HIGH DIMENSIONAL M-ESTIMATION WITH MISSING OUTCOMES: A SEMI-PARAMETRIC FRAMEWORK

By Abhishek Chakrabortty*, Jiarui Lu, T. Tony Cai and Hongzhe Li

University of Pennsylvania

In this paper, we consider high dimensional M-estimation problems in settings where the response Y is possibly missing at random and the covariates $\mathbf{X} \in \mathbb{R}^p$ can be high dimensional compared to the sample size n (including $p \gg n$), settings that are of great relevance in a variety of modern studies. The parameter of interest $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is defined simply as the minimizer of the risk of a convex loss, under a fully non-parametric model, and $\boldsymbol{\theta}_0$ *itself* is *high dimensional* which is a key distinction from existing works in the relevant literature (e.g. estimation of means or average treatment effects in high dimensional settings). As special cases, our framework includes all standard high dimensional regression and series estimation problems with possibly misspecified models and missing Y. Under an equivalent formulation of this setting based on 'potential' outcomes in causal inference, these parameters also have important applications in heterogeneous treatment effects estimation that are of interest in precision medicine.

Assuming $\boldsymbol{\theta}_0$ is s-sparse ($s \ll n$), we propose to estimate $\boldsymbol{\theta}_0$ via an L_1-regularized debiased and doubly robust (DDR) estimator based on a high dimensional adaptation of traditional double robust (DR) estimators' construction along with careful use of debiasing and sample splitting. Under mild tail assumptions and arbitrarily chosen (working) models for the propensity score (PS) and the outcome regression (OR) estimators, satisfying *only* some high-level consistency conditions, we establish *finite sample* performance bounds for the DDR estimator showing its (optimal) L_2 error rate to be $\sqrt{s(\log d)/n}$ when both working models are correct, and its consistency and DR properties when only one of them is correct. Further, when both models are correct, we propose a *desparsified* version of our DDR estimator that satisfies an *asymptotic linear expansion* and facilitates *inference* on low dimensional components of $\boldsymbol{\theta}_0$. Finally, we discuss various choices of high dimensional parametric and semi-parametric working models for the PS and OR estimators and establish their properties needed for our main results. All results are validated via detailed simulations.

---

*Corresponding author. Date of this version: August 2, 2019. This research was partially supported by the National Institutes of Health grants R01-GM123056 and R01-GM129781.

MSC 2010 subject classifications: 62F10, 62F12, 62J07, 62F25, 62F35, 62G08, 62J02.

Keywords and phrases: Missing data, Causal inference, Regularized M-estimation, Double robustness, Debiasing, Sparsity, High dimensional inference, Nuisance functions.

1. Introduction. Large and complex observational data are commonplace in the modern 'big data' era. Statistical analyses of such datasets often poses unique challenges that has led to a plethora of work in recent times. In particular, two such frequently encountered challenges include: (a) *high dimensional settings,* wherein the dimension of the observed covariates is often comparable to or far exceeds the available sample size, and (b) potential *incompleteness in the data,* especially in the outcome (or response) variable of interest. Both these issues arise naturally (and often concurrently) whenever observations are easily available for several covariates but the corresponding response is difficult and/or expensive to obtain. The latter could be due to practical constraints (e.g. logistics, time, cost etc.) or simply by 'design' (e.g. any treatment-response data setting in causal inference, where the response is automatically unobserved for any untreated individual). All these scenarios are routinely encountered in a variety of modern studies involving large databases, including biomedical data like electronic health records, or eQTL mapping studies in integrative genomics involving gene expression data, as well as in econometrics (e.g. policy evaluation). Further, owing to the very *observational nature* of the data, the underlying missingness (or 'treatment' assignment) mechanism is often informative (i.e. not randomized) and depends on the covariates which leads to further complexities of *selection bias* and confounding issues. Appropriate accounting of such biases is *essential* to ensure the validity of any subsequent statistical analyses and inference.

For issue (a) above, both estimation and inference under high dimensional settings, *but* with complete data, are by now quite well studied and equipped with a vast and growing literature centered around regularized methods and sparsity; see Bühlmann and Van De Geer (2011) and Wainwright (2019) for an overview. For issue (b) as well, under classical (low dimensional) settings, there has been substantial work leading to a rich body of literature on semiparametric inference for incomplete response data. We refer to Tsiatis (2007) and Bang and Robins (2005) for a review, as well as the fundamental works of Robins, Rotnitzky and Zhao (1994) and Robins and Rotnitzky (1995). Even under high dimensional settings, there has been a recent surge of work aimed at an analogous treatment of these problems but mostly in cases where the parameter of interest is still low dimensional (typically, the mean response) (Farrell, 2015; Belloni et al., 2017; Chernozhukov et al., 2018a). In this paper, we consider a more challenging and unique setting that essentially represents a confluence of all the issues highlighted above, combined with the fact the parameter of interest *itself* is high dimensional, something that has received relatively limited attention as of now. We first formalize our basic setup and the problem of interest, followed by an overview of our contributions.

1.1. *Problem Setup, Available Data and the Basic Assumptions.* Let $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$ denote an outcome variable and a covariate vector of interest respectively, with supports $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^p$ neither of which necessarily need to be continuous. In practice, however, Y may not always be observed and let $T \in \{0,1\}$ denotes the indicator of Y being observed. $\mathbb{Z} := (T, Y, \mathbf{X})$ is assumed to be defined jointly under some probability measure $\mathbb{P}(\cdot)$, while the *observable* random vector is: $\mathbf{Z} := (T, TY, \mathbf{X})$. The *observed data* $\mathcal{D}_n := \{\mathbf{Z}_i \equiv (T_i, T_i Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ consists of n independent and identically distributed (i.i.d.) realizations of $\mathbf{Z}$ with joint distribution defined via $\mathbb{P}(\cdot)$. We emphasize here that our focus is on *high dimensional* settings, where the covariate dimension p is allowed to diverge with n (possibly, faster than n).

Assumption 1.1 (Basic assumptions). We assume throughout two basic conditions which are both fairly standard in the literature (Imbens, 2004).

(a) *Ignorability:* $T \perp\!\!\!\perp Y \,|\, \mathbf{X}$, so that the missingness mechanism may depend on $\mathbf{X}$, but is conditionally independent of Y given $\mathbf{X}$. This is also referred to often as the missing at random (MAR) assumption in the literature.

(b) *Positivity/overlap:* let $\pi(\mathbf{X}) := \mathbb{P}(T = 1|\mathbf{X})$ denote the *propensity score* (Rosenbaum and Rubin, 1983), and let $\pi := \mathbb{P}(T = 1)$. Then, we assume:

$$\pi(\mathbf{x}) \;\geq\; \delta_\pi > 0 \;\; \forall\, \mathbf{x} \in \mathcal{X}, \quad \text{for some constant } \; \delta_\pi \in (0,1]. \tag{1.1}$$

Hence, the probability of observing Y given $\mathbf{X}$ is always strictly positive.

The MAR assumption in 1.1 (a) also includes the special case $T \perp\!\!\!\perp (Y, \mathbf{X})$, commonly known as missing completely at random (MCAR). In such cases, $\pi(\cdot)$ simply equals the constant π from part (b). In general, $\pi(\cdot)$ is allowed to depend on $\mathbf{X}$ and may be unknown in practice when it needs to be estimated.

The framework and notations above are in accordance with the standard treatment in the missing data literature (Tsiatis, 2007). However, the setting also encompasses problems in causal inference under the 'potential' outcome framework. These may be equivalently formulated as missing data problems, a fact well known in the literature. We briefly discuss this equivalence below.

Causal inference under 'potential' outcomes framework. In this setting, the observable vector is $\mathbf{Z} := (T, \mathbb{Y}, \mathbf{X})$, where $T \in \{0,1\}$ denotes a binary 'treatment' assignment indicator (can be any kind of assignment or intervention) and $\mathbb{Y} := TY^{(1)} + (1-T)Y^{(0)}$ denotes the observed outcome with $(Y^{(1)}, Y^{(0)})$ being the true 'potential' outcomes (Rubin, 1974; Imbens and Rubin, 2015) for $T = 1$ and $T = 0$ respectively. Thus, for each potential outcome, this corresponds to our setting if we set $(Y, T) \equiv (Y^{(1)}, T)$ or $(Y, T) \equiv (Y^{(0)}, 1 - T)$.

It is also worth noting that in the causal inference (CI) literature, $\mathbf{X}$ is often referred to as 'confounders' (in observational studies) or 'adjustment' variables (in randomized trials), while the MAR assumption is often known as no unmeasured confounding (NUC) and MCAR as complete randomization.

1.2. *High Dimensional M-Estimation.* We next introduce our *main problem* of interest under this setting. Let $L(Y, \mathbf{X}, \boldsymbol{\theta}) : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ be any 'loss' function that is convex and differentiable in $\boldsymbol{\theta}$, and we assume that $[\mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\}^2] < \infty \; \forall \, \boldsymbol{\theta} \in \mathbb{R}^d$. Then, the M-estimation problem considers the estimation of the minimizer $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ of the risk function defined by $L(\cdot)$. Specifically, we aim to estimate the functional $\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P}) \in \mathbb{R}^d$ defined as:

$$\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(L, \mathbb{P}) := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \; \mathbb{L}(\boldsymbol{\theta}), \quad \text{where} \; \mathbb{L}(\boldsymbol{\theta}) := \mathbb{E}\left\{L(Y, \mathbf{X}, \boldsymbol{\theta})\right\}. \tag{1.2}$$

Here, d is allowed to be high dimensional, i.e. d can diverge with n (possibly faster). We assume without loss of generality (w.l.o.g.) that $d \geq 2$. The existence and uniqueness of $\boldsymbol{\theta}_0$ is implicitly assumed given the generality of the framework considered. For most standard examples, this is fairly straightforward to establish with $L(\cdot)$ being convex and sufficiently smooth in $\boldsymbol{\theta}$. For convenience of further discussion, let us define: $\forall \; y \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\phi(\mathbf{x}, \boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}\} \; \text{ and } \; \boldsymbol{\nabla} L(y, \mathbf{x}, \boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}} L(y, \mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^d.$$

REMARK 1.1. It is important to note that $\boldsymbol{\theta}_0$ in (1.2) is defined under a fully non-parametric family of $\mathbb{P}$ without any restrictions (upto Assumption 1.1 and basic moment conditions). Hence, the framework is *semi-parametric* and *model free* in this sense with $\boldsymbol{\theta}_0(\mathbb{P})$ well-defined for every $\mathbb{P}$ without any model assumptions for $Y|\mathbf{X}$ (even though $\boldsymbol{\theta}_0$ may sometimes be 'motivated' by such 'working' models for $Y|\mathbf{X}$, as in the case of regression problems).

Further, the framework also highlights the *necessity of accounting for the incompleteness* of $\mathcal{D}_n$. If one simply ignores it and chooses to estimate $\boldsymbol{\theta}_0$ via risk minimization in the complete part of the data (i.e. observations with $T = 1$), then the corresponding 'complete case' (CC) estimator will, in general, be *inconsistent* for $\boldsymbol{\theta}_0$ since the target parameter for this estimator is simply the minimizer of $\mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})|T = 1\}$ which bears no direct relation to the unconditional minimizer $\boldsymbol{\theta}_0$ in (1.2). The only cases when the CC estimator will just so happen to be consistent for $\boldsymbol{\theta}_0$ is if either $T \perp\!\!\!\perp (Y, \mathbf{X})$, i.e. MCAR holds (hence, there is no selection bias), or if $\mathbb{E}\{\boldsymbol{\nabla} L(Y, \mathbf{X}, \boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$ almost surely (a.s.) $[\mathbb{P}_{\mathbf{X}}]$. In case of regression problems, the latter implies a correctly specified parametric model holds for $\mathbb{E}(Y|\mathbf{X})$ with the 'true' parameter being

$\boldsymbol{\theta}_0$. Both these cases, however, correspond to additional restrictions on $\mathbb{P}$. In general, for consistent estimation of $\boldsymbol{\theta}_0$ over the *entire* large family of $\mathbb{P}$ where it is defined, appropriate accounting of the missingness is thus necessary.

Finally, it is worth mentioning that a special low dimensional case of (1.2) is the *mean estimation* problem where $\boldsymbol{\theta}_0 = \mathbb{E}(Y)$ with $L(Y, \mathbf{X}, \boldsymbol{\theta}) = (Y - \boldsymbol{\theta})^2$ and $d = 1$. In causal inference under the 'potential' outcome framework, this also corresponds to the *average treatment effect* (ATE) estimation problem. Both versions of this problem have by now been extensively studied in classical as well as high dimensional settings, especially the latter in recent times. We defer a detailed literature review to Section 1.4 and only point out here that the *key distinction* between this literature and our setting is that our parameter of interest $\boldsymbol{\theta}_0$ in (1.2) is *itself high dimensional* (apart from $\mathbf{X}$).

1.3. *Some Applications.* The framework (1.2) encompasses a broad range of important problems. We enlist below a few useful examples for illustration.

1. High dimensional regression with possibly misspecified models and missing outcomes. (1.2) includes all standard high dimensional regression problems, where we further allow for: (i) potentially misspecified (working) models and (ii) Y to be partly unobserved. For instance, set $\boldsymbol{\theta} = (a, \mathbf{b})$ and $L(Y, \mathbf{X}, \boldsymbol{\theta}) := l(Y, a + \mathbf{b}'\mathbf{X})$ in (1.2), with $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^p$ and $l(u, v) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ being some loss function convex and differentiable in v. Typical choices of $l(\cdot, \cdot)$ include the 'canonical' losses leading to standard regression problems as follows.

(a) The *squared loss:* $l(u, v) \equiv l_{\text{sq}}(u, v) := (u - v)^2$ (for linear regression).
(b) The *logistic loss*: $l(u, v) \equiv l_{\log}(u, v) := -uv + \log\{1 + \exp(v)\}$ (for logistic regression) and *exponential loss*: $l(u, v) \equiv l_{\exp}(u, v) := -uv + \exp(v)$ (for Poisson regression), used often for binary or count valued Y respectively.

In all examples, $\boldsymbol{\theta}_0$ is *model free* and is well defined *regardless* of the validity of any motivating parametric (working) model for $Y \mid \mathbf{X}$. In general, it simply corresponds to the 'projection' of $\mathbb{E}(Y \mid \mathbf{X})$ onto that working model space.

As an extension, one may also consider any (model free) *series estimation problem* by replacing $\mathbf{X}$ above with $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_j(\mathbf{X})\}_{j=1}^d$, a vector (possibly high dimensional) of d basis functions comprising transformations (possibly non-linear) of $\mathbf{X}$. We may analogously set $L(Y, \mathbf{X}, \boldsymbol{\theta}) := l\{Y, \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}\}$ with the same choices of $l(\cdot, \cdot)$ as above. A frequently used choice of $\boldsymbol{\Psi}(\cdot)$ includes the polynomial bases: $\boldsymbol{\Psi}(\mathbf{X}) := \{1, \mathbf{x}_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\}$, for any fixed degree $d_0 \geq 1$ whereby $d = pd_0 + 1$. The special case of $d_0 = 1$ (linear bases) leads to all the earlier examples, while $d_0 = 3$ leads to the cubic spline bases.

2. High dimensional single index models (SIMs) with elliptically symmetric designs. Another interesting application of (1.2) lies in signal recovery in

SIMs with elliptically symmetric designs that satisfy a certain 'linearity condition'. To this end, consider the SIM $Y = f(\beta_0'\mathbf{X}, \epsilon)$, where $f(\cdot): \mathbb{R}^2 \to \mathcal{Y}$ is an *unknown* link function, $\epsilon \perp\!\!\!\perp \mathbf{X}$ is a random noise (so that $Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}_0'\mathbf{X}$) and $\boldsymbol{\beta}_0$ denotes the unknown index parameter (identifiable *only* upto scalar multiples). Now, consider any of the regression problems introduced in Example 1 and assume further that $\mathbf{X}$ has an elliptically symmetric distribution (e.g. Gaussian). Then, $\boldsymbol{\theta}_0 \equiv (a_0, \mathbf{b}_0)$ defined therein satisfies: $\mathbf{b}_0 \propto \boldsymbol{\beta}_0$. This result, first noted by Li and Duan (1989), provides an 'easy' route to signal recovery in SIMs, especially in high dimensional settings and with missing outcomes. This also serves as a classic example where the parameter $\boldsymbol{\theta}_0$ is defined based on a misspecified parametric model and yet, it has direct interpretability relating it to a parameter characterizing a larger semi-parametric model and allows one to still simply use (1.2) for signal recovery in a SIM.

3. Applications in causal inference (heterogeneous treatment effects). All the problems in Examples 1 and 2 also have equivalent counterparts in causal inference under the 'potential' outcome framework discussed in Section 1.1. In this setting, these problems have important applications in the estimation of *heterogeneous treatment effects* which is of great interest in personalized medicine. Fundamentally, this problem relates to estimation of the *average conditional treatment effect* (ACTE): $\Delta(\mathbf{X}) := \mathbb{E}\{Y_{(1)} - Y_{(0)} | \mathbf{X}\}$. In classical settings, estimation of $\Delta(\mathbf{X})$ via non-parametric machine learning methods has received considerable attention in recent times, including use of random forests or neural networks (Wager and Athey, 2017; Farrell, Liang and Misra, 2018). However, in a 'truly' high dimensional setting, wherein p diverges with n (possibly, at a comparable or faster rate), fully non-parametric approaches may not be feasible and/or efficient. In such cases, it is often more reasonable to focus on (model free) projections of $\Delta(\mathbf{X})$ on finite (but high) dimensional function spaces. For the space of linear functions of $\mathbf{X}$, this leads to the *linear heterogeneous treatment effects* estimation problem. Such ideas and problems have indeed been advocated and considered in the recent works of Chernozhukov et al. (2017a) and Chernozhukov and Semenova (2017).

In our framework, this simply corresponds to the linear regression problem discussed in Example 1 (when adapted to the CI setup). Furthermore, under our setting, one can consider even more general problems by focussing on non-linear basis function spaces (e.g. series estimation) and/or other loss functions (e.g. logistic regression). These problems precisely correspond to the other illustrations in Example 1. On the other hand, using the illustration in Example 2, one may also consider ACTE estimation based on SIMs which provide clear generalizations over standard parametric models and yet, to the best of our knowledge, has received relatively less attention in the literature.

1.4. *Overview of Related Literature and Summary of Our Contributions.* Our work contributes to two distinct lines of literature: (i) high dimensional M-estimation *and* inference, and (ii) semi-parametric 'doubly robust' inference for incomplete (and high dimensional) data. As regards the first line of work, for a *complete* data, M-estimation problems are quite well studied in both classical and high dimensional settings; see Van der Vaart (2000) for an overview of the vast classical literature and Negahban et al. (2012); Loh and Wainwright (2012, 2015); Loh (2017) for some of the more recent advances in high dimensional settings. Relatively little work, however, has been done for the case of incomplete (in the response) data, especially in high dimensional settings. In classical low dimensional settings, inference with incomplete data has a rich literature on semi-parametric methods and so called 'doubly robust' inference. We refer to Bang and Robins (2005); Tsiatis (2007); Kang and Schafer (2007) and Graham (2011) for a review. Some of the pioneering works in this area were by Robins, Rotnitzky and Zhao (1994); Robins and Rotnitzky (1995) and some of their ensuing works on other related problems which we skip here for brevity. In recent times, there has also been substantial interest in the extension of these approaches to high dimensional settings leading to a flurry of papers, including Belloni, Chernozhukov and Hansen (2014), Farrell (2015), Belloni et al. (2017), Chernozhukov et al. (2018a) and Athey, Imbens and Wager (2016), among many other notable ones which we don't attempt to enlist here. However, their focus has still mostly been on simple low dimensional parameters like the mean (or the ATE) and less on cases where the *parameter itself is high dimensional.* This is one of the key distinctions of our framework. To our best knowledge, only Chernozhukov and Semenova (2017) and Chernozhukov et al. (2018b) have recently considered settings of a similar sort. While the former considers only the special case of linear regression and that too under a moderate dimensional setting (with $d \ll \sqrt{n}$), the latter certainly allows for a more general framework but their approach is also somewhat abstract. Our approach is more detailed and targeted specifically towards the missing data setting, where we provide a complete hands-on solution to the problem (1.2). Further, another key contribution of our work is to provide inferential tools for our estimator which hasn't been considered therein or any other existing work for that matter.

Our *main contributions* can be summarized in *three* different facets: (i) *estimation*, (ii) *inference* and (iii) *estimation of the nuisance functions.* Adopting a semi-perspective (as in Remark 1.1) and assuming $\boldsymbol{\theta}_0$ is s-sparse (with $s \ll n$), we propose to estimate $\boldsymbol{\theta}_0$ via an *L_1-regularized debiased and doubly robust (DDR) estimator* based on a high dimensional adaptation of the traditional double robust (DR) estimator's construction, along with careful use of debiasing and sample splitting techniques. The DDR estimator serves as the appropriate *generalization* of standard (low dimensional) DR estimators (Bang and Robins, 2005; Chernozhukov et al., 2018a) for high dimensional parameters. We also present a simple *user friendly implementation algorithm* for these estimators which can be achieved with standard software packages. The ambient high dimensionality (of both $\mathbf{X}$ and $\boldsymbol{\theta}_0$) coupled with the missingness of Y and the unavoidable presence of other nuisance function estimators (possibly also high dimensional) makes the analyses challenging and substantially *nuanced* compared to the low dimensional case. Under mild tail assumptions and arbitrarily chosen (working) models for estimating the two *nuisance functions*, the propensity score (PS) and the outcome regression (OR) function, satisfying *only* some *high-level (pointwise) consistency conditions*, we establish *finite sample performance bounds* for the DDR estimator showing its (optimal) L_2 error rate to be $\sqrt{s(\log d)/n}$ when both working models are correct, and its consistency and DR properties when only one of them is correct. Further, the estimators are first order *insensitive* to any estimation errors or knowledge of construction of the PS and OR estimators, thus allowing the use of non-smooth high dimensional and/or adhoc non/semi-parametric estimators with unclear first order properties. Further, when both models are correct, we propose a *desparsified version* of our DDR estimator that satisfies an *asymptotic linear expansion* (ALE) and facilitates *inference* on low dimensional components of $\boldsymbol{\theta}_0$. The desparsified DDR estimator is similar (in spirit) to a Debiased Lasso type approach (van de Geer et al., 2014; Javanmard and Montanari, 2014) and serves as its appropriate generalization in the missing data setting. Furthermore, the ALE it achieves is *semi-parametric optimal* and matches the 'efficient' influence function for this problem. Finally, we also discuss various novel and flexible *choices of the nuisance function estimators*, including common high dimensional parametric models, as well as more general semi-parametric models based on series estimators and single index models. We also establish general results for all these estimators under high dimensional settings that verify their properties needed for our main results and may also be of independent interest. All our results regarding estimation, inference and the DR properties are validated via extensive simulation studies over various data settings, nuisance function (working) models and comparisons with other (optimal) oracle estimators.

Organization. The rest of this paper is organized as follows. In Section 2, we detail our estimation strategy, including preliminaries on DR estimation, followed by construction and implementation of the DDR estimator as well as deterministic deviation bounds on its performance. Section 3 contains our main results (Theorems 3.1-3.4), and the associated high-level assumptions, regarding convergence rates of the DDR estimator via non-asymptotic probabilistic bounds for various error terms. In Section 4, we discuss inference via the deparsified DDR estimator and establish all its properties in Theorem 4.1. In Section 5, we discuss various choices of the nuisance function estimators and also establish their properties through Theorems 5.1-5.3. Finally, the simulation results are presented in Section 6, followed by a concluding discussion in Section 7. In the Supplementary Material (Appendices A-K), we collect several important materials that could not be accommodated in the main manuscript, including discussions on DR properties of the estimator, additional numerical results and all technical materials and discussions, including proofs for all the main results and associated supporting lemmas.

2. Estimation Strategy: A General Approach Based on L_1- Regularized Debiased and Doubly Robust (DDR) Loss Minimization.

Notations. We use the following general notations throughout. For any $\mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{v}\|_r$ denotes the L_r vector norm of $\mathbf{v}$ for any $r \geq 0$, $\overrightarrow{\mathbf{v}}$ denotes $(1, \mathbf{v}')' \in \mathbb{R}^{d+1}$, $\mathbf{v}_{[j]}$ denotes the j^{th} coordinate of $\mathbf{v}$ $\forall\, 1 \leq j \leq d$, $\mathcal{A}(\mathbf{v}) := \{j : \mathbf{v}_{[j]} \neq 0\}$ denotes the support of $\mathbf{v}$ and $s_{\mathbf{v}} := |\mathcal{A}(\mathbf{v})|$ denotes the cardinality of $\mathcal{A}(\mathbf{v})$. For any $\mathcal{J} \subseteq \{1, \ldots, d\}$ and $\mathbf{v} \in \mathbb{R}^d$, we let $\Pi_{\mathcal{J}}(\mathbf{v}) := [\mathbf{v}_{[j]} 1\{j \in \mathcal{J}\}]_{j=1}^d \in \mathbb{R}^d$, $\mathcal{M}_{\mathcal{J}} := \{\mathbf{v} \in \mathbb{R}^d : \mathcal{A}(\mathbf{v}) \subseteq \mathcal{J}\}$ and $\mathcal{M}_{\mathcal{J}}^{\perp} := \{\mathbf{v} \in \mathbb{R}^d : \mathcal{A}(\mathbf{v}) \subseteq \mathcal{J}^c\}$, where $\mathcal{J}^c := \{1, \ldots, d\} \backslash \mathcal{J}$ denotes the complement of $\mathcal{J}$. We use the shorthand $\Pi_{\mathbf{v}}(\cdot)$ and $\Pi_{\mathbf{v}}^c(\cdot)$ to denote $\Pi_{\mathcal{A}(\mathbf{v})}(\cdot)$ and $\Pi_{\mathcal{A}^c(\mathbf{v})}(\cdot)$ respectively. Further, for any measurable (and possibly random) function $f(\cdot)$ of $\mathbf{X}$, we let $\|f(\cdot)\|_r := [\mathbb{E}_{\mathbf{X}}\{|f(\mathbf{X})|^r\}]^{1/r}$ denote the L_r norm of $f(\cdot)$ with respect to (w.r.t.) $\mathbb{P}_{\mathbf{X}}$ for any $r \geq 1$ and $\|f(\cdot)\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$ denote the L_∞ norm w.r.t. $\mathbb{P}_{\mathbf{X}}$. For any sequences $a_n, b_n \geq 0$, we use $a_n \lesssim b_n$ to denote $a_n \leq C b_n$ and $a_n \asymp b_n$ to denote $c b_n \leq a_n \leq C b_n$ for all $n \geq 1$ and some constants $c, C > 0$. Finally, $a_n \ll b_n$ denotes $a_n = o(b_n)$ and $a_n \gg b_n$ denotes $b_n = o(a_n)$ as $n \to \infty$.

2.1. *Identification and Alternative Representations of the Expected Loss.*

We next provide three alternative representations of $\mathbb{L}(\cdot)$ in terms of the observables $(T, TY, \mathbf{X})$ and some *nuisance functions* identifiable through them. These representations also underlie three fundamental estimation strategies typically adopted in the literature for these problems, namely inverse probability weighting (IPW) involving the propensity score $\pi(\cdot)$, regression based imputation (REG) involving the conditional mean $\phi(\cdot,\cdot)$, and *'doubly robust'* (DR) methods that use both IPW as well as regression based imputation and provide the benefits of (double) robustness against model misspecification in the estimation of either one of the two nuisance functions $\pi(\cdot)$ and $\phi(\cdot,\cdot)$. DR estimators are also known to be (locally) semi-parametric optimal when both nuisance function estimation models are correctly specified. We refer to Robins, Rotnitzky and Zhao (1994); Robins and Rotnitzky (1995); Imbens (2004); Bang and Robins (2005); Kang and Schafer (2007); Tsiatis (2007) and Graham (2011) for a detailed review of the related classical literature.

IPW and regression based representations of $\mathbb{L}(\cdot)$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, we have:

$$\begin{aligned} \mathbb{L}(\boldsymbol{\theta}) &\equiv \mathbb{E}\{L(Y,\mathbf{X},\boldsymbol{\theta})\} = \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X},\boldsymbol{\theta})\} =: \mathbb{L}_{\text{REG}}(\boldsymbol{\theta}) \ \ (\text{say}), \quad \text{and} \\ \mathbb{L}(\boldsymbol{\theta}) &\equiv \mathbb{E}\{L(Y,\mathbf{X},\boldsymbol{\theta})\} = \mathbb{E}\left\{\frac{T}{\pi(\mathbf{X})}L(Y,\mathbf{X},\boldsymbol{\theta})\right\} =: \mathbb{L}_{\text{IPW}}(\boldsymbol{\theta}) \ \ (\text{say}). \end{aligned}$$

Debiased and doubly robust (DDR) representation of $\mathbb{L}(\cdot)$. It also holds that:

$$\begin{aligned} \mathbb{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X},\boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi(\mathbf{X})}\{L(Y,\mathbf{X},\boldsymbol{\theta}) - \phi(\mathbf{X},\boldsymbol{\theta})\}\right] \\ &=: \mathbb{L}_{\text{DDR}}(\boldsymbol{\theta}) \ \ (\text{say}) \quad \forall\, \boldsymbol{\theta} \in \mathbb{R}^d. \end{aligned} \tag{2.1}$$

Further, for any functions $\phi^*(\mathbf{X},\boldsymbol{\theta})$ and $\pi^*(\mathbf{X})$ such that $\phi^*(\cdot,\cdot) = \phi(\cdot,\cdot)$ or $\pi^*(\cdot) = \pi(\cdot)$ holds, but *not* necessarily both, it continues to hold that:

$$\mathbb{L}_{\text{DDR}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}\{\phi^*(\mathbf{X},\boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi^*(\mathbf{X})}\{L(Y,\mathbf{X},\boldsymbol{\theta}) - \phi^*(\mathbf{X},\boldsymbol{\theta})\}\right]. \tag{2.2}$$

$\mathbb{L}_{\text{DDR}}(\cdot)$, unlike $\mathbb{L}_{\text{IPW}}(\cdot)$ and $\mathbb{L}_{\text{REG}}(\cdot)$, is thus DR as it is 'protected' against misspecification of either $\pi(\cdot)$ or $\phi(\cdot,\cdot)$, as in (2.2). Further, even when both are correctly specified, it has a naturally 'debiased' form owing to the second term in (2.1). While this term is simply 0 in the population version, it leads to *crucial* first order benefits in the empirical version of the loss involving the nuisance function estimators, where it has a debiasing effect making the loss first order insensitive to any estimation errors of the nuisance functions. Approaches based on other representations don't enjoy these benefits which can be especially crucial in high dimensional settings. Further discussions on these nuances in a more general context can be found in the recent works of Chernozhukov et al. (2016, 2017b, 2018a,b) and Chernozhukov, Newey and Robins (2018) on the use of *Neyman orthogonal* scores for semi-parametric inference in the presence of (unknown) high dimensional nuisance functions.

Finally, note that all three identifications above are fully non-parametric. They follow from simple uses of Assumption 1.1 (and iterated expectations) and require no further assumptions on $\mathbb{P}$. The nuisance functions $\pi(\mathbf{X})$ and $\phi(\mathbf{X},\boldsymbol{\theta})$ are both estimable from the observed data. $\pi(\mathbf{X})$ is estimable from the data on $(T,\mathbf{X})$, while under MAR, $\phi(\mathbf{X},\boldsymbol{\theta}) = \mathbb{E}\{L(Y,\mathbf{X},\boldsymbol{\theta})|\mathbf{X},T=1)$ is estimable from the 'complete case' data. Note that in some cases, $\phi(\mathbf{X},\boldsymbol{\theta})$

may itself involve $\mathbb{E}(Y\,|\,\mathbf{X})$. While the latter may sometimes also 'motivate' the definition of $\boldsymbol{\theta}_0$ in (1.2), as in regression problems based on parametric (working) models for $\mathbb{E}(Y\,|\,\mathbf{X})$, this should *not* be confused in any way with its role as a nuisance function in the identifications of $\mathbb{L}(\cdot)$ above. In fact, it plays the *same* role as a nuisance function here as it does for the special case of the mean/ATE estimation problem, where this role (and its importance) is very well understood and it is common practice to estimate these nuisance functions and use them to implement the DR type estimators. We emphasize that the same principle (and practice) continue to apply here for the general problem (1.2) and it should not be confused with the other (unrelated) issue.

2.2. *Simplifying Structural Assumptions.* For simplicity, we shall assume henceforth a structure on the derivative of $L(Y, \mathbf{X}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ as follows. For some functions $\mathbf{h}(\mathbf{X}) \in \mathbb{R}^d$ and $g(\mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}$, we assume it takes the form:

$$\boldsymbol{\nabla} L(Y, \mathbf{X}, \boldsymbol{\theta}) \;\equiv\; \frac{\partial}{\partial \boldsymbol{\theta}} L(Y, \mathbf{X}, \boldsymbol{\theta}) \;=\; \mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}. \tag{2.3}$$

The structural assumption in (2.3) is mostly for simplicity in the theoretical analyses of our proposed estimator. This form is satisfied by most standard loss functions used in practice, including the examples given in Section 1.2. Extensions of our results to loss functions with more general structures may also be obtained easily albeit at the cost of less tractable technical conditions.

Under (2.3), the loss function $L(Y, \mathbf{X}, \boldsymbol{\theta})$ therefore takes the form:

$$L(Y, \mathbf{X}, \boldsymbol{\theta}) \;=\; \{\mathbf{h}(\mathbf{X})'\boldsymbol{\theta}\}Y - f(\mathbf{X}, \boldsymbol{\theta}) + C(Y, \mathbf{X}), \;\; \text{where} \tag{2.4}$$

$f(\mathbf{X}, \boldsymbol{\theta})$ is the anti-derivative of $\mathbf{h}(\mathbf{X})g(\mathbf{X}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ and $C(Y, \mathbf{X})$ is some function independent of $\boldsymbol{\theta}$, e.g. $C(Y, \mathbf{X}) := Y^2$ for the squared loss. Hence, $\phi(\mathbf{X}, \boldsymbol{\theta}) = \{\mathbf{h}(\mathbf{X})'\boldsymbol{\theta}\}\mathbb{E}(Y\,|\,\mathbf{X}) - f(\mathbf{X}, \boldsymbol{\theta}) + m_C(\mathbf{X})$ is convex and differentiable, where $m_C(\mathbf{X}) := \mathbb{E}\{C(Y, \mathbf{X})|\mathbf{X}\}$, and $\boldsymbol{\nabla}\phi(\mathbf{X}, \boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}}\phi(\mathbf{X}, \boldsymbol{\theta})$ is given by:

$$\boldsymbol{\nabla}\phi(\mathbf{X}, \boldsymbol{\theta}) \;=\; \mathbf{h}(\mathbf{X})\{m(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta})\}, \;\; \text{where} \;\; m(\mathbf{X}) \;:=\; \mathbb{E}(Y\,|\,\mathbf{X}). \tag{2.5}$$

Thus, given any estimates $\{\widehat{m}(\mathbf{X}), \widehat{m}_C(\mathbf{X})\}$ of $\{m(\mathbf{X}), m_C(\mathbf{X})\}$, one can estimate $\phi(\mathbf{X}, \boldsymbol{\theta})$ as: $\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta}) := \{\mathbf{h}(\mathbf{X})'\boldsymbol{\theta}\}\widehat{m}(\mathbf{X}) - f(\mathbf{X}, \boldsymbol{\theta}) + \widehat{m}_C(\mathbf{X})$. Further, $\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta})$ is also convex and differentiable in $\boldsymbol{\theta}$ and we have:

$$\boldsymbol{\nabla}\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta}) \;:=\; \frac{\partial}{\partial \boldsymbol{\theta}}\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta}) \;=\; \mathbf{h}(\mathbf{X})\{\widehat{m}(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta})\}. \tag{2.6}$$

Note that to compute $\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta})$ explicitly, one needs both the estimates $\widehat{m}(\cdot)$ and $\widehat{m}_C(\cdot)$. However, the part of $\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta})$ involving $\widehat{m}_C(\cdot)$ is *free* of $\boldsymbol{\theta}$. Our proposed estimator of $\boldsymbol{\theta}$ in Section 2.3 is constructed based on an L_1-regularized minimization (w.r.t. $\boldsymbol{\theta}$) of an objective function involving $\widehat{\phi}(\cdot)$, whereby only its gradient $\boldsymbol{\nabla}\widehat{\phi}(\mathbf{X}, \boldsymbol{\theta})$ is of interest and that depends only on $\widehat{m}(\mathbf{X})$ due to (2.6). Thus, the part of $\widehat{\phi}(\cdot)$ involving $\widehat{m}_C(\cdot)$ may be ignored for all practical implementation purposes wherein we *only* require an estimator $\widehat{m}(\cdot)$ of $m(\cdot)$ and an arbitrary choice of $\widehat{m}_C(\cdot)$ to plug in and obtain the estimator $\widehat{\phi}(\cdot)$.

2.3. *The L_1-Regularized DDR Estimator.* Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be *any* reasonable estimators of $\{\pi(\cdot), m(\cdot)\}$, and we assume that $\widehat{\pi}(\cdot)$ is obtained solely from the data $\{(T_i, \mathbf{X}_i)\}_{i=1}^n$ (see Appendix F for more discussions). Let $\widehat{\phi}(\cdot, \cdot)$ be the corresponding estimator of $\phi(\cdot, \cdot)$ based on $\widehat{m}(\cdot)$. We use sample splitting to further construct 'cross-fitted' versions of $\widehat{m}(\cdot)$ and $\widehat{\phi}(\cdot, \cdot)$ as follows.

Cross-fitted versions of $\widehat{m}(\cdot)$ and $\widehat{\phi}(\cdot, \cdot)$ based on sample splitting. Let $\{\mathcal{D}_n^{(1)}, \mathcal{D}_n^{(2)}\}$ denote a random partition (or split) of the original data $\mathcal{D}_n$ into $\mathbb{K} = 2$ equal parts of size $\bar{n} := n/2$, where we assume w.l.o.g. that n is even. Further, let $\mathcal{I}_1$ and $\mathcal{I}_2$ respectively denote the index sets for the observations in $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$. Hence, we have $\bigcup_{k=1}^{\mathbb{K}} \mathcal{I}_k = \mathcal{I} := \{1, \ldots, n\}$ and $\bigcup_{k=1}^{\mathbb{K}} \mathcal{D}_n^{(k)} = \mathcal{D}_n$.

Given any general procedure for obtaining $\widehat{m}(\cdot)$ and $\widehat{\phi}(\cdot, \cdot)$ based on the full observed data $\mathcal{D}_n$, let $\{\widehat{m}^{(1)}(\cdot), \widehat{\phi}^{(1)}(\cdot, \cdot)\}$ and $\{\widehat{m}^{(2)}(\cdot), \widehat{\phi}^{(2)}(\cdot, \cdot)\}$ denote the corresponding versions of these estimators based on $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$ respectively. Then, we define the *cross-fitted* estimates $\{\widetilde{m}(\mathbf{X}_i), \widetilde{\phi}(\mathbf{X}_i, \boldsymbol{\theta})\}_{i=1}^n$ of $\{m(\mathbf{X}_i), \phi(\mathbf{X}_i, \boldsymbol{\theta})\}_{i=1}^n$ at the n training points in $\mathcal{D}_n$ as follows:

$$\{\widetilde{m}(\mathbf{X}_i), \widetilde{\phi}(\mathbf{X}_i, \boldsymbol{\theta})\} = \begin{cases} \{\widehat{m}^{(2)}(\mathbf{X}_i), \widehat{\phi}^{(2)}(\mathbf{X}_i, \boldsymbol{\theta})\} & \forall\, i \in \mathcal{I}_1, \quad \text{and} \\ \{\widehat{m}^{(1)}(\mathbf{X}_i), \widehat{\phi}^{(1)}(\mathbf{X}_i, \boldsymbol{\theta})\} & \forall\, i \in \mathcal{I}_2. \end{cases} \tag{2.7}$$

A detailed discussion regarding the benefits (and virtual necessity) of considering these cross-fitted estimators is given in Appendix F. Further insights regarding the benefits of cross-fitting for general semi-parametric estimation problems in the presence of nuisance components can also be found in Chernozhukov et al. (2016, 2018a,b) and Newey and Robins (2018). However, note also that we do *not* require sample splitting for constructing the estimates $\{\widehat{\pi}(\mathbf{X}_i)\}_{i=1}^n$ as long as $\widehat{\pi}(\cdot)$ is obtained only from the data on $\{(T_i, \mathbf{X}_i)\}_{i=1}^n$.

REMARK 2.1. While we focus here on the simple case of sample splitting with $\mathbb{K} = 2$, our notations and analyses are designed to easily accommodate the general case of $\mathbb{K}$-fold cross fitting for any fixed $\mathbb{K} \geq 2$. We stick to $\mathbb{K} = 2$ for simplicity and brevity of our arguments. Finally, note that the estimator $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ obtained via this cross-fitting procedure can also be replicated several times over different splittings of $\mathcal{D}_n$, and then suitably combined over these replications to average out the (minor) randomness due to sample splitting.

The estimator. Recall the DDR representation of the expected loss $\mathbb{L}(\boldsymbol{\theta})$:

$$\mathbb{L}_{\mathrm{DDR}}(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X};\boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi(\mathbf{X})}\left\{L(Y,\mathbf{X},\boldsymbol{\theta}) - \phi(\mathbf{X};\boldsymbol{\theta})\right\}\right],$$

and define its empirical version, based on the estimates $\{\widetilde{\phi}(\mathbf{X},\boldsymbol{\theta}), \widehat{\pi}(\mathbf{X}_i)\}_{i=1}^n$ plugged in, as follows. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, let us define the *empirical DDR loss*

$$\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) \;:=\; \frac{1}{n}\sum_{i=1}^n \widetilde{\phi}(\mathbf{X}_i,\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)}\left\{L(Y_i,\mathbf{X}_i,\boldsymbol{\theta}_i) - \widetilde{\phi}(\mathbf{X}_i,\boldsymbol{\theta})\right\}. \tag{2.8}$$

With $\boldsymbol{\theta}_0$ (and $\mathbf{X}$) possibly high dimensional, we shall need to assume that $\boldsymbol{\theta}_0$ is sparse with sparsity much smaller than d when $d \gg n$. In general, we denote the sparsity of $\boldsymbol{\theta}_0$ as $s := \|\boldsymbol{\theta}_0\|_0$ with $1 \leq s \leq d$. We now propose to estimate $\boldsymbol{\theta}_0$ using the L_1*-regularized DDR estimator*, $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$, given by:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} \;\equiv\; \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) \;=\; \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\;\left\{\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) + \lambda_n\|\boldsymbol{\theta}\|_1\right\}, \tag{2.9}$$

where $\mathcal{L}_n^{\mathrm{DDR}}(\cdot)$ is as in (2.8) and $\lambda_n \geq 0$ denotes the regularization (or tuning) parameter. (For a classical setting with $d \ll n$, λ_n may be set to 0 if desired).

2.4. *Simple Algorithm for Implementation.* The estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ in (2.9) can be implemented using a simple user-friendly imputation type algorithm.

Given the observed data $\mathcal{D}_n$ and the estimates $\{\widehat{\pi}(\mathbf{X}_i), \widetilde{m}(\mathbf{X}_i)\}_{i=1}^n$, define a set of *pseudo outcomes* $\{\widetilde{Y}_i\}_{i=1}^n$ and the *pseudo loss* $\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ as follows:

$$\widetilde{Y}_i := \widetilde{m}(\mathbf{X}_i) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)}\{Y_i - \widetilde{m}(\mathbf{X}_i)\} \text{ and } \widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n L(\widetilde{Y}_i,\mathbf{X}_i,\boldsymbol{\theta}). \tag{2.10}$$

Clearly $\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\cdot)$ is convex and differentiable, and under (2.3)-(2.6), it is easy to see that $\boldsymbol{\nabla}\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) = \boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$, where for any $f(\cdot)$, $\boldsymbol{\nabla}f(\boldsymbol{\theta}) := \frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{\theta})$.

Further, observe that the solution for the minimization in (2.9) is uniquely determined by the underlying normal equations (the KKT conditions) which *only* depend on the gradient of $\mathcal{L}_n^{\mathrm{DDR}}(\cdot)$ and the subgradient of $\|\cdot\|_1$. Hence, the solution stays *unchanged* if $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ in (2.9) is replaced by $\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ which has the same gradient. Consequently, $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ in (2.9) may also be defined as:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} \;\equiv\; \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) \;:=\; \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\{\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) + \lambda_n\|\boldsymbol{\theta}\|_1\}. \tag{2.11}$$

Thus, if one 'pretends' to have a fully observed data $\widetilde{\mathcal{D}}_n := \{(\widetilde{Y}_i,\mathbf{X}_i)\}_{i=1}^n$ in terms of the pseudo outcomes $\widetilde{Y}_i$, then $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ can be simply obtained by a

L_1-penalized minimization of the corresponding empirical risk for $L(\cdot)$ based on $\widetilde{\mathcal{D}}_n$. This minimization is quite straightforward to implement and can be done so using standard statistical software packages (e.g. `'glmnet'` in `R`).

Note also that (2.11) confirms our earlier claim that although the estimator $\widetilde{\phi}(\mathbf{X}, \boldsymbol{\theta})$ involved in the definition (2.8) of $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ may require estimation of other nuisance functions (independent of $\boldsymbol{\theta}$) apart from $m(\mathbf{X})$, the implementation of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ via the minimization in (2.9), or equivalently the one in (2.11), requires *only* an estimator of $m(\mathbf{X})$, along with that of $\pi(\mathbf{X})$.

2.5. *Performance Guarantees: Deviation Bounds.* We next provide a *deterministic* deviation bound regarding the finite sample performance of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ that serves as the backbone for most of our main theoretical analyses. We begin with an assumption. Recall the notations introduced in Section 2.

