds for $q = \infty$, i.e., $\gamma_N(\mathbf{X})/\gamma_N$ is bounded almost surely. Assumption 5 (b), is an implication of a T-overlap condition. Consider $R \equiv 1$ and let $\widetilde{\gamma}^*_N(\cdot)$ be correctly specified. Then, the T-overlap implies $|\mathbf{x}^T\boldsymbol{\beta}^*_{p,1}| < c'$ almost surely with $c' > 0$. This in turn implies $\|\boldsymbol{\beta}^*_{p,1}\|_1 \leq \infty$, provided that the marginal distribution of $\mathbf{X}$ fulfills: $\mathbb{P}(\mathbf{X}^T\mathbf{e}_j < -c_1) > c_2$ and $\mathbb{P}(\mathbf{X}^T\mathbf{e}_j > c_1) > c_2$, $\forall j \leq d$, with any $c_1, c_2 > 0$. These inequalities hold for example when the marginal distribution of $\mathbf{X}$ is Uniform with mean zero and bounded (possibly different) supports. In general decaying MAR, a bounded $\|\boldsymbol{\beta}^*_{p,1}\|_1$ is implied by the same condition together with the ratio $\gamma_N(\mathbf{X})/\gamma_N$ being bounded almost surely.

THEOREM 4.1 (PS estimator). *Let Assumptions 1-5 hold. If $N\gamma_N \gg \max(\log N, s_{p,1})\log d$, then with some $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log d/(N\gamma_N)}$, as $N, d \to \infty$,*

$$\|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_2 = O_p\Big(\sqrt{\frac{s_{p,1}\log d}{N\gamma_N}} + \zeta_N\Big),\ \|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_1 = O_p\Big(s_{p,1}\sqrt{\frac{\log d}{N\gamma_N}} + \zeta_N^2\sqrt{\frac{N\gamma_N}{\log d}}\Big).$$

*Moreover, if $\zeta_N = o(\{\log d/(N\gamma_N)\}^{1/4})$ then as $N \to \infty$, $P\Big(\mathcal{E}_{\widehat{\boldsymbol{\beta}}} := \{\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 \leq 8C_{\boldsymbol{\beta}}\}\Big) \to 1$.*

The estimation of $\widehat{\pi}(\cdot)$ shifts $\widehat{\boldsymbol{\beta}}_{p,1}$ away from the conventional sparse cone set, $\mathcal{C}(s_{p,1}, k_0)$. This set is typically described as: $\mathcal{C}(s_{p,1}, k_0) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \exists S \subseteq \{1, \ldots, d\}$ with $|S| \leq s_{p,1}$ and $\|\boldsymbol{\Delta}_{S^c}\|_1 \leq k_0\|\boldsymbol{\Delta}_S\|_1\}$. Due to this deviation, we cannot rely on standard techniques to prove estimation quality. Our solution involves a more encompassing cone set, $\widetilde{\mathcal{C}}(r_N)$: $\widetilde{\mathcal{C}}(r_N) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \leq r_N\|\boldsymbol{\Delta}\|_2\}$, where $r_N$ is of the order of $\sqrt{s_{p,1}} + \zeta_N\sqrt{N\gamma_N/\log d}$. More details can be found in (F.18) of the Supplement. Adopting this broader cone set necessitated new restricted strong convexity properties, which cater to this extended sparsity framework. Additionally, it required us to develop uniform bounds for quadratic forms, like $|\boldsymbol{\Delta}^T\boldsymbol{\Omega}\boldsymbol{\Delta}|$, as well as uniform gradient control over what are now more complex sets. For further insights, see Lemmas F.4-F.7 in the Supplement. Importantly, these new techniques are of potential interest for other estimators that deviate from traditional cone set constraints. Notably, achieving an 'effective sample size' measure as $N\gamma_N$ depends crucially on ensuring the adaptability to $\gamma_N$.

THEOREM 4.2 (Correctly specified OR). *Let $m(\mathbf{x}) = m^*(\mathbf{x}) = \mathbf{x}^T\widetilde{\boldsymbol{\alpha}}^*$ for all $\mathbf{x} \in \mathcal{X}$. Let Assumptions 1-5 hold. Then, for*

$$N\gamma_N \gg \max(\log N, s_{p,1}, s_{\widetilde{\boldsymbol{\alpha}}})\log d \quad \textit{and} \quad \zeta_N = o(\{\log d/(N\gamma_N)\}^{1/4}), \tag{4.8}$$

*with some $\lambda_{\boldsymbol{\beta}} \asymp \lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/(N\gamma_N)}$, we have as $N, d \to \infty$,*

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_2 = O_p\left(\sqrt{\frac{s_{\widetilde{\boldsymbol{\alpha}}}\log d}{N\gamma_N}}\right), \quad \|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_1 = O_p\left(s_{\widetilde{\boldsymbol{\alpha}}}\sqrt{\frac{\log d}{N\gamma_N}}\right). \tag{4.9}$$

If the OR model is misspecified, we assume the following additional condition.

ASSUMPTION 6 (Conditional sub-Gaussian noise). Let, conditional on $\mathbf{X}$, $\varepsilon := Y(1) - \mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^*$ be a sub-Gaussian random variable with a constant $\psi_2$-norm $\sigma_\varepsilon > 0$.

THEOREM 4.3 (Misspecified OR). *Consider the general case of (possibly) $m(\mathbf{x}) \neq m^*(\mathbf{x}) = \mathbf{x}^T \widetilde{\boldsymbol{\alpha}}^*$. Let Assumptions 1-6 hold. Then, as long as (4.8) holds, with $\lambda_{\boldsymbol{\beta}} \asymp \lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/(N\gamma_N)}$ and $\tilde{s} = s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1}$, we have as $N, d \to \infty$,*

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_2 = O_p\left(\sqrt{\frac{\tilde{s}\log d}{N\gamma_N}} + \zeta_N\right), \ \|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_1 = O_p\left(\tilde{s}\sqrt{\frac{\log d}{N\gamma_N}} + \zeta_N^2\sqrt{\frac{N\gamma_N}{\log d}}\right).$$

For $R \equiv 1$ and $\zeta_N = 0$, Tan (2020); Smucler, Rotnitzky and Robins (2019) showed the consistency rate is influenced by both $s_{\boldsymbol{\alpha}}$ and $s_{\boldsymbol{\beta}}$, as in the above. In contrast, Bradic, Wager and Zhu (2019) found the rate only depends on $s_{\boldsymbol{\alpha}}$, but always assume accurate PS and OR, $\gamma_N(\cdot) = \widetilde{\gamma}_N^*(\cdot)$ and $m(\cdot) = \widetilde{m}^*(\cdot)$. Our study adopts a distinctive approach. Instead of relying on out-of-sample $\widehat{\boldsymbol{\beta}}_{p,1}$, we utilize in-sample estimates. This choice allows us to fully exploit the dataset for enhanced efficiency, achieve reduced variance, and simplify the complexities of data partitioning. However, this approach also brings forward unique challenges. For instance, we handle the intricate dependencies of imputation errors and our training samples, leading to dependencies that are not straightforward. Yet, we achieve rates that are adaptive to the correctness of the OR model. On another front, we've integrated an innovative nuisance estimation step centered around $\pi(\cdot)$ and achieved rates independent of the correctness of such estimates.

4.5. *Asymptotic properties of the bias-reduced DR-DMAR SS estimators.* In this section, we provide theoretical properties of the semi-parametric approach while also providing corollaries for the BRSS estimator.

THEOREM 4.4 (Correctly specified OR but not PS). *Let the OR model be correctly specified, i.e., $m(\cdot) = \widetilde{m}^*(\cdot)$. Let Assumptions 1- 6 hold. Let $N\gamma_N \gg \log(d \vee N)\log N$ and $s_{p,1} + s_{\widetilde{\boldsymbol{\alpha}}} = o(N\gamma_N/(\log d \log^{1/2} N))$. Let either one of the following hold:*

*(a)* $s_{p,1} = o\left(\frac{\sqrt{N\gamma_N}}{\log d}\right), \ \frac{s_{\widetilde{\boldsymbol{\alpha}}}\log d\left(\sqrt{s_{p,1}\log d} \wedge s_{p,1}\log d\right)}{N\gamma_N} = o(1), \ \zeta_N = o\left(\frac{1}{\sqrt{N\gamma_N}}\right);$

*(b)* $s_{\widetilde{\boldsymbol{\alpha}}} = o\left(\frac{\sqrt{N\gamma_N}}{\log d}\right), \ s_{p,1} = o\left(\frac{N\gamma_N}{\log d}\right), \ \zeta_N = o\left(\left(\frac{\log d}{N\gamma_N}\right)^{1/4}\right).$

*Then, with some $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log d/(N\gamma_N)}$ and $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log d/(N\gamma_N)}$, as $N, d \to \infty$,*

$$\widehat{\theta}_{1,\text{SP-BRSS}} - \theta_1 \ = \ N^{-1}\sum_{i\in\mathcal{I}}\widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) + o_p\left((N\gamma_N)^{-1/2}\right), \tag{4.10}$$

*where $\widetilde{\psi}_{N,1}^*(\mathbf{Z}) := \widetilde{m}^*(\mathbf{X}) - \theta_1 + \Gamma\{Y - \widetilde{m}^*(\mathbf{X})\}/\widetilde{\gamma}_N^*(\mathbf{X})$ with $\mathbb{E}\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\} = 0$, $\widetilde{\Sigma}_{N,1}^* := Var\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\} = O(\gamma_N^{-1})$, and $N^{-1}\sum_{i\in\mathcal{I}}\widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) = O_p((N\gamma_N)^{-1/2})$. If further Assumption 7 holds, then $\widetilde{\Sigma}_{N,1}^* \asymp \gamma_N^{-1} \asymp a_{N,1}^{-1}$, and as $N, d \to \infty$,*

$$\sqrt{N}\left(\widetilde{\Sigma}_{N,1}^*\right)^{-1/2}\left(\widehat{\theta}_{1,\text{SP-BRSS}} - \theta_1\right) \xrightarrow{d} \mathcal{N}(0,1), \tag{4.11}$$

$$\widehat{\Sigma}_{1,\text{SP-BRSS}} = \widetilde{\Sigma}_{N,1}^*\{1 + o_p(1)\}, \ \textit{and} \ \sqrt{N}\left(\widehat{\Sigma}_{1,\text{SP-BRSS}}\right)^{-1/2}\left(\widehat{\theta}_{1,\text{SP-BRSS}} - \theta_1\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{4.12}$$

For misspecified OR models, the following condition is further assumed.

ASSUMPTION 7 (Lower bound). Let $\mathbb{E}(\varepsilon^2 \mid \mathbf{X}) \geq c_{\min}$ almost surely.

THEOREM 4.5 (Correctly specified PS but not OR). *Let the PS model be correctly specified, i.e., $\gamma_N(\mathbf{x}) = \widetilde{\gamma}_N^*(\mathbf{x})$. Let Assumptions 1- 7 hold. Let $N\gamma_N \gg \log(d \vee N)\log N$ and $s_{p,1} + s_{\widetilde{\boldsymbol{\alpha}}} = o(N\gamma_N/(\log d \log^{1/2} N))$. If Condition (a) of Theorem 4.4 holds, then all of the conclusions of Theorem 4.4 also hold:* (4.10)-(4.12).

REMARK 2 (Robust inference). As shown in Theorems 4.4 and 4.5 above, the estimator $\widehat{\theta}_{1,\text{SP-BRSS}}$ is *model DR*. In Theorem 4.5 above, we only require the product PS model $\widetilde{\gamma}_N^*(\cdot) = \pi^*(\cdot)p_{N,1}^*(\cdot)$ to be correctly specified, which occurs if and only if $\gamma_N(\cdot)/\pi^*(\cdot)$ is logistic, i.e., there exists $\boldsymbol{\beta}_{p,1}^0$, such that $\gamma_N(\mathbf{x})/\pi^*(\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\beta}_{p,1}^0 + \log p_{N,1})$. The treatment PS model $\pi^*(\cdot)$ can be parametric or non-parametric – all we need is it satisfies the 'high-level' conditions assumed in Theorems 4.4 and 4.5. Moreover, the treatment PS is also not necessarily well specified, and $\pi(\cdot) \neq \pi^*(\cdot)$ is allowed. We can therefore use many dimension reduction and non-parametric methods to estimate $\pi^*(\cdot)$ and utilize the full sample size $N$. If for instance we shrink the dimension to some $s < d$, when $\pi^*(\cdot)$ is in a Hölder class with parameter $\alpha > 0$, $\pi^*(\cdot)$ can be estimated through kernel methods with $\zeta_N \asymp N^{-\alpha/(2\alpha+s)}$. Theorem 4.4 Conditions (a) and (b) are achieved for $\gamma_N = o(N^{-s/(2\alpha+s)})$ and $\gamma_N = o(N^{(2\alpha-s)/(2\alpha+s)}\log d)$, respectively, with the later holding for $\alpha \geq s/2$ even when $\gamma_N \asymp 1$. We can also directly implement non-parametric methods without dimension reduction techniques. For instance, with a diverging number of covariates satisfying $d = O(N^c)$ ($c > 0$ is a constant), Chi et al. (2022) showed that a random forest estimate leads to an estimation error $\zeta_N \asymp N^{-1/(11.1\alpha_1 \vee 16.1)}$, where $\alpha_1 > 0$ is the 'sufficient impurity decrease' parameter; see Condition 1 therein. Hence, Conditions (a) and (b) above are reached when $\gamma_N = o(N^{2/(11.1\alpha_1 \vee 16.1)-1})$ and $\gamma_N = o(N^{4/(11.1\alpha_1 \vee 16.1)-1}\log d)$, respectively.

As we allow $\pi(\cdot) \neq \pi^*(\cdot)$, we can set $\widehat{\pi}^{(k)}(\cdot) \equiv \pi^*(\cdot) \equiv 1$, and the estimator $\widehat{\theta}_{1,\text{SP-BRSS}}$ will degenerates to the parametric version $\widehat{\theta}_{1,\text{BRSS}}$, (4.5), apart from difference between the offset terms $\log\widehat{\gamma}_N^{(k)}$ and $\log\widehat{p}_{N,1}^{(k)}$. Whenever $\pi^*(\cdot) \equiv 1$, we have $\widetilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$, $\boldsymbol{\beta}_{p,1}^* = \boldsymbol{\beta}^*$, and the following result for the BRSS estimator.

COROLLARY 4.6 (BRSS). *Let Assumptions 1- 3 and 5- 7 hold with $\widetilde{\boldsymbol{\alpha}}^* = \boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}_{p,1}^* = \boldsymbol{\beta}^*$. Let $N\gamma_N \gg \log(d \vee N)\log N$ and $s_{\boldsymbol{\beta}} + s_{\boldsymbol{\alpha}} = o(N\gamma_N/(\log d \log^{1/2} N))$. Let either (i) or (ii) hold with (i) $m(\cdot) = m^*(\cdot)$ and either (a) or (b) below holds, and (ii) $\gamma_N(\cdot) = \gamma_N^*(\cdot)$ and (a) below holds, where*

$$(a)\ s_{\boldsymbol{\beta}} = o\left(\frac{\sqrt{N\gamma_N}}{\log d}\right),\ \frac{s_{\boldsymbol{\alpha}}\log d\left(\sqrt{s_{\boldsymbol{\beta}}\log d} \wedge s_{\boldsymbol{\beta}}\log d\right)}{N\gamma_N} = o(1);$$

$$(b)\ s_{\boldsymbol{\alpha}} = o\left(\frac{\sqrt{N\gamma_N}}{\log d}\right),\ s_{\boldsymbol{\beta}} = o\left(\frac{N\gamma_N}{\log d}\right).$$

*With $\psi_{N,1}^*(\mathbf{Z}) := m^*(\mathbf{X}) - \theta_1 + \Gamma\{Y - m(\mathbf{X})\}/\gamma_N^*(\mathbf{X})$, as $N, d \to \infty$, $\Sigma_{N,1}^* := Var\{\psi_{N,1}^*(\mathbf{Z})\} = O(\gamma_N^{-1})$ and*

$$\sqrt{N}\left(\Sigma_{N,1}^*\right)^{-1/2}\left(\widehat{\theta}_{1,BRSS} - \theta_1\right) \xrightarrow{d} \mathcal{N}(0,1).$$

REMARK 3 (Comparative Analysis). We detail comparisons with existing literature.

*Decaying MAR causal.* Zhang, Chakrabortty and Bradic (2023) provided *rate DR* results only for the estimation of the non-causal mean response, equivalent to estimating $\theta_1 := \mathbb{E}\{Y(1)\}$ in causal scenarios when $T \equiv 1$. In contrast, our method offers superior robustness with *model DR* results and weaker sparsity conditions achieving the *sparsity DR*. While Kallus and Mao (2020) introduced a SS estimator addressing selection bias, they did not allow $N \gg n$ (see Section 1.1) and achieve the *rate DR* only.

TABLE 4.1

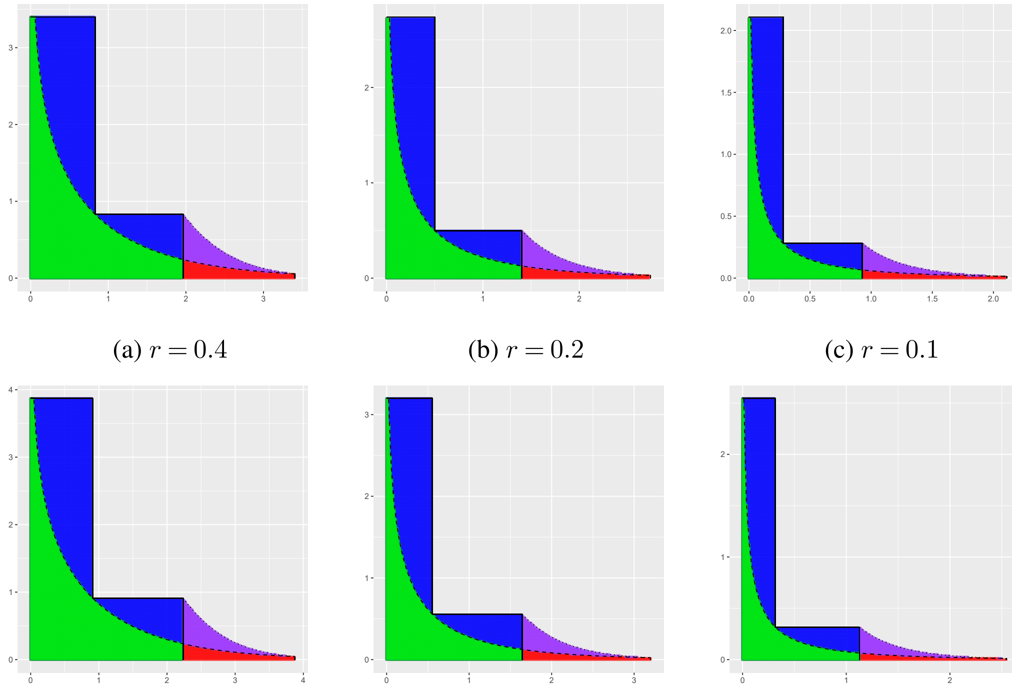
*Comparison of the sparsity conditions required for the asymptotic normality of the DR estimators with $R \equiv 1$. (C1) denotes the sparse outcome $s_{\boldsymbol{\alpha}} = o(N/\log d)$, (C2) the sparse propensity $s_{\boldsymbol{\beta}} = o(N/\log d)$, (C3) the product rate condition $s_{\boldsymbol{\alpha}} s_{\boldsymbol{\beta}} = o(N/\log^2 d)$ , (C4) the ultra-sparse outcome $s_{\boldsymbol{\alpha}} = o(\sqrt{N}/\log d)$, (C5) the ultra-sparse propensity $s_{\boldsymbol{\beta}} = o(\sqrt{N}/\log d)$. For brevity and without prejudice, we've abbreviated 'Avagyan' as 'Av.', 'Bradic' as 'Br.', 'Chakrabortty' as 'Chakr.', 'Chernozhukov' as 'Chern.', 'Kallus' as 'Kal.', and 'Vansteelandt' as 'Vdt.'.*

| Literature | M1<br>both models are correct | M2<br>missed PS | M3<br>missed OR |
|---|---|---|---|
| Farell (2015) | (C4) and (C5) | $\times$ | $\times$ |
| Chern. et al.(2018)<br>Kal. and Mao (2020)<br>Hou et al. (2021)<br>Chakr. et al. (2022)<br>Zhang and Br. (2022) | (C1), (C2), and (C3) | $\times$ | $\times$ |
| Athey et al. (2018) | (C4) and (C5) | (M1) | $\times$ |
| Tan (2020)<br>Ning et al. (2020)<br>Av. and Vdt. (2021) | (C4) and (C5) | (M1) | (M1) |
| Dukes et al.(2020)<br>Dukes and Vdt. (2021) | (C1), (C3), and (C5) | (C4) and (C5) | (M2) |
| Smucler et al. (2019) | (C1), (C2), and (C3) | (M1) | (C1), (C3), and (C5) |
| Br. et al. (2019) | $s_{\boldsymbol{\alpha}} = o(N^{3/4}/\log d)$ and (C5)<br>or<br>(C4) and (C2) | $\times$ | $\times$ |
| This paper | (C1), (C2), and (C3)<br>or<br>Br. et al. (2019)(M1)<br>or<br>$s_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\beta}} = o(N^2/\log^3 d)$ and (C5) | (M1) | (C1), (C3), and (C5)<br>or<br>$s_{\boldsymbol{\alpha}}^2 s_{\boldsymbol{\beta}} = o(N^2/\log^3 d)$ and (C5) |

*Regular MAR causal.* Wei et al. (2022) introduced ***model DR*** method for $N \asymp n$, focusing solely on low dimensions, where nuisance estimates reliably achieve root-$n$ rate. They estimate $\pi(\cdot)$ and $p_{N,1}(\cdot)$ separately, and require either (a) the OR function is correctly parametrized, or (b) both the labeling and treatment PS functions are correctly parametrized. Our DR-DMAR SS estimators achieve ***model DR*** in high dimensions, where achieving a root-$n$ rate for nuisances is not guaranteed, even under correct model assumptions. Moreover, our semi-parametric approach allows fully non-parametric estimates for $\pi(\cdot)$ and, different from condition (b), we only require the fraction $\gamma_N(\cdot)/\pi^*(\cdot)$ to be correctly parametrized and $\pi^*(\cdot) \neq \pi(\cdot)$ is allowed; see Remark 2.

*Regular SS.* Hou, Mukherjee and Cai (2021) presented ATE estimators with surrogates, ensuring normality under a more restrictive MCAR condition $R \perp\!\!\!\perp (Y, T, \mathbf{X})$. Zhang and Bradic (2022) and Chakrabortty, Dai and Tchetgen (2022) introduced SS estimates that retain normality despite misspecified OR models, especially as $p_{N,1}(\mathbf{X})$ stabilizes. In their method, $R_i$ and $n$ are non-random, and the error for $\pi(\cdot)$ depends on total $N$ rather than labeled size $n$. However, accurate PS models with consistent labeling remain essential.

*Supervised causal.* With $R \equiv 1$, our approach outperforms existing methods by relaxing model correctness and sparsity; see Table 4.1. Our proposal is closest to ***model DR*** estimators of Smucler, Rotnitzky and Robins (2019), Tan (2020) and Avagyan and Vansteelandt (2021). Our approach introduces an additional parameter tailored to the decaying MAR setting, and we utilize asymmetric cross-fitting, leading to more relaxed sparsity conditions.



(a) $r = 0.4$ (b) $r = 0.2$ (c) $r = 0.1$

(d) $r = 0.4$ (e) $r = 0.2$ (f) $r = 0.1$

Fig 1: Plots of $\log(1+\mathbf{s}) = (\log(1+s_{\boldsymbol{\alpha}}), \log(1+s_{\boldsymbol{\beta}}))$ satisfying sparsity conditions with $N = 500$, $d = 1000$ for (a)-(c) and $N = 1000$, $d = 5000$ for (d)-(f) : Green = $\{f_1(\mathbf{s}) \le r$, $f_2(\mathbf{s}) \le r$, and $f_3(\mathbf{s}) \le r\}$; Red = $\{f_1(\mathbf{s}) \le r$, $f_3(\mathbf{s}) \le r$, and $f_2(\mathbf{s}) > r\}$; Blue = $\{f_2(\mathbf{s}) \le r$, $f_3(\mathbf{s}) \le r$, and $f_1(\mathbf{s}) > r\}$; Purple = $\{f_3(\mathbf{s}) \le r$, $f_1(\mathbf{s}) > r$, and $f_2(\mathbf{s}) > r\}$. Lines $\{f_1(\mathbf{s}) = r\}$, $\{f_2(\mathbf{s}) = r\}$, and $\{f_3(\mathbf{s}) = r\}$ are dashed, solid, and dotted. With $R \equiv 1$ and all models well specified, Chernozhukov et al. (2018) requires $f_1(\mathbf{s}) = o(1)$ (green + red), Bradic, Wager and Zhu (2019) requires $f_2(\mathbf{s}) = o(1)$ (green + blue), and the proposed method only requires $f_3(\mathbf{s}) = o(1)$ (green + red + blue + purple).

While previous work required the product of two sparsity levels to be on the order of $n$, our approach allows for this product to be on the order of $n^{3/2}$ or $n^{5/4}$ depending on which model is correctly specified. Although Bradic, Wager and Zhu (2019) shares a similar cross-fitting strategy, their estimator's implementation is challenging due to non-convex constraints. Furthermore, our proposed estimator exhibits *model DR* and *rate DR* with weaker sparsity.

In the following, we clarify the above and consider the following cases: **(M1)** both nuisance models are correctly specified, **(M2)** only the OR model is correctly specified, and **(M3)** only the PS model is correctly specified. Important quantities with $\mathbf{s} = (s_{\boldsymbol{\alpha}}, s_{\boldsymbol{\beta}})$ are

$$f_1(\mathbf{s}) := s_{\boldsymbol{\alpha}} \log d/N \vee s_{\boldsymbol{\beta}} \log d/N \vee \sqrt{s_{\boldsymbol{\alpha}} s_{\boldsymbol{\beta}}} \log d/\sqrt{N},$$

$$f_2(\mathbf{s}) := \left(s_{\boldsymbol{\alpha}} \log d/N^{3/4} \vee s_{\boldsymbol{\beta}} \log d/\sqrt{N}\right) \wedge \left(s_{\boldsymbol{\alpha}} \log d/\sqrt{N} \vee s_{\boldsymbol{\beta}} \log d/N\right), \text{ and}$$

$$f_3(\mathbf{s}) := f_1(\mathbf{s}) \wedge f_2(\mathbf{s}) \wedge \left\{(s_{\boldsymbol{\alpha}} \log d/N) \vee \left(s_{\boldsymbol{\beta}} \log d/\sqrt{N}\right) \vee \left(s_{\boldsymbol{\alpha}}^{2/3} s_{\boldsymbol{\beta}}^{1/3} \log d/N^{2/3}\right)\right\}.$$

**(M1)** $m(\cdot) = m^*(\cdot)$ and $\gamma_N(\cdot) = \gamma_N^*(\cdot)$. While Chernozhukov et al. (2018) requires $f_1(\mathbf{s}) = o(1)$ for *rate DR*, and Bradic, Wager and Zhu (2019) relies on $f_2(\mathbf{s}) = o(1)$ for *sparsity DR*, our method allows for a more flexible and general sparsity setting $f_3(\mathbf{s}) = o(1)$, as illustrated in Figure 1 – our work combines all four colors, where the area colored in purple denotes the sparsity scenario new to the literature.

**(M2)** $m(\cdot) = m^*(\cdot)$ and $\gamma_N(\cdot) \neq \gamma_N^*(\cdot)$. We achieve the coveted *model DR* property while requiring the same sparsity as in (M1), i.e., $f_3(\mathbf{s}) = o(1)$. Comparatively, the best result in the existing literature, Smucler, Rotnitzky and Robins (2019), still necessitates a stronger condition of $f_1(\mathbf{s}) = o(1)$, while Bradic, Wager and Zhu (2019) is valid only for correctly specified models. Further details and comparisons can be found in Table 4.1.

**(M3)** $m(\cdot) \neq m^*(\cdot)$ and $\gamma_N(\cdot) = \gamma_N^*(\cdot)$. Our sparsity conditions are once again weaker. The correctness of OR and PS models affects the required sparsity conditions differently. When the PS model is misspecified, we do not require any additional assumptions compared with (M1). However, when the OR model is misspecified, Smucler, Rotnitzky and Robins (2019) and our proposed method further require an ultra-sparse PS – this is originated from the linear approximation of the non-linear PS model; see $\Delta_3$ in Section 4.2. By using the non-cross-fitted PS estimate, we allow a weaker product rate condition $s_{\boldsymbol{\alpha}}\sqrt{s_{\boldsymbol{\beta}}} = o(N)$ (omitting the logarithm terms) instead of the usual product rate condition $s_{\boldsymbol{\alpha}} s_{\boldsymbol{\beta}} = o(N)$ – a condition that is always required in Smucler, Rotnitzky and Robins (2019). Although Tan (2020); Ning, Sida and Imai (2020); Avagyan and Vansteelandt (2021); Dukes and Vansteelandt (2021); Dukes, Avagyan and Vansteelandt (2020) also provide robust inference for the ATE when the OR model is misspecified, they require ultra-sparse conditions for both nuisances, whereas we only need the PS model to be ultra-sparse.

## 5. Simulation studies.

We evaluate the general DR-DMAR SS estimator and the bias-reduced DR-DMAR SS estimators using various data-generating processes (DGPs).

### 5.1. *Results under the decaying MAR setting.*

We consider three types of DGPs in our simulation studies: (a) linear OR models, logistic (product) PS models; (b) linear OR models, non-logistic PS models; and (c) non-linear OR models, logistic PS models. We consider i.i.d. truncated normal covariates $\mathbf{X}_{ij} \sim^{\text{iid}} Z_{\text{trun},2}$ and $\mathbf{X}_{i1} = 1$ for each $i \in \{1, \dots, N\}$ and $j \in \{2, \dots, d\}$, where $Z_{\text{trun},2} \sim Z \mid \{|Z| < 2\}$ and $Z \sim \mathcal{N}(0,1)$.

*(a) Linear OR models, logistic PS models.* For $j \in \{0,1\}$ and any $\mathbf{x} \in \mathbb{R}^d$, we set

$$\gamma_N(j, \mathbf{x}) = g(\mathbf{x}^T \boldsymbol{\beta}(j)), \;\; \pi(\mathbf{x}) = \frac{\gamma_N(1,\mathbf{x}) + 1 - \gamma_N(0,\mathbf{x})}{2}, \tag{5.1}$$

$$p_N(1,\mathbf{x}) = \frac{\gamma_N(1,\mathbf{x})}{\pi(\mathbf{x})}, \;\text{ and }\; p_N(0,\mathbf{x}) = \frac{\gamma_N(0,\mathbf{x})}{1 - \pi(\mathbf{x})}. \tag{5.2}$$

We then consider

$$T_i \mid \mathbf{X}_i \sim \text{Bernoulli}(\pi(\mathbf{X}_i)) \;\text{ and }\; R_i \mid (\mathbf{X}_i, T_i = j) \sim \text{Bernoulli}(p_N(j, \mathbf{X}_i)). \tag{5.3}$$

Finally, we set linear OR models as follows:

$$Y_i(j) = \mathbf{X}_i^T \boldsymbol{\alpha}(j) + \delta_i, \;\text{ and }\; R_i Y_i = R_i Y_i(T_i), \;\text{ where }\; \delta_i \sim \mathcal{N}(0,1). \tag{5.4}$$

*(b) Linear OR models, non-logistic PS models.* We use $\pi(\mathbf{x}) = 0.3\sin(\mathbf{x}^T\boldsymbol{\omega}) + 0.5$ and $p_N(j,\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\beta}(j))$. The treatment and missingness indicators $(T_i, R_i)$ follow (5.3), and the outcomes $Y_i(j)$ are determined as in (5.4).

TABLE 5.1

*Simulation results for DGP (a). 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. Bias, RMSE, Length, ESD, and ASD are calculated based on medians.*

| Estimator | Bias | RMSE | Length | Coverage | ESD | ASD |
|---|---|---|---|---|---|---|
| | DGP (a) $N=10000$, $d=51$, $s_{\boldsymbol{\alpha}}=3$, $s_{\boldsymbol{\beta}}=3$, $\gamma_{N,j}=0.05$ $(N\gamma_{N,j}=500)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | 0.007 | 0.102 | 0.540 | 0.948 | 0.154 | 0.138 |
| $\widehat{\mu}_{\text{MCAR}}$ | 0.003 | 0.090 | 0.374 | 0.828 | 0.136 | 0.095 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | 0.008 | 0.100 | 0.470 | 0.906 | 0.147 | 0.120 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.026 | 0.151 | 0.619 | 0.832 | 0.214 | 0.158 |
| $\widehat{\mu}_{\text{BRSS}}$ | 0.004 | 0.102 | 0.503 | 0.908 | 0.152 | 0.128 |
| | DGP (a) $N=10000$, $d=51$, $s_{\boldsymbol{\alpha}}=3$, $s_{\boldsymbol{\beta}}=3$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=1000)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | 0.006 | 0.075 | 0.428 | 0.956 | 0.111 | 0.109 |
| $\widehat{\mu}_{\text{MCAR}}$ | 0.003 | 0.067 | 0.329 | 0.902 | 0.101 | 0.084 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | 0.004 | 0.073 | 0.391 | 0.934 | 0.107 | 0.100 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.023 | 0.097 | 0.422 | 0.848 | 0.139 | 0.108 |
| $\widehat{\mu}_{\text{BRSS}}$ | 0.008 | 0.073 | 0.409 | 0.942 | 0.111 | 0.104 |
| | DGP (a) $N=5000$, $d=201$, $s_{\boldsymbol{\alpha}}=3$, $s_{\boldsymbol{\beta}}=3$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=500)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.014 | 0.098 | 0.597 | 0.954 | 0.145 | 0.152 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.011 | 0.091 | 0.466 | 0.878 | 0.136 | 0.119 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.011 | 0.095 | 0.522 | 0.936 | 0.144 | 0.133 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.012 | 0.099 | 0.555 | 0.924 | 0.145 | 0.142 |
| | DGP (a) $N=10000$, $d=201$, $s_{\boldsymbol{\alpha}}=3$, $s_{\boldsymbol{\beta}}=3$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=1000)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.013 | 0.080 | 0.424 | 0.964 | 0.112 | 0.108 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.010 | 0.069 | 0.329 | 0.900 | 0.103 | 0.084 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.012 | 0.074 | 0.379 | 0.946 | 0.109 | 0.097 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.014 | 0.079 | 0.402 | 0.958 | 0.114 | 0.103 |

*(c) Non-linear OR models, logistic PS models.* We consider (5.1)-(5.3) as in part (a) but set quadratic OR models with $\mathbf{X}_i^2 := (\mathbf{X}_{i1}^2, \ldots, \mathbf{X}_{id}^2)^T$ and

$$Y_i(j) = \mathbf{X}_i^T \boldsymbol{\alpha}(j) + (\mathbf{X}_i^2)^T \boldsymbol{\eta}(j) + \delta_i, \;\; R_i Y_i = R_i Y_i(T_i). \tag{5.5}$$

The parameter values across all the DGPs above are chosen as follows: $\boldsymbol{\alpha}(1) := 3(1,1,\mathbf{1}_{s_{\boldsymbol{\alpha}}-1}^T/\sqrt{s_{\boldsymbol{\alpha}}-1}, 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\alpha}(0) := -\boldsymbol{\alpha}(1)$, $\boldsymbol{\beta}(1) := (\beta_N(1), 1, \mathbf{1}_{s_{\boldsymbol{\beta}}-1}^T/(s_{\boldsymbol{\beta}}-1), 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\beta}(0) := (\beta_N(0), -1, -\mathbf{1}_{s_{\boldsymbol{\beta}}-1}^T/(s_{\boldsymbol{\beta}}-1), 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\omega} := (0, 1, \mathbf{1}_{s_{\boldsymbol{\beta}}-1}^T/(s_{\boldsymbol{\beta}}-1), 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\eta}(1) := (0, 1, \mathbf{1}_{s_{\boldsymbol{\alpha}}-1}^T/\sqrt{s_{\boldsymbol{\alpha}}-1}, 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\eta}(0) := -\boldsymbol{\eta}(1)$, where for any positive integer $s \geq 1$, $\mathbf{1}_s := \{1, \ldots, 1\} \in \mathbb{R}^s$ and $\mathbf{1}_0 := \emptyset$, and $\beta_N(1), \beta_N(0)$ are chosen such that $\mathbb{E}(RT) = \gamma_{N,1}$ and $\mathbb{E}\{R(1-T)\} = \gamma_{N,0}$.

We consider: (1) Oracle estimator $\widehat{\mu}_{\text{oracle}}$: DR-DMAR SS estimator with true values as nuisances. (2) $\widehat{\mu}_{\text{MCAR}}$: SS estimator treating missingness as MCAR (the selection bias is ignored), estimated using Lasso for OR and PS. (3) $\widehat{\mu}_{\text{SS-Lasso}}$: DR-DMAR SS estimator with Lasso for OR and PS. (4) $\widehat{\mu}_{\text{SS-RF}}$: DR-DMAR SS estimator with random forest estimates (up to $d = 51$ dimensions). (5) The *bias-reduced* $\widehat{\mu}_{\text{BRSS}}$ of (4.6). The tuning parameters are chosen using 5-fold cross-validation and results, repeated 500 times, are presented in Tables 5.1-5.3.

Among the four estimators, $\widehat{\mu}_{\text{SS-RF}}$ exhibits larger bias and RMSE due to slower convergence rates, while the remaining estimators demonstrate smaller biases compared to RMSE in Tables 5.1 and 5.2. In contrast, in Table 5.3, notable biases and larger RMSEs are observed for $\widehat{\mu}_{\text{SS-RF}}$, $\widehat{\mu}_{\text{SS-Lasso}}$, and $\widehat{\mu}_{\text{MCAR}}$, while $\widehat{\mu}_{\text{BRSS}}$ outperforms them significantly. The poor performance of $\widehat{\mu}_{\text{MCAR}}$ in all DGPs is attributed to its incorrect treatment of the labeling indicator's PS as a constant, resulting in significant undercoverage, particularly in Table 5.3. On

TABLE 5.2
*Simulation results for DGP (b). The rest of the caption details remain the same as those in Table 5.1.*

| Estimator | Bias | RMSE | Length | Coverage | ESD | ASD |
|---|---|---|---|---|---|---|
| | DGP (b) $N=10000$, $d=51$, $s_{\boldsymbol{\alpha}}=2$, $s_{\boldsymbol{\beta}}=6$, $\gamma_{N,j}=0.05$ $(N\gamma_{N,j}=500)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.012 | 0.084 | 0.565 | 0.964 | 0.125 | 0.144 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.005 | 0.067 | 0.286 | 0.812 | 0.099 | 0.073 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.004 | 0.079 | 0.390 | 0.912 | 0.118 | 0.099 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.023 | 0.101 | 0.403 | 0.824 | 0.147 | 0.103 |
| $\widehat{\mu}_{\text{BRSS}}$ | 0.001 | 0.075 | 0.470 | 0.946 | 0.111 | 0.120 |
| | DGP (b) $N=10000$, $d=51$, $s_{\boldsymbol{\alpha}}=2$, $s_{\boldsymbol{\beta}}=6$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=1000)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.007 | 0.075 | 0.412 | 0.954 | 0.110 | 0.105 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.001 | 0.061 | 0.250 | 0.856 | 0.090 | 0.064 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.005 | 0.065 | 0.312 | 0.916 | 0.096 | 0.080 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.015 | 0.070 | 0.297 | 0.852 | 0.097 | 0.076 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.005 | 0.068 | 0.366 | 0.942 | 0.100 | 0.093 |
| | DGP (b) $N=5000$, $d=201$, $s_{\boldsymbol{\alpha}}=2$, $s_{\boldsymbol{\beta}}=6$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=500)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.008 | 0.106 | 0.572 | 0.936 | 0.156 | 0.146 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.008 | 0.078 | 0.354 | 0.872 | 0.118 | 0.090 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.006 | 0.089 | 0.418 | 0.884 | 0.135 | 0.107 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.010 | 0.089 | 0.480 | 0.918 | 0.129 | 0.122 |
| | DGP (b) $N=10000$, $d=201$, $s_{\boldsymbol{\alpha}}=2$, $s_{\boldsymbol{\beta}}=6$, $\gamma_{N,j}=0.1$ $(N\gamma_{N,j}=1000)$ $\forall j\in\{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | 0.001 | 0.078 | 0.412 | 0.952 | 0.115 | 0.105 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.004 | 0.062 | 0.270 | 0.880 | 0.094 | 0.069 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.001 | 0.066 | 0.300 | 0.920 | 0.099 | 0.077 |
| $\widehat{\mu}_{\text{BRSS}}$ | 0.001 | 0.074 | 0.357 | 0.952 | 0.110 | 0.091 |

the other hand, $\widehat{\mu}_{\text{SS-Lasso}}$ provides valid inference with correct nuisance model specification (as per Theorem A.2), but its reliability diminishes when model misspecification occurs, leading to underestimation of variance and large bias in Tables 5.2 and 5.3. In contrast, the $\widehat{\mu}_{\text{SS-RF}}$ estimator, relying on non-parametric nuisance estimators, fails to satisfy the 'product-rate' condition, resulting in undercoverage in all DGPs. The coverage results in Tables 5.1-5.3 support the inference quality of $\widehat{\mu}_{\text{BRSS}}$ when one nuisance is misspecified, as per Corollary 4.6. Notably, Table 5.2 exhibits excellent coverage with minor deviations for smaller effective sample size and higher dimension, while in Table 5.3, $\widehat{\mu}_{\text{BRSS}}$ consistently delivers strong performance for higher sample sizes where $\widehat{\mu}_{\text{MCAR}}$, $\widehat{\mu}_{\text{SS-Lasso}}$, and $\widehat{\mu}_{\text{SS-RF}}$ fail, respectively.

