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

BY ABHISHEK CHAKRABORTTY, T. TONY CAI AND HONGZHE LI

University of Pennsylvania

October 20, 2018

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. The Problem Setup. Let $Y \in \mathbb{R}$ be an outcome and $X \in \mathbb{R}^p$ be a corresponding vector of covariates of interest. Let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^p$ denote the respective supports of $Y$ and $X$, both of which are allowed to be of arbitrary nature, i.e. continuous and/or discrete (both finite or infinite). However, the outcome $Y$ may not always be observed and the random vector that can be actually observed in practice 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: $TY = T Y$, i.e. $(Y | T = 1) \equiv Y$ almost surely (a.s.), 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 $Y := TY^{(1)} + (1 - T)Y^{(0)}$ denotes the observed outcome with $(Y^{(1)}, Y^{(0)})$ being the true ‘potential’ outcomes (Rubin, 1974; Imbens and Rubin, 2015) for $T = 1$ and $T = 0$ respectively. 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: $D_n := \{Z_i \equiv \ldots$
(Ti, TiYi, Xi) : i = 1, . . . , n} consisting of n independent and identically distributed (i.i.d.) realizations of Z with joint distribution defined via P(·).

**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 ⊥⊥ Y | 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 π(x) := P(T = 1 | X = x) ∀ x ∈ X, denote the ‘propensity score’ (Rosenbaum and Rubin, 1983) of the outcome being observed given the covariates X, and let π := P(T = 1). Then, we assume: ∃ a universal constant δπ with 0 < δπ ≤ 1, such that

\[(1.1) \quad π(x) ≥ δπ > 0 \quad ∀ x ∈ X, \quad \text{(and hence, } π ≥ δπ > 0 \text{ as well).}\]

Note that Assumption 1.1 (a) includes the special case: T ⊥⊥ (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, π(·) simply becomes the constant π defined above. In general, π(·) may depend on X and moreover, 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 ≪ n, p ∼ n or p ≫ 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, θ) : ℝ × ℝp × ℝd → ℝ be any ‘loss function’ that is convex and differentiable in θ ∈ ℝd, and we assume that [E{L(Y, X, θ)}]2 < ∞ for each θ ∈ ℝd. Then, the M-estimation problem considers the estimation of the parameter vector θ0 that corresponds to the minimizer of the risk function defined by L(·). Specifically, we aim to estimate the parameter vector θ0 ≡ θ0(π) ∈ ℝd, a functional of the probability measure P(·) underlying the unobserved Z, defined as:

\[(1.2) \quad θ0 ≡ θ0(L, P) := \arg \min_{θ ∈ ℝ^d} L(θ), \quad \text{where } L(θ) := E\{L(Y, X, θ)\}.\]
A. CHAKRABORTTY, T. T. CAI AND H. LI

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 $\theta \in \mathbb{R}^d$,

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

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 $l(\cdot)$ 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
\( L(Y, X, \theta) := l\{Y, \Psi(X)^\theta\} \) with the same of choices of \( l(\cdot, \cdot) \) as above. One frequently used choice of \( \Psi(\cdot) \) includes the polynomial bases of any (fixed) degree \( d_0 \geq 1 \), given by: 
\[
\Psi(X) := \{1, x^k_j : 1 \leq j \leq p, 1 \leq k \leq d_0\}
\]
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 \( Y = f(\beta_0^\prime X, \epsilon) \) 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. \( Y \perp \perp X | \beta_0^\prime X \)) and \( \beta_0 \) denotes the unknown index parameter (identifiable only upto 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 \( L(\cdot) \) 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
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

\[
\mathbb{L}(\theta) \equiv \mathbb{E}\{L(Y, X, \theta)\} = \mathbb{E}_X\{\phi(X, \theta)\} =: \mathbb{L}_{\text{REG}}(\theta) \quad \text{(say), and}
\]

\[
\mathbb{L}(\theta) \equiv \mathbb{E}\{L(Y, X, \theta)\} = \mathbb{E}\left\{ \frac{T}{\pi(X)}L(Y, X, \theta) \right\} =: \mathbb{L}_{\text{IPW}}(\theta) \quad \text{(say)}.
\]

**Debiased and doubly robust (DDR) representation of** $\mathbb{L}(\cdot)$. We also have

\[
\mathbb{L}(\theta) = \mathbb{E}_X\{\phi(X, \theta)\} + \mathbb{E}\left[ \frac{T}{\pi(X)}\{L(Y, X, \theta) - \phi(X, \theta)\} \right] =: \mathbb{L}_{\text{DDR}}(\theta) \quad \forall \theta \in \mathbb{R}^d.
\]

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:

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

$\mathbb{L}_{\text{DDR}}(\cdot)$, unlike $\mathbb{L}_{\text{IPW}}(\cdot)$ and $\mathbb{L}_{\text{REG}}(\cdot)$, 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.

2. **High Dimensional M-Estimation and Sparse Signal Recovery: A General Framework via Regularized DDR Loss Minimization.**

**Notations.** We first introduce some notations to be used throughout. For any $v \in \mathbb{R}^d$, $j \in \{1, \ldots, d\}$ and $J \subseteq \{1, \ldots, d\}$, $\hat{V}$ 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 $r \geq 0$, $A(v) := \{j : v_{[j]} \neq 0\}$ denotes the support of $v$, $s_v := |A(v)|$. 
denotes the cardinality of $A(v)$, $\Pi_J(v)$ denotes $\{v[j]1\{j \in J\}\}_{j=1}^d \in \mathbb{R}^d$, and $J^c := \{1, \ldots, d\} \setminus J$ denotes the complement of $J$. We use the shorthand $\Pi_v(\cdot)$ and $\Pi_x(\cdot)$ to denote $\Pi_{A(v)}(\cdot)$ and $\Pi_{A^c(v)}(\cdot)$ respectively. We further let $M_J = \{v \in \mathbb{R}^d : A(v) \subseteq J\}$ and $M^J_J = \{v \in \mathbb{R}^d : A(v) \subseteq J^c\}$. Lastly, for any measurable (and possibly random) function $f(\cdot)$ of $X$, $\|f(\cdot)\|_r := [E_X\{f(X)^r\}]^{1/r}$ denotes the $L_r$ norm of $f(\cdot)$ with respect to (w.r.t.) $P_X$ for any $r \geq 1$, and $\|f(\cdot)\|_\infty := \sup_{x \in X} |f(x)|$ denotes the $L_\infty$ norm w.r.t. $P_X$.

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

\begin{equation}
\nabla L(Y, X, \theta) \equiv \frac{\partial}{\partial \theta} L(Y, X, \theta) = h(X)\{Y - g(X, \theta)\}.
\end{equation}

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

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

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

\begin{equation}
\nabla \hat{\phi}(X, \theta) := \frac{\partial}{\partial \theta} \hat{\phi}(X, \theta) = h(X)\{\hat{m}(X) - g(X, \theta)\}.
\end{equation}

Note that to compute $\hat{\phi}(X, \theta)$ explicitly, one needs both the estimates $\hat{m}(\cdot)$ and $\hat{m}_C(\cdot)$. However, the part of $\hat{\phi}(X, \theta)$ involving $\hat{m}_C(\cdot)$ is free of $\theta$. Our proposed estimator, discussed in Section 2.2, is constructed based on a $L_1$-regularized minimization (w.r.t. $\theta$) of an objective function involving $\hat{\phi}(\cdot)$, whereby only its gradient $\nabla \hat{\phi}(X, \theta)$ is of interest and that depends only on $\hat{m}(X)$ 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 $\hat{m}_C(\cdot)$ 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 $\{D_n^{(1)}, D_n^{(2)}\}$ denote a random partition (or split) of the original data $D_n$ into $K = 2$ equal parts. Let $\bar{n} := n/2$ denote the size of $D_n^{(1)}$ and $D_n^{(2)}$ where, without loss of generality (w.l.o.g.), we assume that $n$ is even. Further, let $I_1$ and $I_2$ respectively denote the index sets $\{1, \ldots, n\} =: I$ for the observations in $D_n^{(1)}$ and $D_n^{(2)}$. Hence, with $K \equiv 2$, we have: $\{D_n^{(k)}\}_{k=1}^K$ and $\{I_k\}_{k=1}^K$ are disjoint and exhaustive partitions of $D$ and $I$ respectively, i.e. $\bigcup_{k=1}^K D_n^{(k)} = D$, $\bigcup_{k=1}^K I_k = I$ and $|I_k| = \bar{n} = n/K$ $\forall$ $k \in \{1, \ldots, K \equiv 2\}$.