ASSUMPTION 2.1 (Restricted strong convexity). We assume that the loss function $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ satisfies a restricted strong convexity (RSC) property at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, as follows: $\exists$ a (non-random) constant $\kappa_{\mathrm{DDR}} > 0$ such that

$$
\begin{aligned}
\delta\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0; \mathbf{v}) \ &\geq \ \kappa_{\mathrm{DDR}} \|\mathbf{v}\|_2^2 \quad \forall\, \mathbf{v} \in \mathbb{C}(\boldsymbol{\theta}_0), \quad \text{where } \forall\, \boldsymbol{\theta}, \mathbf{v} \in \mathbb{R}^d, \\
\delta\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}; \mathbf{v}) \ &:= \ \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta} + \mathbf{v}) - \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) - \mathbf{v}'\{\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})\} \\
\text{and} \quad \mathbb{C}(\boldsymbol{\theta}_0) \ &:= \ \{\mathbf{v} \in \mathbb{R}^d : \ \|\Pi_{\boldsymbol{\theta}_0}^c(\mathbf{v})\|_1 \leq 3\|\Pi_{\boldsymbol{\theta}_0}(\mathbf{v})\|_1\} \ \subseteq \ \mathbb{R}^d.
\end{aligned}
\tag{2.12}
$$

Assumption 2.1, largely adopted from Negahban et al. (2012), is one of the several restricted eigenvalue type assumptions that are standard in the high dimensional statistics literature. While we assume (2.12) deterministically for any realization of $\mathcal{D}_n$, it can be relaxed with appropriate modifications to only hold with high probability (w.h.p.). It is important to note that owing to the very structure of $\mathcal{L}_n^{\mathrm{DDR}}(\cdot)$ in (2.8) and the assumed structures in (2.3)-(2.6) for $L(\cdot)$ and $\widetilde{\phi}(\cdot)$, the RSC condition (2.12) is completely *independent* of the quantities depending on the missingness aspect of the problem, i.e. $\delta\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0; \mathbf{v})$ in (2.12) is independent of $\{T_i, Y_i\}_{i=1}^n$ as well as the nuisance function estimates $\{\widehat{\pi}(\mathbf{X}_i), \widetilde{m}(\mathbf{X}_i)\}_{i=1}^n$. In fact, it is the *same* as the corresponding version one would obtain in the case of a fully observed data. This fact also follows from the alternative definition of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ in (2.11) based on the pseudo outcomes and the pseudo loss $\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\cdot)$ in (2.10). Thus, verifying (2.12) is *equivalent* to verifying the same for a fully observed data which is quite well studied (Negahban et al., 2012; Rudelson and Zhou, 2013; Lecué and Mendelson, 2014; Kuchibhotla and Chakrabortty, 2018; Vershynin, 2018) for several standard problems under fairly mild conditions. This thereby provides an easy route to verifying the RSC condition (2.12) under our setting.

LEMMA 2.1 (Deterministic deviation bounds for $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$). *Assume $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ and satisfies the form (2.3). Let Assumption 2.1 hold, with $\kappa_{DDR} > 0$ as defined therein, and recall that $s := \|\boldsymbol{\theta}_0\|_0$. Then, for any realization of $\mathcal{D}_n$ and for any choice of $\lambda \equiv \lambda_n \geq 2\,\|\boldsymbol{\nabla}\mathcal{L}_n^{DDR}(\boldsymbol{\theta}_0)\|_\infty$,*

$$\|\widehat{\boldsymbol{\theta}}_{DDR} - \boldsymbol{\theta}_0\|_2 \;\leq\; 3\sqrt{s}\frac{\lambda_n}{\kappa_{DDR}} \;\; \textit{and} \;\; \|\widehat{\boldsymbol{\theta}}_{DDR} - \boldsymbol{\theta}_0\|_1 \;\leq\; 12s\frac{\lambda_n}{\kappa_{DDR}}. \tag{2.13}$$

Convergence rates (informal statement). *We establish via Theorems 3.1-3.4 later that under suitable assumptions (given in Section 3.2), $\|\boldsymbol{\nabla}\mathcal{L}_n^{DDR}(\boldsymbol{\theta}_0)\|_\infty \lesssim \sqrt{(\log d)/n}$ w.h.p. Hence, choosing $\lambda \equiv \lambda_n \asymp \sqrt{(\log d)/n}$, it follows that*

$$\|\widehat{\boldsymbol{\theta}}_{DDR} - \boldsymbol{\theta}_0\|_2 \;\lesssim\; \sqrt{\frac{s\log d}{n}} \;\; \textit{and} \;\; \|\widehat{\boldsymbol{\theta}}_{DDR} - \boldsymbol{\theta}_0\|_1 \;\lesssim\; s\sqrt{\frac{\log d}{n}} \;\; \textit{w.h.p.}$$

The deviation bounds (2.13), essentially an easy consequence of the results of Negahban et al. (2012), deterministically relate the L_2 and L_1 error rates of the estimator to the chosen λ_n and provides an easy recipe for establishing its convergence rates by studying the same for the (random) lower bound of λ_n given in Lemma 2.1. This is the main goal of Section 3, where we obtain sharp non-asymptotic upper bounds for $\|\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ converging to 0 at satisfactory rates w.h.p. A choice of λ_n of the order of this bound guarantees the requirement of $\lambda_n \geq 2\,\|\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ in Lemma 2.1 to hold w.h.p. and establishes the convergence rates, defined by the λ_n, for the bounds in (2.13).

Finally, note also that the (informal) bounds in the second part of Lemma 2.1 establish the obvious *rate optimality* of the estimator since it matches the (well known) optimal estimation error rate for a fully observed data.

3. The Main Results for the DDR Estimator: Convergence Rate and Probabilistic Bounds for $\left\|\boldsymbol{\nabla}\mathcal{L}_{\boldsymbol{n}}^{\mathbf{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty$. For most of our theoretical analyses of $\|\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$, we will assume that $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ are both correctly specified estimators of $\{\pi(\cdot), m(\cdot)\}$. The analysis even for this case is involved (and necessarily non-asymptotic) due to the presence of the nuisance function estimators and the inherent high dimensional setting.

Under possible misspecification of one of the estimators, the DR property (in terms of consistency) of $\|\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ and that of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n)$, for a suitably chosen λ_n under Lemma 2.1, indeed follows due to the very nature of construction of $\mathcal{L}_n^{\mathrm{DDR}}(\cdot)$ and its population version $\mathbb{L}_{\mathrm{DDR}}(\cdot)$ outlined in (2.1)-(2.2). This DR property is well known in classical settings (Robins, Rotnitzky and Zhao, 1994; Robins and Rotnitzky, 1995; Bang and Robins, 2005) and should also be expected to hold in high-dimensional settings under suitable conditions. We discuss these DR properties further in Appendix A.

One of the reasons behind considering the DDR representations $\mathbb{L}_{\text{DDR}}(\boldsymbol{\theta})$ and $\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta})$ is that apart from the obvious benefits of double robustness, the DDR loss has a naturally 'debiased' form that provides *crucial* technical benefits in controlling the associated error terms which are naturally 'centered' (in a certain sense) when both $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ are correctly specified, a setting where other approaches such as IPW and REG type estimators are also applicable in principle, but they don't enjoy such technical benefits. The advantages of such debiased representations, especially in high dimensional settings, have also been studied in a more general context under the name of *Neyman orthogonalization* in the recent works of Chernozhukov et al. (2016, 2017b, 2018a,b) and Chernozhukov, Newey and Robins (2018). The DDR representation indeed (naturally) satisfies such an 'orthogonal' structure.

3.1. *The Basic Decomposition.* Let $\mathbf{T}_n := \boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) \in \mathbb{R}^d$ with $\|\mathbf{T}_n\|_\infty$ being our quantity of interest. We first note a decomposition of $\mathbf{T}_n$ as follows.

$$\begin{aligned}\mathbf{T}_n \;&=\; \mathbf{T}_{0,n} + \mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n} \\ &:= \frac{1}{n}\sum_{i=1}^n \mathbf{T}_0(\mathbf{Z}_i) + \frac{1}{n}\sum_{i=1}^n \mathbf{T}_\pi(\mathbf{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbf{T}_m(\mathbf{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbf{R}_{\pi,m}(\mathbf{Z}_i), \qquad (3.1)\end{aligned}$$

where $\mathbf{T}_0(\mathbf{Z})$, $\mathbf{T}_\pi(\mathbf{Z})$, $\mathbf{T}_m(\mathbf{Z})$ and $\mathbf{R}_{\pi,m}(\mathbf{Z})$ with $\mathbf{Z} = (T, \mathbb{Y}, \mathbf{X})$ are given by:

$$\mathbf{T}_0(\mathbf{Z}) := \{m(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}_0)\}\mathbf{h}(\mathbf{X}) + \frac{T}{\pi(\mathbf{X})}\{Y - m(\mathbf{X})\}\mathbf{h}(\mathbf{X}) \qquad (3.2)$$

$$\mathbf{T}_\pi(\mathbf{Z}) := \left\{\frac{T}{\widehat{\pi}(\mathbf{X})} - \frac{T}{\pi(\mathbf{X})}\right\}\{Y - m(\mathbf{X})\}\mathbf{h}(\mathbf{X}), \qquad (3.3)$$

$$\mathbf{T}_m(\mathbf{Z}) := \left\{\frac{T}{\pi(\mathbf{X})} - 1\right\}\{\widetilde{m}(\mathbf{X}) - m(\mathbf{X})\}\mathbf{h}(\mathbf{X}), \quad \text{and} \qquad (3.4)$$

$$\mathbf{R}_{\pi,m}(\mathbf{Z}) := \left\{\frac{T}{\widehat{\pi}(\mathbf{X})} - \frac{T}{\pi(\mathbf{X})}\right\}\{\widetilde{m}(\mathbf{X}) - m(\mathbf{X})\}\mathbf{h}(\mathbf{X}). \qquad (3.5)$$

In the decomposition (3.1), $\mathbf{T}_{0,n}$ denotes the leading (first order) term, while $\mathbf{T}_{\pi,n}$ and $\mathbf{T}_{m,n}$ denote the main error terms accounting for the estimation errors of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ respectively, and $\mathbf{R}_{\pi,m,n}$ is a second order bias term involving the product of the estimation errors of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$.

Summary of results. We control $\|\mathbf{T}_n\|_\infty$ by separately controlling $\|\mathbf{T}_{0,n}\|_\infty$, $\|\mathbf{T}_{\pi,n}\|_\infty$, $\|\mathbf{T}_{m,n}\|_\infty$ and $\|\mathbf{R}_{\pi,m,n}\|_\infty$ through Theorems 3.1-3.4. The results show that the convergence rate of $\|\mathbf{T}_n\|_\infty$ is determined primarily by that of the leading term $\|\mathbf{T}_{0,n}\|_\infty$ while the rates of the other three terms are of a (faster) lower order. In particular, we show that under suitable assumptions,

$$\|\mathbf{T}_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \text{ and } \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty + \|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} o(1)$$

w.h.p. The results (proved in Appendices E-I) are all non-asymptotic (with precise constants) and involve careful analyses via concentration inequalities to account for the nuisance function estimators and the high dimensionality.

REMARK 3.1 (Generality of the results). It is important to note that our results here are completely *free* in terms of choice of the nuisance function estimators. The analysis and the convergence rates are *first order insensitive* to any estimation errors of the nuisance functions and hold *regardless* of any knowledge of the construction and/or first order properties of the estimators, as long as they satisfy some basic high-level conditions on their convergence rates. This is also largely an artifact of the debiased form of the DDR loss.

3.2. *The Assumptions Required.* We first summarize the main assumptions required for controlling the various terms in (3.1). We begin with a few standard assumptions on the tail behaviors of some key random variables.

ASSUMPTION 3.1 (Sub-Gaussian tail behaviors). (a) We assume that $\varepsilon(\mathbb{Z}) := Y - m(\mathbf{X})$, $\psi(\mathbf{X}) := m(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}_0)$ and $\mathbf{h}(\mathbf{X})$ are sub-Gaussian (as per Definition C.1 with $\alpha = 2$ therein) with $\|\varepsilon\|_{\psi_2} \le \sigma_\varepsilon$, $\|\psi(\mathbf{X})\|_{\psi_2} \le \sigma_\psi$ and $\|\mathbf{h}(\mathbf{X})\|_{\psi_2} \le \sigma_{\mathbf{h}}$ for some constants $\sigma_\varepsilon, \sigma_\psi, \sigma_{\mathbf{h}} \ge 0$.

(b) For controlling $\mathbf{T}_{\pi,n}$, we additionally assume that $\{\varepsilon(\mathbb{Z})|\mathbf{X}\}$ is (conditionally) sub-Gaussian with $\|\varepsilon(\mathbb{Z})|\mathbf{X}\|_{\psi_2} \le \sigma_\varepsilon(\mathbf{X})$ for some function $\sigma_\varepsilon(\cdot) \ge 0$ such that $\|\sigma_\varepsilon(\cdot)\|_\infty \le \sigma_\varepsilon < \infty$ with σ_ε being as in part (a) above.

Next, we discuss the basic high-level conditions we require regarding the behavior and convergence rates of the nuisance function estimators $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$. Further discussions on the assumptions are given in Remarks 3.2-3.4.

ASSUMPTION 3.2 (Tail bounds on the pointwise behavior of $\widehat{\pi}(\cdot) - \pi(\cdot)$). We assume that $\widehat{\pi}(\cdot)$ is obtained solely from the data $\mathcal{X}_n := \{(T_i, \mathbf{X}_i)\}_{i=1}^n \subseteq \mathcal{D}_n$, and for some sequences $v_{n,\pi} \ge 0$ with $v_{n,\pi} = o(1)$ and $q_{n,\pi} \in [0,1]$ with $q_{n,\pi} = o(1)$, $\widehat{\pi}(\cdot) - \pi(\cdot)$ satisfies a (pointwise) tail bound at the n training points $\{\mathbf{X}_i\}_{i=1}^n$ as follows: for any $t \ge 0$ and for some constant $C \ge 0$,

$$\mathbb{P}\{|\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)| > t v_{n,\pi}\} \le C \exp(-t^2) + q_{n,\pi} \ \ \forall\, 1 \le i \le n, \tag{3.6}$$

and we further assume that $v_{n,\pi}\sqrt{\log(nd)} = o(1)$ and $q_{n,\pi} = o(n^{-1}d^{-1})$.

ASSUMPTION 3.3 (Pointwise tail bounds on $\widehat{m}(\cdot) - m(\cdot)$). For a generic version of $\widehat{m}(\cdot)$ obtained from a data of size n (e.g. $\mathcal{D}_n$), we assume that for some sequences $v_{n,m} \geq 0$ with $v_{n,m} = o(1)$ and $q_{n,m} \in [0,1]$ with $q_{n,m} = o(1)$, $\widehat{m}(\cdot) - m(\cdot)$ satisfies a (pointwise) tail bound at any fixed $\mathbf{x} \in \mathcal{X}$ as follows: for any $t \geq 0$ and for some constant $C > 0$,

$$\mathbb{P}\{|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| > t v_{n,m}\} \leq C \exp(-t^2) + q_{n,m}, \text{ so that} \tag{3.7}$$

$$\mathbb{P}\{|\widehat{m}^{(k)}(\mathbf{X}_i) - m(\mathbf{X}_i)| > t v_{\bar{n},m}\} \leq C \exp(-t^2) + q_{\bar{n},m}, \ \forall\, k = 1, 2 \tag{3.8}$$

and $\mathbf{X}_i \in \mathcal{D}_n^{(k')} \perp\!\!\!\perp \mathcal{D}_n^{(k)}$ with $k' \neq k \in \{1, 2\}$, where $\bar{n} := n/2$ and $\widehat{m}^{(k)}(\cdot)$ denotes the version of $\widehat{m}(\cdot)$ obtained from $\mathcal{D}_n^{(k)}$ with size $\bar{n} \equiv n/2$. Further, we assume that $v_{\bar{n},m}\sqrt{\log(nd)} = o(1)$ and $q_{\bar{n},m} = o(n^{-1}d^{-1})$.

REMARK 3.2. Assumptions 3.2 and 3.3 are both fairly mild and general (high-level) conditions that should be expected to hold for most reasonable estimators $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ of $\{\pi(\cdot), m(\cdot)\}$. Note that (3.6), (3.7) and (3.8) are all conditions on the *pointwise* behaviors of $\widehat{\pi}(\cdot) - \pi(\cdot)$ and $\widehat{m}(\cdot) - m(\cdot)$, and do *not* require any uniform tail bounds over all $\mathbf{x} \in \mathcal{X}$, such as bounds on the L_∞ or L_2 errors of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Such conditions are much stronger and also generally harder to verify in high dimensional settings. We simply require pointwise tail bounds for the errors $\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)$ and $\widehat{m}(\mathbf{x}) - m(\mathbf{x})$, ensuring that they have well-behaved tails. The sequences $\{v_{n,\pi}, v_{n,m}\}$ indicate the convergence rates of the estimators, while $\{q_{n,\pi}, q_{n,m}\}$ in the probability bounds allow to rigourously account for potential lower order terms.

REMARK 3.3 (Sufficient conditions for Assumptions 3.2 and 3.3). In general, for any estimator $\widehat{\pi}(\cdot)$ satisfying a high probability guarantee of the form: $|\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)| \leq v_n$ with probability at least $1 - q_n$, the bound (3.6) can be shown to hold with $\{v_{n,\pi}, q_{n,\pi}\} \propto \{v_n, q_n\}$, through a simple use of Hoeffding's inequality (see Lemma C.7 in this regard). Similarly, for any estimator $\widehat{m}(\cdot)$ satisfying a high probability bound: $|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \leq v_n$ with probability at least $1 - q_n$, the bounds (3.7)-(3.8) can be shown to hold with $\{v_{n,m}, q_{n,m}\} \propto \{v_n, q_n\}$. These high probability bounds are expected to be satisfied by most reasonable estimators and hence, so are our assumptions.

REMARK 3.4 (Examples). In Section 5, we discuss several choices of the estimators $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ based on parametric families, 'extended' parametric families (series estimators) and semi-parametric single index families. For all these estimators, we establish precise tail bounds (see Theorems 5.1, 5.2 and 5.3) that are generally useful and should be of independent interest. Among other implications, they also verify the bounds in Assumptions 3.2 and 3.3.

3.3. *Controlling the Leading Order Term.* The following result quantifies the behavior and convergence rate of the first order term $\|\mathbf{T}_{0,n}\|_\infty$ in (3.1).

THEOREM 3.1 (Control of $\|\mathbf{T}_{0,n}\|_\infty$). *Under Assumptions 1.1 and 3.1 (a),*

$$\mathbb{P}\left(\|\mathbf{T}_{0,n}\|_\infty > \sqrt{2}\sigma_0\epsilon + K_0\epsilon^2\right) \;\le\; 4\exp\left(-n\epsilon^2 + \log d\right) \quad \textit{for any } \epsilon \ge 0,$$

where $\sigma_0 := 2\sqrt{2}\sigma_{\mathbf{h}}(\sigma_\psi + \sigma_\varepsilon\delta_\pi^{-1})$, $K_0 := 2\sigma_{\mathbf{h}}(\sigma_\psi + \sigma_\varepsilon\delta_\pi^{-1})$ *and* $(\delta_\pi, \sigma_\varepsilon, \sigma_{\mathbf{h}}, \sigma_\psi)$ *are as defined in the assumptions. In particular, setting* $\epsilon = c\sqrt{(\log d)/n}$ *for any constant* $c > 1$*, we have: with probability at least* $1 - 4d^{-(c^2-1)}$,

$$\|\mathbf{T}_{0,n}\|_\infty \;\le\; c\sqrt{\frac{\log d}{n}}\sqrt{2}\sigma_0 + c^2\frac{\log d}{n}K_0 \;\lesssim\; \sqrt{\frac{\log d}{n}}.$$

3.4. *Controlling the Error Term from the Propensity Score's Estimation.* Next, we propose the following result to control the error term $\mathbf{T}_{\pi,n}$ in (3.1).

THEOREM 3.2 (Control of $\|\mathbf{T}_{\pi,n}\|_\infty$). *Let Assumptions 1.1, 3.1 and 3.2 hold with* $(v_{n,\pi}, q_{n,\pi})$ *and* $(\delta_\pi, \sigma_\varepsilon, \sigma_{\mathbf{h}}, C)$ *as defined therein. Then, for any constants* $c, c_1, c_2, c_3 > 1$*, where we assume* $c_2 v_{n,\pi}\sqrt{\log(nd)} \le \delta_\pi/2 < \delta_\pi$ *and* $c_3\sqrt{(\log d)/n} < 1$ *w.l.o.g., we have: with probability at least* $1 - 2d^{-(c^2-1)} - 4d^{-(c_3^2-1)} - 2C(nd)^{-(c_1^2-1)} - 2C(nd)^{-(c_2^2-1)} - 4q_{n,\pi}(nd)$,

$$\|\mathbf{T}_{\pi,n}\|_\infty \;\le\; c\sqrt{\frac{\log d}{n}}\{v_{n,\pi}\sqrt{\log(nd)}\}C_1\left(\frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + C_2\sqrt{\frac{\log d}{n}}\right)^{\frac{1}{2}},$$

where $\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty := \max_{1\le j\le d}\mathbb{E}\{\mathbf{h}_{[j]}^2(\mathbf{X})\}$, $C_1 := c_1(4\sqrt{2}\sigma_\varepsilon/\delta_\pi)$ *and* $C_2 := c_3(\sqrt{2}\sigma_\pi + K_\pi)$ *with* $\sigma_\pi := 2\sqrt{2}\sigma_{\mathbf{h}}^2\delta_\pi^{-2}$ *and* $K_\pi := 2\sigma_{\mathbf{h}}^2\delta_\pi^{-2}$ *being constants.*

REMARK 3.5. Theorem 3.2 therefore shows that $\|\mathbf{T}_{\pi,n}\|_\infty \lesssim \sqrt{(\log d)/n}\,\{v_{n,\pi}\sqrt{\log(nd)}\} = o\{\sqrt{(\log d)/n}\}$ w.h.p. In the proof of Theorem 3.2, we also provide a general result (Theorem G.1) on tail bounds for $\mathbf{T}_{\pi,n}$.

3.5. *Controlling the Error Term from the Conditional Mean's Estimation.* We now control the error term $\mathbf{T}_{m,n}$ in (3.1) involving the cross-fitted estimates $\{\widetilde{m}(\mathbf{X}_i)\}_{i=1}^n$ obtained via sample splitting, through the result below.

THEOREM 3.3 (Control of $\|\mathbf{T}_{m,n}\|_\infty$). *Let Assumptions 1.1, 3.1 (a) and 3.3 hold, with* $(v_{\bar{n},m}, q_{\bar{n},m})$, $\bar{n} \equiv n/2$ *and* $(\delta_\pi, \sigma_{\mathbf{h}}, C)$ *as defined therein. Then,* *for any constants* $c, c_1, c_2 > 1$*, where we assume* $c_2\sqrt{(\log d)/\bar{n}} < 1$ *w.l.o.g., with probability at least* $1-4d^{-(c^2-1)}-8d^{-(c_2^2-1)}-4C(\bar{n}d)^{-(c_1^2-1)}-4q_{\bar{n},m}(\bar{n}d)$*,*

$$\|\mathbf{T}_{m,n}\|_\infty \ \leq \ c\sqrt{\frac{\log d}{n}}\{v_{\bar{n},m}\sqrt{\log(nd)}\}C_1^*\left(\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + C_2^*\sqrt{\frac{\log d}{n}}\right)^{\frac{1}{2}},$$

where $\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty$ *is as in Theorem 3.2,* $C_1^* := 4c_1\bar{\delta}_\pi$ *and* $C_2^* := \sqrt{2}c_2(\sqrt{2}\sigma_m + K_m)$*, with* $\sigma_m := 2\sqrt{2}\sigma_{\mathbf{h}}^2$*,* $K_m := 2\sigma_{\mathbf{h}}^2$ *and* $\bar{\delta}_\pi \leq \delta_\pi^{-1}$ *being constants.*

REMARK 3.6. Theorem 3.3 therefore shows that $\|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{(\log d)/n}$ $\{v_{\bar{n},m}\sqrt{\log(nd)}\} = o\{\sqrt{(\log d)/n}\}$ w.h.p. In the proof of Theorem 3.3, we also provide a general result (Theorem H.1) on tail bounds for $\mathbf{T}_{m,n}$.

3.6. *Controlling The Lower Order Term.* Finally, we control the second order error (or bias) term $\mathbf{R}_{\pi,m,n}$ in (3.1) through the following result.

THEOREM 3.4 (Control of $\|\mathbf{R}_{\pi,m,n}\|_\infty$). *Let Assumptions 1.1, 3.1, 3.2 and 3.3 hold with* $(v_{n,\pi}, q_{n,\pi})$*,* $(v_{\bar{n},m}, q_{\bar{n},m}, \bar{n})$ *and* (δ_π, C) *as defined therein, and assume that* $v_{n,\pi}v_{\bar{n},m}(\log n) = o\{\sqrt{(\log d)/n}\}$*. Then, for any constants* $c_1, c_2, c_3, c_4 > 1$ *with* $c_2 v_{n,\pi}\sqrt{\log n} \leq \delta_\pi/2 < \delta_\pi$ *and* $c_4\sqrt{(\log d)/n} < 1$*, we have: with probability at least* $1-\sum_{j=1}^{3} Cn^{-(c_j^2-1)}-2d^{-(c_4^2-1)}-2nq_{n,\pi}-nq_{\bar{n},m}$*,*

$$\|\mathbf{R}_{\pi,m,n}\|_\infty \ \leq \ c_1c_3\bar{C}_1\{v_{n,\pi}v_{\bar{n},m}(\log n)\}\left(\|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty + c_4\bar{C}_2\sqrt{\frac{\log d}{n}}\right), \ \ \textit{where}$$

$\|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty := \max_{1\leq j\leq d}\mathbb{E}\{|\mathbf{h}_{[j]}(\mathbf{X})|\}$ *and* $\bar{C}_1 := 2/\delta_\pi$*,* $\bar{C}_2 := \sqrt{2}\sigma_{\pi,m} + K_{\pi,m}$ *are constants with* $\sigma_{\pi,m} := 4\sigma_{\mathbf{h}}\delta_\pi^{-1}$ *and* $K_{\pi,m} := 2\sqrt{2}\sigma_{\mathbf{h}}\delta_\pi^{-1}$*.*

REMARK 3.7. Thus, Theorem 3.4 shows $\|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim v_{n,\pi}v_{\bar{n},m}(\log n) = o\{\sqrt{(\log d)/n}\}$ w.h.p. where the last step is by assumption, a sufficient condition for which is $\max\{v_{n,\pi}, v_{\bar{n},m}\}(\log n)^{1/2} \lesssim \{(\log d)/n\}^{1/4}$. Conditions of this flavor are well known and standard in the mean (or ATE) estimation literature, where they are routinely adopted to control these kind of second order (product-type) bias terms (Farrell, 2015; Chernozhukov et al., 2018a). In Theorem I.1, we provide a more general result on tail bounds for $\mathbf{R}_{\pi,m,n}$.

4. High Dimensional Inference via the DDR Estimator: Desparsification and Asymptotic Linear Expansion. We next discuss a debiasing/desparsification approach for the DDR estimator $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ which is useful for establishing an estimator with an *asymptotic linear expansion* (ALE), a property not possessed by the L_1-regularized shrinkage type estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$. Such expansions form the fundamental ingredients for high dimensional inference as they automatically lead to asymptotic normality (and hence confidence intervals, p-values, tests etc.) for low dimensional components of $\boldsymbol{\theta}_0$, thus paving the way for inference on $\boldsymbol{\theta}_0$ among many other implications. For a fully observed data (much unlike the setting we have here), such problems have received substantial attention in recent times (van de Geer et al., 2014; Javanmard and Montanari, 2014, 2018; Cai and Guo, 2017).

For simplicity, we restrict our discussion here to the case of the squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) = \{Y - \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}\}^2$ where $\boldsymbol{\Psi}(\mathbf{X}) \equiv \{\boldsymbol{\Psi}_{[j]}(\mathbf{X})\}_{j=1}^d \in \mathbb{R}^d$ denotes some basis functions (possibly high dimensional) of $\mathbf{X}$. While more general loss functions can also be handled similarly, the corresponding results and conditions can be technically more involved. We choose to skip such analyses here given the scope and content of the current work. Note that $L(\cdot)$ satisfies (2.3) with $\mathbf{h}(\mathbf{x}) = -2\boldsymbol{\Psi}(\mathbf{x})$ and $g(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\Psi}(\mathbf{x})'\boldsymbol{\theta}$. The case $\boldsymbol{\Psi}(\mathbf{X}) = (1, \mathbf{X}')'$ corresponds to standard linear regression. For convenience, let us also define:

$$\boldsymbol{\Sigma} \;:=\; \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X})'\}, \;\; \widehat{\boldsymbol{\Sigma}} \;:=\; \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{X}_i)\boldsymbol{\Psi}(\mathbf{X}_i)' \;\text{ and }\; \boldsymbol{\Omega} \;:=\; \boldsymbol{\Sigma}^{-1},$$

where we assume that $\mathbb{E}\{\|\boldsymbol{\Psi}(\mathbf{X})\|_2^2\} < \infty$ and $\boldsymbol{\Sigma}$ is positive definite, so that $\boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Omega}$ are both well-defined and well-conditioned. With $L(\cdot)$ as above, note that we have: $\mathbb{E}\{\boldsymbol{\nabla}^2 L(Y, \mathbf{X}, \boldsymbol{\theta})\} = 2\boldsymbol{\Sigma}$ and its inverse is $\frac{1}{2}\boldsymbol{\Omega}$ for any $\boldsymbol{\theta}$, where for any function $f(\boldsymbol{\theta})$, $\boldsymbol{\nabla}^2 f(\boldsymbol{\theta})$ denotes its Hessian matrix w.r.t. $\boldsymbol{\theta}$. Further, we also have: $\boldsymbol{\nabla}^2\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) = \boldsymbol{\nabla}^2\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) = 2\widehat{\boldsymbol{\Sigma}}$.

4.1. *The Desparsified DDR Estimator.* Let $\widehat{\boldsymbol{\Omega}}$ be *any* reasonable estimator of the precision matrix $\boldsymbol{\Omega}$ based on the observed data $\mathcal{D}_n$. Then, given the original L_1-regularized DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ in (2.9) or equivalently in (2.11), we define the corresponding *desparsified DDR estimator* $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ as follows.

$$\begin{aligned}(4.1)\qquad \widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}} \;&:=\; \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \frac{1}{2}\widehat{\boldsymbol{\Omega}}\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}) \;\equiv\; \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \frac{1}{2}\widehat{\boldsymbol{\Omega}}\boldsymbol{\nabla}\widetilde{\mathcal{L}}_n^{\mathrm{DDR}}(\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}})\\ &=\; \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} + \widehat{\boldsymbol{\Omega}}\frac{1}{n}\sum_{i=1}^n\{\widetilde{Y}_i - \boldsymbol{\Psi}(\mathbf{X}_i)'\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}\}\boldsymbol{\Psi}(\mathbf{X}_i), \;\text{ where}\end{aligned}$$

$\widetilde{Y}_i \equiv \widetilde{m}(\mathbf{X}_i) + \{T_i/\widetilde{\pi}(\mathbf{X}_i)\}\{Y_i - \widetilde{m}(\mathbf{X}_i)\}$ are the *pseudo outcomes* as in (2.10).

The desparsification step in (4.1) is similar in spirit to that of van de Geer et al. (2014), while accounting for a more general and complex setting here involving missing responses. It serves as the appropriate *generalization* of their approach when adapted to this setting. As seen from the representation in the final step, the debiasing step *still* uses the full data *but* with the pseudo outcomes $\widetilde{Y}_i$ instead of the true Y_i. For a fully observed data with $\widetilde{Y}_i = Y_i$, this indeed reduces to the usual Debiased Lasso estimator of Javanmard and Montanari (2014). In addition, we also allow for misspecified models, non-Gaussian settings and covariate transformations, unlike most of the relevant existing literature (with the exception of Bühlmann and van de Geer (2015)).

It should be noted that the principle of debiasing has also been used extensively in the classical semi-parametric inference literature, where it is often called *one-step update* (Van der Vaart, 2000) and is used to obtain efficient estimators starting from an initial (inefficient) estimator. In our setting, the 'update' is used more as a bias correction to obtain an estimator with an ALE starting from a shrinkage estimator that has no such desirable properties. In classical settings, such ALEs are also known as *Bahadur representations.*

Choice of $\widehat{\mathbf{\Omega}}$*.* Since the debiasing still involves the full data (with the pseudo outcomes), the estimator $\widehat{\mathbf{\Omega}}$ is exactly the *same* as that used for a standard fully observed data. This is again largely due to the structure of the DDR loss (and the debiasing term therein). Consequently, one pays no price for the missing outcomes as far as the estimation of $\mathbf{\Omega}$ and the associated conditions are concerned, and can borrow any standard precision matrix estimator from the literature. Several such examples exist depending on the setting (low or high dimensional). In the former case, one can simply choose $\widehat{\mathbf{\Sigma}}^{-1}$, while for the latter, under sparsity assumptions on $\mathbf{\Omega}$, one can use the Nodewise Lasso estimator of van de Geer et al. (2014), among other choices. For our results on $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$, we only assume some high-level conditions on $\{\widehat{\mathbf{\Omega}}, \mathbf{\Omega}\}$ and one is free to use *any* estimator of $\mathbf{\Omega}$ as long as those conditions are satisfied. We next discuss these conditions (and some notations) followed by our results.

For any matrix $\mathbf{M}_{d\times d}$, let $\mathbf{M}_{[i\cdot]} \in \mathbb{R}^d$ denote its i^{th} row and $\mathbf{M}_{[ij]}$ denote its $(i,j)^{th}$ entry. Let $\|\mathbf{M}\|_1 := \max_{1\le i\le d} \sum_{j=1}^d |\mathbf{M}_{[ij]}|$, $\|\mathbf{M}\|_2 = \lambda_{\max}^{1/2}(\mathbf{M}'\mathbf{M})$ and $\|\mathbf{M}\|_{\max} := \max_{1\le i,j\le d} |\mathbf{M}_{[ij]}|$ denote the maximum rowwise L_1 norm, the spectral norm and the elementwise maximum norm of $\mathbf{M}$ respectively, where $\lambda_{\max(\cdot)}$ denotes the maximum eigenvalue. Finally, recall the notations $\mathbf{T}_{0,n}, \mathbf{T}_{\pi,n}, \mathbf{T}_{m,n}$ and $\mathbf{R}_{\pi,m,n}$ defined in the decomposition (3.1) of $\mathbf{T}_n \equiv \boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)$ and for convenience of further discussion, define:

$$\mathbf{R}_{n,1} := -\frac{1}{2}(\widehat{\mathbf{\Omega}} - \mathbf{\Omega})\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0), \ \ \mathbf{R}_{n,2} := -\frac{\mathbf{\Omega}}{2}(\mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n})$$
$$\mathbf{R}_{n,3} := (I_d - \widehat{\mathbf{\Omega}}\widehat{\mathbf{\Sigma}})(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) \ \text{ and let } \ \boldsymbol{\Delta}_n := (\mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,3}). \tag{4.2}$$

ASSUMPTION 4.1 (High-level conditions on $\mathbf{\Omega}$ and $\widehat{\mathbf{\Omega}}$}). We assume that: (a) $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 = O_{\mathbb{P}}(r_n)$ and $\|I_d - \widehat{\mathbf{\Omega}}\widehat{\mathbf{\Sigma}}\|_{\max} = O_P(\omega_n)$ for some sequences $\{r_n, \omega_n\} \equiv \{r_{n,\mathbf{\Omega}}, \omega_{n,\mathbf{\Omega}}\} \geq 0$ with $r_n\sqrt{\log d} = o_{\mathbb{P}}(1)$ and $\omega_n(s\sqrt{\log d}) = o_{\mathbb{P}}(1)$, where $s = \|\boldsymbol{\theta}_0\|_0$ and I_d denotes the $d \times d$ identity matrix.

(b) $\mathbf{\Upsilon}(\mathbf{X}) := \mathbf{\Omega}\mathbf{\Psi}(\mathbf{X})$ is sub-Gaussian (as per Definition C.1 with $\alpha = 2$) with $\|\mathbf{\Upsilon}(\mathbf{X})\|_{\psi_2} \leq \sigma_{\mathbf{\Upsilon}} < \infty$, for some constant $\sigma_{\mathbf{\Upsilon}} \geq 0$. Further, we assume that $v_n^* = o_{\mathbb{P}}(1)$, where $v_n^* := (v_{n,\pi} + v_{\bar{n},m})\sqrt{(\log d)\log(nd)} + n^{\frac{1}{2}} v_{n,\pi} v_{\bar{n},m}(\log n)$ and $\{v_{n,\pi}, v_{\bar{n},m}\}$ are the rates of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ defined in Assumptions 3.2-3.3.

Assumption 4.1 (a) imposes some general rate conditions on $\widehat{\mathbf{\Omega}}$. For most common choices of $\widehat{\mathbf{\Omega}}$, including those discussed earlier, these lead to fairly standard conditions. Under a low dimensional setting with $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Sigma}}^{-1}$, $\omega_n = 0$ trivially and $r_n = d/\sqrt{n}$ under suitable assumptions; see Vershynin (2018) for relevant results. Under high dimensional settings, with $\mathbf{\Omega}$ assumed to be sparse and $\widehat{\mathbf{\Omega}}$ chosen to be the Nodewise Lasso estimator, $\omega_n = \sqrt{(\log d)/n}$ and $r_n = s_{\mathbf{\Omega}}\sqrt{(\log d)/n}$; see van de Geer et al. (2014) for relevant results. In this case, the conditions read as: $s_{\mathbf{\Omega}}(\log d) = o(\sqrt{n})$ and $s(\log d) = o(\sqrt{n})$. These are all familiar (often unavoidable) conditions in the high dimensional inference literature (Cai and Guo, 2017; Javanmard and Montanari, 2018).

The sub-Gaussianity condition on $\mathbf{\Upsilon}(\mathbf{X})$ in Assumption 4.1 (b) is needed to control the term $\mathbf{R}_{n,2}$ in (4.2). Conditions of a similar flavor have also been adopted implicitly or explicitly in van de Geer et al. (2014) and Javanmard and Montanari (2014). The condition holds with $\sigma_{\mathbf{\Upsilon}}$ to be a constant if either $\|\mathbf{\Omega}\|_2 = O(1)$ and $\mathbf{\Psi}(\mathbf{X})$ is (vector) sub-Gaussian in the sense of Vershynin (2018) with a $O(1)$ norm, or if $\|\mathbf{\Omega}\|_1 = O(1)$ and $\mathbf{\Psi}(\mathbf{X})$ is sub-Gaussian in the (weaker) sense of Definition C.1 with a $O(1)$ norm. Finally, the condition on v_n^* is the same (upto a $\sqrt{\log d}$ factor) as those needed for Theorems 3.2-3.4.

THEOREM 4.1 (ALE and entrywise asymptotic normality of $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$). *Under Assumptions 1.1, 2.1, 3.1-3.3 and 4.1, and with $\mathbf{\Delta}_n$ as defined in (4.2), $L(\cdot)$ assumed to be the squared loss and $\widehat{\boldsymbol{\theta}}_{DDR}$ constructed using a choice of $\lambda_n \asymp \sqrt{(\log d)/n}$, the desparsified DDR estimator $\widetilde{\boldsymbol{\theta}}_{DDR}$ satisfies the ALE:*

$$(4.3)\quad (\widetilde{\boldsymbol{\theta}}_{DDR} - \boldsymbol{\theta}_0) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Omega}\{\boldsymbol{\psi}_0(\mathbf{Z}_i)\} + \mathbf{\Delta}_n, \quad \textit{where } \mathbb{E}\{\boldsymbol{\psi}_0(\mathbf{Z})\} = \mathbf{0} \textit{ with}$$

$$\boldsymbol{\psi}_0(\mathbf{Z}) \;=\; \{m(\mathbf{X}) - \mathbf{\Psi}(\mathbf{X})'\boldsymbol{\theta}_0\}\mathbf{\Psi}(\mathbf{X}) + \frac{T}{\pi(\mathbf{X})}\{Y - m(\mathbf{X})\}\mathbf{\Psi}(\mathbf{X}), \quad \textit{and}$$

$$\|\mathbf{\Delta}_n\|_\infty \;=\; O_{\mathbb{P}}\left(r_n\sqrt{\frac{\log d}{n}} + v_n^* n^{-\frac{1}{2}} + \omega_n s\sqrt{\frac{\log d}{n}}\right) \;=\; o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

Consequently, letting $\boldsymbol{\Gamma}_0(\mathbf{Z}) := \boldsymbol{\Omega}\boldsymbol{\psi}_0(\mathbf{Z})$*,* $\sigma_{0,j}^2 := \mathbb{E}\{\boldsymbol{\Gamma}_{0[j]}^2(\mathbf{Z})\}$ *and assuming that* $\sigma_{0,j} > c_0 \ \forall\, j$*, for some constant* $c_0 > 0$*, we have: for each* $1 \leq j \leq d$*,*

$$\sqrt{n}\sigma_{0,j}^{-1}(\widetilde{\boldsymbol{\theta}}_{\scriptscriptstyle DDR[j]} - \boldsymbol{\theta}_{0[j]}) \xrightarrow{d} \mathcal{N}(0,1) \ \ \textit{and} \ \ \sqrt{n}\widehat{\sigma}_{0,j}^{-1}(\widetilde{\boldsymbol{\theta}}_{\scriptscriptstyle DDR[j]} - \boldsymbol{\theta}_{0[j]}) \xrightarrow{d} \mathcal{N}(0,1),$$

$$\textit{where} \ \ \widehat{\sigma}_{0,j}^2 := \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\Gamma}}_{0[j]}^2(\mathbf{Z}_i) \ \ \textit{satisfying} \ \ \max_{1\leq j\leq d} |\widehat{\sigma}_{0,j}^2 - \sigma_{0,j}^2| = o_{\mathbb{P}}(1).$$

Here $\widehat{\boldsymbol{\Gamma}}_{0[j]}(\mathbf{Z}_i) := \widehat{\boldsymbol{\Omega}}'_{[j\cdot]}\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}_i)$*, where* $\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}_i)$ *denotes the estimated version of* $\boldsymbol{\psi}_0(\mathbf{Z})$ *in* (4.3) *with* $\{\pi(\mathbf{X}_i), m(\mathbf{X}_i), \boldsymbol{\theta}_0\}$ *plugged in as* $\{\widehat{\pi}(\mathbf{X}_i), \widetilde{m}(\mathbf{X}_i), \widehat{\boldsymbol{\theta}}_{\scriptscriptstyle DDR}\}$*.*

Theorem 4.1 therefore provides all the necessary inferential tools for $\widetilde{\boldsymbol{\theta}}_{\rm DDR}$. The ALE (4.3) is also *optimal* in a certain sense since the function $\boldsymbol{\Gamma}_0(\mathbf{Z}) \equiv \boldsymbol{\Omega}\boldsymbol{\psi}_0(\mathbf{Z})$ defining the i.i.d. summand (also known as the influence function) in the ALE is known to be the *efficient influence function* for estimating $\boldsymbol{\theta}_0$ in a classical setting (d fixed) under a fully non-parametric (i.e. unrestricted, upto Assumption 1.1) family of $\mathbb{P}$ and its variance equals the semi-parametric optimal variance (Robins, Rotnitzky and Zhao, 1994; Robins and Rotnitzky, 1995; Graham, 2011). The same conclusions continue to hold in high dimensional settings for low-dimensional components (e.g. each coordinate) of $\boldsymbol{\theta}_0$. Thus, $\widetilde{\boldsymbol{\theta}}_{\rm DDR}$ achieves the (coordinatewise) *semi-parametric efficiency bound* and is optimal among all achievable estimators of $\boldsymbol{\theta}_0$ admitting ALEs under a non-parametric family of $\mathbb{P}$. Furthermore, the asymptotic normality results allow one to construct asymptotically valid $(1-\alpha)$ level confidence intervals: $CI_j := \widetilde{\boldsymbol{\theta}}_{{\rm DDR}[j]} \pm z_{\alpha/2}\widehat{\sigma}_{0,j}$ for each coordinate $\boldsymbol{\theta}_{0[j]}$ of $\boldsymbol{\theta}_0$, where $z_{\alpha/2}$ denotes the $(1-\alpha/2)^{th}$ quantile of the $\mathcal{N}(0,1)$ distribution with $\alpha \in (0,1)$.

5. Estimation of the Nuisance Functions.

In Sections 5.1-5.2, we discuss various choices for the nuisance function estimators $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ required for implementing our proposed methods. Our entire approach so far does *not* require any specific knowledge of the construction or properties of these estimators as long as they satisfy the high-level conditions in Assumption 3.2-3.3. Hence, one is free to use *any* choice of these estimators based on high dimensional parametric or semi-parametric models, or even non-parametric machine learning based estimators, as has been advocated in many recent works for other related problems in similar settings (Farrell, 2015; Chernozhukov et al., 2018a; Farrell, Liang and Misra, 2018). However, a fully non-parametric and/or machine learning based approach may not be feasible or efficient in 'truly' high dimensional settings where p diverges with n. In this section, we discuss a few novel, principled, and yet, flexible families of choices for $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$, including common parametric models, as well as series estimators and single index models. In Sections 5.3-5.4, we establish general results for all these estimators under high dimensional settings that verify our basic assumptions and may also be of independent interest.

5.1. *Propensity Score Estimation: A Few Choices and Their Properties.* In some cases, $\pi(\cdot)$ may be known whereby $\widehat{\pi}(\cdot) \equiv \pi(\cdot)$ trivially. When $\pi(\cdot)$ is unknown, we consider the following (class of) choices for estimating $\pi(\cdot)$.

'Extended' parametric families (or high dimensional series estimators). We assume that $\pi(\cdot)$ belongs to the family: $\pi(\mathbf{x}) \equiv \mathbb{E}(T|\mathbf{X}=\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{x})\}$, where $g(\cdot) \in [0,1]$ is a *known* 'link' function, $\boldsymbol{\Psi}(\mathbf{x}) := \{\psi_k(\mathbf{x})\}_{k=1}^{K}$ is any set of K (known) basis functions, possibly high dimensional with K allowed to depend on n (including $K \gg n$), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an *unknown* parameter vector that is further assumed to be sparse (if required).

Estimator. $\pi(\mathbf{x})$ is then estimated as: $\widehat{\pi}(\mathbf{x}) = g\{\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Psi}(\mathbf{x})\}$, where $\widehat{\boldsymbol{\alpha}}$ denotes some given estimator of $\boldsymbol{\alpha}$ obtained via any suitable estimation procedure based on the observed data for $(T,\mathbf{X})$ that satisfies a basic high-level requirement that $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \leq a_n$ w.h.p. for some sequence $a_n = o(1)$.