5.2. *A degenerate setting with outcomes fully observed.* In the setting of fully observed outcomes ($R_i=1$ for all $i\in 1,\dots,N$), we examine the cases where one of the nuisance models is misspecified and highlight what is different from Section 5.1.

*(d) Linear OR models, non-logistic PS model.*

$$\pi(\mathbf{x})=\exp(\mathbf{x}^T\boldsymbol{\beta}_d)/\{1+\exp(\mathbf{x}^T\boldsymbol{\beta}_d)\}\{0.3\sin(\mathbf{x}^T\boldsymbol{\beta}_d)+0.7\} \text{ and } Y_i(j)=\mathbf{X}_i^T\boldsymbol{\alpha}_d(j)+\delta_i.$$

*(e) Non-linear OR models, logistic PS model.*

$$\pi(\mathbf{x})=\exp(\mathbf{x}^T\boldsymbol{\beta}_e)/\{1+\exp(\mathbf{x}^T\boldsymbol{\beta}_e)\} \text{ and } Y_i(j)=\mathbf{X}_i^T\boldsymbol{\alpha}_e(j)+(\mathbf{X}_i^2)^T\boldsymbol{\eta}_e(j)+\delta_i.$$

The parameters for DGPs (d) and (e) are $\boldsymbol{\alpha}_d(1):=3(1,0.9^1,\dots,0.9^{d-1})^T\in\mathbb{R}^d$, $\boldsymbol{\alpha}_d(0)=-\boldsymbol{\alpha}_d(1)$, $\boldsymbol{\alpha}_e(j)=\boldsymbol{\alpha}_d(j)$, $\boldsymbol{\eta}_e(1):=(0,0.9^1,\dots,0.9^{d-1})^T\in\mathbb{R}^d$, $\boldsymbol{\eta}_e(0)=-\boldsymbol{\eta}_e(1)$, $\boldsymbol{\beta}_d:=(0.99,0.5\cdot 0.7^1,\dots,0.5\cdot 0.7^{d-1})^T\in\mathbb{R}^d$, $\boldsymbol{\beta}_e:=(0.2247,0.7^1,\dots,0.7^{d-1})^T\in\mathbb{R}^d$. Here, we compare the numerical performance of our proposed bias-reduced estimator $\widehat{\mu}_{\text{BRSS}}$ with $\widehat{\mu}_{\text{Smucler}}$ by Smucler, Rotnitzky and Robins (2019); as per Table 4.1 their estimator is the most competitive in the existing literature. The nuisance parameters in DGPs (d) and (e) exhibit a 'weakly

TABLE 5.3
*Simulation results for DGP (c). The rest of the caption details remain the same as those in Table 5.1.*

| Estimator | Bias | RMSE | Length | Coverage | ESD | ASD |
|---|---|---|---|---|---|---|
| | DGP (c) $N = 10000$, $d = 51$, $s_{\boldsymbol{\alpha}} = 6$, $s_{\boldsymbol{\beta}} = 2$, $\gamma_{N,j} = 0.05$ $(N\gamma_{N,j} = 500)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.013 | 0.077 | 0.463 | 0.948 | 0.110 | 0.118 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.722 | 0.722 | 0.484 | 0.002 | 0.168 | 0.124 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.255 | 0.263 | 0.752 | 0.712 | 0.193 | 0.192 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.282 | 0.288 | 0.749 | 0.668 | 0.221 | 0.191 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.116 | 0.139 | 0.637 | 0.884 | 0.153 | 0.163 |
| | DGP (c) $N = 10000$, $d = 51$, $s_{\boldsymbol{\alpha}} = 6$, $s_{\boldsymbol{\beta}} = 2$, $\gamma_{N,j} = 0.1$ $(N\gamma_{N,j} = 1000)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.010 | 0.063 | 0.390 | 0.954 | 0.092 | 0.099 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.642 | 0.642 | 0.394 | 0.000 | 0.132 | 0.101 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.181 | 0.182 | 0.576 | 0.740 | 0.151 | 0.147 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.186 | 0.188 | 0.525 | 0.674 | 0.164 | 0.134 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.044 | 0.095 | 0.493 | 0.930 | 0.133 | 0.126 |
| | DGP (c) $N = 5000$, $d = 201$, $s_{\boldsymbol{\alpha}} = 6$, $s_{\boldsymbol{\beta}} = 2$, $\gamma_{N,j} = 0.1$ $(N\gamma_{N,j} = 500)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.009 | 0.096 | 0.550 | 0.956 | 0.143 | 0.140 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.652 | 0.652 | 0.566 | 0.026 | 0.186 | 0.144 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.292 | 0.292 | 0.745 | 0.648 | 0.199 | 0.190 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.134 | 0.160 | 0.685 | 0.862 | 0.179 | 0.175 |
| | DGP (c) $N = 10000$, $d = 201$, $s_{\boldsymbol{\alpha}} = 6$, $s_{\boldsymbol{\beta}} = 2$, $\gamma_{N,j} = 0.1$ $(N\gamma_{N,j} = 1000)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.009 | 0.064 | 0.391 | 0.964 | 0.090 | 0.100 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.637 | 0.637 | 0.395 | 0.000 | 0.128 | 0.101 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.223 | 0.223 | 0.556 | 0.660 | 0.134 | 0.142 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.065 | 0.091 | 0.488 | 0.894 | 0.117 | 0.125 |
| | DGP (c) $N = 20000$, $d = 201$, $s_{\boldsymbol{\alpha}} = 6$, $s_{\boldsymbol{\beta}} = 2$, $\gamma_{N,j} = 0.1$ $(N\gamma_{N,j} = 2000)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.010 | 0.050 | 0.276 | 0.948 | 0.071 | 0.070 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.628 | 0.628 | 0.278 | 0.000 | 0.089 | 0.071 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.168 | 0.168 | 0.407 | 0.622 | 0.103 | 0.104 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.033 | 0.063 | 0.349 | 0.930 | 0.090 | 0.089 |

sparse' nature, with bounded $\ell_1$-norms but $\ell_0$-norms equal to the dimension. In Table 5.4, it is observed that the estimator $\widehat{\mu}_{\text{Smucler}}$ suffers from substantial biases. Consequently, the coverages based on $\widehat{\mu}_{\text{Smucler}}$ are relatively poor. In contrast, our proposed estimator $\widehat{\mu}_{\text{BRSS}}$ exhibits significantly smaller biases, leading to improved coverages. Additionally, $\widehat{\mu}_{\text{BRSS}}$ achieves smaller RMSEs compared to $\widehat{\mu}_{\text{Smucler}}$ for both DGPs (d) and (e).

TABLE 5.4
*Simulation results for DGPs (d) and (e). The rest of the caption details remain the same as those in Table 5.1.*

| Estimator | Bias | RMSE | Length | Coverage | ESD | ASD |
|---|---|---|---|---|---|---|
| | DGP (d) $N = 300$, $d = 51$, $s_{\boldsymbol{\alpha}} = 51$, $s_{\boldsymbol{\beta}} = 51$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | 0.002 | 0.411 | 2.518 | 0.948 | 0.608 | 0.642 |
| $\widehat{\mu}_{\text{Smucler}}$ | -0.560 | 0.637 | 2.626 | 0.870 | 0.645 | 0.670 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.208 | 0.460 | 2.471 | 0.928 | 0.662 | 0.630 |
| | DGP (e) $N = 400$, $d = 51$, $s_{\boldsymbol{\alpha}} = 51$, $s_{\boldsymbol{\beta}} = 51$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.012 | 0.375 | 2.305 | 0.958 | 0.564 | 0.588 |
| $\widehat{\mu}_{\text{Smucler}}$ | 0.484 | 0.584 | 2.479 | 0.896 | 0.628 | 0.632 |
| $\widehat{\mu}_{\text{BRSS}}$ | 0.104 | 0.471 | 2.315 | 0.942 | 0.655 | 0.591 |

5.3. *Results based on the semi-parametric approach.* In this section, we further examine the behavior of the semi-parametric bias-reduced DR-DMAR SS estimator, $\widehat{\mu}_{\text{SP-BRSS}}$, proposed in Section 4.3. The following case is considered.

*(f) Non-linear OR models, non-logistic treatment PS model, logistic labeling PS model.* Generate $\pi(\mathbf{x}) = 0.3\cos(\mathbf{x}^T\boldsymbol{\beta}) + 0.5$, $p_N(j,\mathbf{x}) = \exp\{\mathbf{x}^T\boldsymbol{\beta}'(j)\}/[1+\exp\{\mathbf{x}^T\boldsymbol{\beta}'(j)\}]$, and (5.5). Choose $\boldsymbol{\alpha}(j)$ and $\boldsymbol{\eta}(j)$ as in Section 5.1 with $s_{\boldsymbol{\alpha}} = 5$, $\boldsymbol{\beta} := 2(0, 0.9^1, \ldots, 0.9^{d-1})^T$, $\boldsymbol{\beta}'(1) := (\boldsymbol{\beta}'_N(1), 1, \mathbf{1}_4^T/4, 0, \ldots, 0)^T \in \mathbb{R}^d$, and $\boldsymbol{\beta}'(0) := (\boldsymbol{\beta}'_N(0), -1, -\mathbf{1}_4^T/4, 0, \ldots, 0)^T \in \mathbb{R}^d$, where $\boldsymbol{\beta}'_N(1)$ and $\boldsymbol{\beta}'_N(0)$ are chosen such that $\mathbb{E}(RT) = \gamma_{N,1}$ and $\mathbb{E}\{R(1-T)\} = \gamma_{N,0}$.

We compare the numerical performance of the estimators considered in Section 5.1 with $\widehat{\mu}_{\text{SP-BRSS}}$, where the treatment PS function $\pi(\cdot)$ is estimated using random forests. As shown in Table 5.5, $\widehat{\mu}_{\text{MCAR}}$ provides large biases and very poor coverages since the true labeling mechanism is not MCAR. The performance of $\widehat{\mu}_{\text{SS-Lasso}}$ is also relatively poor since both nuisance models are misspecified – the OR models are non-linear and the product PS models are non-logistic. The parametric BRSS estimator $\widehat{\mu}_{\text{BRSS}}$ provides slightly smaller biases and RMSEs, although the working models are still misspecified. The fully non-parametric estimator $\widehat{\mu}_{\text{SS-RF}}$ performs similarly as, better than, and worse than $\widehat{\mu}_{\text{BRSS}}$ when $(d, \gamma_{N,j})$ are chosen as $(31, 0.05)$, $(51, 0.05)$, and $(51, 0.1)$, respectively. Although the non-parametric nuisance estimates are consistent, the convergence rates are relatively slow as the 'effective sample size' for the OR and product PS estimation is only $500$ when $\gamma_{N,j} = 0.05$ and $1000$ when $\gamma_{N,j} = 0.1$ (with a moderate dimension $d = 31$ or $d = 51$), resulting in relatively large biases for the final ATE estimation. Lastly, by making full use of the large sized $(T_i, \mathbf{X}_i)$ and estimate $\pi(\cdot)$ non-parametrically while keeping other working models parametrically, the proposed semi-parametric BRSS estimator $\widehat{\mu}_{\text{SP-BRSS}}$ outperforms all the ATE estimators above, although the OR models are misspecified.

TABLE 5.5
*Simulation results for DGP (f). The rest of the caption details remain the same as those in Table 5.1.*

| Estimator | Bias | RMSE | Length | Coverage | ESD | ASD |
|---|---|---|---|---|---|---|
| | DGP (f) $N = 10000$, $d = 31$, $\gamma_{N,j} = 0.05$ $(N\gamma_{N,j} = 500)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.007 | 0.083 | 0.500 | 0.954 | 0.122 | 0.127 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.724 | 0.724 | 0.493 | 0.004 | 0.186 | 0.126 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.224 | 0.233 | 0.774 | 0.736 | 0.217 | 0.198 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.172 | 0.198 | 0.736 | 0.774 | 0.246 | 0.188 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.170 | 0.180 | 0.640 | 0.800 | 0.182 | 0.163 |
| $\widehat{\mu}_{\text{SP-BRSS}}$ | -0.127 | 0.158 | 1.144 | 0.988 | 0.195 | 0.292 |
| | DGP (f) $N = 10000$, $d = 51$, $\gamma_{N,j} = 0.05$ $(N\gamma_{N,j} = 500)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.011 | 0.085 | 0.499 | 0.942 | 0.126 | 0.127 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.727 | 0.727 | 0.495 | 0.006 | 0.170 | 0.126 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.284 | 0.290 | 0.746 | 0.664 | 0.213 | 0.190 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.165 | 0.206 | 0.732 | 0.828 | 0.216 | 0.187 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.190 | 0.200 | 0.634 | 0.780 | 0.184 | 0.162 |
| $\widehat{\mu}_{\text{SP-BRSS}}$ | -0.139 | 0.179 | 1.136 | 0.988 | 0.197 | 0.290 |
| | DGP (f) $N = 10000$, $d = 51$, $\gamma_{N,j} = 0.1$ $(N\gamma_{N,j} = 1000)$ $\forall j \in \{0,1\}$ | | | | | |
| $\widehat{\mu}_{\text{oracle}}$ | -0.018 | 0.074 | 0.405 | 0.944 | 0.104 | 0.103 |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.614 | 0.614 | 0.402 | 0.002 | 0.137 | 0.103 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.255 | 0.255 | 0.521 | 0.534 | 0.141 | 0.133 |
| $\widehat{\mu}_{\text{SS-RF}}$ | -0.151 | 0.161 | 0.498 | 0.732 | 0.154 | 0.127 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.135 | 0.143 | 0.478 | 0.760 | 0.140 | 0.122 |
| $\widehat{\mu}_{\text{SP-BRSS}}$ | -0.084 | 0.109 | 0.799 | 0.992 | 0.142 | 0.204 |

**6. Applications to a pseudo-random dataset.** We compare the performance of the proposed estimators using a synthetic dataset obtained from the Atlantic Causal Inference Conference (ACIC) 2019 Data Challenge. [2] We focus on 14 scenarios from the ACIC 2019 dataset, as two scenarios share the same covariate matrix. To examine the performance under model misspecification, we construct $T_i$ and $R_i$ as in (5.1)-(5.3) and generate the outcome variable $Y_i$ as in (5.5) – that is, we consider a correctly specified (logistic) PS model and a misspecified (quadratic) OR model. We set $\boldsymbol{\alpha}(1) := (2,2,1,1,1,1,0,\ldots,0)^T \in \mathbb{R}^d$, $\boldsymbol{\alpha}(0) := -\boldsymbol{\alpha}(1)$, $\boldsymbol{\beta}(1) := (-1.5, 0.5, 0, \ldots, 0)^T \in \mathbb{R}^d$, $\boldsymbol{\beta}(0) := (-1.5, -0.5, 0, \ldots, 0)^T \in \mathbb{R}^d$, and $\boldsymbol{\eta}(1) := 0.35(0,2,1,1,1,1,0,\ldots,0)^T \in \mathbb{R}^d$, $\boldsymbol{\eta}(0) := -\boldsymbol{\eta}(1)$. We generate 100 sets for each scenario resulting in $14 * 100 = 1400$ pseudo-random datasets in total. The covariate matrices have dimensions of $(N,d) = (1000, 201)$ in Scenarios 1, 2, 3, 6, 8, and 14, and $(N,d) = (2000, 201)$ in the other scenarios. We consider $\widehat{\mu}_{\text{MCAR}}$, $\widehat{\mu}_{\text{SS-Lasso}}$, and $\widehat{\mu}_{\text{BRSS}}$. The results are reported in Table 6.1.

Except for Scenario 3, the MCAR estimator exhibits poor performance with large biases and inadequate coverage. In Scenarios 2, 3, 4, 6, 8, and 14, both the SS-Lasso and BRSS estimators perform well, showing similar biases, RMSEs, and coverages close to 95%. However,

[2]The high-dimensional datasets (with continuous outcomes) provided by ACIC 2019 are available at: https://sites.google.com/view/acic2019datachallenge/data-challenge, and these are constructed based on 16 scenarios.

TABLE 6.1

*Results for the pseudo-random dataset. Bias: empirical bias; RMSE: root mean square error; Length: average length of the 95% confidence intervals; Coverage: average coverage of the 95% confidence intervals.*

| Estimator | Bias | RMSE | Length | Coverage | Bias | RMSE | Length | Coverage |
|---|---|---|---|---|---|---|---|---|
| | Scenario 1 $N = 1000$, $d = 201$ | | | | Scenario 2 $N = 1000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.341 | 0.403 | 1.209 | 0.880 | -0.231 | 0.307 | 0.994 | 0.900 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.257 | 0.398 | 1.248 | 0.900 | -0.157 | 0.260 | 1.057 | 0.960 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.263 | 0.322 | 1.142 | 0.940 | -0.164 | 0.244 | 1.004 | 0.940 |
| | Scenario 3 $N = 1000$, $d = 201$ | | | | Scenario 4 $N = 2000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.080 | 0.400 | 1.879 | 0.960 | -0.237 | 0.273 | 0.652 | 0.760 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.102 | 0.408 | 1.958 | 0.920 | -0.144 | 0.200 | 0.696 | 0.930 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.114 | 0.411 | 1.906 | 0.930 | -0.095 | 0.161 | 0.672 | 0.960 |
| | Scenario 5 $N = 2000$, $d = 201$ | | | | Scenario 6 $N = 1000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.290 | 0.316 | 0.665 | 0.650 | -0.247 | 0.313 | 0.997 | 0.860 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.187 | 0.231 | 0.717 | 0.870 | -0.168 | 0.266 | 1.058 | 0.920 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.118 | 0.167 | 0.686 | 0.980 | -0.171 | 0.245 | 1.006 | 0.950 |
| | Scenario 7 $N = 2000$, $d = 201$ | | | | Scenario 8 $N = 1000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.254 | 0.290 | 0.656 | 0.670 | -0.272 | 0.337 | 0.956 | 0.810 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.152 | 0.217 | 0.709 | 0.880 | -0.180 | 0.285 | 1.019 | 0.910 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.095 | 0.158 | 0.679 | 0.980 | -0.186 | 0.270 | 0.975 | 0.940 |
| | Scenario 9 $N = 2000$, $d = 201$ | | | | Scenario 12 $N = 2000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.247 | 0.289 | 0.666 | 0.710 | -0.250 | 0.283 | 0.662 | 0.710 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.140 | 0.209 | 0.718 | 0.900 | -0.147 | 0.198 | 0.711 | 0.890 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.082 | 0.156 | 0.687 | 0.960 | -0.093 | 0.159 | 0.678 | 0.990 |
| | Scenario 13 $N = 2000$, $d = 201$ | | | | Scenario 14 $N = 1000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.236 | 0.275 | 0.665 | 0.720 | -0.211 | 0.297 | 0.956 | 0.910 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.132 | 0.199 | 0.714 | 0.860 | -0.122 | 0.259 | 1.020 | 0.950 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.084 | 0.155 | 0.682 | 0.990 | -0.134 | 0.242 | 0.970 | 0.960 |
| | Scenario 15 $N = 2000$, $d = 201$ | | | | Scenario 16 $N = 2000$, $d = 201$ | | | |
| $\widehat{\mu}_{\text{MCAR}}$ | -0.245 | 0.278 | 0.661 | 0.730 | -0.356 | 0.388 | 0.832 | 0.590 |
| $\widehat{\mu}_{\text{SS-Lasso}}$ | -0.148 | 0.202 | 0.708 | 0.900 | -0.228 | 0.321 | 0.911 | 0.820 |
| $\widehat{\mu}_{\text{BRSS}}$ | -0.087 | 0.149 | 0.679 | 0.990 | -0.159 | 0.204 | 0.812 | 0.980 |

in Scenarios 5, 7, 9, 12, 13, 15, and 16, BRSS outperforms SS-Lasso with smaller biases, RMSEs, and improved coverages. Additionally, in Scenario 1, BRSS achieves a smaller RMSE and a coverage closer to 95%, while its bias remains comparable to SS-Lasso. Overall, the MCAR estimator's poor performance is attributed to the mischaracterization of the labeling PS function, while the BRSS estimator exhibits greater stability compared to SS-Lasso due to the misspecified OR model.

**7. Discussion.** This paper addresses the estimation of the average treatment effect (ATE) in settings where selection bias may occur. We introduce a novel framework called the *'decaying MAR setting,'* which encompasses both regular selection bias and missing outcome scenarios, providing a more general approach. Within this framework, we propose flexible ATE estimators based on flexible nuisance model estimators, including non-parametric ones. To tackle the challenges of model misspecification, we propose bias-reduced ATE estimators that incorporate carefully designed nuisance estimates and an asymmetric cross-fitting strategy. Our results highlight the crucial role of these design choices in ensuring robust estimation and inference of the ATE, particularly in high-dimensional settings.

In addition to the ATE estimation problem, our framework opens up possibilities for studying estimation and inference of other parameters of interest in the decaying MAR setting. One such parameter is the decaying PS function, which serves as a vital intermediate step for ATE estimation. Further research is needed to explore the validity of non-parametric PS estimators in the decaying PS setup. Additionally, alternative methods like inverse probability weighting and residual balancing can be employed in degenerate supervised settings, but their validity and theoretical properties in this context require further investigation for future research.

**Acknowledgement.** AC acknowledges valuable early discussions with Rajarshi Mukherjee. This work was supported by the National Science Foundation grants NSF DMS-2113768 (AC) and NSF DMS-1712481 (JB).

## APPENDIX: SUPPLEMENTARY MATERIAL

**Supplement to 'The Decaying Missing-at-Random Framework: Doubly Robust Causal Inference with Partially Labeled Data'**. In the Supplement, we provide additional theoretical results, discussions, and proofs related to our main findings. In Section A, we present theoretical results for the estimation of the counterfactual mean and the ATE. Further discussions on estimating PS when the treatment variable $T$ is missing are included in Section B. The proofs of all the main results can be found in Sections C-G.

## REFERENCES

AVAGYAN, V. and VANSTEELANDT, S. (2021). High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics* & *Epidemiology* 1–18.

AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2022). Semi-supervised linear regression. *Journal of the American Statistical Association* **117** 2238–2251.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

BRADIC, J., WAGER, S. and ZHU, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*.

CAI, T. T. and GUO, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** 391–419.

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics* **46** 1541–1572.

Chakrabortty, A., Dai, G. and Carroll, R. J. (2022). Semi-Supervised Quantile Estimation: Robust and Efficient Inference in High Dimensional Settings. *arXiv preprint arXiv:2201.10208*.

Chakrabortty, A., Dai, G. and Tchetgen, E. T. (2022). A General Framework for Treatment Effect Estimation in Semi-Supervised and High Dimensional Settings. *arXiv preprint arXiv:2201.00468*.

Chakrabortty, A., Lu, J., Cai, T. T. and Li, H. (2019). High Dimensional M-Estimation with Missing Outcomes: A Semi-Parametric Framework. *arXiv preprint arXiv:1911.11345*.

Chan, S. F., Hejblum, B. P., Chakrabortty, A. and Cai, T. (2020). Semi-supervised estimation of covariance with application to phenome-wide association studies with electronic medical records data. *Statistical Methods in Medical Research* **29** 455–465.

Cheng, D., Ananthakrishnan, A. N. and Cai, T. (2021). Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics* **77** 413–423.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68.

Chi, C.-M., Vossler, P., Fan, Y. and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics* **50** 3415–3438.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A. and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694.

Dukes, O., Avagyan, V. and Vansteelandt, S. (2020). Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics* **76** 1190–1200.

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

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

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **79** 437–452.

Hou, J., Mukherjee, R. and Cai, T. (2021). Efficient and Robust Semi-supervised Estimation of ATE with Partially Annotated Treatment and Response. *arXiv preprint arXiv:2110.12336*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kallus, N. and Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22** 523–539.

Kawakita, M. and Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine Learning* **91** 189–209.

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

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G. and Cole, S. R. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)* **28** 553.

Manski, C. F. (1997). Monotone Treatment Response. *Econometrica* **65** 1311.

Manski, C. F. and Pepper, J. V. (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica* **68** 997–1010.

Mebane, W. R. and Poast, P. (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis* **21** 233–251.

Molinari, F. (2010). Missing Treatments. *Journal of Business & Economic Statistics* **28** 82–95.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2010). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *arXiv preprint arXiv:1010.2731*.

Ning, Y., Sida, P. and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika* **107** 533–554.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90** 122–129.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RUDELSON, M. and ZHOU, S. (2012). Reconstruction from Anisotropic Random Measurements. In *Proceedings of the 25th Annual Conference on Learning Theory* (S. MANNOR, N. SREBRO and R. C. WILLIAMSON, eds.). *Proceedings of Machine Learning Research* **23** 10.1–10.24. JMLR Workshop and Conference Proceedings, Edinburgh, Scotland.

SHI, X., PAN, Z. and MIAO, W. (2023). Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics* **15** e1581.

SMUCLER, E., ROTNITZKY, A. and ROBINS, J. M. (2019). A unifying approach for doubly-robust $\ell_1$ regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.

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

TSIATIS, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

VAN DE GEER, S. and LEDERER, J. (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields* **157** 225–250.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

WEI, K., QIN, G., ZHANG, J. and SUI, X. (2022). Doubly robust estimation in causal inference with missing outcomes: With an application to the Aerobics Center Longitudinal Study. *Computational Statistics & Data Analysis* **168** 107399.

XUE, F., MA, R. and LI, H. (2021). Semi-Supervised Statistical Inference for High-Dimensional Linear Regression with Blockwise Missing Data. *arXiv preprint arXiv:2106.03344*.

ZHANG, Y. and BRADIC, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika* **109** 387–403.

ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics* **47** 2538–2566.

ZHANG, Y., CHAKRABORTTY, A. and BRADIC, J. (2023). Double robust semi-supervised inference for the mean: Selection bias under mar labeling with decaying overlap. *Information and Inference: A Journal of the IMA* **12** 2066–2159.

ZHANG, Z., LIU, W., ZHANG, B., TANG, L. and ZHANG, J. (2016). Causal inference with missing exposure information: Methods and applications to an obstetric study. *Statistical Methods in Medical Research* **25** 2053–2066.

## SUPPLEMENT TO 'THE DECAYING MISSING-AT-RANDOM FRAMEWORK: DOUBLY ROBUST CAUSAL INFERENCE WITH PARTIALLY LABELED DATA'

This supplementary document contains further discussions, additional theoretical results, and proofs of the main results that could not be accommodated in the main paper. All results and notations are numbered and used as in the main text unless stated otherwise. We summarize the key notations used throughout the main paper and the supplement in Table S.1.

*Organization.* The rest of the document is organized as follows. We first provide some additional theoretical results for the estimation of the counterfactual mean and the ATE (as extensions of the results in Section 3.1) in Section A. Then, we provide further discussions on the PS's estimation when the treatment variable $T$ is missing in Section B. In Section C, we prove Lemma 2.1, which demonstrates several representations of $\theta_1$. In Section D, we show the theoretical results for the general DR-DMAR SS estimator proposed in Sections 3 and A. Before we show the results in Section 4, we first provide some important preliminary lemmas in Section E. In Section F, we demonstrate the theoretical properties of the targeted bias-reducing PS and OR nuisance estimators based on the semi-parametric approach, as in Section 4.3. Lastly, we prove the asymptotic results (in Section 4.5) for the semi-parametric bias-reduced DR-DMAR SS estimator in Section G.

## APPENDIX A: ADDITIONAL THEORETICAL RESULTS UNDER THE GENERAL FRAMEWORK

As the ATE parameter can be represented as the difference between two counterfactual means that $\mu_0 = \mathbb{E}\{Y(1)\} - \mathbb{E}\{Y(0)\} = \theta_1 - \theta_0$, we first establish the theoretical results for the estimation (and inference) of $\theta_1$ (Theorem A.1). Analogous results also hold for the estimation of $\theta_0$, and the final asymptotic theory for the estimation of $\mu_0$ (Corollaries A.2 and A.3) follows if we combine the results for the estimation of $\theta_1$ and $\theta_0$ together.

In the following, we demonstrate the asymptotic results for the DR-DMAR SS estimator $\widehat{\theta}_{1,\text{ss}}$ of $\theta_1$. For the sake of notational simplicity, we let

$$\Gamma := \Gamma^{(1)},\ \gamma_N := \gamma_{N,1},\ \gamma_N(\cdot) := \gamma_N(1,\cdot),\ m(\cdot) := m(1,\cdot),$$
$$\gamma_N^*(\cdot) := \gamma_N^*(1,\cdot),\ m^*(\cdot) := m^*(1,\cdot),\ \widetilde{\gamma}_N(\cdot) := \widetilde{\gamma}_N(1,\cdot),\ \widetilde{m}(\cdot) := \widetilde{m}(1,\cdot).$$

We assume the following conditions.

ASSUMPTION 8 (High-level conditions on the nuisance function estimators). For each $j \in \{0,1\}$, consider the full data estimators $\widehat{m}(j,\cdot)$ and $\widehat{\gamma}_N(j,\cdot)$ of $m(j,\cdot)$ and $\gamma_N(j,\cdot)$, and suppose they have some limits $m^*(j,\cdot)$ and $\gamma_N^*(j,\cdot)$ with, either $m^*(j,\cdot) = m(j,\cdot)$ or $\gamma_N^*(j,\cdot) = \gamma_N(j,\cdot)$ but not necessarily both.

Assume the following conditions hold (recall $\mathbb{E}_{\mathbf{X}}(\cdot)$ from Section 1.4): for each $j \in \{0,1\}$,

$$\text{(A.1)}\quad \mathbb{E}_{\mathbf{X}}\left[\frac{a_{N,j}}{\gamma_N(j,\mathbf{X})}\{\widehat{m}(j,\mathbf{X}) - m^*(j,\mathbf{X})\}^2\right] = O_p(\alpha_{N,j}^2) \quad \text{with some } \alpha_{N,j} = o(1),$$

$$\text{(A.2)}\quad \mathbb{E}_{\mathbf{X}}\left[\frac{a_{N,j}}{\gamma_N(j,\mathbf{X})}\left\{1 - \frac{\gamma_N^*(\mathbf{X})}{\widehat{\gamma}_N(j,\mathbf{X})}\right\}^2\right] = O_p(\beta_{N,j}^2) \quad \text{with some } \beta_{N,j} = o(1),$$

$$\text{(A.3)}\quad \mathbb{E}_{\mathbf{X}}\{\widehat{m}(j,\mathbf{X}) - m(j,\mathbf{X})\}^2 = O_p(c_{N,j}^2) \quad \text{with some } c_{N,j} = o(1), \text{ and}$$

$$\text{(A.4)}\quad \mathbb{E}_{\mathbf{X}}\left\{1 - \frac{\gamma_N(j,\mathbf{X})}{\widehat{\gamma}_N(j,\mathbf{X})}\right\}^2 = O_p(d_{N,j}^2) \quad \text{with some } d_{N,j} = o(1).$$

TABLE S.1
*Table of notations (we let $j \in \{0,1\}$ and $k \in \{1,2\}$ in the following table)*

| Notation | Description |
|---|---|
| $\mathbf{X}_i, \mathbf{X}$ | The vector of covariates |
| $T_i, T$ | The treatment indicators |
| $R_i, R$ | The labeling indicators |
| $\Gamma_i^{(j)}, \Gamma^{(j)}, \Gamma_i := \Gamma_i^{(1)}, \Gamma := \Gamma^{(1)}$ | The product indicators |
| $Y_i, Y$ | The observed outcome of interest |
| $Y_i(j), Y(j)$ | The potential outcomes |
| $N$ | The total sample size |
| $n$ | The labeled sample size $n = \sum_{i=1}^{N} R_i$ |
| $d$ | The dimension of the covariates |
| $\mathbb{K}$ | Number of folds |
| $M = N/\mathbb{K}$ | Number of samples in each fold |
| $p_N, p_{N,1}$ | The labeling probabilities $\mathbb{P}(R=1)$ and $\mathbb{P}(R=1 \mid T=1)$ |
| $\gamma_N$ | The product probability $\mathbb{P}(\Gamma=1)$ |
| $a_{N,j}$ | The inverses of the inverse product PS functions' expectations |
| $a_N$ | The smaller rate among $a_{N,1}$ and $a_{N,0}$ |
| $Na_N$ | The 'effective sample size' |
| $\theta_j$ | The counterfactual means $\theta_j := \mathbb{E}\{Y(j)\}$ |
| $\mu_0$ | The ATE parameter $\mu_0 := \theta_1 - \theta_0$ |
| $m(j,\cdot), m(\cdot) := m(1,\cdot)$ | The true OR functions |
| $\pi(\cdot), \pi(j,\cdot)$ | The true treatment PS functions |
| $p_N(\cdot), p_N(j,\cdot), p_{N,1}(\cdot) = p_N(1,\cdot)$ | The true labeling PS functions |
| $\gamma_N(j,\cdot), \gamma_N(\cdot) = \gamma_N(1,\cdot)$ | The true product PS functions |
| $\widehat{m}(j,\cdot), \widehat{m}(\cdot) := \widehat{m}(1,\cdot),$ | The OR estimators |
| $\widehat{\gamma}_N(j,\cdot), \widehat{\gamma}_N(\cdot) := \widehat{\gamma}_N(1,\cdot)$ | The product PS estimators |
| $\widetilde{m}(j,\cdot), \widetilde{\gamma}_N(j,\cdot)$ | The cross-fitted nuisance estimators |
| $m^*(j,\cdot), m^*(\cdot) := m^*(1,\cdot)$ | The limiting OR functions |
| $\gamma_N^*(j,\cdot), \gamma_N^*(\cdot) := \gamma_N^*(1,\cdot)$ | The limiting PS functions |
| $\widehat{\theta}_{j,\mathrm{SS}}$ | The DR-DMAR SS estimator for the counterfactual mean |
| $\widehat{\mu}_{\mathrm{SS}}$ | The DR-DMAR SS estimator for the ATE |
| $g(\cdot)$ | The logistic function |
| $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ | The nuisance parameters based on the parametric approach |
| $\widetilde{\boldsymbol{\alpha}}^*, \boldsymbol{\beta}_{p,1}^*$ | The nuisance parameters based on the semi-parametric approach |
| $s_{\boldsymbol{\alpha}}, s_{\boldsymbol{\beta}}, s_{\widetilde{\boldsymbol{\alpha}}}, s_{p,1}$ | The sparsity levels of $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \widetilde{\boldsymbol{\alpha}}^*, \boldsymbol{\beta}_{p,1}^*$ |
| $\widehat{\gamma}_N^{(k)}, \widehat{\gamma}_N := \widehat{\gamma}_N^{(1)}, \bar{\gamma}_N$ | The estimates of $\gamma_N$ |
| $\widehat{p}_{N,1}^{(k)}, \widehat{p}_{N,1} := \widehat{p}_{N,1}^{(1)}, \bar{p}_{N,1}$ | The estimates of $p_{N,1}$ |
| $\widehat{\boldsymbol{\alpha}}^{(k)}, \widehat{\boldsymbol{\alpha}} := \widehat{\boldsymbol{\alpha}}^{(1)}$ | The OR estimators based on the parametric approach |
| $\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}^{(1)}$ | The PS estimators based on the parametric approach |
| $\widetilde{\boldsymbol{\alpha}}^{(k)}, \widetilde{\boldsymbol{\alpha}} := \widetilde{\boldsymbol{\alpha}}^{(1)}, \overline{\boldsymbol{\alpha}}$ | The OR estimators based on the parametric approach |
| $\widehat{\boldsymbol{\beta}}_{p,1}^{(k)}, \widehat{\boldsymbol{\beta}}_{p,1} := \widehat{\boldsymbol{\beta}}_{p,1}^{(1)}, \bar{\boldsymbol{\beta}}_{p,1}$ | The PS estimators based on the parametric approach |
| $\widehat{\theta}_{\mathrm{j,BRSS}}^{(k)}, \widehat{\theta}_{\mathrm{j,BRSS}}, \widehat{\theta}_{\mathrm{j,SP\text{-}BRSS}}^{(k)}, \widehat{\theta}_{\mathrm{j,SP\text{-}BRSS}}$ | The bias-reduced DR-DMAR SS estimators for $\theta_j$ |
| $\widehat{\mu}_{\mathrm{BRSS}}, \widehat{\mu}_{\mathrm{SP\text{-}BRSS}}$ | The bias-reduced DR-DMAR SS estimator for the ATE |

ASSUMPTION 9 (Tail condition 1). For any $\delta > 0$,

$$a_{N,j}^{-1}\mathbb{E}\left[\left\{\psi_{N,j}^{opt}(\mathbf{Z})\right\}^2 \mathbb{1}_{|\psi_{N,j}^{opt}(\mathbf{Z})|>\delta\sqrt{N/a_{N,j}}}\right] \to 0 \text{ as } N \to \infty.$$

ASSUMPTION 10 (Tail condition 2). For a constant $c > 0$ and each $j \in \{0,1\}$, let

$$N^{-1/c} a_{N,j}^{1+c/2}\mathbb{E}\left\{|\psi_{N,j}^{opt}(\mathbf{Z})|^{2+c}\right\} \to 0 \text{ as } N \to \infty,$$

where $\Sigma_N^{opt} := \mathrm{Var}\{\psi_N^{opt}(\mathbf{Z})\} \asymp a_N^{-1}$ with $\psi_N^{opt}(\mathbf{Z}) := \psi_{N,1}^{opt}(\mathbf{Z}) - \psi_{N,0}^{opt}(\mathbf{Z})$.

Assumptions 8 – 10 are analogous to the required conditions in Theorem 3.2 of Zhang, Chakrabortty and Bradic (2023), where the authors considered the estimation of $\mathbb{E}(Y)$ in a non-causal setup. In our causal setup, if we treat $\widetilde{Y} = Y(1)$ as the outcome of interest, the estimation of $\theta_1 = \mathbb{E}(\widetilde{Y})$ can also be seen as a mean estimation problem with missing outcomes. However, in our causal *and* missing data problem, $\widetilde{Y}$ suffers from two different types of missingness: (a) missing due to the labeling procedure (occurs when $R = 0$) and (b) missing due to the treatment assignment (occurs when $T = 0$). We can only observe $\widetilde{Y}$ if $\Gamma = TR = 1$. Hence, as also indicated in Lemma 2.1, $\Gamma$ can be viewed as the effective labeling indicator. Based on the triple $(\widetilde{Y}, \Gamma, \mathbf{X})$, we show Theorem A.1 and Corollaries A.2 – A.3 applying the results developed in Theorem 3.2 of Zhang, Chakrabortty and Bradic (2023).

