In this paper, we consider high dimensional $M$-estimation problems in settings where the response $Y$ is possibly missing at random and the covariates $X \in \mathbb{R}^p$ 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 biomedical studies. The parameter of interest $\theta_0 \in \mathbb{R}^d$ is defined simply as the risk minimizer of a convex loss under a fully non-parametric model and $\theta_0$ 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 personalized medicine.

Assuming $\theta_0$ is $s$-sparse (with $s \ll n$), we propose to estimate $\theta_0$ using an $L_1$ regularized debiased and doubly robust estimator (DDR) based on a high dimensional adaptation of traditional double robust (DR) estimators’ construction along with careful usage 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 and show its $L_2$ 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 the nuisance function working models are correctly specified, we propose a desparsification method to obtain an asymptotic linear expansion of our DDR estimator which facilitates inference on low dimensional components of $\theta_0$. Finally, we discuss a variety of high dimensional parametric and semi-parametric working models for the PS and OR estimators and establish their properties needed for our main results.

*This is a working paper and the current draft is certainly not complete. The authors accept full responsibilities for any errors in this incomplete and unpublished manuscript.

Keywords and phrases: Missing data, Causal inference, High dimensional M-estimation, Regularized estimation and sparsity, Debiased and doubly robust estimation, Cross-fitting, Desparsification, High dimensional semi-parametric inference, Nuisance functions.

1
1. The Problem Setup. Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ respectively denote an outcome and a set of covariates of interest, with supports $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^p$ both of which are allowed to be of arbitrary nature (not necessarily continuous). In the observable data, however, the outcome $Y$ may not always be observed and the random vector that can be actually observed is given by $Z := (T, Y, X)$, where $T \in \{0, 1\}$ denotes the indicator of the original outcome $Y$ being observed and $Y$ denotes the observed outcome satisfying: $T Y = T Y$, i.e. $(Y | T = 1) \equiv Y$ almost surely (a.s.) which is often also referred to as the consistency assumption. We next discuss the two main settings that are unified by this general set-up.

(a) Missing data setting (with missing $Y$). The observable random vector in this case may be represented as: $Z := (T, TY, X)$, where $T \in \{0, 1\}$ is the indicator of the original outcome $Y$ being observed (Tsiatis, 2007).

(b) Causal inference setting. The observable random vector in this case is $Z := (T, Y, X)$, where $T \in \{0, 1\}$ denotes a binary treatment assignment indicator (here, ‘treatment’ may correspond to any kind of binary assignment or intervention) and $Y := T Y^{(1)} + (1 - T) Y^{(0)}$ 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. This set-up, for each potential outcome, is included in our general framework if we set $(Y, T) \equiv (Y^{(1)}, T)$ or, $(Y, T) \equiv (Y^{(0)}, 1 - T)$ respectively.

As an extension, settings with multi-category treatment assignments are also included with the observable random vector being $Z := (T, Y, X)$, where $T \in \{0, 1, \ldots, K\}$ is a categorical treatment assignment variable, with $K \geq 1$ fixed, and $Y := \sum_{j=0}^{K} Y^{(j)} 1(T = j)$ denotes the observed outcome with $\{Y^{(j)}\}_{j=0}^{K}$ being the potential outcomes corresponding to $\{T = j\}_{j=0}^{K}$. Clearly, by setting $(Y, T) \equiv \{Y^{(j)}, 1(T = j)\}$, this set-up is also included in our general framework for all $j \in \{0, 1, \ldots, K\}$.

For the causal inference setting, it is worth noting that the covariates $X$ are also often referred to as ‘confounders’ (in the case of observational studies) and ‘adjustment variables/features’ (in the case of randomized trials).

1.1. The Framework, Available Data and the Basic Assumptions. Given the similarities and equivalences of the examples above, we now simplify our notations without loss of generality (w.l.o.g.) and consider a set-up where we have a true underlying random vector $Z := (T, Y, X)$ of interest, defined on a common probability space with probability measure $\mathbb{P}(\cdot)$, but in practice one can only observe $Z := (T, TY, X)$, where $T \in \{0, 1\}$ denotes the indicator of $Y$ being observed. The observed data may be represented as: $D_n := \{Z_i \equiv \ldots$
\( (T_i, T_i Y_i, X_i) : i = 1, \ldots, n \) consisting of \( n \) independent and identically distributed (i.i.d.) realizations of \( Z \) with joint distribution defined via \( P(\cdot) \).

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

(a) (Ignorability assumption). We assume that \( T \perp \perp Y \mid X \). This assumption is also familiarly known as the missing at random (MAR) assumption in the missing data literature, and the no unmeasured confounding (NUC) assumption in the causal inference literature.

(b) (Positivity/overlap assumption). Let \( \pi(x) := P(T = 1 \mid X = x) \) \( \forall x \in \mathcal{X} \), denote the ‘propensity score’ (Rosenbaum and Rubin, 1983) of the outcome being observed given the covariates \( X \), and let \( \pi := P(T = 1) \). Then, we assume: \( \exists a \) universal constant \( \delta_\pi \) with \( 0 < \delta_\pi \leq 1 \), such that

\[
\pi(x) \geq \delta_\pi > 0 \quad \forall x \in \mathcal{X},
\]

(and hence, \( \pi \geq \delta_\pi > 0 \) as well).

Assumption 1.1 (a) includes as a special case: \( T \perp \perp (Y, X) \) which is also known as missing completely at random (MCAR) in missing data literature and complete randomization in the causal inference literature where it is mostly encountered in randomized trials. In such cases, \( \pi(\cdot) \) simply becomes the constant \( \pi \) defined above. In general, \( \pi(\cdot) \) may depend on \( X \) and may be unknown in practice when it needs to be estimated.

We wish to highlight here that the setting we consider is allowed to be high dimensional in terms of the covariates \( X \), i.e. \( p \) is allowed to diverge with \( n \) including \( p \ll n, p \sim n \) or \( p \gg n \), the latter scenario being of particular interest throughout, although our methods and the associated theory (which is mostly non-asymptotic) apply generally to any regime for \( (p,n) \). We now formalize the (high dimensional) M-estimation problem, based on convex and differentiable ‘loss functions’, that we wish to address under this setting.

### 1.2. The M-Estimation Problem.

Let \( L(Y, X, \theta) : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R} \) be any ‘loss function’ that is convex and differentiable in \( \theta \in \mathbb{R}^d \), and we assume that \( \mathbb{E}\{L(Y, X, \theta)\}^2 < \infty \) for each \( \theta \in \mathbb{R}^d \). Then, the M-estimation problem considers the estimation of the parameter vector \( \theta_0 \) that corresponds to the minimizer of the risk function defined by \( L(\cdot) \). Specifically, we aim to estimate the parameter vector \( \theta_0 \equiv \theta_0(\mathbb{P}) \in \mathbb{R}^d \), a functional of the probability measure \( \mathbb{P}(\cdot) \) underlying the unobserved \( Z \), defined as:

\[
\theta_0 \equiv \theta_0(L, \mathbb{P}) := \arg \min_{\theta \in \mathbb{R}^d} L(\theta), \quad \text{where} \quad L(\theta) := \mathbb{E}\{L(Y, X, \theta)\}.
\]
Here, $d$ is allowed to be high dimensional, i.e. $d$ can diverge with $n$ including $d \gg n$. We assume w.l.o.g. that $d \geq 2$. The existence and uniqueness of $\theta_0$ 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 $\theta$. In general, this can be guaranteed as long as the risk function $L(\cdot)$ is strongly convex and coercive in $\theta$. For convenience of further discussion, let us define: $\forall y \in Y$, $x \in X$ and $\theta \in \mathbb{R}^d$,
\[
\phi(x, \theta) := \mathbb{E}\{L(Y, X, \theta) \mid X = x\} \quad \text{and} \quad \nabla L(y, x, \theta) := \frac{\partial}{\partial \theta} L(y, x, \theta) \in \mathbb{R}^d.
\]

Some examples. $M$-estimation problems are quite well studied in classical settings and have a vast literature; see Van der Vaart (2000) for a review. We highlight here a few useful illustrative examples of high dimensional $M$-estimation problems, as in (1.2), that are frequently encountered in practice.

1. **High dimensional regression with possibly misspecified models and missing outcomes.** The framework (1.2) includes as special cases the class of all standard high dimensional regression problems, where we additionally allow for potentially misspecified ‘working’ models and the outcomes to be partly unobserved. For instance, set $d = p+1$ and $\theta = (a, b) \in \mathbb{R}^d$ with $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$ in (1.2), and let $L(Y, X, \theta) := l(Y, a + b'X)$ for some function $l(u, v) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $l(\cdot)$ being convex and differentiable in $v \in \mathbb{R}$. A few typical choices of $l(\cdot, \cdot)$ include the standard ‘canonical’ loss functions leading to standard regression problems as follows.

(a) The squared loss: $l(u, v) \equiv l_{sq}(u, v) := (u - v)^2$, the most common choice and usually motivated by a linear regression (working) model.

(b) The logistic loss: $l(u, v) \equiv l_{log}(u, v) := -uv + \log\{1 + \exp(v)\}$, which is a typical choice in binary outcome regression and is usually motivated by an underlying logistic regression (working) model.

(c) The exponential loss: $l(u, v) \equiv l_{exp}(u, v) := -uv + \exp(v)$, which is often used for regressing discrete (count) outcomes and is usually motivated by an underlying poisson regression (working) model.

Note that the definition of $\theta_0$ is fully non-parametric and ‘model-free’, i.e. it holds regardless of the actual validity of any underlying ‘working’ model motivating the construction, thus allowing for its misspecification.

As an extension, one may also consider any (model-free) series estimation problem, where the original covariate $X$ is replaced by a finite, but possibly high dimensional, vector $\Psi(X) := \{\psi_j(X)\}_{j=1}^d$ of $d$ basis functions comprising transformations (possibly non-linear) of $X$. One can then set
One frequently used choice of $\Psi(\cdot)$ includes the polynomial bases of any (fixed) degree $d_0 \geq 1$, given by: $\Psi(X) := \{1, x^k : 1 \leq j \leq p, 1 \leq k \leq d_0\}$ whereby $d = pd_0 + 1$. This leads to polynomial regression problems and the special case $d_0 = 1$ (linear bases) corresponds to the earlier examples.

2. Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design. Another important application of the framework in (1.2) lies in signal recovery in SIMs with elliptically symmetric designs that satisfy a certain ‘linearity condition’. To this end, let $Y = f(\beta_0'X, \epsilon)$ where $f(\cdot, \cdot) : \mathbb{R}^2 \to \mathcal{Y}$ denotes an unknown link function, $\epsilon \perp \perp X$ denotes a random noise (i.e. $Y \perp \perp X \mid \beta_0'X$) and $\beta_0$ denotes the unknown index parameter (identifiable only up to scalar multiples). Consider any of the regression problems introduced in Example 1, and assume that $X$ has an elliptically symmetric distribution. Then, $\theta_0 \equiv (a_0, b_0)$ defined therein satisfies: $b_0 \propto \beta_0$. This remarkable result was first noted in the seminal work of Li and Duan (1989) and 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 $\theta_0$ is defined based on a misspecified parametric model and yet, it has direct interpretability that relates it to a parameter characterizing a larger semi-parametric model and allows one to simply use (1.2) for signal recovery.

1.3. Identification and Alternative Representations of the Expected Loss. We next provide three key identifications and alternative representations of $\mathbb{L}(\cdot)$ in terms of the observables $(T, TY, X)$ and some nuisance functions estimable through them. Note that these identifications are all fully non-parametric, i.e. no further assumption on the underlying data generating mechanism is made to obtain these representations apart from the basic conditions in Assumption 1.1. These representations also underlie three fundamental strategies typically adopted in the literature for these estimation 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 $\phi(\cdot, \cdot)$. Estimators based on the final approach are also known to (locally) achieve the semi-parametric efficiency bound when both estimators 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
comprehensive overview of the related classical literature on these methods. *IPW and regression based representations of* $\mathbb{L}(\cdot)$. For any $\theta \in \mathbb{R}^d$, we have
\[
\mathbb{L}(\theta) \equiv \mathbb{E}\{L(Y, X, \theta)\} = \mathbb{E}_X\{\phi(X, \theta)\} =: \mathbb{L}_{REG}(\theta) \quad \text{(say),}
\]
and
\[
\mathbb{L}(\theta) \equiv \mathbb{E}\{L(Y, X, \theta)\} = \mathbb{E}\left\{\frac{T}{\pi(X)} L(Y, X, \theta)\right\} =: \mathbb{L}_{IPW}(\theta) \quad \text{(say)}.
\]

*Debiased and doubly robust (DDR) representation of* $\mathbb{L}(\cdot)$. We also have
\[
\mathbb{L}(\theta) = \mathbb{E}_X\{\phi(X, \theta)\} + \mathbb{E}\left[\frac{T}{\pi(X)} \{L(Y, X, \theta) - \phi(X, \theta)\}\right] \\
=: \mathbb{L}_{DDR}(\theta) \quad \forall \theta \in \mathbb{R}^d.
\]

Further, for any functions $\phi^*(X, \theta)$ and $\pi^*(X)$ such that $\phi^*(\cdot, \cdot) = \phi(\cdot, \cdot)$ or $\pi^*(\cdot) = \pi(\cdot)$ holds, but not necessarily both, it continues to hold that:
\[
\mathbb{L}_{DDR}(\theta) = \mathbb{E}_X\{\phi^*(X, \theta)\} + \mathbb{E}\left[\frac{T}{\pi^*(X)} \{L(Y, X, \theta) - \phi^*(X, \theta)\}\right].
\]

$\mathbb{L}_{DDR}(\cdot)$, unlike $\mathbb{L}_{IPW}(\cdot)$ and $\mathbb{L}_{REG}(\cdot)$, is therefore DR as it is ‘protected’ against misspecification of either one of $\pi(\cdot)$ or $\phi(\cdot, \cdot)$, as shown by (1.4). Further, even when both are correctly specified, it has a naturally ‘debiased’ form owing to the second term in (1.3), also called the augmented IPW term. While this term is simply 0 in the population version, it leads to crucial first order benefits in the empirical version (with the nuisance function estimators plugged in) wherein it acts as a debiasing term making the loss insensitive to estimation errors of the nuisance functions at the first order. Approaches based on the other representations don’t enjoy these debiasing benefits which can be particularly crucial in high dimensional settings. Further discussions on these nuances, under a more general context, and their importance in high dimensional settings can be found in the recent works of Chernozhukov et al. (2016, 2017, 2018a,b) and Chernozhukov, Newey and Robins (2018) on the use of ‘Neyman orthogonal’ scores for semi-parametric estimation and inference in the presence of (unknown) nuisance components.


Notations. We first introduce some notations to be used throughout. For any $v \in \mathbb{R}^d$, $j \in \{1, \ldots, d\}$ and $J \subseteq \{1, \ldots, d\}$, $\nabla v$ denotes $(1, v)’ \in \mathbb{R}^{d+1}$, $v[j]$ denotes the $j^{th}$ coordinate of $v$, $\|v\|_r$ denotes the $L_r$ vector norm of $v$ for any $r \geq 0$, $\mathcal{A}(v) := \{j : v[j] \neq 0\}$ denotes the support of $v$, $s_v := |\mathcal{A}(v)|$
denotes the cardinality of $\mathcal{A}(v)$, $\Pi_{\mathcal{J}}(v)$ denotes $[v_{[j]}1\{j \in \mathcal{J}\}]_{j=1}^d \in \mathbb{R}^d$, and $\mathcal{J}^c := \{1, \ldots, d\} \setminus \mathcal{J}$ denotes the complement of $\mathcal{J}$. We use the shorthand $\Pi_v(\cdot)$ and $\Pi_{\mathcal{J}}(\cdot)$ to denote $\Pi_{\mathcal{A}(v)}(\cdot)$ and $\Pi_{\mathcal{A}(v)}(\cdot)$ respectively. We further let $\mathcal{M}_{\mathcal{J}} = \{v \in \mathbb{R}^d : \mathcal{A}(v) \subseteq \mathcal{J}\}$ and $\mathcal{M}^{\perp}_{\mathcal{J}} = \{v \in \mathbb{R}^d : \mathcal{A}(v) \subseteq \mathcal{J}^c\}$. Lastly, for any measurable (and possibly random) function $f(\cdot)$ of $\mathbf{X}$, $\|f(\cdot)\|_r := [\mathbb{E}_X \{|f(\mathbf{X})|^r\}]^{1/r}$ denotes the $L_r$ norm of $f(\cdot)$ with respect to (w.r.t.) $\mathbb{P}_X$ for any $r \geq 1$, and $\|f(\cdot)\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$ denotes the $L_\infty$ norm w.r.t. $\mathbb{P}_X$.

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

$$\nabla L(Y, \mathbf{X}, \theta) \equiv \frac{\partial}{\partial \theta} L(Y, \mathbf{X}, \theta) = h(\mathbf{X})\{Y - g(\mathbf{X}, \theta)\}. \tag{2.1}$$

The structural assumption in (2.1) is mostly for simplicity in our theoretical analyses regarding probabilistic bounds for our proposed estimator and this form is satisfied by most standard loss functions used in practice, including the ones highlighted earlier as examples in Section 1.2. Extensions of our results to more generally structured loss functions may also be obtained, albeit at the cost of less tractable technical conditions.

Under (2.1), $L(\cdot)$ takes the form: $L(Y, \mathbf{X}, \theta) = \{h(\mathbf{X})\theta\}Y - f(\mathbf{X}, \theta) + C(Y, \mathbf{X})$ where $f(\mathbf{X}, \theta)$ denotes the anti-derivative of $h(\mathbf{X})g(\mathbf{X}, \theta)$ w.r.t. $\theta$ and $C(Y, \mathbf{X})$ is some function independent of $\theta$ (e.g. $C(Y, \mathbf{X}) := Y^2$ for the squared loss). Thus, $\phi(\mathbf{X}, \theta) = \{h(\mathbf{X})\theta\}m(\mathbf{X}) - f(\mathbf{X}, \theta) + m_C(\mathbf{X})$ where $m(\mathbf{X}) := \mathbb{E}(Y | \mathbf{X})$ and $m_C(\mathbf{X}) := \mathbb{E}\{C(Y, \mathbf{X}) | \mathbf{X}\}$. Further, $\phi(\mathbf{X}, \theta)$ is convex and differentiable in $\theta$ and $\nabla \phi(\mathbf{X}, \theta) := \frac{\partial}{\partial \theta} \phi(\mathbf{X}, \theta)$ is given by:

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

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

$$\nabla \hat{\phi}(\mathbf{X}, \theta) := \frac{\partial}{\partial \theta} \hat{\phi}(\mathbf{X}, \theta) = h(\mathbf{X})\{\hat{m}(\mathbf{X}) - g(\mathbf{X}, \theta)\}. \tag{2.2}$$

Note that to compute $\hat{\phi}(\mathbf{X}, \theta)$ explicitly, one needs both the estimates $\hat{m}(\cdot)$ and $\hat{m}_C(\cdot)$. However, the part of $\hat{\phi}(\mathbf{X}, \theta)$ involving $\hat{m}_C(\cdot)$ is free of $\theta$. Our proposed estimator, discussed in Section 2.2, is constructed based on an $L_1$-regularized minimization (w.r.t. $\theta$) of a objective function involving $\hat{\phi}(\cdot)$, whereby only its gradient $\nabla \hat{\phi}(\mathbf{X}, \theta)$ is of interest and that depends only on $\hat{m}(\mathbf{X})$ due to (2.2). Thus, the part of $\hat{\phi}(\cdot)$ involving $\hat{m}_C(\cdot)$ being free of $\theta$ may
be ignored for all practical purposes, and for implementing our estimator, we only require an estimator \( \hat{m}(\cdot) \) of \( m(\cdot) \) along with an arbitrary choice of \( \hat{m}(\cdot) \) to plug in and obtain a corresponding estimator \( \hat{\phi}(X, \theta) \) of \( \phi(X, \theta) \).

2.2. The \( L_1 \)-Regularized DDR Estimator. Let \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \) be any reasonable estimators of \( \pi(\cdot) \) and \( m(\cdot) \) respectively, such that at least one (but not necessarily both) of them are correctly specified estimators and we also assume that \( \hat{\pi}(\cdot) \) is obtained solely from the data on \( \{(T_i, X_i)\}_{i=1}^n \). Let \( \hat{\phi}(\cdot, \cdot) \) be the corresponding estimator of \( \phi(\cdot, \cdot) \) based on \( \hat{m}(\cdot) \). Further, we use sample splitting to construct ‘cross-fitted’ versions of \( \hat{m}(\cdot) \) and \( \hat{\phi}(\cdot, \cdot) \), as follows.

Cross-fitted versions of \( \hat{m}(\cdot) \) and \( \hat{\phi}(\cdot, \cdot) \) based on sample splitting. Let \( \{\mathcal{D}^{(1)}_n, \mathcal{D}^{(2)}_n\} \) denote a random partition (or split) of the original data \( \mathcal{D}_n \) into \( K = 2 \) equal parts. Let \( n := n/2 \) denote the size of \( \mathcal{D}^{(1)}_n \) and \( \mathcal{D}^{(2)}_n \) where, without loss of generality (w.l.o.g.), we assume that \( n \) is even. Further, let \( \mathcal{I}_1 \) and \( \mathcal{I}_2 \) respectively denote the index sets \( \{1, \ldots, n\} =: \mathcal{I} \) for the observations in \( \mathcal{D}^{(1)}_n \) and \( \mathcal{D}^{(2)}_n \). Hence, with \( K \equiv 2 \), we have: \( \{\mathcal{D}^{(k)}_n\}_{k=1}^K \) and \( \{\mathcal{I}_k\}_{k=1}^K \) are disjoint and exhaustive partitions of \( \mathcal{D} \) and \( \mathcal{I} \) respectively, i.e. \( \bigcup_{k=1}^K \mathcal{D}^{(k)}_n = \mathcal{D} \), \( \bigcup_{k=1}^K \mathcal{I}_k = \mathcal{I} \) and \( |\mathcal{I}_k| = \bar{n} \equiv n/K \) \( \forall \ k \in \{1, \ldots, K \} \equiv 2 \).

Given any general procedure for obtaining \( \hat{\pi}(\cdot) \) and \( \hat{\phi}(\cdot, \cdot) \) based on the full observed data \( \mathcal{D}_n \), let \( \{\hat{m}^{(1)}(\cdot), \hat{\phi}^{(1)}(\cdot, \cdot)\} \) and \( \{\hat{m}^{(2)}(\cdot), \hat{\phi}^{(2)}(\cdot, \cdot)\} \) denote the corresponding versions of these estimators based on only the datasets \( \mathcal{D}^{(1)}_n \) and \( \mathcal{D}^{(2)}_n \) respectively. Let us now define the following cross-fitted estimates \( \{\tilde{m}(X_i), \tilde{\phi}(X_i, \theta)\}_{i=1}^n \) of the conditional means \( \{m(X_i), \phi(X_i, \theta)\}_{i=1}^n \) at the \( n \) training points in \( \mathcal{D}_n \), as follows.

\[
(2.3) \quad \{\tilde{m}(X_i), \tilde{\phi}(X_i, \theta)\} = \begin{cases} 
\{\hat{m}^{(2)}(X_i), \hat{\phi}^{(2)}(X_i, \theta)\} & \forall \ i \in \mathcal{I}_1, \\
\{\hat{m}^{(1)}(X_i), \hat{\phi}^{(1)}(X_i, \theta)\} & \forall \ i \in \mathcal{I}_2.
\end{cases}
\]

A detailed discussion regarding the benefits (and virtual necessity) of considering these cross-fitted estimators is deferred to Section 3.7. 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 \( \{\hat{\pi}(X_i)\}_{i=1}^n \) as long as \( \hat{\pi}(\cdot) \) is obtained only from the data on \( \{(T_i, X_i)\}_{i=1}^n \).