Examples. The models above include, as a special case, any logistic regression model for $T|\mathbf{X}$ given by: $\pi(\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{x})\}$, where $g(u) = g_{\text{expit}}(a) := \exp(a)/\{1+\exp(a)\}$. The estimator $\widehat{\boldsymbol{\alpha}}$ in this case maybe obtained using a simple L_1-penalized logistic regression of T vs. $\boldsymbol{\Psi}(\mathbf{X})$ based on the observed data $\{T_i, \boldsymbol{\Psi}(\mathbf{X}_i)\}_{i=1}^n$. Using standard results from high dimensional regression theory (Bühlmann and Van De Geer, 2011; Negahban et al., 2012; Wainwright, 2019), it can be shown that under suitable assumptions (RSC and exponential tail conditions), $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \lesssim a_n \equiv a_n(s_{\boldsymbol{\alpha}}, K) := s_{\boldsymbol{\alpha}}\sqrt{(\log K)/n}$ w.h.p., where $s_{\boldsymbol{\alpha}} := \|\boldsymbol{\alpha}\|_0$ denotes the sparsity of $\boldsymbol{\alpha}$.

As for the basis functions $\boldsymbol{\Psi}(\mathbf{x})$, some reasonable choices include the polynomial bases given by: $\boldsymbol{\Psi}(\mathbf{x}) := \{1, \mathbf{x}_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\}$ for any degree $d_0 \geq 1$. The special case $d_0 = 1$ corresponds to the linear bases which leads to all standard parametric models that are commonly used in practice.

The case when $\pi(\cdot)$ is constant. Note that the extended parametric framework above also includes the special case where $\pi(\cdot)$ is unknown but constant (i.e. the case of MCAR or complete randomization), in which case $g(\boldsymbol{\alpha}'\mathbf{X})$ simply equals the constant π and $\boldsymbol{\alpha}$ is just an unknown parameter in $\mathbb{R}$ that can be estimated at the rate of $O(n^{-1/2})$ via the usual sample mean of T.

5.2. *Estimation of the Conditional Mean: Choices and Their Properties.* We consider the following two (class of) choices for estimating $m(\cdot)$.

1. 'Extended' parametric families (high dimensional series estimators). We assume that $m(\cdot)$ belongs to the family: $g\{\boldsymbol{\gamma}'\boldsymbol{\Psi}(\mathbf{X})\}$ where $g(\cdot)$ is a (known) 'link' function (e.g. 'canonical' links functions), $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^K$ is any set of K (known) basis functions, with K possibly high dimensional and allowed to depend on n (including $K \gg n$), and $\boldsymbol{\gamma} \in \mathbb{R}^K$ is an unknown parameter vector that is further assumed to be sparse (if required).

Estimator. We estimate $m(\mathbf{x}) \equiv \mathbb{E}(Y|\mathbf{X}) \equiv \mathbb{E}(Y|\mathbf{X}, T = 1) = g\{\boldsymbol{\gamma}'\boldsymbol{\Psi}(\mathbf{X})\}$ as: $\widehat{m}(\mathbf{x}) = g\{\widehat{\boldsymbol{\gamma}}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $\widehat{\boldsymbol{\gamma}}$ denotes some given estimator of $\boldsymbol{\gamma}$ obtained via any suitable estimation procedure based on the 'complete case' data $\mathcal{D}_n^{(c)} := \{(Y_i, \mathbf{X}_i) \mid T_i = 1\}_{i=1}^n$ that satisfies a basic high-level requirement that $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \leq a_n$ w.h.p. for some sequence $a_n = o(1)$.

Examples. These models include, as special cases, all standard parametric regression models with 'canonical' link functions, through suitable choices of $g(\cdot)$ depending on the nature of Y (continuous, binary or discrete). Specifically, $g(u) \equiv g_{\text{id}} = u$ (identity link), $g(u) \equiv g_{\text{expit}} = \exp(u)/\{1 + \exp(u)\}$ (expit/logit link) and $g(u) \equiv g_{\exp} = \exp(u)$ (exponential/log link) correspond to the linear, logistic and Poisson regression models respectively.

As for the basis functions $\boldsymbol{\Psi}(\mathbf{x})$, some reasonable choices include the polynomial bases given by: $\boldsymbol{\Psi}(\mathbf{x}) := \{1, \mathbf{x}_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\}$ for any degree $d_0 \geq 1$. The special case $d_0 = 1$ corresponds to the linear bases which leads to all standard parametric models, while $d_0 = 3$ leads to cubic splines.

Examples of $\widehat{\boldsymbol{\gamma}}$. For all the examples above, with $g(\cdot)$ being any 'canonical' link function, the estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ may be simply obtained through a corresponding L_1 penalized 'canonical' link based regression (e.g. linear, logistic or Poisson regression) of Y vs. $\mathbf{X}$ in the 'complete case' data $\mathcal{D}_n^{(c)}$ under Assumption 1.1 (a). Using standard results from high dimensional regression (Bühlmann and Van De Geer, 2011; Negahban et al., 2012; Wainwright, 2019), it can be shown that under suitable assumptions (e.g. RSC and exponential tail conditions) and Assumption 1.1, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \lesssim a_n \equiv a_n(s_{\boldsymbol{\gamma}}, K) := s_{\boldsymbol{\gamma}}\sqrt{(\log K)/n}$ w.h.p., where $s_{\boldsymbol{\gamma}} := \|\boldsymbol{\gamma}\|_0$ denotes the sparsity of $\boldsymbol{\gamma}$.

2. Semi-parametric single index models. We assume that $m(\cdot)$ satisfies the SIM: $m(\mathbf{X}) \equiv \mathbb{E}(Y|\mathbf{X}) \equiv \mathbb{E}(Y|\mathbf{X}, T = 1) = g(\boldsymbol{\gamma}'\mathbf{X})$, where $g(\cdot) \in \mathbb{R}$ is some *unknown* 'link' function and $\boldsymbol{\gamma} \in \mathbb{R}^p$ is an unknown parameter (identifiable only upto scalar multiples) that is further assumed to be sparse (if required).

Estimator. Given any reasonable estimator $\widehat{\boldsymbol{\gamma}}$ of the $\boldsymbol{\gamma}$ 'direction' obtained from $\mathcal{D}_n$, we estimate $m(\mathbf{X}) \equiv \mathbb{E}(Y|\boldsymbol{\gamma}'\mathbf{X}) \equiv \mathbb{E}(Y|\boldsymbol{\gamma}'\mathbf{X}, T = 1) = g(\boldsymbol{\gamma}'\mathbf{X})$ via a one-dimensional kernel smoothing (KS) over the estimated scores $\{\widehat{\boldsymbol{\gamma}}'\mathbf{X}_i\}_{i=1}^n$, under appropriate smoothness and regularity assumptions, as follows.

$$\widehat{m}(\mathbf{x}) \;\equiv\; \widehat{m}(\widehat{\boldsymbol{\gamma}}'\mathbf{x}) \;\equiv\; \widehat{m}(\widehat{\boldsymbol{\gamma}}, \mathbf{x}) \;:=\; \frac{\frac{1}{nh}\sum_{i=1}^{n} T_i Y_i K\left(\frac{\widehat{\boldsymbol{\gamma}}'\mathbf{X}_i - \widehat{\boldsymbol{\gamma}}'\mathbf{x}}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} T_i K\left(\frac{\widehat{\boldsymbol{\gamma}}'\mathbf{X}_i - \widehat{\boldsymbol{\gamma}}'\mathbf{x}}{h}\right)} \quad \forall\, \mathbf{x} \in \mathcal{X},$$

where $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is some suitable 'kernel' function and $h \equiv h_n > 0$ denotes a bandwidth sequence with $h_n = o(1)$. Here, we only assume that $\widehat{\boldsymbol{\gamma}}$ is *some* reasonable estimator of the $\boldsymbol{\gamma}$ direction satisfying a basic high-level condition: $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \leq a_n$ w.h.p. for some $\boldsymbol{\gamma}_0 \propto \boldsymbol{\gamma}$ and $a_n = o(1)$.

Estimation of $\widehat{\boldsymbol{\gamma}}$. Under Assumption 1.1 (a) and the SIM framework we have adopted here, $\mathbb{E}(Y|\mathbf{X}) \equiv \mathbb{E}(Y|\mathbf{X}, T = 1) = g(\boldsymbol{\gamma}'\mathbf{X})$. Hence, in general, one may use any standard method available in the literature for signal recovery in SIMs (Horowitz, 2009; Alquier and Biau, 2013; Radchenko, 2015) and apply it to the 'complete case' data $\mathcal{D}_n^{(c)}$ to obtain a reasonable estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$. Under some additional design restrictions and model assumptions, however, one may also estimate $\boldsymbol{\gamma}$ by even simpler approaches, as follows.

(a) Suppose Y satisfies the (slightly) stronger SIM formulation: $(Y|\mathbf{X}) \equiv (Y|\mathbf{X}, T = 1) = f(\boldsymbol{\gamma}'\mathbf{X}; \epsilon)$ for some unknown function $f : \mathbb{R}^2 \to \mathcal{Y}$ and some noise $\epsilon \perp\!\!\!\perp (T, \mathbf{X})$, and assume further that the distribution of $(\mathbf{X}|T = 1)$ is elliptically symmetric. Then, owing to the results of Li and Duan (1989), one can *still* estimate $\boldsymbol{\gamma}$ with a rate guarantee of $a_n = s_{\boldsymbol{\gamma}}\sqrt{(\log p)/n}$ using a simple L_1 penalized 'canonical' link based regression (e.g. linear, logistic or Poisson regression) of Y vs. $\mathbf{X}$ in the 'complete case' data $\mathcal{D}_n^{(c)}$, as discussed in the previous example. Similar approaches have been used extensively in recent years for sparse signal recovery in high dimensional SIMs with fully observed data and elliptically symmetric designs (Plan and Vershynin, 2013, 2016; Goldstein, Minsker and Wei, 2016; Genzel, 2017; Wei, 2018)

(b) Suppose Y satisfies the same SIM as in part (a) above, and assume now that the distribution of $\mathbf{X}$ is elliptically symmetric. Then, combining the results of Li and Duan (1989) along with those in Section 2.1 regarding IPW representations, it follows that one can estimate $\boldsymbol{\gamma}$ using an L_1-penalized *weighted* regression based on any 'canonical' link (e.g. linear, logistic or Poisson regression) of Y vs. $\mathbf{X}$ in the 'complete case' data $\mathcal{D}_n^{(c)}$. The weights are given by $\pi^{-1}(\mathbf{X})$, if $\pi(\cdot)$ is known, or $\widehat{\pi}^{-1}(\mathbf{X})$ if $\pi(\cdot)$ is unknown and estimated via $\widehat{\pi}(\cdot)$ (assumed to be correctly specified) through any of the choices discussed in Section 5.2. Using the results of Negahban et al. (2012) along with the techniques used in our proofs of Lemma 2.1 and Theorems 3.1 and 3.4, it can be shown that the resulting IPW estimator $\widehat{\boldsymbol{\gamma}}$ satisfies an

L_1 norm bound $\|\widehat{\gamma} - \gamma\|_1 \lesssim a_n \equiv s_{\gamma}\sqrt{(\log p)/n}$ w.h.p. in the case when $\pi(\cdot)$ is known, and $\|\widehat{\gamma} - \gamma\|_1 \lesssim a_n \equiv s_{\gamma} \max\{\sqrt{(\log p)/n}, v_{n,\pi}\sqrt{\log n}\}$ when $\pi(\cdot)$ is unknown, where $v_{n,\pi} = o(1)$ denotes the (pointwise) convergence rate of $\widehat{\pi}(\cdot)$ as given in Assumption 3.2. Given the main goals of this paper, we skip the technical details and proofs of these claims for the sake of brevity.

5.3. *Convergence Rates for the 'Extended' Parametric Families.* We establish here tail bounds and convergence rates for estimators based on the 'extended' parametric families discussed in Sections 5.1-5.2. For notational simplicity, we derive the results for a general outcome which may be assigned to be T for estimation of $\pi(\cdot)$, or TY for estimation of $m(\cdot)$. Let $(Z, \mathbf{X})$ denote a generic random vector where $Z \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$ with support $\mathcal{X} \subseteq \mathbb{R}^p$. Consider an 'extended' parametric family of (working) models for estimating $\mathbb{E}(Z|\mathbf{X})$ given by: $g\{\boldsymbol{\beta}'\boldsymbol{\Psi}(\mathbf{X})\}$ where $\boldsymbol{\Psi}(\mathbf{X}) \in \mathbb{R}^K$ is some vector of basis functions. Let $\boldsymbol{\beta}_0$ denote the 'target' parameter corresponding to this working model and let $\widehat{\boldsymbol{\beta}}$ be *any* estimator of $\boldsymbol{\beta}_0$ based on any suitable procedure applied to the observed data: $\{Z_i, \mathbf{X}_i\}_{i=1}^n$. Then, we estimate $\mathbb{E}(Z|\mathbf{X} = \mathbf{x})$ based on the working model as: $g\{\widehat{\boldsymbol{\beta}}'\boldsymbol{\Psi}(\mathbf{x})\}$. The result below establishes a tail bound for this estimator w.r.t. its target $g\{\boldsymbol{\beta}_0'\boldsymbol{\Psi}(\mathbf{x})\}$.

THEOREM 5.1. *Suppose $\widehat{\boldsymbol{\beta}}$ satisfies a basic high-level L_1 error guarantee:*

$$\mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > a_n) \;\leq\; q_n \;\; \textit{for some} \;\; a_n, q_n = o(1), \;\; a_n \geq 0, \; q_n \in [0,1].$$

Suppose further that $g(\cdot)$ is Lipschitz continuous with $|g(u) - g(v)| \leq C_g|u-v|$ $\forall\, u, v \in \mathbb{R}$ and that $\boldsymbol{\Psi}(\mathbf{X})$ is uniformly bounded, i.e. $\max_{1 \leq j \leq K} |\boldsymbol{\Psi}_{[j]}(\mathbf{X})| \leq C_{\boldsymbol{\Psi}} < \infty$ a.s. $[\mathbb{P}]$, for some constants $C_g, C_{\boldsymbol{\Psi}} \geq 0$. Then, for any $t \geq 0$,

$$\mathbb{P}\left[\sup_{\mathbf{x} \in \mathcal{X}} |g\{\widehat{\boldsymbol{\beta}}'\boldsymbol{\Psi}(\mathbf{x})\} - g\{\boldsymbol{\beta}_0'\boldsymbol{\Psi}(\mathbf{x})\}| \;>\; (\sqrt{2}C_g C_{\boldsymbol{\Psi}})a_n t\right] \;\leq\; 2\exp(-t^2) + q_n.$$

Theorem 5.1 establishes a bound for the supremum which is much stronger than what we need to verify our basic assumptions. Nevertheless, as a consequence, it establishes that when one uses any of these 'extended' parametric families for constructing $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$, then the pointwise tail bounds required in our basic Assumptions 3.2-3.3 hold with the choices of $\{v_{n,\pi}, v_{n,m}\} \propto a_n$ and $\{q_{n,\pi}, q_{n,m}\} \propto q_n$. Further, as discussed in Sections 5.1 and 5.2, for most common choices of $\widehat{\boldsymbol{\beta}}$ based on penalized estimators from high dimensional models, the L_1 error rate a_n should behave as: $a_n \propto s_{\boldsymbol{\beta}_0}\sqrt{(\log K)/n}$ w.h.p.

5.4. *High Dimensional Single Index Models: Non-Asymptotic Bounds and Rates for KS over Estimated Index Parameters.* In this section, we study the properties of single index KS estimators involving high dimensional covariates with the index parameter being (possibly) unknown and estimated. The underlying high dimensionality and the non-ignorable index estimation error makes the analyses nuanced and different from most existing results in the literature under classical settings. We consider both linear kernel average estimators (e.g. density estimators) as well as ratio form estimators (e.g. conditional mean estimators) and develop a non-asymptotic theory that establishes concrete tail bounds and pointwise convergence rates for such estimators. The results apply equally to both classical and high dimensional regimes, and while obtained in course of characterizing our nuisance function estimators' properties, may also be useful in other applications and should be of independent interest. We therefore present the results under a generic framework and a set of notations that is independent of the rest of the paper.

Let $\{(Z_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ denote a sample of $n \geq 2$ i.i.d. realizations of a generic random vector $(Z, \mathbf{X})$ assumed to have finite 2^{nd} moments, where $Z \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$ with support $\mathcal{X} \subseteq \mathbb{R}^p$ and $p \geq 1$ is allowed to be high dimensional compared to the sample size, i.e. p is allowed to diverge with n.

Let $\boldsymbol{\beta} \in \mathbb{R}^p$ be any (unknown) 'parameter' of interest and let $\widehat{\boldsymbol{\beta}}$ denote *any* reasonable estimator of $\boldsymbol{\beta}$ that satisfies a basic high-level L_1 error guarantee:

$$(5.1)\quad \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > a_n) \ \leq \ q_n \ \text{ for some } \ a_n, q_n = o(1), \ \ a_n \geq 0, \ q_n \in [0,1].$$

(5.1) is a reasonable high-level requirement that should hold in most cases. It is important to note that (5.1) is the *only* condition we require on $\{\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}\}$ for all our results and nothing specific regarding their construction or properties.

Let $W \equiv W_{\boldsymbol{\beta}} := \boldsymbol{\beta}'\mathbf{X}$ and $\widehat{W} := \widehat{\boldsymbol{\beta}}'\mathbf{X}$. For any $\mathbf{x} \in \mathbb{R}^p$, let $w_{\mathbf{x}} \equiv w_{\mathbf{x},\boldsymbol{\beta}} := \boldsymbol{\beta}'\mathbf{x}$ and $\widehat{w}_x := \widehat{\boldsymbol{\beta}}'\mathbf{x}$. For any $w \in \mathbb{R}$, let $m_{\boldsymbol{\beta}}(w) := \mathbb{E}(Z|W = w)$ and $l_{\boldsymbol{\beta}}(w) := m_{\boldsymbol{\beta}}(w) f_{\boldsymbol{\beta}}(w)$, where $f_{\boldsymbol{\beta}}(\cdot)$ denotes the density of $W \equiv \boldsymbol{\beta}'\mathbf{X}$. Finally, for any $\mathbf{x} \in \mathcal{X}$, let $m(\boldsymbol{\beta}, \mathbf{x}) := m_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x})$, $f(\boldsymbol{\beta}, \mathbf{x}) := f_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x})$ and $l(\boldsymbol{\beta}, \mathbf{x}) := l_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{x})$.

Given *any* estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfying (5.1), consider the following single index KS estimators of $l(\boldsymbol{\beta}, \mathbf{x})$, $f(\boldsymbol{\beta}, \mathbf{x})$ and $m(\boldsymbol{\beta}, \mathbf{x})$ for any *fixed* $\mathbf{x} \in \mathcal{X}$,

$$\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) \ := \ \frac{1}{nh}\sum_{i=1}^{n} Z_i K\left(\frac{\widehat{\boldsymbol{\beta}}'\mathbf{X}_i - \widehat{\boldsymbol{\beta}}'\mathbf{x}}{h}\right) \ \equiv \ \frac{1}{nh}\sum_{i=1}^{n} Z_i K\left(\frac{\widehat{W}_i - \widehat{w}_{\mathbf{x}}}{h}\right),$$

$$\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) \ := \ \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\widehat{\boldsymbol{\beta}}'\mathbf{X}_i - \widehat{\boldsymbol{\beta}}'\mathbf{x}}{h}\right) \quad \text{and} \quad \widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) \ := \ \frac{\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})}{\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x})},$$

where $K(\cdot) : \mathbb{R} \to \mathbb{R}$ denotes any suitable kernel function (e.g. the Gaussian kernel) and $h \equiv h_n > 0$ denotes the bandwidth sequence with $h_n = o(1)$.

$\widehat{l}(\cdot)$ and $\widehat{f}(\cdot)$ are both linear kernel average (LKA) estimators while $\widehat{m}(\cdot)$ is a ratio type KS estimator. We obtain non-asymptotic tail bounds and (pointwise) convergence rates for these estimators in Theorems 5.2-5.3 below. The Assumptions K.1-K.2 for these results are given separately in Appendix K.2.

THEOREM 5.2 (Tail bounds for LKA estimators). *Consider the estimator $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$ of $l(\boldsymbol{\beta}, \mathbf{x})$. Assume (5.1) and Assumptions K.1-K.2 (in Appendix K.2) and that $h = o(1)$, $\log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, for any fixed $\mathbf{x} \in \mathcal{X}$ and any $t \geq 0$, with probability at least $1 - 9\exp(-t^2) - 2q_n$,*

$$|l(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| \;\leq\; C_1\left(\frac{t+1}{\sqrt{nh}} + \frac{t^2\sqrt{\log n}}{nh}\right) + C_2\left(h^2 + a_n + \frac{a_n^2}{h^2} + \frac{\log(np)}{nh}\right)$$

for some constants $C_1, C_2 > 0$ depending only on those in the assumptions.

Apart from an explicit tail bound, Theorem 5.2 also establishes the convergence rate of $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$ to be $O(nh^{-\frac{1}{2}} + h^2 + a_n + a_n^2 h^{-2})$ which quantifies the additional price one pays for estimating the high dimensional index parameter $\boldsymbol{\beta}$ apart from the error rate of a standard one dimensional KS. This is highlighted through all the terms in the bound involving the L_1 error rate a_n of $\widehat{\boldsymbol{\beta}}$. For a given a_n, one can also optimize the choice of $h = O(n^{-a})$ over $a > 0$ by minimizing the convergence rate above whose terms behave differently with h, similar to a variance-bias tradeoff phenomenon typically observed in KS regression. We skip these technical discussions here for brevity.

THEOREM 5.3 (Tail bounds for ratio type KS estimators). *Consider the ratio type KS estimator $\widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$ of $m(\boldsymbol{\beta}, \mathbf{x})$ and assume that $|m(\boldsymbol{\beta}, \mathbf{x})| \leq \delta_m$ and $f(\boldsymbol{\beta}, \mathbf{x}) \geq \delta_f > 0$ for some constants $\delta_m, \delta_f > 0$. For any $t \geq 0$, define:*

$$\epsilon_n(t) := C_1\frac{t+1}{\sqrt{nh}} + C_2\frac{t^2\sqrt{\log n}}{nh} + C_3 b_n, \quad \textit{where}\;\; b_n := h^2 + a_n + \frac{a_n^2}{h^2} + \frac{\log(np)}{nh}$$

and $C_1, C_2, C_3 > 0$ are the same constants as in Theorem 5.2. Assume (5.1), Assumptions K.1-K.2 (in Appendix K.2) and that $h = o(1)$, $\log(np)/(nh) = o(1)$, $(a_n/h)\sqrt{\log p} = o(1)$ and $b_n = o(1)$. Then, for any fixed $\mathbf{x} \in \mathcal{X}$ and any $t, t_* \geq 0$ with t_* further assumed w.l.o.g. to satisfy $\epsilon_n(t_*) \leq \delta_f/2 < \delta_f$, we have: with probability at least $1 - 18\exp(-t^2) - 9\exp(-t_*^2) - 6q_n$,

$$|\widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - m(\boldsymbol{\beta}, \mathbf{x})| \;\leq\; \frac{2(1+\delta_m)}{\delta_f}\epsilon_n(t) \;\lesssim\; \frac{t+1}{\sqrt{nh}} + \frac{t^2\sqrt{\log n}}{nh} + b_n,$$

where '$\lesssim$' denotes inequality upto multiplicative contants (possibly depending on those introduced in the assumptions). In particular, assuming further that

$\{\log(np)\log n\}/(nh) = o(1)$ *and choosing* $t = t_* = c\sqrt{\log np}$ *for any* $c > 0$ *(assuming w.l.o.g. the chosen* t_* *satisfies the required condition), we have:*

$$|\widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - m(\boldsymbol{\beta}, \mathbf{x})| \lesssim (c+1)\sqrt{\frac{\log(np)}{nh}}\left(1 + c\sqrt{\frac{\log(np)\log n}{nh}}\right) + b_n$$
$$\lesssim c\sqrt{\frac{\log(np)}{nh}} + b_n \quad \textit{with probability at least } 1 - 27(np)^{-c^2} - 6q_n.$$

Theorem 5.3 establishes explicit tail bounds and convergence rates for the ratio-type KS estimator $\widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$. As a consequence, it also verifies our basic Assumption 3.3 regarding $\widehat{m}(\cdot)$ when one chooses to estimate it using SIMs. In particular, in view of Remark 3.3, it establishes that the tail bound (3.7) holds with the choices $v_{n,m} \propto \sqrt{\log(np)/(nh)} + b_n$ and $q_{n,m} \propto (np)^{-c} + q_n$, for some $c > 0$, with b_n and q_n as above. Finally, as discussed in Sections 5.1 and 5.2, for most common choices of the estimator $\widehat{\boldsymbol{\beta}}$, the L_1 error rate a_n is expected to behave as: $a_n \propto s_{\boldsymbol{\beta}}\sqrt{(\log p)/n}$ w.h.p., where $s_{\boldsymbol{\beta}} := \|\boldsymbol{\beta}\|_0$.

6. Simulation Studies. In this section, we perform a group of simulations to examine the performances of our method under different data generating processes (DGPs) and parameter settings. We set $n = 1000$, $p = 50$ and $p = 500$, which correspond to moderate dimensional and high dimensional settings. We also set $n = 50000$ to study the double-robustness of our proposed estimator and the performance of the complete case estimator (see Section 6.4 for details). For DGPs, the observed data $\mathcal{D}_n := \{\mathbf{Z}_i \equiv (T_i, T_iY_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ is given by $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (the choices of $\boldsymbol{\Sigma}_p$ will be discussed later) and three models for $Y|\mathbf{X}$ and $T|\mathbf{X}$: a logistic model for $T|\mathbf{X}$ and a linear model for $Y|\mathbf{X}$ (denoted as "linear-linear" DGP), a logistic model with both linear and quadratic terms for $T|\mathbf{X}$ and a linear model with both linear and quadratic terms for $Y|\mathbf{X}$ (denoted as "quad-quad" DGP) and a single index model (SIM) for both $T|\mathbf{X}$ and $Y|\mathbf{X}$ (denoted as "SIM-SIM" DGP). These models are formalized as following:

(a) "Linear-linear" DGP:

$$Y = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}'\mathbf{X} + \varepsilon, \quad \varepsilon|\mathbf{X} \sim N(0,1),$$
$$\text{logit}\{\pi(\mathbf{X})\} = \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}'\mathbf{X}.$$

(b) "Quad-quad" DGP:

$$Y = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}'\mathbf{X} + \sum_{j=1}^{p} \boldsymbol{\gamma}_j^* \mathbf{X}_j^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim N(0,1),$$

$$\text{logit}\{\pi(\mathbf{X})\} = \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}'\mathbf{X} + \sum_{j=1}^{p} \boldsymbol{\alpha}_j^* \mathbf{X}_j^2.$$

(c) "SIM-SIM" DGP:

$$Y = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}'\mathbf{X} + c_Y(\boldsymbol{\gamma}'\mathbf{X})^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim N(0,1),$$
$$\text{logit}\{\pi(\mathbf{X})\} = \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}'\mathbf{X} + c_T(\boldsymbol{\alpha}'\mathbf{X})^2.$$

The covariance matrix $\boldsymbol{\Sigma}_p$ is chosen to be: $\boldsymbol{\Sigma}_p = \mathbf{I}_p$ (identity matrix) or $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$ (first order autoregressive AR(1) covariance matrix) or $\boldsymbol{\Sigma}_p = \rho \mathbf{1}_p \mathbf{1}_p' + (1-\rho)\mathbf{I}_p$ (compound symmetric matrix) with $\rho = 0.2$. These choices of the covariance structure of $\mathbf{X}$ correspond to different correlations among the coordinates of $\mathbf{X}$ and different sparsities of the covariance matrices, ranging from independent and sparse (identity matrix) to correlated and not sparse (compound symmetric matrix). The simulation is replicated for 500 times (100 times for $n = 50000$). The tuning parameter in the penalized regression for $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ is selected using 10-fold cross validation using minimizing mean squared errors (MSE) as criterion. Further details on parameter choices and other implementation details are given in Appendix B.1.

To avoid getting extreme values for $\pi(\cdot)$, we manually truncated $\pi(\cdot)$ to 0.1-0.9. By the choice of our parameters, the proportion of data being truncated is roughly 1% and the proportion of objects with missing outcomes turns out to be 40%.

6.1. *Target Parameter and choices of working nuisance models.* We consider the linear regression problem with missing outcome Y. Our target parameter $\boldsymbol{\theta}_0$ is the best linear estimator:

$$\boldsymbol{\theta}_0 := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}(Y - \boldsymbol{\theta}'\overrightarrow{\mathbf{X}})^2 = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\overrightarrow{\mathbf{X}}Y),$$

with $d = p + 1$ and the definition of $\overrightarrow{\mathbf{v}}$ is given in Section 2. This $\boldsymbol{\theta}_0$ is a model-free parameter that is always the target for linear regression problems regardless of whether $\mathbb{E}(Y|\mathbf{X})$ is truly linear or not. For "linear-linear" DGP, $\boldsymbol{\theta}_0$ matches with $\boldsymbol{\gamma}$ introduced previously. For other non-linear DGPs, $\boldsymbol{\theta}_0$ is in general different from the parameters we introduced in the working nuisance models. By the choice of our parameters, this $\boldsymbol{\theta}_0$ is sparse. For all

DGPs (linear or non-linear), we compute (and fix) $\boldsymbol{\theta}_0$ based on a large data with size 200000.

To obtain the DDR estimator $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ for $\boldsymbol{\theta}_0$, two choices of the working nuisance model for $\widehat{\pi}(\cdot)$ and three choices of $\widehat{m}(\cdot)$ are considered. For the *choices of the propensity score* $\widehat{\pi}(\cdot)$:

(1) Fit an L_1 penalized logistic regression with linear covariates based on the observed data $\{T_i, \mathbf{X}_i\}_{i=1}^n$ (denoted as "$\widehat{\pi}$: linear").
(2) Fit an L_1 penalized logistic regression with both linear and quadratic covariates based on the observed data $\{T_i, \mathbf{X}_i\}_{i=1}^n$ ("$\widehat{\pi}$: quad").

For the *choices of the conditional mean* $\widehat{m}(\cdot)$:

(1) Fit an L_1 penalized linear regression using the 'complete case' data $\mathcal{D}_n^{(c)}$ ("$\widehat{m}$: linear").
(2) Fit an L_1 penalized linear regression with both linear and quadratic covariates using the 'complete case' data $\mathcal{D}_n^{(c)}$("$\widehat{m}$: quad").
(3) Fitting a SIM using the 'complete case' data $\mathcal{D}_n^{(c)}$ with the index parameter estimated via inverse-probability-weighted Lasso ("$\widehat{m}$: SIM").

In total, 6 different working nuisance models for $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ are considered. One should notice that these notations are used for estimating the nuisance functions via working models, and have no relation with (and are independent of) the true DPGs for $\pi(\cdot)$ and $m(\cdot)$. For the SIM, the direction $\boldsymbol{\gamma}$ is estimated by a weighted L_1 penalized linear regression in the 'complete case' data $\mathcal{D}_n^{(c)}$, using the inverse estimated propensity score $\widehat{\pi}(\cdot)$ as weights.

For each DGP, there exists a combination of working nuisance models that at least correctly specifies one of $\pi(\cdot)$ and $m(\cdot)$. For "linear-linear" DGP, all the working nuisance models correctly specify the DGP. For "quad-quad" DGP, only the combination "$\widehat{\pi}$: quad-$\widehat{m}$: quad" correctly specifies both of the nuisance models. There are some combinations that correctly specifies one of $\pi(\cdot)$ and $m(\cdot)$. For example, the combination "$\widehat{\pi}$: linear-$\widehat{m}$: quad" correctly specifies the working nuisance function $m(\cdot)$ but misspecifies $\pi(\cdot)$. For "SIM-SIM" DGP, we include the case that $m(\cdot)$ is correctly specified but omit the case when $\pi(\cdot)$ is correctly specified.

From the results presented in Section 6.3, we would see that correctly specifying the conditional mean $m(\cdot)$ would largely reduce the estimation errors comparing to that of $\pi(\cdot)$.

6.2. *Estimators implemented.* Aside from $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$, we also consider two other estimators for comparison:

(a) $\widehat{\boldsymbol{\theta}}_{orac}$ (oracle): An estimator obtained assuming $\pi(\cdot)$ and $m(\cdot)$ are known.

(b) $\widehat{\boldsymbol{\theta}}_{full}$ (super oracle): An estimator obtained assuming the full dataset is observed (no missing outcomes).

The implementation of these estimators are the same as our proposed estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$, using the corresponding working nuisance functions or dataset. The oracle estimator $\widehat{\boldsymbol{\theta}}_{orac}$ is considered to examine the impact of estimating the nuisance function by comparing the performance between $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ and $\widehat{\boldsymbol{\theta}}_{orac}$. The super oracle estimator $\widehat{\boldsymbol{\theta}}_{full}$ is computed assuming full data is observed which would never happen under our missing outcomes framework. This ideal-case estimator is used as a benchmark value to indicate the best one can achieve.

We evaluate the simulation performances by:

(1) Measuring the L_2 norms of the differences between the estimators and the true parameter $\boldsymbol{\theta}_0$. This is defined as $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2$, where $\widetilde{\boldsymbol{\theta}}$ is any candidate estimator. The reported values are the average L_2 norms over all replications with the standard errors in the parentheses.
(2) Calculating the average coverage probabilities (CovP) and average lengths of the confidence intervals (CIs). We calculate the empirical coverage probability and average length of the CI for each coefficient and the reported values are means and medians together with standard errors and median absolute deviation (MAD) as subscripts taken over the truly zero and non-zero coefficients. The default level of the CIs is 95%.
(3) We investigate double-robustness of our estimator and the performance of the complete case estimator (samples with $T = 1$) in Section 6.4.

6.3. *Simulation Results.* The simulation results using identity covariance matrix are provided in this section. The results of using AR(1) and compound symmetric covariance matrices share the similar results and hence are included in the supplementary material (Appendix B).

Table 6.1 and 6.2 provide the estimation errors for $n = 1000$ and $p = 50$, 500. From the tables we could see that when both working nuisance models are correctly specified, the estimation errors of our proposed DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ are closed to those of the oracle estimators. Examples are given in Table 6.1(a) and Table 6.2(a). For the “linear-linear” DGP, all the working models are correctly specified as we mentioned in Section 6.1. With the increasing of p, the estimation errors stay the same pattern but are in general larger than the case when $p = 50$. When at least one of the working nuisance models is not correctly specified, the estimation errors are considerably larger. We also notice that when the conditional mean $m(\cdot)$ is correctly specified, the errors are closer to the oracle comparing when $\pi(\cdot)$ is correctly specified. For example in Table 6.1(b), under the “quad-quad” DGP, the error for “$\widehat{\pi}$:

Table 6.1 The L_2 errors of the estimators comparing with oracle values under the setting of $n = 1000$ using identity covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ and different estimators are compared.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.222 (0.035)	0.223 (0.036)	0.168 (0.027)
	$\widehat{\pi}$: quad	0.221 (0.035)	0.223 (0.036)	0.168 (0.027)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.224 (0.035)	0.223 (0.036)	0.168 (0.027)
	$\widehat{\pi}$: quad	0.224 (0.035)	0.223 (0.036)	0.168 (0.027)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.222 (0.036)	0.223 (0.036)	0.168 (0.027)
	$\widehat{\pi}$: quad	0.222 (0.036)	0.223 (0.036)	0.168 (0.027)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.682 (0.115)	0.478 (0.076)	0.453 (0.074)
	$\widehat{\pi}$: quad	0.638 (0.105)	0.478 (0.076)	0.453 (0.074)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.475 (0.077)	0.478 (0.076)	0.453 (0.074)
	$\widehat{\pi}$: quad	0.475 (0.077)	0.478 (0.076)	0.453 (0.074)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.683 (0.116)	0.478 (0.076)	0.453 (0.074)
	$\widehat{\pi}$: quad	0.64 (0.108)	0.478 (0.076)	0.453 (0.074)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.618 (0.138)	0.517 (0.125)	0.499 (0.121)
	$\widehat{\pi}$: quad	0.613 (0.137)	0.517 (0.125)	0.499 (0.121)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.616 (0.141)	0.517 (0.125)	0.499 (0.121)
	$\widehat{\pi}$: quad	0.612 (0.14)	0.517 (0.125)	0.499 (0.121)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.553 (0.132)	0.517 (0.125)	0.499 (0.121)
	$\widehat{\pi}$: quad	0.55 (0.131)	0.517 (0.125)	0.499 (0.121)

Table 6.2 See caption of Table 6.1.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.448 (0.047)	0.424 (0.042)	0.317 (0.028)
	$\widehat{\pi}$: quad	0.448 (0.046)	0.424 (0.042)	0.317 (0.028)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.461 (0.05)	0.424 (0.042)	0.317 (0.028)
	$\widehat{\pi}$: quad	0.461 (0.05)	0.424 (0.042)	0.317 (0.028)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.436 (0.045)	0.424 (0.042)	0.317 (0.028)
	$\widehat{\pi}$: quad	0.436 (0.045)	0.424 (0.042)	0.317 (0.028)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	1.153 (0.122)	0.866 (0.082)	0.811 (0.078)
	$\widehat{\pi}$: quad	1.141 (0.121)	0.866 (0.082)	0.811 (0.078)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.887 (0.088)	0.866 (0.082)	0.811 (0.078)
	$\widehat{\pi}$: quad	0.887 (0.088)	0.866 (0.082)	0.811 (0.078)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	1.151 (0.117)	0.866 (0.082)	0.811 (0.078)
	$\widehat{\pi}$: quad	1.136 (0.117)	0.866 (0.082)	0.811 (0.078)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	1.103 (0.158)	1.116 (0.168)	1.087 (0.165)
	$\widehat{\pi}$: quad	1.09 (0.149)	1.116 (0.168)	1.087 (0.165)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	1.108 (0.159)	1.116 (0.168)	1.087 (0.165)
	$\widehat{\pi}$: quad	1.095 (0.151)	1.116 (0.168)	1.087 (0.165)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	1.034 (0.161)	1.116 (0.168)	1.087 (0.165)
	$\widehat{\pi}$: quad	1.021 (0.153)	1.116 (0.168)	1.087 (0.165)

quad-$\widehat{m}$: linear" is larger than that of "$\widehat{\pi}$: linear-$\widehat{m}$: quad". Similar patterns can also be observed from "SIM-SIM" DGP. In fact, when the working model for $m(\cdot)$ is correct, the estimation performances are equally good for both choices of the working model for $\pi(\cdot)$. In the "SIM-SIM" DGP, the estimation errors for using different working nuisance model are relatively close to each other. This is due to the parameter that we choose for the "SIM-SIM" DGP. In addition, comparing the estimation errors of $\boldsymbol{\theta}_{orac}$ and $\boldsymbol{\theta}_{full}$ across different DGPs, we notice that they are relatively closer for "quad-quad" and "SIM-SIM" DGPs than those of the "linear-linear" DGPs. The joint effect of missing outcomes and non-linear DPGs reduces the gap between the estimation errors using oracle and super oracle estimator.

Table 6.3 Average coverage probabilities and lengths of the CIs built upon the desparsified estimator under the setting of $n = 1000$ using identity covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ are compared. We report the means and medians together with standard errors and MAD as subscripts. The reported values are separated into truly zero and non-zero coefficients.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.93_{0.01}$	$(0.93_{0.01})$	0.16_0
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP:Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.41_0	$0.88_{0.16}$	$(0.93_{0.02})$	$0.46_{0.08}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.41_0	$0.89_{0.12}$	$(0.93_{0.02})$	$0.46_{0.07}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.34_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.38_{0.06}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.34_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.38_{0.06}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.94_{0.01})$	0.41_0	$0.88_{0.16}$	$(0.94_{0.02})$	$0.46_{0.08}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.41_0	$0.89_{0.12}$	$(0.93_{0.03})$	$0.47_{0.07}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.46_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.52_{0.04}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.45_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.52_{0.04}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.45_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.52_{0.04}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.45_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.52_{0.04}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.4_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.46_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.4_0	$0.94_{0.01}$	$(0.94_{0.01})$	$0.45_{0.03}$

Table 6.4 See caption of Table 6.3.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP:Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.16_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.91_{0.02}$	$(0.92_{0.01})$	0.16_0
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.17_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.17_0
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.16_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.16_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.16_0

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.44_0	$0.91_{0.03}$	$(0.92_{0.02})$	$0.46_{0.07}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.43_0	$0.91_{0.03}$	$(0.92_{0.01})$	$0.46_{0.06}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.95_{0.01})$	0.33_0	$0.92_{0.01}$	$(0.92_{0.01})$	$0.35_{0.04}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.33_0	$0.92_{0.01}$	$(0.92_{0.01})$	$0.35_{0.04}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.44_0	$0.91_{0.05}$	$(0.93_{0.02})$	$0.47_{0.07}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.43_0	$0.91_{0.04}$	$(0.92_{0.01})$	$0.46_{0.06}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.95_{0.01})$	0.53_0	$0.87_{0.05}$	$(0.88_{0.06})$	$0.57_{0.03}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.53_0	$0.87_{0.05}$	$(0.86_{0.07})$	$0.57_{0.03}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.53_0	$0.88_{0.04}$	$(0.88_{0.05})$	$0.57_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.53_0	$0.87_{0.05}$	$(0.87_{0.06})$	$0.57_{0.03}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.5_0	$0.93_{0.02}$	$(0.93_{0.01})$	$0.54_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.5_0	$0.93_{0.02}$	$(0.93_{0.01})$	$0.54_{0.03}$

Table 6.3 and 6.4 provide summary statistics of the coverage probabilities and lengths of the CIs. From the results, we could see that the coverage probabilities of the truly zero coefficients are always at a desired level regardless of the working nuisance models and DGPs. For the truly non-zero coefficients, the results are determined by the working models and n, p. When $m(\cdot)$ is correctly specified, the CIs would provide correct coverage probabilities when $p = 50$. When $m(\cdot)$ is not correctly specified, the performance of the CIs is in general not good in terms of mean coverage probabilities, but these CIs still have a reasonable median coverage probabilities (see Table 6.3(b)). This implies that in the moderate dimensional case, there are some particular coefficients whose corresponding CIs have very bad coverage probabilities when the model is misspecified while some others still have desired value. When the true DGP is "SIM-SIM", different working models provide similar results. The pattern is the same as the one we observe in the estimation error table.

When $p = 500$, the coverage probabilities of the CIs are slightly below

95% even when the working models are correctly specified. This is the price we have to pay for estimating the influence function for $\boldsymbol{\theta}_0$ and the precision matrix under high dimensional setting. These finite sample biases would be reduced with larger sample sizes. When $m(\cdot)$ is not correctly specified, the performance of the CIs is not good (see Table 6.4(c) and Table B.4(c)). These provide examples showing the benefits of correctly specifying $m(\cdot)$ under the "SIM-SIM" DGP. However, due to the limited sample size, the differences between the coverage probabilities of the CIs obtained using correct and incorrect working models are not distinguishable. The pattern is much clearer when we examine the large sample setting ($n = 50000$). In this scenario, the median coverage probabilities for the misspecified working models are also below 95%, which is different from $p = 50$.

In addition, the average lengths of the CIs are in general shorter for the models that correctly specifies $m(\cdot)$ (see Table 6.3(b)(c) and Table 6.4(b)(c)). Those misspecified working models provide CIs that have poor coverage probabilities even with larger lengths.

6.4. *Investigating double-robustness of the proposed estimator and the performance of complete case estimator via large sample results.* To investigate whether our proposed estimator has the desired double-robustness property, we study a large sample setting where $n = 50000$, $p = 50$ and 500. This aims to study the asymptotic properties of our estimator. When either the propensity score $\pi(\cdot)$ or the conditional mean $m(\cdot)$ is correctly specified, our estimator should be consistent. In addition, aside from the oracle and super oracle estimators, we also consider the complete case estimator $\widehat{\boldsymbol{\theta}}_{cc}$, which is an estimator obtained by using only the complete data (samples with $T = 1$). This estimator is known to be consistent only when the DGP is "linear-linear".

Table 6.5 and Table 6.6 provide the results of estimation errors of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$, coverage probabilities and lengths of the CIs. We could see the double-robustness of our estimator through the estimation errors. When both working nuisance models are correctly specified, the estimation errors achieve the oracle estimator and are close to the super oracle estimator. In addition, when only the conditional mean $m(\cdot)$ is correctly specified, it could have similar performance comparing to both correctly specified. This is consistent with the case when $n = 1000$. When only the propensity score is correctly specified, the errors are smaller than the case when both are misspecified but cannot reach the same level as the correctly specified case. This is due to the convergence rate is slow in such cases (see the discussion in Section A). Another thing to notice is the complete case estimator. As we

Table 6.5 Table (a): The L_2 errors of the estimator comparing with oracle values under the setting of $n = 50000$ using identity covariance matrix and "quad-quad" DGP. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ and different estimators are compared. Table (b): Average coverage probabilities and lengths of the CIs built upon the desparsified estimator. We report the means and medians together with standard errors and MADs as subscripts. The reported values are separated into truly zero and non-zero coefficients.

(I) $p = 50$.

(a) L_2 errors of the estimator.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.46 (0.026)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.204 (0.137)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.071 (0.01)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.072 (0.011)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.323 (0.019)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.175 (0.079)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)

(b) Average coverage probabilities and lengths of the CIs.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.03})$	0.06_{0}	$0.68_{0.39}$	$(0.84_{0.19})$	$0.07_{0.02}$
	$\widehat{\pi}$: quad	$0.96_{0.02}$	$(0.96_{0.01})$	$0.12_{0.01}$	$0.96_{0.02}$	$(0.96_{0.03})$	$0.14_{0.08}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.02})$	0.05_{0}	$0.93_{0.03}$	$(0.95_{0.01})$	$0.05_{0.01}$
	$\widehat{\pi}$: quad	$0.94_{0.03}$	$(0.95_{0.03})$	0.05_{0}	$0.94_{0.02}$	$(0.95_{0.01})$	$0.05_{0.01}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.94_{0.01})$	0.06_{0}	$0.8_{0.19}$	$(0.88_{0.13})$	$0.07_{0.01}$
	$\widehat{\pi}$: quad	$0.95_{0.02}$	$(0.95_{0.03})$	0.1_{0}	$0.95_{0.02}$	$(0.95_{0.02})$	$0.12_{0.06}$

Table 6.6 See caption of Table 6.5.