THEOREM A.1 (Asymptotic results of $\widehat{\theta}_{1,\mathrm{ss}}$). *Let Assumptions 1 and 8 hold. Let $a_{N,1} = [\mathbb{E}\{\gamma_N^{-1}(\mathbf{X})\}]^{-1}$, assume $Na_{N,1} \to \infty$ and possibly, $a_{N,1} \to 0$, as $N \to \infty$. Further, with $m^*(\cdot)$ and $\gamma_N^*(\cdot)$ as in Assumption 8, define the function $\psi_{N,1}^*(\mathbf{Z})$ as in (A.10) and $\psi_{N,1}^{opt}(\mathbf{Z}) := m(j,\mathbf{X}) - \theta_j + \Gamma^{(j)}\{Y(j) - m(j,\mathbf{X})\}/\gamma_N(j,\mathbf{X})$. Assume that $\psi_{N,1}^*(\mathbf{Z}) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ and note that $\mathbb{E}\{\psi_{N,1}^*(\mathbf{Z})\} = 0$ whenever $m^*(\cdot) = m(\cdot)$ or $\gamma_N^*(\cdot) = \gamma_N(\cdot)$ but not necessarily both. Let $\Sigma_{N,1}^* := Var\{\psi_{N,1}^*(\mathbf{Z})\}$ and $\Sigma_{N,1}^{opt} := Var\{\psi_{N,1}^{opt}(\mathbf{Z})\}$. Then the properties of $\widehat{\theta}_{1,ss}$ as defined in (3.1) under different cases are as follows.*

*(a) Suppose $\gamma_N^*(\cdot) = \gamma_N(\cdot)$ and $m^*(\cdot) = m(\cdot)$. Assume $\mathbb{E}[\{Y(1) - m(\mathbf{X})\}^2|\mathbf{X}] \le C < \infty$. Then, $\widehat{\theta}_{1,ss}$ satisfies the following ('optimal') asymptotic linear expansion:*

$$(\widehat{\theta}_{1,ss} - \theta_1) = \frac{1}{N}\sum_{i=1}^N \psi_{N,1}^{opt}(\mathbf{Z}_i) + O_p\left(\frac{\alpha_{N,1}}{\sqrt{Na_{N,1}}} + \frac{\beta_{N,1}}{\sqrt{Na_{N,1}}}\right) + D_N,$$

$$\textit{where } |D_N| := \left|\frac{1}{N}\sum_{i=1}^N \left\{\frac{\Gamma_i}{\widetilde{\gamma}_N(\mathbf{X}_i)} - \frac{\Gamma_i}{\gamma_N^*(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m^*(\mathbf{X}_i)\}\right| = O_p\left(c_{N,1}d_{N,1}\right),$$

*and we note further that $\mathbb{E}\{\psi_{N,1}^{opt}(\mathbf{Z})\} = 0$ and $\Sigma_{N,1}^{opt} := Var\{\psi_{N,1}^{opt}(\mathbf{Z})\} \asymp a_{N,1}^{-1}$. Further, assume $\mathbb{E}[\{Y(1) - m(\mathbf{X})\}^2|\mathbf{X}] \ge c > 0$ and $Var\{Y(1)\} \le C < \infty$. Then, under Assumption 9, and as long as the product rate $c_{N,1}d_{N,1}$ from (A.3) and (A.4) additionally satisfies: $c_{N,1}d_{N,1} = o(1/\sqrt{Na_{N,1}})$, we have:*

$$\sqrt{Na_{N,1}}(\widehat{\theta}_{1,ss} - \theta_1) = O_p(1) \ \textit{and} \ \sqrt{N}\left(\Sigma_{N,1}^{opt}\right)^{-1/2}(\widehat{\theta}_{1,ss} - \theta_1) \xrightarrow{d} \mathcal{N}(0,1).$$

*Further, let Assumption 10 hold. Then, as $N, d \to \infty$,*

$$\widehat{\Sigma}_{N,1} = \Sigma_{N,1}^{opt}\{1 + o_p(1)\} \ \textit{and} \ \sqrt{N}\left(\widehat{\Sigma}_{N,1}^{opt}\right)^{-1/2}(\widehat{\theta}_{1,ss} - \theta_1) \xrightarrow{d} \mathcal{N}(0,1).$$

*(b) Suppose either $\gamma_N^*(\cdot) = \gamma_N(\cdot)$ or $m^*(\cdot) = m(\cdot)$, but not necessarily both. Assume $\gamma_N^*(\cdot) \ge c\gamma_N(\cdot)$ with some $c > 0$. Then, $\widehat{\theta}_{1,ss}$ satisfies the following expansion:*

$$(\widehat{\theta}_{1,ss} - \theta_1) = \frac{1}{N}\sum_{i=1}^N \psi_{N,1}^*(\mathbf{Z}_i) + O_p\left(\frac{\alpha_{N,1}}{\sqrt{Na_{N,1}}} + \frac{\beta_{N,1}}{\sqrt{Na_{N,1}}}\right) + D_N + \widehat{\Delta}_{N,1},$$

*where*

$$\text{(A.5)} \quad \widehat{\Delta}_{N,1} := \frac{1}{N}\sum_{i=1}^N \frac{\Gamma_i}{\gamma_N(\mathbf{X}_i)}\left\{1 - \frac{\gamma_N(\mathbf{X}_i)}{\widetilde{\gamma}_N(\mathbf{X}_i)}\right\}\{m^*(\mathbf{X}_i) - m(\mathbf{X}_i)\} \ \textit{if} \ \gamma_N^*(\cdot) = \gamma_N(\cdot),$$

$$\text{(A.6)} \quad \text{or } \widehat{\Delta}_{N,1} := \frac{1}{N}\sum_{i=1}^{N} \frac{\Gamma_i}{\gamma_N(\mathbf{X}_i)}\left\{1 - \frac{\gamma_N(\mathbf{X}_i)}{\gamma_N^*(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m(\mathbf{X}_i)\} \text{ if } m^*(\cdot) = m(\cdot)$$

*Assume* $0 < c \leq \mathbb{E}[\{Y(1) - m^*(\mathbf{X})\}^2|\mathbf{X}] \leq C < \infty$*,* $Var\{Y(1)\} \leq C < \infty$*, and* $\gamma_N^*(\cdot) \leq C\gamma_N(\cdot)$*. We note further that* $\mathbb{E}\{\psi_{N,1}^*(\mathbf{Z}_i)\} = 0$ *and* $\Sigma_{N,1}^* := Var\{\psi_{N,1}^*(\mathbf{Z})\} \asymp a_{N,1}^{-1}$*, and in general, under Assumption 8 and accounting for both cases (A.5) and (A.6) above,* $\widehat{\Delta}_{N,1}$ *always satisfies:*

$$\widehat{\Delta}_{N,1} = O_p\left(d_{N,1}\mathbb{1}_{m^*(\cdot)\neq m(\cdot)} + c_{N,1}\mathbb{1}_{\gamma_N^*(\cdot)\neq\gamma_N(\cdot)}\right).$$

Corollary A.1 characterizes the asymptotic behavior of the counterfactual mean estimator $\widehat{\theta}_{1,\mathrm{ss}}$, and similar results also hold for $\widehat{\theta}_{0,\mathrm{ss}}$. Combining the results for $\widehat{\theta}_{1,\mathrm{ss}}$ and $\widehat{\theta}_{0,\mathrm{ss}}$, we can establish the theoretical properties for the general DR-DMAR SS ATE estimator $\widehat{\mu}_{\mathrm{ss}} = \widehat{\theta}_{1,\mathrm{ss}} - \widehat{\theta}_{0,\mathrm{ss}}$. When all the models are correctly specified, we provide the asymptotic normality of $\widehat{\mu}_{\mathrm{ss}}$ in Corollary A.2 below.

COROLLARY A.2 (Asymptotic results when both nuisance models are correctly specified). *Let Assumptions 1 and 8 hold and assume* $Na_N \to \infty$ *and possibly,* $a_N \to 0$*, as* $N \to \infty$*. Let* $\mathbb{E}[\{Y(j) - m(j,\mathbf{X})\}^2 \mid \mathbf{X}] \leq C < \infty$*. Then* $\widehat{\mu}_{ss}$ *satisfies the following ('optimal') asymptotic linear expansion:*

$$(\widehat{\mu}_{ss} - \mu_0) = \frac{1}{N}\sum_{i=1}^{N}\psi_N^{opt}(\mathbf{Z}_i) + O_p\left(\frac{\alpha_{N,1}+\beta_{N,1}}{\sqrt{Na_{N,1}}} + \frac{\alpha_{N,0}+\beta_{N,0}}{\sqrt{Na_{N,0}}}\right) + D_{N,1} + D_{N,0}, \text{ where}$$

$$|D_{N,j}| := \left|\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\Gamma_i^{(j)}}{\widetilde{\gamma}_N(j,\mathbf{X}_i)} - \frac{\Gamma_i^{(j)}}{\gamma_N^*(j,\mathbf{X}_i)}\right\}\{\widetilde{m}(j,\mathbf{X}_i) - m^*(j,\mathbf{X}_i)\}\right| = O_p\left(c_{N,j}d_{N,j}\right).$$

*Note that* $\mathbb{E}\{\psi_N^{opt}(\mathbf{Z})\} = 0$ *and* $\Sigma_N^{opt} := Var\{\psi_N^{opt}(\mathbf{Z})\} \asymp a_N^{-1}$*.*

*Assume further that* $\mathbb{E}[\{Y(j) - m(j,\mathbf{X})\}^2 \mid \mathbf{X}] \geq c > 0$ *and* $Var\{Y(j)\} \leq C < \infty$*. Then under Assumption 9, as long as the product-rates satisfy*

$$c_{N,1}d_{N,1} + c_{N,0}d_{N,0} = o(1/\sqrt{Na_N}),$$

*we have:*

$$\text{(A.7)} \qquad \sqrt{Na_N}(\widehat{\mu}_{ss} - \mu_0) = O_p(1) \text{ and } \sqrt{N}\left(\Sigma_N^{opt}\right)^{-1/2}(\widehat{\mu}_{ss} - \mu_0) \xrightarrow{d} \mathcal{N}(0,1).$$

*If we assume further that* $\mathbb{E}_{\mathbf{X}}\left[\frac{a_{N,j}}{\gamma_N(j,\mathbf{X})}\left\{1 - \frac{\gamma_N(j,\mathbf{X})}{\widetilde{\gamma}_N(j,\mathbf{X})}\right\}^2\{\widehat{m}(j,\mathbf{X}) - m(j,\mathbf{X})\}^2\right] = o_p(1)$ *and Assumption 10 holds, then, as* $N, d \to \infty$*,*

$$\text{(A.8)} \qquad \widehat{\Sigma}_N = \Sigma_N^{opt}\{1 + o_p(1)\} \text{ and } \sqrt{N}\left(\widehat{\Sigma}_N^{opt}\right)^{-1/2}(\widehat{\mu}_{ss} - \mu_0) \xrightarrow{d} \mathcal{N}(0,1),$$

*where we propose the following plug-in estimate of the asymptotic variance,* $\Sigma_N^{opt}$*:*

(A.9)

$$\widehat{\Sigma}_N := \frac{1}{N}\sum_{i=1}^{N}\left[\widetilde{m}(1,\mathbf{X}_i) - \widetilde{m}(0,\mathbf{X}_i) - \widehat{\mu}_{ss} + \frac{\Gamma_i^{(1)}\{Y_i - \widetilde{m}(1,\mathbf{X}_i)\}}{\widetilde{\gamma}_N(1,\mathbf{X}_i)} - \frac{\Gamma_i^{(0)}\{Y_i - \widetilde{m}(0,\mathbf{X}_i)\}}{\widetilde{\gamma}_N(0,\mathbf{X}_i)}\right]^2.$$

In the following, we further provide a full characterization of $\widehat{\mu}_{\mathrm{ss}}$ when at least one nuisance model is correctly specified (but not necessarily both).

COROLLARY A.3 (Consistency of $\widehat{\mu}_{\mathrm{SS}}$ when one nuisance model is correctly specified). *For each $j \in \{0,1\}$, with $m^*(j,\cdot)$ and $\gamma_N^*(j,\cdot)$ as in Assumption 8, suppose either $\gamma_N^*(j,\cdot) = \gamma_N(j,\cdot)$ or $m^*(j,\cdot) = m(j,\cdot)$, but not necessarily both. Let Assumptions 1 and 8 hold. Assume $Na_N \to \infty$ and possibly, $a_N \to 0$, as $N \to \infty$. Define*

$$\psi_{N,j}^*(\mathbf{Z}) := m^*(j,\mathbf{X}) - \theta_j + \frac{\Gamma^{(j)}}{\gamma_N^*(j,\mathbf{X})}\{Y(j) - m^*(j,\mathbf{X})\} \tag{A.10}$$

*Let $\psi_N^*(\mathbf{Z}) = \psi_{N,1}^*(\mathbf{Z}) - \psi_{N,0}^*(\mathbf{Z})$ and $\Sigma_N^* := Var\{\psi_N^*(\mathbf{Z})\}$. Assume $\gamma_N^*(j,\cdot) \geq c\gamma_N(j,\cdot)$ with some $c > 0$. Then, $\widehat{\mu}_{ss}$ satisfies the following expansion:*

$$\widehat{\mu}_{ss} - \mu_0 = \frac{1}{N}\sum_{i=1}^{N} \psi_N^*(\mathbf{Z}_i) + O_p\left(\frac{\alpha_{N,1} + \beta_{N,1}}{\sqrt{Na_{N,1}}} + \frac{\alpha_{N,0} + \beta_{N,0}}{\sqrt{Na_{N,0}}}\right) + D_N + \widehat{\Delta}_{N,1} + \widehat{\Delta}_{N,0},$$

*where $\mathbb{E}\{\psi_N^*(\mathbf{Z})\} = 0$ and $|D_N| = |D_{N,1} + D_{N,0}| = O_p(c_{N,1}d_{N,1} + c_{N,0}d_{N,0})$. If $\gamma_N^*(j,\cdot) = \gamma_N(j,\cdot)$,*

$$\widehat{\Delta}_{N,j} := \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\Gamma_i^{(j)}}{\gamma_N(j,\mathbf{X}_i)} - \frac{\Gamma_i^{(j)}}{\widetilde{\gamma}_N(j,\mathbf{X}_i)}\right\}\{m^*(j,\mathbf{X}_i) - m(j,\mathbf{X}_i)\};$$

*If $m^*(j,\cdot) = m(j,\cdot)$,*

$$\widehat{\Delta}_{N,j} := \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\Gamma_i^{(j)}}{\gamma_N(j,\mathbf{X}_i)} - \frac{\Gamma_i^{(j)}}{\gamma_N^*(j,\mathbf{X}_i)}\right\}\{\widetilde{m}(j,\mathbf{X}_i) - m(j,\mathbf{X}_i)\}.$$

*Assume $c \leq \mathbb{E}[\{Y(j) - m^*(j,\mathbf{X})\}^2 \mid \mathbf{X}] \leq C$, $Var\{Y(j)\} \leq C$, and $\gamma_N^*(j,\cdot) \leq C\gamma_N(j,\cdot)$. We note further that $\mathbb{E}\{\psi_N^*(\mathbf{Z}_i)\} = 0$ and $\Sigma_N^* := Var\{\psi_N^*(\mathbf{Z})\} \asymp a_N^{-1}$, and in general, accounting for both cases $\gamma_N^*(j,\cdot) = \gamma_N(j,\cdot)$ or $m^*(j,\cdot) = m(j,\cdot)$ above, $\widehat{\Delta}_N := \widehat{\Delta}_{N,1} + \widehat{\Delta}_{N,0}$ always satisfies:*

$$\widehat{\Delta}_N = O_p\left(\sum_{j=0}^{1} c_{N,j}\mathbb{1}_{\gamma_N^*(j,\cdot)\neq\gamma_N(j,\cdot)} + \sum_{j=0}^{1} d_{N,j}\mathbb{1}_{m^*(j,\cdot)\neq m(j,\cdot)}\right).$$

## APPENDIX B: ESTIMATION OF THE PS WHEN $T$ IS MISSING

In this section, we discuss the estimation of the product PS function $\gamma_N(\mathbf{x}) = \mathbb{P}(\Gamma = 1 \mid \mathbf{X} = \mathbf{x})$ ($\Gamma = TR = TR_TR_Y$) under the case that $T$ and $Y$ are both possibly missing (Section 3.3). Since $T$ is no longer always observable, we consider the following representations: for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}\gamma_N(\mathbf{x}) &= \mathbb{P}(TR_T = 1 \mid \mathbf{X} = \mathbf{x})\mathbb{P}(R_Y = 1 \mid TR_T = 1, \mathbf{X} = \mathbf{x}) := \gamma_{T,N}(\mathbf{x})\gamma_{Y,N}(1,\mathbf{x}) \\ &= \mathbb{P}(R_Y = 1 \mid \mathbf{X} = \mathbf{x})\mathbb{P}(TR_T = 1 \mid R_Y = 1, \mathbf{X} = \mathbf{x}) := \gamma_{Y,N}(\mathbf{x})\gamma_{T,N}(1,\mathbf{x}).\end{aligned}$$

We first focus on Setting d of Section 3.3 and consider the following three models for $\gamma_N(\cdot)$.

*Model 1': a logistic $\gamma_N(\cdot)$.* This is the same as in Model 1 of Section 3.2.

*Model 2': a logistic $\gamma_{T,N}(\cdot)$ and a logistic $\gamma_{Y,N}(1,\cdot)$.* Consider the following offset logistic models: with some $\boldsymbol{\beta}_T, \boldsymbol{\beta}_{Y,1} \in \mathbb{R}^d$,

$$\gamma_{T,N}(\mathbf{x}) = \frac{\gamma_{T,N}\exp(\mathbf{x}^T\boldsymbol{\beta}_T)}{1 + \gamma_{T,N}\exp(\mathbf{x}^T\boldsymbol{\beta}_T)}, \quad \gamma_{Y,N}(1,\mathbf{x}) = \frac{\gamma_{Y,N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{Y,1})}{1 + \gamma_{Y,N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{Y,1})},$$

where $\gamma_{T,N} := \mathbb{E}(TR_T)$ and $\gamma_{Y,N,1} := \mathbb{E}(R_Y \mid TR_T = 1) = \gamma_N/\gamma_{T,N}$, where $\gamma_N = \mathbb{E}(\Gamma)$. Here, we allow decaying PS functions with $\gamma_{T,N}, \gamma_{Y,N,1}, \gamma_N \to 0$ as $N, d \to \infty$, and how fast the PS functions decay may differ from each other. For instance, when $Y$ is always observable (i.e., the degenerate Setting b of Section 3.3) or $Y$ and $T$ are always observable simultaneously (i.e., the degenerate Setting c of Section 3.3), we have $\gamma_{Y,N,1} = 1$ and $\gamma_{T,N} = \gamma_N$. As in Section 3.2, $\gamma_{T,N}(\cdot)$ and $\gamma_{Y,N}(1,\cdot)$ can be estimated by offset logistic estimators (Zhang, Chakrabortty and Bradic, 2023) with consistency rates $O_p(\sqrt{\|\boldsymbol{\beta}_T\|_0 \log(d)/(N\gamma_{T,N})})$ and $O_p(\sqrt{\|\boldsymbol{\beta}_{Y,1}\|_0 \log(d)/(N\gamma_N)})$. As a result, we can choose $d_{N,1}$, defined as (A.4), as $d_{N,1} = O(\sqrt{\|\boldsymbol{\beta}_T\|_0 \log(d)/(N\gamma_{T,N})} + \sqrt{\|\boldsymbol{\beta}_{Y,1}\|_0 \log(d)/(N\gamma_N)})$ for the estimation of $\gamma_N(\cdot)$. Consider a linear OR model with sparsity level $s_{\boldsymbol{\alpha},1}$, to achieve (A.7), we need the product-sparsity conditions $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_T\|_0 = o(N\gamma_{T,N}/\{\log(d)\}^2)$ and $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_{Y,1}\|_0 = o(N\gamma_N/\{\log(d)\}^2)$.

*Model 3': a logistic $\gamma_{T,N}(1,\cdot)$ and a logistic $\gamma_{Y,N}(\cdot)$.* Consider the following offset logistic models: with some $\boldsymbol{\beta}_{T,1}, \boldsymbol{\beta}_Y \in \mathbb{R}^d$,

$$\gamma_{T,N}(1,\mathbf{x}) = \frac{\gamma_{T,N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{T,1})}{1+\gamma_{T,N,1}\exp(\mathbf{x}^T\boldsymbol{\beta}_{T,1})}, \quad \gamma_{Y,N}(\mathbf{x}) = \frac{\gamma_{Y,N}\exp(\mathbf{x}^T\boldsymbol{\beta}_Y)}{1+\gamma_{Y,N}\exp(\mathbf{x}^T\boldsymbol{\beta}_Y)},$$

where $\gamma_{Y,N} := \mathbb{E}(R_Y)$ and $\gamma_{T,N,1} := \mathbb{E}(TR_T \mid R_Y = 1) = \gamma_N/\gamma_{Y,N}$. Similarly as in Model 2', to achieve (A.7), we need the product-sparsity conditions $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_Y\|_0 = o(N\gamma_{Y,N}/\{\log(d)\}^2)$ and $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_{T,1}\|_0 = o(N\gamma_N/\{\log(d)\}^2)$.

As discussed in Section 3.3, Settings a, b, c (see definitions therein) are all special cases of Setting d. In the following, we further discuss the estimation of the product PS function under the degenerate Settings a, b, and c.

The estimation of $\gamma_N(\cdot)$ under Setting a has been discussed in Section 3.2. Under the degenerate Setting a, we have $R_T \equiv 1$ and hence $\gamma_{T,N} \asymp 1$, and Models 1' and 3' above are the same as Models 1 and 3 of Section 3.2, respectively. In the following, we revisit Setting a under Model 2'. Recall that, to achieve (A.7), we need the product-sparsity conditions $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_T\|_0 = o(N\gamma_{T,N}/\{\log(d)\}^2)$ (where $\gamma_{T,N} \asymp 1$) and $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_{Y,1}\|_0 = o(N\gamma_N/\{\log(d)\}^2)$. Note that, in our decaying MAR setting, the 'effect sample size' is $Na_N \asymp N\gamma_N$ for sub-Gaussian $\mathbf{X}$ (Zhang, Chakrabortty and Bradic, 2023). Hence, we refer the 'high-dimensional' setting as the case that $N\gamma_N \ll d$, where $N \gg d$ is still possible since $\gamma_N$ is potentially very small. When the total sample size is large enough in that $N \gg s_{\boldsymbol{\alpha},1} d\{\log(d)\}^2$, the first product-sparsity condition $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_T\|_0 = o(N\gamma_{T,N}/\{\log(d)\}^2)$ is always satisfied even for a dense $\boldsymbol{\beta}_T$. As a result, here we only require the second product-rate condition that $s_{\boldsymbol{\alpha},1}\|\boldsymbol{\beta}_{Y,1}\|_0 = o(N\gamma_N/\{\log(d)\}^2)$.

For the degenerate Setting b, we have $R_Y \equiv 1$ and hence $\Gamma = TR_T$, $\gamma_{T,N}(1,\cdot) \equiv \gamma_{Y,N}(\cdot) \equiv 1$. That is, Models 2' and 3' degenerate to Model 1' where $\gamma_N(\cdot)$ is modeled through a logistic function directly. We can see that the observations $(Y_i)_{i:R_{T,i}=0}$ are redundant for the ATE estimation. This is because if $T_i$ is missing, we can not tell the observed outcome $Y_i = Y_i(T_i)$ corresponds to which potential outcome among $Y_i(1)$ and $Y_i(0)$.

For the degenerate Setting c, we have $R_T \equiv R_Y$ and hence $\gamma_{Y,N}(1,\cdot) \equiv 1$ that Model 2' degenerates to Model 1' (and hence the same as Model 1 of Section 3.2). Additionally, we also have $\gamma_{Y,N}(\mathbf{x}) = \mathbb{P}(R = 1 \mid \mathbf{X} = \mathbf{x}) = p_N(\mathbf{x})$ and $\gamma_{T,N}(1,\mathbf{x}) = \mathbb{P}(T = 1 \mid R = 1, \mathbf{X} = \mathbf{x}) = \pi(1,\mathbf{x})$, i.e., Model 3' is the same as Model 3 of Section 3.2.

Note that, the general formula for the estimator $\widehat{\mu}_{\text{ss}}$ (as a function of $\widehat{\gamma}_N(\cdot)$) and the general asymptotic results are the same among Settings a – d. The difference is only on the approaches for the PS's estimation and the conditions required for the PS models, e.g., the sparsity conditions for logistic parameters.

## APPENDIX C: PROOF OF LEMMA 2.1

Proof of Lemma 2.1. The regression representation can be denoted as

$$\theta_1 = \mathbb{E}\{Y(1)\} = \mathbb{E}[\mathbb{E}\{Y(1) \mid \mathbf{X}\}] = \mathbb{E}\{m(\mathbf{X})\},$$

and since $Y(1) \perp\!\!\!\perp T \mid \mathbf{X}$, we have

$$m(\mathbf{X}) = \mathbb{E}\{Y(1) \mid \mathbf{X}\} = \mathbb{E}\{Y(1) \mid \mathbf{X}, T = 1\}.$$

Additionally, since $R \perp\!\!\!\perp Y \mid (T, \mathbf{X})$, we have $R \perp\!\!\!\perp Y(1) \mid (\mathbf{X}, T = 1)$. Therefore,

$$\mathbb{E}(\Gamma Y \mid \mathbf{X}) = \mathbb{E}(RTY \mid \mathbf{X}) = \mathbb{E}\{Y(1) \mid \mathbf{X}, T = 1, R = 1\}\mathbb{P}\{RT = 1 \mid \mathbf{X}\} \tag{C.1}$$

$$= \mathbb{E}\{Y(1) \mid \mathbf{X}, T = 1\}\mathbb{P}(\Gamma = 1 \mid \mathbf{X}) = m(\mathbf{X})\gamma_N(\mathbf{X}). \tag{C.2}$$

Hence, we have the following IPW representation:

$$\theta_1 = \mathbb{E}\{m(\mathbf{X})\} = \mathbb{E}\left\{\frac{m(\mathbf{X})\gamma_N(\mathbf{X})}{\gamma_N(\mathbf{X})}\right\} = \mathbb{E}\left\{\frac{\mathbb{E}(\Gamma Y \mid \mathbf{X})}{\gamma_N(\mathbf{X})}\right\} = \mathbb{E}\left\{\frac{\Gamma Y}{\gamma_N(\mathbf{X})}\right\}.$$

Let either $m^*(\cdot) = m(\cdot)$ or $\gamma_N^*(\cdot) = \gamma_N(\cdot)$ hold. Then, it follows that

$$\begin{aligned}
&\mathbb{E}\left[m^*(\mathbf{X}) + \frac{\Gamma}{\gamma_N^*(\mathbf{X})}\{Y - m^*(\mathbf{X})\}\right] - \theta_1 \\
&= \mathbb{E}\left\{\mathbb{E}\left(\left[m^*(\mathbf{X}) + \frac{\Gamma}{\gamma_N^*(\mathbf{X})}\{Y(1) - m^*(\mathbf{X})\}\right] \mid \mathbf{X}\right)\right\} - \theta_1 \\
&\overset{(i)}{=} \mathbb{E}\left[m^*(\mathbf{X}) + \frac{\gamma_N(\mathbf{X})}{\gamma_N^*(\mathbf{X})}\{m(\mathbf{X}) - m^*(\mathbf{X})\}\right] - \mathbb{E}\{m(\mathbf{X})\} \\
&= \mathbb{E}\left[\left\{\frac{\gamma_N(\mathbf{X})}{\gamma_N^*(\mathbf{X})} - 1\right\}\{m(\mathbf{X}) - m^*(\mathbf{X})\}\right] = 0,
\end{aligned}$$

where (i) holds by (C.2) and $\mathbb{E}(\Gamma \mid \mathbf{X}) = \gamma_N(\mathbf{X})$. ■

## APPENDIX D: PROOF OF RESULTS IN SECTION A

Proof of Theorem A.1. We first show that $Y(1) \perp\!\!\!\perp \Gamma \mid \mathbf{X}$ under Assumption 1. Observe that

$$\begin{aligned}
\mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 1\} &\overset{(i)}{=} \mathbb{E}(Y \mid \mathbf{X}, R = T = 1) \overset{(ii)}{=} \mathbb{E}(Y \mid \mathbf{X}, T = 1) \\
&\overset{(iii)}{=} \mathbb{E}\{Y(1) \mid \mathbf{X}, T = 1\} \overset{(iv)}{=} \mathbb{E}\{Y(1) \mid \mathbf{X}\},
\end{aligned}$$

where (i) holds since $Y = Y(T)$ and $\Gamma = TR$; (ii) holds since $R \perp\!\!\!\perp Y \mid (T, X)$; (iii) holds since $Y = Y(T)$; (iv) holds since $T \perp\!\!\!\perp Y(1) \mid \mathbf{X}$. In addition, by the tower rule,

$$\begin{aligned}
\mathbb{E}\{Y(1) \mid \mathbf{X}\} &= \mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 1\}\gamma_N(\mathbf{X}) + \mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 0\}\{1 - \gamma_N(\mathbf{X})\} \\
&= \mathbb{E}\{Y(1) \mid \mathbf{X}\}\gamma_N(\mathbf{X}) + \mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 0\}\{1 - \gamma_N(\mathbf{X})\}.
\end{aligned}$$

It follows that

$$\mathbb{E}\{Y(1) \mid \mathbf{X}\}\{1 - \gamma_N(\mathbf{X})\} = \mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 0\}\{1 - \gamma_N(\mathbf{X})\}.$$

Since $\gamma_N(\mathbf{X}) = \pi(\mathbf{X})p_N(1, \mathbf{X}) > 0$ almost surely under Assumption 1, we have

$$\mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 0\} = \mathbb{E}\{Y(1) \mid \mathbf{X}\} = \mathbb{E}\{Y(1) \mid \mathbf{X}, \Gamma = 1\} \text{ almost surely.}$$

That is, $Y(1) \perp\!\!\!\perp \Gamma \mid \mathbf{X}$, i.e., the missing at random (MAR) condition holds if we consider $\widetilde{Y} = Y(1)$ and $\Gamma$ as the outcome variable and labeling indicator. The remaining results hold as long as we apply the results in Theorem 3.2 of Zhang, Chakrabortty and Bradic (2023), with $(Y, R, \mathbf{X})$ replaced with the triple $(\widetilde{Y}, \Gamma, \mathbf{X})$. ■

Repeating the proofs above, we obtain analogous results for $\widehat{\theta}_{0,\mathrm{ss}}$ as in Theorems A.1. Together with the results for $\widehat{\theta}_{0,\mathrm{ss}}$, Corollary A.2 is a corollary of part (a) of Theorem A.1, and Corollary A.3 is a corollary of part (b) of Theorem A.1.

## APPENDIX E: PRELIMINARY LEMMAS

We first provide some preliminary lemmas before we analyze the biased reduced DR-DMAR SS estimators and the considered nuisance estimators.

LEMMA E.1 (Lemma D.1 of Zhang, Chakrabortty and Bradic (2023)). *Let $(X_N)_{N\geq 1}$ and $(Y_N)_{N\geq 1}$ be sequences of random variables in $\mathbb{R}$. If $\mathbb{E}(|X_N|^r \mid Y_N) = O_p(1)$ for any $r \geq 1$, then $X_N = O_p(1)$.*

LEMMA E.2. *The following are some useful properties regarding the $\psi_\alpha$-norms.*

*(a) If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_2} \leq \|Y\|_{\psi_2}$. If $|X| \leq C$ a.s. for some constant $C > 0$, then $\|X\|_{\psi_2} \leq \{\log(2)\}^{-1/2}C$.*

*(b) If $\|X\|_{\psi_2} \leq \sigma$, then $\mathbb{P}(|X| > t) \leq 2\exp(-t^2/\sigma^2)$ for all $t \geq 0$.*

*(c) If $\|X\|_{\psi_\alpha} \leq \sigma$ for some $(\alpha, \sigma) > 0$, then $\mathbb{E}(|X|^m) \leq C_\alpha^m \sigma^m m^{m/\alpha}$ for all $m \geq 1$, for some constant $C_\alpha$ depending only on $\alpha$. In particular, if $\|X\|_{\psi_2} \leq \sigma$, $\mathbb{E}(|X|^m) \leq 2\sigma^m \Gamma(m/2+1)$, for all $m \geq 1$, where $\Gamma(a) := \int_0^\infty x^{a-1}\exp(-x)dx$ denotes the Gamma function. Hence, $\mathbb{E}(|X|) \leq \sigma\sqrt{\pi}$ and $\mathbb{E}(|X|^m) \leq 2\sigma^m(m/2)^{m/2}$ for $m \geq 2$.*

*(d) For any $\alpha, \beta > 0$, let $\gamma := (\alpha^{-1} + \beta^{-1})^{-1}$. Then, for any $X, Y$ with $\|X\|_{\psi_\alpha} < \infty$ and $\|Y\|_{\psi_\beta} < \infty$, $\|XY\|_{\psi_\gamma} < \infty$ and $\|XY\|_{\psi_\gamma} < \|X\|_{\psi_\alpha}\|Y\|_{\psi_\beta}$.*

*(e) Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector with $\sup_{1\leq j\leq p}\|\mathbf{X}(j)\|_{\psi_\alpha} \leq \sigma$. Then, $\|\|\mathbf{X}\|_\infty\|_{\psi_\alpha} \leq \sigma\{\log(p) + 2\}^{1/\alpha}$.*

Lemma E.2 follows from Lemma D.1 of Chakrabortty et al. (2019).

LEMMA E.3. *If $X \in \mathbb{R}$ is a random variable and there exists constants $a_1, a_2, a_3, a_4 > 0$, such that*

$$\mathbb{P}(|X| > a_1u^2 + a_2u + a_3) \leq a_4\exp(-u^2), \quad \forall u > 0.$$

*Then,*

$$\mathbb{E}(|X|) \leq a_3 + a_4(4a_1 + \sqrt{\pi}a_2).$$

PROOF OF LEMMA E.3. Observe that

$$\begin{aligned}
\mathbb{E}(|X|) &= \int_0^\infty \mathbb{P}(|X| > t)dt \\
&= \int_0^{a_3} \mathbb{P}(|X| > t)dt + \int_{a_3}^\infty \mathbb{P}(|X| > t)dt \\
&\leq a_3 + \int_0^\infty \mathbb{P}(|X| > t + a_3)dt.
\end{aligned}$$

For any $t > 0$, let $u > 0$ satisfies $a_1u^2 + a_2u = t$. Then,

$$u = \frac{\sqrt{a_2^2 + 4a_1t} - a_2}{2a_1} = \frac{2t}{\sqrt{a_2^2 + 4a_1t} + a_2} \geq \frac{2t}{\sqrt{4a_1t} + 2a_2}.$$

Hence,

$$\int_0^\infty \mathbb{P}(|X|>t+a_3)dt = \int_0^\infty \mathbb{P}(|X|>a_1u^2+a_2u+a_3)dt \le \int_0^\infty a_4\exp(-u^2)dt$$
$$\le a_4\int_0^\infty \exp\left\{-\frac{4t^2}{(\sqrt{4a_1t}+2a_2)^2}\right\} \le a_4\int_0^\infty \exp\left(-\frac{t^2}{2a_1t+2a_2^2}\right)$$
$$= a_4\int_0^{a_2^2/a_1} \exp\left(-\frac{t^2}{2a_1t+2a_2^2}\right) + a_4\int_{a_2^2/a_1}^\infty \exp\left(-\frac{t^2}{2a_1t+2a_2^2}\right)$$
$$\le a_4\int_0^{a_2^2/a_1} \exp\left(-\frac{t^2}{4a_2^2}\right) + a_4\int_{a_2^2/a_1}^\infty \exp\left(-\frac{t}{4a_1}\right).$$

Notice that

$$\int_0^{a_2^2/a_1} \exp\left(-\frac{t^2}{4a_2^2}\right) \le 2\sqrt{\pi}a_2\frac{1}{2}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sqrt{2}a_2}\exp\left(-\frac{t^2}{4a_2^2}\right) = \sqrt{\pi}a_2$$

and

$$\int_{a_2^2/a_1}^\infty \exp\left(-\frac{t}{4a_1}\right) = 4a_1\exp\left(-\frac{a_2^2}{4a_1^2}\right) \le 4a_1.$$

Therefore,

$$\mathbb{E}(|X|) \le a_3 + a_4(4a_1+\sqrt{\pi}a_2).$$

■

We establish some empirical process results in Lemmas E.4, E.5, and E.6.

LEMMA E.4. *Let $\mathbf{\Omega}\in\mathbb{R}^{n\times n}$, $k_0>0$, and $s\ge 1$. Then, for any $\delta\in(0,1)$, there exists some $k\asymp s$ such that*

$$\sup_{\mathbf{\Delta}_1,\mathbf{\Delta}_2\in\mathcal{C}(s,k_0)\cap\mathbb{S}^{d-1}} |\mathbf{\Delta}_1^T\mathbf{\Omega}\mathbf{\Delta}_2| \le (1-\delta)^{-1}\sup_{\mathbf{\Delta}_1,\mathbf{\Delta}_2\in\Theta_k}|\mathbf{\Delta}_1^T\mathbf{\Omega}\mathbf{\Delta}_2|, \tag{E.1}$$

*where* $\mathbb{S}^{d-1}:=\{\mathbf{\Delta}\in\mathbb{R}^d:\|\mathbf{\Delta}\|_2=1\}$, $\mathcal{C}(s,k_0):=\{\mathbf{\Delta}\in\mathbb{R}^d:\exists S\subset\{1,\dots,d\},\ |S|\le s,\ s.t.\ \|\mathbf{\Delta}_{S^c}\|_1\le k_0\|\mathbf{\Delta}_S\|_1\}$, *and* $\Theta_k:=\{\mathbf{\Delta}\in\mathbb{R}^d:\|\mathbf{\Delta}\|_0\le k,\|\mathbf{\Delta}\|_2=1\}$.

Lemma E.4 holds by repeating the proof of Lemma 16 of Bradic, Wager and Zhu (2019).