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

\[
\mathbb{L}_{\text{DDR}}(\theta) = \mathbb{E}_X \{\phi(X; \theta)\} + \mathbb{E} \left[ \frac{T}{\pi(X)} \{L(Y, X, \theta) - \phi(X; \theta)\} \right],
\]
and define its empirical version, based on the estimates \( \{ \tilde{\phi}(\mathbf{X}, \theta), \tilde{\pi}(\mathbf{X}_i) \}_{i=1}^n \) plugged in, as follows. For any \( \theta \in \mathbb{R}^d \), let us define the empirical DDR loss

\[
(2.4) \quad L_{n, \text{DDR}}^{\text{DDR}}(\theta) := \frac{1}{n} \sum_{i=1}^n \tilde{\phi}(\mathbf{X}_i, \theta) + \frac{1}{n} \sum_{i=1}^n \frac{T_i}{\tilde{\pi}(\mathbf{X}_i)} \left\{ L(Y_i, \mathbf{X}_i, \theta_i) - \tilde{\phi}(\mathbf{X}_i, \theta) \right\}.
\]

With \( \theta_0 \) (and \( \mathbf{X} \)) possibly high dimensional, when \( d \gg n \), we shall need to assume that \( \theta_0 \) is sparse with sparsity much smaller than \( d \). In general, we assume that \( \theta_0 \) is \( s \)-sparse i.e. \( \| \theta_0 \|_0 = s \) with \( 1 \leq s \leq d \). We now propose to estimate \( \theta_0 \) using the \( L_1 \)-regularized DDR estimator, \( \hat{\theta}_{\text{DDR}} \), given by:

\[
(2.5) \quad \hat{\theta}_{\text{DDR}} = \hat{\theta}_{\text{DDR}}(\lambda_n) = \arg \min_{\theta \in \mathbb{R}^d} \{ L_n^{\text{DDR}}(\theta) + \lambda_n \| \theta \|_1 \},
\]

where \( L_n^{\text{DDR}}(\cdot) \) is as in (2.4) and \( \lambda_n \geq 0 \) denotes the regularization (or tuning) parameter. (Under a classical setting with \( p \ll n \), \( \lambda_n \) may be set to 0, if desired, although we do not pursue the theoretical analysis for this case).

2.3. Simple Algorithm for Implementation. The estimator \( \hat{\theta}_{\text{DDR}} \) in (2.5) can be implemented using a simple user-friendly imputation type algorithm as follows. Given the observed data \( \mathcal{D}_n \) and the estimates \( \{ \tilde{\pi}(\mathbf{X}_i), \tilde{m}(\mathbf{X}_i) \}_{i=1}^n \), let us first define a set of pseudo outcomes \( \tilde{Y}_i \), for all \( 1 \leq i \leq n \), given by:

\[
\tilde{Y}_i := \tilde{m}(\mathbf{X}_i) + \frac{T_i}{\tilde{\pi}(\mathbf{X}_i)} \{ Y_i - \tilde{m}(\mathbf{X}_i) \}, \quad \text{and let} \quad \tilde{L}_n^{\text{DDR}}(\theta) := \frac{1}{n} \sum_{i=1}^n L(\tilde{Y}_i, \mathbf{X}_i, \theta).
\]

Clearly \( \tilde{L}_n^{\text{DDR}}(\cdot) \) is convex and differentiable, and under (2.1) and (2.2), it is straightforward to see that \( \nabla \tilde{L}_n^{\text{DDR}}(\theta) = \nabla L_n^{\text{DDR}}(\theta) \), where \( \nabla \tilde{L}_n^{\text{DDR}}(\theta) := \frac{\partial}{\partial \theta} \tilde{L}_n^{\text{DDR}}(\theta) \). Further, observe that the solution for the minimization in (2.5) is uniquely determined by the underlying KKT conditions which only depend on the gradient of \( L_n^{\text{DDR}}(\cdot) \) and the subgradient of the \( \| \cdot \|_1 \) norm. Hence, the solution stays unchanged if \( L_n^{\text{DDR}}(\theta) \) in (2.5) is replaced by \( \tilde{L}_n^{\text{DDR}}(\theta) \) which has the same gradient. Consequently, \( \hat{\theta}_{\text{DDR}} \) in (2.5) may also be defined as:

\[
(2.6) \quad \hat{\theta}_{\text{DDR}} = \hat{\theta}_{\text{DDR}}(\lambda_n) := \arg \min_{\theta \in \mathbb{R}^d} \{ \tilde{L}_n^{\text{DDR}}(\theta) + \lambda_n \| \theta \|_1 \}.
\]

Thus, if one ‘pretends’ to have a fully observed data \( \tilde{\mathcal{D}}_n := \{ (\tilde{Y}_i, \mathbf{X}_i) \}_{i=1}^n \) in terms of the pseudo outcomes, then \( \hat{\theta}_{\text{DDR}} \) may be obtained by a simple \( L_1 \)-penalized minimization of the corresponding empirical risk for \( L(\cdot) \) based on \( \tilde{\mathcal{D}}_n \). This minimization is quite easy to implement and can be done so using standard statistical software packages, including ‘glmnet’ in R. Finally, note

\[
\hat{\theta}(\lambda) = \hat{\theta}_{\text{DDR}}(\lambda_n) = \arg \min_{\theta \in \mathbb{R}^d} \{ \tilde{L}_n^{\text{DDR}}(\theta) + \lambda \| \theta \|_1 \},
\]

where \( \lambda \) and \( \lambda_n \) are the regularization (or tuning) parameter and its empirical version, respectively.
also that (2.6) confirms our earlier claim that even though the estimator  
\( \tilde{\phi}(X, \theta) \) involved in the definition (2.4) of  
\( L_{\text{DDR}}^n(\theta) \) may require estimation of further nuisance functions (independent of \( \theta \)) apart from \( m(X) \), but the implementation of  
\( \hat{\theta}_{\text{DDR}} \) via the minimization in (2.5), or equivalently the one in (2.6), requires only an estimator of \( m(X) \), along with that of \( \pi(X) \).

2.4. Performance Guarantees: Deviation Bounds. We next provide a deterministic deviation bound regarding the finite sample performance of  
\( \hat{\theta}_{\text{DDR}} \) that will serve as the backbone for most of our main theoretical analyses. We first introduce some notations and assumptions. Define:

\[ \nabla L_{\text{DDR}}^n(\theta) := \frac{\partial}{\partial \theta} L_{\text{DDR}}^n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla \tilde{\phi}(X_i, \theta) + \frac{1}{n} \sum_{i=1}^{n} \nabla \frac{T_i}{\hat{\pi}(X_i)} \left\{ \nabla L(Y_i, X_i, \theta) - \nabla \tilde{\phi}(X_i, \theta) \right\}, \]

and  
\[ \delta L_{\text{DDR}}^n(\theta; v) := L_{\text{DDR}}^n(\theta + v) - L_{\text{DDR}}^n(\theta) - v' \{ \nabla L_{\text{DDR}}^n(\theta) \}. \]

Assumption 2.1 (Restricted strong convexity). Let  
\( \delta L_{\text{DDR}}^n(\theta; v) \) be as above for any  \( \theta, v \in \mathbb{R}^d \). We assume that at  \( \theta = \theta_0 \), the loss function  
\( L_{\text{DDR}}^n(\theta) \) satisfies a restricted strong convexity (RSC) property as follows:  \( \exists \) a (non-random) constant  \( \kappa_{\text{DDR}} > 0 \) such that  
(2.8)  
\[ \delta L_{\text{DDR}}^n(\theta_0; v) \geq \kappa_{\text{DDR}}\|v\|_2^2 \quad \forall v \in \mathbb{C}(\theta_0), \]

where  
\[ \mathbb{C}(\theta_0) := \left\{ v \in \mathbb{R}^d : \|\Pi_{\theta_0}(v)\|_1 \leq 3\|\Pi_{\theta_0}(v)\|_1 \right\} \subseteq \mathbb{R}^d. \]

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 (2.8) to hold deterministically for any realization of  \( D_n \), it only needs to hold a.s.  \( \mathbb{P} \) for some \( \kappa_{\text{DDR}} > 0 \). With appropriate modifications, it can also be generalized further whereby it only needs to hold with high probability. It is important to note that owing to the very structure of  \( L_{\text{DDR}}^n(\cdot) \) in (2.4) and (2.7) and the assumed structures (2.1) and (2.2) for  \( L(\cdot) \) and  \( \tilde{\phi}(\cdot) \), the RSC condition (2.8) is completely independent of the quantities depending on the missingness aspect of the problem, i.e.  \( \delta L_{\text{DDR}}^n(\theta_0; v) \) in (2.8) is independent of  \( \{T_i, Y_i\}_{i=1}^n \) as well as the nuisance function estimates  \( \{\tilde{\pi}(X_i), m(X_i)\}_{i=1}^n \). 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  \( \hat{\theta}_{\text{DDR}} \) in (2.6) based on loss minimization involving the ‘pseudo’ outcomes. Thus, verifying (2.8) is equivalent to verifying the same for a fully observed data.
Since verification of the RSC condition under a fully observed data 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 therefore provides an easy route to verifying the RSC condition (2.8) under our setting.

**Lemma 2.1** (Deterministic deviation bounds for \( \hat{\theta}_{DDR} \)). Let \( \hat{\theta}_{DDR} \) be as defined in (2.5) and assume that \( \mathcal{L}_{n}^{DDR}(\theta) \) in (2.4) is convex and differentiable in \( \theta \) a.s. \( [P] \). Let Assumption 2.1 hold with the constant \( \kappa_{DDR} > 0 \) as defined in (2.8) therein and recall that \( s := \|\theta_{0}\|_{0} \). Then, given any realization of \( D_{n} \) and for any given choice of \( \lambda \equiv \lambda_{n} \geq 2 \|\nabla \mathcal{L}_{n}^{DDR}(\theta_{0})\|_{\infty} \), we have:

\[
\|\hat{\theta}_{DDR} - \theta_{0}\|_{2} \leq 3\sqrt{s} \frac{\lambda_{n}}{\kappa_{DDR}} \quad \text{and} \quad \|\hat{\theta}_{DDR} - \theta_{0}\|_{1} \leq 12s \frac{\lambda_{n}}{\kappa_{DDR}}.
\]

Convergence rates (informal statement). In Section 3, we further establish (see Theorems 3.1-3.4) that under suitable assumptions (given in Section 3.2), \( \|\nabla \mathcal{L}_{n}^{DDR}(\theta_{0})\|_{\infty} \lesssim \sqrt{(\log d)/n} \) with high probability. Hence, by choosing \( \lambda \equiv \lambda_{n} \approx \sqrt{(\log d)/n} \) and using (2.9), it follows that with high probability,

\[
\|\hat{\theta}_{DDR} - \theta_{0}\|_{2} \lesssim \sqrt{s \log d/n} \quad \text{and} \quad \|\hat{\theta}_{DDR} - \theta_{0}\|_{1} \lesssim s \sqrt{(\log d)/n}.
\]

The deviation bound (2.9), essentially an easy consequence of the results of Negahban et al. (2012), deterministically establishes the rates of the estimator in terms of the chosen \( \lambda_{n} \) and provides an easy recipe for establishing its convergence rates by studying the same for the lower bound of \( \lambda_{n} \) given in Lemma 2.1. Hence, the main goal from hereon is to analyze the (random) lower bound \( 2 \|\nabla \mathcal{L}_{n}^{DDR}(\theta_{0})\|_{\infty} \) in Lemma 2.1 regarding the choice of \( \lambda_{n} \).

3. The Core Analyses for the DDR Estimator: Probabilistic Upper Bounds for \( \|\nabla \mathcal{L}_{n}^{DDR}(\theta_{0})\|_{\infty} \) and Convergence Rates. For most of our theoretical analyses of \( \|\nabla \mathcal{L}_{n}^{DDR}(\theta_{0})\|_{\infty} \), we will assume that \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \) are both correctly specified estimators of \( \{\pi(\cdot), m(\cdot)\} \). The analyses even for this case are quite involved and nuanced, with the key technical challenges being the presence of the nuisance function estimators (that leads
to controlling averages of dependent variables) and the inherent high dimen-
sional setting (which necessitates sharp non-asymptotic bounds).

Under possible misspecification of one of the estimators, the DR property
(in terms of consistency) of \( \| \nabla L_n^{\text{DDR}}(\theta_0) \|_\infty \) and hence, that of \( \hat{\theta}_{\text{DDR}}(\lambda_n) \) for
a suitably chosen \( \lambda_n \) sequence according to Lemma 2.1, should indeed follow
owing to the very nature of construction of \( L_n^{\text{DDR}}(\cdot) \) and more fundamentally,
its population version \( L_{\text{DDR}}(\cdot) \), as shown in (1.3)-(1.4). This DR property is
fairly 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.

One of the reasons behind considering the DDR representations \( L_{\text{DDR}}(\theta) \)
and \( L_n^{\text{DDR}}(\theta) \) is that apart from the obvious benefits of double robustness
which is quite well known (at least in classical settings), the construction
of the DDR loss is such that it has a naturally ‘debiased’ form that pro-
vides remarkable technical benefits in controlling the associated error terms
which are naturally ‘centered’ (in some sense) when both \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \) are
correctly specified, a setting when other approaches such as IPW and REG
type estimators are also applicable (albeit possibly less efficient than the
DDR estimator) but these approaches do not enjoy such technical benefits.
The advantages of such debiased representations, especially in high dimen-
sional settings, have been studied in more general contexts under the name
of ‘Neyman orthogonalization’ in the recent works of Chernozhukov et al.
(2016, 2017, 2018a,b) and Chernozhukov, Newey and Robins (2018). The
DDR representation indeed (naturally) satisfies such an orthogonal struc-
ture.

Furthermore, it is important to note that our analyses here are completely
free in terms of the choice of the nuisance function estimators. The results
and the convergence rates we obtain are first order insensitive to any estima-
tion errors of the nuisance functions and hold regardless of any knowledge of
how these estimators are obtained and what their first order properties are,
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.

Lastly, even under possible misspecification of one of \( \{ \hat{\pi}(\cdot), \hat{m}(\cdot) \} \), we still
sketch the DR property of \( \| \nabla L_n^{\text{DDR}}(\theta_0) \|_\infty \) in Section 7 later, but only in
terms of consistency with the convergence rates being obtained based on a
general but crude analysis. To obtain possibly faster convergence rates, one
needs a more nuanced and careful analysis which, even in a classical setting,
has to be done in a case-specific manner as the analysis and the first order
properties (and convergence rates) now will depend on the exact nature of
construction of the estimators and their corresponding first order properties.
Considering the main goals and scope of this paper, we suppress such finer analysis under those cases for the sake of simplicity and brevity.

### 3.1. The Basic Decomposition

Let $T_n := \nabla L_{\theta_0}^{\text{DDR}}(\theta_0) \in \mathbb{R}^d$ with $\|T_n\|_\infty$ being our quantity of interest. We first note a decomposition of $T_n$ as follows.

$$T_n = T_{0,n} + T_{\pi,n} - T_{m,n} - R_{\pi,m,n}$$

where

\[ T_{0,n} = \frac{1}{n} \sum_{i=1}^{n} T_0(Z_i), \quad T_{\pi,n} = \frac{1}{n} \sum_{i=1}^{n} T_{\pi}(Z_i), \quad T_{m,n} = \frac{1}{n} \sum_{i=1}^{n} T_m(Z_i), \quad R_{\pi,m,n} = \frac{1}{n} \sum_{i=1}^{n} R_{\pi,m}(Z_i), \]

and

\[ T_{0}(Z) := \{m(X) - g(X, \theta_0)\} h(X) + \frac{T}{\pi(X)} \{Y - m(X)\} h(X), \]

\[ T_{\pi}(Z) := \left\{ \frac{T}{\hat{\pi}(X)} - \frac{T}{\pi(X)} \right\} \{Y - m(X)\} h(X), \]

\[ T_{m}(Z) := \left\{ \frac{T}{\hat{\pi}(X)} - 1 \right\} \{\tilde{m}(X) - m(X)\} h(X), \quad \text{and} \]

\[ R_{\pi,m}(Z) := \left\{ \frac{T}{\hat{\pi}(X)} - \frac{T}{\pi(X)} \right\} \{\tilde{m}(X) - m(X)\} h(X). \]

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

We next proceed towards our main goal of obtaining non-asymptotic probabilistic bounds for $\|T_n\|_\infty$. In order to control $\|T_n\|_\infty$, we separately control each of $\|T_{0,n}\|_\infty$, $\|T_{\pi,n}\|_\infty$, $\|T_{m,n}\|_\infty$ and $\|R_{\pi,m,n}\|_\infty$. The results (Theorems 3.1-3.4) are given in a stepwise manner, and show that the convergence rate of $\|T_n\|_\infty$ is determined primarily by that of the leading order term $\|T_{0,n}\|_\infty$, while the rates contributed by the other three terms are all of a (faster) lower order. The proofs of Theorems 3.1-3.4 are given in Sections C-F respectively. In general, they involve careful analyses based on concentration inequalities and other techniques. These technical tools are all collected in Section A.

### 3.2. The Assumptions Required

In this section, we summarize the main assumptions we require for controlling the various terms in the decomposition (3.1) for $T_n$. We begin with a few standard assumptions on the tail behaviors of some key random variables involved in our analyses.
Assumption 3.1 (Sub-Gaussian tail behaviors). (a) We assume that 
\[ \varepsilon(Z) := Y - m(X), \psi(X) := m(X) - g(X, \theta_0) \] and \( h(X) \) are sub-Gaussian (as per Definition A.1 with \( \alpha = 2 \) therein) with \( \|\varepsilon\|_{\psi_2} \leq \sigma_{\varepsilon}, \|\psi(X)\|_{\psi_2} \leq \sigma_{\psi} \) and \( \|h(X)\|_{\psi_2} \leq \sigma_{h} \) for some constants \( \sigma_{\varepsilon}, \sigma_{\psi}, \sigma_{h} \geq 0 \).

(b) For controlling \( T_{\pi,n} \), we additionally assume that \( \{\varepsilon(Z)|X\} \) is (conditionally) sub-Gaussian with \( \|\varepsilon(Z)|X\|_{\psi_2} \leq \sigma_{\varepsilon}(X) \) for some function \( \sigma_{\varepsilon}(\cdot) \geq 0 \) such that \( \|\sigma_{\varepsilon}(\cdot)\|_{\infty} \leq \sigma_{\varepsilon} < \infty \) with \( \sigma_{\varepsilon} \) 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 \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \). Further discussions on the assumptions are given in Remarks 3.1-3.2.

Assumption 3.2 (Tail bounds on the pointwise behavior of \( \hat{\pi}(\cdot) - \pi(\cdot) \)). We assume that the estimator \( \hat{\pi}(\cdot) \) is obtained based solely on the data subset \( X_n := \{(T_i, X_i)\}_{i=1}^n \). Further, for some non-negative sequences \( b_{n,\pi} = o(1) \) and \( v_{n,\pi} = o(1) \), and some sequence \( q_{n,\pi} \in [0, 1] \) with \( q_{n,\pi} = o(1) \), we assume that \( \hat{\pi}(\cdot) - \pi(\cdot) \) satisfies a (pointwise) tail bound only at the \( n \) training points \( \{X_i\}_{i=1}^n \) as follows. For any \( t \geq 0 \) and for some constant \( C \geq 0 \), we have:

\[ \mathbb{P}\{|\hat{\pi}(X_i) - \pi(X_i)| > tv_{n,\pi} + b_{n,\pi}\} \leq C \exp(-t^2) + q_{n,\pi} \quad \forall \ 1 \leq i \leq n. \]  

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

Assumption 3.3 (Pointwise tail bounds on \( \hat{m}(\cdot) - m(\cdot) \)). Let \( \hat{m}(\cdot) \) denote a generic version of the estimator of \( m(\cdot) \) obtained from a dataset of size \( n \), e.g. the original data \( D_n \). Then, we assume that for some non-negative sequences \( b_{n,m} = o(1) \) and \( v_{n,m} = o(1) \) and some sequence \( q_{n,m} \in [0, 1] \) with \( q_{n,m} = o(1) \), \( \hat{m}(\cdot) - m(\cdot) \) satisfies a (pointwise) tail bound as follows. For any fixed \( x \in X \), any \( t \geq 0 \) and for some constant \( C \geq 0 \), we have:

\[ \mathbb{P}\{|\hat{m}(x) - m(x)| > tv_{n,m} + b_{n,m}\} \leq C \exp(-t^2) + q_{n,m}. \]

Consequently, \( \forall \ k \in \{1, 2\} \), \( \hat{m}(k)(\cdot) \) obtained from \( D_n^{(k)} \) with sample size \( \bar{n} \equiv n/2 \) and rates \( \{v_{\bar{n},m}, b_{\bar{n},m}, q_{\bar{n},m}\} \), satisfies a pointwise tailbound as follows. For all \( k' \neq k \) \in \{1, 2\} and for each \( X_i \in D_n^{(k')} \perp D_n^{(k)} \),

\[ \mathbb{P}\{|\hat{m}(k)(X_i) - m(X_i)| > tv_{n,m} + b_{n,m}\} \leq C \exp(-t^2) + q_{\bar{n},m}. \]

Further, we assume that \( v_{n,m}\sqrt{\log(nd)} + b_{n,m} = o(1) \), \( q_{\bar{n},m} = o(n^{-1}d^{-1}) \) and that \( (v_{n,\pi}\sqrt{\log n} + b_{n,\pi})(v_{\bar{n},m}\sqrt{\log \bar{n}} + b_{\bar{n},m}) = o\{\sqrt{(\log d)/n}\} \).
Remark 3.1. 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 \(\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}\) of \(\{\pi(\cdot), m(\cdot)\}\). Note that the bounds (3.6) and (3.7) are both conditions on the pointwise behaviors of \(\hat{\pi}(\cdot) - \pi(\cdot)\) and \(\hat{m}(\cdot) - m(\cdot)\) respectively and do not require any uniform tail bounds over all \(x \in \mathcal{X}\), such as bounds on the \(L_{\infty}\) errors \(\|\hat{\pi}(\cdot) - \pi(\cdot)\|_{\infty}\) and \(\|\hat{m}(\cdot) - m(\cdot)\|_{\infty}\) or \(L_2\) errors \(\|\hat{\pi}(\cdot) - \pi(\cdot)\|_2\) and \(\|\hat{m}(\cdot) - m(\cdot)\|_2\). Such bounds are generally much stronger requirements and also harder to verify in high dimensional settings.

The assumptions simply require pointwise tail bounds for the error terms \(\hat{\pi}(X_i) - \pi(X_i)\), at each training point \(X_i\), and \(\hat{m}(x) - m(x)\), for any fixed \(x \in \mathcal{X}\), ensuring that they have sufficiently well-behaved tails. The sequences \(\{v_{n,\pi}, v_{n,m}\}\) indicate the convergence rates of the stochastic components of the estimators, while \(\{b_{n,\pi}, b_{n,m}\}\) account for the rates of any (deterministic) bias terms. Further, the sequences \(\{q_{n,\pi}, q_{n,m}\}\) in the probability bounds allow to rigourously account for potential lower order terms that may be encountered in the analysis of the estimators. Finally, note that we require explicit tail bounds, as opposed to assumptions only on the asymptotic rates, since \(T_n\) is high dimensional and these error terms are directly involved in analyzing the (non-asymptotic) behavior of \(\|T_n\|_{\infty}\) that we need to control.

Remark 3.2. In Section 5, we discuss several choices of the estimators \(\hat{\pi}(\cdot)\) and \(\hat{m}(\cdot)\) 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 quite useful and are of independent interest. Among other implications, they also verify the bounds in Assumptions 3.2 and 3.3.