(I) $p = 500$.

(a) L_2 errors of the estimator.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.297 (0.017)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.282 (0.113)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.177 (0.008)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.18 (0.01)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.407 (0.022)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.294 (0.045)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)

(b) Average coverage probabilities and lengths of the CIs.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.02}$	$(0.95_{0.03})$	0.07_{0}	$0.78_{0.32}$	$(0.94_{0.04})$	$0.07_{0.01}$
	$\widehat{\pi}$: quad	$0.95_{0.02}$	$(0.96_{0.01})$	0.09_{0}	$0.94_{0.04}$	$(0.96_{0.03})$	$0.1_{0.03}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.02}$	$(0.95_{0.01})$	0.05_{0}	$0.94_{0.02}$	$(0.94_{0.02})$	$0.05_{0.01}$
	$\widehat{\pi}$: quad	$0.95_{0.02}$	$(0.95_{0.01})$	0.05_{0}	$0.94_{0.02}$	$(0.94_{0.02})$	$0.05_{0.01}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.02}$	$(0.95_{0.03})$	0.08_{0}	$0.75_{0.38}$	$(0.94_{0.05})$	$0.09_{0.01}$
	$\widehat{\pi}$: quad	$0.95_{0.02}$	$(0.95_{0.01})$	0.08_{0}	$0.88_{0.12}$	$(0.92_{0.04})$	$0.09_{0.02}$

stated, only in "linear-linear" DGP is the complete case estimator consistent. This is clearly revealed in the last column of the Table 6.5 and 6.6. The lengths of the CIs show similar pattern as in the case of $n = 1000$. For those working models that only specify one of $\pi(\cdot)$ and $m(\cdot)$ correct, the CIs have desired coverage probabilities but are wider than the CIs obtained when both are correct. For the working models that misspecify both, the coverage probabilities are very poor not only in mean, but also in median.

7. Discussion. In this paper, we study the high dimensional M-estima– tion problem with missing outcomes. With the response Y possibly missing at random and high dimensional covariates, we consider estimation and inference problem for the target parameter $\boldsymbol{\theta}_0$, which is defined as the minimizer of the risk of a convex loss. This parameter of interest is defined in such a way that it is a high dimensional parameter under a fully non-parametric model. This framework includes standard regression problem with missing outcomes and is also applicable to causal inference literature such as heterogeneous treatment effects estimation.

We propose an L_1-regularized debiased and doubly robust (DDR) estimator for $\boldsymbol{\theta}_0$ and carefully study its properties. Under proper assumptions and high-order consistency in estimating the working models for the propensity score and the outcome model, we provide the finite sample non-asymptotic estimation error bounds for $\boldsymbol{\theta}_0$. When both working models are correctly specified, we propose a desparsified estimator using an one-step update procedure that achieves the semi-parametric efficiency bound. Meanwhile, we provide theoretical results on estimating the nuisance functions (propensity score and outcome models) using linear and non-linear, parametric and semi-parametric models, which expands the current literature.

We also investigate the double robustness of our estimator showing its consistency even if only one of the working models on the propensity score and the outcome models is correctly specified. We include both the theoretical results and the simulations that examine this property in the appendix. The sharp rates of our proposed estimator under more general settings requires case-by-case studies and remains an open problem.

SUPPLEMENTARY MATERIAL

Supplementary Materials for "High Dimensional M-Estimation with Missing Outcomes: A Semi-Parametric Framework" (.pdf file). In the Supplementary Material (Appendices A-K), we collect several important materials that could not be accommodated in the main manuscript.

REFERENCES

Alquier, P. and Biau, G. (2013). Sparse Single-Index Model. *Journal of Machine Learning Research* **14** 243–280.

Andrews, D. W. K. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* **11** 560-586.

Athey, S., Imbens, G. W. and Wager, S. (2016). Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. (ArXiv version: arXiv:1604.07125v5).

Avagyan, V. and Vansteelandt, S. (2017). Honest Data-Adaptive Inference for the Average Treatment Effect under Model Misspecification using Penalised Bias-Reduced Double-Robust Estimation. *ArXiv preprint arXiv:1708.03787*.

Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference models. *Biometrics* **61** 962–973.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies* **81** 608–650.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica* **85** 233–298.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Bühlmann, P. and van de Geer, S. (2015). High-Dimensional Inference in Misspecified Linear Models. *Electronic Journal of Statistics* **9** 1449–1473.

Buldygin, V. V. and Moskvichova, K. K. (2013). The Sub-Gaussian Norm of a Binary Random Variable. *Theory of Probability and Mathematical Statistics* **86** 33-49.

Cai, T. T. and Guo, Z. (2017). Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity. *The Annals of Statistics* **45** 615–646.

Chernozhukov, V., Newey, W. and Robins, J. (2018). Double/De-Biased Machine Learning using Regularized Riesz Representers. *ArXiv preprint arXiv:1802.08667*.

Chernozhukov, V. and Semenova, V. (2017). Simultaneous Inference for Best Linear Predictor of the Conditional Average Treatment Effect and Other Structural Functions. *ArXiv preprint arXiv:1702.06240v2*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2016). Locally Robust Semiparametric Estimation. *ArXiv preprint arXiv:1608.00033v2*.

Chernozhukov, V., Demirer, M., Duflo, E. and Fernandez-Val, I. (2017a). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *ArXiv preprint arXiv:1712.04802*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017b). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review* **107** 261–65. (ArXiv version: arXiv:1701.08687).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018a). Double/Debiased Machine Learning for Treatment and Structural parameters. *The Econometrics Journal* **21** C1–C68. (ArXiv version: arXiv:1608.00060v5).

Chernozhukov, V., Nekipelov, D., Semenova, V. and Syrgkanis, V. (2018b). Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models. *ArXiv preprint arXiv:1806.04823*.

Farrell, M. H. (2015). Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics* **189** 1–23.

Farrell, M. H., Liang, T. and Misra, S. (2018). Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands. *ArXiv preprint arXiv:1809.09953*.

Genzel, M. (2017). High-Dimensional Estimation of Structured Signals from Non-Linear Observations with General Convex Loss Functions. *IEEE Transactions on Information Theory* **63** 1601–1619.

Goldstein, L., Minsker, S. and Wei, X. (2016). Structured Signal Recovery from Non-Linear and Heavy-Tailed measurements. *ArXiv preprint arXiv:1609.01025*.

Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica* **79** 437–452.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* **24** 726-748.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics* **12**. Springer.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and statistics* **86** 4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Javanmard, A. and Montanari, A. (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research* **15** 2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal Sample Size for Gaussian Designs. *The Annals of Statistics* **46** 2593–2622.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data (with Discussions and Rejoinder). *Statistical Science* **22** 523–580.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving Beyond Sub-Gaussianity in High Dimensional Statistics: Applications in Covariance Estimation and Linear Regression. *ArXiv preprint arXiv:1804.02605v2*.

Lecué, G. and Mendelson, S. (2014). Sparse Recovery under Weak Moment Assumptions. *ArXiv preprint arXiv:1401.2188*.

Li, K.-C. and Duan, N. (1989). Regression Analysis under Link Violation. *The Annals of Statistics* **17** 1009-1052.

Loh, P.-L. (2017). Statistical Consistency and Asymptotic Normality for High-Dimensional Robust M-Estimators. *The Annals of Statistics* **45** 866–896.

Loh, P.-L. and Wainwright, M. J. (2012). High-Dimensional Regression With Noisy and Missing Data: Provable Guarantees with Nonconvexity. *The Annals of Statistics* **40** 1637.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-Estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *Journal of Machine Learning Research* **16** 559–616.

Masry, E. (1996). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* **17** 571-600.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science* **27** 538-557. (Extended ArXiv version: arXiv:1010.2731v1).

Newey, W. K. and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics* **4** 2111–2245.

Newey, W. K. and Robins, J. M. (2018). Cross-Fitting and Fast Remainder Rates for

Semiparametric Estimation. *ArXiv preprint arXiv:1801.09138*.

Plan, Y. and Vershynin, R. (2013). Robust 1-Bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach. *IEEE Transactions on Information Theory* **59** 482-494.

Plan, Y. and Vershynin, R. (2016). The Generalized Lasso with Non-Linear Observations. *IEEE Transactions on Information Theory* **62** 1528–1537.

Pollard, D. (2015). A Few Good Inequalities Book Chapter, Department of Statistics, Yale University. (Available at www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf).

Radchenko, P. (2015). High Dimensional Single Index Models. *Journal of Multivariate Analysis* **139** 266–282.

Rigollet, P. and Hütter, J.-C. (2017). Sub-Gaussian Random Variables. In *High Dimensional Statistics* 1 Massachusetts Institute of Technology OpenCourseWare Lecture notes. (Available at http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf).

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.

Rudelson, M. and Zhou, S. (2013). Reconstruction from Anisotropic Random Measurements. *IEEE Transactions on Information Theory* **59** 3434–3447.

Smucler, E., Rotnitzky, A. and Robins, J. M. (2019). A Unifying Approach for Doubly-Robust l_1 Regularized Estimation of Causal Contrasts. *ArXiv preprint arXiv:1904.03737v1*.

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 de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On Asymptotically Optimal Cnfidence Regions and Tests for High-Dimensional Models. *The Annals of Statistics* **42** 1166–1202.

Van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.

Vershynin, R. (2012). Introduction to the Non-Asymptotic Analysis of Random Matrices. In *Compressed Sensing: Theory and Applications* 210-268. Cambridge University Press (ArXiv preprint available at arXiv:1011.3027v7).

Vershynin, R. (2018). *High Dimensional Probability. An Introduction with Applications in Data Science* **47**. Cambridge University Press (Available from the author's website).

Wager, S. and Athey, S. (2017). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*. (To appear). (ArXiv version: arXiv:1510.04342v2).

Wainwright, M. J. (2019). *High Dimensional Statistics: A Non-Asymptotic Viewpoint* **48**. Cambridge University Press.

Wei, X. (2018). Structured Recovery with Heavy-tailed Measurements: A Thresholding Procedure and Optimal Rates. *ArXiv preprint arXiv:1804.05959*.

SUPPLEMENTARY MATERIALS FOR "HIGH DIMENSIONAL M-ESTIMATION WITH MISSING OUTCOMES: A SEMI-PARAMETRIC FRAMEWORK"

BY ABHISHEK CHAKRABORTTY, JIARUI LU, T. TONY CAI AND HONGZHE LI

University of Pennsylvania

APPENDIX A: DOUBLE ROBUSTNESS OF THE DDR ESTIMATOR

Our probabilistic analysis of $\|\mathbf{T}_n\|_\infty$ for establishing the convergence rate of $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ (in the light of Lemma 2.1) has so far assumed that both the nuisance functions $\{\pi(\cdot), m(\cdot)\}$ are correctly estimated via $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ satisfying Assumptions 3.2-3.3. As noted in (2.2), the nature of the population DDR loss $\mathbb{L}_{\text{DDR}}(\cdot)$ and the empirical version $\mathcal{L}_n^{\text{DDR}}(\cdot)$ is such that consistency of $\|\mathbf{T}_n\|_\infty$ (and hence $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$) should hold even if only one of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ is correct.

In this section, we briefly sketch the arguments that ensure *consistency* of $\|\mathbf{T}_n\|_\infty$ even if *only one* of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ is correctly specified but *not necessarily both.* The convergence rates underlying this consistency, while reasonable, are not necessarily sharp however. To obtain sharper rates (if possible at all) under these general situations, one needs a far more nuanced case-by-case analysis which *will depend* now on the construction of the estimators and their first order properties and rates, unlike the case when both the estimators are correctly specified and the results are first order insensitive (see Remark 3.1) requiring no specific knowledge about the estimators except for some high-level convergence properties. This is true even for classical settings and the high dimensional setting here only lends further complexity and subtlety to the issue. Considering the main goals and scope of this paper, we suppress such finer analysis under those cases for the sake of simplicity.

Case 1. Suppose that $\widehat{\pi}(\cdot)$ is misspecified, such that $\widehat{\pi}(\mathbf{x}) \xrightarrow{\mathbb{P}} \pi^*(\mathbf{x}) \neq \pi(\mathbf{x})$ following Assumption 3.2 with $\pi(\cdot)$ therein replaced by a general $\pi^*(\cdot)$, while $\widehat{m}(\cdot)$ is still correctly specified with $\widehat{m}(\mathbf{x}) \xrightarrow{\mathbb{P}} m(\mathbf{x})$ following Assumption 3.3. In this case, the terms $\mathbf{T}_{0,n}$ and $\mathbf{T}_{m,n}$ in the decomposition (3.1) of $\mathbf{T}_n$ will stay unaffected and their properties still governed by the results of Theorems 3.1 and 3.3 respectively, while the error terms $\mathbf{T}_{\pi,n}$ and $\mathbf{R}_{\pi,m,n}$ involving $\widehat{\pi}(\cdot)$ would be affected and need to be appropriately analyzed as follows.

$\mathbf{T}_{\pi,n}$ should be further decomposed into two terms as: $\mathbf{T}_{\pi,n} = \widetilde{\mathbf{T}}_{\pi,n} + \mathbf{T}^*_{\pi,n}$,

$$\text{where } \widetilde{\mathbf{T}}_{\pi,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{T_i}{\pi^*(\mathbf{X}_i)}\right\}\{Y_i - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i)$$
$$\text{and } \mathbf{T}^*_{\pi,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi^*(\mathbf{X}_i)} - \frac{T_i}{\pi(\mathbf{X}_i)}\right\}\{Y_i - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i),$$

while $\mathbf{R}_{\pi,m,n}$ should be decomposed further as: $\mathbf{R}_{\pi,m,n} = \widetilde{\mathbf{R}}_{\pi,m,n} + \mathbf{R}^*_{\pi,m,n}$,

$$\text{where } \widetilde{\mathbf{R}}_{\pi,m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{T_i}{\pi^*(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i)$$
$$\text{and } \mathbf{R}^*_{\pi,m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi^*(\mathbf{X}_i)} - \frac{T_i}{\pi(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i).$$

Suppose Assumption 3.2 is modified appropriately with $\pi(\cdot)$ therein replaced throughout by $\pi^*(\cdot)$, the true target function of $\widehat{\pi}(\cdot)$ in this case, and assume also that $\pi^*(\mathbf{X}) > \delta^*_\pi > 0$ for some constant δ^*_π, and $\pi^*(\mathbf{X}) - \pi(\mathbf{X})$ is bounded (or sub-Gaussian). Then, under Assumptions 1.1 and 3.1-3.3, using similar arguments as those used in the proofs of Theorems 3.1-3.2 (for $\mathbf{T}^*_{\pi,n}$ and $\widetilde{\mathbf{T}}_{\pi,n}$ respectively) and Theorem 3.4 (for $\widetilde{\mathbf{R}}_{\pi,m,n}$ and $\mathbf{R}^*_{\pi,m,n}$), it can be shown that

$$\|\widetilde{\mathbf{T}}_{\pi,n}\|_\infty \lesssim v_{n,\pi}\sqrt{\log(nd)}\sqrt{\frac{\log d}{n}} \text{ and } \|\mathbf{T}^*_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \text{ w.h.p., and}$$
$$\|\widetilde{\mathbf{R}}_{\pi,m,n}\|_\infty \lesssim v_{n,\pi}v_{\bar{n},m}(\log n) \text{ and } \|\mathbf{R}^*_{\pi,m,n}\|_\infty \lesssim v_{\bar{n},m}\sqrt{\log n} \text{ w.h.p.}$$

Case 2. Suppose $\widehat{m}(\cdot)$ is misspecified instead with $\widehat{m}(\mathbf{x}) \xrightarrow{\mathbb{P}} m^*(\mathbf{x}) \neq m(\mathbf{x})$ according to Assumption 3.3 with $m(\cdot)$ replaced by a general $m^*(\cdot)$ therein, while $\widehat{\pi}(\cdot)$ is still correctly specified with $\widehat{\pi}(\mathbf{x}) \xrightarrow{\mathbb{P}} \pi(\mathbf{x})$ following Assumption 3.2. In this case, the terms $\mathbf{T}_{0,n}$ and $\mathbf{T}_{\pi,n}$ in the decomposition (3.1) of $\mathbf{T}_n$ stay unaffected and their properties still governed by the results of Theorems 3.1 and 3.2 respectively, while the error terms $\mathbf{T}_{m,n}$ and $\mathbf{R}_{\pi,m,n}$ involving $\widehat{m}(\cdot)$ would be affected and need to be appropriately analyzed as follows.

$\mathbf{T}_{m,n}$ may be further decomposed into two terms as: $\mathbf{T}_{m,n} = \widetilde{\mathbf{T}}_{m,n} + \mathbf{T}^*_{m,n}$,

$$\text{where } \widetilde{\mathbf{T}}_{m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\mathbf{X}_i)} - 1\right\}\{\widetilde{m}(\mathbf{X}_i) - m^*(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i)$$
$$\text{and } \mathbf{T}^*_{m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\mathbf{X}_i)} - 1\right\}\{m^*(\mathbf{X}_i) - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i),$$

while $\mathbf{R}_{\pi,m,n}$ should be decomposed further as: $\mathbf{R}_{\pi,m,n} = \mathbf{R}^{\dagger}_{m,n} + \mathbf{R}^{**}_{\pi,m,n}$,

$$\text{where } \mathbf{R}^{\dagger}_{m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{T_i}{\pi(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m^*(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i)$$

$$\text{and } \mathbf{R}^{**}_{\pi,m,n} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{T_i}{\pi(\mathbf{X}_i)}\right\}\{m^*(\mathbf{X}_i) - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i).$$

Suppose Assumption 3.3 is modified appropriately whereby $m(\cdot)$ is replaced throughout by $m^*(\cdot)$, the true target function of $\widehat{m}(\cdot)$ in this case. Further, assume also that $m^*(\mathbf{X}) - m(\mathbf{X})$ is sub-Gaussian. Then, under Assumptions 1.1 and 3.1-3.3, using similar arguments as those in the proofs of Theorems 3.1 and 3.3 (for $\mathbf{T}^*_{m,n}$ and $\widetilde{\mathbf{T}}_{m,n}$ respectively) and Theorem 3.4 (for $\mathbf{R}^{\dagger}_{m,n}$ and $\mathbf{R}^{**}_{\pi,m,n}$), it is not difficult to show that the following hold:

$$\|\widetilde{\mathbf{T}}_{m,n}\|_\infty \lesssim v_{\bar{n},m}\sqrt{\log(nd)}\sqrt{\frac{\log d}{n}} \text{ and } \|\mathbf{T}^*_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \text{ w.h.p., and}$$
$$\|\mathbf{R}^{\dagger}_{m,n}\|_\infty \lesssim v_{n,\pi}v_{\bar{n},m}(\log n) \text{ and } \|\mathbf{R}^{**}_{\pi,m,n}\|_\infty \lesssim v_{n,\pi}\sqrt{\log n} \text{ w.h.p.}$$

Combining the results over the two cases, under a general setting allowing for misspecification of either $\widehat{\pi}(\cdot)$ or $\widehat{m}(\cdot)$, the terms in (3.1) therefore satisfy:

(A.1)

$$\|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}\{1 + 1_{(\pi^*,m^*)\neq(\pi,m)} + o(1)\}$$
$$\text{and } \|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \{v_{n,\pi}1_{(m^*\neq m)} + v_{\bar{n},m}1_{(\pi^*\neq\pi)}\}\sqrt{\log n} + v_{n,\pi}v_{\bar{n},m}(\log n).$$

Hence, even under possible misspecification of one of the nuisance function estimators, $\|\mathbf{T}_n\|_\infty$ is certainly $o_{\mathbb{P}}(1)$ and thus double robust (in terms of consistency). Consequently, $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ is also double robust (in terms of consistency) in the light of Lemma 2.1 for an appropriately chosen $\lambda_n \geq 2\|\mathbf{T}_n\|_\infty = o_{\mathbb{P}}(1)$ as long as the corresponding the deviation bounds in (2.13) involving $\sqrt{s\lambda_n}$ (for L_2 consistency) and $s\sqrt{\lambda_n}$ (for L_1 consistency) are assumed to be $o(1)$.

It is important to note from (A.1) that under the misspecification of either $\widehat{\pi}(\cdot)$ or $\widehat{m}(\cdot)$, at least one among $\|\mathbf{T}_{\pi,n}\|_\infty$ and $\|\mathbf{T}_{m,n}\|_\infty$ is no longer a lower order term, but instead contributes an extra term of order $\sqrt{(\log d)/n}$, same as the main term $\mathbf{T}_{0,n}$, while the other one stays to be of lower order. More importantly, however, the behavior of the product-type bias (or 'drift') term $\mathbf{R}_{\pi,m,n}$ changes dramatically! From being a lower order term involving the products of the rates of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$, it now involves the individual rates themselves appearing as leading order terms in a complementary manner, i.e.

$v_{\bar{n},m}$ appears if $\widehat{\pi}(\cdot)$ is misspecified and $v_{n,\pi}$ appears if $\widehat{m}(\cdot)$ is misspecified. This is mainly due to the unavoidable appearance of the additional terms $\mathbf{R}^*_{\pi,m,n}$ or $\mathbf{R}^{**}_{\pi,m,n}$, and their control inevitably requires use of the first order properties and rates of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. In general, these rates are not necessarily of faster (or even same) order than $\sqrt{(\log d)/n}$. In fact, they are quite likely to be slower in most cases, especially if $\widehat{\pi}(\cdot)$ and/or $\widehat{m}(\cdot)$ are obtained based on non/semi-parametric models or high dimensional parametric models, in all of which cases the convergence rates are typically slower than $\sqrt{(\log d)/n}$.

Hence, under misspecification of $\widehat{\pi}(\cdot)$ or $\widehat{m}(\cdot)$, the L_2 convergence rate of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ is likely to be slower than the usual benchmark rate of $\sqrt{s(\log d)/n}$. To achieve estimators with faster rates, one needs to carefully incorporate further bias corrections while constructing the estimator itself given a choice of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. This is quite a challenging problem in high dimensional settings, even for the simple case of mean (or ATE) estimation and with $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ obtained using standard high dimensional sparse parametric models. This case has been considered only recently by Avagyan and Vansteelandt (2017) and Smucler, Rotnitzky and Robins (2019), where the methods and the associated analyses are evidently quite involved. We refer the interested reader to these papers for further insights on the problem and the ensuing challenges and nuances. However, given the scope of this paper, we do not delve further into such analyses for brevity, especially since in our case, the parameter is also high dimensional which lends further complexity to the problem. But we do empirically investigate in detail and validate the double robustness of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ and $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ in our simulation studies; see Section 6 for all the results.

APPENDIX B: SUPPLEMENTARY NUMERICAL RESULTS

B.1. Simulation Setting: Technical Details. In this section we provide more technical details of the simulation. The parameters in the DGPs are specified as the following:

(a) Choices of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}_0$:

   (i) When $p = 50$, we set $\|\boldsymbol{\alpha}\|_0 = 5$ and $\|\boldsymbol{\alpha}^*\|_0 = 2$ with:

$$\boldsymbol{\alpha} = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, \mathbf{0}_{p-5}),$$
$$\boldsymbol{\alpha}^* = (0.25, -0.25, \mathbf{0}_{p-2}).$$

   (ii) When $p = 500$, we set $\|\boldsymbol{\alpha}\|_0 = 10$ and $\|\boldsymbol{\alpha}^*\|_0 = 4$ with:

$$\boldsymbol{\alpha} = 1/\sqrt{10}(\mathbf{1}_3, -\mathbf{1}_2, \mathbf{0.5}_2, -\mathbf{0.5}_3, \mathbf{0}_{p-10}),$$
$$\boldsymbol{\alpha}^* = (\mathbf{0.25}_2, -\mathbf{0.25}_2, \mathbf{0}_{p-4}).$$

(b) Choices of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^*$ and $\boldsymbol{\gamma}_0$:

(i) When $p = 50$, we set $\|\boldsymbol{\gamma}\|_0 = 10$ and $\|\boldsymbol{\gamma}^*\|_0 = 5$ with:

$$\boldsymbol{\gamma} = (\mathbf{1}_3, -\mathbf{1}_2, \mathbf{0.5}_2, -\mathbf{0.5}_3, \mathbf{0}_{p-10}),$$
$$\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, \mathbf{0}_{p-5}).$$

(ii) When $p = 500$, we set $\|\boldsymbol{\gamma}\|_0 = 20$, $\|\boldsymbol{\gamma}^*\|_0 = 5$ with:

$$\boldsymbol{\gamma} = (\mathbf{1}_3, -\mathbf{1}_2, \mathbf{0.5}_5, -\mathbf{0.5}_5, \mathbf{0.25}_2, -\mathbf{0.25}_3, \mathbf{0}_{p-20}),$$
$$\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, \mathbf{0}_{p-5}).$$

In the SIM, c_T and c_Y are set to be 0.2 and $\dfrac{0.3}{\sqrt{\lambda_{max}(\boldsymbol{\Sigma}_p)}}$ and the intercepts $\boldsymbol{\gamma}_0$ and $\boldsymbol{\alpha}_0$ are set to be 1 and 0.5. Here $\lambda_{max}(\boldsymbol{\Sigma}_p)$ is the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_p$ and we define the notation $\mathbf{a}_d := (\underbrace{a, a, \ldots, a}_{d})$. The parameters in the DGPs for $T|\mathbf{X}$ are normalized by $\sqrt{\|\boldsymbol{\alpha}\|_0}$ so that the proportion of $\pi(\mathbf{X})$ that is close to 0 or 1 is small.

The number of folds in the sample splitting is 2. The tuning parameters in fitting the penalized logistic regression for $\pi(\cdot)$ are selected using Bayesian information criterion (BIC) and the tuning parameters in the penalized regression for $m(\cdot)$ are selected using 10-fold cross validation with minimizing mean squared errors (MSE) as criterion. The band-width in the nonparametric regression for SIM is chosen using least square cross-validation as suggested in the "np" package in R. All the codes are implemented in R and will be provided upon request.

As a summary, the algorithm for obtaining the *DDR estimator* $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$, the *desparsified DDR estimator* $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ and its confidence interval is given in Algorithm 1:

B.2. Simulations with non-identity covariance structures. In this section we provide some additional simulation results. Aside from identity covariance matrix, we also study the case when the covariance matrix $\boldsymbol{\Sigma}_p$ is AR(1) and compound symmetric.

When the covariance matrix is AR(1), the results are given in Table B.1, B.2, B.3 and B.4. Overall, the results are consistent with the scenario when $\boldsymbol{\Sigma}_p$ is identity matrix. The estimation errors for two choices of the covariance matrix are close to each other, drawing the similar conclusions as the identity case. This is because the AR(1) covariance matrix with a relatively small $\rho = 0.2$ is very close to an identity matrix. One thing to notice is that in

Algorithm 1 Summarized Algorithm for Obtaining $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ and $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$

Input: Generated data $\mathcal{D}_n := \{\mathbf{Z}_i \equiv (T_i, T_iY_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ based on one of the DGPs.

Output: $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$,$\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ and corresponding confidence intervals.

1: **Estimation of nuisance functions $\pi(\cdot)$ and $m(\cdot)$ given observed data.**
   - (i) Estimation of $\pi(\cdot)$: One of the two working models for $\pi(\cdot)$.
   - (ii) Estimation of $m(\cdot)$: One of the three working models for $m(\cdot)$.

2: **Obtaining $\widehat{\boldsymbol{\theta}}_{\mathbf{DDR}}$ using the L_1 regularized DDR estimator**
   - (i) Obtain the sample-split version $\widetilde{m}(\cdot)$ of $\widehat{m}(\cdot)$ based on (2.7).
   - (ii) Obtain the pseudo outcome $\widetilde{Y}$ using $\widetilde{m}(\cdot)$ and $\widehat{\pi}(\cdot)$ based on (2.10).
   - (iii) Fit an L_1 penalized linear regression using pseudo outcomes $\{\widetilde{Y}_i, \mathbf{X}_i\}_{i=1}^n$ to obtain $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$.

3: **Desparsified estimator of $\widehat{\boldsymbol{\theta}}_{\mathbf{DDR}}$**
   - (i) Obtain an estimator for $\boldsymbol{\Omega}$, denoted as $\widehat{\boldsymbol{\Omega}}$. There are two possible methods for estimating $\boldsymbol{\Omega}$:
     - (a) When p is relatively small comparing with n: invert $\widehat{\boldsymbol{\Sigma}}$ directly.
     - (b) Otherwise, use node-wise Lasso (ref).
   - (ii) Compute the *desparsified DDR estimator* $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ by (4.1).
   - (iii) Compute the confidence intervals for $\boldsymbol{\theta}_0$ as discussed in Section 4.1.

Table B.1 The L_2 errors of the estimator comparing with oracle values under the setting of $n = 1000$ using AR(1) covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ and different estimators are compared.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.222 (0.038)	0.223 (0.038)	0.169 (0.028)
	π: quad	0.222 (0.038)	0.223 (0.038)	0.169 (0.028)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.224 (0.038)	0.223 (0.038)	0.169 (0.028)
	$\widehat{\pi}$: quad	0.223 (0.038)	0.223 (0.038)	0.169 (0.028)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.222 (0.038)	0.223 (0.038)	0.169 (0.028)
	$\widehat{\pi}$: quad	0.222 (0.038)	0.223 (0.038)	0.169 (0.028)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.664 (0.107)	0.469 (0.075)	0.445 (0.074)
	π: quad	0.625 (0.104)	0.469 (0.075)	0.445 (0.074)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.464 (0.075)	0.469 (0.075)	0.445 (0.074)
	$\widehat{\pi}$: quad	0.464 (0.075)	0.469 (0.075)	0.445 (0.074)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.671 (0.109)	0.469 (0.075)	0.445 (0.074)
	$\widehat{\pi}$: quad	0.631 (0.106)	0.469 (0.075)	0.445 (0.074)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.569 (0.127)	0.478 (0.112)	0.459 (0.109)
	π: quad	0.567 (0.127)	0.478 (0.112)	0.459 (0.109)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.562 (0.126)	0.478 (0.112)	0.459 (0.109)
	π: quad	0.562 (0.126)	0.478 (0.112)	0.459 (0.109)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.499 (0.119)	0.478 (0.112)	0.459 (0.109)
	$\widehat{\pi}$: quad	0.498 (0.12)	0.478 (0.112)	0.459 (0.109)

Table B.2 See caption of Table B.1.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.42 (0.045)	0.401 (0.043)	0.295 (0.029)
	$\widehat{\pi}$: quad	0.419 (0.044)	0.401 (0.043)	0.295 (0.029)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.43 (0.046)	0.401 (0.043)	0.295 (0.029)
	$\widehat{\pi}$: quad	0.43 (0.046)	0.401 (0.043)	0.295 (0.029)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.409 (0.044)	0.401 (0.043)	0.295 (0.029)
	$\widehat{\pi}$: quad	0.408 (0.044)	0.401 (0.043)	0.295 (0.029)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	1.06 (0.112)	0.797 (0.084)	0.743 (0.077)
	$\widehat{\pi}$: quad	1.049 (0.109)	0.797 (0.084)	0.743 (0.077)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.814 (0.083)	0.797 (0.084)	0.743 (0.077)
	$\widehat{\pi}$: quad	0.814 (0.083)	0.797 (0.084)	0.743 (0.077)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	1.05 (0.11)	0.797 (0.084)	0.743 (0.077)
	$\widehat{\pi}$: quad	1.038 (0.109)	0.797 (0.084)	0.743 (0.077)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	1.026 (0.166)	1.001 (0.153)	0.974 (0.151)
	$\widehat{\pi}$: quad	1.016 (0.159)	1.001 (0.153)	0.974 (0.151)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	1.029 (0.162)	1.001 (0.153)	0.974 (0.151)
	$\widehat{\pi}$: quad	1.019 (0.157)	1.001 (0.153)	0.974 (0.151)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.961 (0.162)	1.001 (0.153)	0.974 (0.151)
	$\widehat{\pi}$: quad	0.952 (0.158)	1.001 (0.153)	0.974 (0.151)

Table B.3 Average coverage probabilities and lengths of the CIs built upon the desparsified estimator under the setting of $n = 1000$ using AR(1) covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ are compared. We report the means and medians together with standard errors and MAD as subscripts. The reported values are separated into truly zero and non-zero coefficients.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.94_{0.02})$	0.17_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.94_{0.02})$	0.17_{0}
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.94_{0.02})$	0.17_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.95_{0.02})$	0.17_{0}
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.17_{0}

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.42_{0}	$0.89_{0.14}$	$(0.94_{0.02})$	$0.47_{0.08}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.42_{0}	$0.9_{0.12}$	$(0.94_{0.02})$	$0.47_{0.07}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.34_{0}	$0.95_{0.01}$	$(0.95_{0.01})$	$0.38_{0.05}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.34_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.38_{0.05}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.42_{0}	$0.89_{0.14}$	$(0.94_{0.01})$	$0.47_{0.07}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.42_{0}	$0.9_{0.12}$	$(0.94_{0.02})$	$0.47_{0.07}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.43_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.48_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.94_{0.01})$	0.43_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.48_{0.03}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.42_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.47_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.42_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.47_{0.03}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.37_{0}	$0.94_{0.01}$	$(0.95_{0.01})$	$0.41_{0.02}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.37_{0}	$0.94_{0.01}$	$(0.95_{0.01})$	$0.41_{0.02}$

Table B.4 See caption of table B.3.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.17_0
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.17_0
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.02})$	0.17_0
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.17_0	$0.91_{0.02}$	$(0.91_{0.02})$	0.17_0
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.91_{0.01}$	$(0.91_{0.01})$	0.17_0
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.16_0	$0.91_{0.01}$	$(0.91_{0.02})$	0.17_0

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.44_0	$0.92_{0.03}$	$(0.93_{0.02})$	$0.46_{0.07}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.43_0	$0.91_{0.03}$	$(0.92_{0.02})$	$0.46_{0.06}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.33_0	$0.92_{0.02}$	$(0.92_{0.02})$	$0.35_{0.04}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.33_0	$0.92_{0.02}$	$(0.92_{0.02})$	$0.35_{0.04}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.44_0	$0.91_{0.03}$	$(0.92_{0.02})$	$0.46_{0.07}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.43_0	$0.91_{0.03}$	$(0.91_{0.02})$	$0.46_{0.06}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.52_0	$0.88_{0.04}$	$(0.88_{0.04})$	$0.55_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.52_0	$0.87_{0.04}$	$(0.87_{0.04})$	$0.55_{0.03}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.52_0	$0.88_{0.03}$	$(0.88_{0.03})$	$0.55_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.52_0	$0.88_{0.03}$	$(0.88_{0.04})$	$0.55_{0.03}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.95_{0.01})$	0.48_0	$0.93_{0.01}$	$(0.94_{0.01})$	$0.51_{0.03}$
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.95_{0.01})$	0.48_0	$0.93_{0.01}$	$(0.94_{0.01})$	$0.51_{0.03}$

Table B.2(c), the estimation errors of the "$\widehat{m}$: SIM" are slight better than the oracle and super oracle estimator. This is in general not the case.

When the covariance matrix is compound symmetric matrix, the results are given in table B.5 and B.7 for $p = 50$ and in table B.6 and B.8 for $p = 500$. Notice that a compound symmetric matrix is not sparse, so the node-wise Lasso method is not theoretically guaranteed to work when $p = 500$. The general pattern stays the same as identity and AR(1) covariance matrix structure. When $p = 500$, having errors in estimating the precision matrices and the influence function leads to slightly lower coverage probabilities even when the working models are correctly specified.

Table B.5 The L_2 errors of the estimator comparing with oracle values under the setting of $n = 1000$ using compound symmetric covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ and different estimators are compared.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.245 (0.039)	0.247 (0.04)	0.185 (0.03)
	$\widehat{\pi}$: quad	0.245 (0.038)	0.247 (0.04)	0.185 (0.03)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.248 (0.039)	0.247 (0.04)	0.185 (0.03)
	$\widehat{\pi}$: quad	0.247 (0.039)	0.247 (0.04)	0.185 (0.03)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.246 (0.039)	0.247 (0.04)	0.185 (0.03)
	$\widehat{\pi}$: quad	0.246 (0.038)	0.247 (0.04)	0.185 (0.03)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.701 (0.126)	0.513 (0.088)	0.483 (0.083)
	$\widehat{\pi}$: quad	0.657 (0.118)	0.513 (0.088)	0.483 (0.083)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.509 (0.087)	0.513 (0.088)	0.483 (0.083)
	$\widehat{\pi}$: quad	0.509 (0.088)	0.513 (0.088)	0.483 (0.083)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.704 (0.126)	0.513 (0.088)	0.483 (0.083)
	$\widehat{\pi}$: quad	0.662 (0.119)	0.513 (0.088)	0.483 (0.083)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.284 (0.052)	0.272 (0.047)	0.224 (0.042)
	$\widehat{\pi}$: quad	0.282 (0.052)	0.272 (0.047)	0.224 (0.042)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.287 (0.052)	0.272 (0.047)	0.224 (0.042)
	$\widehat{\pi}$: quad	0.285 (0.052)	0.272 (0.047)	0.224 (0.042)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.275 (0.048)	0.272 (0.047)	0.224 (0.042)
	$\widehat{\pi}$: quad	0.274 (0.048)	0.272 (0.047)	0.224 (0.042)

Table B.6 See caption of table B.5.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.492 (0.055)	0.466 (0.05)	0.35 (0.032)
	$\widehat{\pi}$: quad	0.492 (0.055)	0.466 (0.05)	0.35 (0.032)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.509 (0.059)	0.466 (0.05)	0.35 (0.032)
	$\widehat{\pi}$: quad	0.508 (0.059)	0.466 (0.05)	0.35 (0.032)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.483 (0.053)	0.466 (0.05)	0.35 (0.032)
	$\widehat{\pi}$: quad	0.483 (0.053)	0.466 (0.05)	0.35 (0.032)

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	1.245 (0.131)	0.949 (0.094)	0.89 (0.087)
	$\widehat{\pi}$: quad	1.236 (0.13)	0.949 (0.094)	0.89 (0.087)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.972 (0.1)	0.949 (0.094)	0.89 (0.087)
	$\widehat{\pi}$: quad	0.973 (0.1)	0.949 (0.094)	0.89 (0.087)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	1.251 (0.128)	0.949 (0.094)	0.89 (0.087)
	$\widehat{\pi}$: quad	1.24 (0.128)	0.949 (0.094)	0.89 (0.087)

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.46 (0.055)	0.463 (0.051)	0.364 (0.036)
	$\widehat{\pi}$: quad	0.458 (0.055)	0.463 (0.051)	0.364 (0.036)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.473 (0.057)	0.463 (0.051)	0.364 (0.036)
	$\widehat{\pi}$: quad	0.472 (0.057)	0.463 (0.051)	0.364 (0.036)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.466 (0.054)	0.463 (0.051)	0.364 (0.036)
	$\widehat{\pi}$: quad	0.465 (0.054)	0.463 (0.051)	0.364 (0.036)

Table B.7 Average coverage probabilities and lengths of the CIs built upon the desparsified estimator under the setting of $n = 1000$ using compound symmetric covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ are compared. We report the means and medians together with standard errors and MAD as subscripts. The reported values are separated into truly zero and non-zero coefficients.

(I) $p = 50$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	(0.94_{0})	0.18_{0}
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_{0}

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.45_{0}	$0.9_{0.11}$	$(0.94_{0.01})$	$0.49_{0.07}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.45_{0}	$0.9_{0.1}$	$(0.93_{0.02})$	$0.49_{0.06}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.37_{0}	$0.95_{0.01}$	$(0.95_{0.01})$	$0.41_{0.05}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.37_{0}	$0.95_{0.01}$	$(0.95_{0.01})$	$0.41_{0.05}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.95_{0.01})$	0.45_{0}	$0.9_{0.11}$	$(0.94_{0.02})$	$0.5_{0.07}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.95_{0.01})$	0.45_{0}	$0.9_{0.09}$	$(0.93_{0.02})$	$0.5_{0.06}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.21_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.22_{0.01}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.21_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.22_{0.01}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.21_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.23_{0.01}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.21_{0}	$0.94_{0.01}$	$(0.94_{0.01})$	$0.22_{0.01}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.02})$	0.2_{0}	$0.94_{0.01}$	$(0.95_{0.01})$	$0.21_{0.01}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.2_{0}	$0.94_{0.01}$	$(0.95_{0.01})$	$0.21_{0.01}$

Table B.8 See caption of table B.7.

(II) $p = 500$.

(a) DGP: "Linear-linear" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.91_{0.02}$	$(0.92_{0.01})$	0.18_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.91_{0.02}$	$(0.92_{0.01})$	0.18_0
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.19_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.19_0	$0.91_{0.02}$	$(0.91_{0.01})$	0.19_0
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.91_{0.01}$	$(0.91_{0.01})$	0.18_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.91_{0.01}$	$(0.91_{0.01})$	0.18_0

(b) DGP: "Quad-quad" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.95_{0.01})$	0.47_0	$0.91_{0.02}$	$(0.92_{0.01})$	$0.5_{0.06}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.47_0	$0.91_{0.03}$	$(0.91_{0.03})$	$0.49_{0.06}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.36_0	$0.92_{0.02}$	$(0.92_{0.01})$	$0.38_{0.04}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.36_0	$0.92_{0.02}$	$(0.92_{0.02})$	$0.38_{0.04}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.47_0	$0.91_{0.03}$	$(0.92_{0.02})$	$0.5_{0.06}$
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.47_0	$0.91_{0.03}$	$(0.91_{0.02})$	$0.49_{0.06}$

(c) DGP: "SIM-SIM" for $\pi(\cdot)$ and $m(\cdot)$.

Working nuisance model		Zero			Non-zero		
		CovP: Mean	(Median)	Length	CovP: Mean	(Median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.18_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.18_0
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.19_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.19_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.19_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.19_0
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.18_0
	$\widehat{\pi}$: quad	$0.94_{0.01}$	$(0.94_{0.01})$	0.18_0	$0.92_{0.01}$	$(0.92_{0.01})$	0.18_0

Table B.9 The L_2 errors of the estimator comparing with oracle values under the setting of $n = 50000$ and $p = 50$ using identity covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ and different estimators are compared.

(a) DGP: "Linear-linear" model for $\pi(\cdot)$ and $m(\cdot)$.

working nuisance model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.033 (0.005)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)
	$\widehat{\pi}$ quad	0.033 (0.005)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.033 (0.005)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)
	$\widehat{\pi}$: quad	0.033 (0.005)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.066 (0.011)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)
	$\widehat{\pi}$: quad	0.067 (0.011)	0.033 (0.005)	0.025 (0.003)	0.032 (0.004)

(b)DGP: "Quad-quad" model for $\pi(\cdot)$ and $m(\cdot)$.

working nuisance model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.46 (0.026)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.204 (0.137)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.071 (0.01)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.072 (0.011)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.323 (0.019)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.175 (0.079)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)

APPENDIX C: TECHNICAL TOOLS

We collect here some useful definitions and supporting lemmas that serve throughout as key technical ingredients in the proofs of all our main results.

C.1. Orlicz Norms, Sub-Gaussians and Sub-Exponentials. We first introduce a few definitions and results regarding concentration bounds.

Definition C.1 (Orlicz norms). For any $\alpha > 0$, let $\psi_\alpha(\cdot)$ denote the function given by: $\psi_\alpha(x) = \exp(x^\alpha) - 1 \ \forall \ x \geq 0$. Then, for any random variable X and any $\alpha > 0$, the ψ_α*-Orlicz norm* $\|X\|_{\psi_\alpha}$ of X is defined as:

$$\|X\|_{\psi_\alpha} := \inf \{c > 0 : \ \mathbb{E}\{\psi_\alpha(|X|/c)\} \ \leq \ 1\},$$

and X is said to have a finite ψ_α-Orlicz norm, denoted as $\|X\|_{\psi_\alpha} < \infty$ (if the set above is empty, then the infimum is simply defined to be ∞).

For a *random vector* $\mathbf{X} \in \mathbb{R}^d$ $(d \geq 1)$, we define $\mathbf{X}$ to have finite ψ_α-Orlicz norm if each coordinate of $\mathbf{X}$ does and we let $\|\mathbf{X}\|_{\psi_\alpha} := \max_{1 \leq j \leq d} \|\mathbf{X}_{[j]}\|_{\psi_\alpha}$.

A random variable (or random vector) is said to be *sub-Gaussian* or *sub-exponential* if it has finite ψ_α-Orlicz norm with $\alpha = 2$ or $\alpha = 1$ respectively.

Table B.10 Average coverage probabilities and lengths of the CIs built upon the desparsified estimator under the setting of $n = 50000$ and $p = 50$ using identity covariance matrix. Different working nuisance models for $\pi(\cdot)$ and $m(\cdot)$ are compared. We report the means and medians together with standard errors and MADs as subscripts. The reported values are separated into truly zero and non-zero coefficients.