Given any general procedure for obtaining $\hat{\pi}(\cdot)$ and $\hat{\phi}(\cdot, \cdot)$ based on the full observed data $D_n$, 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 $D_n^{(1)}$ and $D_n^{(2)}$ respectively. Let us now define the following cross-fitted estimates $\{\hat{m}(X_i), \hat{\phi}(X_i, \theta)\}_{i=1}^n$ of the conditional means $\{m(X_i), \phi(X_i, \theta)\}_{i=1}^n$ at the $n$ training points in $D_n$, as follows.

\begin{equation}
(2.3) \quad \{\hat{m}(X_i), \hat{\phi}(X_i, \theta)\} = \begin{cases} 
\{\hat{m}(2)(X_i), \hat{\phi}(2)(X_i, \theta)\} & \forall \ i \in I_1, \\
\{\hat{m}(1)(X_i), \hat{\phi}(1)(X_i, \theta)\} & \forall \ i \in I_2.
\end{cases}
\end{equation}

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 $L(\theta)$:

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

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

With \( \theta_0 \) (and \( X \)) 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, \( \hat{\theta}_{DDR} \), given by:

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

where \( L_{n, DDR}^{\text{DDR}}(\cdot) \) 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 \( \hat{\theta}_{DDR} \) in (2.5) can be implemented using a simple user-friendly imputation type algorithm as follows. Given the observed data \( D_n \) and the estimates \( \{ \hat{\pi}(X_i), \hat{m}(X_i) \} \) \( i = 1 \), let us first define a set of pseudo outcomes \( \tilde{Y}_i \), for all \( 1 \leq i \leq n \), given by:

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

Clearly \( \tilde{L}_{n, DDR}^{\text{DDR}}(\cdot) \) is convex and differentiable, and under (2.1) and (2.2), it is straightforward to see that \( \nabla \tilde{L}_{n, DDR}^{\text{DDR}}(\theta) = \nabla L_{n, DDR}^{\text{DDR}}(\theta) \), where \( \nabla L_{n, DDR}^{\text{DDR}}(\theta) := \frac{\partial}{\partial \theta} L_{n, DDR}^{\text{DDR}}(\theta) \). 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 \( L_{n, DDR}^{\text{DDR}}(\cdot) \) and the subgradient of the \( \| \cdot \|_1 \) norm. Hence, the solution stays unchanged if \( L_{n, DDR}^{\text{DDR}}(\theta) \) in (2.5) is replaced by \( \tilde{L}_{n, DDR}^{\text{DDR}}(\theta) \) which has the same gradient. Consequently, \( \hat{\theta}_{DDR} \) in (2.5) may also be defined as:

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

Thus, if one ‘pretends’ to have a fully observed data \( \bar{D}_n := \{ (\tilde{Y}_i, X_i) \} \) in terms of the pseudo outcomes, then \( \hat{\theta}_{DDR} \) may be obtained by a simple \( L_1 \)-penalized minimization of the corresponding empirical risk for \( L(\cdot) \) based on \( \bar{D}_n \). This minimization is quite easy to implement and can be done so using standard statistical software packages, including ‘glmnet’ in R. Finally, note
also that (2.6) confirms our earlier claim that even though the estimator $	ilde{\phi}(X, \theta)$ involved in the definition (2.4) of $L_{DDR}^n(\theta)$ may require estimation of further nuisance functions (independent of $\theta$) apart from $m(X)$, but the implementation of $\hat{\theta}_{DDR}$ 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 $\hat{\theta}_{DDR}$ that will serve as the backbone for most of our main theoretical analyses. We first introduce some notations and assumptions. Define: \[
\forall \theta, v \in \mathbb{R}^d, \\
\nabla L_{DDR}^n(\theta) := \frac{\partial}{\partial \theta} L_{DDR}^n(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla \tilde{\phi}(X_i, \theta) + \frac{1}{n} \sum_{i=1}^n T_i \hat{\pi}(X_i) \left\{ \nabla L(Y_i, X_i, \theta) - \nabla \tilde{\phi}(X_i, \theta) \right\},
\]
and $\delta L_{DDR}^n(\theta; v) := L_{DDR}^n(\theta + v) - L_{DDR}^n(\theta) - v' \{ \nabla L_{DDR}^n(\theta) \}$.

Assumption 2.1 (Restricted strong convexity). Let $\delta L_{DDR}^n(\theta; v)$ be as above for any $\theta, v \in \mathbb{R}^d$. We assume that at $\theta = \theta_0$, the loss function $L_{DDR}^n(\theta)$ satisfies a restricted strong convexity (RSC) property as follows: \[ \exists a \text{ (non-random) constant } \kappa_{DDR} > 0 \text{ such that } \delta L_{DDR}^n(\theta_0; v) \geq \kappa_{DDR} \|v\|^2 \forall v \in C(\theta_0), \] where
\[ C(\theta_0) := \left\{ v \in \mathbb{R}^d : \|\Pi_{\theta_0}(v)\|_1 \leq 3\|\Pi_{\theta_0}(v)\|_1 \right\} \subseteq \mathbb{R}^d. \]

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. $[P]$ for some $\kappa_{DDR} > 0$. 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 $L_{DDR}^n(\cdot)$ 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. $\delta L_{DDR}^n(\theta_0; v)$ in (2.8) is independent of $\{T_i, Y_i\}_{i=1}^n$ as well as the nuisance function estimates $\{\tilde{\pi}(X_i), \tilde{m}(X_i)\}_{i=1}^n$. 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 $\hat{\theta}_{DDR}$ 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 \( \hat{\theta}_{\text{DDR}} \)). Let \( \hat{\theta}_{\text{DDR}} \) be as defined in (2.5) and assume that \( \mathcal{L}_{\text{DDR}}(\theta) \) in (2.4) is convex and differentiable in \( \theta \) a.s. \([\mathbb{P}]\). Let Assumption 2.1 hold with the constant \( \kappa_{\text{DDR}} > 0 \) 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 \( \lambda \equiv \lambda_n \geq 2 \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \), we have:

\[
\| \hat{\theta}_{\text{DDR}} - \theta_0 \|_2 \leq 3 \sqrt{s} \lambda_n \kappa_{\text{DDR}} \text{ and } \| \hat{\theta}_{\text{DDR}} - \theta_0 \|_1 \leq 12 s \lambda_n \kappa_{\text{DDR}}.
\]

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), \( \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \lesssim \sqrt{(\log d)/n} \) with high probability. Hence, by choosing \( \lambda \equiv \lambda_n \asymp \sqrt{(\log d)/n} \) and using (2.9), it follows that with high probability,

\[
\| \hat{\theta}_{\text{DDR}} - \theta_0 \|_2 \lesssim \sqrt{s \log d/n} \text{ and } \| \hat{\theta}_{\text{DDR}} - \theta_0 \|_1 \lesssim s \sqrt{\log d/n}.
\]

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 \( 2 \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \) in Lemma 2.1 regarding the choice of \( \lambda_n \). This is the focus of Section 3, where we obtain sharp non-asymptotic upper bounds for \( \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \) that hold with high probability (w.h.p.) and converge to 0 at satisfactory rates. A choice of \( \lambda_n \) of the order of this upper bound guarantees that the lower bound requirement \( \lambda_n \geq 2 \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \) in Lemma 2.1 holds w.h.p. and establishes the convergence rates, defined by the rate of the chosen \( \lambda_n \), holding w.h.p. for the deviation bound in (2.9).