In general, for any estimator \(\hat{\pi}(\cdot)\) of \(\pi(\cdot)\) that satisfies a high probability (pointwise) bound of the form: \(|\hat{\pi}(X_i) - \pi(X_i)| \leq v_n\) with probability at least \(1 - q_n\), for any \(1 \leq i \leq n\), the bound in Assumption 3.2 can be shown to hold with \(\{v_{n,\pi}, q_{n,\pi}, b_{n,\pi}\} \equiv \{\sqrt{2}v_n, q_n, 0\}\), through a simple application of Hoeffding’s inequality. Similarly, for any estimator \(\hat{m}(\cdot)\) that satisfies a high probability bound of the form: \(|\hat{m}(x) - m(x)| \leq v_n\) with probability at least \(1 - q_n\), for any fixed \(x \in \mathcal{X}\), the bounds in Assumption 3.2 can be shown to hold with \(\{v_{n,m}, q_{n,m}, b_{n,m}\} \equiv \{\sqrt{2}v_n, q_n, 0\}\). These high probability bounds should be expected to be satisfied by most reasonable estimators and consequently, the assumptions are also expected to hold in most cases.

3.3. Controlling the Leading Order Term. We first aim to control the term \(\|T_{0.n}\|_{\infty}\) in (3.1) which is essentially the ‘leading order’ term but also has the simplest structure and is easiest to control among all terms in (3.1).
THEOREM 3.1 (Control of $\|T_{0,n}\|_\infty$). Under Assumptions 1.1 and 3.1 (a),

$$\mathbb{P}\left(\|T_{0,n}\|_\infty > \sqrt{2} \sigma_0 \epsilon + K_0 \epsilon^2 \right) \leq 4 \exp\left(-n \epsilon^2 + \log d\right) \quad \text{for any } \epsilon \geq 0,$$

where $\sigma_0 := 2\sqrt{2} \sigma_h (\sigma_\psi + \sigma_\pi \delta_\pi^{-1}) \geq 0$ and $K_0 := 2\sigma_h (\sigma_\psi + \sigma_\pi \delta_\pi^{-1}) \geq 0$ are constants. Hence, with $\epsilon = c\sqrt{\log d}/n$, for any constant $c > 1$, we have:

With probability $\geq 1 - \frac{4}{d^{c^2-1}}$,

$$\|T_{0,n}\|_\infty \leq c\sqrt{\frac{\log d}{n}} \sqrt{2\sigma_0 + c^2 \log d} K_0 \leq \frac{\log d}{n}.$$

3.4. Controlling the Error Term from the Estimation of Propensity Score.

We next focus on controlling the term $T_{\pi,n}$ in the decomposition (3.1).

THEOREM 3.2 (Control of $\|T_{\pi,n}\|_\infty$). Let Assumptions 1.1, 3.1 and 3.2 hold, with the sequences $(v_{n,\pi}, b_{n,\pi}, q_{n,\pi})$ and the constants $(\delta_\pi, \sigma_\pi, C)$ being as defined therein, and let $\|\mu_h^{(2)}\|_\infty := \max\{\mathbb{E}\{h_{ij}^2(X)\} : j = 1, \ldots, d\}$. Then, for any constants $c_1, c_2, c_3 > 1$, where we further assume w.l.o.g. that $c_2 v_{n,\pi} \sqrt{\log (nd)} + b_{n,\pi} \leq \delta_\pi/2 < \delta_\pi$ and $c_3 \sqrt{\log d}/n < 1$, we have:

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

$$\|T_{\pi,n}\|_\infty \leq c\sqrt{\frac{\log d}{n}} \left\{v_{n,\pi} \sqrt{\log (nd)} + b_{n,\pi}\right\} \left(\frac{\|\mu_h^{(2)}\|_\infty}{\delta_\pi} + C_2 \sqrt{\frac{\log d}{n}}\right)^{\frac{1}{2}},$$

where $C_1 := c_1 (4\sqrt{2} \sigma_\pi / \delta_\pi)$ and $C_2 := c_3 (\sqrt{2} \sigma_\pi + K_\pi)$, with $\sigma_\pi := 2\sqrt{2} \sigma_h^2 \delta_\pi^{-2}$ and $K_\pi := 2\sigma_h^2 \delta_\pi^{-2}$ being constants.

REMARK 3.3. Theorem 3.2 therefore shows that $\|T_{\pi,n}\|_\infty \leq \sqrt{\log (nd)/n}$ \{v_{n,\pi} \sqrt{\log (nd)} + b_{n,\pi}\} = o\{\sqrt{\log (nd)/n}\} w.h.p. In the proof of Theorem 3.2, we also provide a general result (Theorem D.1) on tail bounds for $T_{\pi,n}$.

3.5. Controlling the Error Term from the Conditional Mean’s Estimation.

We next focus on controlling the term $T_{m,n}$ in the decomposition (3.1) involving the cross-fitted estimates $\{\hat{m}(X_i)\}_{i=1}^n$ obtained via sample splitting.

THEOREM 3.3 (Control of $\|T_{m,n}\|_\infty$). Let Assumptions 1.1, 3.1 (a) and 3.3 hold, with the sequences $(v_{n,m}, b_{n,m}, q_{n,m})$, $\bar{n} \equiv n/2$ and the constants
(δ_π, C) being as defined therein. Then, for any constants c, c_1, c_2 > 1, where we further assume w.l.o.g. that c_2 \sqrt{(\log d)/n} < 1, we have:

With probability \geq 1 - \frac{4}{d^2-1} - \frac{8}{d^2-1} - \frac{4C}{(\bar{n}d)^{c_2-1}} - 4q_{\bar{n},m}(\bar{n}d),

\|T_{m,n}\|_\infty \leq c\sqrt{\frac{\log d}{n}}\{v_{\bar{n},m}\sqrt{\log(\bar{n}d)} + b_{\bar{n},m}\}C_1^* \left( \|\mu_h^{(2)}\|_\infty + C_2^* \sqrt{\frac{\log d}{n}} \right)^{\frac{1}{2}},

where \|\mu_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_h^2, K_m := 2\sigma_h and \bar{\delta}_\pi \leq \delta_\pi being constants.

Remark 3.4. Theorem 3.3 therefore shows that \|T_{m,n}\|_\infty \lesssim \sqrt{(\log d)/n}
\{v_{\bar{n},m}\sqrt{\log(\bar{n}d)} + b_{\bar{n},m}\} = o\{\sqrt{(\log d)/n}\} w.h.p. In the proof of Theorem 3.3, we also provide a general result (Theorem E.1) on tail bounds for T_{m,n}.

3.6. Controlling The Lower Order Term. Finally, we now control the term R_{\pi,m,n} in (3.1) involving the random variable R_{\pi,m}(Z) defined in (3.5).

Theorem 3.4 (Control of \|R_{\pi,m,n}\|_\infty). Let Assumptions 1.1, 3.1, 3.2 and 3.3 hold, with the sequences \{(v_{n,\pi}, b_{n,\pi}, q_{n,\pi}), (v_{\bar{n},m}, b_{\bar{n},m}, q_{\bar{n},m}, \bar{n})\} and the constants (δ_\pi, C) as defined therein. Then, for any constants c_1, c_2, c_3, c_4 > 1 with c_2 v_{n,\pi}\sqrt{\log n} + b_{n,\pi} \leq \delta_\pi/2 < \delta_\pi and c_4\sqrt{(\log d)/n} < 1, we have:

With probability \geq 1 - \sum_{j=1}^3 \frac{C}{n^{c_j-1}} - \frac{2}{d^2-1} - 2nq_{\pi,\pi} - nq_{\bar{n},m},

\|R_{\pi,m,n}\|_\infty \leq c_1c_3 \bar{C}_1 r_{\pi,n} r_{m,\bar{n}} \left( \|\mu_{\pi,\pi,\pi}\|_\infty + c_4 \bar{C}_2 \sqrt{\frac{\log d}{n}} \right),

where r_{\pi,n} := v_{n,\pi}\sqrt{\log n} + b_{n,\pi} and r_{m,\bar{n}} := v_{\bar{n},m}\sqrt{\log \bar{n}} + b_{\bar{n},m}, with r_{\pi,n} r_{m,\bar{n}} = o\{\sqrt{(\log d)/n}\}, \|\mu_{\pi,\pi,\pi}\|_\infty := \max_{1 \leq j \leq d} \|\mu_{h,j}(X)\| \text{ 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_h\delta_\pi^{-1} and K_{\pi,m} := 2\sqrt{2}\sigma_h\delta_\pi^{-1}.

Remark 3.5. Theorem 3.4 therefore shows that \|R_{\pi,m,n}\|_\infty \lesssim r_{\pi,n} r_{m,\bar{n}} = o\{\sqrt{(\log d)/n}\} w.h.p., where the last step is by assumption a sufficient condition for which is max\{r_{\pi,n}, r_{m,\bar{n}}\} \lesssim (\log d)/n)^{0.25}. In the proof of Theorem 3.4, we provide a general result (Theorem F.1) on tail bounds for R_{\pi,m,n}.

3.7. Some Key Technical Discussions on the Structure of the Error Terms.
(a) The structure of $T_{\pi,n}$ and reasons for obtaining $\hat{\pi}(\cdot)$ solely from $X_n$.

Note that $T_{\pi,n}$ is simply the sample average of the random variables $T_{\pi}(Z)$ in (3.3). However, this average is not an i.i.d. average due to the presence of the estimator $\hat{\pi}(\cdot)$ which depends on all observations in $D_n$. In this regard, an important feature of $\hat{\pi}(\cdot)$ that is quite useful for our purposes here is that $\hat{\pi}(\cdot)$ is obtained based on only the sub-part $X_n := \{(T_i, X_i) : i = 1, \ldots, n\}$ of the full observed data $D_n$, as assumed. We then have: $E \{T_{\pi}(Z_i)\} = E[E \{T_{\pi}(Z_i) | \hat{\pi}(\cdot), X_i\}] = E[E \{T_{\pi}(Z_i) | X_n\}] = 0$, owing to the definitions of the underlying quantities involved, the nature of the construction of $\hat{\pi}(\cdot)$, and Assumption 1.1 (a). The conditioning on $X_n$ ensures that $\hat{\pi}(\cdot)$, as well as all other components in $T_{\pi}(Z_i)$, which are functions of $(T_i, X_i)$ only, can now be treated as fixed, and further, the conditional expectation being 0 follows from the fact that $E \{Y_i - m(X_i)\} | X_n) \equiv E \{\varepsilon(Z_i) | X_n\} = E \{\varepsilon(Z_i) | T_i, X_i\} = E \{\varepsilon(Z_i) | X_i\} = 0$, where the final step is due to Assumption 1.1 (a).

Thus, $T_{\pi,n}$ is a centred average of (conditionally) independent variables. We exploit this fact and the structure of $T_{\pi}(Z_i)$ in order to control $T_{\pi,n}$.

(b) The structure of $T_{m,n}$ and the benefits of sample splitting/cross-fitting.

Note that $T_{m,n}$ is essentially the sample average of the variables $T_m(Z)$ in (3.4). However, in the absence of sample splitting, this average is not an i.i.d. average due to the presence of the estimator $\hat{m}(\cdot)$ which depends on all observations in $D_n$. Further, unlike the case of $T_{\pi,n}$, where $\{T_{\pi}(Z_i)\}_{i=1}^n | X_n$ were (conditionally) independent and centered variables, $T_{m,n}$ possesses no such desirable features even if $\hat{m}(\cdot)$ is obtained based on only the subpart

$D_{n,Y} \equiv \{(Y_i, X_i) : T_i = 1, 1 \leq i \leq n\} \equiv \{(Y_i^{(1)}, X_i) : T_i = 1, 1 \leq i \leq n\}$

of the full data $D_n$, as $D_{n,Y}$ still (implicitly) depends on $\{T_i\}_{i=1}^n$ (due to the restriction to the set with $T_i = 1$) and not just on $\{Y_i, X_i\}_{i=1}^n$.

Thus, unlike $T_{\pi,n}$, $T_{m,n}$ (in the absence of sample splitting) 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 $T_{m,n}$ (without sample splitting), one needs tools from empirical process theory and the corresponding analyses can be substantially involved, especially in high dimensional settings. However, these technical complications can be avoided through the sample splitting based construction of the estimates $\{\hat{m}(X_i), n_{i=1}\}$ which ‘induces’ a natural independence.

To see this, note that for a $Z \equiv (T, Y, X) \perp \hat{m}(\cdot)$, or more specifically, $Z \perp \{\text{data used to obtain } \hat{m}(\cdot)\}$, $E \{T_m(Z) | \hat{m}(\cdot), X\} = E \{T_m(Z) | X\} = 0$ due to Assumption 1.1 (a). Hence, $E \{T_m(Z) | \hat{m}(\cdot)\} = 0$ and $E \{T_m(Z)\} = 0$. Further, for any i.i.d. collection $\{Z_k\}_{k=1}^K$ of $Z \perp \hat{m}(\cdot)$, $\{T_m(Z_k)\}_{k=1}^K \perp \hat{m}(\cdot)$ are (conditionally) independent and centered random variables. These serve as the main motivations behind the sample splitting based construction.
In contrast to the ‘in-sample’ estimates \( \{ \hat{m}(X_i) \}_{i=1}^n \) wherein \( \hat{m}(\cdot) \) is obtained from \( D_n \) and also evaluated at the same time at the training points \( \{ X_i \}_{i=1}^n \in D_n \), thereby making them intractably dependent on \( \hat{m}(\cdot) \), the cross-fitted estimates \( \{ \tilde{m}(X_i) \}_{i=1}^n \) ensure that for each \( k \neq k' \in \{1, 2\} \), the evaluation points \( \{ X_i \in D_n^{(k)} \} \) used are independent of the estimator \( \hat{m}(k')(\cdot) \) obtained from \( D_n^{(k')} \perp D_n^{(k)} \), thus inducing a desirable independence structure between the training and the evaluation points. This has substantial technical as well as practical benefits in reducing over-fitting bias.

Remark 3.6. While we focus here on the simple case of sample splitting, i.e. \( K \)-fold ‘cross-fitting’ with \( K = 2 \), our notations as well as the theoretical analyses are designed to easily accommodate the general case of \( K \)-fold cross fitting for any fixed \( K \geq 2 \). The extension to such cases is seamless and we stick to the case of \( K = 2 \) for simplicity and brevity of the technical arguments. Note further that the estimator \( \hat{\theta}_{DDR} \) obtained from the cross-fitting procedure can also be replicated several times, from several partitions of \( D_n \), and then appropriately combined over all those replications to average out the (minor) randomness induced by the sample splitting.

4. Desparsifying the DDR Estimator: Asymptotic Linear Expansion and Inference. We next discuss a desparsification approach for our estimator \( \hat{\theta}_{DDR} \). The desparsification is useful to establish regular estimators with an asymptotic linear expansion, a property that is not possessed by the \( L_1 \) regularized shrinkage estimator \( \hat{\theta}_{DDR} \). Such an expansion then automatically leads to an asymptotic normal distribution and a corresponding confidence interval for any low dimensional component of \( \theta_0 \) (e.g. each coordinate of \( \theta_0 \)) and therefore is useful for inference (e.g. confidence intervals).

For simplicity, we will restrict our discussion to the case of the squared loss, given by: \( L(Y, X, \theta) = \left\{ Y - \Psi(X)\theta \right\}^2 \), where \( \Psi(X) \equiv \{ \Psi_j(X) \}_{j=1}^d \in \mathbb{R}^d \) denotes some vector of basis functions of \( X \) with \( d \geq 1 \) possibly high dimensional. The special case \( \Psi(X) = (1, X)' \), with \( d = p + 1 \), corresponds to standard linear regression. For convenience of discussion, let us define:

\[
\Sigma := \mathbb{E}\{\Psi(X)\Psi(X)'\}, \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \Psi(X_i)\Psi(X_i)' \quad \text{and} \quad \Omega := \Sigma^{-1},
\]

where we assume that \( \mathbb{E}\{\|\Psi(X)\|_2^2\} < \infty \) and \( \Sigma \) is positive definite, so that \( \Sigma \) and \( \Omega \) are both well-defined. For any matrix \( M \in \mathbb{R}^{d \times d} \) and \( \forall 1 \leq i, j \leq d \), let \( M_{ij} \in \mathbb{R} \), \( M_{[i]} \in \mathbb{R}^d \) and \( M_{[j]} \in \mathbb{R}^d \) respectively denote the \((i, j)^{th}\) entry, the \(i^{th}\) row and the \(j^{th}\) column of \( M \). Further, let \( \|M\|_1 := \sum_{i=1}^d |M_{ii}| \).
\[ \max_1 \leq j \leq d \sum_{i=1}^d |M_{ij}|, \quad \|M\|_\infty := \max_1 \leq i \leq d \sum_{j=1}^d |M_{ij}| \quad \text{and} \quad \|M\|_{\max} := \max_1 \leq i, j \leq d |M_{ij}| \]\denote the matrix \(L_1\) norm, the matrix \(L_\infty\) norm and the elementwise maximum norm of \(M\) respectively. Finally, let us recall all notations defined in the basic decomposition (3.1) of \(T_n \equiv \|\nabla L_n^{\text{DDR}}(\theta_0)\|_{\infty}\) and also observe that under the specific choice of \(L(\cdot)\) here, we have:

\[ (4.1) \quad -\frac{1}{2}T_0(Z) = \{m(X) - \Psi(X) \theta_0\} \Psi(X) + \frac{T}{\pi(X)} \{Y - m(X)\} \Psi(X). \]

4.1. The Desparsified DDR Estimator. Let \(\hat{\Omega}\) be any reasonable estimator of the precision matrix \(\Omega\) based on the observed data \(D_n\). Then given the original \(L_1\) penalized DDR estimator \(\hat{\theta}_{\text{DDR}}\) in (2.5), or equivalently in (2.6), the corresponding desparsified DDR estimator \(\tilde{\theta}_{\text{DDR}}\) is defined as follows.

\[ (4.2) \quad \tilde{\theta}_{\text{DDR}} := \hat{\theta}_{\text{DDR}} - \frac{1}{2} \hat{\Omega} \nabla L_n^{\text{DDR}}(\hat{\theta}_{\text{DDR}}) \equiv \hat{\theta}_{\text{DDR}} - \frac{1}{2} \hat{\Omega} \nabla L_n^{\text{DDR}}(\hat{\theta}_{\text{DDR}}) \]

\[ = \hat{\theta}_{\text{DDR}} + \hat{\Omega} \frac{1}{n} \sum_{i=1}^n \{\tilde{Y}_i - \Psi(X_i) \hat{\theta}_{\text{DDR}}\} \Psi(X_i), \quad \text{where} \]

\[ \tilde{Y}_i \equiv \tilde{m}(X_i) + \{T_i/\hat{\pi}(X_i)\}\{Y_i - \tilde{m}(X_i)\}, \quad \forall \ 1 \leq i \leq n, \quad \text{denotes the pseudo outcomes defined in Section 2.3.} \quad \text{Using (4.2), it then follows easily that} \]

\[ (4.3) \quad \tilde{\theta}_{\text{DDR}} - \theta_0 = \frac{1}{n} \sum_{i=1}^n \psi_0(Z_i) + R_{n,1} + R_{n,2} + \Delta_n, \quad \text{where} \]

\[ 2\psi_0(Z) := -\Omega T_0(Z) \quad \text{with} \ E\{\psi_0(Z)\} = 0, \quad 2R_{n,1} := (\Omega - \hat{\Omega}) \nabla L_n^{\text{DDR}}(\theta_0), \]

\[ 2R_{n,2} := -\Omega(T_{\pi,n} + T_{m,n} + R_{\pi,m,n}) \quad \text{and} \quad \Delta_n := (I - \hat{\Omega} \hat{\Sigma})(\hat{\theta}_{\text{DDR}} - \theta_0). \]

Formal results and the rest of the details for this section to be added soon. Please check the slides for an overview of the results.

5. Estimation of the Nuisance Functions. In this section, we discuss various choices for estimating the propensity score \(\pi(\cdot)\) and the conditional mean \(m(\cdot)\) which are required for implementing our proposed methods. Our entire approach and theory so far is completely free of any specific knowledge of the construction and/or first order properties of these nuisance function estimators as long as they satisfy the basic high level conditions in Assumption 3.2 and 3.3. Consequently, 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.
However, a fully non-parametric and/or machine learning based approach may not be feasible or practicable in truly high dimensional settings where \( p \) diverges with \( n \). In this section, we extensively discuss a few novel, principled and yet, flexible, choices of \( \hat{\pi}(\cdot) \) and \( \hat{\pi}(\cdot) \), including commonly used parametric models, as well as series estimators and single index models. We establish a general set of results for all these estimators under high dimensional settings which, apart from verifying our basic assumptions, may also be of independent interest.


In some cases, \( \pi(\cdot) \) may be known whereby \( \hat{\pi}(\cdot) \equiv \pi(\cdot) \) 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: \( \pi(x) \equiv \mathbb{E}(T|X = x) = g(\alpha' \Psi(x)) \), where \( g(\cdot) \in [0, 1] \) is a known ‘link’ function, \( \Psi(x) := \{\psi_k(x)\}_{k=1}^K \) is any set of \( K \) (known) basis functions, possibly high dimensional with \( K \) allowed to depend on \( n \) (including \( K \gg n \)), and \( \alpha \in \mathbb{R}^K \) is an unknown parameter vector that is further assumed to be sparse (if required).

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

Examples. These models above include, for instance, any logistic regression model for \( T|X \) given by: \( \pi(x) = g(\alpha' \Psi(x)) \), where \( g(u) = \text{expit}(a) := \exp(a)/(1 + \exp(a)) \). The estimator \( \hat{\alpha} \) in this case maybe obtained using a simple \( L_1 \)-penalized logistic regression of \( T \) vs. \( \Psi(X) \) based on the observed data \( \{(T_i, \Psi(X_i))\}_{i=1}^n \). Using standard results from high dimensional regression theory (Bühlmann and Van De Geer, 2011; Negahban et al., 2012; Hastie, Tibshirani and Wainwright, 2015), it can be shown that under suitable assumptions, \( \|\hat{\alpha} - \alpha\|_1 \lesssim a_n \equiv a_n(s_\alpha, K) := s_\alpha \sqrt{\log K}/n \) with high probability (w.h.p.), where \( s_\alpha := \|\alpha\|_0 \) denotes the sparsity of \( \alpha \).

As for the basis functions \( \Psi(x) \), some reasonable choices include the polynomial bases given by: \( \Psi(x) := \{1, x_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0 \} \) 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 \( g(\alpha' X) \)
simply equals the constant \( \pi \) and \( \alpha \) 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: \( g\{\gamma'\Psi(X)\} \) where \( g(\cdot) \) is a (known) ‘link’ function (e.g. ‘canonical’ links functions), \( \Psi(X) := \{\psi_k(X)\}_{k=1}^{K} \) is any set of \( K \) (known) basis functions, with \( K \) possibly high dimensional and allowed to depend on \( n \) (including \( K \gg n \)), and \( \gamma \in \mathbb{R}^{K} \) is an unknown parameter vector that is further assumed to be sparse (if required).

Estimator. We estimate \( m(x) \equiv \mathbb{E}(Y|X) \equiv \mathbb{E}(Y|X, T = 1) = g\{\gamma'\Psi(X)\} \) via any suitable estimation procedure based on the ‘complete case’ data \( \mathcal{D}^{(c)}_n := \{(Y_i, X_i) | T_i = 1\}_{i=1}^{n} \) that satisfies a basic ‘high level’ requirement that \( \|\tilde{\gamma} - \gamma\|_1 \leq a_n \) w.h.p. for some sequence \( a_n = o(1) \).