(a) DGP: "Linear-linear" model for $\pi(\cdot)$ and $m(\cdot)$.

working nuisance model		zero			non-zero		
		CovP: mean	(median))	Length	CovP: mean	(median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.01})$	0.02_{0}	$0.93_{0.03}$	$(0.94_{0.02})$	0.02_{0}
	$\widehat{\pi}$: quad	$0.94_{0.03}$	$(0.95_{0.02})$	0.02_{0}	$0.93_{0.03}$	$(0.94_{0.02})$	0.02_{0}
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.01})$	0.02_{0}	$0.93_{0.03}$	$(0.94_{0.02})$	0.02_{0}
	$\widehat{\pi}$: quad	$0.94_{0.03}$	$(0.95_{0.01})$	0.02_{0}	$0.93_{0.03}$	$(0.94_{0.02})$	0.02_{0}
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.95_{0.01}$	$(0.96_{0.01})$	0.04_{0}	$0.94_{0.03}$	$(0.94_{0.04})$	0.05_{0}
	$\widehat{\pi}$: quad	$0.95_{0.01}$	$(0.96_{0.01})$	0.04_{0}	$0.94_{0.03}$	$(0.93_{0.04})$	0.05_{0}

(b) DGP: "Quad-Quad" model for $\pi(\cdot)$ and $m(\cdot)$.

working nuisance model		zero			non-zero		
		CovP: mean	(median)	Length	CovP: mean	(median)	Length
$\widehat{m}$: linear	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.03})$	0.06_{0}	$0.68_{0.39}$	$(0.84_{0.19})$	$0.07_{0.02}$
	$\widehat{\pi}$: quad	$0.96_{0.02}$	$(0.96_{0.01})$	$0.12_{0.01}$	$0.96_{0.02}$	$(0.96_{0.03})$	$0.14_{0.08}$
$\widehat{m}$: quad	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.95_{0.02})$	0.05_{0}	$0.93_{0.03}$	$(0.95_{0.01})$	$0.05_{0.01}$
	$\widehat{\pi}$: quad	$0.94_{0.03}$	$(0.95_{0.03})$	0.05_{0}	$0.94_{0.02}$	$(0.95_{0.01})$	$0.05_{0.01}$
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	$0.94_{0.03}$	$(0.94_{0.01})$	0.06_{0}	$0.8_{0.19}$	$(0.88_{0.13})$	$0.07_{0.01}$
	$\widehat{\pi}$: quad	$0.95_{0.02}$	$(0.95_{0.03})$	0.1_{0}	$0.95_{0.02}$	$(0.95_{0.02})$	$0.12_{0.06}$

Note that sub-Gaussians and sub-exponentials also possess other alternative definitions in terms of tail bounds, moment bounds or moment generating functions that are standard in the literature. All these definitions may be shown to be equivalent, upto constant factors in the parameters, to the one above. The ψ_α-Orlicz norms are more general norms allowing for any $\alpha > 0$ (not just 1 or 2), and hence, weaker tail behaviors. It is also worth noting that a bounded random variable X has $\|X\|_{\psi_\alpha} < \infty$ for *any* $\alpha \in (0, \infty]$.

C.2. Properties of Orlicz Norms and Concentration Bounds. We enlist here some useful general properties of Orlicz norms along with a few specific ones for sub-Gaussians and sub-exponentials. These are all quite well known and routinely used. Their statements (possibly with slightly different constants) and proofs can be found in several relevant references, including Van der Vaart and Wellner (1996); Pollard (2015); Vershynin (2012, 2018); Rigollet and Hütter (2017) and Wainwright (2019) among others. We therefore skip their proofs here for the sake of brevity.

LEMMA C.1 (General properties of Orlicz norms, sub-Gaussians and sub–exponentials). *Let X, Y denote generic random variables and let $\mu := \mathbb{E}(X)$.*

(i) (Basic properties). *For $\alpha \geq 1$, $\|\cdot\|_{\psi_\alpha}$ is a norm satisfying: (a) $\|X\|_{\psi_\alpha} \geq 0$ and $\|X\|_{\psi_\alpha} = 0 \Leftrightarrow X = 0$ a.s., (b) $\|cX\|_{\psi_\alpha} = |c|\|X\|_{\psi_\alpha}$ $\forall\, c \in \mathbb{R}$ and $\||X|\|_{\psi_\alpha} = \|X\|_{\psi_\alpha}$, and (c) $\|X + Y\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}$.*

(ii) (Monotonicities). *(a) For any $0 < \alpha \leq \beta$, $(\log 2)^{1/\alpha}\|X\|_{\psi_\alpha} \leq (\log 2)^{1/\beta}$ $\|X\|_{\psi_\beta}$. (b) For any $\alpha > 0$, $\||X|^\alpha\|_{\psi_1} \leq \|X\|_{\psi_\alpha}^\alpha$. (c) If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_\alpha} \leq \|Y\|_{\psi_\alpha}$ $\forall\, \alpha > 0$. (d) If X is bounded, i.e. $|X| \leq M$ a.s. for some constant M, then $\|X\|_{\psi_\alpha} \leq (\log 2)^{-1/\alpha} M$ for each $\alpha \in (0, \infty]$.*

(iii) (Tail bounds and equivalences). *(a) If $\|X\|_{\psi_\alpha} \leq \sigma$, then $\mathbb{P}(|X| > \epsilon) \leq 2\exp(-\epsilon^\alpha/\sigma^\alpha)$ $\forall\, \epsilon \geq 0$. (b) Conversely if $\mathbb{P}(|X| > \epsilon) \leq C\exp(-\epsilon^\alpha/\sigma^\alpha)$ $\forall\, \epsilon \geq 0$, for some $(C, \sigma, \alpha) > 0$, then $\|X\|_{\psi_\alpha} \leq \sigma(1 + C/2)^{1/\alpha}$.*

(iv) (Moment bounds). *If $\|X\|_{\psi_\alpha} \leq \sigma$ for some $(\alpha, \sigma) > 0$, then $\mathbb{E}(|X|^m) \leq C_\alpha^m \sigma^m m^{m/\alpha}$ $\forall\, m \geq 1$, for some constant C_α depending only on α. (A converse also holds although not presented here). In particular,*

  (a) If $\|X\|_{\psi_1} \leq \sigma$, then for each $m \geq 1$, $\mathbb{E}(|X|^m) \leq \sigma^m m! \leq \sigma^m m^m$.

  (b) If $\|X\|_{\psi_2} \leq \sigma$, then $\mathbb{E}(|X|^m) \leq 2\sigma^m \Gamma(m/2 + 1)$ $\forall\, m \geq 1$, where $\Gamma(a) := \int_0^\infty x^{a-1} exp(-x) dx$ $\forall\, a > 0$ denotes the Gamma function. Hence, $\mathbb{E}(|X|) \leq \sigma\sqrt{\pi}$ and $\mathbb{E}(|X|^m) \leq 2\sigma^m (m/2)^{m/2}$ $\forall\, m \geq 2$.

(v) (Hölder-type inequality for the Orlicz norm of products). *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* $\|XY\|_{\psi_\gamma} \leq \|X\|_{\psi_\alpha} \|Y\|_{\psi_\beta}$. *In particular, if* X *and* Y *are sub-Gaussian, then* XY *is sub-exponential and* $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$. *Further, if* Y *is bounded with* $Y \leq M$ *a.s. and* $\|X\|_{\psi_\alpha} < \infty$ *for any* $\alpha > 0$, *then* $\|XY\|_{\psi_\alpha} \leq M \|X\|_{\psi_\alpha}$.

(vi) (Orlicz norms and tail bounds for maximums). *Let* $\{X_i\}_{i=1}^n$ $(n \geq 1)$ *be random variables (possibly dependent) with* $\max_{1 \leq i \leq n} \|X_i\|_{\psi_\alpha} \leq \sigma$ *for some* (α, σ) *and let* $Z_n := \max_{1 \leq i \leq n} |X_i|$. *Then,* $\|Z_n\|_{\psi_\alpha} \leq \sigma(\log n + 2)^{1/\alpha} \leq \sigma\{3 \log(n+1)\}^{1/\alpha}$ *and* $\mathbb{P}\{Z_n > c\sigma(\log n)^{1/\alpha}\} \leq 2n^{-(c^\alpha - 1)}\ \forall\, c > 1$.

(vii) (MGF related properties of sub-Gaussians). *Let* $\mathbb{E}[\exp\{t(X-\mu)\}]$ *denote the moment generating function (MGF) of* $X - \mu$ *at* $t \in \mathbb{R}$. *Then:*

(a) *If* $\|X - \mu\|_{\psi_2} \leq \sigma$, *then* $\mathbb{E}[\exp\{t(X-\mu)\}] \leq \exp(2\sigma^2 t^2)\ \forall\, t \in \mathbb{R}$.

(b) *Conversely, if* $\mathbb{E}[\exp\{t(X-\mu)\}] \leq \exp(\sigma^2 t^2)\ \forall\, t \in \mathbb{R}$, *then* $\forall\, \epsilon \geq 0$, $\mathbb{P}(|X - \mu| > \epsilon) \leq 2\exp(-\epsilon^2/4\sigma^2)$ *and hence,* $\|X - \mu\|_{\psi_2} \leq 2\sqrt{2}\sigma$.

LEMMA C.2 (Concentration bounds for sums of independent sub-Gaussian variables). *Let* $\{X_i\}_{i=1}^n$ $(n \geq 1)$ *be independent (but not necessarily i.i.d.) random variables with means* $\{\mu_i\}_{i=1}^n$ *such that* $\|X_i - \mu_i\|_{\psi_2} \leq \sigma_i$ *for some* $\{\sigma_i\}_{i=1}^n \geq 0$. *Then, for any set of real numbers* $\{a_i\}_{i=1}^n$, *we have*

$$\mathbb{E}\left[\exp\left\{t\sum_{i=1}^n a_i(X_i - \mu_i)\right\}\right] \;\leq\; \exp\left(2t^2 \sum_{i=1}^n \sigma_i^2 a_i^2\right) \quad \forall\, t \in \mathbb{R}, \quad \textit{and}$$

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i(X_i - \mu_i)\right| \;>\; \epsilon\right\} \;\leq\; 2\exp\left(\frac{-\epsilon^2}{8\sum_{i=1}^n \sigma_i^2 a_i^2}\right) \quad \forall\, \epsilon \geq 0.$$

This further implies that $\|a_i(X_i - \mu_i)\|_{\psi_2} \leq 4(\sum_{i=1}^n \sigma_i^2 a_i^2)^{1/2}$. *In particular, when* $a_i = 1/n$ *and* $\sigma_i = \sigma$, $\|\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)\|_{\psi_2} \leq (4\sigma)/\sqrt{n}$.

LEMMA C.3 (Sub-Gaussian properties of binary random variables). *Let* $Z \in \{0, 1\}$ *be a binary random variable with* $\mathbb{E}(Z) \equiv \mathbb{P}(Z = 1) = p \in [0, 1]$ *and let* $\widetilde{Z} = (Z - p)$. *Then,* $\|\widetilde{Z}\|_{\psi_2} \leq 2\widetilde{p}$, *where* $\widetilde{p} = 0$ *if* $p \in \{0, 1\}$, $\widetilde{p} = 1/2$ *if* $p = 1/2$, *and* $\widetilde{p} = [(p - 1/2)/\log\{p/(1-p)\}]^{1/2}$ *if* $p \notin \{0, 1, 1/2\}$.

Lemma C.3 explicitly characterizes the sub-Gaussian properties of (centered) binary random variables and its proof can be found in Buldygin and Moskvichova (2013). The statement therein uses a MGF based definition of sub-Gaussians. The statement above is appropriately modified with the factor 2 multiplied in the $\|\cdot\|_{\psi_2}$ norm bound to adapt to our definition.

Next, we present a version of the well known Bernstein's inequality. While Lemma C.2 is useful, it applies only to sub-Gaussians. However, Bersntein's inequality applies more generally to sub-exponentials that include as special cases: sub-guassian variables, bounded variables, as well as products of two sub-Gaussian and/or bounded variables (see Lemma C.5).

LEMMA C.4 (Bernstein's inequality - adopted from Van de Geer and Lederer (2013)). *Let $\{Z_i\}_{i=1}^n$ be independent (but not necessarily i.i.d.) random variables and let $\mu_i := \mathbb{E}(Z_i)\ \forall\ 1 \leq i \leq n$. Suppose $\exists$ constants $\sigma, K \geq 0$ such that $n^{-1}\sum_{i=1}^n \mathbb{E}(|Z_i - \mu_i|^m) \leq (m!/2)\sigma^2 K^{m-2}$ for each $m \geq 2$. Then,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i - \mu_i)\right| \geq \sqrt{2}\sigma\epsilon + K\epsilon^2\right) \leq 2\exp\left(-n\epsilon^2\right) \quad \textit{for any } \epsilon \geq 0.$$

In particular, if $\{Z_i\}_{i=1}^n$ are i.i.d. realizations of a sub-exponential variable Z with $\mathbb{E}(Z) = \mu$ and $\|Z\|_{\psi_1} \leq \sigma_Z$ for some $\sigma_Z \geq 0$, then $\|Z - \mu\|_{\psi_1} \leq 2\sigma_Z$ and the bound above holds with $\sigma \equiv 2\sqrt{2}\sigma_Z$ and $K \equiv 2\sigma_Z$. Two important special cases of such a setting include: (a) $Z = XY$ with X and Y sub-Gaussian, in which case $\sigma_Z \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$, and (b) $Z = XY$ with X sub-exponential and $|Y| \leq M$ a.s. for some $M > 0$, in which case $\sigma_Z \leq M\|X\|_{\psi_1}$.

LEMMA C.5 (The Bernstein moment conditions and their verification). *Consider the moment conditions required in Bernstein's inequality in Lemma C.4. Define a random variable Z to satisfy the Bernstein moment conditions (BMC) with parameters $(\sigma, K) \geq 0$, denoted as $Z \sim BMC(\sigma, K)$, if for each $m \geq 2$, $\mathbb{E}(|Z - \mu|^m) \leq (m!/2)\sigma^2 K^{m-2}$ where $\mu := E(Z)$. Then,*

(a) If Z is sub-exponential with $\|Z\|_{\psi_1} \leq \sigma_Z$, then $Z \sim BMC(2\sqrt{2}\sigma_Z, 2\sigma_Z)$ and $|Z| \sim BMC(2\sqrt{2}\sigma_Z, 2\sigma_Z)$.

(b) If X and Y sub-Gaussian variables, then $Z := XY \sim BMC(2\sqrt{2}\sigma_Z, 2\sigma_Z)$ with $\sigma_Z = \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

(c) If X is sub-exponential and Y is a bounded random variable with $|Y| \leq M$ a.s., then $Z := XY \sim BMC(2\sqrt{2}\sigma_Z, 2\sigma_Z)$ with $\sigma_Z = M\|X\|_{\psi_1}$.

PROOF. If $\|Z\|_{\psi_1} \leq \sigma_Z$, then using Lemma C.1 (i)(c) and (iv)(a), $\|Z - \mu\|_{\psi_1} \leq 2\sigma_Z$ and $\mathbb{E}(|Z - \mu|^m) \leq (2\sigma_Z)^m m! \equiv (m!/2)(2\sqrt{2}\sigma_Z)^2(2\sigma_Z)^{m-2}$ for each $m \geq 1$. Hence, by definition, $Z \sim \mathrm{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$. ∎

Similarly, $\||Z|\|_{\psi_1} = \|Z\|_{\psi_1} \leq \sigma_Z$ and $\||Z| - \mathbb{E}\{|Z|\}\|_{\psi_1} \leq 2\sigma_Z$. Therefore, by identical arguments as above we again have: $|Z| \sim \mathrm{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$. ∎

Finally, using Lemma C.1, we have: for case (b), $\|Z\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2} \equiv \sigma_Z$, while for case (c), $\|Z\|_{\psi_1} \leq M\|X\|_{\psi_1} \equiv \sigma_Z$. The desired results then follow by using the same arguments used for proving the first result above. ∎

The following lemma is a useful concentration inequality that applies generally to any random variables with finite ψ_α-Orlicz norm, preserves the right rate and tail behaviors and involves only the variance in the leading term.

LEMMA C.6 (Concentration bounds with variance in the leading term - adopted from Theorem 3.4 of Kuchibhotla and Chakrabortty (2018)). *Suppose $\{\mathbf{X}_i\}_{i=1}^n$ are independent mean zero random vectors in $\mathbb{R}^p$, for any $p \geq 1$ and $n \geq 2$, such that for some $\alpha > 0$ and some $K_n > 0$,*

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \|\mathbf{X}_{i[j]}\|_{\psi_\alpha} \leq K_n, \quad \text{and define } \Gamma_n := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\mathbf{X}_{i[j]}^2\right).$$

Then for any $t \geq 0$, with probability at least $1 - 3e^{-t}$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \right\|_\infty \leq 7\sqrt{\frac{\Gamma_n(t + \log p)}{n}} + \frac{C_\alpha K_n (\log n)^{1/\alpha}(t + \log p)^{1/\alpha^*}}{n},$$

where $\alpha^* := \min\{\alpha, 1\}$ and $C_\alpha > 0$ is some constant depending only on α.

Finally, we end with a simple lemma that relates high probability bounds to sub-Gaussian type tail bounds with an extra probability correction term.

LEMMA C.7 (High probability bounds to sub-Gaussian type tail bounds). *Let X_n be any sequence of random variables satisfying $|X_n| \leq a_n$ with probability at least $1 - q_n$ for some $a_n \in [0, \infty)$ and $q_n \in [0, 1]$, $\forall\, n \geq 1$. Then,*

$$\mathbb{P}(|X_n| > t) \;\leq\; 2\exp\left\{-t^2/(2a_n^2)\right\} + q_n \;\; \text{for any } \; t \geq 0.$$

PROOF. Define the event $\mathcal{A}_n := \{X_n \leq a_n\}$ and let $\mathcal{A}_n^c$ denote its complement event. Then, $\mathbb{P}(\mathcal{A}_n^c) \leq q_n$ by assumption. Furthermore, note that $|X_n 1(\mathcal{A}_n)| \leq a_n$ a.s. $[\mathbb{P}]$, where $1(\cdot)$ denotes the indicator function. Hence, using Lemma C.1 (ii) (d), we have: $\|X_n 1(\mathcal{A}_n)\|_{\psi_2} \leq (\log 2)^{-1/2} a_n \leq \sqrt{2} a_n$.

Hence, using Lemma C.1 (iii) (a), $\mathbb{P}\{|X_n 1(\mathcal{A}_n)| > t\} \leq 2\exp\{-t^2/(2a_n^2)\}$ for any $t \geq 0$. Consequently, we have: for any $t \geq 0$,

$$\begin{aligned} \mathbb{P}(|X_n| > t) \;&=\; \mathbb{P}(|X_n| > t, \mathcal{A}_n) + \mathbb{P}(|X_n| > t, \mathcal{A}_n^c) \\ &\leq\; \mathbb{P}(|X_n 1(\mathcal{A}_n)| > t) + \mathbb{P}(\mathcal{A}_n^c) \;\leq\; 2\exp\{-t^2/(2a_n^2)\} + q_n. \end{aligned}$$

This establishes the desired tail bound and completes the proof. ∎

APPENDIX D: PROOF OF LEMMA 2.1

The proof relies substantially on a useful result of Negahban et al. (2012). We therefore adopt some of their basic notations and terminology at the beginning of the proof in order to facilitate the use of that result.

For any $\mathbf{u} \in \mathbb{R}^p$, let $\mathcal{R}(\mathbf{u}) = \|\mathbf{u}\|_1$ and let $\mathcal{R}^*(\mathbf{u}) \equiv \sup_{\mathbf{v} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} \{\mathbf{u}'\mathbf{v}/\mathcal{R}(\mathbf{v})\}$ be the 'dual norm' for $\mathcal{R}(\cdot)$. Further, for any subspace $\mathcal{M} \subseteq \mathbb{R}^p$, let $\Psi(\mathcal{M}) \equiv \sup_{\mathbf{u} \in \mathcal{M} \setminus \{\mathbf{0}\}} \{\mathcal{R}(\mathbf{u})/\|\mathbf{u}\|_2\}$ denote its 'subspace compatibility constant' with respect to $\mathcal{R}(\cdot)$. Then, with $\mathcal{J}, \mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{J}}^{\perp}$ as defined in Section 2, it is not difficult to show that: (i) $\mathcal{R}(\cdot)$ is *decomposable* with respect to the orthogonal subspace pair $(\mathcal{M}_{\mathcal{J}}, \mathcal{M}_{\mathcal{J}}^{\perp})$ for any $\mathcal{J} \subseteq \{1, \ldots, p\}$, in the sense that $\mathcal{R}(\mathbf{u}+\mathbf{v}) = \mathcal{R}(\mathbf{u}) + \mathcal{R}(\mathbf{v})\ \forall\ \mathbf{u} \in \mathcal{M}_{\mathcal{J}}, \mathbf{v} \in \mathcal{M}_{\mathcal{J}}^{\perp}$; (ii) $\mathcal{R}^*(\mathbf{u}) = \|\mathbf{u}\|_\infty\ \forall\ \mathbf{u} \in \mathbb{R}^p$; and (iii) with $\mathcal{J} = \mathcal{A}(\mathbf{v})$ for any $\mathbf{v} \in \mathbb{R}^p$, $\Psi^2(\mathcal{M}_{\mathcal{J}}) = s_{\mathbf{v}}$. (We refer to Negahban et al. (2012) for further discussions and/or proofs of these facts). Lastly, let $P_{\mathcal{J}}(\mathbf{v})$ and $P_{\mathcal{J}}^{\perp}(\mathbf{v})$ respectively denote the orthogonal projections of any $\mathbf{v} \in \mathbb{R}^p$ onto $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{J}}^{\perp}$, for any $\mathcal{J}$ as above.

To establish the result, we consider the alternative representation (2.11) of $\boldsymbol{\theta}_0$ based on regularized minimization of the pseudo loss $\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$ defined in (2.10). Clearly, since $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ as assumed, so is $\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$. Further, owing to (2.3)-(2.6), we have: for any $\boldsymbol{\theta}, \mathbf{v} \in \mathbb{R}^d$,

$$\text{(D.1)} \qquad \boldsymbol{\nabla}\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta}) \;=\; \boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}) \;\text{ and }\; \delta\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta}, \mathbf{v}) \;=\; \delta\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}, \mathbf{v}),$$

where $\delta\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta}, \mathbf{v}) := \widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta} + \mathbf{v}) - \widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta}) - \mathbf{v}'\boldsymbol{\nabla}\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$. Thus, under Assumption 2.1, $\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$ also satisfies the RSC property (2.12) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Hence, using the decomposability of $\mathcal{R}(\cdot)$ over $(\mathcal{M}_{\mathcal{J}}, \mathcal{M}_{\mathcal{J}}^{\perp})$ with $\mathcal{J}$ chosen to be $\mathcal{A}(\boldsymbol{\theta}_0)$, and the RSC property of $\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ under Assumption 2.1 and (D.1), we have: by Theorem 1 of Negahban et al. (2012), for any realization of $\mathcal{D}_n$ and any choice of $\lambda \equiv \lambda_n \geq 2\|\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0\|_\infty$,

$$\text{(D.2)} \qquad \left\|\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0\right\|_2 \;\equiv\; \left\|\widehat{\boldsymbol{\theta}}_{\text{DDR}}(\lambda_n; \mathcal{D}_n) - \boldsymbol{\theta}_0\right\|_2 \;\leq\; 3\sqrt{s}\frac{\lambda}{\kappa_{\text{DDR}}}$$

where, while applying the result from Negahban et al. (2012), we chose the parameter $\boldsymbol{\theta}^*$, in their notation, as $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, $\{\mathcal{R}(\cdot), \mathcal{R}^*(\cdot)\}$ as $\{\|\cdot\|_1, \|\cdot\|_\infty\}$, and used: $\Psi^2(\mathcal{M}_{\mathcal{J}}) = \|\boldsymbol{\theta}_0\|_0 \equiv s$, $\mathcal{R}^*[\boldsymbol{\nabla}\{\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})\}] = \mathcal{R}^*[\boldsymbol{\nabla}\{\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta})\}] \equiv \|\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ and $P_{\mathcal{A}(\boldsymbol{\theta}_0)}^{\perp}(\boldsymbol{\theta}_0) = \Pi_{\mathcal{A}^c(\boldsymbol{\theta}_0)}(\boldsymbol{\theta}_0) \equiv \Pi_{\boldsymbol{\theta}_0}^c(\boldsymbol{\theta}_0) = \mathbf{0}$. ∎

Further, using Lemma 1 of Negahban et al. (2012), we also have that for λ chosen as above, the error $\widehat{\boldsymbol{\Delta}} := (\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0)$ belongs to the set $\mathbb{C}(\boldsymbol{\theta}_0)$ as defined in (2.12). Consequently, $\|\Pi^c_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1 \leq 3\|\Pi_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1$. Hence we have:

$$\begin{aligned} \left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0\right\|_1 &\equiv \|\widehat{\boldsymbol{\Delta}}\|_1 \;=\; \|\Pi_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1 + \|\Pi^c_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1 \;\leq\; 4\|\Pi_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1 \\ &\leq\; 4\sqrt{s}\|\Pi_{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\Delta}})\|_1 \;\leq\; 4\sqrt{s}\left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0\right\|_2 \;\leq\; 12s\frac{\lambda}{\kappa_{\mathrm{DDR}}}, \end{aligned}$$

where the final step follows from using (D.2). This, along with (D.2), establishes the desired L_2 and L_1 error bounds for $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$. The rest of the informal claims in the second part of Lemma 2.1 are straightforward consequences of combining the deterministic error bounds proved above with the results of Theorems 3.1-3.4. This completes the proof of Lemma 2.1. ■

APPENDIX E: PROOF OF THEOREM 3.1

Recalling from (3.1) and (3.2), we note that $\mathbf{T}_{0,n}$ is simply a sum of two centered i.i.d. averages given by:

$$\text{(E.1)}\quad \mathbf{T}_{0,n} \;=\; \mathbf{T}^{(1)}_{0,n} + \mathbf{T}^{(2)}_{0,n} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}^{(1)}_0(\mathbf{Z}_i) \;+\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}^{(2)}_0(\mathbf{Z}_i), \quad \text{where}$$

$$\mathbf{T}^{(1)}_0(\mathbf{Z}) := \{m(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}_0)\}\mathbf{h}(\mathbf{X}) \;\text{ and }\; \mathbf{T}^{(2)}_0(\mathbf{Z}) := \frac{T}{\pi(\mathbf{X})}\{Y - m(\mathbf{X})\}\mathbf{h}(\mathbf{X}),$$

with $\mathbb{E}\{\mathbf{T}^{(1)}_0(\mathbf{Z})\} = \mathbf{0}$ and $\mathbb{E}\{\mathbf{T}^{(2)}_0(\mathbf{Z})\} = \mathbf{0}$ since $\mathbb{E}\{\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0)\} = \mathbf{0}$ and $\mathbb{E}\{\epsilon(\mathbb{Z})|\mathbf{X}\} = 0$, by definition, and $\epsilon(\mathbb{Z}) \perp\!\!\!\perp T \,|\, \mathbf{X}$ due to Assumption 1.1 (a).

Now, using Assumption 3.1 (a) and Lemma C.5 (a), we have:

$$\text{(E.2)}\qquad \mathbf{T}^{(1)}_{0[j]}(\mathbf{Z}) \;\equiv\; \psi(\mathbf{X})\mathbf{h}_{[j]}(\mathbf{X}) \;\sim\; \mathrm{BMC}(\bar{\sigma}_1, \bar{K}_1) \quad \forall\, j \in \{1, \ldots, d\},$$

for some constants $\bar{\sigma}_1 := 2\sqrt{2}\sigma_\psi\sigma_{\mathbf{h}} \geq 0$ and $\bar{K}_1 := 2\sigma_\psi\sigma_{\mathbf{h}} \geq 0$.

Next, using Assumption 3.1 (a) and Lemma C.1 (v), $\|\varepsilon(\mathbb{Z})\mathbf{h}_{[j]}(\mathbf{X})\|_{\psi_1} \leq \sigma_\varepsilon\sigma_{\mathbf{h}}$ for each $j \in \{1, \ldots, d\}$. Further, owing to Assumption 1.1 (b) and (1.1), $T/\pi(\mathbf{X}) \leq \delta_\pi^{-1}$ a.s. $[\mathbb{P}]$. Hence, using Lemma C.5 (b), we have

$$\text{(E.3)}\quad \mathbf{T}^{(2)}_{0[j]}(\mathbf{Z}) \;\equiv\; \frac{T}{\pi(\mathbf{X})}\varepsilon(\mathbb{Z})\mathbf{h}_{[j]}(\mathbf{X}) \;\sim\; \mathrm{BMC}(\bar{\sigma}_2, \bar{K}_2) \quad \forall\, j \in \{1, \ldots, d\},$$

for some constants $\bar{\sigma}_2 := 2\sqrt{2}\sigma_\varepsilon\sigma_{\mathbf{h}}\delta_\pi^{-1} \geq 0$ and $\bar{K}_2 := 2\sigma_\varepsilon\sigma_{\mathbf{h}}\delta_\pi^{-1} \geq 0$

Hence, (E.2) and (E.3) ensure that for each $j \in \{1, \ldots, d\}$, $\mathbf{T}^{(1)}_{0[j]}(\mathbf{Z})$ and $\mathbf{T}^{(2)}_{0[j]}(\mathbf{Z})$ satisfy the required moment conditions for Bernstein's inequality

(Lemma C.4) to apply. Using Lemma C.4, we then have: for any $\epsilon_1 \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \left\| \mathbf{T}_{0,n}^{(1)} \right\|_\infty \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{T}_0^{(1)}(\mathbf{Z}_i) \right\|_\infty > \sqrt{2}\bar{\sigma}_1\epsilon_1 + \bar{K}_1\epsilon_1^2 \right\} \\
&\quad \leq \sum_{j=1}^{d} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{T}_{0[j]}^{(1)}(\mathbf{Z}_i) \right| > \sqrt{2}\bar{\sigma}_1\epsilon_1 + \bar{K}_1\epsilon_1^2 \right\} \\
&\quad \leq \sum_{j=1}^{d} 2\exp\left(-n\epsilon_1^2\right) = 2d\exp\left(-n\epsilon_1^2\right) \equiv 2\exp\left(-n\epsilon_1^2 + \log d\right),
\end{aligned} \tag{E.4}$$

where the second step uses the union bound (u.b.). Similarly, for any $\epsilon_2 \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \left\| \mathbf{T}_{0,n}^{(2)} \right\|_\infty \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{T}_0^{(2)}(\mathbf{Z}_i) \right\|_\infty > \sqrt{2}\bar{\sigma}_2\epsilon_2 + \bar{K}_2\epsilon_2^2 \right\} \\
&\quad \leq \sum_{j=1}^{d} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{T}_{0[j]}^{(2)}(\mathbf{Z}_i) \right| > \sqrt{2}\bar{\sigma}_2\epsilon_2 + \bar{K}_2\epsilon_2^2 \right\} \\
&\quad \leq \sum_{j=1}^{d} 2\exp\left(-n\epsilon_2^2\right) = 2d\exp\left(-n\epsilon_2^2\right) \equiv 2\exp\left(-n\epsilon_2^2 + \log d\right).
\end{aligned} \tag{E.5}$$

Hence, setting $\epsilon_1 = \epsilon_2 \equiv \epsilon$ for any $\epsilon \geq 0$, letting $\sigma_0 := \bar{\sigma}_1 + \bar{\sigma}_2$ and $K_0 := \bar{K}_1 + \bar{K}_2$, and using (E.4)-(E.5) in the original decomposition (E.1) of $\mathbf{T}_{0,n}$, we have a tail bound for $\|\mathbf{T}_{0,n}\|_\infty$, as follows. For any $\epsilon \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|\mathbf{T}_{0,n}\|_\infty \equiv \left\| \mathbf{T}_{0,n}^{(1)} + \mathbf{T}_{0,n}^{(2)} \right\|_\infty > \sqrt{2}\sigma_0\epsilon + K_0\epsilon^2 \right) \\
&\quad \leq \mathbb{P}\left(\left\| \mathbf{T}_{0,n}^{(1)} \right\|_\infty > \sqrt{2}\bar{\sigma}_1\epsilon + \bar{K}_1\epsilon^2 \right) + \mathbb{P}\left(\left\| \mathbf{T}_{0,n}^{(2)} \right\|_\infty > \sqrt{2}\bar{\sigma}_2\epsilon + \bar{K}_2\epsilon^2 \right) \\
&\quad \leq 4\exp\left(-n\epsilon^2 + \log d\right).
\end{aligned} \tag{E.6}$$

(E.6) therefore establishes a general tail bound for $\|\mathbf{T}_{0,n}\|_\infty$ and also establishes its rate of convergence. This completes the proof of Theorem 3.1. ∎

APPENDIX F: TECHNICAL DISCUSSIONS ON THE ERROR TERMS

We note here a few useful details regarding the structure and techniques for controlling the error terms $\mathbf{T}_{\pi,n}$, $\mathbf{T}_{m,n}$ and $\mathbf{R}_{\pi,m,n}$ accounting for the nuisance function estimators $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ in the decomposition (3.1) of $\mathbf{T}_n$.

(a) The structure of $\mathbf{T}_{\pi,n}$ and reasons for obtaining $\widehat{\pi}(\cdot)$ solely from $\mathcal{X}_n$. $\mathbf{T}_{\pi,n}$ is simply the sample average of the random variables $\{\mathbf{T}_\pi(\mathbf{Z}_i)\}_{i=1}^n$ in (3.3). However, this average is *not* an i.i.d. average due to the presence of $\widehat{\pi}(\cdot)$ which depends on all observations in $\mathcal{D}_n$. A key property that is quite useful in this regard is that, by assumption, $\widehat{\pi}(\cdot)$ is obtained solely from the subset $\mathcal{X}_n := \{(T_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ of $\mathcal{D}_n$. Hence, $\{\mathbf{T}_\pi(Z_i)\}_{i=1}^n | \mathcal{X}_n$ are *conditionally* independent and centered with $\mathbb{E}\{\mathbf{T}_\pi(\mathbf{Z}_i)\} = \mathbb{E}[\mathbb{E}\{\mathbf{T}_\pi(\mathbf{Z}_i) \mid \widehat{\pi}(\cdot), \mathbf{X}_i\}]$ $= \mathbb{E}[\mathbb{E}\{\mathbf{T}_\pi(\mathbf{Z}_i) \mid \mathcal{X}_n\}] = \mathbf{0}$. The conditioning on $\mathcal{X}_n$ ensures that $\widehat{\pi}(\cdot)$, as well as all other components in $\mathbf{T}_\pi(\mathbf{Z}_i)$ which are functions of $(T_i, \mathbf{X}_i)$ only, can now be treated as fixed and further, the conditional expectation being $\mathbf{0}$ follows from the fact that $\mathbb{E}[\{Y_i - m(\mathbf{X}_i)\}|\mathcal{X}_n] \equiv \mathbb{E}\{\varepsilon(\mathbb{Z}_i)|\mathcal{X}_n\} = \mathbb{E}\{\varepsilon(\mathbb{Z}_i)|\mathbb{T}_i, \mathbf{X}_i\}$ $= \mathbb{E}\{\varepsilon(\mathbb{Z}_i) \mid \mathbf{X}_i\} = 0$, where the final step is due to Assumption 1.1 (a).

Thus, $\mathbf{T}_{\pi,n}$ is a centred average of (conditionally) independent variables. We exploit this and the structure of $\mathbf{T}_\pi(\mathbf{Z})$ in Theorem 3.2 to control $\mathbf{T}_{\pi,n}$.

(b) The structure of $\mathbf{T}_{m,n}$ and the benefits of sample splitting/cross-fitting. $\mathbf{T}_{m,n}$ is simply the sample average of the random variables $\{\mathbf{T}_m(\mathbf{Z})\}_{i=1}^n$ in (3.4). However, in the absence of sample splitting, this is *not* an i.i.d. average due to the presence of $\widehat{m}(\cdot)$ which depends on all observations in $\mathcal{D}_n$. Further, unlike $\mathbf{T}_{\pi,n}$ where $\{\mathbf{T}_\pi(\mathbf{Z}_i)\}_{i=1}^n | \mathcal{X}_n$ were at least (conditionally) independent and centered, $\mathbf{T}_{m,n}$ possesses no such desirable features even if $\widehat{m}(\cdot)$ is obtained solely from the subset $\mathcal{D}_n^{(c)} := \{(Y_i, \mathbf{X}_i) : T_i = 1, 1 \leq i \leq n\}$ of 'complete cases' in $\mathcal{D}_n$, as $\mathcal{D}_n^{(c)}$ still (implicitly) depends on $\{T_i\}_{i=1}^n$ due to the restriction to the set with $T_i = 1$, and not just on $\{Y_i, \mathbf{X}_i\}_{i=1}^n$.

Thus, in the absence of sample splitting, $\mathbf{T}_{m,n}$ has no additional 'structure' readily available that may lead to averages of variables which can be treated as conditionally independent and centered. In general, to control $\mathbf{T}_{m,n}$ without sample splitting, one needs tools from empirical process theory. The corresponding analyses can be substantially involved and the conditions necessary can be quite strong, especially in high dimensional settings. However, these technical issues can be avoided through the sample splitting based estimates $\{\widetilde{m}(\mathbf{X}_i)\}_{i=1}^n$ which 'induces' a natural independence.

For any $\mathbf{Z} \perp\!\!\!\perp \widehat{m}(\cdot)$, or more specifically, $\mathbf{Z} \perp\!\!\!\perp \{\text{data used to obtain } \widehat{m}(\cdot)\}$, $\mathbb{E}\{\mathbf{T}_m(\mathbf{Z}) \,|\, \widehat{m}(\cdot), \mathbf{X}\} = \mathbb{E}\{\mathbf{T}_m(\mathbf{Z}) | \mathbf{X}\} = \mathbf{0}$ due to Assumption 1.1 (a). Hence, $\mathbb{E}\{\mathbf{T}_m(\mathbf{Z}) \mid \widehat{m}(\cdot)\} = \mathbf{0}$ and for any i.i.d. collection $\{\mathbf{Z}_k\}_{k=1}^K$ of $\mathbf{Z} \perp\!\!\!\perp \widehat{m}(\cdot)$, $\{\mathbf{T}_m(\mathbf{Z}_k)\}_{k=1}^K \mid \widehat{m}(\cdot)$ are (conditionally) independent and centered random variables. These serve as the main motivations behind the sample splitting.

In contrast to the 'in-sample' estimates $\{\widehat{m}(\mathbf{X}_i\}_{i=1}^n$, wherein $\widehat{m}(\cdot)$ is obtained from $\mathcal{D}_n$ and also evaluated at the same training points $\{\mathbf{X}_i\}_{i=1}^n \in \mathcal{D}_n$, thereby making them intractably dependent on $\widehat{m}(\cdot)$, the cross-fitted estimates $\{\widetilde{m}(\mathbf{X}_i)\}_{i=1}^n$ ensure that for each $k \neq k' \in \{1,2\}$, the evaluation points $\{\mathbf{X}_i \in \mathcal{D}_n^{(k)}\}$ used are independent of the estimator $\widehat{m}^{(k')}(\cdot)$ obtained from $\mathcal{D}_n^{(k')} \perp\!\!\!\perp \mathcal{D}_n^{(k)}$, thus inducing a desirable 'independence structure'. This has substantial technical as well as practical benefits in reducing over-fitting. We exploit the technical benefits greatly in Theorem 3.3 to control $\mathbf{T}_{m,n}$.

(b) The structure of $\mathbf{R}_{\pi,m,n}$. Finally, note that $\mathbf{R}_{\pi,m,n}$ is essentially a second order (product-type) bias term involving the product of two error terms arising from the estimation of $\{\pi(\cdot), m(\cdot)\}$. Under reasonable assumptions on the convergence rates of the estimators $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$, one can try to control the behavior of this term by 'naive' techniques, as opposed to the more sophisticated analyses required for controlling $\mathbf{T}_{\pi,n}$ and $\mathbf{T}_{m,n}$. Such techniques and associated conditions are well known and standard in the literature for the special case of the mean estimation problem (or ATE estimation problem in CI), where a commonly adopted assumption is to have the product of the two convergence rates to be faster than $n^{-0.5}$ (Farrell, 2015; Chernozhukov et al., 2018a). In general, such product conditions are typically reasonable and allows for much weaker (slower) convergence rates for one estimator as long as the other one has sufficiently fast enough rates. A stronger but familiar sufficient condition however is to have the convergence rates of both estimators to be faster than $n^{-0.25}$. In Theorem 3.4, we control $\mathbf{R}_{\pi,m,n}$ by adopting a similar condition with an additional logarithmic factor involved to account for the inherent high dimensionality of our error terms.

APPENDIX G: PROOF OF THEOREM 3.2

To establish Theorem 3.2, we first state and prove a more general result that gives an explicit tail bound for $\|\mathbf{T}_{\pi,n}\|_\infty$.

THEOREM G.1 (Tail bound for $\|\mathbf{T}_{\pi,n}\|_\infty$). *Let Assumptions 1.1, 3.1 and 3.2 hold with the sequences* $(v_{n,\pi}, q_{n,\pi})$ *and the constants* $(\delta_\pi, \sigma_\varepsilon, \sigma_{\mathbf{h}}, C)$ *as defined therein, Then, for any* $\epsilon, \epsilon_1, \epsilon_2, \epsilon_3 \geq 0$, *with* $\epsilon_2 < \delta_\pi$ *small enough,*

$$\mathbb{P}\left(\|\mathbf{T}_{\pi,n}\|_\infty > \epsilon\right) \;\leq\; 2\exp\left\{\frac{-n\epsilon^2}{d_n(\epsilon_1,\epsilon_2,\epsilon_3)} + \log d\right\} + 4\exp\left(-n\epsilon_3^2 + \log d\right)$$
$$+ 2C\exp\left\{\frac{-\epsilon_1^2}{v_{n,\pi}^2} + \log(nd)\right\} + 2C\exp\left\{\frac{-\epsilon_2^2}{v_{n,\pi}^2} + \log(nd)\right\} + 4q_{n,\pi}(nd),$$

where, for any $(\epsilon_1,\epsilon_2,\epsilon_3) \geq 0$ *as above,* $d_n(\epsilon_1,\epsilon_2,\epsilon_3) \geq 0$ *is given by:*

$$d_n(\epsilon_1,\epsilon_2,\epsilon_3) \;:=\; \frac{8\sigma_\varepsilon^2\epsilon_1^2}{(\delta_\pi-\epsilon_2)^2}\left(\frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + \sqrt{2}\sigma_\pi\epsilon_3 + K_\pi\epsilon_3^2\right), \quad \textit{with}$$

$\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty := \max_{1\le j\le d}\mathbb{E}\{\mathbf{h}_{[j]}^2(\mathbf{X})\}$, $\sigma_\pi := 2\sqrt{2}\sigma_{\mathbf{h}}^2\delta_\pi^{-2}$ *and* $K_\pi := 2\sigma_{\mathbf{h}}^2\delta_\pi^{-2}$.

G.1. Proof of Theorem G.1. Let $\mathcal{X}_n := \{(T_i, \mathbf{X}_i) : i = 1, \ldots, n\}$. Let $\mathbb{E}_{\mathcal{X}_n}(\cdot)$ and $\mathbb{P}_{\mathcal{X}_n}(\cdot)$ respectively denote expectation and probability w.r.t. $\mathcal{X}_n$ and $\mathbb{P}(\cdot \mid \mathcal{X}_n)$ denote conditional probability given $\mathcal{X}_n$. Next, let us define:

$$\Delta_{\pi,n}(\mathbf{X}) := \widehat{\pi}(\mathbf{X}) - \pi(\mathbf{X}), \quad \|\Delta_{\pi,n}\|_{\infty,n} := \max_{1\le i\le n}|\Delta_{\pi,n}(\mathbf{X}_i)|, \tag{G.1}$$

$$\widetilde{\pi}_n(\mathbf{X}) := -\frac{1}{\widehat{\pi}(\mathbf{X})} \quad \text{and} \quad \|\widetilde{\pi}_n\|_{\infty,n} := \max_{1\le i\le n}|\widetilde{\pi}_n(\mathbf{X}_i)|. \tag{G.2}$$

Further, for each $j \in \{1, \ldots, d\}$, let us define:

$$\boldsymbol{\varphi}_{[j]}(T, \mathbf{X}) := \frac{T}{\pi(\mathbf{X})}\mathbf{h}_{[j]}(\mathbf{X}), \quad \bar{\boldsymbol{\varphi}}_{n[j]}^{(2)} \equiv \bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}(\mathcal{X}_n) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varphi}_{[j]}^2(T_i, \mathbf{X}_i), \tag{G.3}$$

$$\boldsymbol{\mu}_{\boldsymbol{\varphi}[j]}^{(2)} := \mathbb{E}\left\{\boldsymbol{\varphi}_{[j]}^2(T, \mathbf{X})\right\} \equiv \mathbb{E}\left\{\bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}(\mathcal{X}_n)\right\} \text{ and } \boldsymbol{\mu}_{\mathbf{h}[j]}^{(2)} := \mathbb{E}\left\{\mathbf{h}_{[j]}^2(\mathbf{X})\right\}. \tag{G.4}$$

Using (G.1)-(G.3) in (3.3) and recalling that $\varepsilon(\mathbb{Z}) = Y - m(\mathbf{X})$, we have:

$$\mathbf{T}_\pi(\mathbf{Z}) = \Delta_{\pi,n}(\mathbf{X})\widetilde{\pi}_n(\mathbf{X})\boldsymbol{\varphi}(T, \mathbf{X})\varepsilon(\mathbb{Z}), \quad \text{where} \tag{G.5}$$

$\boldsymbol{\varphi}(T, \mathbf{X}) \in \mathbb{R}^d$ denotes the vector with j^{th} entry $= \boldsymbol{\varphi}_{[j]}(T, \mathbf{X})\ \ \forall\, 1 \le j \le d$.