3. The Core Analyses for the DDR Estimator: Probabilistic Upper Bounds for \( \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \) and Convergence Rates. For most of our theoretical analyses of \( \| \nabla \mathcal{L}_{\text{DDR}}(\theta_0) \|_\infty \), 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 dimensional setting (which necessitates sharp non-asymptotic bounds).

Under possible misspecification of one of the estimators, the DR property (in terms of consistency) of \( \| \nabla L^{\text{DDR}}_n(\theta_0) \|_\infty \) 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 \( L^{\text{DDR}}_n(\cdot) \) 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 \( L^{\text{DDR}}_n(\theta) \) 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 provides 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 dimensional 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 structure.

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

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

(3.1)

where with $Z \equiv (T, Y, X)$, the individual random variables $T_{0}(Z)$, $T_{\pi}(Z)$, $T_{m}(Z)$ and $R_{\pi,m}(Z)$ appearing in the respective sums above are given by:

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

(3.2)

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

(3.3)

$$T_{m}(Z) := \left\{\frac{T}{\tilde{\pi}(X)} - 1\right\}\{\tilde{m}(X) - m(X)\}h(X),$$

(3.4)

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

(3.5)

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 $\tilde{\pi}(\cdot)$ and $\tilde{m}(\cdot)$ respectively, and $R_{\pi,m,n}$ is a second order bias term involving the product of the estimation errors of $\tilde{\pi}(\cdot)$ 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 9-12 respectively. In general, they involve careful analyses based on concentration inequalities and other techniques. These technical tools are all collected in Section 13.

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 13.1 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 \( \|\varepsilon(Z)\|_{\psi_2} \leq \sigma_{\varepsilon}(X) \) 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 \( \pi(\cdot) - \pi(\cdot) \)). 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 \( \pi(\cdot) - \pi(\cdot) \) satisfies a (pointwise) tail bound only at the \( n \) training points \( \{X_i\}_{i=1}^n \) as follows. For any \( c \geq 0 \) and for some constant \( C \geq 0 \), we have:

\[
\mathbb{P}\{|\pi(X_i) - \pi(X_i)| > cv_{n,\pi} + b_{n,\pi}\} \leq C \exp(-c^2) + q_{n,\pi} \quad \forall \ 1 \leq i \leq n. \tag{3.6}
\]

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 \( x \in \mathcal{X} \), any \( c \geq 0 \) and for some constant \( C \geq 0 \), we have:

\[
\mathbb{P}\{|\hat{m}(x) - m(x)| > cv_{n,m} + b_{n,m}\} \leq C \exp(-c^2) + q_{n,m}. \tag{3.7}
\]

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 \( X_i \in D_n^{(k')} \perp D_n^{(k)} \),

\[
\mathbb{P}\{|\hat{m}^{(k)}(X_i) - m(X_i)| > cv_{n,m} + b_{n,m}\} \leq C \exp(-c^2) + q_{\bar{n},m}. \tag{3.8}
\]

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 \( (v_{n,\pi}\sqrt{\log n} + b_{n,\pi})(v_{n,m}\sqrt{\log nd} + b_{n,m}) = o\{\sqrt{(\log d)/n}\} \).
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 rigorously 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 Sections 5-6, 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, 6.1 and 6.2) 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 \(\sigma_0 := 2\sqrt{2} \sigma_h (\sigma_\psi + \sigma_\varepsilon \delta_\pi^{-1}) \geq 0\) and \(K_0 := 2\sigma_h (\sigma_\psi + \sigma_\varepsilon \delta_\pi^{-1}) \geq 0\) are constants. Hence, with \(\epsilon = c\sqrt{\log d}/n\), for any constant \(c > 1\), we have:

With probability \(\geq 1 - \frac{4}{d^{c^2-1}}\),
\[
\|T_{0,n}\|_{\infty} \leq c \sqrt{\frac{\log d}{n}} \sqrt{2} \sigma_0 + c^2 \frac{\log d}{n} K_0 \lesssim \sqrt{\frac{\log d}{n}}.
\]

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 \((\delta_\pi, \sigma_\varepsilon, C)\) being as defined therein, and let \(\|\mu_h^{(2)}\|_{\infty} := \max\{\mathbb{E}\{h_j^2(X)\} : j = 1, \ldots, d\}\). Then, for any constants \(c_1, c_2, c_3 > 1\), where we further assume w.l.o.g. that \(c_2 v_{n,\pi} \sqrt{\log(n d)} + b_{n,\pi} \leq \delta_\pi/2 < \delta_\pi\) and \(c_3 \sqrt{\log d}/n < 1\), we have:

With probability \(\geq 1 - \frac{2}{d^{c^2-1}} - \frac{4}{d^{c^2-1}} - \sum_{j=1}^{2} \frac{2C}{nd} \left( \frac{\log d}{n} \right) \frac{1}{c^2-1} - 4q_{n,\pi}(nd),\)
\[
\|T_{\pi,n}\|_{\infty} \leq c \sqrt{\frac{\log d}{n}} \left\{ v_{n,\pi} \sqrt{\log(n d)} + b_{n,\pi} \right\} C_1 \left( \frac{\|\mu_h^{(2)}\|_{\infty}}{\delta_\pi} + C_2 \sqrt{\frac{\log d}{n}} \right)^{1/2},
\]
where \(C_1 := c_1 (4\sqrt{2} \sigma_\varepsilon/\delta_\pi)\) 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 \(\|T_{\pi,n}\|_{\infty} \lesssim \sqrt{\log d}/n\), \(\{v_{n,\pi} \sqrt{\log(n d)} + b_{n,\pi}\} = o\{\sqrt{\log d}/n\}\) w.h.p. In the proof of Theorem 3.2, we also provide a general result (Theorem 10.1) 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 \(\{\tilde{m}(X_i)\}_{i=1}^{n}\) 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
(δπ, C) being as defined therein. Then, for any constants c, c1, c2 > 1, where we further assume w.l.o.g. that c2√(log d)/n < 1, we have:

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

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

REMARK 3.4. Theorem 3.3 therefore shows that \(\|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} (\log d)/n\) \(\{v_{\bar{n},m} \sqrt{\log(\bar{n}d)} + b_{\bar{n},m}\} = o\{\sqrt{\log(\bar{n}d)/n}\} \) w.h.p. In the proof of Theorem 3.3, we also provide a general result (Theorem 11.1) 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 \(R_{\pi,m,n}(Z)\) 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 \((v_{\bar{n},m}, b_{\bar{n},m}, q_{\bar{n},m})\), \((v_{\bar{n},m}, b_{\bar{n},m}, q_{\bar{n},m}, \bar{m})\) and the constants \(\{\delta_\pi, C\}\) as defined therein. Then, for any constants \(c_1, c_2, c_3, c_4 > 1\) with \(c_2v_{\bar{n},m}\sqrt{\log \bar{n}} + b_{\bar{n},m} \leq \delta_\pi/2 < \delta_\pi\) and \(c_4\sqrt{(\log d)/n} < 1\), we have:

\[
\|R_{\pi,m,n}\|_\infty \leq c_1c_3\bar{C}_1r_{\pi,n}r_{\bar{m},\bar{n}} \left( \|\mu_h\|_\infty + c_4\bar{C}_2 \sqrt{\frac{\log d}{n}} \right),
\]

where \(r_{\pi,n} := v_{\bar{n},m}\sqrt{\log \bar{n}} + b_{\bar{n},m}\) and \(r_{\bar{m},\bar{n}} := v_{\bar{n},m}\sqrt{\log \bar{n}} + b_{\bar{n},m}\), with \(r_{\pi,n}r_{\bar{m},\bar{n}} = o\{\sqrt{(\log d)/n}\}\), \(\|\mu_h\|_\infty := \max_{1 \leq j \leq d} \mathbb{E}\{h_{\bar{m},\bar{n}}(X)\}\) and \(\bar{C}_1 := 2/\delta_\pi\), \(\bar{C}_2 := \sqrt{2}\sigma_{\bar{m}} + K_{\bar{m}}\) are constants with \(\sigma_{\bar{m}} := 4\sigma_h\delta_\pi^{-1}\) and \(K_{\bar{m}} := 2\sqrt{2}\sigma_h\delta_\pi^{-1} - 1\).