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

As for the basis functions \( \Psi(x) \), some reasonable choices include the polynomial bases given by: \( \Psi(x) := \{1, x^j : 1 \leq j \leq p, 1 \leq k \leq d_0\} \) 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.

Examples of \( \tilde{\gamma} \). For all the examples above, with \( g(\cdot) \) being any ‘canonical’ link function, the estimator \( \tilde{\gamma} \) of \( \gamma \) may be simply obtained through a corresponding \( L_1 \) penalized ‘canonical’ link based regression (e.g. linear, logistic or Poisson regression) of \( Y \) vs. \( X \) in the ‘complete case’ data \( \mathcal{D}^{(c)}_n \) under Assumption 1.1 (a). Using standard results from high dimensional regression (Bühlmann and Van De Geer, 2011; Negahban et al., 2012; Hastie, Tibshirani and Wainwright, 2015), it can be shown that under suitable assumptions and Assumption 1.1, \( \|\tilde{\gamma} - \gamma\|_1 \lesssim a_n \equiv a_n(s_\gamma, K) := s_\gamma \sqrt{(\log K)/n} \) w.h.p., where \( s_\gamma := \|\gamma\|_0 \) denotes the sparsity of \( \gamma \).

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

**Estimator.** Given any reasonable estimator $\hat{\gamma}$ of the $\gamma$ ‘direction’, obtained based on some suitable procedure on the observed data $D_n$, we then estimate

$$m(X) \equiv \mathbb{E}(Y \mid \gamma'X) \equiv \mathbb{E}(Y \mid \gamma'X, T = 1) = g(\gamma'X)$$

via a one-dimensional KS over the estimated scores $\{\hat{\gamma}'X_i\}_{i=1}^n$, under appropriate smoothness and regularity assumptions, as follows.

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

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 $\hat{\gamma}$ is some reasonable estimator of the $\gamma$ direction satisfying a basic ‘high level’ condition: $||\hat{\gamma} - \gamma_0||_1 \leq a_n$ w.h.p. for some $\gamma_0 \propto \gamma$ and $a_n = o(1)$.

**Estimation of $\hat{\gamma}$.** Under Assumption 1.1 (a) and the SIM framework we have adopted here, $\mathbb{E}(Y \mid X) \equiv \mathbb{E}(Y \mid X, T = 1) = g(\gamma'X)$. 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; Yi et al., 2015; Yang, Balasubramanian and Liu, 2017) and apply it to the ‘complete case’ data $D_n^{(c)}$ to obtain a reasonable estimator $\hat{\gamma}$ of $\gamma$. Under some additional design restrictions and model assumptions, however, one may also estimate $\gamma$ by even simpler approaches, as follows.

(a) Suppose $Y$ satisfies the (slightly) stronger SIM formulation: $(Y \mid X) \equiv (Y \mid X, T = 1) = f(\gamma'X; \epsilon)$ for some unknown function $f : \mathbb{R}^2 \to \mathcal{Y}$ and some noise $\epsilon \perp \perp (T, X)$, and assume further that the distribution of $(X \mid T = 1)$ is elliptically symmetric. Then, owing to the results of Li and Duan (1989), one can still estimate $\gamma$ with a rate guarantee of $a_n = s_\gamma \sqrt{(\log p)/n}$ using a simple $L_1$ penalized ‘canonical’ link based regression (e.g. linear, logistic or Poisson regression) of $Y$ vs. $X$ in the ‘complete case’ data $D_n^{(c)}$, as discussed in the previous example. Approaches based on similar ideas 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; Plan, Vershynin and Yudovina, 2017; Wei, 2018).

(b) Suppose $Y$ satisfies the same SIM as in part (a) above, and assume now that the distribution of $X$ is elliptically symmetric. Then, combining the results of Li and Duan (1989) along with those in Section 1.3 regarding IPW
representations, it follows that one can estimate \( \gamma \) using an \( L_1 \)-penalized weighted regression based on any ‘canonical’ link (e.g. linear, logistic or Poisson regression) of \( Y \) vs. \( X \) in the ‘complete case’ data \( D_{n}^{(c)} \). The weights are given by \( \pi^{-1}(X) \), if \( \pi(\cdot) \) is known, or \( \hat{\pi}^{-1}(X) \) if \( \pi(\cdot) \) is unknown and estimated via \( \hat{\pi}(\cdot) \) (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 \( \hat{\gamma} \) satisfies an \( L_1 \) norm bound \( |\hat{\gamma} - \gamma|_1 \leq a_n \equiv s_\gamma \sqrt{(\log p)/n} \) w.h.p. in the case when \( \pi(\cdot) \) is known, and \( |\hat{\gamma} - \gamma|_1 \leq a_n \equiv s_\gamma \max\{ \sqrt{(\log p)/n}, \pi_n \sqrt{\log n} \} \) when \( \pi(\cdot) \) is unknown, where \( \pi_n = o(1) \) denotes the (pointwise) convergence rate of \( \hat{\pi}(\cdot) \). 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 the tail bounds and convergence rates for the estimators based on ‘extended’ parametric families, as discussed in Sections 5.1 and 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 \( Z \in \mathbb{R} \) be a generic random variable and \( X \in \mathbb{R}^p \) be a random vector of covariates with support \( \mathcal{X} \subseteq \mathbb{R}^p \). Consider an ‘extended’ parametric family of (working) models for estimating \( \mathbb{E}(Z|X) \) given by: \( g\{\beta(\Psi(X))\} \), where \( \Psi(X) \in \mathbb{R}^K \) denotes some vector of basis functions. Let \( \beta_0 \) denote the ‘target’ parameter corresponding to this working model and let \( \hat{\beta} \) be any corresponding estimator of \( \beta_0 \) based on any suitable estimation procedure applied to the observed data on \( \{Z_i, X_i\}_{i=1}^n \). Then we estimate \( \mathbb{E}(Z|X = x) \) based on the working model as: \( g\{\beta(\Psi(x))\} \). The following result establishes a tail bound and convergence rates for this estimator w.r.t. its target \( g\{\beta_0(\Psi(x))\} \).

**Theorem 5.1.** Suppose \( \hat{\beta} \) satisfies a high level guarantee: \( \|\hat{\beta} - \beta_0\|_1 \leq a_n \) with probability at least \( 1 - q_n \) for some sequences \( a_n \geq 0 \) and \( q_n \in [0, 1] \) such that \( a_n, q_n = o(1) \). Suppose further that \( g(\cdot) \) is Lipschitz continuous with \( |g(u) - g(v)| \leq C_g|u - v| \) for any \( u, v \in \mathbb{R} \) for some constant \( C_g \geq 0 \) and that \( \Psi(X) \) is uniformly bounded, i.e. \( \max_{1 \leq j \leq K} |\Psi_{ij}(X)| \leq C_{\Psi} < \infty \) a.s. \( \mathbb{P}_X \) for some constant \( C_{\Psi} \geq 0 \). Then, for any \( t \geq 0 \),

\[
\mathbb{P}\left[ \sup_{x \in \mathcal{X}} |g\{\beta(\Psi(x))\} - g\{\beta_0(\Psi(x))\}| > (\sqrt{2C_gC_\Psi})a_n t \right] \leq 2 \exp(-t^2) + q_n.
\]

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 the nuisance function estimators \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \), then the pointwise tail bounds required in our basic Assumptions 3.2 and 3.3 hold with the choices of \( \{v_{n,\pi}, v_{n,m}\} \propto a_n \), \( \{b_{n,\pi}, b_{n,m}\} = 0 \) 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 \( \hat{\beta} \) based on penalized estimators from high dimensional models, the \( L_1 \) error rate \( a_n \) should behave as: \( a_n \propto s_{\beta_0} \sqrt{(\log K)/n} \) with high probability.

5.4. High Dimensional Single Index Models: Non-Asymptotic Bounds and Rates for Kernel Smoothing over Estimated Index Parameters. In this section, we study the properties of single index kernel smoothing estimators with high dimensional covariates, where the index parameter is further allowed to be estimated. The underlying high dimensionality and the possibly non-ignorable index estimation error makes the analyses quite nuanced and different from most existing results available 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 they are obtained in course of characterizing our nuisance function estimators’ properties, they 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 \( \{(Z_i, X_i) : i = 1, \ldots, n\} \) denote a sample of \( n \geq 2 \) i.i.d. realizations of a generic random vector \((Z, X)\) assumed to have finite 2
d moments, where \( Z \in \mathbb{R}, X \in \mathbb{R}^p \) 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 \( \beta \in \mathbb{R}^p \) be any (unknown) ‘parameter’ vector of interest and let \( \hat{\beta} \) denote any reasonable estimator of \( \beta \) that satisfies a basic (‘high level’) \( L_1 \) error guarantee: \( \|\hat{\beta} - \beta\|_1 \leq a_n \) with high probability for some \( a_n = o(1) \).

Let \( W \equiv W_\beta := \beta'X, \hat{W} := \hat{\beta}'X, W_i \equiv W_{\beta,i} := \beta'X_i \) and \( \hat{W}_i := \hat{\beta}'X_i \) for \( i = 1, \ldots, n \), and for any \( x \in \mathbb{R}^p \), let \( w_x \equiv w_{x,\beta} := \beta'x \) and \( \hat{w}_x := \hat{\beta}'x \). Next, for any \( w \in \mathbb{R} \), let \( l_\beta(w) := m_\beta(w)f_\beta(w) \), where \( m_\beta(w) := \mathbb{E}(Z|W = w) \) and \( f_\beta(\cdot) \) denotes the density of \( W \equiv \beta'X \). Finally, for any \( x \in \mathcal{X} \), let \( m(\beta, x) := m_\beta(\beta'x), f(\beta, x) := f_\beta(\beta'x) \) and \( l(\beta, x) := l_\beta(\beta'x) \equiv m(\beta, x)f(\beta, x) \).

Given any estimator \( \hat{\beta} \) of \( \beta \), consider the following kernel smoothing (KS)
estimators of \(l(\beta, x), f(\beta, x)\) and \(m(\beta, x)\), for any fixed \(x \in X\), given by:

\[
\hat{\eta}(\beta, x) := \frac{1}{nh} \sum_{i=1}^{n} Z_i K \left( \frac{\beta' X_i - \hat{\beta}'}{h} \right) = \frac{1}{nh} \sum_{i=1}^{n} Z_i K \left( \frac{\hat{W}_i - \hat{w}_x}{h} \right),
\]

\[
\hat{f}(\beta, x) := \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{\beta' X_i - \hat{\beta}' X_i}{h} \right) \quad \text{and} \quad \hat{m}(\beta, x) := \frac{\hat{\eta}(\beta, x)}{\hat{f}(\beta, x)},
\]

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 \(h_n = o(1)\).

Note that \(\hat{f}(\cdot)\) is a special case of \(\hat{\eta}(\cdot)\) with \(Z \equiv 1\), and \(\hat{\eta}(\cdot)\) and \(\hat{f}(\cdot)\) are both linear kernel average (LKA) estimators while \(\hat{m}(\cdot)\) is a ratio type (Nadaraya-Watson) KS estimator. Below we obtain non-asymptotic tail bounds and (pointwise) convergence rates for these estimators. We mainly focus on the analysis of the LKA estimator \(\hat{\eta}(\cdot)\) from which the results for \(\hat{f}(\cdot)\) follow as a special case. These can be then combined to obtain results for the ratio estimator \(\hat{m}(\cdot)\). We summarize our assumptions first, followed by our results.

**Assumption 5.1 (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 \(m^{(2)}_\beta(w) := \mathbb{E}\{Z^2 | \beta' X = w\}\) for any \(w \in \mathbb{R}\). Then, \(m^{(2)}_\beta(w)f_\beta(w)\) is bounded in \(w \in \mathbb{R}\) and \(\|m^{(2)}_\beta(\cdot)f_\beta(\cdot)\|_{\infty} \leq B_1\) for some constant \(B_1 \geq 0\).

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

**Assumption 5.2 (Further conditions on \(K(\cdot)\) and other assumptions to account for the estimation error of \(\beta\)).** 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 \(\eta_\beta(w) := \mathbb{E}(ZX | \beta' X = w)f_\beta(w)\) for any \(w \in \mathbb{R}\), and let \(\eta_\beta[j](\cdot)\) denote the \(j^{th}\) coordinate of \(\eta_\beta(\cdot)\) for \(j = 1, \ldots, d\). Then, for each \(j\), \(\eta_\beta[j](\cdot)\)
is continuously differentiable with derivative \( \eta_{\beta j}(\cdot) \) that is bounded uniformly in \( j = 1, \ldots, d \). Further, \( l_\beta(\cdot) \) is also continuously differentiable with a bounded derivative \( l'_\beta(\cdot) \). Thus, \( \max_{1 \leq j \leq d} \| \eta_{\beta j}(\cdot) \|_\infty \leq B_1^* \) and \( \| l'_\beta(\cdot) \|_\infty \leq B_2^* \) 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 \( \varphi(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^+ \) 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) \( X \) is bounded, i.e. \( \| X \|_\infty \leq M_X \) a.s. \( \mathbb{P} \) for some constant \( M_X \geq 0 \), and the estimator \( \hat{\beta} \) satisfies the ‘high level’ guarantee: \( \| \hat{\beta} - \beta \|_1 \leq a_n \) with probability \( \geq 1 - q_n \), for some \( a_n, q_n \geq 0 \) with \( a_n = o(1) \) and \( q_n = o(1) \). Further, \( a_n/h = o(1) \) and \( 2M_X(a_n/h) \leq L \) where \( L \) is as in (c) above.

Most of the smoothness assumptions and the conditions on \( K(\cdot) \) in Assumptions 5.1 and 5.2 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 5.2 (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. \( |K''(u)| \leq |u|^{-\rho} \) 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 \( X \) 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 \( X \) (e.g. \( X \) is sub-Gaussian), although the corresponding technical analyses can be more involved.

**Theorem 5.2 (Tail bounds for the LKA estimators).** Consider the LKA estimator \( \hat{l}(\beta, x) \) of \( l(\beta, x) \). Suppose Assumptions 5.1 and 5.2 hold and that \( h = o(1), \log(np)/(nh) = o(1) \) and \( (a_n/h)\sqrt{\log p} = o(1) \). Then, for each fixed \( x \in \mathcal{X} \) and any \( t \geq 0 \), we have: with probability at least \( 1 - 9 \exp(-t^2) - 2q_n \),

\[
|\hat{l}(\beta, x) - l(\beta, x)| \leq C_1 \left( \frac{t + 1}{\sqrt{nh}} + \frac{t^2 \sqrt{\log n}}{nh} \right) + C_2 \left( h^2 + a_n + \frac{a_n^2}{h^2} + \frac{\log np}{nh} \right)
\]

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 \( \hat{l}(\beta, x) \) 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 \( \beta \) 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 \( \hat{\beta} \). 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 \( \hat{m}(\hat{\beta}, x) \) of \( m(\beta, x) \) and assume that \( |m(\beta, x)| \leq \delta_m \) and \( f(\beta, x) \geq \delta_f > 0 \) for some constants \( \delta_m, \delta_f > 0 \). For any \( t \geq 0 \), define:

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

and \( C_1, C_2, C_3 > 0 \) are the same constants as in Theorem 5.2. Assume that \( h = o(1), \log(np)/(nh) = o(1), (a_n/h)\sqrt{\log p} = o(1) \) and \( b_n = o(1) \). Then, under Assumptions 5.1 and 5.2, for any fixed \( x \in X \) 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 \( 1 - 18 \exp(-t^2) - 9 \exp(-t_*^2) - 6q_n \),

\[
|\hat{m}(\hat{\beta}, x) - m(\beta, x)| \leq \frac{2(1 + \delta_m)}{\delta_f} \epsilon_n(t) \lesssim \frac{t + 1}{\sqrt{nh}} + \frac{t^2}{nh} \log n + b_n,
\]

where ‘\( \lesssim \)' denotes inequality upto multiplicative constants (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:

\[
|\hat{m}(\hat{\beta}, x) - m(\beta, x)| \lesssim (c + 1) \sqrt{\log np} \left( 1 + c \sqrt{\log(np)\log n} \right) + b_n
\]

\[
\lesssim \frac{c\sqrt{\log np}}{\sqrt{nh}} + b_n \quad \text{with probability} \quad 1 - 27(np)^{-c^2} - 6q_n.
\]

Theorem 5.3 establishes explicit tail bounds and convergence rates for the ratio-type KS estimator \( \hat{m}(\hat{\beta}, x) \). As a consequence, it also verifies our basic Assumption 3.3 regarding \( \hat{m}(\cdot) \) when one chooses to estimate it based on a single index model. In particular, in view of the second part of Remark 3.2, it establishes that the tail bound in Assumption 3.3 holds with the choices \( v_{n,m} \propto v_n \equiv \sqrt{\log(np)/(nh)} + b_n, b_{n,m} \equiv 0 \) 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 \( \hat{\beta} \), the \( L_1 \) error rate \( a_n \) is expected to behave as: \( a_n \propto s_\beta \sqrt{\log p}/n \) with high probability, where \( s_\beta := \|\beta\|_0 \).
6. Simulation Studies. Results are available and will be added soon. Please check the slides for an overview of all the results.

7. Double Robustness of the DDR Estimator: Consistency Even under Misspecification of one of the Nuisance Function Estimates.

Our entire probabilistic analysis of \( \| T_n \|_\infty \equiv \| \nabla L^\text{DDR} \|_\infty \) (for establishing the convergence rates of \( \hat{\theta}_{\text{DDR}} \) in the light of Lemma 2.1) has so far assumed that the nuisance components \( \pi(\cdot) \) and \( m(\cdot) \) are both correctly estimated by \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \) respectively, as characterized in Assumptions 3.2 and 3.3. As noted in (1.4), the population version \( L^\text{DDR}(\cdot) \) of the empirical DDR loss \( L_n^\text{DDR}(\cdot) \) is designed in such a way that consistency of the resulting estimator should hold even if only one of \( \{ \hat{\pi}(\cdot), \hat{m}(\cdot) \} \), and not necessarily both, is correctly specified. But as mentioned earlier, the theoretical analyses (especially non-asymptotic bounds with sharp rate guarantees) under these general situations are quite involved and more importantly, will depend on the exact nature of the estimators and their first order properties (and rates), unlike the case when both are correctly specified and the analyses are general and ‘free’ (as discussed in Section 3) requiring no specific knowledge about the estimators except for some basic high level convergence properties. In this section, we briefly sketch the arguments that ensure consistency (at some reasonable rate, but not necessarily sharp) of \( \| T_n \|_\infty \) even if only one of \( \{ \hat{\pi}(\cdot), \hat{m}(\cdot) \} \) is correctly specified, but not necessarily both.

If \( \hat{\pi}(\cdot) \) is misspecified but \( \hat{m}(\cdot) \) is correctly specified. Suppose \( \hat{\pi}(\cdot) \overset{p}{\to} \pi^*(\cdot) \neq \pi(\cdot) \), while \( \hat{m}(\cdot) \overset{p}{\to} m(\cdot) \) still. In this case, the terms \( T_{\pi,n} \) and \( R_{\pi,m,n} \) in the basic decomposition (3.1) of \( T_n \) would be affected and need to be further decomposed appropriately into two terms each and analysed as follows.

\[
T_{\pi,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\hat{\pi}(X_i)} - \frac{T_i}{\pi^*(X_i)} \right\} \{ Y_i - m(X_i) \} h(X_i)
\]

(7.1)

\[
= \hat{T}_{\pi,n} \quad \text{(say)}
\]

\[
= \hat{T}_{\pi,n} \quad \text{(say)}
\]

\[
= : T_{\pi,n} \quad \text{(say)}
\]
\[ R_{\pi,m,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\hat{\pi}(X_i)} - \frac{T_i}{\pi^*(X_i)} \right\} \left\{ \hat{m}(X_i) - m(X_i) \right\} h(X_i) \]

\[ =: \tilde{R}_{\pi,m,n} \text{ (say)} \]

\[ \begin{align*}
  &+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi^*(X_i)} - \frac{T_i}{\pi(X_i)} \right\} \left\{ \hat{m}(X_i) - m(X_i) \right\} h(X_i) \\
  &=: R_{\pi,m,n}^* \text{ (say)} \]
\]

If \( \hat{m}(\cdot) \) is misspecified but \( \hat{\pi}(\cdot) \) is correctly specified. Now if \( \hat{m}(\cdot) \xrightarrow{p} m^*(\cdot) \neq m(\cdot), \) while \( \hat{\pi}(\cdot) \xrightarrow{p} \pi(\cdot) \) still. In this case, the terms \( T_{m,n} \) and \( R_{\pi,m,n} \) in the decomposition (3.1) of \( T_n \) would be affected and need to be further decomposed appropriately into two terms each and analysed as follows.

\[ T_{\pi,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi^*(X_i)} - 1 \right\} \left\{ \hat{m}(X_i) - m^*(X_i) \right\} h(X_i) \]

\[ =: T_{m,n} \text{ (say)} \]

\[ \begin{align*}
  &+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi^*(X_i)} - 1 \right\} \left\{ m^*(X_i) - m(X_i) \right\} h(X_i), \quad \text{and} \\
  &=: T_{m,n}^* \text{ (say)} \]
\]

\[ R_{\pi,m,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\hat{\pi}(X_i)} - \frac{T_i}{\pi(X_i)} \right\} \left\{ \hat{m}(X_i) - m^*(X_i) \right\} h(X_i) \]

\[ =: R_{m,n}^* \text{ (say)} \]

\[ \begin{align*}
  &+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\hat{\pi}(X_i)} - \frac{T_i}{\pi(X_i)} \right\} \left\{ m^*(X_i) - m(X_i) \right\} h(X_i) \\
  &=: R_{m,n}^* \text{ (say)} \]
\]

Formal results and the rest of the details for this section to be added soon. Please check the slides for an overview of the results.

**APPENDIX A: TECHNICAL TOOLS**

In this section, we collect some useful technical results, definitions and supporting lemmas, that would serve throughout as essential ingredients in the proofs of all our main results.
A.1. Orlicz Norms, Sub-Gaussians and Sub-Exponentials. We first introduce a few definitions and results regarding concentration bounds.

**Definition A.1 (Orlicz norms).** For any $\alpha > 0$, let $\psi_\alpha(x) = \exp(x^\alpha) - 1 \forall x \geq 0$. Then, for any random variable $X$ and any $\alpha > 0$, the $\psi_\alpha$-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 $\psi_\alpha$-Orlicz norm, denoted as $\|X\|_{\psi_\alpha} < \infty$ (if the set above is empty, then the infimum is simply defined to be $\infty$).

Further, for any random vector $X \in \mathbb{R}^d$ for any $d \geq 1$, we define $X$ to have finite $\psi_\alpha$-Orlicz norm if each coordinate of $X$ has finite $\psi_\alpha$-Orlicz norm, and we let $\|X\|_{\psi_\alpha} := \max\{\|X_j\|_{\psi_\alpha} : 1 \leq j \leq d\}$.

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

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 other definitions may be shown to be equivalent, up to constant factors in the parameters, to the one above. The $\psi_\alpha$-Orlicz norms are much more general norms allowing for any $\alpha > 0$ (not just $\alpha = 1$ 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 $\alpha > 0$ and hence, has finite $\psi_\alpha$-Orlicz norms with $\alpha = \infty$.

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

**Lemma A.1 (General properties of Orlicz norms, sub-Gaussians and sub-exponentials).** In the following, $X,Y \in \mathbb{R}$ denote generic random variables and $\mu$ denotes $\mathbb{E}(X) \in \mathbb{R}$.