LEMMA E.5. *Let $\mathbf{\Omega}\in\mathbb{R}^{d\times d}$. Then, for any $\mathbf{\Delta}\in\mathbb{R}^d$ and $k_0>0$,*

$$|\mathbf{\Delta}^T\mathbf{\Omega}\mathbf{\Delta}| \le \inf_{s\ge 1}\left\{\left(\frac{6\|\mathbf{\Delta}\|_1^2}{k_0^2 s}+4\|\mathbf{\Delta}\|_2^2\right)\sup_{\boldsymbol{\delta}\in\mathcal{C}(s,k_0)\cap\mathbb{S}^{d-1}}|\boldsymbol{\delta}^T\mathbf{\Omega}\boldsymbol{\delta}|\right\}. \tag{E.2}$$

PROOF OF LEMMA E.5. For any $s\le d$, choose some $S\subset\{1,2,\dots,d\}$ satisfying $|S|=s$. For any $\mathbf{\Delta}\in\mathbb{R}^d$, define $\widetilde{\boldsymbol{\delta}}=(\widetilde{\boldsymbol{\delta}}_S^T,\widetilde{\boldsymbol{\delta}}_{S^c}^T)^T\in\mathbb{R}^d$ as

$$\widetilde{\boldsymbol{\delta}}_S=(k_0s)^{-1}\|\mathbf{\Delta}\|_1(1,\dots,1)^T\in\mathbb{R}^s,\quad \widetilde{\boldsymbol{\delta}}_{S^c}=\mathbf{\Delta}_{S^c}\in\mathbb{R}^{d-s}. \tag{E.3}$$

Then, $\|\widetilde{\boldsymbol{\delta}}_S\|_1=\|\mathbf{\Delta}\|_1/k_0$ and $\|\widetilde{\boldsymbol{\delta}}_S\|_2=\|\mathbf{\Delta}\|_1/(k_0\sqrt{s})$. Hence,

$$\|\widetilde{\boldsymbol{\delta}}_{S^c}\|_1=\|\mathbf{\Delta}_{S^c}\|_1\le\|\mathbf{\Delta}\|_1=k_0\|\widetilde{\boldsymbol{\delta}}_S\|_1.$$

That is, $\widetilde{\boldsymbol{\delta}} \in \mathcal{C}(s, k_0)$. Also, note that $\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}} \in \mathcal{C}(s, k_0)$ since $(\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}})_{S^c} = \mathbf{0}$. Hence,

$$\begin{aligned}
|\boldsymbol{\Delta}^T \boldsymbol{\Omega} \boldsymbol{\Delta}| &\le |\widetilde{\boldsymbol{\delta}}^T \boldsymbol{\Omega} \widetilde{\boldsymbol{\delta}}| + |(\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}})^T \boldsymbol{\Omega} (\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}})| + 2|\widetilde{\boldsymbol{\delta}}^T \boldsymbol{\Omega} (\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}})| \\
&\le (\|\widetilde{\boldsymbol{\delta}}\|_2^2 + \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2^2 + 2\|\widetilde{\boldsymbol{\delta}}\|_2 \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2) \sup_{\boldsymbol{\delta} \in \mathcal{C}(s,k_0) \cap \mathbb{S}^{d-1}} \boldsymbol{\delta}^T \boldsymbol{\Omega} \boldsymbol{\delta} \\
&\le \left( \frac{6\|\boldsymbol{\Delta}\|_1^2}{k_0^2 s} + 4\|\boldsymbol{\Delta}\|_2^2 \right) \sup_{\boldsymbol{\delta} \in \mathcal{C}(s,k_0) \cap \mathbb{S}^{d-1}} \sum_{i \in \mathcal{I}_2} (\mathbf{U}_i^T \boldsymbol{\delta})^2 v_i,
\end{aligned}$$

since

$$\begin{aligned}
&\|\widetilde{\boldsymbol{\delta}}\|_2^2 + \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2^2 + 2\|\widetilde{\boldsymbol{\delta}}\|_2 \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2 \le 2\|\widetilde{\boldsymbol{\delta}}\|_2^2 + 2\|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2^2 \\
&\quad = 2\|\widetilde{\boldsymbol{\delta}}_S\|_2^2 + 2\|\widetilde{\boldsymbol{\delta}}_{S^c}\|_2^2 + 2\|\boldsymbol{\Delta}_S - \widetilde{\boldsymbol{\delta}}_S\|_2^2 \le 2\|\widetilde{\boldsymbol{\delta}}_S\|_2^2 + 2\|\boldsymbol{\Delta}_{S^c}\|_2^2 + 4\|\boldsymbol{\Delta}_S\|_2^2 + 4\|\widetilde{\boldsymbol{\delta}}_S\|_2^2 \\
&\quad \le 6\|\widetilde{\boldsymbol{\delta}}_S\|_2^2 + 4\|\boldsymbol{\Delta}\|_2^2 = 6\|\boldsymbol{\Delta}\|_1^2/(k_0^2 s) + 4\|\boldsymbol{\Delta}\|_2^2.
\end{aligned}$$

When $s > d$, since $\boldsymbol{\Delta} \in \mathbb{R}^d = \mathcal{C}(d, k_0) = \mathcal{C}(s, k_0)$, we also have

$$|\boldsymbol{\Delta}^T \boldsymbol{\Omega} \boldsymbol{\Delta}| \le \|\boldsymbol{\Delta}\|_2^2 \sup_{\boldsymbol{\delta} \in \mathcal{C}(s,k_0) \cap \mathbb{S}^{d-1}} |\boldsymbol{\delta}^T \boldsymbol{\Omega} \boldsymbol{\delta}|.$$

To sum up, we conclude that (E.2) holds. ∎

LEMMA E.6. *Let Assumptions 2 and 3 hold. Let $j \in \{1, 2\}$.*

*(a) For any $s = o(M/\log(d))$, we have*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{M^{-1} \sum_{i \in \mathcal{I}_j} (\mathbf{X}_i^T \boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_1^2/s + \|\boldsymbol{\Delta}\|_2^2} = O_p(1).$$

*(b) Let $s_0 := \lceil N\gamma_N / \{\log(d)\log(N)\} \rceil$, then*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{(M\gamma_N)^{-1} \sum_{i \in \mathcal{I}_j} \Gamma_i (\mathbf{X}_i^T \boldsymbol{\Delta})^2}{\|\boldsymbol{\Delta}\|_1^2/s_0 + \|\boldsymbol{\Delta}\|_2^2} \le c \left\{ 1 + \sqrt{\frac{t}{N\gamma_N}} + \frac{t \log(N)}{N\gamma_N} \right\} \tag{E.4}$$

*with probability at least $1 - 3\exp(-t)$ and some constant $c > 0$.*

*(c) Let $f(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ and $|f(\mathbf{x})| \le c$ for all $\mathbf{x} \in \mathcal{X}$ with some constant $c > 0$. Then, for any $s \ge 1$, as $N, d \to \infty$,*

$$\begin{aligned}
&\sup_{\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{(M\gamma_N)^{-1} \left| \boldsymbol{\Delta}_1^T \left[ \sum_{i \in \mathcal{I}_1} \Gamma_i f(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^T - \mathbb{E}\{\Gamma f(\mathbf{X}) \mathbf{X}\mathbf{X}^T\} \right] \boldsymbol{\Delta}_2 \right|}{\|\boldsymbol{\Delta}_1\|_2 \|\boldsymbol{\Delta}_2\|_2 \{\|\boldsymbol{\Delta}_1\|_1^2/(s\|\boldsymbol{\Delta}_1\|_2^2) + \|\boldsymbol{\Delta}_2\|_1^2/(s\|\boldsymbol{\Delta}_2\|_2^2) + 1\}} \\
&\quad = O_p\left( \sqrt{\frac{s \log(d)}{N\gamma_N}} + \frac{s \log(d) \log(N)}{N\gamma_N} \right).
\end{aligned}$$

*(d) Let $r_0 \ge 2$ be a constant. Then, for any $s \ge 1$,*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{(M\gamma_N)^{-1} \sum_{i \in \mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T \boldsymbol{\Delta}|^{r_0}}{\|\boldsymbol{\Delta}\|_1^{r_0}/s^{\frac{r_0}{2}} + \|\boldsymbol{\Delta}\|_2^{r_0}} \le c \left( 1 + \sqrt{\frac{s \log(d)}{M\gamma_N}} + \frac{\{s \log(M) \log(d)\}^{r_0/2}}{M\gamma_N} \right)$$

*with probability at least $1 - 2\exp(-t)$ and some constant $c > 0$. Hence, as $N, d \to \infty$,*

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{(M\gamma_N)^{-1} \sum_{i \in \mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T \boldsymbol{\Delta}|^{r_0}}{\|\boldsymbol{\Delta}\|_1^{r_0}/s^{\frac{r_0}{2}} + \|\boldsymbol{\Delta}\|_2^{r_0}} = O_p\left( 1 + \sqrt{\frac{s \log(d)}{M\gamma_N}} + \frac{\{s \log(M) \log(d)\}^{r_0/2}}{M\gamma_N} \right).$$

*(e) Further assume that* $\|\mathbf{X}\|_\infty \le C_{\mathbf{X}}$. *Let* $(\epsilon_i)_{i\in\mathcal{I}_j} \in \mathbb{R}$ *be i.i.d. random variables satisfying* $\mathbb{E}(|\epsilon_i| \mid \mathbf{X}_i, \Gamma_i = 1) = \mathbb{E}(|\epsilon_i| \mid \mathbf{X}_i)$ *and* $\|\epsilon\|_{\psi_\alpha} \le \sigma_\epsilon$ *with some constants* $\alpha, \epsilon > 0$. *Then, for any* $s \ge 1$,

$$
\begin{aligned}
&(M\gamma_N)^{-1} \sum_{i\in\mathcal{I}_j} \Gamma_i |\epsilon_i| (\mathbf{X}_i^T \mathbf{\Delta})^2 \\
&\qquad \le c\|\mathbf{\Delta}\|_2^2 + c\left[\sqrt{\frac{t + s\log(d)}{N\gamma_N}} + \frac{s\log^{1/\alpha}(M)\{t+s\log(d)\}^{1/\alpha}}{N\gamma_N}\right]\left(\frac{\|\mathbf{\Delta}\|_1^2}{s} + \|\mathbf{\Delta}\|_2^2\right),
\end{aligned}
$$

*uniformly for all* $\mathbf{\Delta} \in \mathbb{R}^d$ *with probability at least* $1 - 3\exp(-t)$ *and some constant* $c > 0$.

*(f) Let* $(\epsilon_i)_{i\in\mathcal{I}_j} \in \mathbb{R}$ *be i.i.d. random variables satisfying* $\mathbb{E}\{\epsilon_i^2 \mid (\Gamma_i, \mathbf{X}_i)\} < C$. *Then,*

$$
\sup_{\mathbf{\Delta}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\epsilon_i^2(\mathbf{X}_i^T\mathbf{\Delta})^2}{\|\mathbf{\Delta}\|_1^2/s_0 + \|\mathbf{\Delta}\|_2^2} \le c\left\{1 + \sqrt{\frac{t}{N\gamma_N}} + \frac{t\log(N)}{N\gamma_N}\right\}
$$

*with probability at least* $1 - 3\exp(-t) - t^{-1}$ *and some constant* $c > 0$, *where* $s_0$ *is defined as in part (b).*

Proof of Lemma E.6. (a) By Theorem 15 of Rudelson and Zhou (2012), for any $s = o(M/\log(d))$, we have

$$
\sup_{\mathbf{\Delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} \frac{\sum_{i\in\mathcal{I}_j}(\mathbf{X}_i^T\mathbf{\Delta})^2}{M\mathbb{E}(\mathbf{X}^T\mathbf{\Delta})^2} = O_p(1),
$$

where $\sup_{\mathbf{\Delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} \mathbb{E}(\mathbf{X}^T\mathbf{\Delta})^2 = O(1)$ under Assumption 2. Therefore,

$$
\sup_{\mathbf{\Delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} M^{-1}\sum_{i\in\mathcal{I}_j}(\mathbf{X}_i^T\mathbf{\Delta})^2 = O_p(1).
$$

Together with Lemma E.5,

$$
\sup_{\mathbf{\Delta}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{M^{-1}\sum_{i\in\mathcal{I}_j}(\mathbf{X}_i^T\mathbf{\Delta})^2}{\|\mathbf{\Delta}\|_1^2/s + \|\mathbf{\Delta}\|_2^2} = O_p(1).
$$

(b) Fix $\delta \in (0,1)$ and let $\mathbf{\Omega} = M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\mathbf{X}_i\mathbf{X}_i^T - \mathbb{E}(\Gamma\mathbf{X}\mathbf{X}^T)$. For any $s \ge 1$, by Lemma E.4, (E.1) holds with some $k \asymp s$. Since $\mathbf{X}$ is sub-Gaussian, under Assumption 3, we have

$$
\sup_{\mathbf{\Delta}\in\Theta_k} \text{Var}\left\{(\Gamma\mathbf{X}^T\mathbf{\Delta})^2\right\} \le \sup_{\mathbf{\Delta}\in\Theta_k} \mathbb{E}\left\{\Gamma(\mathbf{X}^T\mathbf{\Delta})^4\right\} = \sup_{\mathbf{\Delta}\in\Theta_k} \mathbb{E}\left\{\gamma_N(\mathbf{X})(\mathbf{X}^T\mathbf{\Delta})^4\right\} = O(\gamma_N).
$$

By Theorem 4.3 of Kuchibhotla and Chakrabortty (2022),

$$
\mathbb{P}_{\mathcal{D}'_N}\left(\sup_{\mathbf{\Delta}\in\Theta_k}\left|\mathbf{\Delta}^T\mathbf{\Omega}\mathbf{\Delta}\right| > c\left[\sqrt{\frac{\gamma_N\{t+k\log(d)\}}{M}} + \frac{\log(N)\{t+k\log(d)\}}{M}\right]\right) \ge 3\exp(-t),
$$

with some constant $c > 0$. Together with (E.1), Lemma E.5 and note that $k \asymp s$ and $M \asymp N$, we have for any $s \ge 1$,

$$
\sup_{\mathbf{\Delta}\in\mathbb{R}^d} \frac{|\mathbf{\Delta}^T\mathbf{\Omega}\mathbf{\Delta}|}{\|\mathbf{\Delta}\|_1^2/s + \|\mathbf{\Delta}\|_2^2} \le c\left[\sqrt{\frac{\gamma_N\{t+s\log(d)\}}{N}} + \frac{\log(N)\{t+s\log(d)\}}{N}\right] \tag{E.5}
$$

with probability at least $1 - 3\exp(-t)$ and some constant $c > 0$. Note that

$$
\gamma_N^{-1}\mathbb{E}\{\Gamma(\mathbf{X}^T\mathbf{\Delta})^2\} \le \gamma_N^{-1}\|\gamma_N(\mathbf{X})\|_{\mathbb{P},q}\|\mathbf{X}^T\mathbf{\Delta}\|_{\mathbb{P},2r}^2 = O(1).
$$

Hence, we conclude that uniformly for all $\mathbf{\Delta} \in \mathbb{R}^d$,

$$\frac{\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^2}{M\gamma_N} \le c\left[1+\sqrt{\frac{t+s\log(d)}{N\gamma_N}}+\frac{\log(N)\{t+s\log(d)\}}{N\gamma_N}\right]\left(\frac{\|\mathbf{\Delta}\|_1^2}{s}+\|\mathbf{\Delta}\|_2^2\right)$$

with probability at least $1-3\exp(-t)$ and some constant $c>0$. Choose $s=s_0 := \lceil N\gamma_N/\{\log(N)\log(d)\}\rceil$, then it follows that (E.4) holds with probability at least $1-3\exp(-t)$ and some constant $c>0$.

(c) Fix $\delta\in(0,1)$ and let $\mathbf{\Omega} = M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i f(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^T - \mathbb{E}\{\Gamma f(\mathbf{X})\mathbf{X}\mathbf{X}^T\}$. For any $s\ge 1$, by Lemma E.4, (E.1) holds with some $k\asymp s$. Since $\mathbf{X}$ is sub-Gaussian and the function $f(\cdot)$ is bounded, under Assumption 3, we also have

$$\begin{aligned}\sup_{\mathbf{\Delta}\in\Theta_k}\mathrm{Var}\left[\{\Gamma f(\mathbf{X})\mathbf{X}^T\mathbf{\Delta})^2\}\right] &\le \sup_{\mathbf{\Delta}\in\Theta_k}\mathbb{E}\left\{\Gamma f(\mathbf{X})(\mathbf{X}^T\mathbf{\Delta})^4\right\}\\ &= \sup_{\mathbf{\Delta}\in\Theta_k}\mathbb{E}\left\{\gamma_N(\mathbf{X})f^2(\mathbf{X})(\mathbf{X}^T\mathbf{\Delta})^4\right\} = O(\gamma_N).\end{aligned}$$

By Theorem 4.3 of Kuchibhotla and Chakrabortty (2022),

$$\mathbb{P}_{\mathcal{D}'_N}\left(\sup_{\mathbf{\Delta}\in\Theta_k}\left|\mathbf{\Delta}^T\mathbf{\Omega}\mathbf{\Delta}\right| > c\left[\sqrt{\frac{\gamma_N\{t+k\log(d)\}}{M}}+\frac{\log(N)\{t+k\log(d)\}}{M}\right]\right) \ge 3\exp(-t),$$

with some constant $c>0$. Together with (E.1), Lemma E.5 and note that $k\asymp s$ and $M\asymp N$, we have for any $s\ge 1$,

$$\sup_{\mathbf{\Delta}\in\mathbb{R}^d}\frac{\left|\mathbf{\Delta}^T\mathbf{\Omega}\mathbf{\Delta}\right|}{\|\mathbf{\Delta}\|_1^2/s+\|\mathbf{\Delta}\|_2^2} = O_p\left(\sqrt{\frac{\gamma_N\{t+s\log(d)\}}{N}}+\frac{\log(N)\{t+s\log(d)\}}{N}\right).$$

For any $\mathbf{\Delta}_1,\mathbf{\Delta}_2\in\mathbb{R}^d\setminus\{\mathbf{0}\}$, note that

$$\begin{aligned}\left\|\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right\|_1 &\le \frac{\|\mathbf{\Delta}_1\|_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\|\mathbf{\Delta}_2\|_1}{\|\mathbf{\Delta}_2\|_2},\\ \left\|\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}-\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right\|_1 &\le \frac{\|\mathbf{\Delta}_1\|_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\|\mathbf{\Delta}_2\|_1}{\|\mathbf{\Delta}_2\|_2},\\ \left\|\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right\|_2 &\le \frac{\|\mathbf{\Delta}_1\|_2}{\|\mathbf{\Delta}_1\|_2}+\frac{\|\mathbf{\Delta}_2\|_2}{\|\mathbf{\Delta}_2\|_2}=2,\\ \left\|\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}-\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right\|_2 &\le \frac{\|\mathbf{\Delta}_1\|_2}{\|\mathbf{\Delta}_1\|_2}+\frac{\|\mathbf{\Delta}_2\|_2}{\|\mathbf{\Delta}_2\|_2}=2.\end{aligned}$$

Hence, uniformly for all $\mathbf{\Delta}_1,\mathbf{\Delta}_2\in\mathbb{R}^d$,

$$\begin{aligned}&\gamma_N^{-1}|\mathbf{\Delta}_1^T\mathbf{\Omega}\mathbf{\Delta}_2| = \gamma_N^{-1}\|\mathbf{\Delta}_1\|_2\|\mathbf{\Delta}_2\|_2\left|\frac{\mathbf{\Delta}_1^T}{\|\mathbf{\Delta}_1\|_2}\mathbf{\Omega}\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right|\\ &\le (2\gamma_N)^{-1}\|\mathbf{\Delta}_1\|_2\|\mathbf{\Delta}_2\|_2\left|\left(\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right)^T\mathbf{\Omega}\left(\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}+\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right)\right|\\ &\quad+(2\gamma_N)^{-1}\|\mathbf{\Delta}_1\|_2\|\mathbf{\Delta}_2\|_2\left|\left(\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}-\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right)^T\mathbf{\Omega}\left(\frac{\mathbf{\Delta}_1}{\|\mathbf{\Delta}_1\|_2}-\frac{\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|_2}\right)\right|\\ &= O_p\left(\left(\sqrt{\frac{s\log(d)}{N\gamma_N}}+\frac{s\log(d)\log(N)}{N\gamma_N}\right)\|\mathbf{\Delta}_1\|_2\|\mathbf{\Delta}_2\|_2\left(\frac{\|\mathbf{\Delta}_1\|_1^2}{s\|\mathbf{\Delta}_1\|_2^2}+\frac{\|\mathbf{\Delta}_2\|_1^2}{s\|\mathbf{\Delta}_2\|_2^2}+1\right)\right).\end{aligned}$$

(d) For any $r_0 \geq 2$, we have $\sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \|\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\|_{\psi_{2/r_0}} \leq \sigma^{r_0}$ using part (d) of Lemma E.2. By the Hölder inequality and part (c) of Lemma E.2,

$$\begin{aligned}\sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \mathbb{E}\{\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\} &= \sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \mathbb{E}\{\gamma_N(\mathbf{X})|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\} \\ &\leq \|\gamma_N(\mathbf{X})\|_{\mathbb{P},q} \sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \|\mathbf{X}^T\mathbf{\Delta}\|^{r_0}_{\mathbb{P},r_0 r} = O(\gamma_N),\end{aligned} \tag{E.6}$$

with $r > 0$ satisfying $1/r + 1/q = 1$. Let $W = W(\mathbf{\Delta}) = \Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0} - \mathbb{E}\{\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\}$. By part (d) of Lemma E.2 and note that $\gamma_N \leq 1$, we have $\sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \|W\|_{\psi_{2/r_0}} = O(1)$. Additionally, we have

$$\sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \mathbb{E}(W^2) \leq \sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \mathbb{E}\left\{\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{2r_0}\right\} \leq \|\gamma(\mathbf{X})\|_{\mathbb{P},q} \sup_{\mathbf{\Delta}\in\mathbb{S}^{d-1}} \|\mathbf{X}^T\mathbf{\Delta}\|^{2r_0}_{\mathbb{P},2r_0 r} = O(\gamma_N).$$

By Theorem 3.2 and Proposition A.3 of Kuchibhotla and Chakrabortty (2022), for any $\mathbf{\Delta} \in \mathbb{S}^{d-1}$ and $u > 0$,

$$\begin{aligned}\mathbb{P}_{\mathcal{D}'_N}\Bigg(&\left|(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\mathbf{\Delta}|^{r_0} - \gamma_N^{-1}\mathbb{E}\{\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\}\right| \\ &> c\left[\sqrt{\frac{u}{M\gamma_N}} + \frac{\{\log(M)\}^{r_0/2}u^{r_0/2}}{M\gamma_N}\right]\Bigg) \leq 2\exp(-u),\end{aligned} \tag{E.7}$$

with some constant $c > 0$ independent of $\mathbf{\Delta}$. For any $s \geq 1$, repeating Step 2 of proof of Lemma 17 in Bradic, Wager and Zhu (2019) (with $\|\cdot\|_4$ replaced by $\|\cdot\|_{r_0}$), we obtain

$$\sup_{\mathbf{\Delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} \left\{\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^{r_0}\right\}^{1/r_0} \leq (1-\delta)^{-1}\max_{\mathbf{\Delta}\in\mathcal{T}}\left\{\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^{r_0}\right\}^{1/r_0}, \tag{E.8}$$

with some constant $\delta \in (0,1)$ and a set $\mathcal{T} \subset \mathbb{S}^{d-1}$ satisfying

$$|\mathcal{T}| \leq (c\delta^{-1}d)^{cs},$$

where $c > 0$ is a constant. By the union bound and (E.7), choosing $u = cs\log(c\delta^{-1}d) + t$ with $t > 0$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}'_N}\Bigg(&\max_{\mathbf{\Delta}\in\mathcal{T}}\left|(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\mathbf{\Delta}|^{r_0} - \gamma_N^{-1}\mathbb{E}\{\Gamma|\mathbf{X}^T\mathbf{\Delta}|^{r_0}\}\right| \\ &> c'\left[\sqrt{\frac{s\log(d)+t}{M\gamma_N}} + \frac{\{\log(M)\}^{r_0/2}\{s\log(d)+t\}^{r_0/2}}{M\gamma_N}\right]\Bigg) \\ &\leq 2(c\delta^{-1}d)^{cs}\exp(-u) = 2\exp(-t),\end{aligned} \tag{E.9}$$

with some constant $c' > 0$. Together with (E.6) and (E.8), we have

$$\sup_{\mathbf{\Delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} (M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\mathbf{\Delta}|^{r_0} \leq c\left(1 + \sqrt{\frac{s\log(d)}{N\gamma_N}} + \frac{\{s\log(N)\log(d)\}^{r_0/2}}{N\gamma_N}\right), \tag{E.10}$$

with probability at least $1 - 2\exp(-t)$ and some constant $c > 0$. Now, for any $\mathbf{\Delta} \in \mathbb{R}^d$, define $\widetilde{\boldsymbol{\delta}}$ as in (E.3) with $k_0 = 3$. Then, we also have $\|\widetilde{\boldsymbol{\delta}}_S\|_1 = \|\mathbf{\Delta}\|_1/3$, $\|\widetilde{\boldsymbol{\delta}}_S\|_2 = \|\mathbf{\Delta}\|_1/(3\sqrt{s})$, and

$\widetilde{\boldsymbol{\delta}}, \boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}} \in \mathcal{C}(s,3)$. Therefore,

$$\sum_{i\in\mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T \boldsymbol{\Delta}|^{r_0} \le 2^{r_0-1} \left( \sum_{i\in\mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T \widetilde{\boldsymbol{\delta}}|^{r_0} + \sum_{i\in\mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T (\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}})|^{r_0} \right)$$

$$\le 2^{r_0-1} \left( \|\widetilde{\boldsymbol{\delta}}\|_2^{r_0} + \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2^{r_0} \right) \sup_{\boldsymbol{\delta}\in\mathcal{C}(s,3)\cap\mathbb{S}^{d-1}} \sum_{i\in\mathcal{I}_j} \Gamma_i |\mathbf{X}_i^T \boldsymbol{\delta}|^{r_0}. \tag{E.11}$$

Note that

$$\begin{aligned}
\|\widetilde{\boldsymbol{\delta}}\|_2^{r_0} + \|\boldsymbol{\Delta} - \widetilde{\boldsymbol{\delta}}\|_2^{r_0} &= \|\widetilde{\boldsymbol{\delta}}\|_2^{r_0} + (\|\boldsymbol{\Delta}_S - \widetilde{\boldsymbol{\delta}}_S\|_2)^{r_0} \\
&\le (\|\widetilde{\boldsymbol{\delta}}_S\|_2 + \|\widetilde{\boldsymbol{\delta}}_{S^c}\|_2)^{r_0} + (\|\boldsymbol{\Delta}_S\|_2 + \|\widetilde{\boldsymbol{\delta}}_S\|_2)^{r_0} \\
&= \left\{\|\boldsymbol{\Delta}\|_1/(3\sqrt{s}) + \|\boldsymbol{\Delta}_{S^c}\|_2\right\}^{r_0} + \left\{\|\boldsymbol{\Delta}_S\|_2 + \|\boldsymbol{\Delta}\|_1/(3\sqrt{s})\right\}^{r_0} \\
&\le 2^{r_0-1}\left\{\|\boldsymbol{\Delta}\|_1^{r_0}/(3\sqrt{s})^{r_0} + \|\boldsymbol{\Delta}_{S^c}\|_2^{r_0} + \|\boldsymbol{\Delta}_S\|_2^{r_0} + \|\boldsymbol{\Delta}\|_1^{r_0}/(3\sqrt{s})^{r_0}\right\} \\
&\le 2^{r_0}\left\{\|\boldsymbol{\Delta}\|_1^{r_0}/(3\sqrt{s})^{r_0} + \|\boldsymbol{\Delta}\|_2^{r_0}\right\}.
\end{aligned}$$

Together with (E.10) and (E.11), we have

$$\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^{r_0}}{\|\boldsymbol{\Delta}\|_1^{r_0}/s^{\frac{r_0}{2}} + \|\boldsymbol{\Delta}\|_2^{r_0}} \le c\left(1 + \sqrt{\frac{s\log(d)}{M\gamma_N}} + \frac{\{s\log(M)\log(d)\}^{r_0/2}}{M\gamma_N}\right),$$

with probability at least $1-2\exp(-t)$ and some constant $c>0$.

(e) By the tower rule, for any $\boldsymbol{\Delta}\in\mathbb{R}^d$,

$$\mathbb{E}\{\Gamma|\epsilon|(\mathbf{X}^T\boldsymbol{\Delta})^2\} = \mathbb{E}[\mathbb{E}\{\Gamma|\epsilon|(\mathbf{X}^T\boldsymbol{\Delta})^2 \mid \mathbf{X}\}] = \mathbb{E}\{\gamma_N(\mathbf{X})|\epsilon|(\mathbf{X}^T\boldsymbol{\Delta})^2\} \le c\gamma_N\|\boldsymbol{\Delta}\|_2^2, \tag{E.12}$$

with some constant $c>0$ under Assumption 3, as $\mathbf{X}$ is sub-Gaussian and $\epsilon$ has a bounded $\psi_\alpha$-norm. Let $\boldsymbol{\Omega} = M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\epsilon_i|\mathbf{X}_i\mathbf{X}_i^T - \mathbb{E}(\Gamma|\epsilon|\mathbf{X}\mathbf{X}^T)$. For any $\delta\in(0,1)$ and $s\ge1$, by Lemma E.4, we have (E.1) holds with $k_0=3$ and some $k\asymp s$. Similar to (E.12), we also have

$$\sup_{\boldsymbol{\Delta}\in\Theta_k}\mathrm{Var}\left\{\Gamma|\epsilon|(\mathbf{X}^T\boldsymbol{\Delta})^2\right\} \le \sup_{\boldsymbol{\Delta}\in\Theta_k}\mathbb{E}\left\{\Gamma\epsilon^2(\mathbf{X}^T\boldsymbol{\Delta})^4\right\} = \sup_{\boldsymbol{\Delta}\in\Theta_k}\mathbb{E}\left\{\gamma_N(\mathbf{X})\epsilon^2(\mathbf{X}^T\boldsymbol{\Delta})^4\right\} \le c\gamma_N,$$

with some constant $c>0$. Note that $|\epsilon| = \max(\epsilon,-\epsilon)$. By part (e) of Lemma E.1, $\||\epsilon|\|_{\psi_\alpha} \le \sigma_\epsilon\{\log(2)+2\}^{1/\alpha}$. Together with the definition of $\psi_\alpha$-norms, we further have $\|\sqrt{|\epsilon|}\|_{\psi_{2\alpha}} \le \sqrt{\sigma_\epsilon}\{\log(2)+2\}^{1/(2\alpha)}$. By part (a) of Lemma E.2, $\|\sqrt{|\epsilon|}\mathbf{X}^T\mathbf{e}_j\|_{\psi_{2\alpha}} \le C_{\mathbf{X}}\sqrt{\sigma_\epsilon}\{\log(2)+2\}^{1/(2\alpha)}$, where $\mathbf{e}_j\in\mathbb{R}^d$ denotes the $j$-th column of an identity matrix. By Theorem 4.3 of Kuchibhotla and Chakrabortty (2022) and note that $k\asymp s$ and $M\asymp N$, we have

$$\sup_{\boldsymbol{\Delta}\in\Theta_k}\left|\boldsymbol{\Delta}^T\boldsymbol{\Omega}\boldsymbol{\Delta}\right| \le c\left[\sqrt{\frac{\gamma_N\{t+s\log(d)\}}{N}} + \frac{s\log^{2/(2\alpha)}(M)\{t+s\log(d)\}^{2/(2\alpha)}}{N}\right] \tag{E.13}$$

with probability at least $1-3\exp(-t)$ and some constant $c>0$. Together with (E.1), (E.12), and Lemma E.5,

$$\begin{aligned}
&(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\epsilon_i|(\mathbf{X}_i^T\boldsymbol{\Delta})^2 \\
&\quad\le c\|\boldsymbol{\Delta}\|_2^2 + c\left[\sqrt{\frac{t+s\log(d)}{N\gamma_N}} + \frac{s\log^{1/\alpha}(M)\{t+s\log(d)\}^{1/\alpha}}{N\gamma_N}\right]\left(\frac{\|\boldsymbol{\Delta}\|_1^2}{s} + \|\boldsymbol{\Delta}\|_2^2\right),
\end{aligned}$$

uniformly for all $\mathbf{\Delta} \in \mathbb{R}^d$ with probability at least $1 - 3\exp(-t)$ and some constant $c > 0$.

(f) For any $\mathbf{\Delta} \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}'_N}\left\{M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\epsilon_i^2(\mathbf{X}_i^T\mathbf{\Delta})^2 \mid (\Gamma_i, \mathbf{X}_i)_{i\in\mathcal{I}_j}\right\} \\
&\qquad = M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^2\mathbb{E}_{\mathcal{D}'_N}\left\{\epsilon_i^2 \mid (\Gamma_i, \mathbf{X}_i)_{i\in\mathcal{I}_j}\right\} \le CM^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^2.
\end{aligned}$$

By Markov's inequality,

$$\text{(E.14)}\qquad \mathbb{P}_{\mathcal{D}'_N}\left(M^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\epsilon_i^2(\mathbf{X}_i^T\mathbf{\Delta})^2 \ge tCM^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\mathbf{\Delta})^2 \mid (\Gamma_i, \mathbf{X}_i)_{i\in\mathcal{I}_j}\right) \le t^{-1}.$$

Together with (E.4), we conclude that

$$\sup_{\mathbf{\Delta}\in\mathbb{R}^d\setminus\{\mathbf{0}\}}\frac{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\epsilon_i^2(\mathbf{X}_i^T\mathbf{\Delta})^2}{\|\mathbf{\Delta}\|_1^2/s_0 + \|\mathbf{\Delta}\|_2^2} \le c\left\{1 + \sqrt{\frac{t}{N\gamma_N}} + \frac{t\log(N)}{N\gamma_N}\right\},$$

with probability at least $1 - 3\exp(-t) - t^{-1}$, $s_0 = \lceil N\gamma_N/\{\log(d)\log(N)\}\rceil$, and some constant $c > 0$. ∎

## APPENDIX F: PROOF OF THE PROPERTIES OF THE NUISANCE ESTIMATOR FOR THE SP-BRSS ESTIMATOR

In this section, we analyze the properties of $\widehat{p}_{N,1} = \widehat{p}_{N,1}^{(1)}$, $\widehat{\boldsymbol{\beta}}_{p,1} = \widehat{\boldsymbol{\beta}}_{p,1}^{(1)}$, $\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}}^{(1)}$, with analogous results applicable to $\widehat{p}_{N,1}^{(2)}$, $\widehat{\boldsymbol{\beta}}_{p,1}^{(2)}$, and $\widetilde{\boldsymbol{\alpha}}^{(2)}$. We also denote $\widehat{\pi}(\cdot) = \widehat{\pi}^{(1)}$, $\mathcal{D}'_N := \mathcal{D}_N^{(1)}$, $\mathcal{J} := \mathcal{I}_1$, and $M := |\mathcal{J}| = N/2$ throughout.

**F.1. Preliminary analysis of the semi-parametric PS estimator.** We first study a general RSC property. For any $v \in [0,1]$, $\boldsymbol{\beta}^*, \mathbf{\Delta} \in \mathbb{R}^d$, and $\phi : \mathbb{R} \to (0,\infty)$, we define

$$f(\mathbf{\Delta}, a, v, \boldsymbol{\beta}^*, \phi(\cdot)) := (aM)^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\phi\left(\mathbf{X}_i^T(\boldsymbol{\beta}^* + v\mathbf{\Delta})\right)(\mathbf{X}_i^T\mathbf{\Delta})^2.$$

LEMMA F.1. *Let Assumptions 2 and 3 hold and $\phi(\cdot)$ is a continuous function. Let $\kappa_0 > 0$ be any fixed number, let $\kappa_1, \kappa_2, C_1, C_2, c_1, c_2 > 0$ be some constants depending only on $(\sigma, \kappa_0, \kappa_l, v, \phi(\cdot))$. For any (possibly random) $a \in (0,1]$, when $M\gamma_N > \max\{C_2, C_1\log(M)\log(d)\}$ and $k \le 2$,*

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}'_N}\left(f(\mathbf{\Delta}, a, v, \boldsymbol{\beta}^*, \phi(\cdot)) \ge \frac{\gamma_N}{a}\left\{\kappa_1\|\mathbf{\Delta}\|_2^2 - \kappa_2\frac{\log(d)}{M\gamma_N}\|\mathbf{\Delta}\|_1^2\right\},\ \forall\|\mathbf{\Delta}\|_2 \le \kappa_0\right) \\
&\qquad \ge 1 - c_1\exp(-c_2M\gamma_N).
\end{aligned}$$

The following lemma demonstrates the properties of $\widehat{\gamma}_N$.

LEMMA F.2. *For any $t > 0$, define*

$$\text{(F.1)}\qquad \mathcal{E}_\gamma := \left\{|\widehat{\gamma}_N - \gamma_N| \le 2\sqrt{\frac{t\gamma_N}{M}} + \frac{t}{M}\right\}.$$

*Then,*

$$\mathbb{P}_{\mathcal{D}'_N}(\mathcal{E}_\gamma) \le 1 - 2\exp(-t).$$

*On $\mathcal{E}_\gamma$, when $0 < t < 0.01M\gamma_N$, we have*

$$\left|\frac{\gamma_N}{\widehat{\gamma}_N} - 1\right| \leq 2.66\sqrt{\frac{t}{M\gamma_N}}, \quad 0.79\gamma_N \leq \widehat{\gamma}_N \leq 1.21\gamma_N.$$

The following lemma demonstrates the properties of $\widehat{p}_{N,1}$.

LEMMA F.3. *For any $t > 0$, define*

$$\mathcal{E}_p := \left\{ |\widehat{p}_{N,1} - p_{N,1}| \leq \frac{5.32}{\mathbb{E}(T)}\sqrt{\frac{t\gamma_N}{M}} \right\}. \tag{F.2}$$

*Then,*

$$\mathbb{P}_{\mathcal{D}'_N}(\mathcal{E}_p) \leq 1 - 4\exp(-t).$$

*On $\mathcal{E}_p$, when $0 < t < 0.01M\gamma_N$, we have*

$$\left|\frac{p_{N,1}}{\widehat{p}_{N,1}} - 1\right| \leq 12\sqrt{\frac{t}{M\gamma_N}}, \quad 0.46p_{N,1} \leq \widehat{p}_{N,1} \leq 1.54p_{N,1}.$$

For any $\boldsymbol{\beta} \in \mathbb{R}^d$, $a \in (0,1]$, and $\widetilde{\pi}(\cdot) : \mathcal{X} \mapsto [0,1]$, define the following loss function

$$\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}; a; \widetilde{\pi}) := M^{-1}\sum_{i \in \mathcal{J}}\left[\left\{1 - \frac{\Gamma_i}{\widetilde{\pi}(\mathbf{X}_i)}\right\}\mathbf{X}_i^T\boldsymbol{\beta} + \frac{\Gamma_i\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{a\widetilde{\pi}(\mathbf{X}_i)}\right].$$

In addition, for any $\boldsymbol{\beta}, \boldsymbol{\Delta} \in \mathbb{R}^d$, $a \in (0,1]$, and $\widetilde{\pi}(\cdot) : \mathcal{X} \mapsto [0,1]$, further define

$$\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; a; \boldsymbol{\beta}; \widetilde{\pi}) \;:=\; \widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta} + \boldsymbol{\Delta}; a; \widetilde{\pi}) - \widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}; a; \widetilde{\pi}) - \nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}; a; \widetilde{\pi})^T\boldsymbol{\Delta}.$$

In the following, we show the RSC property required for the labeling PS estimation and control the gradient $\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)\|_\infty$.

LEMMA F.4 (The PS model's RSC property). *Let Assumption 1 hold. Then, $p_{N,1} \asymp \gamma_N$. Further let Assumptions 2 and 3 hold. For any constant $\kappa_0 > 1$ and some $\kappa_1, \kappa_2 > 0$, define*

$$\widetilde{\mathcal{B}}_1 := \left\{\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) \geq \frac{p_{N,1}}{\widehat{p}_{N,1}}\left\{\kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\}, \quad \forall\|\boldsymbol{\Delta}\|_2 \leq \kappa_0\right\}. \tag{F.3}$$

*Then, with some constants $C_1, C_2 > 0$, when $M\gamma_N > \max\{C_2, C_1\log(M)\log(d)\}$,*

$$\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{B}}_1 \mid \mathcal{E}_\pi) \geq 1 - c_1\exp(-c_2M\gamma_N),$$

*with some constants $c_1, c_2 > 0$.*

REMARK 4 (Technical challenges of showing Lemma F.4). The proof of Lemma F.4 is an analog of showing the RSC property of a generalized linear model (GLM) (Negahban et al., 2010; Wainwright, 2019). However, we have additional challenges because of the non-standard loss function and also the decaying PS. By Taylor's theorem, we have

$$\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) = (M\widehat{p}_{N,1})^{-1}\sum_{i \in \mathcal{J}}\frac{\Gamma_i}{\widehat{\pi}(\mathbf{X}_i)}\exp\{-\mathbf{X}_i^T(\boldsymbol{\beta}^*_{p,1} + v\boldsymbol{\Delta})\}(\mathbf{X}_i^T\boldsymbol{\Delta})^2,$$

with some $v \in (0,1)$. Unlike a generalized linear model (GLM), $\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi})$ is also a function of $\Gamma_i$. Without the presence of $\Gamma_i$ (or when $\Gamma_i \equiv 1$), $\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi})$ can be directly lower bounded by, e.g., Proposition 2 of Negahban et al. (2010) or Theorem 9.36 of Wainwright (2019), as $\widehat{\pi}(\mathbf{X}_i)$ is bounded below under Assumption 4. However, with the presence of $\Gamma_i$ and the decaying MAR setting mechanism, we need to carefully track on the impact of $\Gamma_i$ since only a few of them are non-zero that $\mathbb{E}(\Gamma) = \gamma_N \to 0$ as $N, d \to \infty$. The usual Bernstein inequality with the usage of $\psi_\alpha$-norms is suboptimal here. To construct a tight lower bound for $\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi})$, we utilize concentration inequalities that involve random variables' $\psi_2$-norm and also second moment (Kuchibhotla and Chakrabortty, 2022). Here, the second moment helps us to capture the decaying value $\gamma_N$; see more details in Section F. In addition, the same challenges also arise in the proofs of Lemmas F.5 – F.10 below.