REMARK 3.5. Theorem 3.4 therefore shows that \(\|R_{\pi,m,n}\|_\infty \lesssim r_{\pi,n}r_{\bar{m},\bar{n}} = o\{\sqrt{(\log d)/n}\}\) w.h.p., where the last step is by assumption a sufficient condition for which is \(max\{r_{\pi,n}, r_{\bar{m},\bar{n}}\} \lesssim \{(\log d)/n\}^{0.25}\). In the proof of Theorem 3.4, we provide a general result (Theorem 12.1) on tail bounds for \(R_{\pi,m,n}\).

3.7. Discussions on the Structure of the Main Error Terms and the Benefits of Cross-Fitting the Conditional Mean Estimator.
The structure of $T_{\pi,n}$ and the benefits of 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: $E\{T_{\pi}(Z_i)\} = E\{E\{T_{\pi}(Z_i) | \hat{\pi}(\cdot), X_i\} \} = E\{E\{T_{\pi}(Z_i) | \hat{\pi}(\cdot), X_i\} \} = 0$, 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 $E\{Y_i - m(X_i)\} | X_n \equiv E\{\varepsilon(Z_i) | X_n\}$ is a centred average of (conditionally) independent variables, $T_{\pi,n}$ 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 $D_{n,Y}$ (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, $T_{\pi,n}$ is a centred average of (conditionally) independent variables. We exploit this fact and the structure of $T_{\pi}(Z_i)$ in order to control $T_{\pi,n}$.

The structure of $T_{m,n}$ and the benefits of cross-fitting for estimating $m(\cdot)$. 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 $\{T_{\pi}(Z_i)\}_{i=1}^n \equiv X_n$ 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 $\{\hat{m}(X_i)\}_{i=1}^n$ 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 $Z \perp \hat{m}(\cdot)$, $\{T_{m}(Z_k)\}_{k=1}^K$ 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 \( \{ X_i \in D_n^{(k)} \}_{i=1}^{n} \) 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 \( \hat{\theta}_{\text{DDR}} \) 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 \( \hat{\theta}_{\text{DDR}} \). 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 \( \hat{\theta}_{\text{DDR}} \). 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: \( L(Y, X, \theta) = (Y - \Psi(X)'\theta)^2 \), 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:

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

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 \( M_{ij} \in \mathbb{R} \), \( M_{i[\cdot]} \in \mathbb{R}^d \) and \( M_{[\cdot]j} \in \mathbb{R}^d \) respectively denote the \((i, j)^{th}\) entry, the \(i^{th}\) row and the \(j^{th}\) column of \( M \). Further, let \( ||M||_1 := \)
notations defined in the basic decomposition (3.1) of \(T_\tilde{\psi}(4.3)\) outcomes defined in Section 2.3. Using (4.2), it then follows easily that

\[
\max \frac{1}{2} \mathbf{T}_0(Z) = \{m(X) - \Psi(X)'\theta_0\} \Psi(X) + \frac{T}{\pi(X)} \{Y - m(X)\} \Psi(X).
\]

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 \(\hat{\theta}_{DDR}\) in (2.5), or equivalently in (2.6), the corresponding desparsified DDR estimator \(\hat{\theta}_{DDR}\) is defined as follows.

\[
\begin{align*}
\hat{\theta}_{DDR} &= \hat{\theta}_{DDR} - \frac{1}{2} \hat{\Omega} \nabla \mathcal{L}_{DDR}^\Omega(\hat{\theta}_{DDR}) = \hat{\theta}_{DDR} - \frac{1}{2} \hat{\Omega} \nabla \mathcal{L}_{DDR}^\Omega(\hat{\theta}_{DDR}) \\
&= \hat{\theta}_{DDR} + \hat{\Omega} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \Psi(X_i)'\hat{\theta}_{DDR}\} \Psi(X_i), \text{ where}
\end{align*}
\]

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

\[
\begin{align*}
\hat{\theta}_{DDR} - \theta_0 &= \frac{1}{n} \sum_{i=1}^{n} \psi_0(Z_i) + R_{n,1} + R_{n,2} + \Delta_n, \text{ where}
\end{align*}
\]

\[
\begin{align*}
2\psi_0(Z) &:= -\Omega \mathbf{T}_0(Z) \text{ with } \mathbb{E}\{\psi_0(Z)\} = 0, \ 2R_{n,1} := (\Omega - \hat{\Omega}) \nabla \mathcal{L}_{DDR}(\theta_0), \\
2R_{n,2} &:= -\Omega (\mathbf{T}_{\pi,n} + \mathbf{T}_{m,n} + \mathbf{R}_{\pi,m,n}) \text{ and } \Delta_n := (I - \hat{\Omega} \hat{\Sigma})(\hat{\theta}_{DDR} - \theta_0).
\end{align*}
\]

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.

5.1. Estimation of the Propensity Score: Choices and Their Properties. In some cases, \(\pi(\cdot)\) may be known whereby \(\hat{\pi}(\cdot) \equiv \pi(\cdot)\) trivially. When \(\pi(\cdot)\) is unknown, we consider two (class of) choices for estimating \(\pi(\cdot)\) as follows.

1. \(\pi(\cdot)\) belongs to an (extended) parametric family given by \(\pi(x) \equiv \mathbb{E}(T|X = x) = g(\alpha'\Psi(x))\), 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 with \(K\) allowed to depend on \(n\) (and possibly \(K \gg n\)), and \(\alpha \in \mathbb{R}^K\) is an unknown parameter vector that is further assumed to be sparse (if required).
Estimator. $\pi(x)$ 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 $(T, X)$ and 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 $g(u) = g_{\text{expit}}(a) := \exp(a)/(1 + \exp(a))$. 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, $\|\hat{\alpha} - \alpha\|_1 \leq a_n \equiv a_n(s_\alpha, K) := s_\alpha \sqrt{(\log K)/n}$ 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, of any degree $d_0 \geq 1$, given by: $\Psi(x) := \{1, x_j : 1 \leq j \leq p, 1 \leq k \leq d_0\}$. 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 parametric framework also includes the 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 $O(n^{-1/2})$ via the usual sample mean of $T$.

2. $\pi(\cdot)$ belongs to a semi-parametric single index family, wherein the framework is similar to the parametric framework with $\Psi(x) \equiv x$ but is more general and flexible in that the function $g(\cdot)$ itself is further allowed to be unknown. Specifically, we assume here that $\pi(x) \equiv \mathbb{E}(T|X = x) = g(\alpha'x)$ with $\alpha \in \mathbb{R}^p$ and $g(\cdot)$ being both unknown, so that $\alpha$ is identifiable only upto scalar multiples, and again, we assume $\alpha_0$ to be sparse (if required).

Estimator. Given any reasonable estimator $\hat{\alpha}$ of the $\alpha$ ‘direction’, obtained based on some suitable procedure on the observed data for $(T, X)$, we then estimate $\pi(x) \equiv \mathbb{E}(T|\alpha'X = \alpha'x) = g(\alpha'x)$, under appropriate smoothness and regularity assumptions, via a one-dimensional kernel smoothing (KS) over the estimated scores $\{\hat{\alpha}'X_i\}_{i=1}^n$, as follows.

$$\hat{\pi}(x) \equiv \hat{g}(\hat{\alpha}'x) \equiv \hat{g}(\hat{\alpha}, x) := \frac{1}{nh} \sum_{i=1}^n T_i K\left(\frac{\hat{\alpha}'x_i - \hat{\alpha}'x}{h}\right) \quad \forall \ x \in \mathcal{X},$$

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
\( \hat{\alpha} \) is some reasonable estimator of the \( \alpha \) direction satisfying a basic ‘high level’ condition: \( \| \hat{\alpha} - \alpha_0 \|_1 \leq a_n \) w.h.p. for some \( \alpha_0 \propto \alpha \) and \( a_n = o(1) \).