(i) (Basic properties). For $\alpha \geq 1$, $\| \cdot \|_{\psi_\alpha}$ is a norm (and a quasinorm if $\alpha < 1$) satisfying: (a) $\|X\|_{\psi_\alpha} \geq 0$ and $\|X\|_{\psi_\alpha} = 0 \iff X = 0$ 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) (Monotonicity). (a) For any \( 0 < \alpha \leq \beta \), \( (\log 2)^{1/\alpha} \|X\|_{\psi_{\alpha}} \leq (\log 2)^{1/\beta} \|X\|_{\psi_{\beta}} \). In particular, \( \|X\|_{\psi_{1}} \leq (\log 2)^{-1/2} \|X\|_{\psi_{2}} \). (b) For any \( \alpha > 0 \), \( ||X||_{\psi_{\alpha}} \leq ||X||_{\psi_{\beta}} \). (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 \( \|X\|_{\psi_{\alpha}} \leq \sigma \) for some \( (\alpha, \sigma) > 0 \), then \( \forall \epsilon > 0 \), \( \mathbb{P}(|X| > \epsilon) \leq 2 \exp(-\epsilon^{\alpha}/\sigma^{\alpha}) \). (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 \( \mathbb{E}(|X|^{m}) \leq C_{\alpha}^{m} \sigma^{m} \Gamma(m/2 + 1) \) for each \( m \geq 1 \), 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 \( \mathbb{E}(|X|^{m}) \leq 2 \sigma^{m} \Gamma(m/2)^{m/2} \) for any \( m \geq 2 \).

(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 two random variables \( X \) and \( Y \) with \( \|X\|_{\psi_{\beta}} < \infty \) and \( \|Y\|_{\psi_{\gamma}} < \infty \), \( \|XY\|_{\psi_{\alpha}} \leq \|X\|_{\psi_{\alpha}} \|Y\|_{\psi_{\gamma}} \). In particular, for any two sub-Gaussians \( X \) and \( Y \), \( XY \) is sub-exponential and \( \|XY\|_{\psi_{1}} \leq \|X\|_{\psi_{2}} \|Y\|_{\psi_{2}} \). Moreover, if \( Y \leq M \) a.s. and \( \|X\|_{\psi_{\alpha}} < \infty \), 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 \( (\alpha, \sigma) > 0 \), and let \( Z_{n} := \max_{1 \leq i \leq n} |X_{i}| \). Then for any \( n \geq 1 \), \( \|Z_{n}\|_{\psi_{\alpha}} \leq \sigma (\log n + 2)^{1/\alpha} \leq \sigma \{3 \log(n + 1)\}^{1/\alpha} \), \( \mathbb{P}(Z_{n} > \epsilon) \leq 2n \exp(-\epsilon^{\alpha}/\sigma^{\alpha}) \) \( \forall \epsilon \geq 0 \), and \( \mathbb{P}(Z_{n} > c \sigma (\log n)^{1/\alpha}) \leq 2n^{-c^{2}+1} \forall c > 1 \).

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

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

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

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

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 A.3** (Sub-Gaussian properties of binary random variables). Let $Z \in \{0, 1\}$ be a binary random variable with $E(Z) \equiv P(Z = 1) = p \in [0, 1]$. Let $\bar{Z} = (Z - p)$ denote the corresponding centered version of $Z$. Then, $\|\bar{Z}\|_{\psi_2} \leq 2\tilde{p}$, where $\tilde{p} \geq 0$ is given by $\tilde{p} = 0$ if $p \in \{0, 1\}$, $\tilde{p} = 1/2$ if $p = 1/2$, and $\tilde{p} = [(p - 1/2)/\log(p/(1-p))]^{1/2}$ if $p \notin \{0, 1, 1/2\}$.

Lemma A.3 explicitly characterizes the sub-Gaussian properties of (centered) binary random variables and its proof can be found in Buldygin and Moskviychova (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 the bounds in Lemma A.2 are useful, they apply only to sub-Gaussians. However, Bernstein’s inequality applies more generally to sub-exponentials that include as special cases: sub-gaussians, bounded variables, as well as products of two sub-Gaussians and/or bounded variables (see Lemma A.5).

**Lemma A.4** (Bernstein’s inequality - adopted from Van de Geer and Lederer (2013)). Let $\{Z_1, \ldots, Z_n\}$ denote any collection of $n \geq 1$ independent (not necessarily i.i.d.) random variables such that $E(Z_i) = \mu_i \forall 1 \leq i \leq n$. Suppose $\sum$ constants $\sigma \geq 0$ and $K \geq 0$, such that $n^{-1} \sum_{i=1}^n E(|Z_i - \mu_i|^m) \leq (m!/2)\sigma^2 K^{m-2}$, for each positive integer $m \geq 2$. Then,

$$P \left( \left| \frac{1}{n} \sum_{i=1}^n (Z_i - \mu_i) \right| \geq \sqrt{2}\epsilon + K\epsilon^2 \right) \leq 2 \exp \left( -n\epsilon^2 \right) \quad \text{for any} \ \epsilon \geq 0.$$
In particular, if $\{Z_i\}_{i=1}^n$ are i.i.d. realizations of a sub-exponential variable $Z$ with mean $\mu$ 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 A.5** (The Bernstein moment conditions and their verification). Consider the moment conditions required in Bernstein’s inequality in Lemma A.4. Let us define any random variable by identical arguments as above we again have:

(a) For any $Z = XY$ with $X$ and $Y$ sub-Gaussian, $Z \sim \text{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$ with $\sigma_Z = \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

(b) For any $Z = XY$ with $X$ sub-exponential and $Y$ bounded by $M$ a.s., $Z \sim \text{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$ with $\sigma_Z = M\|X\|_{\psi_1}$.

**Proof.** If $\|Z\|_{\psi_1} \leq \sigma_Z$, then using Lemma A.1 (i)(c) and (iv)(a), $\|Z - \mu\|_{\psi_1} \leq 2\sigma_Z$ and $\mathbb{E}(\|Z - \mu\|^m) \leq (2\sigma_Z)^m\text{m!} \equiv (m/2)(2\sqrt{2}\sigma_Z)^2(2\sigma_Z)^{m-2}$ for each $m \geq 1$. Hence, by definition, $Z \sim \text{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$. Similarly, $\|Z\|_{\psi_1} = \|Z\|_{\psi_1} \leq \sigma_Z$ and $\mathbb{E}(\|Z\|_{\psi_1} - \mathbb{E}\{Z\}\|_{\psi_1} \leq 2\sigma_Z$. Therefore, by identical arguments as above we again have: $|Z| \sim \text{BMC}(2\sqrt{2}\sigma_Z, 2\sigma_Z)$.

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

**Lemma A.6** (Concentration bounds with variance in the leading term - adopted from Theorem 3.4 of Kuchibhotla and Chakrabortty (2018)). Suppose $\{X_i\}_{i=1}^n$ 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 $K_n > 0$,

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \|X_{ij}\|_{\psi_1} \leq K_n, \text{ and define } \Gamma_n := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_{ij}^2).$$

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

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

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

APPENDIX C: PROOF OF THEOREM 3.1

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

\[(C.1) \quad T_{0,n} = T_{0,n}^{(1)} + T_{0,n}^{(2)} \equiv \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(1)}(Z_{i}) + \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(2)}(Z_{i}), \quad \text{where} \]

$T_{0}^{(1)}(Z) := \{m(X) - g(X, \theta_{0})\}h(X)$ and $T_{0}^{(2)}(Z) := \frac{T}{\pi(X)}\{Y - m(X)\}h(X)$,

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

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

\[(C.2) \quad T_{0[j]}^{(1)}(Z) \equiv \psi(X)h_{0[j]}(X) \sim \text{BMC}(\bar{\sigma}_{1}, \bar{K}_{1}) \quad \forall j \in \{1, \ldots, d\}, \]

for some constants $\bar{\sigma}_{1} := 2\sqrt{2}\sigma_{\psi}\sigma_{h} \geq 0$ and $\bar{K}_{1} := 2\sigma_{\psi}\sigma_{h} \geq 0$.

Next, using Assumption 3.1 (a) and Lemma A.1 (v), $\|\epsilon(Z)h_{0[j]}(X)\|_{\psi_{1}} \leq \sigma_{\epsilon}\sigma_{h}$ for each $j \in \{1, \ldots, d\}$. Further, owing to Assumption 1.1 (b) and (1.1), $T/\pi(X) \leq \delta_{\pi}^{-1}$ a.s. $[\mathbb{P}]$. Hence, using Lemma A.5 (b), we have

\[(C.3) \quad T_{0[j]}^{(2)}(Z) \equiv \frac{T}{\pi(X)}\epsilon(Z)h_{0[j]}(X) \sim \text{BMC}(\bar{\sigma}_{2}, \bar{K}_{2}) \quad \forall j \in \{1, \ldots, d\}, \]

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

Hence, (C.2) and (C.3) ensure that for each $j \in \{1, \ldots, d\}$, $T_{0[j]}^{(1)}(Z)$ and $T_{0[j]}^{(2)}(Z)$ satisfy the required moment conditions for Bernstein’s inequality (Lemma A.4) to apply. Using Lemma A.4, we then have: for any $\epsilon_{1} \geq 0$,

\[
\mathbb{P}\left\{ \left\| T_{0,n}^{(1)} \right\|_{\infty} \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(1)}(Z_{i}) \right\|_{\infty} > \sqrt{2}\bar{\sigma}_{1} \epsilon_{1} + \bar{K}_{1} \epsilon_{1}^{2} \right\} \leq \sum_{j=1}^{d} \mathbb{P}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0[j]}^{(1)}(Z_{i}) \right\|_{\infty} > \sqrt{2}\bar{\sigma}_{1} \epsilon_{1} + \bar{K}_{1} \epsilon_{1}^{2} \right\}
\]

\[(C.4) \quad \leq \sum_{j=1}^{d} 2\exp\left(-n\epsilon_{1}^{2}\right) = 2d \exp\left(-n\epsilon_{1}^{2}\right) \equiv 2 \exp\left(-n\epsilon_{1}^{2} + \log d\right), \]
where the second step uses the union bound (u.b.). Similarly, for any \( \epsilon_2 \geq 0 \),
\[
\mathbb{P} \left\{ \left\| T_{0,n}^{(2)} \right\|_\infty \equiv \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(2)}(Z_i) \right\}_\infty > \sqrt{2}\sigma_2 \epsilon_2 + \bar{K} \epsilon_2^2 \right\}
\leq \sum_{j=1}^{d} \mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0,j}^{(2)}(Z_i) \right\|_\infty > \sqrt{2}\sigma_2 \epsilon_2 + \bar{K} \epsilon_2^2 \right\}
\]
(C.5)
\[
\leq \sum_{j=1}^{d} 2\exp \left( -n\epsilon_2^2 \right) = 2d \exp \left( -n\epsilon_2^2 \right) \leq 2 \exp \left( -n\epsilon_2^2 + \log d \right).
\]

Hence, setting \( \epsilon_1 = \epsilon_2 = \epsilon \) 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 (C.4)-(C.5) in the original decomposition (C.1) of \( T_{0,n} \), we have a tail bound for \( \| T_{0,n} \|_\infty \), as follows. For any \( \epsilon \geq 0 \),
\[
\mathbb{P} \left( \| T_{0,n} \|_\infty \equiv \left\| T_{0,n}^{(1)} + T_{0,n}^{(2)} \right\|_\infty > \sqrt{2}\sigma_0 \epsilon + K_0 \epsilon^2 \right)
\leq \mathbb{P} \left( \| T_{0,n}^{(1)} \|_\infty > \sqrt{2}\sigma_1 \epsilon + \bar{K}_1 \epsilon^2 \right) + \mathbb{P} \left( \| T_{0,n}^{(2)} \|_\infty > \sqrt{2}\sigma_2 \epsilon + \bar{K}_2 \epsilon^2 \right)
\]
(C.6)
\[
\leq 4 \exp \left( -n\epsilon^2 + \log d \right).
\]

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

**APPENDIX D: 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 \( \| T_{\pi,n} \|_\infty \).

**THEOREM D.1 (Tail bound for \( \| T_{\pi,n} \|_\infty \)).** Let Assumptions 1.1, 3.1 and 3.2 hold with the sequences \((v_{n,\pi}, b_{n,\pi}, q_{n,\pi})\) and the constants \((\delta_\pi, \sigma_\pi, C)\) being as defined therein. Let \( \| \mu_h^{(2)} \|_\infty := \max \{ E[h_j^{(2)}(X)] : j = 1, \ldots, d \} \). Then, for any \( \epsilon, \epsilon_1, \epsilon_2, \epsilon_3 \geq 0 \), with \( \epsilon_2 \) small enough such that \( (\epsilon_2 + b_{n,\pi}) < \delta_\pi \),
\[
\mathbb{P} \left( \| T_{\pi,n} \|_\infty > \epsilon \right) \leq 2 \exp \left\{ \frac{-n\epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} + \log d \right\} + 4 \exp \left( -n\epsilon_3^2 + \log d \right) + 2C \exp \left\{ \frac{-\epsilon^2}{v_{n,\pi}^2} + \log(n^2d) \right\} + 2C \exp \left\{ \frac{-\epsilon^2}{v_{n,\pi}^2} + \log(n^2d) \right\} + 4q_{n,\pi}(nd),
\]
where, for any choice of \((\epsilon_1, \epsilon_2, \epsilon_3)\) as above, \( d_n(\epsilon_1, \epsilon_2, \epsilon_3) \geq 0 \) is given by:
\[
d_n(\epsilon_1, \epsilon_2, \epsilon_3) := \frac{8\sigma_\pi^2(\epsilon_1 + b_{n,\pi})^2}{(\delta_\pi - (\epsilon_2 + b_{n,\pi}))^2} \left( \frac{\| \mu_h^{(2)} \|_\infty}{\delta_\pi} + \sqrt{2}\sigma_\pi \epsilon_3 + K_\pi \epsilon_3^2 \right),
\]
and \( \sigma_\pi, K_\pi \geq 0 \) are constants given by \( \sigma_\pi := 2\sqrt{2}\sigma_h^2 \delta_\pi^{-2} \) and \( K_\pi := 2\sigma_h^2 \delta_\pi^{-2} \).
D.1. Proof of Theorem D.1. Let $\mathcal{X}_n := \{(T_i, X_i) : i = 1, \ldots, n\}$. 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:

\begin{align}
\Delta_{\pi,n}(X) & := \hat{\pi}(X) - \pi(X), \quad \|\Delta_{\pi,n}\|_{\infty,n} := \max_{1 \leq i \leq n} |\Delta_{\pi,n}(X_i)|, \\
\bar{\pi}_n(X) & := - \frac{1}{\pi(X)} \quad \text{and} \quad \|\bar{\pi}_n\|_{\infty,n} := \max_{1 \leq i \leq n} |\bar{\pi}_n(X_i)|.
\end{align}

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

\begin{align}
\varphi_{[j]}(T, X) & := \frac{T}{\pi(X)} h_{[j]}(X), \quad \varphi^{(2)}_{n[j]}(X_n) := \frac{1}{n} \sum_{i=1}^{n} \varphi^{(2)}_{[j]}(T_i, X_i), \\
\mu^{(2)}_{n[j]} & := \mathbb{E} \left\{ \varphi^{(2)}_{[j]}(T, X) \right\} \equiv \mathbb{E} \left\{ \varphi^{(2)}_{n[j]}(X_n) \right\} \quad \text{and} \quad \mu^{(2)}_{n[j]} := \mathbb{E} \left\{ h^2_{[j]}(X) \right\}.
\end{align}

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

\begin{align}
T_{\pi}(Z) = \Delta_{\pi,n}(X) \bar{\pi}_n(X) \varphi(T, X) \varepsilon(Z), \quad \text{where}
\end{align}

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

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

\begin{align}
\|T_{\pi}[j](Z_i) \mid \mathcal{X}_n\|_{\psi_2} & = \|\Delta_{\pi,n}(X_i) \bar{\pi}_n(X_i) \varphi_{[j]}(T_i, X_i) \varepsilon(Z_i) \mid \mathcal{X}_n\|_{\psi_2} \\
& \leq \Delta_{\pi,n}(X_i) \bar{\pi}_n(X_i) \varphi_{[j]}(T_i, X_i) \sigma_\varepsilon(X_i) \leq \sigma_\varepsilon \|\Delta_{\pi,n}\|_{\infty,n} \|\bar{\pi}_n\|_{\infty,n} \varphi\right|^{(2)}_{n[j]}(T_i, X_i).
\end{align}

Further, $\forall 1 \leq j \leq d$, $\{T_{\pi}[j](Z_i)\}_{i=1}^{n} \mid \mathcal{X}_n$ are (conditionally) independent and centered random variables. Hence, using Lemma A.2, we have: $\forall 1 \leq j \leq d,$

\begin{align}
\left\| \frac{1}{n} \sum_{i=1}^{n} T_{\pi}[j](Z_i) \mid \mathcal{X}_n \right\|_{\psi_2} \leq \frac{4c_{n,j}(\mathcal{X}_n)}{\sqrt{n}}, \quad \text{where}
\end{align}

\begin{align}
c_{n,j}(\mathcal{X}_n) & := \sigma_\varepsilon \|\Delta_{\pi,n}\|_{\infty,n} \|\bar{\pi}_n\|_{\infty,n} \left( \varphi\right|^{(2)}_{n[j]} \right)^{1/2}
\end{align}

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

\begin{align}
\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} T_{\pi}[j](Z_i) \right| > \epsilon \mid \mathcal{X}_n \right\} \leq 2 \exp \left\{ - \frac{n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)} \right\} \quad \forall 1 \leq j \leq d.
\end{align}
The fundamental bound for \( \|T_{\pi,n}\|_{\infty} \). Using (D.7), the union bound (u.b.) and the law of iterated expectations (l.i.e.), we then have: for any \( \epsilon \geq 0 \),

\[
\mathbb{P}\left\{ \|T_{\pi,n}\|_{\infty} \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} T_{\pi}(Z_i) \right\|_{\infty} > \epsilon \right\} \\
\leq \sum_{j=1}^{d} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} T_{\pi[j]}(Z_i) \right| > \epsilon \right\} \text{ [using the u.b.]} \\
= \sum_{j=1}^{d} \mathbb{E}_{X_n} \left[ \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} T_{\pi[j]}(Z_i) \right| > \epsilon \left| X_n \right. \right\} \right] \text{ [using the l.i.e.]} ,
\]

(D.8)

\[
\leq \sum_{j=1}^{d} 2 \mathbb{E}_{X_n} \left[ \exp \left\{ \frac{-n \epsilon^2}{8c_{n,j}^2(X_n)} \right\} \right] \text{ [using (D.7)].} \quad \blacksquare
\]

Next, we aim to control the behavior of the random variable \( c_{n,j}^2(X_n) \) appearing in the bound (D.8). Based on the definition of \( c_{n,j}(X_n) \) in (D.6), it suffices to separately control the variables \( \|\Delta_{\pi,n}\|_{\infty,n}, \|\tilde{\pi}_n\|_{\infty,n}^2 \) and \( \varphi_{n[j]}^{(2)} \).

Controlling \( \|\Delta_{\pi,n}\|_{\infty,n}^2 \). Owing to the bound (3.6) in Assumption 3.2 and recalling all notations defined in (D.1)-(D.2), we have: for any \( \epsilon_1 \geq 0 \),

\[
\mathbb{P}\left[ \|\Delta_{\pi,n}\|_{\infty,n}^2 \equiv \left\{ \max_{1 \leq i \leq n} |\Delta_{\pi,n}(X_i)| \right\}^2 > (\epsilon_1 + b_{n,\pi})^2 \right] \\
\leq \sum_{i=1}^{n} \mathbb{P}\left\{ |\hat{\pi}(X_i) - \pi(X_i)| > \epsilon_1 + b_{n,\pi} \right\} \text{ [using the u.b.]} ,
\]

(D.9)

\[
\leq C n \exp \left( \frac{-\epsilon_1^2}{b_{n,\pi}^2} \right) + nq_{n,\pi} \text{ [using (3.6)].} \quad \blacksquare
\]
Controlling $\|\hat{\pi}_n\|_{\infty,n}^2$. Using similar arguments, along with (1.1), we have: \(\forall \epsilon_2 \geq 0\) small enough such that \((\epsilon_2 + b_{n,\pi}) < \delta_{\pi}\) with \(\delta_{\pi}\) as in (1.1),

\[
\mathbb{P}\left[\|\hat{\pi}_n\|_{\infty,n}^2 \geq \left\{\max_{1 \leq i \leq n} |\hat{\pi}_n(X_i)|\right\}_i^2 > \{\delta_{\pi} - (\epsilon_2 + b_{n,\pi})\}^{-2}\right] \\
\leq \sum_{i=1}^{n} \mathbb{P}\left[\hat{\pi}_n(X_i) > \{\delta_{\pi} - (\epsilon_2 + b_{n,\pi})\}^{-1}\right] \text{ [using the u.b.]} \\
\leq \sum_{i=1}^{n} \mathbb{P}\{\hat{\pi}(X_i) < \pi(X_i) - (\epsilon_2 + b_{n,\pi})\} \text{ [using (1.1)]}, \\
\leq \sum_{i=1}^{n} \mathbb{P}\{|\hat{\pi}(X_i) - \pi(X_i)| > \epsilon_2 + b_{n,\pi}\}
\]

\[\text{(D.10)} \quad \leq Cn \exp\left(-\frac{\epsilon_2^2}{\delta_{\pi}^2}\right) + n\epsilon_{n,\pi} \text{ [using (3.6)].} \]

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

\[\text{(D.11)} \quad \varphi_{n[j]}^{(2)}(T, X) \equiv \frac{T^2}{\pi^2(X)}h_{n[j]}^2(X) \sim \text{BMC}(\sigma_{\varphi}, K_{\varphi}) \quad \text{and further,}
\]

\[\text{(D.12)} \quad \mu_{\varphi_{n[j]}^{(2)}} \equiv \mathbb{E}\left\{\varphi_{n[j]}^{(2)}(T, X)\right\} = \mathbb{E}\left\{h_{n[j]}^2(X)\right\} = \frac{\mu_{h_{n[j]}^{(2)}}}{\delta_{\pi}} \leq \frac{\|\mu_{h_{n[j]}^{(2)}}\|_{\infty}}{\delta_{\pi}},
\]

where $\|\mu_{h_{n[j]}^{(2)}}\|_{\infty} := \max\{\mu_{h_{n[j]}^{(2)}} : j = 1, \ldots, d\} < \infty$ and $\mu_{h_{n[j]}^{(2)}}$ is as in (D.4).