Under Assumptions 1.1 (a) and 3.1 (b), $\mathbb{E}\{\varepsilon(\mathbb{Z}) \mid \mathbf{X}\} \equiv \mathbb{E}\{\varepsilon(\mathbb{Z}) \mid T, \mathbf{X}\} = 0$ and $\|\varepsilon(\mathbb{Z})|\mathbf{X}\|_{\psi_2} \equiv \|\varepsilon(\mathbb{Z})|(T, \mathbf{X})\|_{\psi_2} \le \sigma_\varepsilon(\mathbf{X}) \le \sigma_\varepsilon < \infty$. Hence, $\varepsilon(\mathbb{Z}_i)|\mathcal{X}_n$ are (conditionally) independent random variables satisfying: $\mathbb{E}\{\varepsilon(\mathbb{Z}_i) \mid \mathcal{X}_n\} = 0$ and $\|\varepsilon(\mathbf{Z}_i) \mid \mathcal{X}_n\|_{\psi_2} \le \sigma_\varepsilon\ \forall\, 1 \le i \le n$. Further, conditional on $\mathcal{X}_n$, $\phi(T_i, \mathbf{X}_i)$, $\Delta_{\pi,n}(\mathbf{X}_i)$ and $\mathbf{h}_{[j]}(\mathbf{X}_i)$ are all constants $\forall\, i, j$. Using these facts along with (G.1)-(G.3), we have: $\forall\, 1 \le i \le n$ and $1 \le j \le d$,

$$\begin{aligned}\left\|\mathbf{T}_{\pi[j]}(\mathbf{Z}_i)\ \mid \mathcal{X}_n\right\|_{\psi_2} &\equiv \left\|\Delta_{\pi,n}(\mathbf{X}_i)\widetilde{\pi}_n(\mathbf{X}_i)\boldsymbol{\varphi}_{[j]}(T,\mathbf{X}_i)\varepsilon(\mathbb{Z}_i) \mid \mathcal{X}_n\right\|_{\psi_2}\\ &\le \Delta_{\pi,n}(\mathbf{X}_i)\widetilde{\pi}_n(\mathbf{X}_i)\boldsymbol{\varphi}_{[j]}(T_i, \mathbf{X}_i)\sigma_\varepsilon(\mathbf{X}_i) \ \le \sigma_\varepsilon\|\Delta_{\pi,n}\|_{\infty,n}\|\widetilde{\pi}_n\|_{\infty,n}\boldsymbol{\varphi}_{[j]}(T_i, \mathbf{X}_i).\end{aligned}$$

Further, $\forall\, 1 \le j \le d$, $\{\mathbf{T}_{\pi[j]}(\mathbf{Z}_i)\}_{i=1}^n|\mathcal{X}_n$ are (conditionally) independent and centered random variables. Hence, using Lemma C.2, we have: $\forall\, 1 \le j \le d$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}_{\pi[j]}(\mathbf{Z}_i)\ \middle|\ \mathcal{X}_n\right\|_{\psi_2} \le \frac{4c_{n,j}(\mathcal{X}_n)}{\sqrt{n}}, \quad \text{where}$$

$$c_{n,j}(\mathcal{X}_n) := \sigma_\varepsilon\|\Delta_{\pi,n}\|_{\infty,n}\|\widetilde{\pi}_n\|_{\infty,n}\left(\bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}\right)^{1/2} \tag{G.6}$$

and all notations are as defined in (G.1)-(G.3). Using Lemma C.2 again, it now follows that for any $\epsilon \ge 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{T}_{\pi[j]}(\mathbf{Z}_i)\right| > \epsilon\ \middle|\ \mathcal{X}_n\right\} \le 2\exp\left\{\frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)}\right\}\ \ \forall\, 1 \le j \le d. \tag{G.7}$$

The fundamental bound for $\|\mathbf{T}_{\pi,n}\|_\infty$. Using (G.7), the union bound (u.b.) and the law of iterated expectations (l.i.e.), we then have: for any $\epsilon \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \|\mathbf{T}_{\pi,n}\|_\infty \;\equiv\; \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{T}_\pi(\mathbf{Z}_i) \right\|_\infty \;>\; \epsilon \right\} \\
&\quad\leq\; \sum_{j=1}^d \mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^n \mathbf{T}_{\pi[j]}(\mathbf{Z}_i) \right| > \epsilon \right\} \quad \text{[using the u.b.]}, \\
&\quad=\; \sum_{j=1}^d \mathbb{E}_{\mathcal{X}_n}\left[\mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^n \mathbf{T}_{\pi[j]}(\mathbf{Z}_i) \right| > \epsilon \;\middle|\; \mathcal{X}_n \right\} \right] \quad \text{[using the l.i.e.]}, \\
&\quad\leq\; \sum_{j=1}^d 2\,\mathbb{E}_{\mathcal{X}_n}\left[\exp\left\{ \frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)} \right\} \right] \quad \text{[using (G.7)]}. \quad \blacksquare
\end{aligned} \tag{G.8}$$

Next, we aim to control the behavior of the random variable $c_{n,j}^2(\mathcal{X}_n)$ appearing in the bound (G.8). Based on the definition of $c_{n,j}(\mathcal{X}_n)$ in (G.6), it suffices to separately control the variables $\|\Delta_{\pi,n}\|_{\infty,n}^2$, $\|\widetilde{\pi}_n\|_{\infty,n}^2$ and $\bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}$.

Controlling $\|\Delta_{\pi,n}\|_{\infty,n}^2$. Using (3.6) in Assumption 3.2 along with the u.b., and recalling all notations defined in (G.1)-(G.2), we have: for any $\epsilon_1 \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \|\Delta_{\pi,n}\|_{\infty,n}^2 \equiv \max_{1\leq i\leq n} |\Delta_{\pi,n}(\mathbf{X}_i)|^2 \;>\; \epsilon_1^2 \right\} \\
&\leq\; \sum_{i=1}^n \mathbb{P}\left\{ |\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)| > \epsilon_1 \right\} \;\leq\; Cn\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2} \right) + nq_{n,\pi}. \quad \blacksquare
\end{aligned} \tag{G.9}$$

Controlling $\|\widetilde{\pi}_n\|_{\infty,n}^2$. Using similar arguments, along with (1.1), we have: $\forall\, \epsilon_2 \geq 0$ small enough such that $\epsilon_2 < \delta_\pi$ with δ_π as in (1.1),

$$\begin{aligned}
&\mathbb{P}\left[\|\widetilde{\pi}_n\|_{\infty,n}^2 \equiv \max_{1\leq i\leq n} |\widetilde{\pi}_n(\mathbf{X}_i)|^2 \;>\; (\delta_\pi - \epsilon_2)^{-2} \right] \\
&\quad\leq\; \sum_{i=1}^n \mathbb{P}\left\{ \widehat{\pi}^{-1}(\mathbf{X}_i) > (\delta_\pi - \epsilon_2)^{-1} \right\} \;\leq\; \sum_{i=1}^n \mathbb{P}\left\{ \widehat{\pi}(\mathbf{X}_i) \;<\; \pi(\mathbf{X}_i) - \epsilon_2 \right\} \\
&\leq\; \sum_{i=1}^n \mathbb{P}\left\{ |\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)| \;>\; \epsilon_2 \right\} \;\leq\; Cn\exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2} \right) + nq_{n,\pi} \quad \blacksquare
\end{aligned} \tag{G.10}$$

Controlling $\bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}$. Finally, in order to control $\bar{\boldsymbol{\varphi}}_{n[j]}^{(2)}(\mathcal{X}_n)$ which is an average of the i.i.d. random variables $\{\boldsymbol{\varphi}_{[j]}^2(T_i, \mathbf{X}_i)\}_{i=1}^n$, we first recall all notations from (G.3)-(G.4) and note that under Assumption 3.1 (a), $\|\mathbf{h}^2_{[j]}(\mathbf{X})\|_{\psi_1} \leq \sigma^2_{\mathbf{h}}$ $\forall\ j \in \{1,\ldots,d\}$ owing to Lemma C.1 (v). Further, $T^2/\pi^2(\mathbf{X}) \leq \delta_\pi^{-2}$ a.s. $[\mathbb{P}]$. Hence, using Lemma C.5 (b), we have: $\forall\ j \in \{1,\ldots,d\}$, and for some constants $\sigma_\pi \equiv \bar{\sigma}_{\boldsymbol{\varphi}} := 2\sqrt{2}\sigma^2_{\mathbf{h}}\delta_\pi^{-2}$ and $K_\pi \equiv \bar{K}_{\boldsymbol{\varphi}} := 2\sigma^2_{\mathbf{h}}\delta_\pi^{-2}$,

$$\text{(G.11)} \quad \boldsymbol{\varphi}^2_{[j]}(T,\mathbf{X}) \ \equiv \ \frac{T^2}{\pi^2(\mathbf{X})}\mathbf{h}^2_{[j]}(\mathbf{X}) \ \sim \ \mathrm{BMC}(\bar{\sigma}_{\boldsymbol{\varphi}}, \bar{K}_{\boldsymbol{\varphi}}) \quad \text{and further,}$$

$$\text{(G.12)} \ \boldsymbol{\mu}^{(2)}_{\boldsymbol{\varphi}[j]} \ \equiv \ \mathbb{E}\left\{\boldsymbol{\varphi}^2_{[j]}(T,\mathbf{X})\right\} \ = \ \mathbb{E}\left\{\frac{\mathbf{h}^2_{[j]}(\mathbf{X})}{\pi(\mathbf{X})}\right\} \ \leq \ \frac{\boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]}}{\delta_\pi} \ \leq \ \frac{\|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty}{\delta_\pi},$$

where $\|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty := \max\{\boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]} : j = 1,\ldots,d\} < \infty$ and $\boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]}$ is as in (G.4).

Using (G.11)-(G.12) along with Lemma C.4, we then have: for any $\epsilon_3 > 0$ and for each $j \in \{1,\ldots,d\}$,

$$\begin{aligned}
&\mathbb{P}\left\{\bar{\boldsymbol{\varphi}}^{(2)}_{n[j]} \ \equiv \ \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}^2_{[j]}(T_i,\mathbf{X}_i) \ > \ \frac{\|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty}{\delta_\pi} + \sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}}\epsilon_3 + \bar{K}_{\boldsymbol{\varphi}}\epsilon_3^2\right\} \\
&\quad \leq \ \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}^2_{[j]}(T_i,\mathbf{X}_i) - \boldsymbol{\mu}^{(2)}_{\boldsymbol{\varphi}[j]}\right| \ > \ \sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}}\epsilon_3 + \bar{K}_{\boldsymbol{\varphi}}\epsilon_3^2\right\} \\
\text{(G.13)} \quad &\quad \leq \ 2\exp\left(-n\epsilon_3^2\right). \qquad \blacksquare
\end{aligned}$$

For any $\epsilon_1, \epsilon_3 > 0$, and any $\epsilon_2 > 0$ such that $\epsilon_2 < \delta_\pi$, let us now define the event $\mathcal{A}_{\pi,n,j}(\epsilon_1,\epsilon_2,\epsilon_3)$, for each $j \in \{1,\ldots,d\}$, as follows.

$$\text{(G.14)} \ \ \mathcal{A}_{\pi,n,j}(\epsilon_1,\epsilon_2,\epsilon_3) := \left\{8c^2_{n,j}(\mathcal{X}_n) > d_n(\epsilon_1,\epsilon_2,\epsilon_3)\right\}, \ 1 \leq j \leq d, \ \text{where}$$

$$d_n(\epsilon_1,\epsilon_2,\epsilon_3) \ := \ \frac{8\sigma^2_\varepsilon\epsilon_1^2}{(\delta_\pi - \epsilon_2)^2}\left(\frac{\|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty}{\delta_\pi} + \sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}}\epsilon_3 + \bar{K}_{\boldsymbol{\varphi}}\epsilon_3^2\right).$$

Then, recalling from (G.6) that $c^2_{n,j}(\mathcal{X}_n) \equiv \sigma^2_\varepsilon \|\Delta_{\pi,n}\|^2_{\infty,n} \|\widetilde{\pi}_n\|^2_{\infty,n} \bar{\boldsymbol{\varphi}}^{(2)}_{n[j]}$ and using the bounds (G.9), (G.10) and (G.13) for $\|\Delta_{\pi,n}\|^2_{\infty,n}$, $\|\widetilde{\pi}_n\|^2_{\infty,n}$ and $\bar{\boldsymbol{\varphi}}^{(2)}_{n[j]}$ respectively, along with the union bound, we have:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{A}_{\pi,n,j}\right) \ \equiv \ \mathbb{P}_{\mathcal{X}_n}\left(\mathcal{A}_{\pi,n,j}\right) \ \equiv \ \mathbb{P}_{\mathcal{X}_n}\left\{8c^2_{n,j}(\mathcal{X}_n) > d_n(\epsilon_1,\epsilon_2,\epsilon_3)\right\} \\
\text{(G.15)} \quad &\quad \leq \ Cn\exp\left(\frac{-\epsilon_1^2}{v^2_{n,\pi}}\right) + Cn\exp\left(\frac{-\epsilon_2^2}{v^2_{n,\pi}}\right) + 2nq_{n,\pi} + 2\exp\left(-n\epsilon_3^2\right).
\end{aligned}$$

Therefore, it now follows that for each $j \in \{1, \ldots, d\}$ and any $\epsilon \geq 0$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_n}\left[\exp\left\{\frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)}\right\}\right] &= \mathbb{E}\left[\exp\left\{\frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)}\right\} \,\middle|\, \mathcal{A}_{\pi,n,j}^c\right]\mathbb{P}\left(\mathcal{A}_{\pi,n,j}^c\right) \\
&\quad + \mathbb{E}\left[\exp\left\{\frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)}\right\} \,\middle|\, \mathcal{A}_{\pi,n,j}\right]\mathbb{P}\left(\mathcal{A}_{\pi,n,j}\right) \\
&\leq \exp\left\{\frac{-n\epsilon^2}{d_n(\epsilon_1,\epsilon_2,\epsilon_3)}\right\} + 2\exp\left(-n\epsilon_3^2\right) + 2nq_{n,\pi} \\
&\quad + Cn\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2}\right) + Cn\exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2}\right) \quad [\text{using (G.14)-(G.15)}]. \qquad \text{(G.16)}
\end{aligned}$$

The final bound for $\|\mathbf{T}_{\pi,n}\|_\infty$. Using (G.16) in the fundamental bound (G.8) for $\|\mathbf{T}_{\pi,n}\|_\infty$, we finally have: for any $\epsilon \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{T}_{\pi,n}\|_\infty > \epsilon\right) &\leq \sum_{j=1}^{d} 2\,\mathbb{E}_{\mathcal{X}_n}\left[\exp\left\{\frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)}\right\}\right] \\
&\leq 2d\exp\left\{\frac{-n\epsilon^2}{d_n(\epsilon_1,\epsilon_2,\epsilon_3)}\right\} + 4d\exp\left(-n\epsilon_3^2\right) + 4q_{n,\pi}(nd) \\
&\quad + 2C(nd)\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2}\right) + 2C(nd)\exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2}\right) \quad [\text{using (G.16)}], \\
&\equiv 2\exp\left\{\frac{-n\epsilon^2}{d_n(\epsilon_1,\epsilon_2,\epsilon_3)} + \log d\right\} + 4\exp\left(-n\epsilon_3^2 + \log d\right) + 4q_{n,\pi}(nd) \\
&\quad + 2C\exp\left\{\frac{-\epsilon_1^2}{v_{n,\pi}^2} + \log(nd)\right\} + 2C\exp\left\{\frac{-\epsilon_2^2}{v_{n,\pi}^2} + \log(nd)\right\}. \qquad \text{(G.17)}
\end{aligned}$$

This leads to the desired bound and completes the proof of Theorem G.1. ■

G.2. Completing Proof of Theorem 3.2. We next evaluate the general tail bound for $\|\mathbf{T}_{\pi,n}\|_\infty$ in Theorem G.1 under a specific family of choices for $(\epsilon, \epsilon_1, \epsilon_2, \epsilon_3) > 0$ in order to understand its behavior and also establish the convergence rate of $\|\mathbf{T}_{\pi,n}\|_\infty$. Let $(c_1, c_2, c_3) > 1$ be any universal constants and set $\epsilon_1 = c_1 v_{n,\pi}\sqrt{\log(nd)}$, $\epsilon_2 = c_2 v_{n,\pi}\sqrt{\log(nd)}$ and $\epsilon_3 = c_3\sqrt{(\log d)/n}$, where we assume w.l.o.g. that $\epsilon_3 < 1$ and $\epsilon_2 \leq \delta_\pi/2$, so that $(\delta_\pi - \epsilon_2) \geq \delta_\pi/2$. Further with a choice of ϵ_3 as above, note that

$$\frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + \sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}}\epsilon_3 + \bar{K}_{\boldsymbol{\varphi}}\epsilon_3^2 \leq \frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + \left(\sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}} + \bar{K}_{\boldsymbol{\varphi}}\right)c_3\sqrt{\frac{\log d}{n}}.$$

Using these in the definition (G.14) and letting $C_{\boldsymbol{\varphi}} := (\sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}} + \bar{K}_{\boldsymbol{\varphi}})$, we get

$$d_n(\epsilon_1, \epsilon_2, \epsilon_3) \;\leq\; 8\sigma_\varepsilon^2 \frac{4c_1^2}{\delta_\pi^2} \{v_{n,\pi}\sqrt{\log(nd)}\}^2 \left(\frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + c_3 C_{\boldsymbol{\varphi}} \sqrt{\frac{\log d}{n}} \right).$$

Given these choices of $\{\epsilon_j\}_{j=1}^3$, let us now set $\epsilon = c\sqrt{\{(\log d)/n\} d_n(\epsilon_1, \epsilon_2, \epsilon_3)}$ for any universal constant $c > 1$. Using Theorem G.1, we then have:

$$\text{With probability at least } \; 1 - \frac{2}{d^{c^2-1}} - \frac{4}{d^{c_3^2-1}} - \sum_{j=1}^{2} \frac{2C}{(nd)^{c_j^2-1}} - 4q_{n,\pi}(nd),$$

$$\|\mathbf{T}_{\pi,n}\|_\infty \;\leq\; c\sqrt{\frac{\log d}{n}} \{v_{n,\pi}\sqrt{\log(nd)}\} C_1 \left(\frac{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty}{\delta_\pi} + C_2 \sqrt{\frac{\log d}{n}} \right)^{\frac{1}{2}},$$

where $C_1 := c_1(4\sqrt{2}\sigma_\varepsilon/\delta_\pi)$ and $C_2 := c_3 C_{\boldsymbol{\varphi}} \equiv c_3(\sqrt{2}\bar{\sigma}_{\boldsymbol{\varphi}} + \bar{K}_{\boldsymbol{\varphi}})$, with $\bar{\sigma}_{\boldsymbol{\varphi}}$ and $\bar{K}_{\boldsymbol{\varphi}}$ being as in (G.11). This completes the proof of Theorem 3.2. ∎

APPENDIX H: PROOF OF THEOREM 3.3

To show Theorem 3.3, we first state and prove a more general result that gives an explicit tail bound for $\|\mathbf{T}_{m,n}\|_\infty$.

THEOREM H.1 (Tail bound for $\|\mathbf{T}_{m,n}\|_\infty$). *Let Assumptions 1.1, 3.1 (a) and 3.3 hold with the sequences $(v_{\bar{n},m}, q_{\bar{n},m})$, $\bar{n} \equiv n/2$ and the constants $(\delta_\pi, \sigma_{\mathbf{h}}, C)$ as defined therein. Then, for any $\epsilon, \epsilon_1, \epsilon_2 \geq 0$,*

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{T}_{m,n}\|_\infty > \epsilon\right) \;&\leq\; 4\exp\left\{\frac{-\bar{n}\epsilon^2}{t_{\bar{n}}(\epsilon_1,\epsilon_2)} + \log d\right\} \;+\; 8\exp(-\bar{n}\epsilon_2^2 + \log d) \\
&\quad + 4C\exp\left\{\frac{-\epsilon_1^2}{v_{\bar{n},m}^2} + \log(\bar{n}d)\right\} + 4q_{\bar{n},m}(\bar{n}d), \quad \textit{where} \\
t_{\bar{n}}(\epsilon_1,\epsilon_2) \;&:=\; 8\bar{\delta}_\pi^2\epsilon_1^2\left(\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + \sqrt{2}\sigma_m\epsilon_2 + K_m\epsilon_2^2\right), \quad \textit{with}
\end{aligned}$$

$\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty := \max_{1\leq j\leq d}\mathbb{E}\{\mathbf{h}_{[j]}^2(\mathbf{X})\}$, $\bar{\delta}_\pi \leq \delta_\pi^{-1}$, $\sigma_m := 2\sqrt{2}\sigma_{\mathbf{h}}^2$ *and* $K_m := 2\sigma_{\mathbf{h}}^2$.

H.1. Proof of Theorem H.1. We first rewrite $\mathbf{T}_{m,n}$ from (3.1) as:

$$\begin{aligned}
\mathbf{T}_{m,n} &\equiv \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(\mathbf{X}_i)}-1\right\}\{\widetilde{m}(\mathbf{X}_i)-m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i)\\
&= \frac{1}{2\bar{n}}\sum_{k\neq k'=1}^{2}\sum_{i\in\mathcal{I}_{k'}}\left\{\frac{T_i}{\pi(\mathbf{X}_i)}-1\right\}\left\{\widehat{m}^{(k)}(\mathbf{X}_i)-m(\mathbf{X}_i)\right\}\mathbf{h}(\mathbf{X}_i)\\
&=: \frac{1}{2}\sum_{k\neq k'=1}^{2}\mathbf{T}_{m,\bar{n}}^{(k,k')},\ \text{ where }\ \mathbf{T}_{m,\bar{n}}^{(k,k')} := \frac{1}{\bar{n}}\sum_{i\in\mathcal{I}_{k'}}\mathbf{T}_m^{(k)}(\mathbf{Z}_i)\ \text{ and} \qquad \text{(H.1)}\\
\mathbf{T}_m^{(k)}(\mathbf{Z}) &:= \left\{\frac{T}{\pi(\mathbf{X})}-1\right\}\left\{\widehat{m}^{(k)}(\mathbf{X})-m(\mathbf{X})\right\}\mathbf{h}(\mathbf{X}) \quad \forall\, k\neq k'\in\{1,2\}.
\end{aligned}$$

Define $\mathcal{X}_{n,k}^* := \{\mathbf{X}_i : i\in\mathcal{I}_k\}\ \forall\, k\in\{1,2\}$, and let $\mathbb{E}_{\mathcal{X}_{n,k}^*}(\cdot)$ and $\mathbb{P}(\cdot\mid\mathcal{X}_{n,k}^*)$ respectively denote expectation w.r.t. $\mathcal{X}_{n,k}^*$ and conditional probability given $\mathcal{X}_{n,k}^*$. Further, for each $k\neq k'\in\{1,2\}$, let $\mathbb{E}_{\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*}(\cdot)$ and $\mathbb{P}(\cdot\mid\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*)$ respectively denote expectation w.r.t. $\{\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*\}$ and conditional probability given $\{\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*\}$. With $\mathcal{D}_n^{(k)}\perp\!\!\!\perp\mathcal{X}_{n,k'}^*\ \forall\, k\neq k'\in\{1,2\}$, we note that $\mathbb{E}_{\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*}(\cdot)=\mathbb{E}_{\mathcal{X}_{n,k'}^*}\{\mathbb{E}_{\mathcal{D}_n^{(k)}}(\cdot)\}$. Next, let us define: $\forall\, k\neq k'\in\{1,2\}$,

$$\Delta_{m,\bar{n}}^{(k)}(\mathbf{X}) := \widehat{m}^{(k)}(\mathbf{X})-m(\mathbf{X}),\quad \left\|\Delta_{m,\bar{n}}^{(k,k')}\right\|_{\infty,\bar{n}} := \max_{i\in\mathcal{I}_{k'}}\left|\Delta_{m,\bar{n}}^{(k)}(\mathbf{X}_i)\right|, \qquad \text{(H.2)}$$

$$\bar{\mathbf{h}}_{\bar{n}[j]}^{(2,k')} := \frac{1}{\bar{n}}\sum_{i\in\mathcal{I}_{k'}}\mathbf{h}_{[j]}^2(\mathbf{X}_i)\ \text{ and let }\ \psi(T,\mathbf{X}) := \frac{T}{\pi(\mathbf{X})}-1. \qquad \text{(H.3)}$$

Further, for any $a\in(0,1]$, let $\bar{a}:=2\widetilde{a}/a$, where $\widetilde{a}:=1/2$ if $a=1/2$, $\widetilde{a}:=0$ if $a=1$ and $\widetilde{a}:=[(a-1/2)/\log\{a/(1-a)\}]^{1/2}$ if $a\notin\{1/2,1\}$. Let $\{\bar{\pi}(\mathbf{X}),\widetilde{\pi}(\mathbf{X})\}$ and $\{\bar{\delta}_\pi,\widetilde{\delta}_\pi\}$ denote the corresponding versions of $\{\bar{a},\widetilde{a}\}$ for $a\equiv\pi(\mathbf{X})$ and $a\equiv\delta_\pi$ respectively, with δ_π being as in (1.1). We note that $\bar{a}$ is decreasing in $a\in(0,1]$ and $\widetilde{a}\le 1/2$, so that $\bar{a}\le 1/a\ \forall\, a\in(0,1]$. Using this and (1.1), we therefore have: $\bar{\pi}(\mathbf{x})\le\bar{\delta}_\pi\le 1/\delta_\pi\ \forall\,\mathbf{x}\in\mathcal{X}$.

Using the notations from (H.2) and (H.3), we have: for each $k\in\{1,2\}$,

$$\mathbf{T}_m^{(k)}(\mathbf{Z})\equiv\left\{\frac{T}{\pi(\mathbf{X})}-1\right\}\{\widehat{m}^{(k)}(\mathbf{X})-m(\mathbf{X})\}\mathbf{h}(\mathbf{X})=\psi(T,\mathbf{X})\Delta_{m,\bar{n}}^{(k)}(\mathbf{X})\mathbf{h}(\mathbf{X}).$$

Now, for each $k\in\{1,2\}$ and $k'\neq k\in\{1,2\}$, $\mathcal{D}_n^{(k)}\perp\!\!\!\perp\mathcal{X}_{n,k'}^*$ and we have: $\{\psi(T_i,\mathbf{X}_i)\mid\mathcal{D}_n^{(k)},\mathcal{X}_{n,k'}^*\}_{i\in\mathcal{I}_{k'}}\equiv\{\psi(T_i,\mathbf{X}_i)\mid\mathcal{X}_{n,k'}^*\}_{i\in\mathcal{I}_{k'}}\equiv\{\psi(T_i,\mathbf{X}_i)\mid\mathbf{X}_i\}_{i\in\mathcal{I}_{k'}}$

are (conditionally) independent sub-Gaussian random variables that satisfy:

$$\forall\, i \in \mathcal{I}_{k'},\ \ \mathbb{E}\{\psi(T_i, \mathbf{X}_i) \mid \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\} \ \equiv\ \mathbb{E}\{\psi(T_i, \mathbf{X}_i) \mid \mathbf{X}_i\} = 0 \quad \text{and}$$

$$\|\psi(T_i, \mathbf{X}_i) \mid \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\|_{\psi_2} \ \equiv\ \|\psi(T_i, \mathbf{X}_i) \mid \mathbf{X}_i\|_{\psi_2} \ \le\ \bar{\pi}^2(\mathbf{X}_i) \ \le\ \bar{\delta}_\pi^2, \tag{H.4}$$

where the bounds on the $\|\cdot\|_{\psi_2}$ norm follow from using Lemma C.3 and Lemma C.1 (i)(b) along with the definitions of $\bar{\pi}(\cdot)$ and $\bar{\delta}_\pi$ given earlier. Further, conditional on $\mathcal{D}_n^{(k)}$ and $\mathcal{X}_{n,k'}^*$, $\{\Delta_{m,\bar{n}}^{(k)}(\mathbf{X}_i)\}_{i\in\mathcal{I}_{k'}}$ and $\{\mathbf{h}_{[j]}(\mathbf{X}_i)\}_{i\in\mathcal{I}_{k'}}$, for each $j \in \{1,\ldots,d\}$, are all constants. Hence, using Lemma C.2 and (H.4), along with (H.1)-(H.3), we have: $\forall\ k \neq k' \in \{1,2\}$ and $j \in \{1,\ldots,d\}$,

$$\left\| \frac{1}{\bar{n}} \sum_{i\in\mathcal{I}_{k'}} \mathbf{T}_{m[j]}^{(k)}(\mathbf{Z}_i) \,\middle|\, \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^* \right\|_{\psi_2} \ \le\ \frac{4 d_{\bar{n},j}\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)}{\sqrt{\bar{n}}}, \quad \text{where} \tag{H.5}$$

$$d_{\bar{n},j}\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right) \ :=\ \bar{\delta}_\pi \left\| \Delta_{m,\bar{n}}^{(k,k')} \right\|_{\infty,\bar{n}} \left(\bar{\mathbf{h}}_{\bar{n}[j]}^{(2,k')} \right)^{1/2}.$$

Using Lemma C.2, we then have: $\forall\ k \neq k' \in \{1,2\}$, $1 \le j \le d$ and $\epsilon \ge 0$,

$$\mathbb{P}\left\{ \left| \frac{1}{\bar{n}} \sum_{i\in\mathcal{I}_{k'}} \mathbf{T}_{m[j]}^{(k)}(\mathbf{Z}_i) \right| > \epsilon \,\middle|\, \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^* \right\} \ \le\ 2\exp\left\{ \frac{-\bar{n}\epsilon^2}{8 d_{\bar{n},j}^2\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)} \right\}.$$

The fundamental bound for $\|\mathbf{T}_{m,\bar{n}}^{(k,k')}\|_\infty$. Using the bound obtained above for $\mathbf{T}_{m,\bar{n}[j]}^{(k,k')} \mid \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*$, we then have the following (unconditional) probabilistic bound for $\|\mathbf{T}_{m,\bar{n}}^{(k,k')}\|_\infty$. For any $\epsilon \ge 0$ and $k \neq k' \in \{1,2\}$,

$$\begin{aligned}
&\mathbb{P}\left\{ \left\| \mathbf{T}_{m,\bar{n}}^{(k,k')} \right\|_\infty \ \equiv\ \left\| \frac{1}{\bar{n}} \sum_{i\in\mathcal{I}_{k'}} \mathbf{T}_m^{(k)}(\mathbf{Z}_i) \right\|_\infty \ >\ \epsilon \right\} \\
&\quad \le\ \sum_{j=1}^d \mathbb{P}\left\{ \left| \frac{1}{\bar{n}} \sum_{i\in\mathcal{I}_{k'}} \mathbf{T}_{m[j]}^{(k)}(\mathbf{Z}_i) \right| \ >\ \epsilon \right\} \quad \text{[using the u.b.]}, \\
&\quad =\ \sum_{j=1}^d \mathbb{E}_{\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*} \left[\mathbb{P}\left\{ \left| \frac{1}{\bar{n}} \sum_{i\in\mathcal{I}_{k'}} \mathbf{T}_{m[j]}^{(k)}(\mathbf{Z}_i) \right| \ >\ \epsilon \,\middle|\, \mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^* \right\} \right] \\
&\quad \le\ 2 \sum_{j=1}^d \mathbb{E}_{\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*} \left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{8 d_{\bar{n},j}^2\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)} \right\} \right]. \qquad (\text{H.6}) \qquad \blacksquare
\end{aligned}$$

Next, we aim to control the random variable $d^2_{\bar{n},j}(\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'})$ appearing in (H.6). Based on the definition (H.5) of $d^2_{\bar{n},j}(\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'})$, it suffices to separately control $\left\|\Delta^{(k,k')}_{m,\bar{n}}\right\|^2_{\infty,\bar{n}}$ and $\bar{\mathbf{h}}^{(2,k')}_{\bar{n}[j]}$. To this end, let $\mathbb{E}_{\mathcal{D}_n^{(k)}}(\cdot)$ and $\mathbb{P}_{\mathcal{D}_n^{(k)}}(\cdot)$ denote expectation and probability w.r.t $\mathcal{D}_n^{(k)}$ $\forall\, k \in \{1,2\}$.

With $\mathcal{D}_n^{(k)} \perp\!\!\!\perp \mathcal{X}^*_{n,k'}$ for each $k \neq k' \in \{1,2\}$, we note that for any event $A \equiv A(\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'})$, $\mathbb{P}(A) \equiv \mathbb{P}_{\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'}}(A) = \mathbb{E}_{\mathcal{X}^*_{n,k'}}[\mathbb{E}_{\mathcal{D}_n^{(k)}}\{1(A) \mid \mathcal{X}^*_{n,k'}\}] \equiv \mathbb{E}_{\mathcal{X}^*_{n,k'}}[\mathbb{P}_{\mathcal{D}_n^{(k)}}\{A(\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'}) \mid \mathcal{X}^*_{n,k'}\}] = \mathbb{E}_{\mathcal{X}^*_{n,k'}}[\mathbb{P}_{\mathcal{D}_n^{(k)}}\{A(\mathcal{D}_n^{(k)}, \mathcal{X}^*_{n,k'})\}]$, where the final step holds since $\mathbb{P}_{\mathcal{D}_n^{(k)}}(\cdot \mid \mathcal{X}^*_{n,k}) = \mathbb{P}_{\mathcal{D}_n^{(k)}}(\cdot)$ as $\mathcal{D}_n^{(k)} \perp\!\!\!\perp \mathcal{X}^*_{n,k'}$.

Controlling $\left\|\Delta^{(k,k')}_{m,\bar{n}}\right\|^2_{\infty,\bar{n}}$. Using (3.8) in Assumption 3.3 along with the u.b. and the notations and facts discussed above, we have: $\forall\, k \neq k' \in \{1,2\}$,

$$
\begin{aligned}
&\mathbb{P}\left\{\left\|\Delta^{(k,k')}_{m,\bar{n}}\right\|^2_{\infty,\bar{n}} \equiv \max_{i \in \mathcal{I}_{k'}} \left|\Delta^{(k)}_{m,\bar{n}}(\mathbf{X}_i)\right|^2 > \epsilon_1^2\right\} \\
&\quad \leq \sum_{i \in \mathcal{I}_{k'}} \mathbb{P}\left\{\left|\Delta^{(k)}_{m,\bar{n}}(\mathbf{X}_i)\right| > \epsilon_1\right\} \leq \sum_{i \in \mathcal{I}_{k'}} \mathbb{E}_{\mathcal{X}^*_{n,k'}}\left\{C \exp\left(\frac{-\epsilon_1^2}{v^2_{\bar{n},m}}\right) + q_{\bar{n},m}\right\}
\end{aligned}
$$

$$
\equiv C\bar{n} \exp\left(\frac{-\epsilon_1^2}{v^2_{\bar{n},m}}\right) + \bar{n} q_{\bar{n},m} \quad \text{for any } \epsilon_1 \geq 0, \tag{H.7}
$$

where we also used that $\mathcal{D}_n^{(k)} \perp\!\!\!\perp \mathcal{X}^*_{n,k'}$ which ensures $\mathbb{P}_{\mathcal{D}_n^{(k)}}(\cdot \,|\, \mathcal{X}^*_{n,k}) = \mathbb{P}_{\mathcal{D}_n^{(k)}}(\cdot)$ and makes (3.8) in Assumption 3.3 applicable conditional on $\mathcal{X}^*_{n,k'}$. ■

Controlling $\bar{\mathbf{h}}^{(2,k')}_{\bar{n}[j]}$. We first recall that $\|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty = \max_{1 \leq j \leq d} \boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]}$, where $\boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]} \equiv \mathbb{E}\{\mathbf{h}^2_{[j]}(\mathbf{X})\}$. Now, $\forall\, k' \in \{1,2\}$ and $j \in \{1,\ldots,d\}$, $\bar{\mathbf{h}}^{(2,k')}_{\bar{n}[j]}$ is simply an average of the i.i.d. random variables $\{\mathbf{h}^2_{[j]}(\mathbf{X}_i)\}_{i \in \mathcal{I}_{k'}}$. Further, using Assumption 3.1 (a) and Lemma C.5 (a), $\mathbf{h}^2_{[j]}(\mathbf{X}) \sim \mathrm{BMC}(\bar{\sigma}_{\mathbf{h}}, \bar{K}_{\mathbf{h}})$ for some constants $\sigma_m \equiv \bar{\sigma}_{\mathbf{h}} := 2\sqrt{2}\sigma^2_{\mathbf{h}}$ and $K_m \equiv \bar{K}_{\mathbf{h}} := 2\sigma^2_{\mathbf{h}}$. Hence, using Lemma C.4, we have: for each $k' \in \{1,2\}$ and $j \in \{1,\ldots,d\}$, and for any $\epsilon_2 \geq 0$,

$$
\mathbb{P}\left\{\bar{\mathbf{h}}^{(2,k')}_{\bar{n}[j]} \equiv \frac{1}{\bar{n}} \sum_{i \in \mathcal{I}_{k'}} \mathbf{h}^2_{[j]}(\mathbf{X}_i) > \|\boldsymbol{\mu}^{(2)}_{\mathbf{h}}\|_\infty + \sqrt{2}\bar{\sigma}_{\mathbf{h}}\epsilon_2 + \bar{K}_{\mathbf{h}}\epsilon_2^2\right\} \tag{H.8}
$$

$$
\leq \mathbb{P}\left\{\left|\frac{1}{\bar{n}} \sum_{i \in \mathcal{I}_{k'}} \mathbf{h}^2_{[j]}(\mathbf{X}_i) - \boldsymbol{\mu}^{(2)}_{\mathbf{h}[j]}\right| > \sqrt{2}\bar{\sigma}_{\mathbf{h}}\epsilon_2 + \bar{K}_{\mathbf{h}}\epsilon_2^2\right\} \leq 2\exp(-\bar{n}\epsilon_2^2). \quad ■
$$

The final bound for $\left\| \mathbf{T}_{m,\bar{n}}^{(k,k')} \right\|_\infty$. For any $\epsilon_1, \epsilon_2 \geq 0$, let us now define:

$$t_{\bar{n}}(\epsilon_1, \epsilon_2) \;:=\; 8\bar{\delta}_\pi^2 \epsilon_1^2 \left(\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + \sqrt{2}\bar{\sigma}_{\mathbf{h}}\epsilon_2 + \bar{K}_{\mathbf{h}}\epsilon_2^2 \right). \tag{H.9}$$

Then, using the bounds (H.7) and (H.8) in the definition of $d_{\bar{n},j}^2(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*)$ in (H.5), we have: for each $k \neq k' \in \{1,2\}$, $j \in \{1, \ldots, d\}$ and $\epsilon_1, \epsilon_2 \geq 0$,

$$\begin{aligned} &\mathbb{P}\left\{ 8d_{\bar{n},j}^2(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*) \;>\; t_{\bar{n}}\,(\epsilon_1, \epsilon_2) \right\} \\ &\leq\; C\bar{n}\exp\left(\frac{-\epsilon_1^2}{v_{\bar{n},m}^2} \right) + \bar{n}q_{\bar{n},m} + 2\exp(-\bar{n}\epsilon_2^2). \end{aligned} \tag{H.10}$$

Using (H.10) in the fundamental bound (H.6) for $\|\mathbf{T}_{m,\bar{n}}^{(k,k')}\|_\infty$, we then have: for each $k \neq k' \in \{1,2\}$ and for any $\epsilon, \epsilon_1, \epsilon_2 \geq 0$,

$$\begin{aligned} \mathbb{P}\left\{ \left\| \mathbf{T}_{m,\bar{n}}^{(k,k')} \right\|_\infty > \; \epsilon \right\} \;&\leq\; 2\sum_{j=1}^d \mathbb{E}_{\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*} \left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{8d_{\bar{n},j}^2\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)} \right\} \right] \\ &\equiv\; 2\sum_{j=1}^d \mathbb{E}\left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{8d_{\bar{n},j}^2\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)} \right\} 1_{\left\{8d_{\bar{n},j}^2(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*) \,\leq\, t_{\bar{n}}(\epsilon_1,\epsilon_2)\right\}} \right] \\ &\quad + 2\sum_{j=1}^d \mathbb{E}\left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{8d_{\bar{n},j}^2\left(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*\right)} \right\} 1_{\left\{8d_{\bar{n},j}^2(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*) \,>\, t_{\bar{n}}(\epsilon_1,\epsilon_2)\right\}} \right] \\ &\leq\; 2d\left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{t_{\bar{n}}(\epsilon_1,\epsilon_2)} \right\} \;+\; \mathbb{P}\left\{ 8d_{\bar{n},j}^2(\mathcal{D}_n^{(k)}, \mathcal{X}_{n,k'}^*) \;>\; t_{\bar{n}}\,(\epsilon_1, \epsilon_2) \right\} \right] \\ &\leq\; 2d\left[\exp\left\{ \frac{-\bar{n}\epsilon^2}{t_{\bar{n}}(\epsilon_1,\epsilon_2)} \right\} + C\bar{n}\exp\left(\frac{-\epsilon_1^2}{v_{\bar{n},m}^2} \right) + \bar{n}q_{\bar{n},m} + 2\exp(-\bar{n}\epsilon_2^2) \right]. \end{aligned} \tag{H.11}$$

Thus, (H.11) establishes an explicit tail bound for $\left\| \mathbf{T}_{m,\bar{n}}^{(k,k')} \right\|_\infty$. ■

The final bound for $\|\mathbf{T}_{m,n}\|_\infty$. A tail bound for $\|\mathbf{T}_{m,n}\|_\infty$ now follows easily using (H.1) and (H.11) along with the u.b. For any $\epsilon, \epsilon_1, \epsilon_2 \geq 0$, we have:

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{T}_{m,n}\|_\infty \;>\; \epsilon \right) \;&\leq\; \mathbb{P}\left(\left\| \mathbf{T}_{m,\bar{n}}^{(1,2)} \right\|_\infty > \; \epsilon \right) \;+\; \mathbb{P}\left(\left\| \mathbf{T}_{m,\bar{n}}^{(2,1)} \right\|_\infty > \; \epsilon \right) \\ &\leq\; 4d\exp\left\{ \frac{-\bar{n}\epsilon^2}{t_{\bar{n}}(\epsilon_1,\epsilon_2)} \right\} + 4C\bar{n}d\exp\left(\frac{-\epsilon_1^2}{v_{\bar{n},m}^2} \right) + 4\bar{n}dq_{\bar{n},m} + 8d\exp(-\bar{n}\epsilon_2^2). \end{aligned} \tag{H.12}$$

This leads to the desired bound and concludes the proof of Theorem H.1. ■

H.2. Completing the Proof of Theorem 3.3. Given the general tail bound for $\|\mathbf{T}_{m,n}\|_\infty$ in Theorem H.1, we next evaluate it for a specific set of choices of $(\epsilon, \epsilon_1, \epsilon_2) > 0$ in order to understand its behavior and also establish the convergence rate of $\|\mathbf{T}_{m,n}\|_\infty$. To this end, let $(c_1, c_2) > 1$ be any universal constants and set $\epsilon_1 = c_1 v_{\bar{n},m}\sqrt{\log(\bar{n}d)}$ and $\epsilon_2 = c_2\sqrt{(\log d)/\bar{n}}$, where we further assume w.l.o.g. that $\epsilon_2 < 1$ so that

$$\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + \sqrt{2}\bar{\sigma}_{\mathbf{h}}\epsilon_2 + \bar{K}_{\mathbf{h}}\epsilon_2^2 \;\leq\; \|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + \left(\sqrt{2}\bar{\sigma}_{\mathbf{h}} + \bar{K}_{\mathbf{h}}\right) c_2\sqrt{\frac{\log d}{\bar{n}}}.$$

Using these in the definition (H.9) and letting $C_{\mathbf{h}} := (\sqrt{2}\bar{\sigma}_{\mathbf{h}} + \bar{K}_{\mathbf{h}})$, we have:

$$t_{\bar{n}}(\epsilon_1, \epsilon_2) \;\leq\; 8c_1^2\bar{\delta}_\pi^2\{v_{\bar{n},m}\sqrt{\log(\bar{n}d)}\}^2\left\{\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + c_2 C_{\mathbf{h}}\sqrt{\frac{\log d}{\bar{n}}}\right\}.$$

Given these choices of $\{\epsilon_j\}_{j=1}^2$, let us now set $\epsilon = c\sqrt{\{(\log d)/\bar{n}\}t_{\bar{n}}(\epsilon_1, \epsilon_2)}$ for any $c > 1$. Using Theorem H.1 and with $\bar{n} \equiv n/2 \leq n$, we then have:

$$\text{With probability at least } 1 - \frac{4}{d^{c^2-1}} - \frac{8}{d^{c_2^2-1}} - \frac{4C}{(\bar{n}d)^{c_1^2-1}} - 4q_{\bar{n},m}(\bar{n}d),$$

$$\|\mathbf{T}_{m,n}\|_\infty \;\leq\; c\sqrt{\frac{\log d}{n}}\{v_{\bar{n},m}\sqrt{\log(nd)}\}C_1^*\left(\|\boldsymbol{\mu}_{\mathbf{h}}^{(2)}\|_\infty + C_2^*\sqrt{\frac{\log d}{n}}\right)^{\frac{1}{2}},$$

where $C_1^* := 4c_1\bar{\delta}_\pi$ and $C_{2,n}^* := \sqrt{2}c_2 C_{\mathbf{h}} \equiv \sqrt{2}c_2(\sqrt{2}\bar{\sigma}_{\mathbf{h}} + \bar{K}_{\mathbf{h}})$, with $\bar{\sigma}_{\mathbf{h}}$ and $\bar{K}_{\mathbf{h}}$ being as in (H.8). This completes the proof of Theorem 3.3. ■

APPENDIX I: PROOF OF THEOREM 3.4

To show Theorem 3.4, we first state and prove a more general result that gives an explicit tail bound for $\|\mathbf{R}_{\pi,m,n}\|_\infty$.

THEOREM I.1 (Tail bound for $\|\mathbf{R}_{\pi,m,n}\|_\infty$). *Let Assumptions 1.1, 3.1, 3.2 and 3.3 hold with the sequences $(v_{n,\pi}, q_{n,\pi})$, $(v_{\bar{n},m}, q_{\bar{n},m}, \bar{n})$ and the constants $(\delta_\pi, \sigma_{\mathbf{h}}, C)$ as defined therein, and let $\|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty := \max\{\mathbb{E}\{|\mathbf{h}_{[j]}(\mathbf{X})|\} : j = 1, \ldots, d\}$. Then, for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \geq 0$ with $\epsilon_2 < \delta_\pi$ small enough,*

$$\mathbb{P}\left\{\|\mathbf{R}_{\pi,m,n}\|_\infty \;>\; \frac{\epsilon_1\epsilon_3}{\delta_\pi - \epsilon_2}r_*(\epsilon_4)\right\} \;\leq\; 2d\exp(-n\epsilon_4^2)$$
$$+Cn\left\{\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2}\right) + \exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2}\right) + \exp\left(\frac{-\epsilon_3^2}{v_{\bar{n},m}^2}\right)\right\} + 2nq_{n,\pi} + nq_{\bar{n},m},$$

where $r_*(\epsilon_4) := \|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty + \sqrt{2}\sigma_{\pi,m}\epsilon_4 + K_{\pi,m}\epsilon_4^2$ with $\sigma_{\pi,m} := 4\sigma_{\mathbf{h}}\delta_\pi^{-1}$ and $K_{\pi,m} := 2\sqrt{2}\sigma_{\mathbf{h}}\delta_\pi^{-1}$ being constants.