**Estimation of \( \alpha \).** As mentioned above, we only require some reasonable estimator \( \hat{\alpha} \) of the \( \alpha \) direction that satisfies an \( L_1 \) norm bound w.h.p. If the underlying design distribution of \( X \) is elliptically symmetric, then owing to the results of Li and Duan (1989), such an estimator may still be obtained using a simple \( L_1 \)-penalized logistic regression of \( T \) vs. \( X \) with a rate guarantee of \( a_n = s_\alpha \sqrt{(\log p)/n} \), as noted in the previous example, even though \( T|X \) now satisfies a single index model (not necessarily a parametric logistic regression model). Similar approaches have been extensively used in recent years in the high dimensional statistics and compressed sensing literature for sparse signal recovery in high dimensional SIMs (Plan and Vershynin, 2013, 2016; Goldstein, Minsker and Wei, 2016; Genzel, 2017; Plan, Vershynin and Yudovina, 2017; Wei, 2018). More generally, one may also 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 data \( \{(T_i, X_i)\}_{i=1}^n \) to obtain an estimator of \( \alpha \).

5.2. **Estimation of the Conditional Mean: Choices and Their Properties.**

We consider the following two (class of) choices for estimating \( m(\cdot) \).

1. \( m(\cdot) \) belongs to an (extended) parametric family with \( m(x) = g(\gamma' \Psi(X)) \) 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 (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 \( m(x) \equiv \mathbb{E}(Y|X) \equiv \mathbb{E}(Y|X, T = 1) = g(\gamma' \Psi(X)) \) as \( \hat{m}(x) = g(\hat{\gamma}' \Psi(X)) \), where \( \hat{\gamma} \) denotes some given estimator of \( \gamma \) obtained via any suitable estimation procedure based on the ‘complete case’ data \( D_n^{(c)} := \{(Y_i, X_i) \ | \ T_i = 1\}_{i=1}^n \), and satisfies a basic ‘high level’ requirement that \( \| \hat{\gamma} - \gamma \|_1 \leq a_n \) 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, \( g(u) \equiv g_{expit} = \exp(u)/(1 + \exp(u)) \) (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, of
any degree \( d_0 \geq 1 \), given by: \( \Psi(x) := \{1, x_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\} \).

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 \( \hat{\gamma} \).** For all the examples above, with \( g(\cdot) \) being any ‘canonical’ link function, the estimator \( \hat{\gamma} \) 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 \( \mathcal{D}_n^{(c)} \) 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, \( \|\hat{\gamma} - \gamma\|_1 \lesssim a_n \equiv a_n(s_\gamma, K) := s_\gamma \sqrt{(\log K)/n} \) w.h.p., where \( s_\gamma \equiv \|\gamma\|_0 \) denotes the sparsity of \( \gamma \).

2. \( m(\cdot) \) belongs to a semi-parametric single index family given by \( m(X) \equiv E(Y|X) \equiv E(Y|X, T = 1) = g(\gamma'X) \), 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 \( \mathcal{D}_n \), we then estimate \( m(X) \equiv E(Y|\gamma'X) \equiv E(Y|\gamma'X, T = 1) = g(\gamma'X) \), under appropriate smoothness and regularity assumptions, via a one-dimensional KS over the estimated scores \( \{\hat{\gamma}'X_i\}_{i=1}^n \), as follows.

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

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, \( E(Y|X) \equiv E(Y|X, T = 1) = g(\gamma'X) \). 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 \( \mathcal{D}_n^{(c)} \) 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: $(Y | X, T = 1) = f(\gamma' X; \epsilon)$ for some unknown function $f : \mathbb{R}^2 \rightarrow \mathcal{Y}$ and some noise $\epsilon \perp (T, X)$, and assume further that the distribution of $(X | T = 1)$ 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 ‘tar-
Theorem 5.1. Suppose that \( \hat{\beta} \) satisfies a basic 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. \( \max_{1 \leq j \leq K} |\Psi[j](X)| \leq C_\Psi < \infty \) a.s. \([\mathbb{P}_X]\) for some constant \( C_\Psi \geq 0 \). Then, for any \( t \geq 0 \),

\[
\mathbb{P} \left[ \sup_{x \in \mathcal{X}} |g(\hat{\beta}' \Psi(x)) - g(\beta_0' \Psi(x))| > (\sqrt{2} C_g C_\Psi) a_n t \right] \leq 2 \exp(-t^2) + q_n.
\]

The proof of Theorem 5.1 is quite straightforward and follows from a simple application of Hoeffding’s inequality along with the union bound and use of the assumptions made. The proof is therefore skipped here for brevity.

6. Kernel Smoothing over Estimated Transformations of High Dimensional Covariates: Non-Asymptotic Analyses and Rates. In this section, we study the properties of kernel smoothing estimators involving smoothing over single index transformations of high dimensional covariates, where the index parameter is further allowed to be estimated. The underlying high dimensional setting and the possibly non-ignorable estimation error arising from the index estimation makes the analyses challenging, nuanced and different from results that are available in the literature under classical settings. We develop here a non-asymptotic theory that establishes concrete tail bounds and pointwise convergence rates for such estimators, and applies equally to any regime - classical or high dimensional. We consider both linear kernel average estimators (e.g. density estimators) as well as ratio form estimators (e.g. conditional mean estimators). The results herein, while they are obtained in course of characterizing our nuisance function estimators' properties, may also be useful in several other applications and are 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 \( Z \in \mathbb{R} \) be a generic random variable and \( X \in \mathbb{R}^p \) be a random vector of covariates with support \( \mathcal{X} \subseteq \mathbb{R}^p \). We assume that \((Z, X)\) has a well defined joint distribution with finite \( 2^{nd} \) moments. Let \( \{(Z_i, X_i) : i = 1, \ldots, n\} \) be a sample of \( n \geq 2 \) i.i.d. realizations of \((Z, X)\), where we note that \( p \geq 1 \) is
allowed to be high dimensional w.r.t. \( n \), 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’ guarantee (to be made precise shortly) in terms of an \( L_1 \) norm bound given by: \( \| \hat{\beta} - \beta \|_1 \leq a_n \) with high probability for some sequence \( a_n = o(1) \).

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

Given any estimator \( \hat{\beta} \) of \( \beta \), we now consider estimating the functions \( l(\beta, x) \), \( f(\beta, x) \) and \( m(\beta, x) \), for any fixed \( x \in \mathcal{X} \), using kernel smoothing (KS). Let \( K(\cdot) : \mathbb{R} \rightarrow \mathbb{R} \) be any suitable kernel function (e.g. the Gaussian kernel) and let \( h \equiv h_n > 0 \) be any bandwidth sequence with \( h_n = o(1) \). Consider the KS estimators \( \hat{l}(\beta, x) \), \( \hat{f}(\beta, x) \) and \( \hat{m}(\beta, x) \) given by:

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

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

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

**Assumption 6.1** (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 \( m_\beta^{(2)}(w) := \mathbb{E}(Z^2 \mid \beta'X = w) \) for any \( w \in \mathbb{R} \). Then, \( m_\beta^{(2)}(w)f_\beta(w) \) is bounded in \( w \in \mathbb{R} \) and \( \| m_\beta^{(2)}(\cdot)f_\beta(\cdot) \|_{\infty} \leq B_1 \) for some constant \( B_1 \geq 0 \).
(d) $K(\cdot)$ is a second order kernel satisfying: $\int_{\mathbb{R}} K(u)du = 1$, $\int_{\mathbb{R}} uK(u)du = 0$ and $\int_{\mathbb{R}} u^2|K(u)|du \leq R_K < \infty$ for some constant $R_K \geq 0$. $l_\beta(\cdot) \equiv m_\beta f_\beta(\cdot)$ 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 6.2** (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 $u \to \pm \infty$.

(b) Let $\eta_\beta(w) := \mathbb{E}(Z\mathbf{X}|\beta'\mathbf{X}=w)f_\beta(w)$ 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 $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}^+$ with $\|\varphi(\cdot)\|_\infty \leq M_\varphi$ and $\int_{\mathbb{R}} \varphi(u)du \leq C_\varphi$ for some constants $M_\varphi,C_\varphi \geq 0$.