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

\[
\mathbb{P}\left\{\varphi_{n[j]}^{(2)} \equiv \frac{1}{n} \sum_{i=1}^{n} \varphi_{n[j]}^{(2)}(T_i, X_i) > \frac{\|\mu_{h_{n[j]}^{(2)}}\|_{\infty}}{\delta_{\pi}} + \sqrt{2}\sigma_{\varphi}\epsilon_3 + \bar{K}_{\varphi}\epsilon_3^2\right\} \\
\leq \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^{n} \varphi_{n[j]}^{(2)}(T_i, X_i) - \mu_{\varphi_{n[j]}^{(2)}}\right| > \sqrt{2}\sigma_{\varphi}\epsilon_3 + \bar{K}_{\varphi}\epsilon_3^2\right\}
\]

\[\text{(D.13)} \quad \leq 2 \exp\left(-n\epsilon_3^2\right). \]
For any \( \epsilon_1, \epsilon_3 > 0 \), and any \( \epsilon_2 > 0 \) such that \( (\epsilon_2 + b_{n, \pi}) < \delta \), let us now define the event \( A_{\pi, n, j}(\epsilon_1, \epsilon_2, \epsilon_3) \), for each \( j \in \{1, \ldots, d\} \), as follows.

\[
A_{\pi, n, j}(\epsilon_1, \epsilon_2, \epsilon_3) := \left\{ 8c^2_{n,j} (X_n) > d_n(\epsilon_1, \epsilon_2, \epsilon_3) \right\}, \quad 1 \leq j \leq d,
\]

where

\[
d_n(\epsilon_1, \epsilon_2, \epsilon_3) := \frac{8\sigma^2(\epsilon_1 + b_{n, \pi})}{\delta - (\epsilon_2 + b_{n, \pi})} \left( \frac{\|\mu^{(2)}_n\|_\infty}{\delta} + \sqrt{2\varphi \epsilon_3 + K_\varphi \epsilon_3^2} \right).
\]

Then, recalling from (D.6) that \( c^2_{n, j}(X_n) \equiv \sigma^2 \|\Delta_{\pi,n}\|_{\infty,n}^2 \right\| \pi_n \|_{\infty,n}^2 \varphi_{n[j]} \right\| \}

and using the bounds (D.9), (D.10) and (D.13) for \( \|\Delta_{\pi,n}\|_{\infty,n}^2 \), \( \|\pi_n\|_{\infty,n}^2 \), and \( \varphi_{n[j]}^{(2)} \) respectively, along with the union bound, we have:

\[
P(A_{\pi, n, j}) \equiv P_X(n, A_{\pi, n, j}) \equiv P_X(n, \{8c^2_{n,j} (X_n) > d_n(\epsilon_1, \epsilon_2, \epsilon_3) \}) \]

\[
\leq Cn \exp \left( -\frac{\epsilon_1^2}{v_{n, \pi}^2} \right) + Cn \exp \left( -\frac{\epsilon_2^2}{v_{n, \pi}^2} \right) + 2n \epsilon_3 + 2 \epsilon_3 \delta.
\]

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

\[
E_X \left[ \exp \left\{ \frac{-n \epsilon^2}{8c^2_{n,j}(X_n)} \right\} \right] = E \left[ \exp \left\{ \frac{-n \epsilon^2}{8c^2_{n,j}(X_n)} \right\} \mid A_{\pi, n, j}^c \right] P(A_{\pi, n, j}^c)
\]

\[
+ E \left[ \exp \left\{ \frac{-n \epsilon^2}{8c^2_{n,j}(X_n)} \right\} \mid A_{\pi, n, j} \right] P(A_{\pi, n, j})
\]

\[
\leq \exp \left( \frac{-n \epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} \right) + 2 \exp (-n \epsilon_3^2) + 2n \epsilon_3 + Cn \exp \left( \frac{-\epsilon_1^2}{v_{n, \pi}^2} \right) + Cn \exp \left( \frac{-\epsilon_2^2}{v_{n, \pi}^2} \right) \quad \text{[using (D.14)-(D.15)].}
\]

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

\[
P \left( \|T_{\pi, n}\|_{\infty} > \epsilon \right) \leq \sum_{j=1}^{d} 2E_X \left[ \exp \left\{ \frac{-n \epsilon^2}{8c^2_{n,j}(X_n)} \right\} \right]
\]

\[
\leq 2d \exp \left( \frac{-n \epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} \right) + 4d \exp (-n \epsilon_3^2) + 4n \epsilon_3 \delta + 4q_n \epsilon_3 (n d)
\]

\[
+ 2C(n d) \exp \left( \frac{-\epsilon_1^2}{v_{n, \pi}^2} \right) + 2C(n d) \exp \left( \frac{-\epsilon_2^2}{v_{n, \pi}^2} \right) \quad \text{[using (D.16)],}
\]

\[
\leq 2 \exp \left( \frac{-n \epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} \right) + 4 \exp (-n \epsilon_3^2 + \log d) + 4n \epsilon_3 \delta + 4q_n \epsilon_3 \delta
\]

\[
+ 2C \exp \left( \frac{-\epsilon_1^2}{v_{n, \pi}^2} \right) + 4C \exp \left( \frac{-\epsilon_2^2}{v_{n, \pi}^2} + \log(n d) \right).
\]
This proves the desired bound and completes the proof of Theorem D.1.

**D.2. Completing Proof of Theorem 3.2.** We next evaluate the general tail bound for \( \|T_{\pi,n}\|_\infty \) in Theorem D.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 \( \|T_{\pi,n}\|_\infty \). Let \((c_1, c_2, c_3) > 1\) be any universal constants, and set \(\epsilon_1 = c_1 v_n,\pi \sqrt{\log(n)}/d\), \(\epsilon_2 = c_2 v_n,\pi \sqrt{\log(n)}/d\) and \(\epsilon_3 = c_3 \sqrt{\log(d)/n}\). Further, we also assume w.l.o.g. that \(\epsilon_3 < 1\) and \((\epsilon_2 + b_n,\pi) \leq \delta_\pi/2 < \delta_\pi\), so that \(\epsilon_2\) satisfies the minor requirement in Theorem D.1. Then,

\[
\begin{align*}
\epsilon_1 + b_n,\pi &\equiv c_1 v_n,\pi \sqrt{\log(n)/d} + b_n,\pi \leq c_1 \{v_n,\pi \sqrt{\log(n)/d} + b_n,\pi\} \quad \text{and} \\
\epsilon_2 + b_n,\pi &\equiv c_2 v_n,\pi \sqrt{\log(n)/d} + b_n,\pi \leq \delta_\pi/2 \quad \text{(by choice),}
\end{align*}
\]

so that \(\{\delta_\pi - (\epsilon_2 + b_n,\pi)\} \geq \delta_\pi/2\). Further, with a choice of \(\epsilon_3\) as above,

\[
\|\mu_{h}^{(2)}\|_\infty + \sqrt{2\sigma_\varphi \epsilon_3 + K_\varphi \epsilon_3^2} \leq \|\mu_{h}^{(2)}\|_\infty + \left(\sqrt{2\hat{\sigma}_\varphi + K_\varphi}\right) c_3 \sqrt{\frac{\log d}{n}}.
\]

Combining and using all the inequalities above in the definition (D.14) of \(d_n(\epsilon_1, \epsilon_2, \epsilon_3)\), and letting \(C_\varphi := (\sqrt{2\hat{\sigma}_\varphi + K_\varphi})\), we have:

\[
d_n(\epsilon_1, \epsilon_2, \epsilon_3) \leq 8\hat{\sigma}_\varphi^2 \frac{4c_1^2}{\delta_\pi^2} \{v_n,\pi \sqrt{\log(n)/d} + b_n,\pi\}^2 \left(\frac{\|\mu_{h}^{(2)}\|_\infty}{\delta_\pi} + c_3 C_\varphi \sqrt{\frac{\log d}{n}}\right).
\]

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 D.1, we then have:

With probability \(\geq 1 - \frac{2}{d^{c^2 - 1} - 4d^{c^2 - 1} \left(\sum_{j=1}^{2} \frac{2C}{(nd)^{c^2 - 1} - 4q_{n,\pi}(nd)}\right)}\),

\[
\|T_{\pi,n}\|_\infty \leq c \sqrt{\frac{\log d}{n}} \{v_n,\pi \sqrt{\log(n)/d} + b_n,\pi\} C_1 \left(\frac{\|\mu_{h}^{(2)}\|_\infty}{\delta_\pi} + C_2 \sqrt{\frac{\log d}{n}}\right)^{1/2},
\]

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

**APPENDIX E: 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 \(\|T_{m,n}\|_\infty\).
THEOREM E.1 (Tail bound for \(\|T_{m,n}\|_\infty\)). Let Assumptions 1.1, 3.1 (a) and 3.3 hold, with the sequences \((v_{b,m}, b_{\bar{n},m}, q_{\bar{n},m})\), \(\bar{n} = n/2\) and the constants \((\delta_\pi, C)\) being as defined therein. Then, for any \(\epsilon, \epsilon_1, \epsilon_2 \geq 0\),

\[
\mathbb{P}\left(\|T_{m,n}\|_\infty > \epsilon\right) \leq 4 \exp\left\{\frac{-\bar{n} \epsilon^2}{\hat{t}_n(\epsilon_1, \epsilon_2)} + \log d\right\} + 8 \exp(-\bar{n} \epsilon_2^2 + \log d) + 4C \exp\left\{\frac{-\epsilon_2^2}{\hat{v}_n(\epsilon_1, \epsilon_2)} + \log(\bar{n}d)\right\} + 4q_{\bar{n},m}(\bar{n}d),
\]

where

\[
h(\epsilon_1, \epsilon_2) := 8 \delta_\pi^2 (\epsilon_1 + b_{\bar{n},m})^2 \left(\|\mu_\pi(2)\|_\infty + \sqrt{2\sigma_m \epsilon_2} + K_m \epsilon_2^2\right),
\]

with \(\|\mu_\pi(2)\|_\infty := \max\{\mathbb{E}\{h_j^2(X)\} : j = 1, \ldots, d\}, \) and \(\delta_\pi, \sigma_m, K_m \geq 0\) are constants with \(\delta_\pi \leq \delta_\pi^{-1}\) and \(\sigma_m := 2\sqrt{2\sigma_m^2} \) and \(K_m := 2\sigma_m^2\).

E.1. Proof of Theorem E.1. We first rewrite \(T_{m,n}\) from (3.1) as:

\[
T_{m,n} \equiv \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi(X_i)} - 1 \right\} \left\{ \tilde{m}(X_i) - m(X_i) \right\} h(X_i)
\]

\[= \frac{1}{2n} \sum_{k \neq k'} \sum_{i \in I_{k'}} \left\{ \frac{T_i}{\pi(X_i)} - 1 \right\} \left\{ \tilde{m}^{(k)}(X_i) - m(X_i) \right\} h(X_i)
\]

(E.1) \[= \frac{1}{2} \sum_{k \neq k'} T^{(k,k')}_{m,n}, \text{ where } T^{(k,k')}_{m,n} := \frac{1}{n} \sum_{i \in I_{k'}} T^{(k)}_m(Z_i) \text{ and } T^{(k)}_m(Z) := \left\{ \frac{T}{\pi(X)} - 1 \right\} \left\{ \tilde{m}^{(k)}(X) - m(X) \right\} h(X) \quad \forall k \neq k' \in \{1, 2\}.
\]

Define \(X_{n,k}^* := \{X_i : i \in I_k\} \forall k \in \{1, 2\}\), and let \(\mathbb{E}_{X_{n,k}^*}^* (\cdot)\) and \(\mathbb{P}(\cdot \mid X_{n,k}^* )\) respectively denote expectation w.r.t. \(X_{n,k}^*\) and conditional probability given \(X_{n,k}^*\). Further, for each \(k \neq k' \in \{1, 2\}\), let \(\mathbb{E}_{D_n^{(k)}}, X_{n,k'}^* (\cdot)\) and \(\mathbb{P}(\cdot \mid D_n^{(k)}, X_{n,k'}^* )\) respectively denote expectation w.r.t. \(\{D_n^{(k)}, X_{n,k'}^*\}\) and conditional probability given \(\{D_n^{(k)}, X_{n,k'}^*\}\). With \(D_n^{(k)} \perp X_{n,k'}^* \forall k \neq k' \in \{1, 2\}\), we note that \(\mathbb{E}_{D_n^{(k)}, X_{n,k'}^* }^* (\cdot) = \mathbb{E}_{X_{n,k'}^* }^* (\mathbb{E}_{D_n^{(k)}} (\cdot))\). Next, let us define: \(\forall k \neq k' \in \{1, 2\}\),

\[
\Delta_{m,n}^{(k)}(X) := \tilde{m}^{(k)}(X) - m(X), \quad \left\| \Delta_{m,n}^{(k,k')} \right\|_{\infty, n} := \max_{i \in I_{k'}} \left| \Delta_{m,n}^{(k)}(X_i) \right|
\]

(E.2) \[\hat{h}_{\bar{n},[j]}^{(2,k')}(X_i) := \frac{1}{n} \sum_{i \in I_{k'}} h_{\bar{n}}^2(X_i) \] and let \(\psi(T, X) := \frac{T}{\pi(X)} - 1\).
Further, conditional on for each \( \parallel \cdot \parallel \) along with (E.1)-(E.3), we have:

Lemma A.1 (i)(b) along with the definitions of \( \bar{\pi} \) is decreasing in \( a \equiv \pi(X) \) and \( a \equiv \delta_\pi \) respectively, with \( \delta_\pi \) being as in (1.1). We note that \( a \) is decreasing in \( a \in (0,1] \) and \( a \leq 1/2 \), so that \( a \leq 1/a \) \( \forall a \in (0,1] \). Using this and (1.1), we therefore have: \( \bar{\pi}(x) \leq \delta_\pi \leq 1/\delta_\pi \) \( \forall x \in \mathcal{X} \).

Using the notations from (E.2) and (E.3), we have: for each \( \{ \psi \} \),

\[
T_m(k) \equiv \left\{ \frac{T}{\pi(X)} - 1 \right\} \{ \bar{m}(k)(X) - m(X) \} \mathbf{h}(X) = \psi(T, X) \Delta_m(k)(X) \mathbf{h}(X).
\]

Now, for each \( k \in \{ 1,2 \} \) and \( k' \neq k \in \{ 1,2 \} \), \( D_n(k) \perp X_{n,k'} \) and we have:

\( \{ \psi(T_i, X_i) | D_n(k), X_{n,k'} \} \in \mathcal{I}_{k'} \equiv \{ \psi(T_i, X_i) | X_{n,k'} \} \equiv \{ \psi(T_i, X_i) | X_i \}_{i \in \mathcal{I}_{k'}} \) are (conditionally) independent sub-Gaussian random variables that satisfy:

\[
\forall i \in \mathcal{I}_{k'}, \quad \mathbb{E} \{ \psi(T_i, X_i) | D_n(k), X_{n,k'} \} = \mathbb{E} \{ \psi(T_i, X_i) | X_i \} = 0 \quad \text{and} \quad \| \psi(T_i, X_i) | D_n(k), X_{n,k'} \| \psi_2 = \| \psi(T_i, X_i) | X_i \| \psi_2 \leq \bar{\pi}^2 \leq \delta_\pi^2,
\]

where the bounds on the \( \| \cdot \| \psi_2 \) norm follow from using Lemma A.3 and Lemma A.1 (i)(b) along with the definitions of \( \bar{\pi}(\cdot) \) and \( \delta_\pi \) given earlier. Further, conditional on \( D_n(k) \) and \( X_{n,k'} \), \( \{ \Delta_m(k)(X_i) \}_{i \in \mathcal{I}_{k'}} \) and \( \{ \mathbf{h}(X_i) \}_{i \in \mathcal{I}_{k'}} \), for each \( j \in \{ 1, \ldots, d \} \), are all constants. Hence, using Lemma A.2 and (E.4), along with (E.1)-(E.3), we have: \( \forall k \neq k' \in \{ 1,2 \} \) and \( j \in \{ 1, \ldots, d \} \),

\[
\| \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} T_{m[j]}(Z_i) | D_n(k), X_{n,k'} \| \psi_2 \leq \frac{4d_{n,j} (D_n(k), X_{n,k'})}{\sqrt{n}}, \quad \text{where}
\]

\[
d_{n,j} (D_n(k), X_{n,k'}) \equiv \delta_\pi \| \Delta_m(k)(X_i) \|_{\infty, \bar{\pi}} \| \mathbf{h}^{(k,k')}_{n[j]} \|_{1/2}.
\]

Using Lemma A.2, we then have: \( \forall k \neq k' \in \{ 1,2 \} \), \( 1 \leq j \leq d \) and \( \epsilon \geq 0 \),

\[
\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} T_{m[j]}(Z_i) \right| > \epsilon | D_n(k), X_{n,k'} \right\} \leq 2 \exp \left\{ \frac{-\bar{\pi}^2}{8d_{n,j} (D_n(k), X_{n,k'})} \right\}.
\]

The fundamental bound for \( \| T_{m,n}^{(k,k')} \|_{\infty} \). Using the bound obtained above for \( T_{m,n}^{(k,k')} | D_n(k), X_{n,k'} \), we then have the following (unconditional) proba-
bilistic bound for $\|T^{(k,k')}_{m,n}\|_\infty$. For any $\epsilon \geq 0$ and $k \neq k' \in \{1, 2\}$,

$$
\mathbb{P}\left\{ \frac{1}{\bar{n}} \sum_{i \in I_{k'}} T^{(k)}_{m}(Z_i) > \epsilon \right\} \\
\leq \sum_{j=1}^{d} \mathbb{P}\left\{ \frac{1}{\bar{n}} \sum_{i \in I_{k'}} T^{(k)}_{m[j]}(Z_i) > \epsilon \right\} [\text{using the u.b.}],
$$

$$
= \sum_{j=1}^{d} \mathbb{E}_{D^{(k)}_{n}} X_{n,k'}^{*} \left[ \mathbb{P}\left\{ \frac{1}{\bar{n}} \sum_{i \in I_{k'}} T^{(k)}_{m[j]}(Z_i) > \epsilon \right| D_{n}^{(k)}, X_{n,k'}^{*} \right]\right]\right].
$$

(E.6)

Next, we aim to control the random variable $d_{\bar{n},j}^{2}(D^{(k)}_{n}, \chi_{n,k'})$ appearing in (E.6). Based on the definition (E.5) of $d_{\bar{n},j}^{2}(D^{(k)}_{n}, \chi_{n,k'})$, it suffices to separately control $\Delta^{(k,k')}_{m,n}$. To this end, let $\mathbb{E}_{D^{(k)}_{n}}(\cdot)$ and $\mathbb{P}_{D^{(k)}_{n}}(\cdot)$ denote expectation and probability w.r.t. $D^{(k)}_{n}$ for each $k \neq k' \in \{1, 2\}$.

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

Controlling $\|\Delta^{(k,k')}_{m,n}\|_\infty^2$. Using Assumption 3.3 along with the notations and facts discussed above, we have: for any $k \neq k' \in \{1, 2\}$ and any $\epsilon_1 \geq 0$,

$$
\mathbb{P}\left[ \|\Delta^{(k,k')}_{m,n}\|_\infty^2 \geq \left( \epsilon_1 + b_{\bar{n},m} \right)^2 \right] \\
\leq \sum_{i \in I_{k'}} \mathbb{P}\left\{ \max_{i \in I_{k'}} \Delta^{(k)}_{m,n}(X_i) \geq \epsilon_1 + b_{\bar{n},m} \right\} [\text{using the u.b.}],
$$

$$
\leq \sum_{i \in I_{k'}} \mathbb{E}_{X_{n,k'}} \left\{ C \exp \left( -\epsilon_1^2 \frac{1}{\epsilon_1^2} \right) + q_{\bar{n},m} \right\} [\text{using (3.8)}],
$$

(E.7)

$$
\equiv C\bar{n} \exp \left( -\epsilon_1^2 \frac{1}{\epsilon_1^2} \right) + \bar{n}q_{\bar{n},m}.
$$
where we also used that $D_n^{(k)} \perp \mathcal{X}_{n,k}'$ which ensures $\mathbb{P}_{D_n^{(k)}_n}(\cdot | \mathcal{X}_{n,k}') = \mathbb{P}_{D_n^{(k)}_n}(\cdot)$ and makes (3.8) in Assumption 4.3 applicable conditional on $\mathcal{X}_{n,k}'$. \hfill \blacksquare

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

(E.8) $\mathbb{P} \left\{ \bar{h}_{n[i]}^{(2,k')} \equiv \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} h_{j}^{2}(X_i) > \| \mu_h^{(2)} \|_{\infty} + \sqrt{2}\bar{\sigma}_h\epsilon_2 + \bar{K}_h\epsilon_2^2 \right\}
\leq \mathbb{P} \left\{ \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} h_{j}^{2}(X_i) - \mu_{h[j]}^{(2)} > \sqrt{2}\bar{\sigma}_h\epsilon_2 + \bar{K}_h\epsilon_2^2 \right\} \leq 2\exp(-n\epsilon_2^2). \hfill \blacksquare$

The final bound for $\| T_{m,n}^{(k,k')} \|_{\infty}$. For any $\epsilon_1, \epsilon_2 > 0$, let us now define:

(E.9) $t_{\bar{n}}(\epsilon_1, \epsilon_2) := 8d_{\bar{n}}^2 (\epsilon_1 + b_{\bar{n},m})^2 \left( \| \mu_h^{(2)} \|_{\infty} + \sqrt{2}\bar{\sigma}_h\epsilon_2 + \bar{K}_h\epsilon_2^2 \right)$.

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

(E.10) $\mathbb{P} \left\{ 8d_{\bar{n},j}^2(D_n^{(k)}, \mathcal{X}_{n,k}') > t_{\bar{n}}(\epsilon_1, \epsilon_2) \right\}$
\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).
Using (E.10) in the fundamental bound (E.6) for $\|T_{m,n}^{(k,k')}\|_\infty$, we then have: for each $k \neq k' \in \{1, 2\}$ and for any $\epsilon, \epsilon_1, \epsilon_2 \geq 0$,

$$\mathbb{P}\left\{ \|T_{m,n}^{(k,k')}\|_\infty > \epsilon \right\} \leq 2 \sum_{j=1}^{d} \mathbb{E} D_{n,j}^{(k),X^{*}_{n,k'}} \left\{ \exp \left\{ \frac{-\bar{n} \epsilon^2}{8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right)} \right\} \right\}$$

$$\equiv 2 \sum_{j=1}^{d} \mathbb{E} \left\{ \exp \left\{ \frac{-\bar{n} \epsilon^2}{8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right)} \right\} \right\} \left\{ 8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right) \leq t_n(\epsilon_1, \epsilon_2) \right\}$$

$$+ 2 \sum_{j=1}^{d} \mathbb{E} \left\{ \exp \left\{ \frac{-\bar{n} \epsilon^2}{8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right)} \right\} \right\} \left\{ 8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right) > t_n(\epsilon_1, \epsilon_2) \right\}$$

$$\leq 2d \left\{ \exp \left\{ \frac{-\bar{n} \epsilon^2}{t_n(\epsilon_1, \epsilon_2)} \right\} \right\} + \mathbb{P} \left\{ 8d_{n,j}^2 \left( D_{n,j}^{(k),X^{*}_{n,k'}} \right) > t_n(\epsilon_1, \epsilon_2) \right\}$$

(E.11) \( \leq 2d \left\{ \exp \left\{ \frac{-\bar{n} \epsilon^2}{t_n(\epsilon_1, \epsilon_2)} \right\} \right\} + C\bar{n} \exp \left( \frac{-\epsilon_1^2}{v_{n,m}} \right) + \bar{n} q_{n,m} + 2 \exp(-\bar{n}^2 \epsilon_2). \)

Thus, (E.11) establishes an explicit tail bound for $\|T_{m,n}^{(k,k')}\|_\infty$. The final bound for $\|T_{m,n}\|_\infty$. A tail bound for $\|T_{m,n}\|_\infty$ now follows easily using (E.1) and (E.11) along with the u.b. For any $\epsilon, \epsilon_1, \epsilon_2 \geq 0$, we have:

$$\mathbb{P} \left( \|T_{m,n}\|_\infty > \epsilon \right) \leq \mathbb{P} \left( \|T_{m,n}^{(1,2)}\|_\infty > \epsilon \right) + \mathbb{P} \left( \|T_{m,n}^{(2,1)}\|_\infty > \epsilon \right)$$

$$\leq 4d \exp \left\{ \frac{-\bar{n} \epsilon^2}{t_n(\epsilon_1, \epsilon_2)} \right\} + 4C\bar{n}d \exp \left( \frac{-\epsilon_1^2}{v_{n,m}} \right) + 4\bar{n}d q_{n,m} + 8d \exp(-\bar{n}^2 \epsilon_2).$$

This proves the desired bound and concludes the proof of Theorem E.1. ■

**E.2. Completing the Proof of Theorem 3.3.** Given the general tail bound for $\|T_{m,n}\|_\infty$ in Theorem E.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 $\|T_{m,n}\|_\infty$. To this end, let $(c_1, c_2) > 1$ be any universal constants, and set $\epsilon_1 = c_1 v_{n,m} \sqrt{\log(\bar{n}d)}$ and $\epsilon_2 = c_2 \sqrt{\log(\bar{n})}/\bar{n}$, where we further assume w.l.o.g. that $\epsilon_2 < 1$. Then, note that

$$\epsilon_1 + b_{n,m} \equiv c_1 v_{n,m} \sqrt{\log(\bar{n}d)} + b_{n,m} \leq c_1 \{ v_{n,m} \sqrt{\log(\bar{n}d)} + b_{n,m} \}$$

and

$$\|\mu_h^{(2)}\|_\infty + \sqrt{2} \bar{\sigma}_h \epsilon_2 + \bar{K}_h \epsilon_2^2 \leq \|\mu_h^{(2)}\|_\infty + \left( \sqrt{2} \bar{\sigma}_h + \bar{K}_h \right) \epsilon_2 \sqrt{\frac{\log d}{\bar{n}}}.$$
Therefore, combining and using all the inequalities above in the definition (E.9) of $t_\hat{n}(\epsilon_1, \epsilon_2)$, and letting $C_h := \sqrt{2\tilde{\sigma}_h + \tilde{K}_h}$, we have: $t_\hat{n}(\epsilon_1, \epsilon_2)$

$$\leq 8c_1^2\tilde{\sigma}_h^2 \{v_n, m \sqrt{\log(\tilde{n}d) + b_\tilde{n}, m}\}^2 \{\|\mu_h^{(2)}\|_\infty + c_2 C_h \sqrt{\log \frac{d}{n}}\}.$$ 

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

With probability $\geq 1 - \frac{4}{d_2^2-1} - \frac{8}{d_2^2-1} - \frac{4C}{(\tilde{n}d)_{q\delta}^2-1} - 4q_{n, \tilde{n}}(\tilde{n}d)$,

$$\|T_{m,n}\|_\infty \leq c v_n, m \sqrt{\log(\tilde{n}d) + b_\tilde{n}, m}C_1^* \left(\|\mu_h^{(2)}\|_\infty + C_2^* \sqrt{\log \frac{d}{n}}\right)^{1/2},$$

where $C_1^* := 4c_1 \tilde{\sigma}_h$ and $C_2^* := \sqrt{2c_2} C_h \equiv \sqrt{2c_2}(\sqrt{2\tilde{\sigma}_h + \tilde{K}_h})$, with $\tilde{\sigma}_h$ and $\tilde{K}_h$ being as in (E.8). This completes the proof of Theorem 3.3.