I.1. Proof of Theorem I.1. Recalling from the notations in (3.1),

$$\mathbf{R}_{\pi,m,n} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{T_i}{\pi(\mathbf{X}_i)}\right\}\{\widetilde{m}(\mathbf{X}_i) - m(\mathbf{X}_i)\}\,\mathbf{h}(\mathbf{X}_i). \tag{I.1}$$

Hence, with $\|\Delta_{\pi,n}\|_{\infty,n}$ and $\|\widetilde{\pi}_n\|_{\infty,n}$ as in (G.1) and (G.2) respectively, and with $\left\|\Delta_{m,\bar{n}}^{(k,k')}\right\|_{\infty,\bar{n}}$ as in (H.2) for any $k \neq k' \in \{1,2\}$, we have:

$$\|\mathbf{R}_{\pi,m,n}\|_{\infty} \;\le\; \|\widetilde{\pi}_n\|_{\infty,n}\,\|\Delta_{\pi,n}\|_{\infty,n}\left\|\Delta^*_{m,n}\right\|_{\infty,n}\|\bar{\boldsymbol{\xi}}_n\|_{\infty}, \quad \text{where} \tag{I.2}$$

$$\left\|\Delta^*_{m,n}\right\|_{\infty,n} \;:=\; \max\left\{\left\|\Delta_{m,\bar{n}}^{(1,2)}\right\|_{\infty,\bar{n}}, \left\|\Delta_{m,\bar{n}}^{(2,1)}\right\|_{\infty,\bar{n}}\right\} \;\text{ and}$$

$$\bar{\boldsymbol{\xi}}_n \;:=\; \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}(T_i,\mathbf{X}_i), \;\text{ with }\; \boldsymbol{\xi}(T,\mathbf{X}) \;:=\; \left\{\frac{T}{\pi(\mathbf{X})}\left|\mathbf{h}_{[j]}(\mathbf{X})\right|\right\}_{j=1}^{d} \in \mathbb{R}^d.$$

For most of the quantities appearing in the bound (I.2), we already have their explicit tail bounds. Specifically, using (G.9), we have: for any $\epsilon_1 \ge 0$,

$$\mathbb{P}\left\{\|\Delta_{\pi,n}\|_{\infty,n} \;>\; \epsilon_1\right\} \;\le\; Cn\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2}\right) + nq_{n,\pi}, \quad \text{where} \tag{I.3}$$

and using (G.10), for any $\epsilon_2 \ge 0$ small enough such that $\epsilon_2 < \delta_\pi$,

$$\mathbb{P}\left\{\|\widetilde{\pi}_n\|_{\infty,n} \;>\; (\delta_\pi - \epsilon_2)^{-1}\right\} \;\le\; Cn\exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2}\right) + nq_{n,\pi}. \tag{I.4}$$

Next, using (H.7) and recalling that $\bar{n} = n/2$, we have: for any $\epsilon_3 \ge 0$,

$$\mathbb{P}\left\{\left\|\Delta^*_{m,n}\right\|_{\infty,n} > \epsilon_3\right\} \;\le\; \sum_{k\neq k'\in\{1,2\}}\mathbb{P}\left\{\left\|\Delta_{m,\bar{n}}^{(k,k')}\right\|_{\infty,\bar{n}} \;>\; \epsilon_3\right\}$$

$$\le\; 2C\bar{n}\exp\left(\frac{-\epsilon_3^2}{v_{\bar{n},m}^2}\right) + 2\bar{n}q_{\bar{n},m} \;\;\equiv\;\; Cn\exp\left(\frac{-\epsilon_3^2}{v_{\bar{n},m}^2}\right) + nq_{\bar{n},m}. \tag{I.5}$$

Finally, $\bar{\boldsymbol{\xi}}_n$ is a simple i.i.d. average defined by the random vector $\boldsymbol{\xi}(T,\mathbf{X})$ and can be controlled as follows. Under Assumption 3.1 (a) and Lemma C.1 (ii)(a), $\||\mathbf{h}_{[j]}(\mathbf{X})|\|_{\psi_1} = \|\mathbf{h}_{[j]}(\mathbf{X})\|_{\psi_1} \le \sqrt{2}\|\mathbf{h}_{[j]}(\mathbf{X})\|_{\psi_2} \le \sqrt{2}\sigma_{\mathbf{h}}\ \forall\ 1 \le j \le d$. Further, due to (1.1), $T/\pi(\mathbf{X}) \le \delta_\pi^{-1}$ a.s. $[\mathbb{P}]$. Hence, using Lemma C.5 (ii), we have: for constants $\sigma_{\pi,m} \equiv \bar{\sigma}_{\boldsymbol{\xi}} := 4\sigma_{\mathbf{h}}\delta_\pi^{-1}$ and $K_{\pi,m} \equiv \bar{K}_{\boldsymbol{\xi}} := 2\sqrt{2}\sigma_{\mathbf{h}}\delta_\pi^{-1}$,

$$\boldsymbol{\xi}_{[j]}(T,\mathbf{X}) \;\equiv\; \frac{T}{\pi(\mathbf{X})}|\mathbf{h}_{[j]}(\mathbf{X})| \;\sim\; \mathrm{BMC}(\bar{\sigma}_{\boldsymbol{\xi}}, \bar{K}_{\boldsymbol{\xi}}) \quad \forall\, j \in \{1,\ldots,d\}. \tag{I.6}$$

Further, $\mathbb{E}\{\boldsymbol{\xi}_{[j]}(\mathbb{T},\mathbf{X})\} = \mathbb{E}\{|\mathbf{h}_{[j]}(\mathbf{X})|\} \equiv \boldsymbol{\mu}_{|\mathbf{h}_{[j]}|}$ (say) $\forall\ j \in \{1,\ldots,d\}$, and recall that $\|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty = \max\{\boldsymbol{\mu}_{|\mathbf{h}_{[j]}|} : j = 1,\ldots,d\}$. Using (I.6) and Lemma C.4 along with the u.b., we then have: for any $\epsilon_4 \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\left\{\|\bar{\boldsymbol{\xi}}_n\|_\infty \;>\; r_*(\epsilon_4) \;\equiv\; \|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty + \sqrt{2}\bar{\sigma}_{\boldsymbol{\xi}}\epsilon_4 + \bar{K}_{\boldsymbol{\xi}}\epsilon_4^2\right\} \\
&\leq\; \sum_{j=1}^d \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_{[j]}(T_i,\mathbf{X}_i) - \boldsymbol{\mu}_{|\mathbf{h}_{[j]}|}\right| \;>\; \sqrt{2}\bar{\sigma}_{\boldsymbol{\xi}}\epsilon_4 + \bar{K}_{\boldsymbol{\xi}}\epsilon_4^2\right\} \\
&\leq\; 2d\exp(-n\epsilon_4^2) \;\equiv\; 2\exp(-n\epsilon_4^2 + \log d).
\end{aligned} \tag{I.7}
$$

Using the bounds (I.3), (I.4), (I.5) and (I.7), along with the u.b., in the original bound (I.2) for $\|\mathbf{R}_{\pi,m,n}\|_\infty$, we then have: for any $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4 \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\left\{\|\mathbf{R}_{\pi,m,n}\|_\infty \;>\; \frac{\epsilon_1\epsilon_3}{\delta_\pi - \epsilon_2} r_*(\epsilon_4)\right\} \;\leq\; 2d\exp(-n\epsilon_4^2) \\
&+Cn\left\{\exp\left(\frac{-\epsilon_1^2}{v_{n,\pi}^2}\right) + \exp\left(\frac{-\epsilon_2^2}{v_{n,\pi}^2}\right) + \exp\left(\frac{-\epsilon_3^2}{v_{\bar{n},m}^2}\right)\right\} + 2nq_{n,\pi} + nq_{\bar{n},m},
\end{aligned} \tag{I.8}
$$

where we assume that $\epsilon_2 < \delta_\pi$. The proof of Theorem I.1 is complete. ∎

I.2. Completing the Proof of Theorem 3.4. Given the general tail bound for $\|\mathbf{R}_{\pi,m,n}\|_\infty$ in Theorem I.1, we next evaluate it under a specific set of choices for $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4 > 0$ to understand its behavior and to establish the convergence rate of $\|\mathbf{R}_{\pi,m,n}\|_\infty$. Let $c_1,c_2,c_3,c_4 > 1$ be universal constants, and set $\epsilon_1 = c_1 v_{n,\pi}\sqrt{\log n}$, $\epsilon_2 = c_2 v_{n,\pi}\sqrt{\log n}$, $\epsilon_3 = c_3 v_{\bar{n},m}\sqrt{\log n}$ and $\epsilon_4 = c_4\sqrt{(\log d)/n}$, where we assume w.l.o.g. that $\epsilon_2 \leq \delta_\pi/2$ and $\epsilon_4 < 1$, so that

$$
r_*(\epsilon_4) \;\leq\; \|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty + c_4 C_{\boldsymbol{\xi}}\sqrt{\frac{\log d}{n}}, \quad \text{where} \;\; C_{\boldsymbol{\xi}} := \sqrt{2}\bar{\sigma}_{\boldsymbol{\xi}} + \bar{K}_{\boldsymbol{\xi}}
$$

with $\bar{\sigma}_{\boldsymbol{\xi}}$ and $\bar{K}_{\boldsymbol{\xi}}$ as in (I.6). Using Theorem I.1, we then have: with probability at least $1 - \sum_{j=1}^3 Cn^{-(c_j^2-1)} - 2d^{-(c_4^2-1)} - 2nq_{n,\pi} - nq_{\bar{n},m}$,

$$
\|\mathbf{R}_{\pi,m,n}\|_\infty \;\leq\; \frac{2c_1c_3}{\delta_\pi}\{v_{n,\pi}v_{\bar{n},m}(\log n)\}\left(\|\boldsymbol{\mu}_{|\mathbf{h}|}\|_\infty + c_4 C_{\boldsymbol{\xi}}\sqrt{\frac{\log d}{n}}\right), \quad \text{where}
$$

This leads to the desired bound and completes the proof of Theorem 3.4. ∎

APPENDIX J: PROOF OF THEOREM 4.1

Under the assumed form of $L(\cdot)$ and recalling the definition of $\widehat{\boldsymbol{\Sigma}}$ and that $\nabla\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}) = \nabla\widetilde{\mathcal{L}}_n^{\text{DDR}}(\boldsymbol{\theta})$, we first note that the gradient $\nabla\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta})$ satisfies:

$$
\nabla\mathcal{L}_n^{\text{DDR}}(\widehat{\boldsymbol{\theta}}_{\text{DDR}}) - \nabla\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) \;=\; 2\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0).
$$

Using the definition (4.1) of $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$ and the notations in (4.2), we then have:

$$
\begin{aligned}
(\widetilde{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) &= (\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) - \frac{1}{2}\widehat{\boldsymbol{\Omega}}\{\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) + 2\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0)\} \\
&= -\frac{1}{2}\boldsymbol{\Omega}\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) - \frac{1}{2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) + (I_d - \widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) \\
&\equiv -\frac{1}{2}\boldsymbol{\Omega}\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) + \mathbf{R}_{n,1} + \mathbf{R}_{n,3} \qquad \text{[using (4.2)]}.
\end{aligned} \tag{J.1}
$$

Next, recall from (3.1) that $\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0) \equiv \mathbf{T}_n = \mathbf{T}_{0,n} + \mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n}$, with all notations as in (3.2)-(3.5). Further, with our choice of $L(\cdot)$, we have:

$$
\mathbf{T}_{0,n} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}_0(\mathbf{Z}_i) = -\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_0(\mathbf{Z}_i), \quad \text{with } \boldsymbol{\psi}_0(\mathbf{Z}) \text{ as in the ALE (4.3).}
$$

Applying these facts in (J.1) and using the notations in (4.2), we then have:

$$
\begin{aligned}
(\widetilde{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) &= -\frac{1}{2}\boldsymbol{\Omega}(\mathbf{T}_{0,n} + \mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n}) + \mathbf{R}_{n,1} + \mathbf{R}_{n,3} \\
&= -\frac{1}{2}\boldsymbol{\Omega}\mathbf{T}_{0,n} - \frac{1}{2}\boldsymbol{\Omega}(\mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n}) + \mathbf{R}_{n,1} + \mathbf{R}_{n,3} \\
&\equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Omega}\boldsymbol{\psi}_0(\mathbf{Z}_i) + \mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,3} \equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Omega}\boldsymbol{\psi}_0(\mathbf{Z}_i) + \boldsymbol{\Delta}_n.
\end{aligned} \tag{J.2}
$$

Now, under Assumptions 1.1, 3.1, 3.2 and 3.3, all of Theorems 3.1-3.4 apply, and under Assumption 2.1 and with $L(\cdot)$ being convex and differentiable in $\boldsymbol{\theta}$ trivially, Lemma 2.1 applies as well. Using these results, we then have:

$$
\|\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{n}}\right) \text{ and } \|\widehat{\boldsymbol{\theta}}_{\text{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\|_1 = O_{\mathbb{P}}\left(s\sqrt{\frac{\log d}{n}}\right) \tag{J.3}
$$

for any choice of $\lambda_n \asymp \sqrt{(\log d)/n}$, as assumed. Using these facts along with Assumption 4.1 (a) and multiple uses of L_1-L_∞ type bounds, we then have:

$$
\|\mathbf{R}_{n,1}\|_\infty \leq \frac{1}{2}\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 \|\boldsymbol{\nabla}\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty = O_{\mathbb{P}}\left(r_n\sqrt{\frac{\log d}{n}}\right), \text{ and} \tag{J.4}
$$

$$
\|\mathbf{R}_{n,3}\|_\infty \leq \|I_d - \widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}}\|_{\max}\|\widehat{\boldsymbol{\theta}}_{\text{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\|_1 = O_{\mathbb{P}}\left(\omega_n s\sqrt{\frac{\log d}{n}}\right). \tag{J.5}
$$

Next, to control $\mathbf{R}_{n,2} \equiv -\frac{1}{2}\boldsymbol{\Omega}(\mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n})$, observe that each of the variables $-\frac{1}{2}\boldsymbol{\Omega}\mathbf{T}_{\pi,n}$, $-\frac{1}{2}\boldsymbol{\Omega}\mathbf{T}_{m,n}$ and $-\frac{1}{2}\boldsymbol{\Omega}\mathbf{R}_{\pi,m,n}$ admit exactly the *same* *form* as $\mathbf{T}_{\pi,n}$, $\mathbf{T}_{m,n}$ and $\mathbf{R}_{\pi,m,n}$ in (3.1), respectively, but with a *different choice* of the function $\mathbf{h}(\mathbf{X})$ in the definitions (3.3)-(3.5) of the underlying summands for these terms. In this particular case, the summands correspond to the forms (3.3)-(3.5) with $h(\mathbf{X})$ replaced by $\widetilde{h}(\mathbf{X}) = \boldsymbol{\Omega}\boldsymbol{\Psi}(\mathbf{X}) \equiv \boldsymbol{\Upsilon}(\mathbf{X})$.

Further under Assumption 4.1 (b), $\widetilde{h}(\mathbf{X})$ is sub-Gaussian with $\|\widetilde{h}(\mathbf{X})\|_{\psi_2} \leq \sigma_{\boldsymbol{\Upsilon}}$, as required in Assumption 3.1 (a). Hence, under Assumptions 1.1, 3.1, 3.2, 3.3 and 4.1, Theorems 3.2, 3.3 and 3.4 certainly apply to $\boldsymbol{\Omega}\mathbf{T}_{\pi,n}$, $\boldsymbol{\Omega}\mathbf{T}_{m,n}$ and $\boldsymbol{\Omega}\mathbf{R}_{\pi,m,n}$ with this 'modified choice' $\widetilde{h}(\mathbf{X})$ of $\mathbf{h}(\mathbf{X})$, using which we have:

$$\begin{aligned} \|\boldsymbol{\Omega}\mathbf{T}_{\pi,n}\|_\infty + \|\boldsymbol{\Omega}\mathbf{T}_{m,n}\|_\infty &= O_{\mathbb{P}}\left((v_{n,\pi} + v_{\bar{n},m}) \sqrt{\frac{(\log d)\log(nd)}{n}} \right) \\ \text{and} \quad \|\boldsymbol{\Omega}\mathbf{R}_{\pi,m,n}\|_\infty &= O_{\mathbb{P}}\left(v_{n,\pi} v_{\bar{n},m} \log n\right), \end{aligned}$$

where both results follow directly from the non-asymptotic bounds in Theorems 3.2-3.4. Combining these and recalling the definition of v_n^* in Assumption 4.1 (b) along with the rate condition on v_n^* assumed therein, we have:

$$\|\mathbf{R}_{n,3}\|_\infty \equiv \frac{1}{2}\|\boldsymbol{\Omega}(\mathbf{T}_{\pi,n} - \mathbf{T}_{m,n} - \mathbf{R}_{\pi,m,n})\|_\infty = O_{\mathbb{P}}\left(v_n^* n^{-\frac{1}{2}}\right). \tag{J.6}$$

Combining (J.4), (J.5) and (J.6) along with the definition of $\boldsymbol{\Delta}_n$ in (4.2), and using these in the original decomposition (J.2) of $(\widetilde{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0)$, we have:

$$\begin{aligned} (\widetilde{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0) &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Omega}\boldsymbol{\psi}_0(\mathbf{Z}_i) + \boldsymbol{\Delta}_n, \quad \text{where } \boldsymbol{\Delta}_n \text{ satisfies:} \\ \|\boldsymbol{\Delta}_n\|_\infty &\equiv \|\mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,2}\|_\infty \leq \|\mathbf{R}_{n,1}\|_\infty + \|\mathbf{R}_{n,2}\|_\infty + \|\mathbf{R}_{n,3}\|_\infty \\ &= O_{\mathbb{P}}\left(r_n\sqrt{\frac{\log d}{n}} + v_n^* n^{-\frac{1}{2}} + \omega_n s\sqrt{\frac{\log d}{n}} \right) = o_{\mathbb{P}}(n^{-\frac{1}{2}}). \quad \blacksquare \end{aligned} \tag{J.7}$$

(J.7) therefore establishes the desired ALE (4.3). Note further that the claim $\mathbb{E}\{\boldsymbol{\psi}_0(\mathbf{Z})\} = \mathbf{0}$ holds as a simple consequence of the definition of $\boldsymbol{\theta}_0$ and Assumption 1.1 (b). Specifically, recalling the notations $\varepsilon(\mathbb{Z}) = Y - m(\mathbf{X})$ and $\psi(\mathbf{X}) = m(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}_0)$ from Assumption 3.1 (a), with $g(\mathbf{X}, \boldsymbol{\theta}_0) = \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}_0$ for our choice of $L(\cdot)$, we have: $\mathbb{E}\{\varepsilon(\mathbb{Z})|\mathbf{X}\} = 0$, by definition of $m(\mathbf{X})$, and hence, $\mathbb{E}\{\psi(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X})\} = \mathbb{E}[\boldsymbol{\Psi}(\mathbf{X})\{Y - \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}_0\}] - \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\varepsilon(\mathbb{Z})\} = \mathbf{0}$, by definition of $\boldsymbol{\theta}_0$ and $L(\cdot)$. Further, $T \perp\!\!\!\perp Y|\mathbf{X}$ by Assumption 1.1 (a). Thus,

$$\mathbb{E}\{\boldsymbol{\psi}_0(\mathbf{Z})\} \equiv \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\psi(\mathbf{X})\} + \mathbb{E}_{\mathbf{X}}[\mathbb{E}\{T\pi^{-1}(\mathbf{X})|\mathbf{X}\}\mathbb{E}\{\varepsilon(\mathbb{Z})|\mathbf{X}\}] = \mathbf{0}.$$

This therefore completes the proof of the first part of Theorem 4.1. ■

To establish the (coordinatewise) asymptotic normality results claimed in the second part, we simply use the established ALE (4.3) or (J.7) and invoke Lyapunov's Central Limit Theorem (CLT) along with Slutsky's Theorem. To apply Lyapunov's CLT, we need to verify the Lyapunov moment conditions for $\mathbf{\Gamma}_0(\mathbf{Z}) \equiv \mathbf{\Omega}\boldsymbol{\psi}_0(\mathbf{Z})$. We establish this by first showing that $\mathbf{\Gamma}_0(\mathbf{Z})$ is, in fact, sub-exponential (as per Definition C.1 with $\alpha = 1$) under our assumptions.

To this end, under Assumptions 3.1 (a), 1.1 (b) and 4.1 (b), we have:

$$\text{(J.8)} \quad \|\mathbf{\Gamma}_0(\mathbf{Z})\|_{\psi_1} \equiv \|\mathbf{\Omega}\boldsymbol{\psi}_0(\mathbf{Z})\|_{\psi_1} = \|\mathbf{\Omega}\mathbf{\Psi}(\mathbf{X})\{\psi(\mathbf{X}) + T\pi^{-1}(\mathbf{X})\varepsilon(\mathbb{Z})\}\|_{\psi_1}$$
$$\leq \|\mathbf{\Omega}\mathbf{\Psi}(\mathbf{X})\|_{\psi_2}\{\|\psi(\mathbf{X})\|_{\psi_2} + \|\varepsilon(\mathbb{Z})\|_{\psi_2}\delta_\pi^{-1}\} \leq \sigma_{\mathbf{\Upsilon}}(\sigma_\psi + \delta_\pi^{-1}\sigma_\varepsilon) =: \sigma_{\mathbf{\Gamma}_0},$$

where the steps follow from using Lemma C.1 (v) and (i) (c). Consequently, using (J.8) and Lemma C.1 (iv) (a), we have: uniformly in $j \in \{1, \ldots, d\}$,

$$\rho_{\mathbf{\Gamma}_0,j} := \mathbb{E}\{|\mathbf{\Gamma}_{0[j]}(\mathbf{Z})|^3\} \leq 6\sigma_{\mathbf{\Gamma}_0}^3 < \infty \text{ and } \sigma_{0,j}^2 := \mathbb{E}\{|\mathbf{\Gamma}_{0[j]}(\mathbf{Z})|^2\} > c_0^2,$$

where the second result is due to the lower bound condition assumed on $\sigma_{0,j}$ with the constant $c_0 > 0$ as defined there. Hence, $\rho_{\mathbf{\Gamma}_0,j}/\sigma_{0,j}^3 \leq 6\sigma_{\mathbf{\Gamma}_0}^3/c_0^3 < \infty$ uniformly in $j \in \{1, \ldots, d\}$. Thus, the Lyapunov moment conditions are now verified (uniformly) for each coordinate of $\mathbf{\Gamma}_0(\mathbf{Z}) \equiv \mathbf{\Omega}\mathbf{\Psi}_0(\mathbf{Z})$. Note also that $\mathbb{E}\{\mathbf{\Gamma}_0(\mathbf{Z})\} = \mathbf{0}$ since $\mathbb{E}\{\boldsymbol{\psi}_0(\mathbf{Z})\} = \mathbf{0}$, as shown earlier. Finally, observe that $\sigma_{0,j}^{-1}|\mathbf{\Delta}_{n[j]}| \leq c_0^{-1}\|\mathbf{\Delta}_n\|_\infty = o_{\mathbb{P}}(n^{-\frac{1}{2}})$. Hence, by Lyapunov's CLT along with multiple uses of Slutsky's Theorem, we have: for each $1 \leq j \leq d$,

$$\text{(J.9)} \quad \sqrt{n}\sigma_{0,j}^{-1}\left(\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}[j]} - \boldsymbol{\theta}_{0[j]}\right) = \frac{1}{\sqrt{n}\sigma_{0,j}}\sum_{i=1}^n \mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i) + \sqrt{n}\sigma_{0,j}^{-1}\mathbf{\Delta}_{n[j]|}$$
$$= \frac{1}{\sqrt{n}\sigma_{0,j}}\sum_{i=1}^n \mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0,1) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0,1). \quad \blacksquare$$

This establishes the first of the two (coordinatewise) asymptotic normality claims in Theorem 4.1. For the second claim, we mainly need to establish the consistency of the estimator $\widehat{\sigma}_{0,j}^2$ of $\sigma_{0,j}^2$, uniformly in $1 \leq j \leq d$, as claimed. The asymptotic normality then follows directly from Slutsky's Theorem and (J.9). To establish the consistency, we first note that for all $1 \leq j \leq d$,

(J.10)

$$\sigma_{0,j}^2 - \sigma_{0,j}^2 \equiv \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{\Gamma}}_{0[j]}^2(\mathbf{Z}_i) - \mathbb{E}\{\mathbf{\Gamma}_{0[j]}^2(\mathbf{Z})\}$$
$$= \left\{\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{\Gamma}}_{0[j]}^2(\mathbf{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbf{\Gamma}_{0[j]}^2(\mathbf{Z}_i)\right\} + \left\{\frac{1}{n}\sum_{i=1}^n \mathbf{\Gamma}_{0[j]}^2(\mathbf{Z}_i) - \mathbb{E}\{\mathbf{\Gamma}_{0[j]}^2(\mathbf{Z})\}\right\},$$

where $\mathbf{\Gamma}_0(\mathbf{Z}) = \mathbf{\Omega}\boldsymbol{\psi}_0(\mathbf{Z})$ and $\widehat{\mathbf{\Gamma}}_0(\mathbf{Z}) = \widehat{\mathbf{\Omega}}\widehat{\boldsymbol{\psi}}_0(\mathbf{Z})$ with $\widehat{\boldsymbol{\psi}}_0(\mathbf{Z})$ given by:

$$\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}) \ := \ \left[\{\widehat{m}(\mathbf{X}) - \mathbf{\Psi}(\mathbf{X})'\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}\} + \frac{T}{\widehat{\pi}(\mathbf{X})}\{Y - \widehat{m}(\mathbf{X})\}\right]\mathbf{\Psi}(\mathbf{X}).$$

Next, recall from (3.1) the terms $\mathbf{T}_0(\mathbf{Z}), \mathbf{T}_\pi(\mathbf{Z}), \mathbf{T}_m(\mathbf{Z})$ and $\mathbf{R}_{\pi,m}(\mathbf{Z})$ defined in (3.2)-(3.5), with $g(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbf{\Psi}(\mathbf{X})'\boldsymbol{\theta}_0$ and $h(\mathbf{X}) = -2\mathbf{\Psi}(\mathbf{X})$ in this case, and let $\mathbf{T}^*_0(\mathbf{Z}), \mathbf{T}^*_\pi(\mathbf{Z}), \mathbf{T}^*_m(\mathbf{Z})$ and $\mathbf{R}^*_{\pi,m}(\mathbf{Z})$ respectively denote their versions with $h(\mathbf{X})$ replaced by $h^*(\mathbf{X}) = \mathbf{\Psi}(\mathbf{X})$. Then, we have: $\boldsymbol{\psi}_0(\mathbf{Z}) = \mathbf{T}^*_0(\mathbf{Z})$ and

$$\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}) \ = \ \mathbf{T}^*_0(\mathbf{Z}) + \mathbf{T}^*_\pi(\mathbf{Z}) - \mathbf{T}^*_m(\mathbf{Z}) - \mathbf{R}^*_{\pi,m}(\mathbf{Z}) - \mathbf{\Psi}(\mathbf{X})\mathbf{\Psi}(\mathbf{X})'(\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0).$$

Hence for all $1 \le i \le n$, $\widehat{\mathbf{\Gamma}}_0(\mathbf{Z}_i) - \mathbf{\Gamma}_0(\mathbf{Z}_i)$ satisfies:

$$\text{(J.11)} \quad \begin{aligned} \widehat{\mathbf{\Gamma}}_0(\mathbf{Z}_i) - \mathbf{\Gamma}_0(\mathbf{Z}_i) \ &\equiv \ \widehat{\mathbf{\Omega}}\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}_i) - \mathbf{\Omega}\boldsymbol{\psi}_0(\mathbf{Z}_i) \ = \ (\widehat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{T}^*_0(\mathbf{Z}_i) \\ &+ \widehat{\mathbf{\Omega}}\{\mathbf{T}^*_\pi(\mathbf{Z}_i) - \mathbf{T}^*_m(\mathbf{Z}_i) - \mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\} - \widehat{\mathbf{\Omega}}\mathbf{\Psi}(\mathbf{X}_i)\mathbf{\Psi}(\mathbf{X}_i)'(\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0). \end{aligned}$$

Under Assumption 3.1 (a) and 1.1 (b), similar to the proof of (J.8), we have using Lemma C.1 (v) and (i) (c): $\mathbf{T}^*_0(\mathbf{Z}) \equiv -\frac{1}{2}\mathbf{T}_0(\mathbf{Z})$ is sub-exponential with

$$\|\mathbf{T}^*_0(\mathbf{Z})\|_{\psi_1} \ \le \ \|\mathbf{\Psi}(\mathbf{X})\|_{\psi_2}\{\|\psi(\mathbf{X})\|_{\psi_2} + \|\varepsilon(\mathbb{Z})\|_{\psi_2}\delta_\pi^{-1}\} \ \le \ \sigma_{\mathbf{h}}(\sigma_\psi + \delta_\pi^{-1}\sigma_\varepsilon).$$

Hence, $\max_{1\le i\le n}\|\mathbf{T}^*_0(\mathbf{Z}_i)\|_\infty \ \equiv \ \max_{1\le i\le n, 1\le j\le d}|\mathbf{T}^*_{0[j]}(\mathbf{Z}_i)| \ = \ O_{\mathbb{P}}(\log(nd))$ due to Lemma C.1 (vi). Using this along with Assumption 4.1 (a), we have:

$$\text{(J.12)} \quad \max_{1\le i\le n}\|(\widehat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{T}^*_0(\mathbf{Z}_i)\|_\infty \ \le \ \|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 \max_i \|\mathbf{T}^*_0(\mathbf{Z}_i)\|_\infty \ = O_{\mathbb{P}}\left(r_n\log(nd)\right).$$

Now, since $\mathbf{\Psi}(\mathbf{X})$, $\varepsilon(\mathbb{Z})$ and $\mathbf{\Omega}\mathbf{\Psi}(\mathbf{X})$ are all sub-Gaussian due to Assumptions 3.1 (a) and 4.1 (b), using Lemma C.1 (vi), we have:

$$\text{(J.13)} \quad \max_{1\le i\le n}\{\|\mathbf{\Psi}(\mathbf{X}_i)\|_\infty + \|\mathbf{\Omega}\mathbf{\Psi}(\mathbf{X}_i)\|_\infty + |\varepsilon(\mathbb{Z}_i)|\} \ = \ O_{\mathbb{P}}\left(\sqrt{\log(nd)}\right).$$

Next, recalling the proof techniques and notations introduced in the proofs of Theorems 3.2, 3.3 and 3.4, as well as using Assumption 1.1 (b), we have:

$$\text{(J.14)} \quad \begin{aligned} \max_{1\le i\le n}\|\mathbf{T}^*_\pi(\mathbf{Z}_i)\|_\infty \ &\le \ \delta_\pi^{-1}\|\widetilde{\pi}_n\|_{\infty,n}\|\Delta_{\pi,n}\|_{\infty,n}\max_{1\le i\le n}\{\|\mathbf{\Psi}(\mathbf{X}_i)\|_\infty|\varepsilon(\mathbb{Z}_i)|\}, \\ \max_{1\le i\le n}\|\mathbf{T}^*_m(\mathbf{Z}_i)\|_\infty \ &\le \ (1 + \delta_\pi^{-1})\left\|\Delta^*_{m,n}\right\|_{\infty,n}\max_{1\le i\le n}\|\mathbf{\Psi}(\mathbf{X}_i)\|_\infty \ \text{ and} \\ \max_{1\le i\le n}\|\mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\|_\infty &\le \ \delta_\pi^{-1}\|\widetilde{\pi}_n\|_{\infty,n}\|\Delta_{\pi,n}\|_{\infty,n}\left\|\Delta^*_{m,n}\right\|_{\infty,n}\max_{1\le i\le n}\|\mathbf{\Psi}(\mathbf{X}_i)\|_\infty, \end{aligned}$$

where $\|\widetilde{\pi}_n\|_{\infty,n}$ and $\|\Delta_{\pi,n}\|_{\infty,n}$ are as in (G.1)-(G.2) and $\left\|\Delta^*_{m,n}\right\|_{\infty,n}$ is as defined in (H.2) and (I.2). Using (I.3), (I.4) and (I.5), we further have:

$$\|\widetilde{\pi}_n\|_{\infty,n}\,\|\Delta_{\pi,n}\|_{\infty,n} = O_{\mathbb{P}}(v_{n,\pi}\sqrt{\log n}) \;\text{ and }\; \left\|\Delta^*_{m,n}\right\|_{\infty,n} = O_{\mathbb{P}}(v_{\bar{n},m}\sqrt{\log n}). \tag{J.15}$$

Using (J.13) and (J.15) in (J.14), we then have:

$$\begin{aligned}\max_{1\le i\le n}\{\|\mathbf{T}^*_\pi(\mathbf{Z}_i)\|_\infty + \|\mathbf{T}^*_m(\mathbf{Z}_i)\|_\infty + \|\mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\|_\infty\} \;&=\; O_{\mathbb{P}}(\widetilde{v}_n),\\ \text{where }\; \widetilde{v}_n \;:=\; \{(v_{n,\pi}+v_{\bar{n},m})\sqrt{\log n} &+ v_{n,\pi}v_{\bar{n},m}(\log n)\}\log(nd).\end{aligned} \tag{J.16}$$

Using similar arguments as above, with $\boldsymbol{\Psi}(\mathbf{X})$ replaced by $\boldsymbol{\Omega}\boldsymbol{\Psi}(\mathbf{X})$ in (J.14) throughout, and using (J.13) and (J.15), we also have:

$$\max_{1\le i\le n}\{\|\boldsymbol{\Omega}\mathbf{T}^*_\pi(\mathbf{Z}_i)\|_\infty + \|\boldsymbol{\Omega}\mathbf{T}^*_m(\mathbf{Z}_i)\|_\infty + \|\boldsymbol{\Omega}\mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\|_\infty\} \;=\; O_{\mathbb{P}}(\widetilde{v}_n). \tag{J.17}$$

Combining (J.16) and (J.17) along with Assumption 4.1 (a), we have:

$$\begin{aligned}&\max_{1\le i\le n}\|\widehat{\boldsymbol{\Omega}}\{\mathbf{T}^*_\pi(\mathbf{Z}_i) - \mathbf{T}^*_m(\mathbf{Z}_i) - \mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\}\|_\infty\\ &\le \max_{1\le i\le n}\{\|\boldsymbol{\Omega}\mathbf{T}^*_\pi(\mathbf{Z}_i)\|_\infty + \|\boldsymbol{\Omega}\mathbf{T}^*_m(\mathbf{Z}_i)\|_\infty + \|\boldsymbol{\Omega}\mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\|_\infty\}\\ &\quad + \|\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\|_1 \max_{1\le i\le n}\{\|\mathbf{T}^*_\pi(\mathbf{Z}_i)\|_\infty + \|\mathbf{T}^*_m(\mathbf{Z}_i)\|_\infty + \|\mathbf{R}^*_{\pi,m}(\mathbf{Z}_i)\|_\infty\}\\ &= O_{\mathbb{P}}\left(\widetilde{v}_n + r_n\widetilde{v_n}\right) \;=\; O_{\mathbb{P}}\left(\widetilde{v}_n\right), \quad \text{since }\; r_n = o(1).\end{aligned} \tag{J.18}$$

Now turning to the third term in (J.11), under Assumption 4.1, and using (J.3) and (J.13) along with multiple uses of L_1-L_∞ type bounds, we have:

$$\begin{aligned}\|\widehat{\boldsymbol{\Omega}}\boldsymbol{\Psi}(\mathbf{X}_i)\boldsymbol{\Psi}(\mathbf{X}_i)'(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0)\|_\infty \;&\le\; \|\boldsymbol{\Omega}\boldsymbol{\Psi}(\mathbf{X}_i)\|_\infty\|\boldsymbol{\Psi}(\mathbf{X}_i)\|_\infty\|\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0\|_1\\ &+ \|\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\|_1\|\boldsymbol{\Psi}(\mathbf{X}_i)\|_\infty\|\boldsymbol{\Psi}(\mathbf{X}_i)\|_\infty\|\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0\|_1 \quad \forall\, 1\le i\le n, \quad \text{so that}\end{aligned}$$

$$\max_{1\le i\le n}\|\widehat{\boldsymbol{\Omega}}\boldsymbol{\Psi}(\mathbf{X}_i)\boldsymbol{\Psi}(\mathbf{X}_i)'(\widehat{\boldsymbol{\theta}}_{\text{DDR}} - \boldsymbol{\theta}_0)\|_\infty \;\le\; O_{\mathbb{P}}\left(s\sqrt{\frac{\log d}{n}}\log(nd)(1+r_n)\right). \tag{J.19}$$

Applying (J.12), (J.18) and (J.19) in (J.11) via triangle inequality, we get

$$\max_{1\le i\le n}\|\widehat{\boldsymbol{\Gamma}}_0(\mathbf{Z}_i) - \boldsymbol{\Gamma}_0(\mathbf{Z}_i)\|_\infty \;=\; O_{\mathbb{P}}\left(r_n\log(nd) + \widetilde{v}_n + s\sqrt{\frac{\log d}{n}}\log(nd)\right). \tag{J.20}$$

Finally, note that owing to (J.8), $\mathbf{\Gamma}_0(\mathbf{Z})$ is sub-exponential with $\|\mathbf{\Gamma}_0(\mathbf{Z})\|_{\psi_1} \leq \sigma_{\mathbf{\Gamma}_0} < \infty$. Hence, using Bernstein's Inequality (Lemma C.4), we have:

(J.21)
$$\max_{1\leq j\leq d}\left\{\frac{1}{n}\sum_{i=1}^{n}|\mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i)|\right\} \;\leq\; \max_{1\leq j\leq d}\mathbb{E}\{|\mathbf{\Gamma}_{0[j]}(\mathbf{Z})|\} + O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right),$$

which is $O_{\mathbb{P}}(1)$ since $\mathbb{E}\{|\mathbf{\Gamma}_{0[j]}(\mathbf{Z})|\} \leq \sigma_{\mathbf{\Gamma}_0}\ \forall\, j$ owing to Lemma C.1 (iv) (a).

Applying (J.20) and (J.21) to the first term in (J.10) via several uses of the triangle inequality and that $a^2 - b^2 = (a-b)(a+b)\ \forall\, a, b \in \mathbb{R}$, we have:

$$\begin{aligned}
\text{(J.22)}\quad & \max_{1\leq j\leq d}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{\Gamma}}^2_{0[j]}(\mathbf{Z}_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{\Gamma}^2_{0[j]}(\mathbf{Z}_i)\right| \\
&= \max_{1\leq j\leq d}\frac{1}{n}\sum_{i=1}^{n}|\widehat{\mathbf{\Gamma}}_{0[j]}(\mathbf{Z}_i) - \mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i)|\;|\widehat{\mathbf{\Gamma}}_{0[j]}(\mathbf{Z}_i) - \mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i) + 2\mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i)| \\
&\leq \max_{1\leq i\leq n}\|\widehat{\mathbf{\Gamma}}_0(\mathbf{Z}_i) - \mathbf{\Gamma}_0(\mathbf{Z}_i)\|_\infty\left[\max_{1\leq j\leq d}\left\{\frac{2}{n}\sum_{i=1}^{n}|\mathbf{\Gamma}_{0[j]}(\mathbf{Z}_i)|\right\} + o_{\mathbb{P}}(1)\right] \\
&= O_{\mathbb{P}}\left(r_n\log(nd) + \widetilde{v}_n + s\sqrt{\frac{\log d}{n}}\log(nd)\right).
\end{aligned}$$

Furthermore, since $\|\mathbf{\Gamma}_0(\mathbf{Z})\|_{\psi_1} \leq \sigma_{\mathbf{\Gamma}_0}$, we have: $\max_{1\leq j\leq d}\|\Gamma^2_{0[j]}(\mathbf{Z})\|_{\psi_\alpha} \leq \sigma^2_{\mathbf{\Gamma}_0}$ with $\alpha = \frac{1}{2}$ owing to Lemma C.1 (v). Hence, using Lemma C.6, we get

(J.23)
$$\max_{1\leq j\leq d}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Gamma}^2_{0[j]}(\mathbf{Z}_i) - \mathbb{E}\{\mathbf{\Gamma}^2_{0[j]}(\mathbf{Z})\}\right| \;\leq\; O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{n}} + \frac{(\log n)^2(\log d)^2}{n}\right).$$

Hence, combining (J.22) and (J.23) via a triangle inequality and applying them in (J.10), and recalling $\widetilde{v}_n$ from (J.16), we finally have:

$$\begin{aligned}
\text{(J.24)}\qquad & \max\nolimits_{1\leq j\leq d}|\widehat{\sigma}^2_{0,j} - \sigma^2_{0,j}| \;=\; O_{\mathbb{P}}(\tau_n) \;=\; o_{\mathbb{P}}(1), \quad \text{where} \\
& \tau_n \;:=\; r_n\log(nd) + \widetilde{v}_n + s\sqrt{\tfrac{\log d}{n}}\log(nd) + \sqrt{\tfrac{\log d}{n}} + \tfrac{(\log n)^2(\log d)^2}{n}.
\end{aligned}$$

Note that we have implcitly assumed τ_n to be $o(1)$ here. A careful analysis will reveal that this entails essentially the same rate conditions as those needed for the ALE (4.3) in Theorem 4.1 to hold, upto an additional factor of $\sqrt{\log(nd)}$ appearing in the first three terms of τ_n, as well as the presence of the last term in τ_n (which is expected to be of lower order than the rest).

(J.24) therefore establishes the desired (uniform) consistency of the standard error estimators $\{\widehat{\sigma}_{0,j}\}_{j=1}^{d}$, and also establishes the second asymptotic normality result in Theorem 4.1 through use of (J.9), (J.24) and Slutsky's Theorem, as discussed earlier. This completes the proof of Theorem 4.1. ∎

APPENDIX K: PROOFS OF ALL RESULTS IN SECTION 5

We present here the proofs of Theorems 5.1-5.3, as well as the assumptions required for Theorems 5.2 and 5.3. We begin with the proof of Theorem 5.1.

K.1. Proof of Theorem 5.1. Under the assumed conditions, we have:

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}} |g\{\widehat{\boldsymbol{\beta}}'\boldsymbol{\Psi}(\mathbf{x})\} - g\{\boldsymbol{\beta}_0'\boldsymbol{\Psi}(\mathbf{x})\}| \;&\leq\; C_g \sup_{\mathbf{x}\in\mathcal{X}} |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\boldsymbol{\Psi}(\mathbf{x})| \\
&\leq\; C_g \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \sup_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\Psi}(\mathbf{X})\|_\infty \;\leq\; C_g C_{\boldsymbol{\Psi}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1.
\end{aligned} \tag{K.1}$$

where the steps follow from the Lipschitz continuity of $g(\cdot)$ and the boundedness of $\boldsymbol{\Psi}(\cdot)$ along with an L_1-L_∞ bound. Now, under the L_1 error bound assumed for $\widehat{\boldsymbol{\beta}}$ and using a simple union bound argument, we have: $\forall\, \epsilon \geq 0$,

$$\begin{aligned}
&\mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > \epsilon) \\
&= \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > \epsilon, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq a_n) + \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > \epsilon, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > a_n) \\
&\leq \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > \epsilon, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq a_n) + \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > a_n) \\
&\leq \mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 > \epsilon \mid \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq a_n)\mathbb{P}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq a_n) + q_n \\
&\leq 2\exp\{-\epsilon^2/(2a_n^2)\}(1 - q_n) + q_n \;\leq\; 2\exp\{-\epsilon^2/(2a_n^2)\} + q_n,
\end{aligned}$$

where the final bounds follow from an application of Hoeffding's inequality for bounded random variables (or using Lemma C.1 (ii)(d) and (iii)(a)). Using this bound along with that in (K.1), we then have: for any $\epsilon \geq 0$,

$$\mathbb{P}[\sup_{\mathbf{x}\in\mathcal{X}} |g\{\widehat{\boldsymbol{\beta}}'\boldsymbol{\Psi}(\mathbf{x})\} - g\{\boldsymbol{\beta}_0'\boldsymbol{\Psi}(\mathbf{x})\}| > C_g C_{\boldsymbol{\Psi}}\epsilon] \;\leq\; 2\exp\{-\epsilon^2/(2a_n^2)\} + q_n.$$

The desired result then follows by setting $\epsilon = \sqrt{2}a_n t$ for any $t \geq 0$. ∎

K.2. Assumptions for Theorems 5.2 and 5.3. We summarize here the smoothness and regularity assumptions required for Theorems 5.2-5.3.

ASSUMPTION K.1 (Standard smoothness assumptions and conditions on $K(\cdot)$ and the tail behavior of Z). We assume the following conditions.

(a) Z is sub-Gaussian with $\|Z\|_{\psi_2} \leq \sigma_Z$ for some constant $\sigma_Z \geq 0$.

(b) $K(\cdot)$ is bounded and integrable with $\|K(\cdot)\|_\infty \leq M_K$ and $\int_{\mathbb{R}} |K(u)|du \leq C_K$ for some constants $M_K, C_K \geq 0$.

(c) Let $m^{(2)}_{\boldsymbol{\beta}}(w) := \mathbb{E}\{Z^2 \mid \boldsymbol{\beta}'\mathbf{X} = w\}$ for any $w \in \mathbb{R}$. Then, $m^{(2)}_{\boldsymbol{\beta}}(w) f_{\boldsymbol{\beta}}(w)$ is bounded in $w \in \mathbb{R}$ and $\|m^{(2)}_{\boldsymbol{\beta}}(\cdot) f_{\boldsymbol{\beta}}(\cdot)\|_\infty \leq B_1$ for some constant $B_1 \geq 0$.

(d) $K(\cdot)$ is a second order kernel satisfying: $\int_{\mathbb{R}} K(u)d(u) = 1$, $\int_R uK(u)du = 0$ and $\int_{\mathbb{R}} u^2|K(u)|du \leq R_K < \infty$ for some constant $R_K \geq 0$. $l_{\boldsymbol{\beta}}(\cdot) \equiv m_{\boldsymbol{\beta}}(\cdot) f_{\boldsymbol{\beta}}(\cdot)$ is twice continuously differentiable with bounded second derivatives $l''_{\boldsymbol{\beta}}(\cdot)$ satisfying: $\|l''_{\boldsymbol{\beta}}(\cdot)\|_\infty \leq B_2$ for some constant $B_2 \geq 0$.

ASSUMPTION K.2 (Further conditions on $K(\cdot)$ and other assumptions to account for the estimation error of $\boldsymbol{\beta}$). We also assume the following.

(a) $K(\cdot)$ is continuously differentiable with a bounded and integrable derivative $K'(\cdot)$ satisfying $\|K'(\cdot)\|_\infty \leq M_{K'}$ and $\int_{\mathbb{R}} |K'(u)|du \leq C_{K'}$ for some constants $M_{K'}, C_{K'} \geq 0$. Further, $K(u) \to 0$ as $u \to \infty$ or $u \to -\infty$.