(d) $\mathbf{X}$ is bounded, i.e. $\|\mathbf{X}\|_\infty \leq M_\mathbf{X}$ a.s. $[\mathbb{P}]$ for some constant $M_\mathbf{X} \geq 0$, and 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 $2M_\mathbf{X}(a_n/h) \leq L$ where $L$ is as in (c) above.

Most of the smoothness assumptions and the conditions on $K(\cdot)$ in Assumptions 6.1 and 6.2 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 6.2 (c) on $K'(\cdot)$ essentially imposes some kind of a ‘local’ Lipschitz property on $K'(\cdot)$, where the Lipschitz constant is defined locally as a function that is bounded and further decays quickly enough to be integrable. In particular, this is satisfied by the Gaussian kernel. 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) for details).

Finally, the boundedness assumption on $\mathbf{X}$ is mostly for simplicity in
the exposition of the bounds. 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. We therefore stick to a bounded $X$ to avoid such technicalities.

**Theorem 6.1 (Tail bounds for the LKA estimators).** Consider the LKA estimator $\hat{l}(\beta, x)$ of $l(\beta, x)$. Suppose Assumptions 6.1 and 6.2 hold and that $h = o(1)$, $(\log n)/(nh) = o(1)$ and $a_n/h = o(1)$. Then, for each fixed $x \in \mathcal{X}$ and any $t \geq 0$, we have: with probability at least $1 - 9 \exp(-t^2) - 2q_n$,

$$|l(\hat{l}, x) - l(\beta, x)| \leq C_1 \frac{t}{\sqrt{nh}} + C_2 \left( h^2 + a_n + \frac{a_n^2}{h^2} + \frac{a_n \sqrt{\log p}}{\sqrt{nh^3}} \right),$$

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

Further, a similar tail bound holds at the training points $\{X_i\}_{i=1}^n$. For any $1 \leq i \leq n$ and any $t \geq 0$, we have: with probability $\geq 1 - 11 \exp(-t^2) - 2q_n$,

$$|l(\hat{l}, x_i) - l(\beta, x_i)| \leq C'_1 \frac{t}{\sqrt{nh}} + C'_2 \left( h^2 + a_n + \frac{a_n^2}{h^2} + \frac{a_n \sqrt{\log p}}{\sqrt{nh^3}} \right),$$

for some constants $C'_1, C'_2 > 0$ depending only on those in the assumptions.

**Theorem 6.2 (Tail bounds for ratio type KS estimators).** Consider the ratio type KS estimator $\hat{m}(\beta, x)$ of $m(\beta, x)$. Suppose Assumptions 6.1 and 6.2 hold and that $h = o(1)$, $(\log n)/(nh) = o(1)$ and $a_n/h = o(1)$. Define:

$$v_n := \frac{1}{\sqrt{nh}} = o(1) \text{ and } b_n := h^2 + a_n + \frac{a_n^2}{h^2} + \frac{a_n \sqrt{\log p}}{\sqrt{nh^3}} = o(1).$$

Assume further that $|m(\beta, x)| \leq \delta_m$ and $f(\beta, x) \geq \delta_f > 0$ for some constants $\delta_m, \delta_f > 0$, and let $C_1, C_2, C'_1, C'_2 > 0$ be the same constants as in Theorem 6.1. Then, for any fixed $x \in \mathcal{X}$ and any $t, t_* \geq 0$ with $t_*$ further assumed w.l.o.g. to be small enough to satisfy $C_1 t_* v_n + C_2 b_n \leq \delta_f/2 < \delta_f$, we have:

$$P \left\{ |\hat{m}(\beta, x) - m(\beta, x)| > 2\delta_f^{-1}(\delta_m + 1) \left( C_1 \frac{t}{\sqrt{nh}} + C_2 b_n \right) \right\} \leq 18 \exp(-t^2) + 9 \exp(-t_*^2) + 6q_n, \quad \forall x \in \mathcal{X}.$$

Further for any $1 \leq i \leq n$ and any $t, t_* \geq 0$ with $t_*$ assumed to be small enough to satisfy $C'_1 t_* v_n + C'_2 b_n \leq \delta_f/2 < \delta_f$, we have:

$$P \left\{ |\hat{m}(\beta, x_i) - m(\beta, x_i)| > 2\delta_f^{-1}(\delta_m + 1) \left( C'_1 \frac{t}{\sqrt{nh}} + C'_2 b_n \right) \right\} \leq 22 \exp(-t^2) + 11 \exp(-t_*^2) + 6q_n, \quad \forall 1 \leq i \leq n.
6.1. Proof Sketch for Theorems 6.1 and 6.2. To analyze the behavior of \(\widehat{l}(\beta, x)\), we first introduce the corresponding 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 (hypothetical) estimator:

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

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

\[
|\widehat{l}(\beta, x) - l(\beta, x)| \leq |\widehat{l}(\beta, x) - \widetilde{l}(\beta, x)| + |\widetilde{l}(\beta, x) - l(\beta, x)|
\]

\[
\leq |\widehat{l}(\beta, x) - \mathbb{E}[\widetilde{l}(\beta, x)]| + \|\mathbb{E}[\widetilde{l}(\beta, x)] - l(\beta, x)\| + |\widetilde{l}(\beta, x) - l(\beta, x)|
\]

\[
=: |\bar{S}_n(x)| + |\bar{S}_n(x)| + |\bar{R}_n(x)| \quad (\text{say}).
\]

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

**Lemma 6.1** (Characterizing the tail bounds for \(\bar{S}_n(x)\) and \(\bar{S}_n(x)\)). Under Assumption 6.1 (a)-(c), we have: for any fixed \(x \in \mathcal{X}\) and any \(t \geq 0\),

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

where \(C_1 := 7(B_1 C_K M_K)^{1/2}\) and \(C_2 := D \sigma_Z M_K\) for some absolute constant \(D > 0\). Further, under Assumption 6.1 (d), we have:

\[
|\bar{S}_n(x)| \leq C_3 h^2 \quad \text{uniformly in } x \in \mathcal{X}, \quad \text{where } C_3 := B_2 R_K.
\]

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

\[
|\widehat{l}(\beta, x) - l(\beta, x)| \leq t \left( \frac{C_1}{\sqrt{nh}} + C_2 \frac{\sqrt{n \log n}}{nh} \right) + C_3 h^2, \quad \forall x \in \mathcal{X}.
\]

Next, we aim to control the term \(\bar{R}_n(x)\) 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 \(\bar{l}(\beta, x)\) around \(l(\beta, x) \equiv \bar{l}(\beta, x)\), we first rewrite \(\bar{R}_n(x) \equiv \bar{l}(\beta, x) - \bar{l}(\beta, x)\) as:

\[
\bar{R}_n(x) = (\bar{\beta} - \beta)' \left\{ \frac{1}{nh} \sum_{i=1}^{n} Z_i \frac{(X_i - x)}{h} K' \left( \frac{W_i - w}{h} \right) \right\}, \quad \text{where}
\]

\[
(6.2) \quad \bar{R}_n(x) = (\bar{\beta} - \beta)' \left\{ \frac{1}{nh} \sum_{i=1}^{n} Z_i \frac{(X_i - x)}{h} K' \left( \frac{W_i - w}{h} \right) \right\}, \quad \text{where}
\]

\[
(6.1) \quad |\bar{l}(\beta, x) - l(\beta, x)| \leq t \left( \frac{C_1}{\sqrt{nh}} + C_2 \frac{\sqrt{n \log n}}{nh} \right) + C_3 h^2, \quad \forall x \in \mathcal{X}.
\]
\{W_i^*\}_{i=1}^n and \(w_\mathbf{x}^*\) are ‘intermediate’ points that satisfy, for each \(i = 1, \ldots, n\),
\[|(W_i^* - w_\mathbf{x}^*) - (W_i - w_\mathbf{x})| \leq |(\tilde{W}_i - \tilde{w}_\mathbf{x}) - (W_i - w_\mathbf{x})| \equiv |(\beta - \beta')(\mathbf{x}_i - \mathbf{x})/h|.