**APPENDIX F: 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 $\|R_{\pi,m,n}\|_\infty$.

**THEOREM F.1 (Tail bound for $\|R_{\pi,m,n}\|_\infty$).** Let Assumptions 1.1, 3.1, 3.2 and 3.3 hold, with the sequences $(v_n, \pi, b_n, \pi, q_n, \pi)$, $(\tilde{\pi}, m, \tilde{b}_n, m, q_n, \tilde{m})$ and the constants $(\delta_\pi, C)$ being as defined therein. Further, define $\|\mu_{[h]}\|_\infty := \max\{E\{|h_j| |X|\} : j = 1, \ldots, d\}$. Then, for any $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \geq 0$, with $\epsilon_2$ small enough such that $(\epsilon_2 + b_{n, \pi}) < \delta_\pi$, we have:

$$\P \left\{ \|R_{\pi,m,n}\|_\infty > \frac{r_{\pi,n}(\epsilon_1)}{\delta_\pi - r_{\pi,n}(\epsilon_2)}r_{\pi,n}(\epsilon_3)r_\ast(\epsilon_4) \right\} \leq 2d \exp(-n\epsilon_4^2)$$

$$+ Cn \left\{ \exp\left(\frac{-\epsilon_2^2}{v_{n, \pi}^2}\right) + \exp\left(\frac{-\epsilon_3^2}{v_{n, m}^2}\right) + \exp\left(\frac{-\epsilon_4^2}{v_{n, m}^2}\right) \right\} + 2nq_{n, \pi} + nq_{\tilde{n}, m},$$

where $r_{\pi,n}(\epsilon_1) := \epsilon_1 + b_{n, \pi}$, $r_{\pi,n}(\epsilon_2) := \epsilon_2 + b_{n, \pi}$, $r_{\pi,n}(\epsilon_3) := \epsilon_3 + b_{n, m}$ and $r_\ast(\epsilon_4) := \|\mu_{[h]}\|_\infty + \sqrt{2\sigma_{\pi,m} \epsilon_4 + K_{\pi,m} \epsilon_4^2}$ with $\sigma_{\pi,m}, K_{\pi,m} \geq 0$ being constants given by $\sigma_{\pi,m} := 4\sigma_{h} \delta_\pi^{-1}$ and $K_{\pi,m} := 2\sqrt{2}\sigma_{h} \delta_\pi^{-1}$.

**F.1. Proof of Theorem F.1.** Note that $R_{\pi,m,n}$ is essentially a ‘second order’ term since it involves a product of the two error terms arising from the estimation of $\pi(\cdot)$ and $m(\cdot)$ and under reasonable assumptions on $\tilde{\pi}(\cdot) - \pi(\cdot)$ and $\tilde{m}(\cdot) - m(\cdot)$, one can attempt to control the behavior of this term by
‘naive’ techniques, as opposed to the more sophisticated analyses required for controlling $T_{\tau,n}$ and $T_{m,n}$. Recall from (3.1) that

\[(F.1) \quad R_{\tau,m,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi(X_i)} - \frac{T_i}{\bar{\pi}(X_i)} \right\} \{\tilde{m}(X_i) - m(X_i)\} h(X_i).\]

Hence, with $\|\Delta_{\tau,n}\|_{\infty,n}$ and $\|\bar{\pi}_n\|_{\infty,n}$ as in (D.1) and (D.2) respectively, and with $\|\Delta_{m,n}\|_{\infty,n}$ as in (E.2) for any $k \neq k' \in \{1, 2\}$, we have:

\[(F.2) \quad \|R_{\tau,m,n}\|_{\infty} \leq \|\bar{\pi}_n\|_{\infty,n} \|\Delta_{\tau,n}\|_{\infty,n} \|\Delta_{m,n}\|_{\infty,n} \|\bar{\xi}_n\|_{\infty}, \text{ where} \]

\[\|\Delta_{m,n}\|_{\infty,n} := \max \left\{ \|\Delta_{(1.2)}\|_{\infty,n}, \|\Delta_{(2.1)}\|_{\infty,n} \right\} \text{ and} \]

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

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

\[(F.3) \quad \mathbb{P} \left\{ \|\Delta_{\tau,n}\|_{\infty,n} > r_{\tau,n}(\epsilon_1) \right\} \leq Cn \exp \left( -\frac{\epsilon_1^2}{v_{n,\pi}^2} \right) + nq_{n,\pi}, \text{ where} \]

\[r_{\tau,n}(\epsilon) := \epsilon + b_{n,\pi} \quad \text{for any } \epsilon \geq 0,\]

and using (D.10), for any $\epsilon_2 \geq 0$ small enough such that $r_{\tau,n}(\epsilon_2) < \delta_\tau$,

\[(F.4) \quad \mathbb{P} \left[ \|\bar{\pi}_n\|_{\infty,n} > \left(\delta_\tau - r_{\tau,n}(\epsilon_2)\right)^{-1} \right] \leq Cn \exp \left( -\frac{\epsilon_2^2}{v_{n,\pi}^2} \right) + nq_{n,\pi}.\]

Next, let $r_{m,n}(\epsilon) := \epsilon + b_{\bar{\pi}_n,m}$. Using (E.7), we have: for any $\epsilon_3 \geq 0$,

\[(F.5) \quad \mathbb{P} \left\{ \|\Delta_{m,n}\|_{\infty,n} > r_{\bar{m},n}(\epsilon_3) \right\} \leq \sum_{k \neq k' \in \{1, 2\}} \mathbb{P} \left\{ \|\Delta_{m,n}\|_{\infty,n} > r_{m,n}(\epsilon_3) \right\} \leq 2C\bar{n} \exp \left( -\frac{\epsilon_3^2}{v_{n,m}^2} \right) + 2\bar{n}q_{\bar{n},m} \equiv Cn \exp \left( -\frac{\epsilon_3^2}{v_{n,m}^2} \right) + nq_{\bar{n},m}.\]

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

\[(F.6) \quad \xi_{[j]}(T, X) \equiv \frac{T}{\pi(X)} |h_{[j]}(X)| \sim \text{BMC}(\bar{\sigma}_\xi, \bar{K}_\xi) \forall j \in \{1, \ldots, d\}.\]
Further, $E\{\xi\|T, X\} = E\{h\|X\} \equiv \mu_{|h|}$ (say) $\forall j \in \{1, \ldots, d\}$, and recall that $\|\mu_{|h|}\|_\infty = \max\{\mu_{|h|}\}_{j = 1, \ldots, d}$. Using (F.6) and Lemma A.4 along with the u.b., we then have: for any $\epsilon_4 \geq 0$,

$$
P\left\{\|\xi_n\|_\infty > r_*(\epsilon_4) \right\} \equiv \|\mu_{|h|}\|_\infty + \sqrt{2}\bar{\sigma}_\xi \epsilon_4 + \bar{K}_\xi \epsilon_4^2 \right\}
\leq \sum_{j=1}^d P\left\{\left|\frac{1}{n} \sum_{i=1}^n \xi_{j}(T_i, X_i) - \mu_{|h|}\right| > \sqrt{2}\bar{\sigma}_\xi \epsilon_4 + \bar{K}_\xi \epsilon_4^2 \right\}
\leq 2d \exp(-nc_4^2) \equiv 2 \exp(-nc_4^2 + \log d).
$$

(F.7)

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

$$
P\left\{\|R_{\pi,m,n}\|_\infty > \frac{r_{\pi,n}(\epsilon_1)}{\delta_\pi} r_{\pi,n}(\epsilon_3) r_*(\epsilon_4) \right\} \leq 2d \exp(-nc_4^2)
+ Cn \left\{ \exp\left(-\frac{\epsilon_1^2}{v_{n,\pi}^2}\right) + \exp\left(-\frac{\epsilon_3^2}{v_{n,m}^2}\right) + \exp\left(-\frac{\epsilon_4^2}{v_{n,\pi}^2}\right) \right\} + 2nq_{n,\pi} + nq_{n,m},
$$

where $r_{\pi,n}(\cdot), r_{\pi,n}(\cdot), r_*(\cdot) \geq 0$ are as in (F.3), (F.5) and (F.7) respectively, and we assume that $r_{\pi,n}(\epsilon_2) < \delta_\pi$. The proof of Theorem 3.1 is complete.

**F.2. Completing the Proof of Theorem 3.4.** Given the general tail bound for $\|R_{\pi,m,n}\|_\infty$ in Theorem F.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 $\|R_{\pi,m,n}\|_\infty$. Let $c_1, c_2, c_3, c_4 > 1$ be universal constants, and set $\epsilon_1 = c_1 \sqrt{\log n}$, $\epsilon_2 = c_2 \sqrt{\log n}$, $\epsilon_3 = c_3 \sqrt{\log n}$ and $\epsilon_4 = c_4 \sqrt{\log n}$, where we assume w.l.o.g. that $\epsilon_2 + b_{n,\pi} \leq \delta_\pi / 2 < \delta_\pi$ (i.e. $\epsilon_2$ satisfies the requirement in Theorem F.1) and $\epsilon_4 < 1$. We then have:

$$r_{\pi,n}(\epsilon_1) \leq c_1 \sqrt{\log n} + b_{n,\pi}, \quad r_*(\epsilon_4) \leq \|\mu_{|h|}\|_\infty + c_4 C_\xi \sqrt{n \log d},$$

$$r_{\pi,n}(\epsilon_3) \leq c_3 \sqrt{\log n} + b_{n,m}, \quad \text{and} \quad r_{\pi,n}(\epsilon_2) \leq \delta_\pi / 2,$$

where $C_\xi := \sqrt{2}\bar{\sigma}_\xi + \bar{K}_\xi$. Using Theorem F.1, we then have:

$$\|R_{\pi,m,n}\|_\infty \leq 2c_1 c_3 \frac{r_{\pi,n} r_{\pi,n}}{\delta_\pi} \left( \|\mu_{|h|}\|_\infty + c_4 C_\xi \sqrt{n \log d} \right),$$

where $r_{\pi,n} := c_1 \sqrt{\log n} + b_{n,\pi}$, $r_{\pi,n} := c_3 \sqrt{\log n} + b_{n,m}$ and $C_\xi := \sqrt{2}\bar{\sigma}_\xi + \bar{K}_\xi$ with $\bar{\sigma}_\xi$ and $\bar{K}_\xi$ as in (F.6). This completes the proof of Theorem 3.4.
APPENDIX G: PROOFS OF ALL RESULTS IN SECTION 5

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

\[
\sup_{x \in \mathcal{X}} |g(\hat{\beta}' \Psi(x)) - g(\beta_0' \Psi(x))| \leq C_g \sup_{x \in \mathcal{X}} |(\hat{\beta} - \beta_0)' \Psi(x)| \\
\leq C_g \|\hat{\beta} - \beta_0\|_1 \sup_{x \in \mathcal{X}} \|\Psi(X)\|_{\infty} \leq C_g C_{\Psi} \|\hat{\beta} - \beta_0\|_1.
\]

(G.1)

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

\[
P(\|\hat{\beta} - \beta_0\|_1 > \epsilon) \\
= P(\|\hat{\beta} - \beta_0\|_1 > \epsilon, \|\hat{\beta} - \beta_0\|_1 \leq a_n) + P(\|\hat{\beta} - \beta_0\|_1 > \epsilon, \|\hat{\beta} - \beta_0\|_1 > a_n) \\
\leq P(\|\hat{\beta} - \beta_0\|_1 > \epsilon, \|\hat{\beta} - \beta_0\|_1 \leq a_n) + P(\|\hat{\beta} - \beta_0\|_1 > a_n) \\
\leq P(\|\hat{\beta} - \beta_0\|_1 > \epsilon \mid \|\hat{\beta} - \beta_0\|_1 \leq a_n) P(\|\hat{\beta} - \beta_0\|_1 \leq a_n) + q_n \\
\leq 2 \exp(-\epsilon^2/(2a_n^2))(1 - q_n) + q_n \leq 2 \exp(-\epsilon^2/(2a_n^2)) + q_n,
\]

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

\[
P(\sup_{x \in \mathcal{X}} |g(\hat{\beta}' \Psi(x)) - g(\beta_0' \Psi(x))| > C_g C_{\Psi} \epsilon \leq 2 \exp(-\epsilon^2/(2a_n^2)) + q_n.
\]

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

G.2. Proof Sketch for Theorems 5.2 and 5.3. We first introduce two key supporting lemmas regarding tail bounds for \(\hat{l}(\hat{\beta}, x)\) 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 \(\hat{l}(\hat{\beta}, x)\), we first introduce the corresponding \textit{hypothetical} version of the estimator where the index parameter \(\beta\) is treated as known. Specifically, for any \(x \in \mathcal{X}\), let us define the ‘oracle’ ‘estimator’:

\[
\hat{l}(\beta, x) := \frac{1}{nh} \sum_{i=1}^n Z_i K \left( \frac{\beta' X_i - \beta' x}{h} \right) = \frac{1}{nh} \sum_{i=1}^n Z_i K \left( \frac{W_i - w x}{h} \right).
\]

Then, we note that the error \(\hat{l}(\hat{\beta}, x) - l(\beta, x)\) of the original estimator \(\hat{l}(\cdot)\)
admits the following decomposition. For any \( x \in \mathcal{X} \),
\[
|\tilde{l}(\beta, x) - l(\beta, x)| \leq |\tilde{l}(\beta, x) - l(\beta, x)| + |\hat{l}(\beta, x) - l(\beta, x)|
\]
\[
\leq |\tilde{l}(\beta, x) - E[l(\beta, x)]| + |E[\tilde{l}(\beta, x)] - l(\beta, x)| + |\hat{l}(\beta, x) - l(\beta, x)|
\]
\[
=: |\tilde{S}_n(x)| + |\mathcal{S}_n(x)| + |\hat{R}_n(x)| \quad \text{(say)}.
\]

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

**Lemma G.1** (Characterizing the tail bounds for \( \tilde{S}_n(x) \) and \( \mathcal{S}_n(x) \)). **Under Assumption 5.1 (a)-(c), we have:** for any fixed \( x \in \mathcal{X} \) and any \( t \geq 0 \),

\[
\mathbb{P} \left\{ |\tilde{S}_n(x)| > C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2 \sqrt{\log n}}{nh} \right\} \leq 3 \exp(-t^2),
\]

where \( C_1 := 7(B_1C_KM_K)^{1/2} \) and \( C_2 := D\sigma_ZM_K \) for some absolute constant \( D > 0 \). **Further, under Assumption 5.1 (d), we have:**

\[
|\mathcal{S}_n(x)| \leq C_3h^2 \quad \text{uniformly in } x \in \mathcal{X}, \quad \text{where } C_3 := B_2R_K.
\]

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

\[
(G.2) \quad |\tilde{l}(\beta, x) - l(\beta, x)| \leq C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2 \sqrt{\log n}}{nh} + C_3h^2, \quad \forall x \in \mathcal{X}.
\]

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

Using a first order Taylor series expansion of \( \tilde{l}(\beta, x) \) around \( \hat{l}(\beta, x) \equiv l(\beta, x) \), we first rewrite \( \hat{R}_n(x) \equiv \tilde{l}(\beta, x) - l(\beta, x) \) as:

\[
\hat{R}_n(x) = (\bar{\beta} - \beta)^T \left\{ \frac{1}{nh} \sum_{i=1}^{n} \frac{Z_i}{h} \left( \frac{W_i - w_x^*}{h} - \frac{W_i - w_x}{h} \right) \right\},
\]

where \( \{W_i^*\}_{i=1}^{n} \) and \( w_x^* \) are ‘intermediate’ points that satisfy, for each \( i = 1, \ldots, n \),

\[
|(W_i^* - w_x^*) - (W_i - w_x)| \leq \|(\bar{W}_i - \bar{w}_x) - (W_i - w_x)\| = \|\bar{W}_i - \bar{w}_x\| = \|(\bar{\beta} - \beta)'(X_i - x)\|.
\]

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

\[
\hat{R}_{n,1}(x) := (\bar{\beta} - \beta)^T \left\{ \frac{1}{nh} \sum_{i=1}^{n} \frac{Z_i}{h} \left( \frac{W_i - w_x}{h} \right) \right\}
\]

\[
=: (\bar{\beta} - \beta)^T \mathcal{T}_n(x) \quad \text{(say)}, \quad \text{and } \hat{R}_{n,2}(x) := \hat{R}_n(x) - \hat{R}_{n,1}(x).
\]

In the result below, we now characterize the tail bounds for \( \hat{R}_n(x) \).
LEMMA G.2 (Characterizing the tail bounds for \( \hat{R}_{n,1}(x) \) and \( \hat{R}_{n,2}(x) \)).
Under Assumption 5.2 (a), (b) and (d), and Assumption 5.1 (a) and (c), we have: for any \( t \geq 0 \), with probability at least 1 – 3 exp\((-t^2)\) – \( q_n \),
\[
|\hat{R}_{n,1}(x)| \leq C_1^* a_n + C_2^* \frac{a_n(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n(t^2 + \log p)\sqrt{\log n}}{nh^2},
\]
where \( C_1^*, C_2^*, C_3^* > 0 \) are constants depending only on the constants introduced in Assumptions 5.2 and 5.1, and \( x \in X \) is any fixed evaluation point.

Further, under the additional condition in Assumption 5.2 (c), we have: for any \( t \geq 0 \), with probability at least 1 – 3 exp\((-t^2)\) – \( q_n \),
\[
|\hat{R}_{n,2}(x)| \leq 4M^2 K^4 \frac{a_n^2}{h^2} + 4M^2 K \left( C_5^* \frac{a_n t}{\sqrt{nh^3}} + C_6^* \frac{t^2 a_n^2 \sqrt{\log n}}{nh^3} \right),
\]
where \( C_4^*, C_5^*, C_6^* > 0 \) are constants depending only on the constants introduced in Assumptions 5.1 and 5.2, and \( x \in X \) 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 \( \hat{R}_{n,2}(x) \) are each dominated by the respective terms in the bound for \( \hat{R}_{n,1}(x) \) in Lemma G.2. Using this, we obtain a bound for \( \hat{R}_n(x) \) as follows: for any \( t \geq 0 \), with probability at least 1 – 6 exp\((-t^2)\) – 2\( q_n \),
\[
|\hat{R}_n(x)| \equiv |\hat{l}(\beta, x) - l(\beta, x)|
\]
\[
\leq C_1^* (a_n + a_n^2 h^{-2}) + C_2^* \frac{a_n(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n(t^2 + \log p)\sqrt{\log n}}{nh^2},
\]
for some constants \( C_1^*, C_2^*, C_3^* > 0 \) (possibly different from those in Lemma G.2) depending only on the constants defined in Assumptions 5.1 and 5.2.

G.3. Completing the Proof of Theorem 5.2. Combining the bounds (G.2) and (G.3) via a union bound, we then have: for any \( x \in X \) and for any \( t \geq 0 \), with probability at least 1 – 9 exp\((-t^2)\) – 2\( q_n \),
\[
|\hat{l}(\beta, x) - l(\beta, x)| \leq |\hat{l}(\beta, x) - l(\beta, x)| + |\hat{R}_n(x)| \leq C_1 \frac{t}{\sqrt{nh}} + C_2 \frac{t^2 \sqrt{\log n}}{nh} + C_3 h^2 + C_1^* (a_n + a_n^2 h^{-2}) + C_2^* \frac{a_n(t + \sqrt{\log p})}{\sqrt{nh^3}} + C_3^* \frac{a_n(t^2 + \log p)\sqrt{\log n}}{nh^2}
\]
\[
= D_1 \frac{t}{\sqrt{nh}} \left( 1 + \frac{a_n}{h} \right) + D_2 \frac{t^2 \sqrt{\log n}}{nh} \left( 1 + \frac{a_n}{h} \right) + D_3 b_n, \quad \text{where}
\]
r_n := h^2 + a_n \frac{a_n^2}{h^2} + a_n \frac{\sqrt{\log p}}{h} + a_n \frac{\sqrt{\log n \log p}}{nh} = o(1) \quad \text{and}
\]
$D_1, D_2, D_3 > 0$ are some constants depending on the constants $\{C_j, C_j^*, j=1\}$.

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 (G.4) can be bounded as: $(a_n/h)\{(\sqrt{\log p}/(nh)) = o(1/\sqrt{nh})$ and the fifth term can be bounded as:

$$a_n \frac{\sqrt{\log n} \log p}{nh} \leq a_n \frac{\sqrt{\log p} \log(np)}{nh} = o\left(\frac{\log(np)}{nh}\right),$$

where we used that $\sqrt{\log n} \sqrt{\log p} \leq (\log n + \log p)/2 \leq \log(np)$. Using these simplifications in (G.4) and that $a_n/h = o(1)$ by assumption, we finally have: for any $x \in X$, and for any $t \geq 0$, with probability at least $1 - 6 \exp(-t^2) - 2q_n$,

$$\|\hat{l}(\beta, x) - l(\beta, x)\| \leq D_1^* \frac{t}{\sqrt{nh}} + D_2^* \frac{t^2 \sqrt{\log n}}{nh} + D_3^* b_n, \text{ where}$$

$$b_n := h^2 + a_n + \frac{a_n^2}{h^2} + \frac{1}{\sqrt{nh}} + \frac{\log(np)}{nh} \text{ and}$$

$D_1^*, D_2^*, D_3^* > 0$ are some constants depending only on those introduced in the assumptions. This completes the proof of Theorem 5.2. \hfill \blacksquare$

**G.4. Completing the Proof of Theorem 5.3.** Using Theorem 5.2, we have: for any fixed $x \in X$, and for any $t \geq 0$,

$$\mathbb{P}\left\{\|\hat{l}(\beta, x) - l(\beta, x)\| > \epsilon_n(t)\right\} \leq 9 \exp(-t^2) + 2q_n \text{ and}$$

(G.5) $$\mathbb{P}\left\{\|\hat{f}(\beta, x) - f(\beta, x)\| > \epsilon_n(t)\right\} \leq 9 \exp(-t^2) + 2q_n,$$

where we recall that $\{\hat{l}(\beta, x), f(\beta, x)\}$ is a special case of $\{l(\beta, x), l(\beta, x)\}$ with $Z \equiv 1$ so that Theorem 5.2 indeed applies to obtain both bounds above.