LEMMA F.5 (Upper bound for $\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)\|_\infty$). *Let Assumptions 1, 4, and 5 hold. For any $0 < t < M\gamma_N / \{100 + \log(M)\log(d)\}$, define*

$$\widetilde{\mathcal{B}}_2 := \left\{ \|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)\|_\infty \le \kappa_3 \sqrt{\frac{t + \log(d)}{M\gamma_N}} \right\}, \tag{F.4}$$

*with some constant $\kappa_3 > 0$ and $\widehat{\gamma}_N$ is defined as (4.3). Then, when $M\gamma_N > C_1 \log(M)\log(d)$,*

$$\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{B}}_2) \ge 1 - 10\exp(-t).$$

Define

$$\begin{aligned} \mathbf{r}_\pi :=& \nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \widehat{\pi}) - \nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*) \\ =& -M^{-1} \sum_{i \in \mathcal{J}} \left\{ \frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)} \right\} \Gamma_i \left\{ 1 + \widehat{p}_{N,1}^{-1} \exp(-\mathbf{X}_i^T \boldsymbol{\beta}^*_{p,1}) \right\} \mathbf{X}_i. \end{aligned} \tag{F.5}$$

The following lemma controls the treatment PS's estimation error's effect on the labeling PS's estimation.

LEMMA F.6. *Let Assumptions 2, 3, and 4 hold. Then, on the event $\mathcal{E}_\zeta$,*

$$\widetilde{\mathcal{B}}_3 := \left\{ |\mathbf{r}_\pi^T \boldsymbol{\Delta}| \le c\zeta_N \left( \|\boldsymbol{\Delta}\|_1 \sqrt{\frac{\log(d)\log(N)}{M\gamma_N}} + \|\boldsymbol{\Delta}\|_2 \right), \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d \right\} \tag{F.6}$$

*occurs with some constant $c > 0$ with probability at least $1 - 3\exp\{N\gamma_N / \log(N)\}$.*

In order to utilize the RSC property of Lemma F.4, we need to first show that the error $\|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_2$ is bounded. For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, define

$$\begin{aligned} \mathcal{F}(\boldsymbol{\Delta}) :=& \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1} + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \widehat{\pi})^T \boldsymbol{\Delta} - \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1 \\ =& \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1} + \boldsymbol{\Delta}\|_1 + \nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)^T \boldsymbol{\Delta} \\ &+ \mathbf{r}_\pi^T \boldsymbol{\Delta} - \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1. \end{aligned} \tag{F.7}$$

By construction, we have $\mathcal{F}(\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}) \le 0$. For any $r_N > 0$, define $\mathcal{K}(r_N, 1) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_1 \le r_N\|\boldsymbol{\Delta}\|_2, \|\boldsymbol{\Delta}\|_2 = 1\}$. The following Lemma shows that the function $\mathcal{F}(\boldsymbol{\Delta})$ is strictly positive with high probability for any $\boldsymbol{\Delta} \in \mathcal{K}(r_N, 1)$. As we will show that $\|\boldsymbol{\Delta}\|_1 \le r_N\|\boldsymbol{\Delta}\|_2$ with some $r_N > 0$ (see (F.18)), the following Lemma indicates that $\|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_2 \ne 1$. In fact, we can further show that $\|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_2 \le 1$ using the convexity of the function $\mathcal{F}(\cdot)$; see details in the proof of Theorem 4.1.

LEMMA F.7. *Let* $s_{p,1} = o(M\gamma_N/\log(d))$. *For* $0 < t < M\gamma_N/\{100 + \log(M)\log(d)\}$, *choose any* $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(d)/(M\gamma_N)}$ *satisfying* $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\{t+\log(d)\}/(M\gamma_N)}$. *Then, on the event* $\widetilde{\mathcal{B}}_2 \cap \widetilde{\mathcal{B}}_3$, *when* $N$ *is large enough,*

$$\mathcal{F}(\boldsymbol{\Delta}) \geq \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 - (2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2, \ \forall \boldsymbol{\Delta} \in \mathbb{R}^d.$$

*Further condition on the event* $\widetilde{\mathcal{B}}_1 \cap \mathcal{E}_p$. *Let* $r_N = o(\sqrt{M\gamma_N/\log(d)})$, *when* $N$ *is large enough, we have* $\mathcal{F}(\boldsymbol{\Delta}) > 0$, $\forall \boldsymbol{\Delta} \in \mathcal{K}(r_N, 1)$.

**F.2. Preliminary analysis of the OR estimator.** For any $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta} \in \mathbb{R}^d$, $a \in (0,1]$, and function $\widetilde{\pi}(\cdot) : \mathcal{X} \mapsto [0,1]$, define the loss function for the corresponding OR model as

$$\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; a, \boldsymbol{\beta}; \widetilde{\pi}) := M^{-1}\sum_{i\in\mathcal{J}} \frac{\Gamma_i}{a\widetilde{\pi}(\mathbf{X}_i)} \exp(-\mathbf{X}_i^T\boldsymbol{\beta})(Y_i - \mathbf{X}_i^T\boldsymbol{\alpha})^2.$$

In addition, let

$$\begin{aligned}\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta}; a, \boldsymbol{\beta}; \boldsymbol{\alpha}; \widetilde{\pi}) &:= \widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} + \boldsymbol{\Delta}; a, \boldsymbol{\beta}; \widetilde{\pi}) - \widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; a, \boldsymbol{\beta}; \widetilde{\pi}) - \nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; a, \boldsymbol{\beta}; \widetilde{\pi})^T\boldsymbol{\Delta} \\ &= M^{-1}\sum_{i\in\mathcal{J}} \frac{\Gamma_i}{a\widetilde{\pi}(\mathbf{X}_i)} \exp(-\mathbf{X}_i^T\boldsymbol{\beta})(\mathbf{X}_i^T\boldsymbol{\Delta})^2.\end{aligned}$$

We first characterize the RSC property for the OR model in the following Lemma.

LEMMA F.8 (The OR model's RSC property). *Let Assumptions 1, 2, 3, and 5 hold. For some constants* $\kappa_1, \kappa_2 > 0$, *define*

$$\widetilde{\mathcal{A}}_1 := \left\{\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widetilde{\boldsymbol{\alpha}}^*; \widehat{\pi}) \geq \frac{p_{N,1}}{\widehat{p}_{N,1}}\left\{\kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\}, \quad \forall \boldsymbol{\Delta} \in \mathbb{R}^d\right\}.$$

*Then, with some constants* $C_1, C_2 > 0$, *when* $M\gamma_N > \max\{C_2, C_1\log(M)\log(d)\}$,

$$\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{A}}_1 \mid \mathcal{E}_\pi \cap \widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}) \geq 1 - c_1\exp(-c_2 M\gamma_N),$$

*with some constants* $c_1, c_2 > 0$.

If the OR model is correctly specified, we control the gradient $\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi})\|_\infty$ in the following lemma.

LEMMA F.9 (Upper bound for $\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi})\|_\infty$). *Let* $m(\mathbf{x}) = m^*(\mathbf{x}) = \mathbf{x}^T\widetilde{\boldsymbol{\alpha}}^*$ *for all* $\mathbf{x} \in \mathcal{X}$. *Let Assumptions 1, 6, and 5 hold. For any* $0 < t < M\gamma_N/\{100 + \log(M)\log(d)\}$, *define*

$$\widetilde{\mathcal{A}}_2 := \left\{\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi})\|_\infty \leq \kappa_4\sqrt{\frac{t + \log(d)}{M\gamma_N}}\right\},$$

*with some constant* $\kappa_4 > 0$. *Then, when* $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\{t + \log(d)\}(M\gamma_N)}$,

$$\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{A}}_2 \mid \mathcal{E}_\pi \cap \widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}) \geq 1 - 2\exp\{-t\log(d)\} - 4\exp(-t).$$

Now, we further consider the case that the OR model is misspecified. We control the gradient $\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \boldsymbol{\beta}^*_{p,1}; \pi^*)\|_\infty$ in the following lemma.

LEMMA F.10 (Upper bound for $\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*;\widehat{p}_{N,1},\boldsymbol{\beta}^*_{p,1};\pi^*)\|_\infty$). *Consider the general case that $m(\mathbf{x}) \neq m^*(\mathbf{x}) = \mathbf{x}^T\widetilde{\boldsymbol{\alpha}}^*$ is allowed. Let Assumptions 3, 6, 4, and 5 hold. For any $0 < t < M\gamma_N/\{100+\log(M)\log(d)\}$, define*

$$\widetilde{\mathcal{A}}_3 := \left\{\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*;\widehat{p}_{N,1},\boldsymbol{\beta}^*_{p,1};\pi^*)\|_\infty \leq \kappa_4\sqrt{\frac{t+\log(d)}{M\gamma_N}}\right\},$$

*with some constant $\kappa_4 > 0$. Let $M\gamma_N > C_1\log(M)\log(d)$ and $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\{t+\log(d)\}(M\gamma_N)}$, then*

$$\mathbb{P}_{\mathcal{D}'_N}(\mathcal{A}_3) \geq 1 - 7\exp(-t).$$

### F.3. Proof of the results in Section F.1.

PROOF OF LEMMA F.1. For any $\boldsymbol{\Delta} \in \mathbb{R}^d$, define

$$A_i = \Gamma_i\phi\left(\mathbf{X}_i^T(\boldsymbol{\beta}^* + v\boldsymbol{\Delta})\right)(\mathbf{X}_i^T\boldsymbol{\Delta})^2.$$

Then,

$$f(\boldsymbol{\Delta},\gamma_N,v,\boldsymbol{\beta}^*,\phi(\cdot)) = (aM)^{-1}\sum_{i\in\mathcal{J}}A_i.$$

For truncation levels $T \geq \tau > 0$, we define the following truncation functions

$$\varphi_\tau(u) = u^2\mathbb{1}_{|u|\leq\tau/2} + (\tau-u)^2\mathbb{1}_{\tau/2<|u|\leq\tau}, \qquad \alpha_\tau(u) = u\mathbb{1}_{|u|\leq\tau}.$$

Now, we show that, for each $i \in \mathcal{J}$,

$$A_i \geq \Gamma_i\phi\left(\alpha_T(\mathbf{X}_i^T\boldsymbol{\beta}^*) + v\alpha_\tau(\mathbf{X}_i^T\boldsymbol{\Delta})\right)\varphi_\tau\left(\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right). \tag{F.8}$$

**Case 1:** $|\mathbf{X}_i^T\boldsymbol{\beta}^*| > T$ or $|\mathbf{X}_i^T\boldsymbol{\Delta}| > \tau$. We can see that $\varphi_\tau\left(\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right) = 0$ and hence (F.8) follows.

**Case 2:** $|\mathbf{X}_i^T\boldsymbol{\beta}^*| \leq T$, and $|\mathbf{X}_i^T\boldsymbol{\Delta}| \leq \tau$. Then,

$$\alpha_T(\mathbf{X}_i^T\boldsymbol{\beta}^*) = \mathbf{X}_i^T\boldsymbol{\beta}^*, \quad \alpha_\tau(\mathbf{X}_i^T\boldsymbol{\Delta}) = \mathbf{X}_i^T\boldsymbol{\Delta}, \quad (\mathbf{X}_i^T\boldsymbol{\Delta})^2 \geq \varphi_\tau\left(\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right),$$

and hence (F.8) follows. Define

$$K_3 := 8r\sigma^2\log\left\{2^{4+1/(2r)}cr\sigma^2\kappa_l^{-1}\right\}, \qquad L_\phi := \min_{|u|\leq\sqrt{K_3}(1+\kappa_0)}\phi(u).$$

For some $\delta \in (0,\kappa_0]$, we choose the truncation levels as

$$T^2 = K_3, \qquad \tau^2 = \tau^2(\delta) = K_3\delta^2.$$

Then, for each $i \in \mathcal{J}$,

$$\phi\left(\alpha_T(\mathbf{X}_i^T\boldsymbol{\beta}^*) + v\alpha_\tau(\mathbf{X}_i^T\boldsymbol{\Delta})\right) \geq L_\phi.$$

Note that, for any $\kappa_1,\kappa_2 > 0$ and $\boldsymbol{\Delta} \in \mathbb{R}^d$,

$$\kappa_1\|\boldsymbol{\Delta}\|_2^2 - \kappa_2\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2 \leq 2\kappa_1\|\boldsymbol{\Delta}\|_2^2 - 2\sqrt{\frac{\kappa_2\log(d)}{M\gamma_N}}\|\boldsymbol{\Delta}\|_1^2.$$

Hence, it suffices to show that

$$M^{-1}\sum_{i\in\mathcal{J}}\varphi_{\tau(\delta)}\left(\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right)$$
$$\geq 2L_\phi^{-1}\gamma_N\left\{\kappa_1\|\boldsymbol{\Delta}\|_2^2 - \sqrt{\frac{\kappa_2\log(d)}{M\gamma_N}}\|\boldsymbol{\Delta}\|_1^2\right\}, \quad \forall\delta\in(0,\kappa_0],\ \|\boldsymbol{\Delta}\|_2 = \delta,$$

with high probability. By rescaling the vector that $\widetilde{\mathbf{\Delta}} = \mathbf{\Delta}/\|\mathbf{\Delta}\|_2$ and notice that $\varphi_{\tau(\delta)}(\delta^2 u) = \delta^2\varphi_{\tau(1)}(u)$, it suffices to show

(F.9)

$$M^{-1}\sum_{i\in\mathcal{J}}\varphi_{\tau(1)}\left(\mathbf{X}_i^T\mathbf{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\le T}\right) \ge 2L_\phi^{-1}\gamma_N\left\{\kappa_1 - \sqrt{\frac{\kappa_2\log(d)}{M\gamma_N}}\|\mathbf{\Delta}\|_1\right\},\ \ \forall\|\mathbf{\Delta}\|_2 = 1.$$

Hence, we will restrict $\delta = 1$ and $\tau = \tau(1) = \sqrt{K_3}$. Define $g_{\mathbf{\Delta}}(\mathbf{x}) = \varphi_\tau\left(\mathbf{x}^T\mathbf{\Delta}\mathbb{1}_{|\mathbf{x}^T\boldsymbol{\beta}^*|\le T}\right)$, $\kappa_1 = \kappa_l/(2L_\phi)$, and $\kappa_2 = 2c_1'^2/L_\phi$.

**Step 1.** We first demonstrate that, for any $\|\mathbf{\Delta}\|_2 = 1$,

$$\mathbb{E}\left\{\Gamma g_{\mathbf{\Delta}}(\mathbf{X})\right\} \ge \gamma_N\kappa_l/2. \tag{F.10}$$

Under Assumption 2, we have

$$\mathbb{E}\left\{\Gamma(\mathbf{X}^T\mathbf{\Delta})^2\right\} \ge \gamma_N\kappa_l.$$

Besides, observe that, for $r > 0$ satisfying $1/r + 1/q = 1$,

$$\begin{aligned}
\mathbb{E}&\left[\Gamma\left\{(\mathbf{X}^T\mathbf{\Delta})^2 - g_{\mathbf{\Delta}}(\mathbf{X})\right\}\right] = \mathbb{E}\left[\gamma_N(\mathbf{X})\left\{(\mathbf{X}^T\mathbf{\Delta})^2 - g_{\mathbf{\Delta}}(\mathbf{X})\right\}\right]\\
&\le \|\gamma_N(\cdot)\|_{\mathbb{P},q}\left\|(\mathbf{X}^T\mathbf{\Delta})^2 - g_{\mathbf{\Delta}}(\mathbf{X})\right\|_{\mathbb{P},r}\\
&\le c\gamma_N\left\|(\mathbf{X}^T\mathbf{\Delta})^2\mathbb{1}_{|\mathbf{X}^T\boldsymbol{\beta}^*|\ge T}\right\|_{\mathbb{P},r} + c\gamma_N\left\|(\mathbf{X}^T\mathbf{\Delta})^2\mathbb{1}_{|\mathbf{X}^T\mathbf{\Delta}|\ge\tau/2}\right\|_{\mathbb{P},r}\\
&\le c\gamma_N\|(\mathbf{X}^T\mathbf{\Delta})^2\|_{\mathbb{P},2r}\left[\left\{\mathbb{P}(|\mathbf{X}^T\boldsymbol{\beta}^*|\ge T)\right\}^{1/(2r)} + \left\{\mathbb{P}(|\mathbf{X}^T\mathbf{\Delta}|\ge\tau/2)\right\}^{1/(2r)}\right].
\end{aligned}$$

Under Assumption 2, $\|\mathbf{X}^T\mathbf{\Delta}\|_{\psi_2} \le \sigma$ and $\|\mathbf{X}^T\boldsymbol{\beta}^*\|_{\psi_2} \le \sigma$. By part (b) of Lemma E.2,

$$\|(\mathbf{X}^T\mathbf{\Delta})^2\|_{\mathbb{P},2r} \le 2^{1+1/(2r)}r\sigma^2.$$

By part (b) of Lemma E.2,

$$\begin{aligned}
\mathbb{P}(|\mathbf{X}^T\boldsymbol{\beta}^*|\ge T) &\le 2\exp\left(-\frac{K_3}{\sigma^2}\right) \le 2\exp\left(-\frac{K_3}{4\sigma^2}\right),\\
\mathbb{P}(|\mathbf{X}^T\mathbf{\Delta}|\ge\tau/2) &\le 2\exp\left(-\frac{K_3}{4\sigma^2}\right).
\end{aligned}$$

Hence, by the construction of $K_3$,

$$\mathbb{E}\left[\Gamma\left\{(\mathbf{X}^T\mathbf{\Delta})^2 - g_{\mathbf{\Delta}}(\mathbf{X})\right\}\right] \le 2^{3+1/(2r)}cr\sigma^2\exp\left(-\frac{K_3}{8r\sigma^2}\right)\gamma_N \le \gamma_N\kappa_l/2.$$

Therefore,

$$\mathbb{E}\left\{\Gamma g_{\mathbf{\Delta}}(\mathbf{X})\right\} = \mathbb{E}\left\{\Gamma(\mathbf{X}^T\mathbf{\Delta})^2\right\} - \mathbb{E}\left[\Gamma\left\{(\mathbf{X}^T\mathbf{\Delta})^2 - g_{\mathbf{\Delta}}(\mathbf{X})\right\}\right] \ge \gamma_N\kappa_l/2.$$

**Step 2.** Define

$$f_N(\mathbf{\Delta}) = \left|M^{-1}\sum_{i\in\mathcal{J}}\Gamma_i g_{\mathbf{\Delta}}(\mathbf{X}_i) - \mathbb{E}\{\Gamma g_{\mathbf{\Delta}}(\mathbf{X})\}\right|,\quad Z(t) = \sup_{\|\mathbf{\Delta}\|_2=1,\|\mathbf{\Delta}\|_1\le t} f_N(\mathbf{\Delta}).$$

We prove that, with some constant $C > 0$, when $M\gamma_N$ is large enough,

$$\mathbb{E}_{\mathcal{D}_N'}\{Z(t)\} \le 1016K_3r\sigma t\sqrt{\frac{\gamma_N\log(d)}{M}}. \tag{F.11}$$

Let $(\epsilon_i)_{i\in\mathcal{I}}$ be i.i.d. Rademacher variables. With a slight abuse of notation, we let $\mathbf{Z} = (R, T, Y(0), Y(1), \mathbf{X}, \epsilon)$ and $\mathcal{D}_N' = \{\mathbf{Z}_i\}_{i\in\mathcal{J}}$. Then,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_N'}\{Z(t)\} &\leq 2\mathbb{E}_{\mathcal{D}_N'}\left\{\sup_{\|\boldsymbol{\Delta}\|_2=1,\|\boldsymbol{\Delta}\|_1\leq t}\left|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i g_{\boldsymbol{\Delta}}(\mathbf{X}_i)\right|\right\} \\
&= 2\mathbb{E}_{\mathcal{D}_N'}\left\{\sup_{\|\boldsymbol{\Delta}\|_2=1,\|\boldsymbol{\Delta}\|_1\leq t}\left|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i\varphi_\tau\left(\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right)\right|\right\} \\
&\overset{(i)}{\leq} 8K_3\mathbb{E}_{\mathcal{D}_N'}\left\{\sup_{\|\boldsymbol{\Delta}\|_2=1,\|\boldsymbol{\Delta}\|_1\leq t}\left|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i\mathbf{X}_i^T\boldsymbol{\Delta}\mathbb{1}_{|\mathbf{X}_i^T\boldsymbol{\beta}^*|\leq T}\right|\right\} \\
&\leq 8K_3 t\mathbb{E}_{\mathcal{D}_N'}\left\|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i\mathbf{X}_i\right\|_\infty,
\end{aligned}
$$

where (i) holds by utilizing Ledoux-Talagrand contraction inequality in Ledoux and Talagrand (2013), notice that $\varphi(\cdot)$ is a $(2K_3)$-Lipschitz continuous and $\varphi_\tau(0) = 0$. Here, $\mathbb{E}(\epsilon\Gamma\mathbf{X}) = \mathbf{0}$, $\sup_{1\leq j\leq d}\|\epsilon\Gamma\mathbf{X}(j)\|_{\psi_2} \leq \sigma$, and with $r > 0$ satisfying $1/r + 1/q = 1$,

$$
\begin{aligned}
\sup_{1\leq j\leq d}\mathbb{E}\{\epsilon\Gamma\mathbf{X}(j)\}^2 &= \sup_{1\leq j\leq d}\mathbb{E}\left\{\gamma_N(\mathbf{X})\mathbf{X}^2(j)\right\} \\
&\leq \|\gamma_N(\cdot)\|_{\mathbb{P},q}\|\mathbf{X}(j)\|_{\mathbb{P},2r}^2 \overset{(i)}{\leq} c2^{1/r}r^2\sigma^2\gamma_N \leq 2cr^2\sigma^2\gamma_N,
\end{aligned}
$$

where (i) holds by Assumption 3 and $\|\mathbf{X}(j)\|_{\mathbb{P},2r} \leq 2^{1/(2r)}r\sigma$ using part (c) of Lemma E.2. By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022), for any $u \geq 0$, with probability at least $1 - 3\exp(-u)$,

$$
\begin{aligned}
\left\|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i\mathbf{X}_i\right\|_\infty &\leq 7r\sigma\sqrt{\frac{2c\gamma_N\{u+\log(d)\}}{M}} + \frac{c_1'\sigma\sqrt{\log(M)}\{u+\log(d)\}}{M} \\
&\leq a_1 u + a_2\sqrt{u} + a_3,
\end{aligned}
$$

where $c_1' > 0$ is a constant and

$$
\begin{aligned}
a_1 &= \frac{C\sigma\sqrt{\log(M)}}{M}, \\
a_2 &= 7r\sigma\sqrt{\frac{2c\gamma_N}{M}}, \\
a_3 &= 7r\sigma\sqrt{\frac{2c\gamma_N\log(d)}{M}} + \frac{c_1'\sigma\sqrt{\log(M)}\log(d)}{M}.
\end{aligned}
$$

By Lemma E.3,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_N'}\left\|M^{-1}\sum_{i\in\mathcal{J}}\epsilon_i\Gamma_i\mathbf{X}_i\right\|_\infty \leq 12a_1 + 3\sqrt{\pi}a_2 + a_3 \\
&= 7r\sigma\sqrt{\frac{2c\gamma_N}{M}}\{\sqrt{\log(d)} + 3\sqrt{\pi}\} + \frac{c_1'\sigma\sqrt{\log(M)}\{\log(d)+12\}}{M} \\
&\leq 14r\sigma\sqrt{\frac{2c\gamma_N}{M}}\{\sqrt{\log(d)} + 3\sqrt{\pi}\} \leq 127r\sigma\sqrt{\frac{c\gamma_N\log(d)}{M}},
\end{aligned}
$$

when $M\gamma_N > C_1 \log(M)\log(d)$, where $C_1 = c_1'^2/(4cr^2)$. Hence, (F.11) follows.

**Step 3.** We showcase that, with some constants $c_2', c_1, c_2 > 0$,

$$\mathbb{P}_{\mathcal{D}_N'}\left(Z(t) \geq \frac{\gamma_N \kappa_l}{4} + 2c_2'\|\boldsymbol{\Delta}\|_1\sqrt{\frac{\gamma_N \log(d)}{M}}, \forall t > 0\right) \leq c_1 \exp\left(-c_2 M\gamma_N\right). \tag{F.12}$$

Define $z^*(t) = \gamma_N\{\kappa_l/8 + 2K_3\sigma t\sqrt{\log(d)/(M\gamma_N)}\}$ and $\mathcal{F} = \{\pm f(\cdot) : f(\mathbf{x}) = \Gamma g_{\boldsymbol{\Delta}}(\mathbf{x}) - \mathbb{E}\{\Gamma g_{\boldsymbol{\Delta}}(\mathbf{X})\}, \|\boldsymbol{\Delta}\|_2 = 1, \|\boldsymbol{\Delta}\|_1 = t\}$. Notice that $0 \leq g_{\boldsymbol{\Delta}}(\mathbf{x}) \leq K_3$ for all $\mathbf{x} \in \mathbb{R}^p$, it follows that $|f(\mathbf{x})| \leq K_3$ for all $f \in \mathcal{F}$. Besides, notice that

$$\sup_{f\in\mathcal{F}} \mathbb{E}\{f^2(\mathbf{X})\} \leq \sup_{\|\boldsymbol{\Delta}\|_2=1, \|\boldsymbol{\Delta}\|_1=t} \mathbb{E}\left\{\Gamma g^2_{\boldsymbol{\Delta}}(\mathbf{X})\right\} \leq K_3^2\gamma_N.$$

By Theorem 3.27 of Wainwright (2019) and (F.11), we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_N'}\left(Z(t) \geq \mathbb{E}_{\mathcal{D}_N'}\{Z(t)\} + z^*(t)\right) \\
&\quad\leq \exp\left(-\frac{M\{z^*(t)\}^2}{8e[K_3^2\gamma_N + 2K_3\mathbb{E}\{Z(t)\}] + 4K_3 z^*(t)}\right) \\
&\quad\leq \exp\left\{-\frac{M\gamma_N\kappa_l^2/64 + 4K_3^2\sigma^2t^2\log(d)}{22K_3^2 + 2032K_3^2 rt\sigma\sqrt{\frac{\log(d)}{M\gamma_N}} + K_3\kappa_l/2 + 8K_3^2\sigma t\sqrt{\frac{\log(d)}{M\gamma_N}}}\right\} \\
&\quad= \exp\left\{-\frac{M\gamma_N\kappa_l^2/64 + K_3^2\sigma^2t^2\log(d)}{22K_3^2 + K_3\kappa_l/2 + (8+2032r)K_3^2\sigma t\sqrt{\frac{\log(d)}{M\gamma_N}}}\right\}.
\end{aligned}$$

It follows that

$$\mathbb{P}_{\mathcal{D}_N'}\left(\sup_{\|\boldsymbol{\Delta}\|_2=1, \|\boldsymbol{\Delta}\|_1\leq t} f_N(\boldsymbol{\Delta}) \geq g(t)\right) \leq h(t),$$

where $a = K_3^2\sigma^2$, $b = \kappa_l^2/64$, $c = (8 + 2032r)K_3^2\sigma$, $d = 22K_3^2 + K_3\kappa_l/2$, $c_2' = (2 + 1016r)K_3\sigma$, $c_3' = \kappa_l/8$,

$$g(t) = c_2't\sqrt{\frac{\gamma_N\log(d)}{M}} + c_3'\gamma_N, \;\; h(t) = \exp\left\{-\frac{at^2\log(d) + bM\gamma_N}{ct\sqrt{\frac{\log(d)}{M\gamma_N}} + d}\right\}.$$

Now we apply a peeling argument to extend the radii $\|\boldsymbol{\Delta}\|_1$. For each $m \geq 1$, define

$$\mathcal{A}_m = \left\{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_2 = 1, 2^{m-1}c_3'\gamma_N \leq g(\|\boldsymbol{\Delta}\|_1) \leq 2^m c_3'\gamma_N\right\}.$$

Notice that $g(t)$ is a strictly increasing function, $g^{-1}(t) = (c_2')^{-1}(t - c_3'\gamma_N)\sqrt{M/\{\gamma_N\log(d)\}}$, and $g(t) \geq c_3'\gamma_N$ for all $t > 0$. Let $c_4' = \log_2\{c_2'/(\sigma c_3') + 1\}$ and $t_m = c_3'(2^m - 1)/c_2' \geq c_3'm/c_2'$ for all $m \geq 1$. Then,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_N'}\left(\exists\boldsymbol{\Delta} \in \mathbb{R}^d \text{ s.t. } \|\boldsymbol{\Delta}\|_2 = 1,\ f_N(\boldsymbol{\Delta}) \geq 2g(\|\boldsymbol{\Delta}\|_1)\right) \\
&\quad\overset{(i)}{\leq} \sum_{m=1}^{\infty} \mathbb{P}_{\mathcal{D}_N'}\left(\exists\boldsymbol{\Delta} \in \mathcal{A}_m \text{ s.t. } f_N(\boldsymbol{\Delta}) \geq 2g(\|\boldsymbol{\Delta}\|_1)\right) \\
&\quad\overset{(ii)}{\leq} \sum_{m=1}^{\infty} \mathbb{P}_{\mathcal{D}_N'}\left(\sup_{\|\boldsymbol{\Delta}\|_2=1, \|\boldsymbol{\Delta}\|_1\leq g^{-1}(2^m\gamma_N c_2)} f_N(\boldsymbol{\Delta}) \geq 2g(\|\boldsymbol{\Delta}\|_1)\right)
\end{aligned}$$

$$\le \sum_{m=1}^{\infty} h(g^{-1}(2^m\gamma_N\kappa_l/4)) = \sum_{m=1}^{\infty} h\left(t_m\sqrt{\frac{M}{\gamma_N\log(d)}}\right)$$
$$= \sum_{m=1}^{\infty}\exp\left(-M\gamma_N\frac{at_m^2+b}{ct_m+d}\right)$$
$$\le \sum_{t_m\le\sigma^{-1}}\exp\left(-M\gamma_N\frac{b}{c\sigma^{-1}+d}\right) + \sum_{t_m>\sigma^{-1}}\exp\left\{-M\gamma_N\frac{at_m}{c+d\sigma}\right\}$$
$$= \sum_{m\le c_4'}\exp\left(-M\gamma_N\frac{b}{c\sigma^{-1}+d}\right) + \sum_{m>c_4'}\exp\left\{-M\gamma_N\frac{at_m}{c+d\sigma}\right\}$$
$$\le c_4'\exp\left(-M\gamma_N\frac{b}{c\sigma^{-1}+d}\right) + \sum_{m=2}^{\infty}\exp\left\{-M\gamma_N\frac{ac_3'm}{c_2'(c+d\sigma)}\right\}$$
$$= c_4'\exp\left(-M\gamma_N\frac{b}{c\sigma^{-1}+d}\right) + \frac{\exp\left\{-M\gamma_N\frac{2ac_3'}{c_2'(c+d\sigma)}\right\}}{1-\exp\left\{-M\gamma_N\frac{ac_3'}{c_2'(c+d\sigma)}\right\}}$$
$$\le c_1\exp\left(-c_2M\gamma_N\right),$$

when $M\gamma_N > c_2'(c+d\sigma)\log(2)/(ac_3')$, where $c_1 = c_4'+2$ and $c_2 = \max[b/(c\sigma^{-1}+d), 2ac_3'/\{c_2'(c+d\sigma)\}]$. Here, (i) holds since $\mathbb{P}_{\mathcal{D}_N'}(\cup_{m=1}^{\infty}\mathcal{A}_m)=1$, (ii) holds since $\mathcal{A}_m \subseteq \{\mathbf{\Delta}: \|\mathbf{\Delta}\|_2=1, \|\mathbf{\Delta}\|_1\le g^{-1}(2^m\gamma_N c_3')\}$. Hence, (F.12) holds.

Combining (F.10) and (F.12), with probability at least $1-c_1\exp\left(-c_2M\gamma_N\right)$,

$$M^{-1}\sum_{i\in\mathcal{I}_k}\Gamma_i g_{\mathbf{\Delta}}(\mathbf{X}_i) \ge \gamma_N\left\{\kappa_l/4 - 2c_2'\sqrt{\frac{\log(d)}{M\gamma_N}}\|\mathbf{\Delta}\|_1\right\}, \quad \forall\|\mathbf{\Delta}\|_2=1.$$

It follows that

$$\mathbb{P}_{\mathcal{D}_N'}\left(\frac{1}{M}\sum_{i\in\mathcal{I}_k}\Gamma_i g_{\mathbf{\Delta}}(\mathbf{X}_i) \ge \gamma_N\left\{\frac{\kappa_l}{4} - 2c_2'\sqrt{\frac{\log(d)}{M\gamma_N}}\|\mathbf{\Delta}\|_1\right\}\right) \ge 1-c_1\exp\left(-c_2M\gamma_N\right).$$

Choose $a=\gamma_N$, $\kappa_1=\kappa_l/8$ and $\kappa_2=c_2'^2$, when $M\gamma_N > \max\{C_2, c_1'^2\log(M)\log(d)/(4r^2)\}$, where $C_2 = \frac{c_2'(c+d\sigma)}{ac_3'}\log(2)$, we have (F.9) holds with probability at least $1-c_1\exp(-c_2M\gamma_N)$. ∎

PROOF OF LEMMA F.2. By Theorem 1 of van de Geer and Lederer (2013), we have

$$\mathbb{P}_{\mathcal{D}_N'}\left(\mathcal{E}_\gamma\right) \le 1-2\exp(-t).$$

Moreover, On $\mathcal{E}_\gamma$, when $0<t<0.01M\gamma_N$,

$$|\widehat{\gamma}_N-\gamma_N| \le 2\sqrt{\frac{0.01M\gamma_N^2}{M}} + \frac{0.01M\gamma_N}{M} = 0.21\gamma_N.$$

Hence,

$$0.79\gamma_N \le \widehat{\gamma}_N \le 1.21\gamma_N.$$

In addition,

$$\left|\frac{\gamma_N}{\widehat{\gamma}_N}-1\right| = \frac{|\widehat{\gamma}_N-\gamma_N|}{\widehat{\gamma}_N} \le \frac{2\sqrt{t\gamma_N/M}+\sqrt{t\cdot 0.01M\gamma_N}/M}{0.79\gamma_N} \le 2.66\sqrt{\frac{t}{M\gamma_N}}.$$

∎

PROOF OF LEMMA F.3. For any $0 < t < 0.01M\gamma_N \le 0.01M\mathbb{E}(T)$, similar to Lemma F.2, we also have $\mathcal{E}_T := \{|M^{-1}\sum_{i\in\mathcal{J}} T_i - \mathbb{E}(T)| \le 2\sqrt{t\mathbb{E}(T)/M} + t/M\}$ occurs with probability at least $1 - 2\exp(-t)$. On the event $\mathcal{E}_T$,

$$\left|\frac{\mathbb{E}(T)}{M^{-1}\sum_{i\in\mathcal{J}} T_i} - 1\right| \le 2.66\sqrt{\frac{t}{M\mathbb{E}(T)}}, \quad 0.79\mathbb{E}(T) \le M^{-1}\sum_{i\in\mathcal{J}} T_i \le 1.21\mathbb{E}(T).$$

Hence, on the event $\mathcal{E}_T \cap \mathcal{E}_\gamma$,

$$\begin{aligned}
|\widehat{p}_{N,1} - p_{N,1}| &= \left|\frac{M^{-1}\sum_{i\in\mathcal{J}}\Gamma_i}{M^{-1}\sum_{i\in\mathcal{J}} T_i} - \frac{\mathbb{E}(\Gamma)}{\mathbb{E}(T)}\right| \le \frac{|\widehat{\gamma}_N - \gamma_N|}{M^{-1}\sum_{i\in\mathcal{J}} T_i} + p_{N,1}\left|\frac{\mathbb{E}(T)}{M^{-1}\sum_{i\in\mathcal{J}} T_i} - 1\right| \\
&\le \frac{1}{0.79\mathbb{E}(T)}\left(2\sqrt{\frac{t\gamma_N}{M}} + \frac{t}{M}\right) + 2.66 p_{N,1}\sqrt{\frac{t}{M\mathbb{E}(T)}} \\
&\le \frac{2.1}{0.79\mathbb{E}(T)}\sqrt{\frac{t\gamma_N}{M}} + \frac{2.66}{\mathbb{E}(T)}\sqrt{\frac{t\gamma_N p_{N,1}}{M}} \le \frac{5.32}{\mathbb{E}(T)}\sqrt{\frac{t\gamma_N}{M}},
\end{aligned}$$

since $0 < t < 0.01M\gamma_N$, $\gamma_N = p_{N,1}E(T)$, and $p_{N,1} \le 1$. It follows that

$$|\widehat{p}_{N,1} - p_{N,1}| \le \frac{5.32}{\mathbb{E}(T)}\sqrt{\frac{0.01M\gamma_N^2}{M}} \le 0.54 p_{N,1}.$$

Hence,

$$0.46 p_{N,1} \le \widehat{p}_{N,1} \le 1.54 p_{N,1}.$$

Moreover,

$$\left|\frac{p_{N,1}}{\widehat{p}_{N,1}} - 1\right| = \frac{|\widehat{p}_{N,1} - p_{N,1}|}{\widehat{p}_{N,1}} \le \frac{5.32\{\mathbb{E}(T)\}^{-1}\sqrt{t\gamma_N/M}}{0.46 p_{N,1}} \le 12\sqrt{\frac{t}{M\gamma_N}}.$$

■

PROOF OF LEMMA F.4. Under Assumption 1, we have

$$\begin{aligned}
&p_{N,1} = \mathbb{E}\{p_N(1,\mathbf{X})\} = \mathbb{E}\{\gamma_N(\mathbf{X})/\pi(\mathbf{X})\} \ge \mathbb{E}\{\gamma_N(\mathbf{X})\} = \gamma_N \text{ and} \\
&p_{N,1} = \mathbb{E}\{\gamma_N(\mathbf{X})/\pi(\mathbf{X})\} \le \mathbb{E}\{\gamma_N(\mathbf{X})\}/c = \gamma_N/c.
\end{aligned}$$

Therefore, $p_{N,1} \asymp \gamma_N$. By Taylor's theorem, we have

$$\begin{aligned}
\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) &= (M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\frac{\Gamma_i}{\widehat{\pi}(\mathbf{X}_i)}\exp\{-\mathbf{X}_i^T(\boldsymbol{\beta}^*_{p,1} + v\boldsymbol{\Delta})\}(\mathbf{X}_i^T\boldsymbol{\Delta})^2 \\
&\ge (cM\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\exp\{-\mathbf{X}_i^T(\boldsymbol{\beta}^*_{p,1} + v\boldsymbol{\Delta})\}(\mathbf{X}_i^T\boldsymbol{\Delta})^2,
\end{aligned}$$

with some $v \in (0,1)$ on the event $\mathcal{E}_\pi$. Hence, Lemma F.4 follows directly from Lemma F.1, with $a = \widehat{p}_{N,1}$, $v = v$, and $\phi(u) \equiv \exp(-u)$ using the fact that $p_{N,1} \asymp \gamma_N$. ■

PROOF OF LEMMA F.5. For any $a \in (0,1]$,

$$\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; a; \pi^*) = M^{-1}\sum_{i\in\mathcal{J}}\left\{1 - \frac{\Gamma_i}{\pi^*(\mathbf{X}_i)} - \frac{\Gamma_i}{a\pi^*(\mathbf{X}_i)}\exp(-\mathbf{X}_i^T\boldsymbol{\beta}^*_{p,1})\right\}\mathbf{X}_i.$$

Let $\mathbf{U}_i=[1-\Gamma_i/\pi^*(\mathbf{X}_i)-\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)/\{p_{N,1}\pi^*(\mathbf{X}_i)\}]\mathbf{X}_i$ and set $\mathbf{U}$ as an independent copy of $\mathbf{U}_i$. By the construction of $\boldsymbol{\beta}_{p,1}^*$, $\mathbb{E}(\mathbf{U})=\mathbf{0}$. For each $1\leq j\leq d$,

$$\sup_{1\leq j\leq d}\|\mathbf{U}(j)\|_{\psi_2}\leq\{\log(2)\}^{-1/2}\left\{1+c^{-1}+(cp_{N,1})^{-1}\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}\right\}=O(\gamma_N^{-1}),$$

since $p_{N,1}^{-1}\asymp\gamma_N$ as in Lemma F.4. In addition,

$$\begin{aligned}\sup_{1\leq j\leq d}\mathbb{E}\{\mathbf{U}^2(j)\}&\leq 2\mathbb{E}\left\{1+\Gamma/\pi^*(\mathbf{X})\right\}^2+2\mathbb{E}\left\{(cp_{N,1})^{-2}\gamma_N(\mathbf{X})\exp(-\mathbf{X}^T\boldsymbol{\beta}^*)\mathbf{X}^2(j)\right\}\\&\leq(1+c^{-1})^2+2(cp_{N,1})^{-2}\gamma_N\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}^2=O(\gamma_N^{-1}).\end{aligned}$$