We now rewrite the expansion above as:
\[
\hat{R}_n(x) \equiv \hat{R}_{n,1}(x) + \hat{R}_{n,2}(x),
\]
where
\[
(6.3) \quad \hat{R}_{n,1}(x) := (\beta - \beta') \left\{ \frac{1}{nh} \sum_{i=1}^{n} \frac{Z_i}{h} (\mathbf{X}_i - \mathbf{x}) K' \left( \frac{W_i - w_\mathbf{x}}{h} \right) \right\}
\]
\[
\quad =: (\beta - \beta') \hat{T}_n(x) \quad \text{(say), and } \quad \hat{R}_{n,2}(x) := \hat{R}_n(x) - \hat{R}_{n,1}(x).
\]

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

**Lemma 6.2 (Characterizing the tail bounds for \(\hat{S}_n(x)\) and \(\overline{S}_n(x)\)).** Under Assumption 6.2 (a), (b) and (d), and Assumption 6.1 (a) and (c), we have: for any fixed \(x \in \mathcal{X}\) and any given \(t \geq 0\),
\[
\mathbb{P} \left\{ |\hat{R}_{n,1}(x)| > C^*_1 a_n + t \left( C^*_2 \frac{a_n}{\sqrt{nh^3}} + C^*_3 \frac{a_n \sqrt{\log n}}{nh^2} \right) \right\}
\leq 3 \exp(-t^2 + \log p) + q_n, \quad \text{where}
\]
\(C^*_1, C^*_2, C^*_3 > 0\) are constants depending only on the constants introduced in Assumptions 6.2 and 6.1.

Further, under the additional condition in Assumption 6.2 (c), we have:
\[
\mathbb{P} \left\{ |\hat{R}_{n,2}(x)| > 4M_\mathbf{X}^* C^*_4 \frac{a_n^2}{h^2} + 4M_\mathbf{X}^* t \left( C^*_5 \frac{a_n^2}{\sqrt{nh^3}} + C^*_6 \frac{a_n^2 \sqrt{\log n}}{nh^3} \right) \right\}
\leq 3 \exp(-t^2) + q_n, \quad \text{for any fixed } x \in \mathcal{X} \text{ and any given } t \geq 0, \quad \text{where}
\]
\(C^*_4, C^*_5, C^*_6 > 0\) are constants depending only on the constants introduced in Assumptions 6.1 and 6.2.

Assuming that \(h = o(1)\), \((\log n)/(nh) = o(1)\) and \(a_n/h = o(1)\), note that the second and third terms in the tail bound for \(\hat{R}_{n,2}(x)\) are each dominated by the respective terms in the tail bound for \(\hat{R}_{n,1}(x)\) in Lemma 6.2, and further, the third term in the latter bound is dominated by the second term as well. Using these facts we now obtain a tail bound for \(\hat{R}_n(x)\) as follows. For any fixed \(x \in \mathcal{X}\) and any \(t \geq 0\), with probability \(\geq 1 - 6 \exp(-t^2 - 2q_n)\),
\[
\hat{R}_n(x) \equiv \hat{l}(\beta, x) - \tilde{l}(\beta, x) \leq C^*_1 a_n + C^*_2 \frac{a_n^2}{h^2} + C^*_3 \frac{a_n}{\sqrt{nh^3}} (t + \sqrt{\log p}),
\]
for some constants \(C^*_1, C^*_2, C^*_3 > 0\) (possibly different from those in Lemma 6.2) depending only on the constants defined in Assumptions 6.1 and 6.2.
7. The DR Aspect: Consistency Guarantees Even if One of the Two Nuisance Functions’ Estimators are Misspecified. Our entire probabilistic analysis of \( \|T_n\|_\infty \equiv \|\nabla \mathcal{L}^{\text{DDR}}_n(\theta_0)\|_\infty \) (with the larger goal of 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 their respective estimators \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \), as characterized in Assumptions 3.2 and 3.3. However, as highlighted in (1.4), the population version \( \mathcal{L}_{\text{DDR}}(\cdot) \) of the empirical DDR loss \( \mathcal{L}^{\text{DDR}}_n(\cdot) \) 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. 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 construction 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.

Case 1. \( \hat{\pi}(\cdot) \) is misspecified but \( \hat{m}(\cdot) \) is correctly specified. Suppose \( \hat{\pi}(\cdot) \xrightarrow{p} \pi^*(\cdot) \neq \pi(\cdot) \), while \( \hat{m}(\cdot) \xrightarrow{p} m(\cdot) \) still. In this case, there are two terms in the basic decomposition (3.1) of \( T_n \), namely \( T_{\pi,n} \) and \( R_{\pi,m,n} \), that would be affected and need to be further decomposed appropriately into two terms each, and carefully analysed, as follows. \( T_{\pi,n} \) needs to be decomposed as:

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

while \( R_{\pi,m,n} \) needs to be appropriately decomposed as:

\[
R_{\pi,m,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi(X_i)} - \frac{T_i}{\pi^*(X_i)} \right\} \{\hat{m}(X_i) - m(X_i)\} h(X_i) \\
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi^*(X_i)} - \frac{T_i}{\pi(X_i)} \right\} \{\hat{m}(X_i) - m(X_i)\} h(X_i)
\]
(7.2) \[ \tilde{R}_{\pi,m,n} + R^*_{\pi,m,n} \quad \text{(say)}. \]

Case 2. \( \hat{m}(\cdot) \) is misspecified but \( \tilde{\pi}(\cdot) \) is correctly specified. Suppose \( \hat{m}(\cdot) \not\xrightarrow{p} m^*(\cdot) \not\xrightarrow{p} m(\cdot) \), while \( \tilde{\pi}(\cdot) \not\xrightarrow{p} \pi(\cdot) \) still. In this case, there are two terms in the basic decomposition (3.1) of \( T_R \), namely \( T_{m,n} \) and \( R^*_{\pi,m,n} \) that would be affected and need to be further decomposed appropriately into two terms each, and carefully analysed, as follows. \( T_{m,n} \) needs to be decomposed as:

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

(7.3) \[ =: \bar{T}_{m,n} + T^*_{\pi,m,n} \quad \text{(say)}, \]

while \( R^*_{\pi,m,n} \) needs to be appropriately decomposed as:

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

(7.4) \[ =: R^*_{m,n} + R^*_{\pi,m,n} \quad \text{(say)}. \]

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.

8. Simulation Studies. Results are available and will be added soon.

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

\[
(9.1) \quad T_{0,n} = T_{0,n}^{(1)} + T_{0,n}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} T_0^{(1)}(Z_i) + \frac{1}{n} \sum_{i=1}^{n} T_0^{(2)}(Z_i), \quad \text{where}
T_0^{(1)}(Z) := \{ m(X) - g(X, \theta_0) \} h(X) \quad \text{and} \quad T_0^{(2)}(Z) := \frac{T - \pi(X)}{\pi(X)} \{ Y - m(X) \} h(X),
\]

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 13.5 (a), we have:

\[
(9.2) \quad T_{0,j}[j]^{(1)}(Z) \equiv \psi(X) h_{[j]}(X) \sim \text{BMC}(\bar{\sigma}_1, \bar{K}_1) \quad \forall j \in \{1, \ldots, d\},
\]
for some constants $\bar{\sigma}_1 := 2\sqrt{2}\sigma_v \sigma_h \geq 0$ and $\bar{K}_1 := 2\sigma_v \sigma_h \geq 0$.