Next, note that $\hat{m}(\cdot) \equiv \hat{l}(\cdot)/\hat{f}(\cdot)$ and $m(\cdot) \equiv l(\cdot)/f(\cdot)$, so that

$$|\hat{f}(\cdot)\hat{m}(\cdot) - m(\cdot)| = |\{\hat{l}(\cdot) - l(\cdot)\} - m(\cdot)\{\hat{f}(\cdot) - f(\cdot)\}|$$

$$\leq |\hat{l}(\cdot) - l(\cdot)| + |m(\cdot)||\hat{f}(\cdot) - f(\cdot)| \leq |\hat{l}(\cdot) - l(\cdot)| + \delta_m |\hat{f}(\cdot) - f(\cdot)|,$$

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 $x \in X$, and for any $t \geq 0$,

$$\mathbb{P}\left\{|\hat{f}(\beta, x)\{\hat{m}(\beta, x) - m(\beta, x)\}| > (1 + \delta_m)\epsilon_n(t)\right\}$$

$$\leq \mathbb{P}\left\{|\hat{l}(\beta, x) - l(\beta, x)| > \epsilon_n(t)\right\} + \mathbb{P}\left\{|\hat{f}(\beta, x) - f(\beta, x)| > \epsilon_n(t)\right\}$$

(G.6) $$\leq 18 \exp(-t^2) + 4q_n.$$
where the final step follows from using the bounds in (G.5).
Recall further that by assumption, $|f(\beta, x)| \equiv f(\beta, x) \geq \delta_f > 0 \ \forall \ x \in \mathcal{X}$. Then, for any $x \in \mathcal{X}$ and any $t_\ast \geq 0$ such that $\delta_f - \epsilon_n(t_\ast) > 0$, we have:

$$
\mathbb{P}\{|\tilde{f}(\beta, x)| < \delta_f - \epsilon_n(t_\ast)\} \leq \mathbb{P}\{|\tilde{f}(\beta, x)| < |f(\beta, x)| - \epsilon_n(t_\ast)\}
$$

(G.7) \quad \leq \mathbb{P}\{|\tilde{f}(\beta, x) - f(\beta, x)| > \epsilon_n(t_\ast)\} \leq 9\exp(-t_\ast^2) + 2q_\ast,

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 (G.5). In particular, we have:

$$
\mathbb{P}\{|\tilde{f}(\beta, x)| < \frac{\delta_f}{2}\} \leq 9\exp(-t_\ast^2) + 2q_\ast, \ \forall \ t_\ast \geq 0 \ \text{such that} \ \epsilon_n(t_\ast) \leq \frac{\delta_f}{2}.
$$

Combining this bound along with (G.6), we now have: for any $x \in \mathcal{X}$ and for any $t, t_\ast \geq 0$ with $\epsilon_n(t_\ast) \leq \delta_f/2$,

$$
\mathbb{P}\{|\tilde{m}(\tilde{\beta}, x) - m(\beta, x)| > \frac{2(1 + \delta_m)}{\delta_f} \epsilon_n(t)\}
$$

$$
= \mathbb{P}\{|\tilde{m}(\tilde{\beta}, x) - m(\beta, x)| > \frac{2(1 + \delta_m)}{\delta_f} \epsilon_n(t), |\tilde{f}(\tilde{\beta}, x)| \geq \frac{\delta_f}{2}\}
$$

$$
+ \mathbb{P}\{|\tilde{m}(\tilde{\beta}, x) - m(\beta, x)| > \frac{2(1 + \delta_m)}{\delta_f} \epsilon_n(t), |\tilde{f}(\tilde{\beta}, x)| < \frac{\delta_f}{2}\}
$$

$$
\leq \mathbb{P}\{|\tilde{f}(\tilde{\beta}, x)| > (1 + \delta_m)\epsilon_n(t)\} + \mathbb{P}\{|\tilde{f}(\tilde{\beta}, x)| < \frac{\delta_f}{2}\}
$$

$$
\leq 18\exp(-t^2) + 9\exp(-t_\ast^2) + 6q_\ast,
$$

where the final bound follows from using (G.6), (G.7) and the bound noted below (G.7) as a special case. This completes the proof of Theorem 5.3.

---

**G.5. Proof of Lemma G.1.** Let $Z := (Z, X)$ and rewrite $\tilde{l}(\beta, x)$ as:

$$
\tilde{l}(\beta, x) = \frac{1}{n} \sum_{i=1}^{n} T_h(Z_i; x, \beta), \text{ where } T_h(Z; x, \beta) := \frac{1}{h} Z K \left( \frac{W_i - w_x}{h^2} \right).
$$

Under Assumption 5.1 (a)-(b) and using Lemma A.1 (i)(b), (ii)(d) and (v), $T_h(Z; x, \beta)$ is sub-Gaussian with $\|T_h(Z; x, \beta)\|_{\psi_2} \leq h^{-1}\sigma_Z M_K$. Hence, using Lemma A.1 (iv)(b) and (i)(c), we have:

$$
\|T_h(Z; x, \beta) - \mathbb{E}\{T_h(Z; x, \beta)\}\|_{\psi_2} \leq 3h^{-1}\sigma_Z M_K \ \text{uniformly for all } x \in \mathcal{X}.
Further, under Assumption 5.1 (b)-(c), we have: uniformly for all $x \in \mathcal{X}$,

$$\text{Var}\{ T_h(Z; x, \beta) \} \leq \mathbb{E}\{ T_h^2(Z; x, \beta) \} = \mathbb{E}_W[\mathbb{E}\{ T_h^2(Z; x, \beta) | W \}]$$

$$= h^{-2} \int_{\mathbb{R}} \mathbb{E}[Z^2 | W = w)K^2\{(w - w_x)/h\}f_\beta(w)dw$$

$$\equiv h^{-2} \int_{\mathbb{R}} m_\beta^2(w)K^2\{(w - w_x)/h\}f_\beta(w)dw$$

$$= h^{-1} \int_{\mathbb{R}} m_\beta^2(w_x + hu)f_\beta(w_x + hu)K^2(u)du \leq h^{-1}B_1M_KC_K,$$

where the penultimate step follows from a standard change of variable and Taylor series expansion of the lemma. This completes the proof of the first part of Lemma G.1.

For the second part regarding $\mathcal{S}_n(x) \equiv \mathbb{E}\{ \hat{l}(\beta, x) \} - l(\beta, x)$, observe that $\mathbb{E}\{ \hat{l}(\beta, x) \} = \mathbb{E}\{ T_h(Z; x, \beta) \}$ and $l(\beta, x) \equiv l_\beta(w_x)$. We have: $\forall x \in \mathcal{X}$,

$$\mathcal{S}_n(x) \equiv \mathbb{E}\{ T_h(Z; x, \beta) \} - l(\beta, x) = \mathbb{E}_W[\mathbb{E}\{ T_h(Z; x, \beta) | W \}] - l_\beta(w_x)$$

$$= h^{-1} \int_{\mathbb{R}} \mathbb{E}[Z | W = w)K\{(w - w_x)/h\}f_\beta(w)dw - l_\beta(w_x)$$

$$= h^{-1} \int_{\mathbb{R}} l_\beta(w)K\{(w - w_x)/h\}dw - l_\beta(w_x)$$

$$= \int_{\mathbb{R}} l_\beta(w_x + hu)K(u)du - l_\beta(w_x) = \int_{\mathbb{R}} \{l_\beta(w_x + hu) - l_\beta(w_x)\}K(u)du$$

$$= h\beta'(w_x)\int_{\mathbb{R}} uK(u)du + h^2R^*_x(x) := h^2\int_{\mathbb{R}} l_\beta''(w_{x,u})u^2K(u)du,$$

where $w^*_{x,u}$ is some ‘intermediate’ point satisfying $|w_{x,u} - w_x| \leq h|u|$. The first two steps use $\mathbb{E}[Z | W = w) \equiv m_\beta(w)$ and $m_\beta(w)f_\beta(w) \equiv l_\beta(w)$. The next steps follow from a standard change of variable and Taylor series expansion argument under the assumed smoothness of $l_\beta(\cdot)$ in Assumption 5.1 (d) along with the conditions imposed therein on the kernel $K(\cdot)$. Using Assumption 5.1 (d), we further have: $\|l_\beta''(\cdot)\|_\infty \leq B_2$ and $\int |u^2K(u)|du \leq R_K$. Hence,

$$|\mathcal{S}_n(x)| \leq B_2 \int_{\mathbb{R}} u^2|K(u)|du \leq B_2R_K$$ uniformly for all $x \in \mathcal{X}$. 

This establishes the second part of Lemma G.1 and completes the proof.

\textbf{G.6. Proof of Lemma G.2.} To control $\hat{R}_{n,1}(x) \equiv (\hat{\beta} - \beta)\hat{T}_n(x)$, note
\begin{equation}
|\hat{R}_{n,1}(x)| \leq \|\hat{\beta} - \beta\|_1 \left[\|\hat{T}_n(x) - \mathbb{E}\{\hat{T}_n(x)\}\|_\infty + \|\mathbb{E}\{\hat{T}_n(x)\}\|_\infty\right]
\end{equation}

In the light of (G.8) and the assumed high probability bound for Assumptions 5.2 (d), it now suffices to bound $\|\hat{T}_n(x) - \mathbb{E}\{\hat{T}_n(x)\}\|_\infty$ and $\|\mathbb{E}\{\hat{T}_n(x)\}\|_\infty$.

To this end, for each $x \in \mathcal{X}$, define
\[T^*_h(Z; x) := \frac{1}{h^2} Z(X - x)K' \left(\frac{W - w_x}{h}\right)\]
so that $\hat{T}_n(x) = \frac{1}{n} \sum_{i=1}^n T^*_h(Z_i; x)$.

Now under Assumptions 5.1 (a), 5.1(c), 5.2 (a) and 5.2 (d), and using Lemma A.1 (i)(b)-(c), (iv)(b) and (v) at appropriate places, we have: for all $x \in \mathcal{X}$,
\[
\max_{1 \leq j \leq p} \|T^*_{h,j}(Z; x)\|_{\psi_2} \leq 2h^{-2}M_X M_K' \sigma_Z \quad \text{and therefore,}
\max_{1 \leq j \leq p} \|T^*_{h,j}(Z; x) - \mathbb{E}\{T^*_{h,j}(Z; x)\}\|_{\psi_2} \leq 6h^{-2}M_X M_K' \sigma_Z.
\]

Further, under Assumptions 5.2(d), 5.1(c) and 5.2 (a), and with $\mathbb{E}\{Z^2(X_{[j]} - x_{[j]})^2 | W\} \leq 4M_X^2 \mathbb{E}W(Z^2 | W) \equiv 4M_X^2 m^{(2)}_\beta(W) \forall j$, we have: for all $x \in \mathcal{X}$,
\[
\max_{1 \leq j \leq p} \mathbb{E}\left[\|T^*_{h,j}(Z; x)\|^2\right] \leq \frac{4}{h^4} M_X^2 \mathbb{E}W(Z^2 | W) \leq \frac{2}{h^3} M_K' B_1 \int m^{(2)}_\beta(w) K'(w - w_x)/h)^2 f_\beta(w)dw
\leq \frac{4}{h^4} M_X^2 M_K' B_1 \int m^{(2)}_\beta(w_x + hu) f_\beta(w_x + hu) K'(u)^2 du
\leq \frac{4}{h^3} M_X^2 M_K' B_1 \int |K'(u)|du \leq \frac{4}{h^3} B_1 M_X^2 M_K' C_K',
\]
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 A.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 $x \in \mathcal{X}$ and for any $t \geq 0$, with probability at least $1 - 3 \exp(-t^2)$,
\begin{equation}
\|\hat{T}_n(x) - \mathbb{E}\{\hat{T}_n(x)\}\|_\infty \equiv \left\|\frac{1}{n} \sum_{i=1}^n T^*_h(Z_i; x) - \mathbb{E}\{T^*_h(Z; x)\}\right\|_\infty
\leq C_1 \left(\frac{t + \sqrt{\log p}}{\sqrt{nh^3}} + C_2 \frac{(t^2 + \log p) \sqrt{\log n}}{nh^2}\right),
\end{equation}
for some constants $C_1, C_2 > 0$ depending only on those introduced in the assumptions. Here, we further used \( \sqrt{a + b} \leq \sqrt{a} + \sqrt{b} \) for any \( a, b \geq 0 \) to obtain the bound (G.9) from the one originally provided by Lemma A.6.

Next, we focus on controlling \( \|E\{T^*_{h}(Z; x)\}\|_{\infty} \). To this end, recall the definitions of \( \eta_\beta(\cdot) \in \mathbb{R}^p \) and \( l_\beta(\cdot) \in \mathbb{R} \), and let \( \eta_\beta'(w) := \frac{d}{dw} \eta_\beta(w) \in \mathbb{R}^p \).

Then, under Assumption 5.2 (a)-(b), we have: uniformly in \( x \in \mathcal{X} \),

\[
\mathbb{E}\{T^*_{h}(Z; x)\} = \frac{1}{h^2} \mathbb{E}_W[\mathbb{E}\{(ZX - Zx)|W\}K'(W - w_x)/h]\]

\[
= \frac{1}{h^2} \int_{\mathbb{R}} \{\eta_\beta(w) - xl_\beta(w)\}K'(W - w_x)/h\,dw
\]

\[
= \frac{1}{h} \int_{\mathbb{R}} \{\eta_\beta(w + hu) - xl_\beta(w + hu)\}K'(u)\,du
\]

\[
= \int_{\mathbb{R}} \{\eta_\beta'(w + hu) - xl_\beta'(w + hu)\}K(u)\,du,
\]

where the last two steps follow from a change of variable and integration by parts argument, where the latter is applicable under Assumption 5.2 (a)-(b).

Under Assumptions 5.2 (a), 5.2 (b) and 5.2 (d), we then have:

\[
\mathbb{E}\{T^*_{h}(Z; x)\}\|_{\infty} \leq \left\{ \max_{1 \leq j \leq p} \|\eta_\beta'^{(j)}(\cdot)\|_\infty + \|x\|_\infty\|l_\beta'(\cdot)\|_\infty \right\} \int_{\mathbb{R}} |K(u)|\,du
\]

(G.10)

\[
\leq (B_1^* + M_X B_2^*) C_K \quad \text{uniformly in } x \in \mathcal{X}.
\]

Finally, recall that from Assumption 5.2 (d), we have \( \|\hat{\beta} - \beta\|_1 \leq a_n \) with probability at least \( 1 - q_n \). Combining this with the bounds (G.9) and (G.10) and applying them in (G.8) through a simple union bound, we have: for any fixed \( x \in \mathcal{X} \) and for \( t \geq 0 \), with probability at least \( 1 - 3 \exp(-t^2) - q_n \),

\[
|\hat{R}_{n,1}(x)| \leq a_n \left\{ C_1^* + C_2^* \left( \frac{t + \sqrt{\log p}}{\sqrt{nh^3}} \right) + C_3^* \frac{(t^2 + \log p)\sqrt{\log n}}{nh^2} \right\},
\]

for some constants \( C_1^*, C_2^*, C_3^* \) depending only on those introduced in our assumptions. This establishes the first part of Lemma G.2.

To establish the second part of Lemma G.2 regarding bounds for \( \hat{R}_{n,2}(x) \), first recall that for some ‘intermediate’ points \( \{W^*_i\}_{i=1}^n \) and \( w^*_x \) satisfying

...
\(|(W_i^* - w_x^*) - (W_i - w_x)| \leq |(\hat{W}_i - \hat{w}_x) - (W_i - w_x)| \equiv |(\hat{\beta} - \beta)'(X_i - x)|,
\]

\[|\hat{R}_{n,2}(x)| \equiv \left|\left(\frac{\hat{\beta} - \beta}{nh^2}\right) \sum_{i=1}^{n} Z_i(X_i - x) \left\{ K'(\frac{W_i^* - w_x^*}{h}) - K'(\frac{W_i - w_x}{h}) \right\} \right| \leq \frac{\|\hat{\beta} - \beta\|_1}{nh^2} \sum_{i=1}^{n} \|X_i - x\|_\infty |Z_i| \left| K'(\frac{W_i^* - w_x^*}{h}) - K'(\frac{W_i - w_x}{h}) \right| \leq 2M_X \|\hat{\beta} - \beta\|_1 \left\{ \frac{1}{nh^2} \sum_{i=1}^{n} |Z_i| \left| K'(\frac{W_i^* - w_x^*}{h}) - K'(\frac{W_i - w_x}{h}) \right| \right\}, \tag{G.11} \]

where the steps follow from an \(L_1-L_\infty\) bound along with a triangle inequality and using the boundedness of \(X\) from Assumption 5.2 (d).

Let \(A_n\) denote the event \(\|\hat{\beta} - \beta\|_1 \leq a_n\) and let \(A_n^c\) denote the complement event of \(A_n\). Then, from Assumption 5.2 (d), we have \(P(A_n) \geq 1 - q_n\). Further, on the event \(A_n\), \((\hat{\beta} - \beta)'(X_i - x)/h \leq 2M_X(a_n/h) \leq L\) under Assumption 5.2 (d) and consequently, using Assumption 5.2 (c) with the function \(\varphi(\cdot)\) as defined therein, we have: on the event \(A_n\),

\[|K'(\frac{W_i - w_x}{h}) - K'(\frac{W_i^* - w_x^*}{h})| \leq \frac{1}{h}|(\hat{\beta} - \beta)'(X_i - x)| \varphi(\frac{W_i - w_x}{h}) \leq \frac{1}{h} \|\hat{\beta} - \beta\|_1 \|X_i - x\|_\infty \varphi(\frac{W_i - w_x}{h}) \leq 2M_X \frac{a_n}{h} \varphi(\frac{W_i - w_x}{h}), \tag{G.12} \]

and consequently, combining (G.11) and (G.12), we have: on the event \(A_n\),

\[|\hat{R}_{n,2}(x)| \leq \frac{2M_X^2 a_n^2}{h h^3} \sum_{i=1}^{n} |Z_i| \varphi(\frac{W_i - w_x}{h}) \quad \forall \, x \in \mathcal{X}. \tag{G.13} \]

Thus, we have: for any \(\epsilon \geq 0\) and for any \(x \in \mathcal{X}\),

\[P(|\hat{R}_{n,2}(x)| > \epsilon) \leq P(|\hat{R}_{n,2}(x)| > \epsilon, A_n) + P(|\hat{R}_{n,2}(x)| > \epsilon, A_n^c) \leq \left\{ \frac{4M_X^2 a_n^2}{nh^3} \sum_{i=1}^{n} |Z_i| \varphi(\frac{W_i - w_x}{h}) > \epsilon, A_n \right\} + P(A_n^c) \leq \left\{ \frac{4M_X^2 a_n^2}{nh^3} \sum_{i=1}^{n} |Z_i| \varphi(\frac{W_i - w_x}{h}) > \epsilon \right\} + q_n, \tag{G.14} \]

where the steps follow from (G.13) and that \(P(A_n^c) \leq q_n\) by assumption.
Next, define: $T_h(Z; x) \equiv T_h(Z; x, \beta) := h^{-3}|Z|\varphi\{(W - w_x)/h\}$ and recall that $m_\beta^{(2)}(W) \equiv \mathbb{E}(Z^2 | W)$. Then, using the boundedness conditions from Assumptions 5.1 (c) and 5.2(c), along with use of iterated expectations, we bound the first and second moments of $T_h(Z; x) \forall x \in \mathcal{X}$ as follows.

\[
\mathbb{E}\{T_h^2(Z; x)\} = \frac{1}{h^6} \int_{\mathbb{R}} m_\beta^{(2)}(w) \varphi^2\left(\frac{W - w_x}{h}\right) f_\beta(w)dw
\]

\[
= \frac{1}{h^5} \int_{\mathbb{R}} m_\beta^{(2)}(w_x + hu) f_\beta(w_x + hu) \varphi^2(u)du \leq \frac{B_1 M_\varphi C_\varphi}{h^5}, \quad \text{and}
\]

\[
\mathbb{E}\{T_h(Z; x)\} = \frac{1}{h^3} \int_{\mathbb{R}} \mathbb{E}(|Z| | W = w) \varphi\left(\frac{W - w_x}{h}\right) f_\beta(w)dw
\]

\[
\leq \frac{1}{h^2} \int_{\mathbb{R}} \{m_\beta^{(2)}(w_x + hu)\}^{\frac{3}{2}} \varphi(u) f_\beta(w_x + hu)du \leq \frac{(B_1 C_f)^{\frac{1}{2}} C_\varphi}{h^2},
\]

where $C_f > 0$ is a constant such that $\|f_\beta(\cdot)\|_\infty \leq C_f$. Further, under Assumptions 5.1 (a) and 5.2 (c), using various parts of Lemma A.1, we have:

\[
\|T_h(Z; x) - \mathbb{E}\{T_h(Z; x)\}\|_{\psi_2} \leq 3\|T_h(Z; x)\|_{\psi_2} \leq 3h^{-3} \sigma_Z M_\varphi \quad \forall x \in \mathcal{X}.
\]

Hence, using Lemma A.6, with all required conditions verified now, we have: for any $x \in \mathcal{X}$ and for any $t \geq 0$, with probability at least $1 - 3 \exp(-t^2),$

\[
\left|\frac{1}{n} \sum_{i=1}^{n} T_h(Z_i; x)\right| \leq \left|\frac{1}{n} \sum_{i=1}^{n} T_h(Z_i; x) - \mathbb{E}\{T_h(Z; x)\}\right| + |\mathbb{E}\{T_h(Z; x)\}|
\]

\[
\leq C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + C_5 \frac{h^2}{h^2},
\]

for some constants $C_3, C_4, C_5 > 0$ depending only on those in the assumptions. Hence, using (G.15) in (G.14), we now have: for any $t \geq 0,$

\[
P\left\{\left|\hat{R}_{n,2}(x)\right| \geq 4M_\varphi a_n^2 \left(C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + C_5 \frac{h^2}{h^2}\right)\right\}
\]

\[
\leq P\left\{\left|\frac{1}{nh^3} \sum_{i=1}^{n} |Z_i| \varphi\left(\frac{W_i - w_x}{h}\right) > C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + C_5 \frac{h^2}{h^2} + q_n\right\}
\]

\[
\equiv P\left\{\left|\frac{1}{n} \sum_{i=1}^{n} T_h(Z_i; x)\right| > C_3 \frac{t}{nh^5} + C_4 \frac{t^2 \sqrt{\log n}}{nh^3} + C_5 \frac{h^2}{h^2} + q_n\right\}
\]

\[
\leq 3 \exp(-t^2) + q_n \quad \text{for any } x \in \mathcal{X}.
\]

This establishes the desired bound for $\hat{R}_{n,2}(x)$ and completes the proof.
REFERENCES


Abhishek Chakrabortty  
Dept. of Statistics  
University of Pennsylvania  
Philadelphia, PA 19104, USA.  
E-mail: abhich@wharton.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