By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022),

$$\mathbb{P}_{\mathcal{D}_N'}\left(\left\|M^{-1}\sum_{i\in\mathcal{J}}\mathbf{U}_i\right\|_\infty>c\sqrt{\frac{t+\log(d)}{M\gamma_N}}+\frac{c\sqrt{\log(M)}\{t+\log(d)\}}{M\gamma_N}\right)\leq 3\exp(-t),$$

with some constant $c>0$. When $M\gamma_N>C_1\log(M)\log(d)$, we have

$$\mathbb{P}_{\mathcal{D}_N'}\left(\|\nabla_{\boldsymbol{\beta}}\ell_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*;\gamma_N)\|_\infty>c\sqrt{\frac{t+t^2+\log(d)}{M\gamma_N}}\right)\leq 3\exp(-t),$$

with some constant $c>0$. Besides, observe that

$$\begin{aligned}&\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_{p,1}^*;\widehat{p}_{N,1};\pi^*)-\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*;p_{N,1};\pi^*)\|_\infty\\&\qquad=\left|\frac{\widehat{p}_{N,1}}{p_{N,1}}-1\right|\left\|(Mp_{N,1})^{-1}\sum_{i\in\mathcal{I}_k}\frac{\Gamma_i}{\pi^*(\mathbf{X}_i)}\exp(-\mathbf{X}_i^T\boldsymbol{\beta}^*)\mathbf{X}_i\right\|_\infty.\end{aligned}$$

Let $\mathbf{W}_i=\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}^*)\mathbf{X}_i/\pi^*(\mathbf{X}_i)$ set $\mathbf{W}$ as an independent copy of $\mathbf{W}_i$. Note that

$$\sup_{1\leq j\leq d}\|\mathbf{W}(j)\|_{\psi_2}=O(1)\ \text{ and }\ \sup_{1\leq j\leq d}\mathbb{E}\{\mathbf{W}^2(j)\}\leq\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}^2\gamma_N/c=O(\gamma_N).$$

By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022) and note that $p_{N,1}\asymp\gamma_N$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_N'}\Bigg(&\left\|(Mp_{N,1})^{-1}\sum_{i\in\mathcal{J}}\mathbf{W}_i-p_{N,1}^{-1}\mathbb{E}(\mathbf{W})\right\|_\infty\\&>c\sqrt{\frac{t+\log(d)}{M\gamma_N}}+\frac{c\sqrt{\log(M)}\{t+\log(d)\}}{M\gamma_N}\Bigg)\leq 3\exp(-t),\end{aligned}$$

with some constant $c>0$. Notice that, we also have $\|p_{N,1}^{-1}\mathbb{E}(\mathbf{W})\|_\infty\leq\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}/c$. Hence, when $0<t<M\gamma_N/\{\log(M)\log(d)\}$ and $M\gamma_N>C_1\log(M)\log(d)$ with some constant $C_1>0$, we have

$$\mathbb{P}_{\mathcal{D}_N'}\left(\left\|(M\gamma_N)^{-1}\sum_{i\in\mathcal{J}}\mathbf{W}_i\right\|_\infty>c\left\{1+\sqrt{\frac{\log(d)}{M\gamma_N}}\right\}\right)\leq 3\exp(-t),$$

with some constant $c>0$. For any $0<t<0.01M\gamma_N$, by Lemma F.3, we have

$$\left|\frac{p_{N,1}}{\widehat{p}_{N,1}}-1\right|\leq 12\sqrt{\frac{t}{Mp_{N,1}}}\asymp\sqrt{\frac{t}{M\gamma_N}},$$

with probability at least $1-4\exp(-t)$. It follows that, when $0<t<M\gamma_N/\{100+\log(M)\log(d)\}$ and $M\gamma_N > C_1\log(M)\log(d)$,

$$\mathbb{P}_{\mathcal{D}'_N}\left(\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1};\widehat{p}_{N,1};\pi^*)-\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1};p_{N,1};\pi^*)\|_\infty > c\sqrt{\frac{t+\log(d)}{M\gamma_N}}\right)\le 7\exp(-t),$$

with some constant $c>0$. Therefore, we have

$$\mathbb{P}_{\mathcal{D}'_N}\left(\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1};\widehat{p}_{N,1};\pi^*)\|_\infty > \kappa_3\sqrt{\frac{t+\log(d)}{M\gamma_N}}\right)\le 10\exp(-t),$$

with some constant $\kappa_3>0$. ∎

PROOF OF LEMMA F.6. By the Cauchy-Schwarz inequality, for any $\boldsymbol{\Delta}\in\mathbb{R}^d$,

$$\begin{aligned}|\mathbf{r}_\pi^T\boldsymbol{\Delta}|^2 &\le M^{-2}\sum_{i\in\mathcal{J}}\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)}-\frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\sum_{i\in\mathcal{J}}\Gamma_i\left\{1+\widehat{p}_{N,1}^{-1}\exp(-\mathbf{X}_i^T\boldsymbol{\beta}^*_{p,1})\right\}^2|\mathbf{X}_i^T\boldsymbol{\Delta}|^2\\ &\le 2c^{-4}M^{-2}\{1+\widehat{p}_{N,1}^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}\sum_{i\in\mathcal{J}}\Gamma_i\{\widehat{\pi}(\mathbf{X}_i)-\pi^*(\mathbf{X}_i)\}^2\sum_{i\in\mathcal{J}}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^2.\end{aligned}$$

By Lemma F.3 and note that $p_{N,1}\asymp\gamma_N$, we have $\widehat{p}_{N,1}^{-2}=O_p(\gamma_N^{-2})$. Under Assumptions 4, we have $(M\gamma_N)^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\{\widehat{\pi}(\mathbf{X}_i)-\pi^*(\mathbf{X}_i)\}^2\le\zeta_N^2$ on the event $\mathcal{E}_\zeta$, with $\mathbb{P}_{\mathcal{D}'_N}(\mathcal{E}_\zeta)=1-o(1)$. In addition, by part (b) of Lemma E.6, we also have

$$\frac{\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^2}{M\gamma_N}\le c\left\{1+\sqrt{\frac{t}{N\gamma_N}}+\frac{t\log(N)}{N\gamma_N}\right\}\left(\frac{\|\boldsymbol{\Delta}\|_1^2\log(d)\log(N)}{N\gamma_N}+\|\boldsymbol{\Delta}\|_2^2\right),$$

uniformly for all $\boldsymbol{\Delta}\in\mathbb{R}^d$ with probability at least $1-3\exp(-t)$. Set $t=N\gamma_N/\log(N)$ and note that $M\asymp N$, we have

$$\mathbb{P}_{\mathcal{D}'_N}\left(\sup_{\boldsymbol{\Delta}\in\mathbb{R}^d\setminus\{\mathbf{0}\}}\frac{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^2}{\|\boldsymbol{\Delta}\|_1^2\log(d)\log(N)/(M\gamma_N)+\|\boldsymbol{\Delta}\|_2^2}\le c\right)\ge 1-3\exp\{N\gamma_N/\log(N)\},$$

with some constant $c>0$. Combining the results above, on the event $\mathcal{E}_\zeta$, we have uniformly for all $\boldsymbol{\Delta}\in\mathbb{R}^d$,

$$|\mathbf{r}_\pi^T\boldsymbol{\Delta}|\le c\zeta_N\left(\|\boldsymbol{\Delta}\|_1\sqrt{\frac{\log(d)\log(N)}{M\gamma_N}}+\|\boldsymbol{\Delta}\|_2\right)$$

with probability at least $1-3\exp\{N\gamma_N/\log(N)\}$ and some constant $c>0$. ∎

PROOF OF LEMMA F.7. Let $S\subset\{1,\dots,d\}$ be the support set of $\boldsymbol{\beta}^*_{p,1}$. For any $\boldsymbol{\Delta}\in\mathbb{R}^d$, we have $\|\boldsymbol{\beta}^*_{p,1}+\boldsymbol{\Delta}\|_1=\|\boldsymbol{\beta}^*_{p,1,S}+\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1\ge\|\boldsymbol{\beta}^*_{p,1,S}\|_1-\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1=\|\boldsymbol{\beta}^*_{p,1}\|_1-\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1$. On the event $\widetilde{\mathcal{B}}_2$, we also have $\|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1};\widehat{p}_{N,1};\pi^*)\|_\infty\le\kappa_3\sqrt{\{t+\log(d)\}/(M\gamma_N)}$. Choose $\lambda_{\boldsymbol{\beta}}>2\kappa_3\sqrt{\{t+\log(d)\}/(M\gamma_N)}$, then it follows that $|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1};\widehat{p}_{N,1};\pi^*)^T\boldsymbol{\Delta}|\le\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/2\le\lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}_{S^c}\|_1+\|\boldsymbol{\Delta}_S\|_1)/2$. Together with (F.7), we have

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta})&\ge\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta};\widehat{p}_{N,1};\boldsymbol{\beta}^*_{p,1};\widehat{\pi})+\lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}_{S^c}\|_1-\|\boldsymbol{\Delta}_S\|_1-\|\boldsymbol{\Delta}_{S^c}\|_1/2-\|\boldsymbol{\Delta}_S\|_1/2)+\mathbf{r}_\pi^T\boldsymbol{\Delta}\\ &=\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta};\widehat{p}_{N,1};\boldsymbol{\beta}^*_{p,1};\widehat{\pi})+\lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}\|_1-4\|\boldsymbol{\Delta}_S\|_1)/2+\mathbf{r}_\pi^T\boldsymbol{\Delta}.\end{aligned}\tag{F.13}$$

Moreover, on the event $\widetilde{\mathcal{B}}_3$,

$$|\mathbf{r}_\pi^T \boldsymbol{\Delta}| \le c\zeta_N \left( \|\boldsymbol{\Delta}\|_1 \sqrt{\frac{\log(d)\log(N)}{M\gamma_N}} + \|\boldsymbol{\Delta}\|_2 \right) \le \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\Delta}\|_1/4 + c\zeta_N \|\boldsymbol{\Delta}\|_2, \tag{F.14}$$

since $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\log(d)/(M\gamma_N)}$ and $\zeta_N \le \kappa_3/\{2c\sqrt{\log(N)}\}$ under Assumption 4 when $N$ is large enough. Together with (F.13) and (F.16), we have

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}) &\ge \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}\|_1 - 4\|\boldsymbol{\Delta}_S\|_1)/2 - \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 - c\zeta_N\|\boldsymbol{\Delta}\|_2 \\ &\ge \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 - (2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2, \qquad \text{(F.15)}\end{aligned}$$

since $\|\boldsymbol{\Delta}_S\|_1 \le \sqrt{s_{p,1}}\|\boldsymbol{\Delta}_S\|_2 \le \sqrt{s_{p,1}}\|\boldsymbol{\Delta}\|_2$.

Additionally, if $\boldsymbol{\Delta} \in \mathcal{K}(r_N, 1)$, we have $\|\boldsymbol{\Delta}\|_1 \le r_N\|\boldsymbol{\Delta}\|_2$ and $\|\boldsymbol{\Delta}\|_2 = 1$. Hence, on the event $\widetilde{\mathcal{B}}_1 \cap \mathcal{E}_p$,

$$\begin{aligned}\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) &\ge c\left\{\|\boldsymbol{\Delta}\|_2^2 - \frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\} \\ &\ge c\left\{\|\boldsymbol{\Delta}\|_2^2 - \frac{r_N^2\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_2^2\right\} \ge c\|\boldsymbol{\Delta}\|_2^2/2 = c/2, \qquad \text{(F.16)}\end{aligned}$$

when $0 < t < 0.01M\gamma_N$ and $N$ is large enough, with some constant $c > 0$. Besides, we also have $(2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2 \le c/2$ when $N$ is large enough since $\|\boldsymbol{\Delta}\|_2 = 1$ and $2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N \asymp \sqrt{s_{p,1}\log(d)/(M\gamma_N)} + \zeta_N = o(1)$. Together with (F.15) and (F.16), we have

$$\mathcal{F}(\boldsymbol{\Delta}) \ge c/4 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 > 0.$$

■

**F.4. Proof of Theorem 4.1.** In the following proofs, we will consider the events $\mathcal{E}_\gamma$, $\mathcal{E}_p$, $\widetilde{\mathcal{B}}_1$, $\widetilde{\mathcal{B}}_2$, and $\widetilde{\mathcal{B}}_3$, defined as in (F.1) – (F.6). Additionally, the events $\mathcal{E}_\pi$ and $\mathcal{E}_\zeta$ are defined in Assumption 4, and $\widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}$ is defined in Theorem 4.1.

PROOF OF THEOREM 4.1. By the construction of $\widehat{\boldsymbol{\beta}}_{p,1}$, we have

$$\widetilde{\ell}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_{p,1}; \widehat{p}_{N,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 \le \widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1.$$

Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}$, then we have $\mathcal{F}(\boldsymbol{\Delta}) \le 0$. Condition on the event $\widetilde{\mathcal{B}}_2 \cap \widetilde{\mathcal{B}}_3$. Together with Lemma F.7, when $N$ is large enough, we have

$$0 \ge \mathcal{F}(\boldsymbol{\Delta}) \ge \delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 - (2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2. \tag{F.17}$$

Since the loss function $\ell_{\boldsymbol{\beta}}(\cdot; \widehat{p}_{N,1}; \widehat{\pi})$ is convex, $\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) \ge 0$, and it follows that $\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 \le (2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2$. That is,

$$\|\boldsymbol{\Delta}\|_1 \le 4(2\sqrt{s_{p,1}} + c\zeta_N/\lambda_{\boldsymbol{\beta}})\|\boldsymbol{\Delta}\|_2 \le \left(8\sqrt{s_{p,1}} + \frac{2c\zeta_N}{\kappa_3}\sqrt{\frac{M\gamma_N}{\log(d)}}\right)\|\boldsymbol{\Delta}\|_2, \tag{F.18}$$

since $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\log(d)/(M\gamma_N)}$. Let $r_N = 8\sqrt{s_{p,1}} + \kappa_3^{-1}2c\zeta_N\sqrt{M\gamma_N/\log(d)}$. Then $\|\boldsymbol{\Delta}\|_1 \le r_N\|\boldsymbol{\Delta}_2\|_2$ with $r_N = o(\sqrt{M\gamma_N/\log(d)})$ under Assumption 4 and since $s_{p,1} = o(M\gamma_N/\log(d))$.

In the following, we further prove that $\|\boldsymbol{\Delta}\|_2 \leq 1$ by contradiction. Suppose that $\|\boldsymbol{\Delta}\|_2 > 1$. Define $\widetilde{\boldsymbol{\Delta}} := \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_2$, then $\|\widetilde{\boldsymbol{\Delta}}\|_2 = 1$ and $\|\widetilde{\boldsymbol{\Delta}}\|_1 = \|\boldsymbol{\Delta}\|_1/\|\boldsymbol{\Delta}\|_2 \leq r_N = r_N\|\widetilde{\boldsymbol{\Delta}}\|_2$. That is, $\widetilde{\boldsymbol{\Delta}} \in \mathcal{K}(r_N, 1)$. By Lemma F.7, $\mathcal{F}(\widetilde{\boldsymbol{\Delta}}) > 0$. Define $u = 1/\|\boldsymbol{\Delta}\|_2$, then $0 < u < 1$. Note that $\mathcal{F}(\cdot)$ is a convex function, $\mathcal{F}(\mathbf{0}) = 0$, and $\mathcal{F}(\boldsymbol{\Delta}) \leq 0$. Hence,

$$0 < \mathcal{F}(\widetilde{\boldsymbol{\Delta}}) = \mathcal{F}(u\boldsymbol{\Delta} + (1-u)\mathbf{0}) \leq u\mathcal{F}(\boldsymbol{\Delta}) + (1-u)\mathcal{F}(\mathbf{0}) = u\mathcal{F}(\boldsymbol{\Delta}) \leq 0.$$

Therefore, we must have $\|\boldsymbol{\Delta}\|_2 \leq 1$.

Now, further condition on $\widetilde{\mathcal{B}}_1 \cap \mathcal{E}_p$. As in (F.16), we also have

$$\begin{aligned}
\delta\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}; \boldsymbol{\beta}^*_{p,1}; \widehat{\pi}) &\geq c\left\{\|\boldsymbol{\Delta}\|_2^2 - \frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\} \\
&\geq c\left\{\|\boldsymbol{\Delta}\|_2^2 - \frac{r_N^2\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_2^2\right\} \geq c\|\boldsymbol{\Delta}\|_2^2/2, \qquad \text{(F.19)}
\end{aligned}$$

when $N$ is large enough, with some constant $c > 0$. Together with (F.17), we have

$$c\|\boldsymbol{\Delta}\|_2^2/2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 \leq (2\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}} + c\zeta_N)\|\boldsymbol{\Delta}\|_2.$$

It follows that $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}$ satisfies

$$\text{(F.20)} \quad \|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_2 \leq 4\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}}/c + 2\zeta_N,$$

$$\text{(F.21)} \quad \|\widehat{\boldsymbol{\beta}}_{p,1} - \boldsymbol{\beta}^*_{p,1}\|_1 \leq r_N\|\boldsymbol{\Delta}\|_2 \leq \left(8\sqrt{s_{p,1}} + \frac{2c\zeta_N}{\kappa_3}\sqrt{\frac{M\gamma_N}{\log(d)}}\right)\left(4\sqrt{s_{p,1}}\lambda_{\boldsymbol{\beta}}/c + 2\zeta_N\right).$$

Since $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(d)/(M\gamma_N)}$ and $\widetilde{\mathcal{B}}_1 \cap \widetilde{\mathcal{B}}_2 \cap \widetilde{\mathcal{B}}_3 \cap \mathcal{E}_p \cap \mathcal{E}_\pi = 1 - o(1)$, we conclude that

$$\|\boldsymbol{\Delta}\|_2 = O_p\left(\sqrt{\frac{s_{p,1}\log(d)}{M\gamma_N}} + \zeta_N\right), \quad \|\boldsymbol{\Delta}\|_1 = O_p\left(s_{p,1}\sqrt{\frac{\log(d)}{M\gamma_N}} + \zeta_N^2\sqrt{\frac{M\gamma_N}{\log(d)}}\right).$$

In addition, by (F.7), (F.19), and $\mathcal{F}(\boldsymbol{\Delta}) \leq 0$, we also have

$$\begin{aligned}
\lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 &\leq |\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)^T\boldsymbol{\Delta}| + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1 + |\mathbf{r}_\pi^T\boldsymbol{\Delta}| - c\|\boldsymbol{\Delta}\|_2^2/2 \\
&\overset{(i)}{\leq} \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/4 + c\zeta_N\|\boldsymbol{\Delta}\|_2 - c\|\boldsymbol{\Delta}\|_2^2/2 \\
&\overset{(ii)}{\leq} 3\lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1/4 + 7\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1/4 + c\zeta_N\|\boldsymbol{\Delta}\|_2 - c\|\boldsymbol{\Delta}\|_2^2/2,
\end{aligned}$$

where (i) holds by $|\nabla_{\boldsymbol{\beta}}\widetilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*_{p,1}; \widehat{p}_{N,1}; \pi^*)^T\boldsymbol{\Delta}| \leq \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/2$ and (F.14); (ii) holds since $\|\boldsymbol{\Delta}\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 + \|\boldsymbol{\beta}^*_{p,1}\|_1$. Then, it follows that

$$\lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 \leq 7\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*_{p,1}\|_1 + 4c\zeta_N\|\boldsymbol{\Delta}\|_2 - 2c\|\boldsymbol{\Delta}\|_2^2 \leq 7\lambda_{\boldsymbol{\beta}}C_{\boldsymbol{\beta}} + 2c\zeta_N^2,$$

since $\|\boldsymbol{\beta}^*_{p,1}\|_1 \leq C_{\boldsymbol{\beta}}$ under Assumption 5 and $4c\zeta_N\|\boldsymbol{\Delta}\|_2 - 2c\|\boldsymbol{\Delta}\|_2^2 = -2c(\|\boldsymbol{\Delta}\|_2 - \zeta_N)^2 + 2c\zeta_N^2 \leq 2c\zeta_N^2$. Hence, we have $\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1 \leq 8C_{\boldsymbol{\beta}}$ as long as $\zeta_N^2 \leq (2c)^{-1}\lambda_{\boldsymbol{\beta}}C_{\boldsymbol{\beta}} \asymp \sqrt{\log(d)/(M\gamma_N)}$, which occurs when $\zeta_N = o(\{\log(d)/(M\gamma_N)\}^{1/4})$ and $N$ is large enough. ∎

### F.5. Proof of the results in Section F.2.

PROOF OF LEMMA F.8. By Lemma F.4, $p_{N,1} \asymp \gamma_N$. Under Assumptions 4 and on the event $\mathcal{E}_\pi \cap \widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}$, we have

$$\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta}; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}; \widetilde{\boldsymbol{\alpha}}^*; \widehat{\pi}) = (M\widehat{p}_{N,1})^{-1} \sum_{i\in\mathcal{J}} \frac{\Gamma_i}{\widehat{\pi}(\mathbf{X}_i)} \exp(-\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1})(\mathbf{X}_i^T \boldsymbol{\Delta})^2 \tag{F.22}$$

$$\geq \exp(-8C_{\mathbf{X}} C_{\boldsymbol{\beta}})(cM\widehat{p}_{N,1})^{-1} \sum_{i\in\mathcal{J}} \Gamma_i (\mathbf{X}_i^T \boldsymbol{\Delta})^2. \tag{F.23}$$

Hence, Lemma F.8 follows directly from Lemma F.1, with $a = \widehat{p}_{N,1}$, $v = 0$, and $\phi(u) \equiv 1$ using the fact that $p_{N,1} \asymp \gamma_N$. ∎

PROOF OF LEMMA F.9. Assume that $m(\mathbf{x}) = \widetilde{m}^*(\mathbf{x}) = \mathbf{x}^T \widetilde{\boldsymbol{\alpha}}^*$. Note that

$$\nabla_{\boldsymbol{\alpha}} \widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi}) = -\frac{2\sum_{i\in\mathcal{J}} \Gamma_i \exp(-\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1}) \mathbf{X}_i \varepsilon_i}{M\widehat{p}_{N,1} \widehat{\pi}(\mathbf{X}_i)} = (M\widehat{p}_{N,1})^{-1} \sum_{i\in\mathcal{J}} \widehat{\mathbf{V}}_i,$$

where $\widehat{\mathbf{V}}_i = -2\Gamma_i \exp(-\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1}) \mathbf{X}_i \varepsilon_i / \widehat{\pi}(\mathbf{X}_i)$. For each $1 \leq j \leq d$ and $i \in \mathcal{J}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}'_N} \left\{ \widehat{\mathbf{V}}_i(j) \mid (R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}} \right\} \\
&\overset{(i)}{=} -2\Gamma_i \widehat{\pi}^{-1}(\mathbf{X}_i) \exp(-\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1}) \mathbf{X}_i(j) \mathbb{E}_{\mathcal{D}'_N} \{\varepsilon_i \mid (R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}}\} \\
&\overset{(ii)}{=} -2\Gamma_i \widehat{\pi}^{-1}(\mathbf{X}_i) \exp(-\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1}) \mathbf{X}_i(j) \mathbb{E}_{\mathcal{D}'_N} (\varepsilon_i \mid \mathbf{X}_i) = 0,
\end{aligned}$$

where (i) holds since by construction, $\widehat{\pi}(\cdot)$ and $\widehat{\boldsymbol{\beta}}_{p,1}$ are only dependent on $(R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}}$; (ii) holds under Assumption 1. By Proposition 2.5 (Hoeffding bound) of Wainwright (2019), for any $u > 0$,

$$\mathbb{P}_{\mathcal{D}'_N} \left( \left| \sum_{i\in\mathcal{J}} \widehat{\mathbf{V}}_i(j) \right| > u \mid (R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}} \right) \leq 2\exp\left( -\frac{u^2}{2\sum_{i\in\mathcal{J}} \sigma_{ij}^2} \right),$$

if $\widehat{\mathbf{V}}_i(j)$ is sub-gaussian with parameter $\sigma_{ij} > 0$ conditional on $(R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}}$. Condition on $\mathcal{E}_\pi \cap \widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}$. Under Assumption 6, $\sigma_{ij}^2$ can be chosen such that

$$\sigma_{ij}^2 \leq 4\sigma_\varepsilon^2 \Gamma_i \widehat{\pi}^{-2}(\mathbf{X}_i) \exp(-2\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{p,1}) \mathbf{X}_i^2(j) \leq c^2 \Gamma_i,$$

with some constant $c > 0$. Let $u = c(1+t)\sqrt{\log(d) \sum_{i\in\mathcal{J}} \Gamma_i}$. By the union bound,

$$\mathbb{P}_{\mathcal{D}'_N} \left( \left\| \sum_{i\in\mathcal{J}} \widehat{\mathbf{V}}_i \right\|_\infty > u \mid (R_i, T_i, \mathbf{X}_i)_{i\in\mathcal{J}} \right) \leq 2\exp\{-t\log(d)\}.$$

By Lemma F.3, when $0 < t < 0.01Mp_{N,1}$, with probability at least $1 - 4\exp(-t)$, we have $\widehat{p}_{N,1} \geq 0.46p_{N,1}$. Meanwhile, we also note that $\widehat{p}_{N,1} = \sum_{i\in\mathcal{J}} \Gamma_i / \sum_{i\in\mathcal{J}} T_i \leq \sum_{i\in\mathcal{J}} \Gamma_i / M = \widehat{\gamma}_N$. Hence

$$\frac{u}{M\widehat{p}_{N,1}} = \frac{c(1+t)\sqrt{\log(d) M\widehat{\gamma}_N}}{M\widehat{p}_{N,1}} \leq c(1+t)\sqrt{\frac{\log(d)}{M\widehat{p}_{N,1}}} \leq c(1+t)\sqrt{\frac{\log(d)}{0.46Mp_{N,1}}}.$$

In addition, note that $p_{N,1} \asymp \gamma_N$ by Lemma F.4. Therefore, on the event $\mathcal{E}_\pi \cap \widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}$,

$$\|\nabla_{\boldsymbol{\alpha}} \widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi})\|_\infty \le \kappa_4(1+t)\sqrt{\frac{\log(d)}{M\gamma_N}},$$

with probability at least $1 - 2\exp\{-t\log(d)\} - 4\exp(-t)$ and some constant $\kappa_4 > 0$. ∎

PROOF OF LEMMA F.10. Consider the general case that $m(\mathbf{x}) \neq m^*(\mathbf{x}) = \mathbf{x}^T\widetilde{\boldsymbol{\alpha}}^*$ is allowed. Let $\mathbf{V}_i = -2\Gamma_i \exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i\varepsilon_i/\pi^*(\mathbf{X}_i)$ and let $\mathbf{V}$ be its independent copy. Then,

$$\|\nabla_{\boldsymbol{\alpha}} \widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \boldsymbol{\beta}_{p,1}^*; \pi^*)\|_\infty = (M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}} \mathbf{V}_i.$$

By the construction of $\widetilde{\boldsymbol{\alpha}}^*$, we have $\mathbb{E}(\mathbf{V}) = \mathbf{0} \in \mathbb{R}^d$. Under Assumptions 6, 4, and 5, we have $\sup_{1\le j\le d}\|\mathbf{V}(j)\|_{\psi_2} \le 2c^{-1}\sigma_\varepsilon \exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}$ for each $j \le d$. In addition, under Assumption 3, with some $r > 1$ satisfies $1/r + 1/q = 1$,

$$\begin{aligned}\sup_{1\le j\le d} \mathbb{E}\{\mathbf{V}^2(j)\} &\le 4c^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}^2\mathbb{E}\{\gamma_N(\mathbf{X})\varepsilon^2\} \\ &\le 4c^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})C_{\mathbf{X}}^2\|\gamma_N(\cdot)\|_{\mathbb{P},q}\|\varepsilon\|_{\mathbb{P},2r}^2 = O(\gamma_N).\end{aligned}$$

By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022), with some constant $c > 0$,

$$\mathbb{P}\left(\left\|M^{-1}\sum_{i\in\mathcal{J}}\mathbf{V}_i\right\|_\infty > c\sqrt{\frac{\gamma_N\{t+\log(d)\}}{M}} + c\frac{\sqrt{\log(M)}\{t+\log(d)\}}{M}\right) \le 3\exp(-t).$$

Together with Lemma F.3, when $0 < t < M\gamma_N/\{100 + \log(M)\log(d)\}$ and $M\gamma_N > C_1\log(M)\log(d)$,

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{\alpha}} \widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \boldsymbol{\beta}_{p,1}^*; \pi^*)\|_\infty > \kappa_4\sqrt{\frac{t+\log(d)}{M\gamma_N}}\right) \le 7\exp(-t),$$

with some constant $\kappa_4 > 0$. ∎

### F.6. Proof of Theorems 4.2 and 4.3.

PROOF OF THEOREM 4.2. Let the OR model be correctly specified. For any $0 < t < 0.01M\gamma_N$, let $\lambda_{\boldsymbol{\beta}} > 2\kappa_3\sqrt{\{t+\log(d)\}/(M\gamma_N)}$ and $\lambda_{\boldsymbol{\alpha}} > 2\kappa_4\sqrt{\{t+\log(d)\}(M\gamma_N)}$. Then, on the event $\widetilde{\mathcal{A}}_2$, we have $2\|\nabla_{\boldsymbol{\alpha}} \widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*; \widehat{p}_{N,1}, \widehat{\boldsymbol{\beta}}_{p,1}; \widehat{\pi})\|_\infty \le \lambda_{\boldsymbol{\alpha}}$. By Proposition 9.13 of Wainwright (2019), on the event $\mathcal{A}_2$, we have $\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^* \in \mathcal{C}(S,3) := \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}_{S^c}\|_1 \le 3\|\boldsymbol{\Delta}_S\|_1\}$, where $S \subseteq \{1,\dots,d\}$ is the support set of $\widetilde{\boldsymbol{\alpha}}^*$. By Corollary 9.20 of Wainwright (2019), on $\widetilde{\mathcal{A}}_1 \cap \widetilde{\mathcal{A}}_2$, when $M\gamma_N \ge 32\kappa_2 s_{\widetilde{\boldsymbol{\alpha}}}\log(d)/\kappa_1$,

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_1 \le \frac{3s_{\widetilde{\boldsymbol{\alpha}}}\lambda_{\boldsymbol{\alpha}}}{\kappa_1}, \quad \|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_2 \le \frac{3\sqrt{s_{\widetilde{\boldsymbol{\alpha}}}}\lambda_{\boldsymbol{\alpha}}}{2\kappa_1}.$$

Here, by Lemmas F.5, F.8, and F.9, as well as Theorem 4.1 and Assumption 4,

$$\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{A}}_1 \cap \widetilde{\mathcal{A}}_2) \ge 1 - 12\exp(-t) - 2\exp\{-t\log(d)\} - c_1\exp(-c_2M\gamma_N) - o(1).$$

Hence, if $M\gamma_N \gg \max\{\log(M), s_{p,1}, s_{\widetilde{\boldsymbol{\alpha}}}\}\log(d)$, with some $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(d)/(M\gamma_N)}$ and $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log(d)/(M\gamma_N)}$, as $N, d \to \infty$,

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_1 = O_p\left(s_{\widetilde{\boldsymbol{\alpha}}}\sqrt{\frac{\log(d)}{M\gamma_N}}\right), \quad \|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_2 = O_p\left(\sqrt{\frac{s_{\widetilde{\boldsymbol{\alpha}}}\log(d)}{M\gamma_N}}\right).$$

∎

PROOF OF THEOREM 4.3. Consider the case that the OR model is possibly misspecified. For any $0 < t < M\gamma_N/\{100 + \log(M)\log(d)\}$, let $\lambda_{\boldsymbol\beta} > 2\kappa_3\sqrt{\{t + \log(d)\}(M\gamma_N)}$ and $\lambda_{\boldsymbol\alpha} > 4\kappa_4\sqrt{\{t + \log(d)\}(M\gamma_N)}$. Condition on the event $\widetilde{\mathcal{A}}_3$. Then, we have

$$4\|\nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\boldsymbol\beta^*_{p,1};\pi^*)\|_\infty \le \lambda_{\boldsymbol\alpha}. \tag{F.24}$$

By Lemma F.9, $\mathbb{P}_{\mathcal{D}'_N}(\widetilde{\mathcal{A}}_3) \ge 1 - 2\exp\{-t\log(d)\} - 4\exp(-t)$. By the construction of $\widetilde{\boldsymbol\alpha}$, we have

$$\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha};\widehat{p}_{N,1},\widehat{\boldsymbol\beta}_{p,1};\widehat{\pi}) + \lambda_{\boldsymbol\alpha}\|\widetilde{\boldsymbol\alpha}\|_1 \le \widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\widehat{\boldsymbol\beta}^*_{p,1};\widehat{\pi}) + \lambda_{\boldsymbol\alpha}\|\widetilde{\boldsymbol\alpha}^*\|_1.$$

Let $\boldsymbol\Delta = \widetilde{\boldsymbol\alpha} - \widetilde{\boldsymbol\alpha}^*$. Then,

$$\begin{aligned}\delta\widetilde{\ell}_{\boldsymbol\alpha}(\boldsymbol\Delta;\widehat{p}_{N,1},\widehat{\boldsymbol\beta};\widetilde{\boldsymbol\alpha}^*;\widehat{\pi}) + \lambda_{\boldsymbol\alpha}\|\widetilde{\boldsymbol\alpha}\|_1 &\le \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\widehat{\boldsymbol\beta}^*_{p,1};\widehat{\pi})^T\boldsymbol\Delta + \lambda_{\boldsymbol\alpha}\|\widetilde{\boldsymbol\alpha}^*\|_1\\ &= \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\boldsymbol\beta^*_{p,1};\pi^*)^T\boldsymbol\Delta + \lambda_{\boldsymbol\alpha}\|\widetilde{\boldsymbol\alpha}^*\|_1 + (\mathbf{r}_{p,1} + \mathbf{r}_{p,2})^T\boldsymbol\Delta, \qquad \text{(F.25)}\end{aligned}$$

where

$$\begin{aligned}\mathbf{r}_{p,1} :=& \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\widehat{\boldsymbol\beta}_{p,1};\widehat{\pi}) - \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\boldsymbol\beta^*_{p,1};\widehat{\pi})\\ =& -2(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\left\{\exp(-\mathbf{X}_i^T\widehat{\boldsymbol\beta}_{p,1}) - \exp(-\mathbf{X}_i^T\boldsymbol\beta^*_{p,1})\right\}\Gamma_i\mathbf{X}_i\varepsilon_i/\widehat{\pi}(\mathbf{X}_i)\end{aligned}$$

and

$$\begin{aligned}\mathbf{r}_{p,2} :=& \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\boldsymbol\beta^*_{p,1};\widehat{\pi}) - \nabla_{\boldsymbol\alpha}\widetilde{\ell}_{\boldsymbol\alpha}(\widetilde{\boldsymbol\alpha}^*;\widehat{p}_{N,1},\boldsymbol\beta^*_{p,1};\pi^*)\\ =& -2(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}\exp(-\mathbf{X}_i^T\boldsymbol\beta^*_{p,1})\Gamma_i\mathbf{X}_i\varepsilon_i.\end{aligned}$$

We first consider the error term $\mathbf{r}_{p,2}^T\boldsymbol\Delta$. For any $a_1 > 0$, we have

$$|\mathbf{r}_{p,2}^T\boldsymbol\Delta| \le (M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left[a_1\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\varepsilon_i^2 + a_1^{-1}\exp(-2\mathbf{X}_i^T\boldsymbol\beta^*_{p,1})|\mathbf{X}_i^T\boldsymbol\Delta|^2\right]$$

Under Assumptions 1 and 6, we have

$$\begin{aligned}&(M\widehat{p}_{N,1})^{-1}\mathbb{E}_{\mathcal{D}'_N}\left[\sum_{i\in\mathcal{J}}\Gamma_i\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\varepsilon_i^2 \mid (R_i,T_i,\mathbf{X}_i)_{i\in\mathcal{J}}\right]\\ &= (M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\mathbb{E}_{\mathcal{D}'_N}\left\{\varepsilon_i^2 \mid (R_i,T_i,\mathbf{X}_i)_{i\in\mathcal{J}}\right\}\\ &= (M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\mathbb{E}_{\mathcal{D}'_N}(\varepsilon_i^2 \mid \mathbf{X}_i) \le c\zeta_N^2,\end{aligned}$$

on the event $\mathcal{E}_\zeta \cap \mathcal{E}_p$ with some constant $c > 0$. By Markov's inequality, for any $t > 0$, the event

$$\mathcal{E}_1 := \left\{(M\gamma_N)^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left\{\frac{1}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\varepsilon_i^2 \le ct\zeta_N^2\right\}$$

occurs with probability at least $1-t^{-1}$ conditional on $\mathcal{E}_\zeta \cap \mathcal{E}_p$. Additionally, together with (F.23), we have

$$\begin{aligned}(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\exp(-2\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)|\mathbf{X}_i^T\boldsymbol{\Delta}|^2 &\le \exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^2\\ &\le c\exp(10C_{\mathbf{X}}C_{\boldsymbol{\beta}})\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi}).\end{aligned}$$

Choose $a_1=4c\exp(10C_{\mathbf{X}}C_{\boldsymbol{\beta}})$, we obtain that, on the event $\mathcal{E}_1$,

$$|\mathbf{r}_{p,2}^T\boldsymbol{\Delta}|\le a_1ct\zeta_N^2+\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})/4.$$

Now, we control the term $\mathbf{r}_{p,1}^T\boldsymbol{\Delta}$. By Taylor's theorem, with some $\widetilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\beta}_{p,1}^*$ and $\widehat{\boldsymbol{\beta}}_{p,1}$, we have

$$\begin{aligned}|\mathbf{r}_{p,1}^T\boldsymbol{\Delta}| &= \left|2(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\exp(-\mathbf{X}_i^T\widetilde{\boldsymbol{\beta}})\Gamma_i\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}_{p,1}-\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i^T\boldsymbol{\Delta}\varepsilon_i/\widehat{\pi}(\mathbf{X}_i)\right|\\ &\le \exp(8C_{\mathbf{X}}C_{\boldsymbol{\beta}})(cM\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left[a_2\left\{\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}_{p,1}-\boldsymbol{\beta}_{p,1}^*)\right\}^2\varepsilon_i^2+a_2^{-1}|\mathbf{X}_i^T\boldsymbol{\Delta}|^2\right],\end{aligned}$$

for any $a_2>0$ on the event $\widetilde{\mathcal{E}}_{\widetilde{\boldsymbol{\beta}}}$ since $\|\widetilde{\boldsymbol{\beta}}\|_1\le\|\widehat{\boldsymbol{\beta}}_{p,1}\|_1\vee\|\boldsymbol{\beta}_{p,1}^*\|_1\le 8C_{\boldsymbol{\beta}}$. Choose $s_0:=\lceil N\gamma_N/\{\log(d)\log(N)\}\rceil$, then $s_{p,1}=o(s_0)$ and $\zeta_N^2\lambda_{\boldsymbol{\beta}}^2/s_0=o(1)$ since $N\gamma_N\gg\log(d)\log(N)$, $\zeta_N=o(1)$, $\lambda_{\boldsymbol{\beta}}=o(1)$, and $s_{p,1}=o(N\gamma_N/\{\log(d)\log(N)\})$. When (F.20) and (F.21) hold, we have

$$\begin{aligned}\frac{\|\widehat{\boldsymbol{\beta}}_{p,1}-\boldsymbol{\beta}_{p,1}^*\|_1^2}{s_0}+\|\widehat{\boldsymbol{\beta}}_{p,1}-\boldsymbol{\beta}_{p,1}^*\|_2^2 &= O\left((s_{p,1}\lambda_{\boldsymbol{\beta}}^2+\zeta_N^2)(1+s_{p,1}/s_0)+(\lambda_{\boldsymbol{\beta}}^2+\zeta_N^2)\zeta_N^2\lambda_{\boldsymbol{\beta}}^2/s_0\right)\\ &= O(s_{p,1}\lambda_{\boldsymbol{\beta}}^2+\zeta_N^2).\end{aligned}$$

Together with part (f) of Lemma E.6, we also have the event

$$\mathcal{E}_2:=\left\{(M\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i\left\{\mathbf{X}_i^T(\widehat{\boldsymbol{\beta}}_{p,1}-\boldsymbol{\beta}_{p,1}^*)\right\}^2\varepsilon_i^2\le c(s_{p,1}\lambda_{\boldsymbol{\beta}}^2+\zeta_N^2)\right\}$$

occurs with probability at least $1-3\exp(-t)-t^{-1}$ and some constant $c>0$ when $N$ is large enough, conditional on $\widetilde{\mathcal{B}}_1\cap\widetilde{\mathcal{B}}_2\cap\widetilde{\mathcal{B}}_3\cap\mathcal{E}_p\cap\mathcal{E}_\pi$ for any $0<t<M\gamma_N/\{100+\log(M)\log(d)\}$. In addition, chose $a_2=4c\exp(16C_{\mathbf{X}}C_{\boldsymbol{\beta}})$, we also have

$$\exp(8C_{\mathbf{X}}C_{\boldsymbol{\beta}})(a_2cM\widehat{p}_{N,1})^{-1}\sum_{i\in\mathcal{J}}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}|^2\le\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})/4.$$

Condition on the event $\mathcal{E}_1\cap\mathcal{E}_2$. Then,

$$|(\mathbf{r}_{p,1}+\mathbf{r}_{p,2})^T\boldsymbol{\Delta}|\le c(t+1)\zeta_N^2+cs_{p,1}\lambda_{\boldsymbol{\beta}}^2+\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})/2.$$

Besides, with $S\subseteq\{1,\dots,d\}$ denoting the support set of $\widetilde{\boldsymbol{\alpha}}^*$, we have $\|\widetilde{\boldsymbol{\alpha}}\|_1=\|\widetilde{\boldsymbol{\alpha}}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1\ge\|\widetilde{\boldsymbol{\alpha}}_S^*\|_1-\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1=\|\widetilde{\boldsymbol{\alpha}}^*\|_1-\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1$. On the event $\widetilde{\mathcal{A}}_3$, we also have $|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*;\widehat{p}_{N,1},\boldsymbol{\beta}_{p,1}^*;\pi^*)^T\boldsymbol{\Delta}|\le\|\nabla_{\boldsymbol{\alpha}}\widetilde{\ell}_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\alpha}}^*;\widehat{p}_{N,1},\boldsymbol{\beta}_{p,1}^*;\pi^*)\|_\infty\|\boldsymbol{\Delta}\|_1\le\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\Delta}\|_1/2$ as long as $\lambda_{\boldsymbol{\alpha}}>2\kappa_3\sqrt{\frac{t+\log(d)}{M\gamma_N}}$. Together with (F.25), it follows that

$$\begin{aligned}&\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})+\lambda_{\boldsymbol{\alpha}}(\|\boldsymbol{\Delta}_{S^c}\|_1-\|\boldsymbol{\Delta}_S\|_1)\\ &\qquad\le\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}\|_1/2+c(t+1)\zeta_N^2+cs_{p,1}\lambda_{\boldsymbol{\beta}}^2+\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})/2.\end{aligned}$$

That is,

$$\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})+\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_{S^c}\|_1\leq 3\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_S\|_1+2c(t+1)\zeta_N^2+2cs_{p,1}\lambda_{\boldsymbol{\beta}}^2. \tag{F.26}$$

**Case 1.** If $\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_S\|_1\leq 2c(t+1)\zeta_N^2+2cs_{p,1}\lambda_{\boldsymbol{\beta}}^2$. Then, we have

$$\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_{S^c}\|_1\leq 8c\{(t+1)\zeta_N^2+s_{p,1}\lambda_{\boldsymbol{\beta}}^2\},$$

and hence

$$\|\boldsymbol{\Delta}\|_1=\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1\leq c\left\{(t+1)\zeta_N^2+\frac{s_{p,1}\log(d)}{M\gamma_N}\right\}\sqrt{\frac{M\gamma_N}{\log(d)}},$$

with some constant $c>0$ when $\lambda_{\boldsymbol{\alpha}}\asymp\lambda_{\boldsymbol{\beta}}\asymp\sqrt{\frac{t+\log(d)}{M\gamma_N}}$. Additionally, by (F.26) and on the event $\widetilde{\mathcal{A}}_1\cap\mathcal{E}_p$, we have

$$\|\boldsymbol{\Delta}\|_2^2\leq c\{(t+1)\zeta_N^2+s_{p,1}\lambda_{\boldsymbol{\beta}}^2\}+c\left\{(t+1)\zeta_N^2+\frac{s_{p,1}\log(d)}{M\gamma_N}\right\}^2\frac{M\gamma_N}{\log(d)}\cdot\frac{\log(d)}{M\gamma_N},$$

with some constant $c>0$. As $\zeta_N=o(1)$ and $s_{p,1}=o(M\gamma_N/\log(d))$, when $N$ is large enough, we have

$$\|\boldsymbol{\Delta}\|_2\leq c\left\{(t+1)\zeta_N+\sqrt{\frac{s_{p,1}\log(d)}{M\gamma_N}}\right\},$$

with some constant $c>0$.