(b) Let $\boldsymbol{\eta}_{\boldsymbol{\beta}}(w) := \mathbb{E}(Z\mathbf{X} \mid \boldsymbol{\beta}'\mathbf{X} = w) f_{\boldsymbol{\beta}}(w)$ for any $w \in \mathbb{R}$, and let $\boldsymbol{\eta}_{\boldsymbol{\beta}[j]}(\cdot)$ denote the j^{th} coordinate of $\boldsymbol{\eta}_{\boldsymbol{\beta}}(\cdot)$ for $j = 1, \ldots, d$. Then, for each j, $\boldsymbol{\eta}_{\boldsymbol{\beta}[j]}(\cdot)$ is continuously differentiable with derivative $\boldsymbol{\eta}'_{\boldsymbol{\beta}[j]}(\cdot)$ that is bounded uniformly in $j = 1, \ldots, d$. Further, $l_{\boldsymbol{\beta}}(\cdot)$ is also continuously differentiable with a bounded derivative $l'_{\boldsymbol{\beta}}(\cdot)$. Thus, $\max_{1 \leq j \leq d} \|\boldsymbol{\eta}'_{\boldsymbol{\beta}[j]}(\cdot)\|_\infty \leq B_1^*$ and $\|l'_{\boldsymbol{\beta}}(\cdot)\|_\infty \leq B_2^*$ for some constants $B_1^*, B_2^* \geq 0$.

(c) $K'(\cdot)$ satisfies a 'local' Lipschitz property as follows. There exists a constant $L > 0$ such that for all $u, v \in \mathbb{R}$ with $|u-v| \leq L$, $|K'(u) - K'(v)| \leq \varphi(u)|u-v|$ for some bounded and integrable function $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}^+$ with $\|\varphi(\cdot)\|_\infty \leq M_\varphi$ and $\int_{\mathbb{R}} \varphi(u)du \leq C_\varphi$ for some constants $M_\varphi, C_\varphi \geq 0$.

(d) $\mathbf{X}$ is bounded, i.e. $\|\mathbf{X}\|_\infty \leq M_{\mathbf{X}}$ a.s. $[\mathbb{P}]$ for some constant $M_{\mathbf{X}} \geq 0$, and $\widehat{\boldsymbol{\beta}}$ satisfies the high-level guarantee (5.1). Further, we assume $a_n/h = o(1)$ and $2M_{\mathbf{X}}(a_n/h) \leq L$, where L is as in (c) above and a_n is as in (5.1).

Most of the smoothness assumptions and the conditions on $K(\cdot)$ in Assumptions K.1 and K.2 are fairly mild and standard in the non-parametric statistics literature. Similar or equivalent versions of these assumptions can be found in a variety of references including Newey and McFadden (1994); Andrews (1995); Masry (1996) and Hansen (2008), among others.

Assumption K.2 (c) imposes a 'local' Lipschitz property of sorts on $K'(\cdot)$, where the Lipschitz 'constant' is a bounded function that also decays quickly enough to be integrable. This is satisfied by the Gaussian kernel in particular. In general, it holds for any $K(\cdot)$ where $K'(\cdot)$ has a compact support and is Lipschitz continuous, or $K'(\cdot)$ is differentiable with a bounded derivative

$K''(\cdot)$ that has a polynomially integrable tail, i.e. $|K''(u)| \le |u|^{-\rho}$ for some $\rho > 1$ and all $u \in \mathbb{R}$ such that $|u| > L^*$ for some $L^* > 0$ (see Hansen (2008)).

Finally, the boundedness assumption on $\mathbf{X}$ is mostly for the simplicity of our exposition. With appropriate modifications in the proofs, this can be relaxed to allow for more general tail behaviors of $\mathbf{X}$ (e.g. $\mathbf{X}$ is sub-Gaussian), although the corresponding technical analyses can be more involved.

K.3. Proof Sketch for Theorems 5.2 and 5.3. We first introduce two key supporting lemmas regarding tail bounds for $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$ both of which will be useful for proving Theorems 5.2 and 5.3. We begin with a few notations and a sketch of our analysis to set up and prove these lemmas, and subsequently, use them to complete the proofs of the main theorems.

To analyze the behavior of $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$, we first introduce the corresponding *hypothetical* version of the estimator where the index parameter $\boldsymbol{\beta}$ is treated as known. Specifically, for any $\mathbf{x} \in \mathcal{X}$, let us define the 'oracle' 'estimator':

$$\widetilde{l}(\boldsymbol{\beta}, \mathbf{x}) \ := \ \frac{1}{nh} \sum_{i=1}^{n} Z_i K\left(\frac{\boldsymbol{\beta}'\mathbf{X}_i - \boldsymbol{\beta}'\mathbf{x}}{h}\right) \ \equiv \ \frac{1}{nh} \sum_{i=1}^{n} Z_i K\left(\frac{W_i - w_{\mathbf{x}}}{h}\right).$$

Then, we note that the error $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})$ of the original estimator $\widehat{l}(\cdot)$ admits the following decomposition. For any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| \ &\le \ |\widetilde{l}(\boldsymbol{\beta}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| + |\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - \widetilde{l}(\boldsymbol{\beta}, \mathbf{x})| \\
&\le \ |\widetilde{l}(\boldsymbol{\beta}, \mathbf{x}) - \mathbb{E}\{\widetilde{l}(\boldsymbol{\beta}, \mathbf{x})\}| + |\mathbb{E}\{\widetilde{l}(\boldsymbol{\beta}, \mathbf{x})\} - l(\boldsymbol{\beta}, \mathbf{x})| + |\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - \widetilde{l}(\boldsymbol{\beta}, \mathbf{x})| \\
&=: \ |\widetilde{S}_n(\mathbf{x})| + |\overline{S}_n(\mathbf{x})| + |\widehat{R}_n(\mathbf{x})| \ \ \text{(say)}.
\end{aligned}$$

Thus, to analyze the behavior of $|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})|$, it suffices to control each of the quantities $\widetilde{S}_n(\mathbf{x})$, $\overline{S}_n(\mathbf{x})$ and $\widehat{R}_n(\mathbf{x})$. We now proceed towards obtaining non-asymptotic pointwise tail bounds for these quantities. We first focus on $\widetilde{S}_n(\mathbf{x})$ and $\overline{S}_n(\mathbf{x})$ which involve only the hypothetical estimator $\widetilde{l}(\cdot)$.

LEMMA K.1 (Characterizing the tail bounds for $\widetilde{S}_n(\mathbf{x})$ and $\overline{S}_n(\mathbf{x})$). *Under Assumption K.1 (a)-(c), we have: for any fixed $\mathbf{x} \in \mathcal{X}$ and any $t \ge 0$,*

$$\mathbb{P}\left\{|\widetilde{S}_n(\mathbf{x})| \ > \ C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2 \sqrt{\log n}}{nh}\right\} \ \le \ 3\exp(-t^2),$$

where $C_1 := 7(B_1 C_K M_K)^{1/2}$ and $C_2 := D\sigma_Z M_K$ for some absolute constant $D > 0$. Further, under Assumption K.1 (d), we have:

$$|\overline{S}_n(\mathbf{x})| \ \le \ C_3 h^2 \quad \textit{uniformly in } \ \mathbf{x} \in \mathcal{X}, \quad \textit{where } \ C_3 \ := \ B_2 R_K.$$

Hence, for any $\mathbf{x} \in \mathcal{X}$ and $t \geq 0$, with probability at least $1 - 3\exp(-t^2)$,

$$\text{(K.2)} \quad |\widetilde{l}(\boldsymbol{\beta}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| \;\leq\; C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2 \sqrt{\log n}}{nh} + C_3 h^2, \quad \forall\, \mathbf{x} \in \mathcal{X}. \quad \blacksquare$$

Next, we aim to control the term $\widehat{R}_n(\mathbf{x})$ whose behavior signifies the nature and extent of the additional price one pays due to estimation of $\boldsymbol{\beta}$.

Using a first order Taylor series expansion of $\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x})$ around $\widehat{l}(\boldsymbol{\beta}, \mathbf{x}) \equiv \widetilde{l}(\boldsymbol{\beta}, \mathbf{x})$, we first rewrite $\widehat{R}_n(\mathbf{x}) \equiv \widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - \widetilde{l}(\boldsymbol{\beta}, \mathbf{x})$ as:

$$\widehat{R}_n(\mathbf{x}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \left\{ \frac{1}{nh} \sum_{i=1}^{n} Z_i \frac{(\mathbf{X}_i - \mathbf{x})}{h} K' \left(\frac{W_i^* - w_{\mathbf{x}}^*}{h} \right) \right\}, \quad \text{where}$$

$\{W_i^*\}_{i=1}^n$ and $w_{\mathbf{x}}^*$ are 'intermediate' points that satisfy, for each $i = 1, \ldots, n$, $|(W_i^* - w_{\mathbf{x}}^*) - (W_i - w_{\mathbf{x}})| \leq |(\widehat{W}_i - \widehat{w}_{\mathbf{x}}) - (W_i - w_{\mathbf{x}})| \equiv |(\widehat{\boldsymbol{\beta}} - \beta)'(\mathbf{X}_i - \mathbf{x})|$.

We now rewrite the expansion above as: $\widehat{R}_n(\mathbf{x}) \equiv \widehat{R}_{n,1}(\mathbf{x}) + \widehat{R}_{n,2}(\mathbf{x})$, where

$$\begin{aligned} \widehat{R}_{n,1}(\mathbf{x}) \;&:=\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \left\{ \frac{1}{nh} \sum_{i=1}^{n} Z_i \frac{(\mathbf{X}_i - \mathbf{x})}{h} K' \left(\frac{W_i - w_{\mathbf{x}}}{h} \right) \right\} \\ &=:\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \widehat{\mathbf{T}}_n(\mathbf{x}) \;\; \text{(say)}, \quad \text{and} \;\; \widehat{R}_{n,2}(\mathbf{x}) \;:=\; \widehat{R}_n(\mathbf{x}) - \widehat{R}_{n,1}(\mathbf{x}). \end{aligned}$$

In the result below, we now characterize the tail bounds for $\widehat{R}_n(\mathbf{x})$.

Lemma K.2 (Characterizing the tail bounds for $\widehat{R}_{n,1}(\mathbf{x})$ and $\widehat{R}_{n,2}(\mathbf{x})$). *Under Assumption K.2 (a), (b) and (d), and Assumption K.1 (a) and (c), we have: for any $t \geq 0$, with probability at least $1 - 3\exp(-t^2) - q_n$,*

$$|\widehat{R}_{n,1}(\mathbf{x})| \;\leq\; C_1^* a_n + C_2^* \frac{a_n (t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n (t^2 + \log p) \sqrt{\log n}}{nh^2}, \quad \textit{where}$$

$C_1^*, C_2^*, C_3^* > 0$ are constants depending only on the constants introduced in Assumptions K.2 and K.1, and $\mathbf{x} \in \mathcal{X}$ is any fixed evaluation point.

Further, under the additional condition in Assumption K.2 (c), we have: for any $t \geq 0$, with probability at least $1 - 3\exp(-t^2) - q_n$,

$$\begin{aligned} |\widehat{R}_{n,2}(\mathbf{x})| \;&\leq\; 4M_{\mathbf{X}}^2 C_4^* \frac{a_n^2}{h^2} + 4M_{\mathbf{X}}^2 \left(C_5^* \frac{t a_n^2}{\sqrt{nh^5}} + C_6^* \frac{t^2 a_n^2 \sqrt{\log n}}{nh^3} \right), \quad \textit{where} \\ &\leq\; 3\exp(-t^2) + q_n, \quad \textit{for any fixed } \mathbf{x} \in \mathcal{X} \textit{ and any given } t \geq 0, \quad \textit{where} \end{aligned}$$

$C_4^*, C_5^*, C_6^* > 0$ are constants depending only on the constants introduced in Assumptions K.1 and K.2, and $\mathbf{x} \in \mathcal{X}$ is any fixed evaluation point.

With $a_n/h = o(1)$ as assumed, note that the second and the third terms in the bound for $\widehat{R}_{n,2}(\mathbf{x})$ are each dominated by the respective terms in the bound for $\widehat{R}_{n,1}(\mathbf{x})$ in Lemma K.2. Using this, we obtain a bound for $\widehat{R}_n(\mathbf{x})$ as follows: for any $t \geq 0$, with probability at least $1 - 6\exp(-t^2) - 2q_n$,

$$
\begin{aligned}
|\widehat{R}_n(\mathbf{x})| \; &\equiv \; |\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - \widetilde{l}(\boldsymbol{\beta}, \mathbf{x})| \\
&\leq \; C_1^*(a_n + a_n^2 h^{-2}) + C_2^* \frac{a_n(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n(t^2 + \log p)\sqrt{\log n}}{nh^2}, \qquad \text{(K.3)}
\end{aligned}
$$

for some constants $C_1^*, C_2^*, C_3^* > 0$ (possibly different from those in Lemma K.2) depending only on the constants defined in Assumptions K.1 and K.2.

■

K.4. Completing the Proof of Theorem 5.2. Combining the bounds (K.2) and (K.3) via a union bound, we then have: for any $\mathbf{x} \in \mathcal{X}$ and for any $t \geq 0$, with probability at least $1 - 9\exp(-t^2) - 2q_n$,

$$
\begin{aligned}
|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| \; &\leq \; |\widetilde{l}(\boldsymbol{\beta}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| + |\widehat{R}_n(\mathbf{x})| \; \leq \; C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2\sqrt{\log n}}{nh} \\
&\quad + C_3 h^2 + C_1^*(a_n + a_n^2 h^{-2}) + C_2^* \frac{a_n(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n(t^2 + \log p)\sqrt{\log n}}{nh^2} \\
\text{(K.4)} \quad &\equiv \; D_1 \frac{t}{\sqrt{nh}} \left(1 + \frac{a_n}{h}\right) + D_2 \frac{t^2\sqrt{\log n}}{nh} \left(1 + \frac{a_n}{h}\right) + D_3 b_n, \quad \text{where} \\
r_n \; &:= \; h^2 + a_n + \frac{a_n^2}{h^2} + \frac{a_n}{h}\sqrt{\frac{\log p}{nh}} + \frac{a_n}{h}\frac{\sqrt{\log n}\log p}{nh} \; = \; o(1) \quad \text{and}
\end{aligned}
$$

$D_1, D_2, D_3 > 0$ are some constants depending on the constants $\{C_j, C_j^*\}_{j=1}^3$.

Further, with $(a_n\sqrt{\log p})/h = o(1)$ and $\{\log(np)\}/(nh) = o(1)$ by assumption, the fourth term in the definition of r_n in (K.4) can be bounded as: $(a_n/h)\{\sqrt{\log p}/(nh)\} = o(1/\sqrt{nh})$ and the fifth term can be bounded as:

$$
\frac{a_n}{h}\frac{\sqrt{\log n}\log p}{nh} \; \leq \; \frac{a_n\sqrt{\log p}}{h}\frac{\log(np)}{nh} = o\left(\frac{\log(np)}{nh}\right),
$$

where we used that $\sqrt{\log n}\sqrt{\log p} \leq (\log n + \log p)/2 \leq \log(np)$. Using these simplifications in (K.4) and that $a_n/h = o(1)$ by assumption, we finally have: for any $\mathbf{x} \in \mathcal{X}$ and for any $t \geq 0$, with probability at least $1 - 6\exp(-t^2) - 2q_n$,

$$
\begin{aligned}
|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| \; &\leq \; D_1^* \frac{t}{\sqrt{nh}} + D_2^* \frac{t^2\sqrt{\log n}}{nh} + D_3^* b_n, \quad \text{where} \\
b_n \; &:= \; h^2 + a_n + \frac{a_n^2}{h^2} + \frac{1}{\sqrt{nh}} + \frac{\log(np)}{nh} \quad \text{and}
\end{aligned}
$$

$D_1^*, D_2^*, D_3^* > 0$ are some constants depending only on those introduced in the assumptions. This completes the proof of Theorem 5.2. ∎

K.5. Completing the Proof of Theorem 5.3. Using Theorem 5.2, we have: for any fixed $\mathbf{x} \in \mathcal{X}$ and for any $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\left\{|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| > \epsilon_n(t)\right\} \leq 9\exp(-t^2) + 2q_n \text{ and} \\ &\mathbb{P}\left\{|\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - f(\boldsymbol{\beta}, \mathbf{x})| > \epsilon_n(t)\right\} \leq 9\exp(-t^2) + 2q_n, \end{aligned} \tag{K.5}$$

where we recall that $\{\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x}), f(\boldsymbol{\beta}, \mathbf{x})\}$ is a special case of $\{\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}), l(\beta, \mathbf{x})\}$ with $Z \equiv 1$ so that Theorem K.2 indeed applies to get both bounds above.

Next, note that $\widehat{m}(\cdot) \equiv \widehat{l}(\cdot)/\widehat{f}(\cdot)$ and $m(\cdot) \equiv l(\cdot)/f(\cdot)$, so that

$$\begin{aligned} &|\widehat{f}(\cdot)\{\widehat{m}(\cdot) - m(\cdot)\}| = |\{\widehat{l}(\cdot) - l(\cdot)\} - m(\cdot)\{\widehat{f}(\cdot) - f(\cdot)\}| \\ &\quad \leq |\widehat{l}(\cdot) - l(\cdot)| + |m(\cdot)||\widehat{f}(\cdot) - f(\cdot)| \leq |\widehat{l}(\cdot) - l(\cdot)| + \delta_m|\widehat{f}(\cdot) - f(\cdot)|, \end{aligned}$$

where in the last step, we used $\|m(\cdot)\|_\infty \leq \delta_m$ by assumption. Using a simple union bound argument, we then have: for any $\mathbf{x} \in \mathcal{X}$ and for any $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\left\{|\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x})\{\widehat{m}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - m(\widehat{\boldsymbol{\beta}}, \mathbf{x})\}| > (1+\delta_m)\epsilon_n(t)\right\} \\ &\quad \leq \mathbb{P}\left\{|\widehat{l}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - l(\boldsymbol{\beta}, \mathbf{x})| > \epsilon_n(t)\right\} + \mathbb{P}\left\{|\widehat{f}(\widehat{\boldsymbol{\beta}}, \mathbf{x}) - f(\boldsymbol{\beta}, \mathbf{x})| > \epsilon_n(t)\right\} \\ &\quad \leq 18\exp(-t^2) + 4q_n, \end{aligned} \tag{K.6}$$

where the final step follows from using the bounds in (K.5).

Recall further that by assumption, $|f(\boldsymbol{\beta}, \mathbf{x})| \equiv f(\boldsymbol{\beta}, \mathbf{x}) \geq \delta_f > 0 \ \forall\, \mathbf{x} \in \mathcal{X}$. Then, for any $\mathbf{x} \in \mathcal{X}$ and any $t_* \geq 0$ such that $\delta_f - \epsilon_n(t_*) > 0$, we have:

$$\begin{aligned} &\mathbb{P}\{|\widehat{f}(\boldsymbol{\beta}, \mathbf{x})| < \delta_f - \epsilon_n(t_*)\} \leq \mathbb{P}\{|\widehat{f}(\boldsymbol{\beta}, \mathbf{x})| < |f(\boldsymbol{\beta}, \mathbf{x})| - \epsilon_n(t_*)\} \\ &\quad \leq \mathbb{P}\{|\widehat{f}(\boldsymbol{\beta}, \mathbf{x}) - f(\boldsymbol{\beta}, \mathbf{x})|| > \epsilon_n(t_*)\} \leq 9\exp(-t_*^2) + 2q_n, \end{aligned} \tag{K.7}$$

where the penultimate bound follows since $|b| - |a| \leq ||a| - |b|| \leq |a - b|$ for any $a, b \in \mathbb{R}$, and the final bound follows from (K.5). In particular, we have:

$$\mathbb{P}\left\{|\widehat{f}(\boldsymbol{\beta}, \mathbf{x})| < \frac{\delta_f}{2}\right\} \leq 9\exp(-t_*^2) + 2q_n, \ \ \forall\, t_* \geq 0 \text{ such that } \epsilon_n(t_*) \leq \frac{\delta_f}{2}.$$

Combining this bound along with (K.6), we now have: for any $\mathbf{x} \in \mathcal{X}$ and for any $t, t_* \geq 0$ with $\epsilon_n(t_*) \leq \delta_f/2$,

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{m}(\widehat{\boldsymbol{\beta}},\mathbf{x}) - m(\boldsymbol{\beta},\mathbf{x})| > \frac{2(1+\delta_m)}{\delta_f}\epsilon_n(t)\right\} \\
&= \mathbb{P}\left\{|\widehat{m}(\widehat{\boldsymbol{\beta}},\mathbf{x}) - m(\boldsymbol{\beta},\mathbf{x})| > \frac{2(1+\delta_m)}{\delta_f}\epsilon_n(t), |\widehat{f}(\widehat{\boldsymbol{\beta}},\mathbf{x})| \geq \frac{\delta_f}{2}\right\} \\
&\quad + \mathbb{P}\left\{|\widehat{m}(\widehat{\boldsymbol{\beta}},\mathbf{x}) - m(\boldsymbol{\beta},\mathbf{x})| > \frac{2(1+\delta_m)}{\delta_f}\epsilon_n(t), |\widehat{f}(\widehat{\boldsymbol{\beta}},\mathbf{x})| < \frac{\delta_f}{2}\right\} \\
&\leq \mathbb{P}\left\{|\widehat{f}(\widehat{\boldsymbol{\beta}},\mathbf{x})||\widehat{m}(\widehat{\boldsymbol{\beta}},\mathbf{x}) - m(\widehat{\boldsymbol{\beta}},\mathbf{x})| > (1+\delta_m)\epsilon_n(t)\right\} + \mathbb{P}\left\{|\widehat{f}(\widehat{\boldsymbol{\beta}},\mathbf{x})| < \frac{\delta_f}{2}\right\} \\
&\leq 18\exp(-t^2) + 9\exp(-t_*^2) + 6q_n,
\end{aligned}$$

where the final bound follows from using (K.6), (K.7) and the bound noted below (K.7) as a special case. This completes the proof of Theorem 5.3. ∎

K.6. Proof of Lemma K.1. Let $\mathbf{Z} := (Z, \mathbf{X})$ and rewrite $\widetilde{l}(\boldsymbol{\beta},\mathbf{x})$ as:

$$\widetilde{l}(\boldsymbol{\beta},\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} T_h(\mathbf{Z}_i;\mathbf{x},\boldsymbol{\beta}), \text{ where } T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta}) := \frac{1}{h} Z K\left(\frac{W_i - w_{\mathbf{x}}}{h}\right).$$

Under Assumption K.1 (a)-(b) and using Lemma C.1 (i)(b), (ii)(d) and (v), $T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})$ is sub-Gaussian with $\|T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\|_{\psi_2} \leq h^{-1}\sigma_Z M_K$. Hence, using Lemma C.1 (iv)(b) and (i)(c), we have:

$$\|T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta}) - \mathbb{E}\{T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\}\|_{\psi_2} \leq 3h^{-1}\sigma_Z M_K \quad \text{uniformly for all } \mathbf{x} \in \mathcal{X}.$$

Further, under Assumption K.1 (b)-(c), we have: uniformly for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\operatorname{Var}\{T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\} &\leq \mathbb{E}\{T_h^2(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\} = \mathbb{E}_W[\mathbb{E}\{T_h^2(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})|W\}] \\
&= h^{-2}\int_{\mathbb{R}} \mathbb{E}(Z^2|W=w) K^2\{(w-w_{\mathbf{x}})/h\} f_{\boldsymbol{\beta}}(w)dw \\
&\equiv h^{-2}\int_{\mathbb{R}} m_{\boldsymbol{\beta}}^{(2)}(w) K^2\{(w-w_{\mathbf{x}})/h\} f_{\boldsymbol{\beta}}(w)dw \\
&= h^{-1}\int_{\mathbb{R}} m_{\boldsymbol{\beta}}^{(2)}(w_{\mathbf{x}}+hu) f_{\boldsymbol{\beta}}(w_{\mathbf{x}}+hu) K^2(u)du \leq h^{-1}B_1 M_K C_K,
\end{aligned}$$

where the penultimate step follows from a standard change of variable argument. We have thus verified all the conditions required for Lemma C.6 using which we now obtain: for any $t \geq 0$, with probability at least $1 - 3\exp(-t^2)$,

$$\widetilde{S}_n(\mathbf{x}) \equiv |\widetilde{l}(\boldsymbol{\beta},\mathbf{x}) - \mathbb{E}\{\widetilde{l}(\boldsymbol{\beta},\mathbf{x})\}| \leq 7t\sqrt{\frac{B_1 M_K C_K}{nh}} + t^2\frac{D\sigma_Z M_K}{nh}\sqrt{\log n},$$

where while using Lemma C.6, we set $\Gamma_n = h^{-1}B_1M_KC_K$, $K_n = h^{-1}\sigma_Z M_K$, $p = 1$, $\alpha = 2$, and D depends on the absolute constant C_α in the statement of the lemma. This completes the proof of the first part of Lemma K.1. ∎

For the second part regarding $\overline{S}_n(\mathbf{x}) \equiv \mathbb{E}\{\widetilde{l}(\boldsymbol{\beta},\mathbf{x})\} - l(\boldsymbol{\beta},\mathbf{x})$, observe that $\mathbb{E}\{\widetilde{l}(\boldsymbol{\beta},\mathbf{x})\} = \mathbb{E}\{T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\}$ and $l(\boldsymbol{\beta},\mathbf{x}) \equiv l_{\boldsymbol{\beta}}(w_{\mathbf{x}})$. We the have: $\forall\ \mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\overline{S}_n(\mathbf{x}) \ &= \ \mathbb{E}\{T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})\} - l(\boldsymbol{\beta},\mathbf{x}) \ = \ \mathbb{E}_W[\mathbb{E}\{T_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta})|W\}] - l_{\boldsymbol{\beta}}(w_{\mathbf{x}}) \\
&= \ h^{-1}\int_{\mathbb{R}} \mathbb{E}(Z|W=w)K\{(w-w_{\mathbf{x}})/h\}f_{\boldsymbol{\beta}}(w)dw - l_{\boldsymbol{\beta}}(w_{\mathbf{x}}) \\
&= \ h^{-1}\int_{\mathbb{R}} l_{\boldsymbol{\beta}}(w)K\{(w-w_{\mathbf{x}})/h\}dw - l_{\boldsymbol{\beta}}(w_{\mathbf{x}}) \\
&= \ \int_{\mathbb{R}} l_{\boldsymbol{\beta}}(w_{\mathbf{x}}+hu)K(u)du - l_{\boldsymbol{\beta}}(w_{\mathbf{x}}) \ = \ \int_{\mathbb{R}}\{l_{\boldsymbol{\beta}}(w_{\mathbf{x}}+hu) - l_{\boldsymbol{\beta}}(w_{\mathbf{x}})\}K(u)du \\
&= \ hl'_{\boldsymbol{\beta}}(w_{\mathbf{x}})\underbrace{\int_{\mathbb{R}} uK(u)du}_{=0} + h^2R^*(\mathbf{x}) \ := \ h^2\int_{\mathbb{R}} l''_{\boldsymbol{\beta}}(w^*_{\mathbf{x},u})u^2K(u)du, \ \text{ where}
\end{aligned}$$

$w^*_{\mathbf{x},u}$ is some 'intermediate' point satisfying $|w_{\mathbf{x},u} - w_{\mathbf{x}}| \leq h|u|$. The first two steps use $\mathbb{E}(Z|W=w) \equiv m_{\boldsymbol{\beta}}(w)$ and $m_{\boldsymbol{\beta}}(w)f_{\boldsymbol{\beta}}(w) \equiv l_{\boldsymbol{\beta}}(w)$. The next steps follow from a standard change of variable and Taylor series expansion argument under the assumed smoothness of $l_{\boldsymbol{\beta}}(\cdot)$ in Assumption K.1 (d) along with the conditions imposed therein on the kernel $K(\cdot)$. Using Assumption K.1 (d), we further have: $\|l''_{\boldsymbol{\beta}}(\cdot)\|_\infty \leq B_2$ and $\int |u^2K(u)|du \leq R_K$. Hence,

$$|\overline{S}_n(\mathbf{x})| \ \leq \ B_2\int_{\mathbb{R}} u^2|K(u)|du \ \leq \ B_2R_K \ \text{ uniformly for all } \ \mathbf{x} \in \mathcal{X}.$$

This establishes the second part of Lemma K.1 and completes the proof. ∎

K.7. Proof of Lemma K.2. To control $\widehat{R}_{n,1}(\mathbf{x}) \equiv (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\widehat{\mathbf{T}}_n(\mathbf{x})$, note

$$|\widehat{R}_{n,1}(\mathbf{x})| \ \leq \ \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1\left[\|\widehat{\mathbf{T}}_n(\mathbf{x}) - \mathbb{E}\{\widehat{\mathbf{T}}_n(\mathbf{x})\}\|_\infty + \|\mathbb{E}\{\widehat{\mathbf{T}}_n(\mathbf{x})\}\|_\infty\right] \tag{K.8}$$

In the light of (K.8) and the assumed high probability bound for $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$ in Assumption K.2 (d), it now suffices to bound $\|\widehat{\mathbf{T}}_n(\mathbf{x}) - \mathbb{E}\{\widehat{\mathbf{T}}_n(\mathbf{x})\}\|_\infty$ and $\|\mathbb{E}\{\widehat{\mathbf{T}}_n(\mathbf{x})\}\|_\infty$. To this end, for each $\mathbf{x} \in \mathcal{X}$, define

$$\mathbf{T}^*_h(\mathbf{Z};\mathbf{x}) \ := \ \frac{1}{h^2}Z(\mathbf{X}-\mathbf{x})K'\left(\frac{W-w_{\mathbf{x}}}{h}\right) \ \text{ so that } \widehat{\mathbf{T}}_n(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{T}^*_h(\mathbf{Z}_i;\mathbf{x}).$$

Now under Assumptions K.1 (a), K.1 (c), K.2 (a), K.2 (d) and using Lemma C.1 (i)(b)-(c), (iv)(b) and (v) at appropriate places, we have: for all $\mathbf{x} \in \mathcal{X}$,

$$\max_{1 \le j \le p} \|\mathbf{T}^*_{h[j]}(\mathbf{Z};\mathbf{x})\|_{\psi_2} \;\le\; 2h^{-2} M_{\mathbf{X}} M_{K'} \sigma_Z \;\text{ and therefore,}$$
$$\max_{1 \le j \le p} \|\mathbf{T}^*_{h[j]}(\mathbf{Z};\mathbf{x}) - \mathbb{E}\{\mathbf{T}^*_h(\mathbf{Z};\mathbf{x})\}\|_{\psi_2} \;\le\; 6h^{-2} M_{\mathbf{X}} M_{K'} \sigma_Z.$$

Further, under Assumptions K.2 (d), K.1 (c), K.2 (a) and with $\mathbb{E}\{Z^2(\mathbf{X}_{[j]} - \mathbf{x}_{[j]})^2 | W\} \le 4M^2_{\mathbf{X}} \mathbb{E}_W(Z^2|W) \equiv 4M^2_{\mathbf{X}} m^{(2)}_{\boldsymbol{\beta}}(W) \; \forall j$, we have: for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\max_{1 \le j \le p} \mathbb{E}[\{\mathbf{T}^*_{h[j]}(\mathbf{Z};\mathbf{x})\}^2] \;\le\; \frac{4}{h^4} M^2_{\mathbf{X}} \int_{\mathbb{R}} m^{(2)}_{\boldsymbol{\beta}}(w)[K'\{(w - w_{\mathbf{x}_j})/h\}]^2 f_{\boldsymbol{\beta}}(w)dw \\
&\le\; \frac{4}{h^3} M^2_{\mathbf{X}} M_{K'} B_1 \int_{\mathbb{R}} m^{(2)}_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu) f_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu)\{K'(u)\}^2 du \\
&\le\; \frac{4}{h^3} M^2_{\mathbf{X}} M_{K'} B_1 \int_{\mathbb{R}} |K'(u)| du \;\;\le\; \frac{4}{h^3} B_1 M^2_{\mathbf{X}} M_{K'} C_{K'},
\end{aligned}$$

where the second step follows from a change of variable argument and the final two bounds follow from using the assumptions mentioned above.

Using Lemma C.6 with the parameters therein set to: $\alpha = 2$, $\Gamma_n \propto h^{-3}$ and $K_n \propto h^{-2}$, all in the light of the two bounds above, we then have: for any fixed $\mathbf{x} \in \mathcal{X}$ and for any $t \ge 0$, with probability at least $1 - 3\exp(-t^2)$,

$$\begin{aligned}
\left\|\widehat{\mathbf{T}}_n(\mathbf{x}) - \mathbb{E}\{\widehat{\mathbf{T}}_n(\mathbf{x})\}\right\|_\infty \;&\equiv\; \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{T}^*_h(\mathbf{Z}_i;\mathbf{x}) - \mathbb{E}\{\mathbf{T}^*_h(\mathbf{Z};\mathbf{x})\}\right\|_\infty \\
&\le\; C_1 \frac{(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_2 \frac{(t^2 + \log p)\sqrt{\log n}}{nh^2}, \qquad \text{(K.9)}
\end{aligned}$$

for some constants $C_1, C_2 > 0$ depending only on those introduced in the assumptions. Here, we further used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$ to obtain the bound (K.9) from the one originally provided by Lemma C.6.

Next, we focus on controlling $\|\mathbb{E}\{\mathbf{T}^*_h(\mathbf{Z};\mathbf{x})\}\|_\infty$. To this end, recall the definitions of $\boldsymbol{\eta}_{\boldsymbol{\beta}}(\cdot) \in \mathbb{R}^p$ and $l_{\boldsymbol{\beta}}(\cdot) \in \mathbb{R}$, and let $\boldsymbol{\eta}'_{\boldsymbol{\beta}}(w) := \frac{d}{dw}\boldsymbol{\eta}_{\boldsymbol{\beta}}(w) \in \mathbb{R}^p$. Then, under Assumption K.2 (a)-(b), we have: uniformly in $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}\{\mathbf{T}^*_h(\mathbf{Z};\mathbf{x})\} \;&=\; \frac{1}{h^2}\mathbb{E}_W[\mathbb{E}\{(ZX - Zx)|W\} K'\{(W - w_{\mathbf{x}})/h\}] \\
&\equiv\; \frac{1}{h^2}\int_{\mathbb{R}} \{\boldsymbol{\eta}_{\boldsymbol{\beta}}(w) - \mathbf{x} l_{\boldsymbol{\beta}}(w)\} K'\{(W - w_{\mathbf{x}})/h\} dw \\
&=\; \frac{1}{h}\int_{\mathbb{R}} \{\boldsymbol{\eta}_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu) - \mathbf{x} l_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu)\} K'(u) du \\
&=\; \int_{\mathbb{R}} \{\boldsymbol{\eta}'_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu) - \mathbf{x} l'_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu)\} K(u) du,
\end{aligned}$$

where the last two steps follow from a change of variable and integration by parts argument, where the latter is applicable under Assumption K.2 (a)-(b). Under Assumptions K.2 (a), K.2 (b) and K.2 (d), we then have:

$$\begin{aligned}\|\mathbb{E}\{\mathbf{T}_h^*(\mathbf{Z};\mathbf{x})\}\|_\infty &\leq \left\{\max_{1\leq j\leq p}\|\boldsymbol{\eta}'_{\boldsymbol{\beta}[j]}(\cdot)\|_\infty + \|\mathbf{x}\|_\infty\|l'_{\boldsymbol{\beta}}(\cdot)\|_\infty\right\}\int_{\mathbb{R}}|K(u)|du \\ &\leq (B_1^* + M_{\mathbf{X}}B_2^*)C_K \quad \text{uniformly in } \mathbf{x}\in\mathcal{X}. \qquad \text{(K.10)}\end{aligned}$$

Finally, recall that from Assumption K.2 (d), we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq a_n$ with probability at least $1-q_n$. Combining this with the bounds (K.9) and (K.10) and applying them in (K.8) through a simple union bound, we have: for any fixed $\mathbf{x}\in\mathcal{X}$ and for $t\geq 0$, with probability at least $1-3\exp(-t^2)-q_n$,

$$|\widehat{R}_{n,1}(\mathbf{x})| \;\leq\; a_n\left\{C_1^* + C_2^*\frac{(t+\sqrt{\log p})}{\sqrt{nh^3}} + C_3^*\frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right\},$$

for some constants C_1^*, C_2^*, C_3^* depending only on those introduced in our assumptions. This establishes the first part of Lemma K.2. ∎

To establish the second part of Lemma K.2 regarding bounds for $\widehat{R}_{n,2}(\mathbf{x})$, first recall that for some 'intermediate' points $\{W_i^*\}_{i=1}^n$ and $w_{\mathbf{x}}^*$ satisfying $|(W_i^* - w_{\mathbf{x}}^*) - (W_i - w_{\mathbf{x}})| \leq |(\widehat{W}_i - \widehat{w}_{\mathbf{x}}) - (W_i - w_{\mathbf{x}})| \equiv |(\widehat{\boldsymbol{\beta}} - \beta)'(\mathbf{X}_i - \mathbf{x})|$,

$$\begin{aligned}|\widehat{R}_{n,2}(\mathbf{x})| &\equiv \left|\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'}{nh^2}\sum_{i=1}^n Z_i(\mathbf{X}_i - \mathbf{x})\left\{K'\left(\frac{W_i^*-w_{\mathbf{x}}^*}{h}\right) - K'\left(\frac{W_i-w_{\mathbf{x}}}{h}\right)\right\}\right| \\ &\leq \frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1}{nh^2}\sum_{i=1}^n\|\mathbf{X}_i-\mathbf{x}\|_\infty|Z_i|\left|K'\left(\frac{W_i^*-w_{\mathbf{x}}^*}{h}\right) - K'\left(\frac{W_i-w_{\mathbf{x}}}{h}\right)\right| \\ &\leq 2M_{\mathbf{X}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1\left\{\frac{1}{nh^2}\sum_{i=1}^n|Z_i|\left|K'\left(\frac{W_i^*-w_{\mathbf{x}}^*}{h}\right) - K'\left(\frac{W_i-w_{\mathbf{x}}}{h}\right)\right|\right\}, \qquad \text{(K.11)}\end{aligned}$$

where the steps follow from an L_1-L_∞ bound along with a triangle inequality and using the boundedness of $\mathbf{X}$ from Assumption K.2 (d).

Let $\mathcal{A}_n$ denote the event $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \leq a_n$ and let $\mathcal{A}_n^c$ denote the complement event of $\mathcal{A}_n$. Then, from Assumption K.2 (d), we have $\mathbb{P}(\mathcal{A}_n) \geq 1-q_n$. Further, on the event $\mathcal{A}_n$, $(\widehat{\boldsymbol{\beta}}-\beta)'(\mathbf{X}_i-\mathbf{x})/h \leq 2M_{\mathbf{X}}(a_n/h) \leq L$ under Assumption K.2 (d) and consequently, using Assumption K.2 (c) with the function $\varphi(\cdot)$ as defined therein, we have: on the event $\mathcal{A}_n$,

$$\left| K'\left(\frac{W_i - w_{\mathbf{x}}}{h}\right) - K'\left(\frac{W_i^* - w_{\mathbf{x}}^*}{h}\right) \right| \;\leq\; \frac{1}{h}|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}_i - \mathbf{x})|\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right)$$

(K.12)

$$\leq\; \frac{1}{h}\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_1\|(\mathbf{X}_i - \mathbf{x})\|_\infty\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right) \;\leq\; \frac{2M_{\mathbf{X}}a_n}{h}\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right),$$

and consequently, combining (K.11) and (K.12), we have: on the event $\mathcal{A}_n$,

$$|\widehat{R}_{n,2}(\mathbf{x})| \;\leq\; \frac{2M_{\mathbf{X}}^2a_n^2}{nh^3}\sum_{i=1}^n |Z_i|\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right) \quad \forall\, \mathbf{x} \in \mathcal{X}. \tag{K.13}$$

Thus, we have: for any $\epsilon \geq 0$ and for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{P}(|\widehat{R}_{n,2}(\mathbf{x})| > \epsilon) \;&\leq\; \mathbb{P}(|\widehat{R}_{n,2}(\mathbf{x})| > \epsilon, \mathcal{A}_n) + \mathbb{P}(|\widehat{R}_{n,2}(\mathbf{x})| > \epsilon, \mathcal{A}_n^c)\\
&\leq\; \mathbb{P}\left\{\frac{4M_{\mathbf{X}}^2a_n^2}{nh^3}\sum_{i=1}^n |Z_i|\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right) > \epsilon, \mathcal{A}_n\right\} + \mathbb{P}(\mathcal{A}_n^c)\\
&\leq\; \mathbb{P}\left\{\frac{4M_{\mathbf{X}}^2a_n^2}{nh^3}\sum_{i=1}^n |Z_i|\varphi\left(\frac{W_i - w_{\mathbf{x}}}{h}\right) > \epsilon\right\} + q_n, \qquad \text{(K.14)}
\end{aligned}$$

where the steps follow from (K.13) and that $\mathbb{P}(\mathcal{A}_n^c) \leq q_n$ by assumption.

Next, define: $\mathcal{T}_h(\mathbf{Z};\mathbf{x}) \equiv \mathcal{T}_h(\mathbf{Z};\mathbf{x},\boldsymbol{\beta}) := h^{-3}|Z|\varphi\{(W - w_{\mathbf{x}})/h\}$ and recall that $m_{\boldsymbol{\beta}}^{(2)}(W) \equiv \mathbb{E}(Z^2|W)$. Then, using the boundedness conditions from Assumptions K.1 (c) and K.2(c), along with use of iterated expectations, we bound the first and second moments of $\mathcal{T}_h(\mathbf{Z};\mathbf{x})$ $\forall\, \mathbf{x} \in \mathcal{X}$ as follows.

$$\begin{aligned}
\mathbb{E}\{\mathcal{T}_h^2(\mathbf{Z};\mathbf{x})\} \;&=\; \frac{1}{h^6}\int_{\mathbb{R}} m_{\boldsymbol{\beta}}^{(2)}(w)\varphi^2\left(\frac{W - w_{\mathbf{x}}}{h}\right) f_{\boldsymbol{\beta}}(w)dw\\
&=\; \frac{1}{h^5}\int_{\mathbb{R}} m_{\boldsymbol{\beta}}^{(2)}(w_{\mathbf{x}} + hu)f_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu)\varphi^2(u)du \;\leq\; \frac{B_1M_\varphi C_\varphi}{h^5}, \;\text{ and}\\
\mathbb{E}\{\mathcal{T}_h(\mathbf{Z};\mathbf{x})\} \;&=\; \frac{1}{h^3}\int_{\mathbb{R}} \mathbb{E}(|Z||W = w)\varphi\left(\frac{W - w_{\mathbf{x}}}{h}\right) f_{\boldsymbol{\beta}}(w)dw\\
&\leq\; \frac{1}{h^3}\int_{\mathbb{R}} \{m_{\boldsymbol{\beta}}^{(2)}(w)\}^{\frac{1}{2}}\varphi\left(\frac{W - w_{\mathbf{x}}}{h}\right) f_{\boldsymbol{\beta}}(w)dw\\
&\leq\; \frac{1}{h^2}\int_{\mathbb{R}} \{m_{\boldsymbol{\beta}}^{(2)}(w_{\mathbf{x}} + hu)\}^{\frac{1}{2}}\varphi(u)f_{\boldsymbol{\beta}}(w_{\mathbf{x}} + hu)du \;\leq\; \frac{(B_1C_f)^{\frac{1}{2}}C_\varphi}{h^2},
\end{aligned}$$

where $C_f > 0$ is a constant such that $\|f_{\boldsymbol{\beta}}(\cdot)\|_\infty \leq C_f$. Further, under Assumptions K.1 (a) and K.2 (c), using various parts of Lemma C.1, we have:

$$\|\mathcal{T}_h(\mathbf{Z};\mathbf{x}) - \mathbb{E}\{\mathcal{T}_h(\mathbf{Z};\mathbf{x})\}\|_{\psi_2} \;\leq\; 3\|\mathcal{T}_h(\mathbf{Z};\mathbf{x})\|_{\psi_2} \;\leq\; 3h^{-3}\sigma_Z M_\varphi \quad \forall \mathbf{x} \in \mathcal{X}.$$

Hence, using Lemma C.6, with all required conditions verified now, we have: for any $\mathbf{x} \in \mathcal{X}$ and for any $t \geq 0$, with probability at least $1 - 3\exp(-t^2)$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}_h(\mathbf{Z}_i; \mathbf{x}) \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}_h(\mathbf{Z}_i; \mathbf{x}) - \mathbb{E}\{\mathcal{T}_h(\mathbf{Z}; \mathbf{x})\} \right| + |\mathbb{E}\{\mathcal{T}_h(\mathbf{Z}; \mathbf{x})|$$

$$\leq C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + \frac{C_5}{h^2}, \tag{K.15}$$

for some constants $C_3, C_4, C_5 > 0$ depending only on those in the assumptions. Hence, using (K.15) in (K.14), we now have: for any $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ |\widehat{R}_{n,2}(\mathbf{x})| \geq 4M_{\mathbf{X}}^2 a_n^2 \left(C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + \frac{C_5}{h^2} \right) \right\} \\
&\leq \mathbb{P}\left\{ \frac{1}{nh^3} \sum_{i=1}^{n} |Z_i| \varphi\left(\frac{W_i - w_{\mathbf{x}}}{h} \right) > C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + \frac{C_5}{h^2} \right\} + q_n \\
&\equiv \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}_h(\mathbf{Z}_i; \mathbf{x}) \right| > C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + \frac{C_5}{h^2} \right) + q_n \\
&\leq 3\exp(-t^2) + q_n \quad \text{for any } \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

This establishes the desired bound for $\widehat{R}_{n,2}(\mathbf{x})$ and completes the proof. ∎

Abhishek Chakrabortty
Dept. of Statistics
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: abhich@wharton.upenn.edu

Jiarui Lu
Dept. of Biostatistics, Epidemiology & Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: jiaruilu@pennmedicine.upenn.edu

T. Tony Cai
Dept. of Statistics
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: tcai@wharton.upenn.edu

Hongzhe Li
Dept. of Biostatistics, Epidemiology & Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: hongzhe@pennmedicine.upenn.edu