**Case 2.** If $\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_S\|_1>2c(t+1)\zeta_N^2+2cs_{p,1}\lambda_{\boldsymbol{\beta}}^2$. Then, (F.26) implies that $\|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}^c}\|_1\leq 4\|\boldsymbol{\Delta}_{S_{\boldsymbol{\alpha}}}\|_1$. Hence,

$$\|\boldsymbol{\Delta}\|_1=\|\boldsymbol{\Delta}_S\|_1+\|\boldsymbol{\Delta}_{S^c}\|_1\leq 5\|\boldsymbol{\Delta}_S\|_1\leq 5\sqrt{s_{\widetilde{\boldsymbol{\alpha}}}}\|\boldsymbol{\Delta}_S\|_2\leq 5\sqrt{s_{\widetilde{\boldsymbol{\alpha}}}}\|\boldsymbol{\Delta}\|_2. \tag{F.27}$$

Besides, on the event $\widetilde{\mathcal{A}}_1\cap\mathcal{E}_p$ and together with (F.26), we also have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2^2&\leq c\left\{\delta\widetilde{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\Delta};\widehat{p}_{N,1},\widehat{\boldsymbol{\beta}};\widetilde{\boldsymbol{\alpha}}^*;\widehat{\pi})+\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\}\\&\leq c\left\{4\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}_S\|_1+\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\}\leq c\left\{4\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\Delta}\|_1+\frac{\log(d)}{M\gamma_N}\|\boldsymbol{\Delta}\|_1^2\right\}.\end{aligned}$$

Meanwhile, by (F.27), we also have $\|\boldsymbol{\Delta}\|_2^2\geq\|\boldsymbol{\Delta}\|_1^2/(25s_{\widetilde{\boldsymbol{\alpha}}})$. Since $s_{\widetilde{\boldsymbol{\alpha}}}=o(M\gamma_N/\log(d))$ and $\lambda_{\boldsymbol{\alpha}}\asymp\sqrt{\log(d)/(M\gamma_N)}$, we have

$$\|\boldsymbol{\Delta}\|_1\leq cs_{\widetilde{\boldsymbol{\alpha}}}\sqrt{\frac{\log(d)}{M\gamma_N}},$$

with some constant $c>0$ when $N$ is large enough. Together with (F.27), we also have

$$\|\boldsymbol{\Delta}\|_2\leq c\sqrt{\frac{s_{\widetilde{\boldsymbol{\alpha}}}\log(d)}{M\gamma_N}},$$

with some constant $c>0$.

Lastly, combining the results in Lemmas F.3, F.4, F.5, F.8, F.10, Theorem 4.1 and under Assumptions 4, we have the event $\mathcal{E}_p\cap\mathcal{E}_\pi\cap\mathcal{E}_\zeta\cap\widetilde{\mathcal{B}}_1\cap\widetilde{\mathcal{B}}_2\cap\widetilde{\mathcal{E}}_{\widehat{\boldsymbol{\beta}}}\cap\widetilde{\mathcal{A}}_1\cap\widetilde{\mathcal{A}}_3\cap\mathcal{E}_1\cap\mathcal{E}_2$ occurs with probability at least $1 - 21\exp(-t) - 2c_1\exp(-c_2 M\gamma_N) - 2/t - o(1)$. Combining Cases 1 and 2 above, we conclude that

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_1 = O_p\left((s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\sqrt{\frac{\log(d)}{M\gamma_N}} + \zeta_N^2\sqrt{\frac{M\gamma_N}{\log(d)}}\right),$$

$$\|\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^*\|_2 = O_p\left(\sqrt{\frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\log(d)}{M\gamma_N}} + \zeta_N\right).$$

■

## APPENDIX G: PROOF OF THE PROPERTIES OF THE SEMI-PARAMETRIC BRSS ESTIMATOR

Recall that $M = N/2$. For each $k \in \{1, 2\}$ and $k' = 3 - k$, we have

$$\widehat{\theta}^{(k)}_{1,\text{SP-BRSS}} = M^{-1}\sum_{i\in\mathcal{I}_k}\left\{\mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(k')} + \frac{\Gamma_i}{\widehat{p}^{(k)}_{N,1}(\mathbf{X}_i)\widehat{\pi}^{(k)}(\mathbf{X}_i)}(Y_i - \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(k')})\right\},$$

where $\widehat{p}^{(k)}_{N,1}(\mathbf{X}_i) = g(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}^{(k)}_{p,1} + \log(\widehat{p}^{(k)}_{N,1}))$. In this section, we consider the case that $k = 1$, the case for $k = 2$ will follow analogously. Here, we have

$$\widehat{\theta}^{(1)}_{1,\text{SP-BRSS}} - \theta_1 = M^{-1}\sum_{i\in\mathcal{I}_1}\widetilde{\psi}^*_{N,1}(\mathbf{Z}_i) + \Delta_1 + \Delta_2 + \Delta_3, \tag{G.1}$$

where

$$\widetilde{\psi}^*_{N,1}(\mathbf{Z}_i) := \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^* + \frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}(Y_i - \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^*) - \theta_1,$$

$$\Delta_1 := -M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)} - \frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*),$$

$$\Delta_2 := M^{-1}\sum_{i\in\mathcal{I}_1}\left\{1 - \frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*), \tag{G.2}$$

$$\Delta_3 := M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)} - \frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}\varepsilon_i,$$

with $\widetilde{\gamma}^{(1)}_N(\cdot) := \widehat{\pi}^{(1)}(\cdot)\widehat{p}^{(1)}_{N,1}(\cdot)$ and $\widetilde{\gamma}^*_N(\cdot) := \pi^*(\cdot)p^*_{N,1}(\cdot)$. Based on the constructions of $\widehat{\boldsymbol{\beta}}^{(1)}_{p,1}$ and $\widetilde{\boldsymbol{\alpha}}^{(2)}$ and by the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\left\|M^{-1}\sum_{i\in\mathcal{I}_1}\left\{1 - \frac{\Gamma_i}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}\right\}\mathbf{X}_i\right\|_\infty \le \lambda_{\boldsymbol{\beta}}, \tag{G.3}$$

$$\left\|M^{-1}\sum_{i\in\mathcal{I}_2}\frac{\Gamma_i\exp(-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(2)}_{p,1})}{\widehat{p}^{(2)}_{N,1}\widehat{\pi}^{(2)}(\mathbf{X}_i)}(Y_i - \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(2)})\mathbf{X}_i\right\|_\infty \le \lambda_{\boldsymbol{\alpha}}. \tag{G.4}$$

**G.1. The regular and asymptotically linear (RAL) expansion for $\widehat{\theta}_{1,\text{SP-BRSS}}$ under a correctly specified OR model.**

LEMMA G.1. *Let $m(\cdot)=\widetilde{m}^*(\cdot)$. Let Assumptions 1, 2, 3, 4, 5, and 6 hold. Suppose that $s_{\widetilde{\boldsymbol{\alpha}}}=o(\sqrt{N\gamma_N}/\log(d))$, $s_{p,1}=o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$, $N\gamma_N\gg\log(N)\log(d)$, and $\zeta_N=o(\{\log(d)/(M\gamma_N)\}^{1/4})$. Then, with some $\lambda_{\boldsymbol{\beta}}\asymp\lambda_{\boldsymbol{\alpha}}\asymp\sqrt{\log(d)/(N\gamma_N)}$, as $N,d\to\infty$,*

$$\widehat{\theta}_{\mathit{1,SP\text{-}BRSS}}-\theta_1=N^{-1}\sum_{i\in\mathcal{I}}\widetilde{\psi}^*_{N,1}(\mathbf{Z}_i)+o_p\left((N\gamma_N)^{-1/2}\right).$$

PROOF OF LEMMA G.1. We first note that $p_{N,1}\asymp\gamma_N$ through Lemma F.4. By Theorems 4.1 and 4.2, as $N,d\to\infty$,

$$\|\widehat{\boldsymbol{\beta}}^{(1)}_{p,1}-\boldsymbol{\beta}^*_{p,1}\|_1=O_p\left(s_{p,1}\sqrt{\frac{\log(d)}{N\gamma_N}}+\zeta_N^2\sqrt{\frac{N\gamma_N}{\log(d)}}\right),\tag{G.5}$$

$$\|\widehat{\boldsymbol{\beta}}^{(1)}_{p,1}-\boldsymbol{\beta}^*_{p,1}\|_2=O_p\left(\sqrt{\frac{s_{p,1}\log(d)}{N\gamma_N}}+\zeta_N\right),\tag{G.6}$$

$$\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_1=O_p\left(s_{\widetilde{\boldsymbol{\alpha}}}\sqrt{\frac{\log(d)}{N\gamma_N}}\right),\quad \|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_2=O_p\left(\sqrt{\frac{s_{\widetilde{\boldsymbol{\alpha}}}\log(d)}{N\gamma_N}}\right).\tag{G.7}$$

Based on the representation (G.1), it suffices to show that

$$\Delta_1+\Delta_2+\Delta_3=o_p\left((N\gamma_N)^{-1/2}\right).$$

By construction, note that both $\widehat{\pi}^{(1)}(\cdot)$ and $\widehat{p}^{(1)}_{N,1}(\cdot)$ are constructed based on training samples $\mathcal{S}:=(R_i,T_i,\mathbf{X}_i)_{i\in\mathcal{D}_N^{(1)}}$. Hence, under Assumption 1,

$$\mathbb{E}_{\mathcal{D}_N^{(1)}}(\Delta_3\mid\mathcal{S})=\mathbb{E}_{\mathcal{D}_N^{(1)}}\left[M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}-\frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}\varepsilon_i\mid\mathcal{S}\right]=0.$$

It follows that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_N^{(1)}}\left(\Delta_3^2\mid\mathcal{S}\right)&=\mathbb{E}_{\mathcal{D}_N^{(1)}}\left(\left[M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}-\frac{\Gamma_i}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}\varepsilon_i\right]^2\mid\mathcal{S}\right)\\&=M^{-2}\mathbb{E}(\varepsilon^2|\mathbf{X})\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\frac{1}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}-\frac{1}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}^2\\&\le cM^{-2}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\frac{1}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}-\frac{1}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}^2\end{aligned}$$

with some constant $c>0$ under Assumption 6. On the event $\mathcal{E}_\pi$ and under Assumptions 4 and 5, we have $1/\widehat{\pi}^{(1)}(\mathbf{X}_i)\le c^{-1}$, $1/\pi^*(\mathbf{X}_i)\ge c^{-1}$, and $1/p^*_{N,1}(\mathbf{X}_i)=1+p_{N,1}^{-1}\exp(-\mathbf{X}_i^T\boldsymbol{\beta}^*_{p,1})\le p_{N,1}^{-1}\{1+\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}$ for all $i\in\mathcal{I}_1$ almost surely. Hence,

$$\begin{aligned}&\left\{\frac{1}{\widetilde{\gamma}^{(1)}_N(\mathbf{X}_i)}-\frac{1}{\widetilde{\gamma}^*_N(\mathbf{X}_i)}\right\}^2\\&\quad\le\frac{2}{\{\widehat{\pi}^{(1)}(\mathbf{X}_i)\}^2}\left\{\frac{1}{\widehat{p}^{(1)}_{N,1}(\mathbf{X}_i)}-\frac{1}{p^*_{N,1}(\mathbf{X}_i)}\right\}^2+\frac{2}{\{p^*_{N,1}(\mathbf{X}_i)\}^2}\left\{\frac{1}{\widehat{\pi}^{(1)}(\mathbf{X}_i)}-\frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2\end{aligned}$$

$$\leq 2c^{-2}\left\{\frac{1}{\widehat{p}_{N,1}^{(1)}(\mathbf{X}_i)}-\frac{1}{p_{N,1}^{*}(\mathbf{X}_i)}\right\}^2+\frac{2\{1+\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}^2}{c^4p_{N,1}^2}\left\{\widehat{\pi}^{(1)}(\mathbf{X}_i)-\pi^*(\mathbf{X}_i)\right\}^2.$$

Let $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)}:=\widehat{\boldsymbol{\beta}}_{p,1}^{(j)}-\boldsymbol{\beta}_{p,1}^{*}+\log(\Delta_{p,1}^{(j)})\mathbf{e}_1$ and $\Delta_{p,1}^{(j)}:=\widehat{p}_{N,1}^{(j)}/p_{N,1}$ for each $j\in\{1,2\}$. By Lemma F.3 and the fact that $1-u^{-1}\leq\log(u)\leq u-1$ for all $u>0$, we have $\widehat{p}_{N,1}^{(j)}=O_p(\gamma_N)$ and

$$\{\log(\Delta_{p,1}^{(j)})\}^2\leq(\Delta_{p,1}^{(j)}-1)^2+(1-1/\Delta_{p,1}^{(j)})^2$$

$$=\left(\frac{\widehat{p}_{N,1}^{(j)}-p_{N,1}}{p_{N,1}}\right)^2+\left(\frac{\widehat{p}_{N,1}^{(j)}-p_{N,1}}{\widehat{p}_{N,1}^{(j)}}\right)^2=O_p\left((N\gamma_N)^{-1}\right). \tag{G.8}$$

Together with (G.5) and (G.6), we have

$$\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)}\|_1=O_p\left(s_{p,1}\sqrt{\frac{\log(d)}{N\gamma_N}}+\zeta_N^2\sqrt{\frac{N\gamma_N}{\log(d)}}\right),\ \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)}\|_2=O_p\left(\sqrt{\frac{s_{p,1}\log(d)}{N\gamma_N}}+\zeta_N\right). \tag{G.9}$$

By Taylor's theorem, for each $i\in\mathcal{I}_1$, with some $v_i\in(0,1)$,

$$\begin{aligned}
\left\{\frac{1}{\widehat{p}_{N,1}^{(1)}(\mathbf{X}_i)}-\frac{1}{p_{N,1}^{*}(\mathbf{X}_i)}\right\}^2&=p_{N,1}^{-2}\exp(-2\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^{*})\left\{\exp(-\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})-1\right\}^2\\
&=p_{N,1}^{-2}\exp(-2\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^{*})\exp(-2v_i\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\\
&=p_{N,1}^{-2}\exp\{-2(1-v_i)\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^{*}-2v_i\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}\}(\Delta_{p,1}^{(1)})^{-2v_i}(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\\
&\leq p_{N,1}^{-2}\exp(2\|\mathbf{X}_i\|_\infty\|\boldsymbol{\beta}_{p,1}^{*}\|_1+2\|\mathbf{X}_i\|_\infty\|\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}\|_1)(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\max\{1,(\Delta_{p,1}^{(1)})^{-2}\}\\
&\leq p_{N,1}^{-2}\exp(18C_{\mathbf{X}}C_{\boldsymbol{\beta}})(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\max\{1,(\Delta_{p,1}^{(1)})^{-2}\},
\end{aligned}$$

conditional on the event that $\|\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}\|_1\leq 8C_{\boldsymbol{\beta}}$, which occurs with probability approaching one. Further conditional on the event $\mathcal{E}_\zeta$, we have

$$\begin{aligned}
&M^{-2}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}_N^{(1)}(\mathbf{X}_i)}-\frac{\Gamma_i}{\widetilde{\gamma}_N^{*}(\mathbf{X}_i)}\right\}^2\\
&\qquad\leq c\max\{1,(\Delta_{p,1}^{(1)})^{-2}\}(M\gamma_N)^{-2}\sum_{i\in\mathcal{I}_1}\Gamma_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2+c(M\gamma_N)^{-1}\zeta_N^2,
\end{aligned} \tag{G.10}$$

with some constant $c>0$. By Lemma F.3, $\max\{1,(\Delta_{p,1}^{(1)})^{-2}\}=O_p(1)$. By part (b) of Lemma E.6 and (G.9),

$$\begin{aligned}
&(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\\
&\qquad=O_p\left(\frac{s_{p,1}^2\log^2(d)\log(N)}{(N\gamma_N)^2}+\frac{s_{p,1}\log(d)}{N\gamma_N}+\zeta_N^4\log(N)+\zeta_N^2\right)=o_p(1),
\end{aligned} \tag{G.11}$$

as long as $s_{p,1} = o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$ and $\zeta_N = o(\log^{-1/4}(N))$. Together with (G.10), we have

$$
\begin{aligned}
&(M\gamma_N)M^{-2}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}_N^{(1)}(\mathbf{X}_i)}-\frac{\Gamma_i}{\widetilde{\gamma}_N^{*}(\mathbf{X}_i)}\right\}^2\\
&=O_p\left(\frac{s_{p,1}^2\log^2(d)\log(N)}{(N\gamma_N)^2}+\frac{s_{p,1}\log(d)}{N\gamma_N}+\zeta_N^4\log(N)+\zeta_N^2\right)=o_p(1),
\end{aligned}
\tag{G.12}
$$

and hence $\mathbb{E}_{\mathcal{D}_N^{(1)}}(\Delta_3^2 \mid \mathcal{S}) = o_p\left((N\gamma_N)^{-1}\right)$. By Lemma E.1,

$$
\Delta_3 = o_p\left((N\gamma_N)^{-1/2}\right).
$$

Now, as for $\Delta_1$ and $\Delta_2$, we have

$$
\begin{aligned}
\Delta_1+\Delta_2 &= M^{-1}\sum_{i\in\mathcal{I}_1}\left\{1-\frac{\Gamma_i}{\widetilde{\gamma}_N^{(1)}(\mathbf{X}_i)}\right\}\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*)\\
&\le \left\|M^{-1}\sum_{i\in\mathcal{I}_1}\left\{1-\frac{\Gamma_i}{\widetilde{\gamma}_N^{(1)}(\mathbf{X}_i)}\right\}\mathbf{X}_i\right\|_\infty \|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_1\\
&\overset{(i)}{\le} \lambda_{\boldsymbol{\beta}}\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_1 \overset{(ii)}{=} O_p\left(\frac{s_{\widetilde{\boldsymbol{\alpha}}}\log(d)}{N\gamma_N}\right)\overset{(iii)}{=} o_p\left((N\gamma_N)^{-1}\right),
\end{aligned}
$$

where (i) holds by (G.3); (ii) holds by (G.7) and $\lambda_{\boldsymbol{\beta}} \asymp \sqrt{\log(d)/(N\gamma_N)}$; (iii) holds if $s_{\widetilde{\boldsymbol{\alpha}}} = o(\sqrt{N\gamma_N}/\log(d))$. ■

### G.2. The RAL expansion for $\widehat{\theta}_{1,\text{SP-BRSS}}$ under a possibly misspecified OR model.

LEMMA G.2. *Let Assumptions 1, 2, 3, 4, 5, and 6 hold. Let* $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1} = o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$, $s_{\widetilde{\boldsymbol{\alpha}}}\sqrt{s_{p,1}} = o(N\gamma_N/\log^{3/2}(d))$, $N\gamma_N \gg \log(d\vee N)\log(N)$ *and* $\zeta_N = o((N\gamma_N)^{-1/2})$. *Then, with some* $\lambda_{\boldsymbol{\beta}} \asymp \lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log(d)/(N\gamma_N)}$, *as* $N, d \to \infty$,

$$
\widehat{\theta}_{\mathit{1,SP\text{-}BRSS}} - \theta_1 = N^{-1}\sum_{i\in\mathcal{I}}\psi_{N,1}^*(\mathbf{Z}_i) + o_p\left((N\gamma_N)^{-1/2}\right).
$$

PROOF OF LEMMA G.2. Define $\widetilde{\varepsilon}_i := Y_i(1) - \mathbf{X}_i^T\widetilde{\boldsymbol{\alpha}}^{(2)}$ and let $\widetilde{\varepsilon}$ be an independent copy of $\widetilde{\varepsilon}_i$. Now we consider the following representation:

$$
\widehat{\theta}_{1,\text{SP-BRSS}}^{(1)} - \theta_1 = M^{-1}\sum_{i\in\mathcal{I}_1}\widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) + \Delta_2 + \Delta_4 + \Delta_5,
\tag{G.13}
$$

where $\Delta_2$ is defined as in (G.2) and

$$
\Delta_4 := M^{-1}\sum_{i\in\mathcal{I}_1}\frac{1}{p_{N,1}^*(\mathbf{X}_i)}\left\{\frac{\Gamma_i}{\widehat{\pi}^{(1)}(\mathbf{X}_i)}-\frac{\Gamma_i}{\pi^*(\mathbf{X}_i)}\right\}\widetilde{\varepsilon}_i,
$$

$$
\Delta_5 := M^{-1}\sum_{i\in\mathcal{I}_1}\frac{1}{\widehat{\pi}^{(1)}(\mathbf{X}_i)}\left\{\frac{\Gamma_i}{\widehat{p}_{N,1}^{(1)}(\mathbf{X}_i)}-\frac{\Gamma_i}{p_{N,1}^*(\mathbf{X}_i)}\right\}\widetilde{\varepsilon}_i.
$$

It suffices to show that

$$\Delta_2 + \Delta_4 + \Delta_5 = o_p\left((N\gamma_N)^{-1/2}\right).$$

By Theorems 4.1 and 4.3, for each $j \in \{1,2\}$, (G.9) hold, and

$$\|\widetilde{\boldsymbol{\alpha}}^{(j)} - \widetilde{\boldsymbol{\alpha}}^*\|_1 = O_p\left((s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\sqrt{\frac{\log(d)}{N\gamma_N}} + \zeta_N^2\sqrt{\frac{N\gamma_N}{\log(d)}}\right), \tag{G.14}$$

$$\|\widetilde{\boldsymbol{\alpha}}^{(j)} - \widetilde{\boldsymbol{\alpha}}^*\|_2 = O_p\left(\sqrt{\frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\log(d)}{N\gamma_N}} + \zeta_N\right). \tag{G.15}$$

For the term $\Delta_2$, since $\widetilde{\boldsymbol{\alpha}}^{(2)}$ is independent of $\mathcal{D}_N^{(1)}$, we have

$$\mathbb{E}_{\mathcal{D}_N^{(1)}}(\Delta_2) = \mathbb{E}_{\mathbf{X},\Gamma}\left[\left\{1 - \frac{\Gamma}{\widetilde{\gamma}_N^*(\mathbf{X})}\right\}\mathbf{X}\right]^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*) = 0,$$

by the construction of $\boldsymbol{\beta}_{p,1}^*$. For any function $f(\cdot)$ and constant $r>0$, denote $\|f(\cdot)\|_{\mathbb{P}_{\mathbf{X}},r} := \{\mathbb{E}_{\mathbf{X}}|f(\mathbf{X})|^r\}^{1/r}$. Then, with $r>0$ satisfying $1/r + 1/q = 1$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_N^{(1)}}(\Delta_2^2) &= M^{-1}\mathbb{E}_{\mathbf{X},\Gamma}\left(\left[\left\{1 - \frac{\Gamma}{\widetilde{\gamma}_N^*(\mathbf{X})}\right\}\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right]^2\right) \\
&\le M^{-1}\mathbb{E}_{\mathbf{X},\Gamma}\left(\left[1 + \frac{\Gamma}{\{\widetilde{\gamma}_N^*(\mathbf{X})\}^2}\right]\left\{\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2\right) \\
&= M^{-1}\mathbb{E}_{\mathbf{X},\Gamma}\left(\left[1 + \frac{\{1 + \gamma_N^{-2}\exp(-2\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*)\}\gamma_N(\mathbf{X})}{\{\pi^*(\mathbf{X})\}^2}\right]\left\{\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2\right) \\
&\le M^{-1}\left\|1 + \frac{\{1 + \gamma_N^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}\gamma_N(\mathbf{X})}{c^2}\right\|_{\mathbb{P},q}\left\|\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\|_{\mathbb{P}_{\mathbf{X}},2r}^2 \\
&= O_p\left((N\gamma_N)^{-1}\left\{\frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\log(d)}{N\gamma_N} + \zeta_N^2\right\}\right),
\end{aligned}$$

under Assumptions 2 and 3 together with (G.15). By Lemma E.1,

$$\Delta_2 = O_p\left((N\gamma_N)^{-1/2}\sqrt{\frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\log(d)}{N\gamma_N}} + \zeta_N\right) = o_p\left((N\gamma_N)^{-1/2}\right), \tag{G.16}$$

as long as $s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1} = o(N\gamma_N/\log(d))$ and $\zeta_N = o(1)$.

Now we consider the term $\Delta_4$. Condition on the event $\mathcal{E}_\pi \cap \mathcal{E}_\zeta$. By the Cauchy-Schwarz inequality, under Assumptions 4 and 5, we have

$$\begin{aligned}
\Delta_4^2 &\le M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\widetilde{\varepsilon}_i^2}{\{p_{N,1}^*(\mathbf{X}_i)\}^2}M^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\frac{1}{\widehat{\pi}^{(1)}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}^2 \\
&\le c\gamma_N^{-2}M^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\widetilde{\varepsilon}_i^2 M^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\widehat{\pi}^{(1)}(\mathbf{X}_i) - \pi^*(\mathbf{X}_i)\right\}^2 \\
&\le c\zeta_N^2(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\widetilde{\varepsilon}_i^2,
\end{aligned}$$

with some constant $c>0$. Denote $\mathbb{P}_{\mathbf{X},\Gamma,Y}$ as the joint distribution of $(\mathbf{X},\Gamma,Y)$ and let $\mathbb{E}_{\mathbf{X},\Gamma,Y}(\cdot)$ be the corresponding expectation. Note that

$$\mathbb{E}_{\mathcal{D}_N^{(1)}}\left\{(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\widetilde{\varepsilon}_i^2\right\}=\gamma_N^{-1}\mathbb{E}_{\mathbf{X},\Gamma,Y}(\Gamma\widetilde{\varepsilon}^2)$$
$$=\gamma_N^{-1}\mathbb{E}_{\mathbf{X},\Gamma,Y}\left[\gamma_N(\mathbf{X})\left\{\varepsilon-\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*)\right\}^2\right]=O_p(1+\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_2^2)=O_p(1),$$

under Assumptions 1, 2, 3, and 6, together with the fact that $\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_2=o_p(1)$. By Lemma E.1, $(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\widetilde{\varepsilon}_i^2=O_p(1)$. Hence, as long as $\zeta_N=o((N\gamma_N)^{-1/2})$,

$$\Delta_4=O_p(\zeta_N)=o_p\left((N\gamma_N)^{-1/2}\right).$$

Lastly, we control the term $\Delta_5$. By Taylor's theorem, there exists some $v\in(0,1)$ such that

$$\Delta_5=M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(1)})\{\exp(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})-1\}\widetilde{\varepsilon}_i}{\widehat{p}_{N,1}^{(1)}\widehat{\pi}^{(1)}(\mathbf{X}_i)}=\Delta_{5,1}+\Delta_{5,2},$$

where

$$\Delta_{5,1}=M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(1)})\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\widetilde{\varepsilon}_i}{\widehat{p}_{N,1}^{(1)}\widehat{\pi}^{(1)}(\mathbf{X}_i)},$$
$$\Delta_{5,2}=M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}+v\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2\widetilde{\varepsilon}_i}{\widehat{p}_{N,1}^{(1)}\widehat{\pi}^{(1)}(\mathbf{X}_i)},$$

with $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}=\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}-\boldsymbol{\beta}_{p,1}^*+\log(\Delta_{p,1}^{(1)})\mathbf{e}_1$ and $\Delta_{p,1}^{(1)}=\widehat{p}_{N,1}^{(1)}/p_{N,1}$. Define $\mathcal{E}_3:=\{0.46p_{N,1}\le\widehat{p}_{N,1}^{(j)}\le1.54p_{N,1},\|\widehat{\boldsymbol{\beta}}_{p,1}^{(j)}\|_1\le8C_{\boldsymbol{\beta}},\forall j\in\{1,2\},\sup_{i\in\mathcal{I}}|\mathbf{X}_i|\le C_{\mathbf{X}}\}$. By Lemmas F.3 and 4.1, $\mathcal{E}_3$ occurs with probability approaching one. Further condition on the event $\mathcal{E}_3$. Then, $\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}+v\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})\le\exp(9C_{\mathbf{X}}C_{\boldsymbol{\beta}})$. It follows that with some constant $c>0$,

$$|\Delta_{5,2}|\le c(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2|\widetilde{\varepsilon}_i|. \tag{G.17}$$

Define $\mathcal{E}_{\widetilde{\boldsymbol{\alpha}}}:=\{\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_2\le1\}$. Then, $\mathbb{P}_{\mathcal{D}_N^{(2)}}(\mathcal{E}_{\widetilde{\boldsymbol{\alpha}}})=1-o(1)$. On the event $\mathcal{E}_{\widetilde{\boldsymbol{\alpha}}}$ and conditional on $\mathcal{D}_N^{(2)}$, we have

$$\|\widetilde{\varepsilon}\|_{\psi_2}=\|\varepsilon-\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*)\|_{\psi_2}\le\|\varepsilon\|_{\psi_2}+\|\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*)\|_2$$
$$\le\sigma_\varepsilon+\sigma\|\widetilde{\boldsymbol{\alpha}}^{(2)}-\widetilde{\boldsymbol{\alpha}}^*\|_2\le\sigma_\varepsilon+\sigma, \tag{G.18}$$

Note that under Assumption 1,

$$\mathbb{E}_{\mathbf{X},\Gamma,Y}(|\widetilde{\varepsilon}|\mid\mathbf{X},\Gamma=1)=\mathbb{E}_{\mathbf{X},\Gamma,Y}(|Y-\mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^{(2)}|\mid\mathbf{X},R=1,T=1)$$
$$\overset{(i)}{=}\mathbb{E}_{\mathbf{X},T,Y}\{|Y(1)-\mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^{(2)}|\mid\mathbf{X},T=1\}\overset{(ii)}{=}\mathbb{E}_{\mathbf{X},Y(1)}(|\widetilde{\varepsilon}|\mid\mathbf{X}),$$

where (i) holds since $R\perp\!\!\!\perp Y\mid(T,\mathbf{X})$ and $Y=Y(T)$; (ii) holds since $T\perp\!\!\!\perp Y(1)\mid\mathbf{X}$. In the above, $\mathbb{E}_{\mathbf{X},\Gamma,Y}(\cdot)$, $\mathbb{E}_{\mathbf{X},T,Y}$, and $\mathbb{E}_{\mathbf{X},Y(1)}$ denote the expectations with respect to the joint distribution of $(\mathbf{X},\Gamma,Y)$, $(\mathbf{X},T,Y)$, and $(\mathbf{X},Y(1))$, respectively. Choose some $s\asymp\sqrt{N\gamma_N/\{\log(d)\log(N)\}}$, then $\sqrt{s\log(d)/(N\gamma_N)}+s^2\sqrt{\log(N)\log(d)/(N\gamma_N)}\asymp$

$[\log(d)/\{N\gamma_N \log(N)\}]^{1/4} + 1/\sqrt{\log(d)\log(N)} = o(1)$ since $N\gamma_N \gg \log(d)\log(N)$. By part (e) of Lemma E.6 and (G.9), we have

$$
\begin{aligned}
\Delta_{5,2} &= O_p\left(\frac{\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_1^2}{s}\left\{\sqrt{\frac{s\log(d)}{N\gamma_N}} + \frac{s^2\sqrt{\log(N)\log(d)}}{N\gamma_N}\right\} + \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_2^2\{1+o(1)\}\right) \\
&= O_p\left(\left(\frac{s_{p,1}^2\log(d)}{N\gamma_N} + \frac{\zeta_N^4 N\gamma_N}{\log(d)}\right)\left\{\frac{\log^{3/4}(d)\log^{1/4}(N)}{(N\gamma_N)^{3/4}} + \frac{1}{\sqrt{N\gamma_N}}\right\}\right) \\
&\quad + O_p\left(\left\{\frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2\right\}\{1+o(1)\}\right) \\
&= O_p\left(\frac{s_{p,1}^2\log^{7/4}(d)\log^{1/4}(N)}{(N\gamma_N)^{7/4}} + \frac{s_{p,1}^2\log(d)}{(N\gamma_N)^{3/2}} + \zeta_N^4 N\gamma_N + \frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2\right) \\
&= o_p\left((N\gamma_N)^{-1/2}\right), \qquad \text{(G.19)}
\end{aligned}
$$

since $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $\zeta_N = o((N\gamma_N)^{-1/2})$, and $N\gamma_N \gg \log(d\vee N)\log(N)$.

Now, we control the term $\Delta_{5,1}$. Define $\boldsymbol{\Delta}_{\boldsymbol{\alpha}}^{(2)} := \widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*$. We consider the following representation:

$$
\begin{aligned}
\Delta_{5,1} =& \Delta_{5,1,1} + \sum_{k=2}^{5}\Delta_{5,1,k}^{(1)} - \sum_{k=2}^{5}\Delta_{5,1,k}^{(2)}, \ \text{ where} \\
\Delta_{5,1,1} :=& M^{-1}\sum_{i\in\mathcal{I}_2}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(2)})\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\widetilde{\varepsilon}_i}{\widehat{p}_{N,1}^{(2)}\widehat{\pi}^{(2)}(\mathbf{X}_i)}, \ \text{ and for each } j\in\{1,2\}, \\
\Delta_{5,1,2}^{(j)} :=& M^{-1}\sum_{i\in\mathcal{I}_j}\frac{\Gamma_i\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\widetilde{\varepsilon}_i}{\widehat{\pi}^{(j)}(\mathbf{X}_i)}\left\{\frac{\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(j)})}{\widehat{p}_{N,1}^{(j)}} - \frac{\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)}{p_{N,1}}\right\}, \\
\Delta_{5,1,3}^{(j)} :=& M^{-1}\sum_{i\in\mathcal{I}_j}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\widetilde{\varepsilon}_i}{p_{N,1}}\left\{\frac{1}{\widehat{\pi}^{(j)}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}, \\
\Delta_{5,1,4}^{(j)} :=& M^{-1}\sum_{i\in\mathcal{I}_j}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\varepsilon_i}{p_{N,1}\pi^*(\mathbf{X}_i)}, \\
\Delta_{5,1,5}^{(j)} :=& M^{-1}\sum_{i\in\mathcal{I}_j}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\alpha}}^{(2)}}{p_{N,1}\pi^*(\mathbf{X}_i)} \\
& - \mathbb{E}_{\mathbf{X},\Gamma}\left\{\frac{\Gamma\exp(-\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\mathbf{X}^T\boldsymbol{\Delta}_{\boldsymbol{\alpha}}^{(2)}}{p_{N,1}\pi^*(\mathbf{X})}\right\}.
\end{aligned}
$$

By (G.4), (G.9), and $\lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log(d)/(N\gamma_N)}$, we have

$$
\Delta_{5,1,1} \le \left\|M^{-1}\sum_{i\in\mathcal{I}_2}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{p,1}^{(2)})}{\widehat{p}_{N,1}^{(2)}\widehat{\pi}^{(2)}(\mathbf{X}_i)}\widetilde{\varepsilon}_i\mathbf{X}_i\right\|_\infty \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_1
$$

$$= O_p\left(\frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2\right) = o_p\left((N\gamma_N)^{-1/2}\right),$$

since $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$ and $\zeta_N = o((N\gamma_N)^{-1/2})$. For each $j \in \{1,2\}$, by Taylor's theorem, with some $v_j \in (0,1)$,

$$\Delta_{5,1,2}^{(j)} = -M^{-1}\sum_{i\in\mathcal{I}_j}\frac{\Gamma_i\exp[-\mathbf{X}_i^T\{v_j\widehat{\boldsymbol{\beta}}_{p,1}^{(j)} + (1-v_j)\boldsymbol{\beta}_{p,1}^*\}]\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)}\widetilde{\varepsilon}_i}{(\widehat{p}_{N,1}^{(j)})^{v_j}p_{N,1}^{1-v_j}\widehat{\pi}^{(j)}(\mathbf{X}_i)}.$$

On the event $\mathcal{E}_3$, with some constant $c>0$

$$\begin{aligned}
|\Delta_{5,1,2}^{(j)}| &\leq c(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i|\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)}\widetilde{\varepsilon}_i| \\
&\leq 2^{-1}c(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2|\widetilde{\varepsilon}_i| + 2^{-1}c(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(j)})^2|\widetilde{\varepsilon}_i| \\
&= o_p\left((N\gamma_N)^{-1/2}\right),
\end{aligned}$$

using part (e) of Lemma E.6 as in (G.19). On the event $\mathcal{E}_3$, we have $|\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}| \leq |\log(\Delta_{p,1}^{(1)})| + \|\mathbf{X}_i\|_\infty(\|\widehat{\boldsymbol{\beta}}_{p,1}^{(1)}\| + \|\boldsymbol{\beta}_{p,1}^*\|) \leq \log(0.46^{-1}\vee 1.54) + 9C_{\mathbf{X}}C_{\boldsymbol{\beta}}$ for all $i \in \mathcal{I}$. Hence, by the Cauchy-Schwarz inequality, with some constant $c>0$,

$$\begin{aligned}
|\Delta_{5,1,3}^{(j)}|^2 &\leq c(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\widetilde{\varepsilon}_i^2(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\left\{\widehat{\pi}^{(j)}(\mathbf{X}_i) - \pi^*(\mathbf{X}_i)\right\}^2 \\
&\leq c\zeta_N^2(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\widetilde{\varepsilon}_i^2 \\
&\leq 2c\zeta_N^2(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\varepsilon_i^2 + 2c\zeta_N^2(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\}^2.
\end{aligned}$$

Since $(M\gamma_N)^{-1}\mathbb{E}(\sum_{i\in\mathcal{I}_j}\Gamma_i\varepsilon_i^2) = \gamma_N^{-1}\mathbb{E}\{\gamma_N(\mathbf{X})\varepsilon^2\} = O(1)$, by Lemma E.1,

$$(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\varepsilon_i^2 = O_p(1).$$

Additionally, by part (b) of Lemma E.6, together with (G.14) and (G.15),

$$\begin{aligned}
&(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_j}\Gamma_i\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\}^2 \\
&= O_p\left(\frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})\log(d)}{N\gamma_N} + \frac{(s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1})^2\log^2(d)\log(N)}{(N\gamma_N)^2} + \zeta_N^2 + \zeta_N^4\log(N)\right) \\
&= O_p(1),
\end{aligned}$$

since $s_{\widetilde{\boldsymbol{\alpha}}} = O(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$, $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $\zeta_N = o(\log^{-1/4}(N))$, and $N\gamma_N \gg \log(d)\log(N)$. Hence,

$$\Delta_{5,1,3} = O_p(\zeta_N) = o_p\left((N\gamma_N)^{-1/2}\right).$$

Let $\mathbf{V}_i = \Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\mathbf{X}_i\varepsilon_i/\pi^*(\mathbf{X}_i)$. By the construction of $\boldsymbol{\beta}_{p,1}^*$, we have $\mathbb{E}(\mathbf{V}_i) = \mathbf{0}$. In addition, $\sup_{1\leq j\leq d}\|\mathbf{V}_i(j)\|_{\psi_2} \leq c$ and $\sup_{1\leq j\leq d}\mathbb{E}\{\mathbf{V}_i(j)\}^2 \leq c\gamma_N$ with some constant

$c > 0$. By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022),

$$\left\| (Mp_{N,1})^{-1} \sum_{i\in\mathcal{I}_j} \mathbf{V}_i \right\|_\infty = O_p\left( \sqrt{\frac{\log(d)}{N\gamma_N}} + \frac{\sqrt{\log(N)\log(d)}}{N\gamma_N} \right) = O_p\left( \sqrt{\frac{\log(d)}{N\gamma_N}} \right),$$

since $N\gamma_N \gg \log(N)$. Together with (G.9), we have

$$\begin{aligned} |\Delta_{5,1,4}| &\le \left\| (Mp_{N,1})^{-1} \sum_{i\in\mathcal{I}_j} \mathbf{V}_i \right\|_\infty \|\mathbf{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_\infty \\ &= O_p\left( \frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2 \right) = o_p\left( (N\gamma_N)^{-1/2} \right), \end{aligned}$$

since $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$ and $\zeta_N = o((N\gamma_N)^{-1/2})$. Lastly, denote $\bar{s} = s_{p,1} + s_{\widetilde{\boldsymbol{\alpha}}}$. By part (c) of Lemma E.6, (G.9), (G.14), and (G.15) with $s \asymp \bar{s} + \zeta_N^2 N\gamma_N/\log(d)$,

$$\begin{aligned} \Delta_{5,1,5}^{(j)} &= \left\{ \sqrt{\frac{s\log(d)}{N\gamma_N}} + \frac{s\log(d)\log(N)}{N\gamma_N} \right\} \\ &\quad \cdot O_p\left( \|\mathbf{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_2 \|\mathbf{\Delta}_{\boldsymbol{\alpha}}^{(2)}\|_2 \left( \frac{\|\mathbf{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_1^2}{s\|\mathbf{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_2^2} + \frac{\|\mathbf{\Delta}_{\boldsymbol{\alpha}}^{(2)}\|_1^2}{s\|\mathbf{\Delta}_{\boldsymbol{\alpha}}^{(2)}\|_2^2} + 1 \right) \right) \\ &= \left\{ \sqrt{\frac{\bar{s}\log(d)}{N\gamma_N}} + \zeta_N + \frac{\bar{s}\log(d)\log(N)}{N\gamma_N} + \zeta_N^2\log(N) \right\} \\ &\quad \cdot O_p\left( \left\{ \sqrt{\frac{s_{p,1}\log(d)}{N\gamma_N}} + \zeta_N \right\} \left\{ \sqrt{\frac{\bar{s}\log(d)}{N\gamma_N}} + \zeta_N \right\} \right) \\ &= o_p\left( (N\gamma_N)^{-1/2} \right), \end{aligned}$$

since $N\gamma_N \gg \log(d)\log(N)$, $\bar{s} = o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$, $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $s_{\widetilde{\boldsymbol{\alpha}}}\sqrt{s_{p,1}} = o(N\gamma_N/\log^{3/2}(d))$, and $\zeta_N = o((N\gamma_N)^{-1/2})$. To sum up, we conclude that

$$\Delta_5 = \Delta_{5,1,1} + \sum_{k=2}^{5} \Delta_{5,1,k}^{(1)} - \sum_{k=2}^{5} \Delta_{5,1,k}^{(2)} + \Delta_{5,2} = o_p\left( (N\gamma_N)^{-1/2} \right).$$

∎

LEMMA G.3. *Let Assumptions 1, 2, 3, 4, 5, and 6 hold. Let* $s_{\widetilde{\boldsymbol{\alpha}}} = o(N\gamma_N/\log(d))$, $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $s_{\widetilde{\boldsymbol{\alpha}}}s_{p,1} = o(N\gamma_N/\{\log(d)\}^2)$, $\zeta_N = o(1/\sqrt{N\gamma_N})$, *and* $N\gamma_N \gg \log(d \vee N)\log(N)$. *Then, with some* $\lambda_{\boldsymbol{\beta}} \asymp \lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log(d)/(N\gamma_N)}$, *as* $N, d \to \infty$,

$$\widehat{\theta}_{1,\textit{SP-BRSS}} - \theta_1 = N^{-1}\sum_{i\in\mathcal{I}} \psi_{N,1}^*(\mathbf{Z}_i) + o_p\left( (N\gamma_N)^{-1/2} \right).$$

PROOF OF LEMMA G.3. Based on the representation (G.1), it suffices to show that

$$\Delta_1 + \Delta_2 + \Delta_3 = o_p\left( (N\gamma_N)^{-1/2} \right).$$

As in the proof of Lemma G.2, we have (G.9), (G.14), and (G.15) hold.

For the term $\Delta_2$, same as in Lemma G.2, (G.16) holds when $s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1} = o(N\gamma_N/\log(d))$ and $\zeta_N = o(1)$.

As for $\Delta_1$, we have

$$|\Delta_1| \leq \sqrt{M^{-1}\sum_{i\in\mathcal{I}_1}\left\{\frac{\Gamma_i}{\widetilde{\gamma}_N^{(1)}(\mathbf{X}_i)} - \frac{\Gamma_i}{\widetilde{\gamma}_N^{*}(\mathbf{X}_i)}\right\}^2}\sqrt{M^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2}.$$

Denote $\bar{s} := s_{\widetilde{\boldsymbol{\alpha}}} + s_{p,1}$, then

$$\mathbb{E}_{\mathcal{D}_N^{(1)}}\left[(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2\right] = \gamma_N^{-1}\mathbb{E}\left[\gamma_N(\mathbf{X})\left\{\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2\right]$$
$$\leq \gamma_N^{-1}\|\gamma_N(\mathbf{X})\|_{\mathbb{P},q}\left\|\mathbf{X}^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\|^2_{\mathbb{P}_{\mathbf{X}},2r} = O_p\left(\frac{\bar{s}\log(d)}{N\gamma_N} + \zeta_N^2\right),$$

under Assumptions 2 and 3, together with (G.15) and part (c) of Lemma E.2. By Lemma E.1,

$$(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i\left\{\mathbf{X}_i^T(\widetilde{\boldsymbol{\alpha}}^{(2)} - \widetilde{\boldsymbol{\alpha}}^*)\right\}^2 = O_p\left(\frac{\bar{s}\log(d)}{N\gamma_N} + \zeta_N^2\right). \tag{G.20}$$

Together with (G.12), it follows that

$$\Delta_1^2 = O_p\left(\left\{\frac{s_{p,1}^2\log^2(d)\log(N)}{(N\gamma_N)^2} + \frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^4\log(N) + \zeta_N^2\right\}\left\{\frac{\bar{s}\log(d)}{N\gamma_N} + \zeta_N^2\right\}\right)$$
$$= o_p\left((N\gamma_N)^{-1}\right),$$

since $s_{\widetilde{\boldsymbol{\alpha}}}s_{p,1} = o(N\gamma_N/\{\log(d)\}^2)$, $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $\zeta_N = o(1/\sqrt{N\gamma_N})$, and $N\gamma_N \gg \log(d \vee N)\log(N)$. Hence, $\boldsymbol{\Delta}_1 = O((N\gamma_N)^{-1/2})$. Lastly, we consider the term $\Delta_3$. Observe that

$$\Delta_3 = \Delta_{3,1} + \Delta_{3,2}, \text{ where}$$
$$\Delta_{3,1} := M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i}{\pi^*(\mathbf{X}_i)}\left\{\frac{1}{\widehat{p}_{N,1}^{(1)}(\mathbf{X}_i)} - \frac{1}{p_{N,1}^*(\mathbf{X}_i)}\right\}\varepsilon_i$$
$$= \sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\{\exp(-\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}) - 1\}\varepsilon_i}{p_{N,1}\pi^*(\mathbf{X}_i)},$$
$$\Delta_{3,2} := M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i}{\widehat{p}_{N,1}^{(1)}(\mathbf{X}_i)}\left\{\frac{1}{\widehat{\pi}^{(1)}(\mathbf{X}_i)} - \frac{1}{\pi^*(\mathbf{X}_i)}\right\}\varepsilon_i.$$

By Taylor's theorem, with some $v \in (0,1)$,

$$|\Delta_{3,1}| \leq \left|M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\varepsilon_i\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}}{p_{N,1}\pi^*(\mathbf{X}_i)}\right|$$
$$+ \left|M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp[-\mathbf{X}_i^T\{\boldsymbol{\beta}_{p,1}^* + v(\widehat{\boldsymbol{\beta}}_{p,1}^{(1)} - \boldsymbol{\beta}_{p,1}^*)\}]\varepsilon_i(\mathbf{X}_i^T\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)})^2}{p_{N,1}^{1-v}\widehat{p}_{N,1}^{v}\pi^*(\mathbf{X}_i)}\right|$$
$$\leq \left\|M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol{\beta}_{p,1}^*)\varepsilon_i\mathbf{X}_i}{p_{N,1}\pi^*(\mathbf{X}_i)}\right\|_\infty\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(1)}\|_1$$

$$\text{(G.21)} \qquad + \frac{\exp(9C_{\mathbf{X}}C_{\boldsymbol\beta})}{0.46cMp_{N,1}} \sum_{i\in\mathcal{I}_1} \Gamma_i|\varepsilon_i|(\mathbf{X}_i^T\boldsymbol\Delta_{\boldsymbol\beta}^{(1)})^2,$$

when $\pi^*(\mathbf{X}_i) \geq c$, $\|\widehat{\boldsymbol\beta}_{p,1}^{(1)}\|_1 \leq 8C_{\boldsymbol\beta}$, and $\widehat{p}_{N,1}^{(1)} \geq 0.46p_{N,1}$, which occurs with probability approaching one as shown in Lemma F.3, Theorem 4.1, and under Assumption 4. Since $Y = Y(T)$ and note that $R \perp\!\!\!\perp Y \mid (T, \mathbf{X})$ and $T \perp\!\!\!\perp Y(1) \mid \mathbf{X}$ under Assumption 1, we have $\mathbb{E}(|\varepsilon| \mid \mathbf{X}, \Gamma = 1) = \mathbb{E}(|Y - \mathbf{X}^T\boldsymbol\alpha^*| \mid \mathbf{X}, R = T = 1) = \mathbb{E}\{|Y(1) - \mathbf{X}^T\boldsymbol\alpha^*| \mid \mathbf{X}, T = 1\} = \mathbb{E}(|\varepsilon| \mid \mathbf{X})$. Choose some $s \asymp \sqrt{N\gamma_N/\{\log(d)\log(N)\}}$, then $\sqrt{s\log(d)/(N\gamma_N)} + s^2\sqrt{\log(N)\log(d)}/(N\gamma_N) \asymp [\log(d)/\{N\gamma_N\log(N)\}]^{1/4} + 1/\sqrt{\log(d)\log(N)} = o(1)$ since $N\gamma_N \gg \log(d)\log(N)$. Since $\varepsilon_i$ is a sub-Gaussian random variable, similarly as in (G.19), by part (e) of Lemma E.6 and (G.9), we have

$$\begin{aligned}
&(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_1}\Gamma_i(\mathbf{X}_i^T\boldsymbol\Delta_{\boldsymbol\beta}^{(1)})^2|\varepsilon_i| \\
&= O_p\left(\frac{\|\boldsymbol\Delta_{\boldsymbol\beta}^{(1)}\|_1^2}{s}\left\{\sqrt{\frac{s\log(d)}{N\gamma_N}} + \frac{s^2\sqrt{\log(N)\log(d)}}{N\gamma_N}\right\} + \|\boldsymbol\Delta_{\boldsymbol\beta}^{(1)}\|_2^2\{1+o(1)\}\right) \\
&= O_p\left(\left(\frac{s_{p,1}^2\log(d)}{N\gamma_N} + \frac{\zeta_N^4 N\gamma_N}{\log(d)}\right)\left\{\frac{\log^{3/4}(d)\log^{1/4}(N)}{(N\gamma_N)^{3/4}} + \frac{1}{\sqrt{N\gamma_N}}\right\}\right) \\
&\quad + O_p\left(\left\{\frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2\right\}\{1+o(1)\}\right) \\
&= O_p\left(\frac{s_{p,1}^2\log^{7/4}(d)\log^{1/4}(N)}{(N\gamma_N)^{7/4}} + \frac{s_{p,1}^2\log(d)}{(N\gamma_N)^{3/2}} + \zeta_N^4 N\gamma_N + \frac{s_{p,1}\log(d)}{N\gamma_N} + \zeta_N^2\right) \\
\text{(G.22)} \quad &= o_p\left((N\gamma_N)^{-1/2}\right),
\end{aligned}$$

since $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $\zeta_N = o((N\gamma_N)^{-1/2})$, and $N\gamma_N \gg \log(d \vee N)\log(N)$. Additionally, let $\mathbf{V} := \Gamma\exp(-\mathbf{X}^T\boldsymbol\beta_{p,1}^*)\varepsilon\mathbf{X}/\pi^*(\mathbf{X})$. By the construction of $\boldsymbol\beta_{p,1}^*$, we have $\mathbb{E}(\mathbf{V}) = \mathbf{0} \in \mathbb{R}^d$. Note that $\sup_{1\leq j\leq d}\|\mathbf{V}(j)\|_{\psi_2} \leq \exp(C_{\mathbf{X}}C_{\boldsymbol\beta})C_{\mathbf{X}}\sigma_\varepsilon/c$ and with $r > 1$ satisfy $1/r + 1/q = 1$,

$$\begin{aligned}
\sup_{1\leq j\leq d}\mathbb{E}\{\mathbf{V}^2(j)\} &\leq c^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol\beta})C_{\mathbf{X}}^2\mathbb{E}\{\gamma_N(\mathbf{X})\varepsilon^2\} \\
&\leq c^{-2}\exp(2C_{\mathbf{X}}C_{\boldsymbol\beta})C_{\mathbf{X}}^2\|\gamma_N(\mathbf{X})\|_{\mathbb{P},q}\|\varepsilon\|_{\mathbb{P},2r}^2 = O(\gamma_N).
\end{aligned}$$

By Theorem 3.4 of Kuchibhotla and Chakrabortty (2022),

$$\left\|M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol\beta_{p,1}^*)\varepsilon_i\mathbf{X}_i}{p_{N,1}\pi^*(\mathbf{X}_i)}\right\|_\infty = O_p\left(\sqrt{\frac{\log d}{N\gamma_N}} + \frac{\sqrt{\log(N)}\log(d)}{N\gamma_N}\right).$$

Together with (G.9), it follows that

$$\begin{aligned}
&\left\|M^{-1}\sum_{i\in\mathcal{I}_1}\frac{\Gamma_i\exp(-\mathbf{X}_i^T\boldsymbol\beta_{p,1}^*)\varepsilon_i\mathbf{X}_i}{p_{N,1}\pi^*(\mathbf{X}_i)}\right\|_\infty\|\boldsymbol\Delta_{\boldsymbol\beta}^{(1)}\|_1 \\
&\quad = O_p\left(\left\{\sqrt{\frac{\log(d)}{N\gamma_N}} + \frac{\sqrt{\log(N)}\log(d)}{N\gamma_N}\right\}\left\{s_{p,1}\sqrt{\frac{\log(d)}{N\gamma_N}} + \zeta_N^2\sqrt{\frac{N\gamma_N}{\log(d)}}\right\}\right)
\end{aligned}$$

$$= o\left((N\gamma_N)^{-1/2}\right),$$

since $s_{p,1} = o(\sqrt{N\gamma_N}/\log(d))$, $\zeta_N = o(1/\sqrt{N\gamma_N})$, and $N\gamma_N \gg \log(d)\log(N)$. Together with (G.21) and (G.22), $\Delta_3 o((N\gamma_N)^{-1/2})$. ■

### G.3. Asymptotic normality of the influence function (IF).

LEMMA G.4. *Let either $m(\cdot) = \widetilde{m}^*(\cdot)$ or $\gamma_N(\cdot) = \widetilde{\gamma}_N^*(\cdot)$ hold. Let Assumptions 1, 2, 3, 4, 5, and 6 hold. Then, $\mathbb{E}\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\} = 0$, $\widetilde{\Sigma}_{N,1}^* = Var\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\} = O(\gamma_N^{-1})$, and*

$$N^{-1}\sum_{i\in\mathcal{I}} \widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) = O_p\left((N\gamma_N)^{-1/2}\right).$$

*Furthermore, let Assumption 7 hold. Then, $\widetilde{\Sigma}_{N,1}^* \asymp \gamma_N^{-1}$, and we have the following asymptotic normality: as $N, d \to \infty$,*

$$\sqrt{N}\left(\widetilde{\Sigma}_{N,1}^*\right)^{-1/2}\left\{N^{-1}\sum_{i\in\mathcal{I}} \widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) - \theta_1\right\} \xrightarrow{d} \mathcal{N}(0,1). \tag{G.23}$$

PROOF OF LEMMA G.4. Notice that, under Assumption 1,

$$\begin{aligned}
\mathbb{E}\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\} &= \mathbb{E}\left[\widetilde{m}^*(\mathbf{X}) + \frac{\Gamma}{\widetilde{\gamma}_N^*(\mathbf{X})}\{Y - \widetilde{m}^*(\mathbf{X})\}\right] - \theta_1 \\
&= \mathbb{E}\left[\left\{\frac{\gamma_N(\mathbf{X})}{\widetilde{\gamma}_N^*(\mathbf{X})} - 1\right\}\{m(\mathbf{X}) - \widetilde{m}^*(\mathbf{X})\}\right] = 0,
\end{aligned}$$

as long as either $m(\cdot) = \widetilde{m}^*(\cdot)$ or $\gamma_N(\cdot) = \widetilde{\gamma}_N^*(\cdot)$. Besides, with $r > 0$ satisfies $1/r + 1/q = 1$,

$$\begin{aligned}
\widetilde{\Sigma}_{N,1}^* &= \mathbb{E}\left[\widetilde{m}^*(\mathbf{X}) + \frac{\Gamma\{Y - \widetilde{m}^*(\mathbf{X})\}}{\widetilde{\gamma}_N^*(\mathbf{X})}\right]^2 - \theta_1^2 \le 2\mathbb{E}\{\widetilde{m}^*(\mathbf{X})\}^2 + 2\mathbb{E}\left[\frac{\Gamma\{Y - \widetilde{m}^*(\mathbf{X})\}^2}{\{\widetilde{\gamma}_N^*(\mathbf{X})\}^2}\right] \\
&= 2\mathbb{E}(\mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^*)^2 + 2\mathbb{E}\left[\gamma_N(\mathbf{X})\{1 + \gamma_N^{-1}\exp(-\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*)\}^2\varepsilon^2\right] \\
&\le 2\mathbb{E}(\mathbf{X}^T\widetilde{\boldsymbol{\alpha}}^*)^2 + 2\{1 + \gamma_N^{-1}\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}^2\|\gamma_N(\cdot)\|_{\mathbb{P},q}\|\varepsilon\|_{\mathbb{P},2r}^2 = O(\gamma_N^{-1}),
\end{aligned}$$

under Assumptions 2, 3, 6, and 5. By Lemma E.1,

$$N^{-1}\sum_{i\in\mathcal{I}} \widetilde{\psi}_{N,1}^*(\mathbf{Z}_i) = O_p\left((N\gamma_N)^{-1/2}\right).$$

Now, we construct a lower bound for $\widetilde{\Sigma}_{N,1}^*$ under an additional Assumption 7.

**Case 1.** Suppose $m(\cdot) = \widetilde{m}^*(\cdot)$. Then,

$$\begin{aligned}
\widetilde{\Sigma}_{N,1}^* &= \mathbb{E}\{\widetilde{m}^*(\mathbf{X}) - \theta_1\}^2 + \mathbb{E}\left[\frac{\Gamma\varepsilon^2}{\{\widetilde{\gamma}_N^*(\mathbf{X})\}^2}\right] + 2\mathbb{E}\left[\frac{\Gamma\varepsilon\{\widetilde{m}^*(\mathbf{X}) - \theta_1\}}{\widetilde{\gamma}_N^*(\mathbf{X})}\right] \\
&\overset{(i)}{=} \mathbb{E}\{\widetilde{m}^*(\mathbf{X}) - \theta_1\}^2 + \mathbb{E}\left[\frac{\gamma_N(\mathbf{X})\varepsilon^2}{\{\widetilde{\gamma}_N^*(\mathbf{X})\}^2}\right] \ge \mathbb{E}\left[\frac{\gamma_N(\mathbf{X})\varepsilon^2}{\{\widetilde{\gamma}_N^*(\mathbf{X})\}^2}\right] \\
&\overset{(ii)}{\ge} \exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})\gamma_N^{-2}\mathbb{E}\left\{\gamma_N(\mathbf{X})\varepsilon^2\right\} \overset{(iii)}{\ge} c_{\min}\exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})\gamma_N^{-1},
\end{aligned}$$

where (i) holds since $\mathbb{E}(\varepsilon \mid \mathbf{X}, \Gamma = 1) = \mathbb{E}(\varepsilon \mid \mathbf{X}, R = T = 1) = \mathbb{E}(\varepsilon \mid \mathbf{X}, T = 1) = \mathbb{E}(\varepsilon \mid \mathbf{X}) = 0$ and $\mathbb{E}(\Gamma\varepsilon^2 \mid \mathbf{X}) = \gamma_N(\mathbf{X})\mathbb{E}(\varepsilon^2 \mid \mathbf{X})$ under Assumption 1; (ii) holds since

$\{\widetilde{\gamma}_N^*(\mathbf{X})\}^{-2} = 1 + \gamma_N^{-2}\exp(-2\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*) \geq \gamma_N^{-2}\exp(-2\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*) \geq \gamma_N^{-2}\exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})$; (iii) holds under Assumption 7 and note that $\mathbb{E}\{\gamma_N(\mathbf{X})\} = \gamma_N$. Therefore,

$$\left(\widetilde{\Sigma}_{N,1}^*\right)^{-1} = O(\gamma_N).$$

**Case 2.** Suppose $\gamma_N(\cdot) = \widetilde{\gamma}_N^*(\cdot)$. Then,

$$\begin{aligned}
\widetilde{\Sigma}_{N,1}^* &= \mathbb{E}\left[\{\widetilde{m}^*(\mathbf{X}) - \theta_1\}^2 + \frac{\varepsilon^2}{\gamma_N(\mathbf{X})} + 2\{\widetilde{m}^*(\mathbf{X}) - \theta_1\}\varepsilon\right] \\
&= \mathbb{E}\left[\{\widetilde{m}^*(\mathbf{X}) - \theta_1 + \varepsilon\}^2 + \frac{\{1-\gamma_N(\mathbf{X})\}\varepsilon^2}{\gamma_N(\mathbf{X})}\right] \geq \mathbb{E}\left[\frac{\{1-\gamma_N(\mathbf{X})\}\varepsilon^2}{\gamma_N(\mathbf{X})}\right] \\
&\overset{(i)}{\geq} (c\gamma_N)^{-1}\exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})\mathbb{E}(\varepsilon^2) \geq (c\gamma_N)^{-1}\exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})c_{\min},
\end{aligned}$$

where (i) holds since $\{1-\gamma_N(\mathbf{X})\}/\gamma_N(\mathbf{X}) = 1/\{\pi^*(\mathbf{X})p_{N,1}^*(\mathbf{X})\} - 1 = \{\gamma_N^{-1}\exp(-\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*) + 1\}/\pi^*(\mathbf{X}) - 1 \geq (c\gamma_N)^{-1}\exp(-C_{\mathbf{X}}C_{\boldsymbol{\beta}})$ under Assumptions 4 and 5 as $1/\pi^*(\mathbf{X}) \geq 1$. Therefore, we also have

$$\left(\widetilde{\Sigma}_{N,1}^*\right)^{-1} = O(\gamma_N).$$

To sum up, as long as either $m(\cdot) = \widetilde{m}^*(\cdot)$ or $\gamma_N(\cdot) = \widetilde{\gamma}_N^*(\cdot)$ holds, we have

$$\widetilde{\Sigma}_{N,1}^* \asymp \gamma_N^{-1}. \tag{G.24}$$

Additionally, with $r > 0$ satisfying $1/r + 1/q = 1$,

$$\begin{aligned}
\|\widetilde{\psi}_{N,1}^*(\cdot)\|_{\mathbb{P},2+c} &= \left\|\widetilde{m}^*(\mathbf{X}) - \theta_1 + \frac{\Gamma\varepsilon}{\widetilde{\gamma}_N^*(\mathbf{X})}\right\|_{\mathbb{P},2+c} \leq \|\widetilde{m}^*(\cdot)\|_{\mathbb{P},2+c} + |\theta_1| + \left\|\frac{\Gamma\varepsilon}{\widetilde{\gamma}_N^*(\mathbf{X})}\right\|_{\mathbb{P},2+c} \\
&\overset{(i)}{\leq} \left\|\mathbf{X}^T\boldsymbol{\alpha}^*\right\|_{\mathbb{P},2+c} + |\theta_1| + \{c^{-1} + (c\gamma_N)^{-1}\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}\left[\mathbb{E}\{\gamma_N(\mathbf{X})|\varepsilon|^{2+c}\}\right]^{1/(2+c)} \\
&\leq \left\|\mathbf{X}^T\boldsymbol{\alpha}^*\right\|_{\mathbb{P},2+c} + |\mathbb{E}(\mathbf{X}^T\boldsymbol{\alpha}^*)| \\
&\quad + \{c^{-1} + (c\gamma_N)^{-1}\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})\}\|\gamma_N(\mathbf{X})\|_{\mathbb{P},q}^{1/(2+c)}\|\varepsilon\|_{\mathbb{P},r(2+c)} \\
&\overset{(ii)}{=} O(\gamma_N^{1/(2+c)-1}),
\end{aligned}$$

where (i) holds since $\mathbb{E}(|\Gamma\varepsilon|^{2+c} \mid \mathbf{X}) = \mathbb{E}(\Gamma|\varepsilon|^{2+c} \mid \mathbf{X}) = \gamma_N(\mathbf{X})\mathbb{E}(|\varepsilon|^{2+c} \mid \mathbf{X})$ and $1/\widetilde{\gamma}_N^*(\mathbf{X}) = \{1 + \gamma_N^{-1}\exp(-\mathbf{X}^T\boldsymbol{\beta}_{p,1}^*)\}/\pi^*(\mathbf{X}) \leq c^{-1} + (c\gamma_N)^{-1}\exp(C_{\mathbf{X}}C_{\boldsymbol{\beta}})$; (ii) holds by part (c) of Lemma E.2 and under Assumption 3. Hence,

$$N^{-c/2}\gamma_N^{1+c/2}\mathbb{E}|\widetilde{\psi}_{N,1}^*(\mathbf{Z})|^{2+c} = O\left((N\gamma_N)^{-c/2}\right) = o(1). \tag{G.25}$$

It follows that, for any $\delta > 0$ as $N, d \to \infty$,

$$\gamma_N^{-1}\mathbb{E}\left[\left\{\widetilde{\psi}_{N,1}^*(\mathbf{Z})\right\}^2 \mathbb{1}_{|\widetilde{\psi}_{N,1}^*(\mathbf{Z})| > \delta\sqrt{N/\gamma_N}}\right] = o(1).$$

By Proposition 2.27 (Lindeberg-Feller theorem) of Van der Vaart (2000), (G.23) holds. ■

### G.4. Asymptotic variance estimation.

LEMMA G.5. *Let the assumptions of Lemma G.4 hold. Let* $N\gamma_N \gg \log(d)\log(N)$, $\bar{s} := s_{p,1} + s_{\widetilde{\boldsymbol{\alpha}}} = o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$, *and* $\zeta_N = o(\{\log(d)/(M\gamma_N)\}^{1/4})$. *Then, with some* $\lambda_{\boldsymbol{\beta}} \asymp \lambda_{\boldsymbol{\alpha}} \asymp \sqrt{\log(d)/(N\gamma_N)}$, *as* $N, d \to \infty$,

$$\widehat{\Sigma}_{\mathit{1,SP\text{-}BRSS}} = \widetilde{\Sigma}^*_{N,1}\{1 + o_p(1)\}. \tag{G.26}$$

PROOF OF LEMMA G.5. For each $i \le N$, define $g_i := \pi^*(\mathbf{X}_i)g(\mathbf{X}_i^T\boldsymbol{\beta}^*_{p,1} + \log(\gamma_N))$, $\bar{g}_i^{(k)} := \widehat{\pi}^{(k)}(\mathbf{X}_i)g(\mathbf{X}_i^T\bar{\boldsymbol{\beta}}_{p,1} + \log(\bar{p}_{N,1}))$, where $g(\cdot)$ is the logistic function and

$$\bar{\psi}^{(k)}_{N,1}(\mathbf{Z}_i) := \mathbf{X}_i^T\overline{\boldsymbol{\alpha}} + \Gamma_i(Y_i - \mathbf{X}_i^T\overline{\boldsymbol{\alpha}})/\bar{g}_i^{(k)} - \widehat{\theta}_{\text{1,SP-BRSS}}.$$

Define $\mathcal{E}_{\bar{\boldsymbol{\beta}}} := \{\|\widehat{\boldsymbol{\beta}}^{(1)}\|_1 \le 8C_{\boldsymbol{\beta}}\} \cap \{\|\widehat{\boldsymbol{\beta}}^{(2)}\|_1 \le 8C_{\boldsymbol{\beta}}\}$. By Theorem 4.1, $\mathbb{P}_{\mathcal{D}_N}(\mathcal{E}_{\bar{\boldsymbol{\beta}}}) = 1 - o(1)$. Condition on the event $\mathcal{E}_{\bar{\boldsymbol{\beta}}} \cap \mathcal{E}_\pi$. Then, for each $i \le N$,

$$1/\bar{g}_i^{(k)} = \{1 + \bar{p}_{N,1}^{-1}\exp(-\mathbf{X}_i^T\bar{\boldsymbol{\beta}}_{p,1})\}/\pi^{(k)}(\mathbf{X}_i) \le \{1 + \exp(8C_{\mathbf{X}}C_{\boldsymbol{\beta}})\bar{p}_{N,1}^{-1}\}/c. \tag{G.27}$$

Observe that

$$\bar{\psi}^{(k)}_{N,1}(\mathbf{Z}_i) - \widetilde{\psi}^*_{N,1}(\mathbf{Z}_i) = \bar{\Delta}^{(k)}_{1,i} + \bar{\Delta}^{(k)}_{2,i},$$

where

$$\bar{\Delta}^{(k)}_{1,i} := (1 - \Gamma_i/\bar{g}_i^{(k)})\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*), \quad \bar{\Delta}^{(k)}_{2,i} := \Gamma_i(1/\bar{g}_i^{(k)} - 1/g_i)\varepsilon_i.$$

Note that

$$\begin{aligned} N^{-1}\sum_{k=1}^2\sum_{i\in\mathcal{I}_k}(\bar{\Delta}^{(k)}_{1,i})^2 &= N^{-1}\sum_{k=1}^2\sum_{i\in\mathcal{I}_k}\left\{1 - 2\Gamma_i/\bar{g}_i^{(k)} + \Gamma_i/(\bar{g}_i^{(k)})^2\right\}\left\{\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2 \\ &\overset{(i)}{\le} N^{-1}\sum_{i=1}^N\left\{\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2 \\ &\quad + c^{-2}\left\{1 + \exp(8C_{\mathbf{X}}C_{\boldsymbol{\beta}})\bar{p}_{N,1}^{-1}\right\}^2 N^{-1}\sum_{i=1}^N\Gamma_i\left\{\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\right\}^2. \end{aligned} \tag{G.28}$$

where (i) holds by (G.27). By part (a) of Lemma E.6 choosing $s = \bar{s} + \zeta_N^2 N\gamma_N/\log(d) = o(N/\log(d))$, together with (G.14) and (G.15),

$$N^{-1}\sum_{i=1}^N\{\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\}^2 = O_p\left(\frac{\bar{s}^2\log(d)}{N^2\gamma_N} + \frac{\zeta_N^4\gamma_N}{\log(d)} + \frac{\bar{s}\log(d)}{N\gamma_N} + \zeta_N^2\right) = o_p(1),$$

as long as $\bar{s} = o(N\gamma_N/\log(d))$, $\zeta_N = o(1)$, and $N\gamma_N \gg \log(d)\log(N)$. Additionally, by part (b) of Lemma E.6, together with (G.14) and (G.15), we also have

$$\begin{aligned} &(N\gamma_N)^{-1}\sum_{i=1}^N\Gamma_i\{\mathbf{X}_i^T(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\}^2 \\ &\quad = O_p\left(\frac{\bar{s}^2\log^2(d)\log(N)}{(N\gamma_N)^2} + \zeta_N^4\log(N) + \frac{\bar{s}\log(d)}{N\gamma_N} + \zeta_N^2\right) = o_p(1), \end{aligned}$$

if $\bar{s} = o(N\gamma_N/\{\log(d)\log^{1/2}(N)\})$ and $\zeta_N = o(\log^{-1/4}(N))$.

By Lemma F.3 and note that $p_{N,1} \asymp \gamma_N$, we have $\bar{p}_{N,1} = O_p(\gamma_N)$ and $\bar{p}_{N,1}^{-1} = O_p(\gamma_N^{-1})$. Now, by (G.28), we obtain

$$N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,i}^{(k)})^2 = o_p(\gamma_N^{-1}).$$

Besides, under Assumption 1, we have for each $k \in \{1,2\}$,

$$\mathbb{E}_{\mathcal{D}_N}\left\{M^{-1}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,2}^{(k)})^2 \mid (R_i,T_i,\mathbf{X}_i)_{i\in\mathcal{D}_N^{(k)}}\right\} = \mathbb{E}(\varepsilon^2|\mathbf{X})M^{-1}\sum_{i\in\mathcal{I}_k}\Gamma_i(1/\bar{g}_i^{(k)} - 1/g_i)^2.$$

Repeat the same procedure as in the proof of Lemma G.1, we also have

$$(M\gamma_N)^{-1}\sum_{i\in\mathcal{I}_k}\Gamma_i(1/\bar{g}_i^{(k)} - 1/g_i)^2 = o_p(1).$$

Under Assumption 6, it follows that

$$\mathbb{E}_{\mathcal{D}_N}\left\{M^{-1}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,2}^{(k)})^2 \mid (R_i,T_i,\mathbf{X}_i)_{i\in\mathcal{D}_N^{(k)}}\right\} = o_p(\gamma_N^{-1}).$$

By Lemma E.1, for each $k \in \{1,2\}$,

$$M^{-1}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,2}^{(k)})^2 = o_p(\gamma_N^{-1}).$$

It follows that

$$N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,2}^{(k)})^2 = o_p(\gamma_N^{-1}).$$

Therefore,

$$\begin{aligned} N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\left\{\bar{\psi}_{N,1}^{(k)}(\mathbf{Z}_i) - \widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2 &\le 2N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,1}^{(k)})^2 + 2N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}(\bar{\Delta}_{1,2}^{(k)})^2 \\ &= o_p(\gamma_N^{-1}). \end{aligned}$$

In addition, by (G.24), (G.25), and Lemma D.3 of Zhang, Chakrabortty and Bradic (2023), we have

$$N^{-1}\sum_{i=1}^{N}\left\{\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2 = \widetilde{\Sigma}_{N,1}^{*}\{1+o_p(1)\} = O_p(\gamma_N^{-1}).$$

Now, we note that

$$\begin{aligned} &\left|N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\{\bar{\psi}_{N,1}^{(k)}(\mathbf{Z}_i)\}^2 - N^{-1}\sum_{i=1}^{N}\left\{\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2\right| \\ &\quad \le N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\left\{\bar{\psi}_{N,1}^{(k)}(\mathbf{Z}_i) - \widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2 \\ &\qquad + 2\left|N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\left\{\widetilde{\psi}_{N,1}^{(k)}(\mathbf{Z}_i) - \widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}\right| \end{aligned}$$

$$
\begin{aligned}
&\le N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\left\{\bar{\psi}_{N,1}^{(k)}(\mathbf{Z}_i)-\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2\\
&\quad +2\sqrt{N^{-1}\sum_{i=1}^{N}\left\{\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2}\sqrt{N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\left\{\bar{\psi}_{N,1}^{(k)}(\mathbf{Z}_i)-\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2}\\
&=o_p(\gamma_N^{-1})+O_p(\gamma_N^{-1/2})o_p(\gamma_N^{-1/2})=o(\gamma_N^{-1})=o_p(\widetilde{\Sigma}_{N,1}^{*}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&|\widehat{\Sigma}_{1,\text{SP-BRSS}}-\widetilde{\Sigma}_{N,1}^{*}|\\
&\quad\le\left|N^{-1}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\{\widetilde{\psi}_{N,1}^{(k)}(\mathbf{Z}_i)\}^2-N^{-1}\sum_{i=1}^{N}\left\{\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2\right|\\
&\qquad+\left|N^{-1}\sum_{i=1}^{N}\left\{\widetilde{\psi}_{N,1}^{*}(\mathbf{Z}_i)\right\}^2-\widetilde{\Sigma}_{N,1}^{*}\right|\\
&\quad=o_p(\widetilde{\Sigma}_{N,1}^{*})+\widetilde{\Sigma}_{N,1}^{*}o_p(1)=o_p(\widetilde{\Sigma}_{N,1}^{*}),
\end{aligned}
$$

and (G.26) follows. ■

### G.5. Proofs of Theorems 4.4 and 4.5.

PROOF OF THEOREM 4.4. Theorem 4.4 follows directly from Lemmas G.1, G.2, G.3, G.4, and G.5. ■

PROOF OF THEOREM 4.5. Theorem 4.5 follows directly from Lemmas G.2, G.3, G.4, and G.5. ■