Next, using Assumption 3.1 (a) and Lemma 13.1 (v), $\|\varepsilon(Z) h_{ij}(X)\|_{\psi_1} \leq \sigma_v \sigma_h$ for each $j \in \{1, \ldots, d\}$. Further, owing to Assumption 1.1 (b) and (1.1), $T/\pi(X) \leq \delta^{-1} \pi$ a.s. [P]. Hence, using Lemma 13.5 (b), we have

$$\text{(9.3)} \quad T_{0[j]}^{(2)}(Z) \equiv \frac{T}{\pi(X)} \varepsilon(Z) h_{ij}(X) \sim \text{BMC}(\bar{\sigma}_2, \bar{K}_2) \quad \forall \ j \in \{1, \ldots, d\},$$

for some constants $\bar{\sigma}_2 := 2\sqrt{2}\sigma_v \sigma_h \delta^{-1}_\pi \geq 0$ and $\bar{K}_2 := 2\sigma_v \sigma_h \delta^{-1}_\pi \geq 0$.

Hence, (9.2) and (9.3) 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 13.4) to apply. Using Lemma 13.4, we then have: for any $\epsilon_1 \geq 0$,

$$\mathbb{P} \left\{ \left\| T_{0,n}^{(1)} \right\| \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(1)}(Z_i) \right\|_\infty > \sqrt{2\bar{\sigma}_1 \epsilon_1 + \bar{K}_1 \epsilon_1^2} \right\}$$

$$\leq \sum_{j=1}^{d} \mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0[j]}^{(1)}(Z_i) \right\| > \sqrt{2\bar{\sigma}_1 \epsilon_1 + \bar{K}_1 \epsilon_1^2} \right\}$$

$$\leq \sum_{j=1}^{d} \left\{ 2 \exp \left( -n\epsilon_1^2 \right) \right\} = 2d \exp \left( -n\epsilon_1^2 \right) \equiv 2 \exp \left( -n\epsilon_1^2 + \log d \right), \quad (9.4)$$

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

$$\mathbb{P} \left\{ \left\| T_{0,n}^{(2)} \right\| \equiv \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0}^{(2)}(Z_i) \right\|_\infty > \sqrt{2\bar{\sigma}_2 \epsilon_2 + \bar{K}_2 \epsilon_2^2} \right\}$$

$$\leq \sum_{j=1}^{d} \mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} T_{0[j]}^{(2)}(Z_i) \right\| > \sqrt{2\bar{\sigma}_2 \epsilon_2 + \bar{K}_2 \epsilon_2^2} \right\}$$

$$\leq \sum_{j=1}^{d} \left\{ 2 \exp \left( -n\epsilon_2^2 \right) \right\} = 2d \exp \left( -n\epsilon_2^2 \right) \equiv 2 \exp \left( -n\epsilon_2^2 + \log d \right). \quad (9.5)$$

Hence, setting $\epsilon_1 = \epsilon_2 \equiv \epsilon$ for any $\epsilon \geq 0$, letting $\sigma_0 := \bar{\sigma}_1 + \bar{\sigma}_2$ and $K_0 := \bar{K}_1 + \bar{K}_2$, and using (9.4)-(9.5) in the original decomposition (9.1) of $T_{0,n}$, we have a tail bound for $\|T_{0,n}\|_\infty$, as follows. For any $\epsilon \geq 0$,

$$\mathbb{P} \left( \|T_{0,n}\|_\infty \equiv \|T_{0,n}^{(1)} + T_{0,n}^{(2)}\|_\infty > \sqrt{2\sigma_0 \epsilon + K_0 \epsilon^2} \right)$$

$$\leq \mathbb{P} \left( \|T_{0,n}^{(1)}\|_\infty > \sqrt{2\sigma_1 \epsilon + \bar{K}_1 \epsilon^2} \right) + \mathbb{P} \left( \|T_{0,n}^{(2)}\|_\infty > \sqrt{2\sigma_2 \epsilon + \bar{K}_2 \epsilon^2} \right)$$

$$\leq 4 \exp \left( -n\epsilon^2 + \log d \right). \quad (9.6)$$
(9.6) 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.

10. 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 10.1 (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 $\|\mu_h^{(2)}\|_\infty := \max \{E\{h_{ij}(X)\} : j = 1, \ldots, d\}$. Then, for any $\epsilon, \epsilon_1, \epsilon_2, \epsilon_3 \geq 0$, with $\epsilon_2$ small enough such that $(\epsilon_2 + b_{n,\pi}) < \delta_\pi$,

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

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 $K_\pi := 2 \sigma_h^2 \delta_\pi^{-2}$.

10.1. Proof of Theorem 10.1. Let $\mathcal{X}_n := \{(T_i, X_i) : i = 1, \ldots, n\}$. Let $E_{\mathcal{X}_n}(\cdot)$ and $P_{\mathcal{X}_n}(\cdot)$ respectively denote expectation and probability w.r.t. $\mathcal{X}_n$ and $P(\cdot | \mathcal{X}_n)$ denote conditional probability given $\mathcal{X}_n$.

Next, let us define the following quantities:

$$\Delta_{\pi,n}(X) := \bar{\pi}(X) - \pi(X), \quad \|\Delta_{\pi,n}\|_{\infty,n} := \max_{1 \leq i \leq n} |\Delta_{\pi,n}(X_i)|,$$

$$\bar{\pi}_n(X) := \frac{1}{\bar{\pi}(X)} \text{ and } \|\bar{\pi}_n\|_{\infty,n} := \max_{1 \leq i \leq n} |\bar{\pi}_n(X_i)|.$$  

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

$$\varphi_{[j]}(T, X) := \frac{T}{\bar{\pi}(X)}h_{[j]}(X), \quad \varphi_{[j]}^{(2)}(X) := \frac{1}{n} \sum_{i=1}^{n} \varphi_{[j]}^{(2)}(T, X_i),$$

$$\mu_{[j]}^{(2)} := E \left\{ \varphi_{[j]}^{(2)}(T, X) \right\} \text{ and } \mu_{[j]}^{(2)} := E \left\{ h_{[j]}^{2}(X) \right\}.$$  

Then, using (10.1)-(10.3) in (3.3) and recalling that $\epsilon(Z) = Y - m(X)$, we have

$$T_{\pi}(Z) = \Delta_{\pi,n}(X)\bar{\pi}_n(X)\varphi(T, X)\epsilon(Z),$$

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

Now, under Assumption 1.1 (a) and Assumption 3.1 (b), \( \mathbb{E}\{\varepsilon(Z) \mid X\} = \mathbb{E}\{\varepsilon(Z) \mid T, X\} = 0 \) and \( \|\varepsilon(Z) \mid X\|_{\psi_2} \equiv \|\varepsilon(Z) \mid (T, X)\|_{\psi_2} \leq \sigma_\varepsilon(X) \leq \sigma_\varepsilon < \infty. \) Consequently, \( \varepsilon(Z_i) \mid X_n \) are (conditionally) independent random variables satisfying: \( \mathbb{E}\{\varepsilon(Z_i) \mid X_n\} = 0 \) and \( \|\varepsilon(Z_i) \mid X_n\|_{\psi_2} \leq \sigma_\varepsilon \forall 1 \leq i \leq n. \) Further, conditional on \( X_n, \phi(T_i, X_i), \Delta_{\pi,n}(X_i) \) and \( h_{[j]}(X_i) \) are all constants \( \forall i, j. \)

Using these facts along with (10.1)-(10.3), we have: \( \forall 1 \leq i \leq n \) and \( 1 \leq j \leq d, \)

\[
\| T_{[j]}(Z_i) \mid X_n \|_{\psi_2} \equiv \left\| \Delta_{\pi,n}(X_i) \bar{\pi}_n(X_i) \varphi_{[j]}(T, X_i) \varepsilon(Z_i) \mid X_n \right\|_{\psi_2} \\
\leq \Delta_{\pi,n}(X_i) \bar{\pi}_n(X_i) \varphi_{[j]}(T, X_i) \sigma_\varepsilon(X_i) \leq \sigma_\varepsilon \|\Delta_{\pi,n}\|_{\infty,n} \|\bar{\pi}_n\|_{\infty,n} \varphi_{[j]}(T, X_i).
\]

Further, \( \forall 1 \leq j \leq d, \{T_{[j]}(Z_i)\}_{i=1}^n \mid X_n \) are (conditionally) independent and centered random variables. Hence, using Lemma 13.2, we have: \( \forall 1 \leq j \leq d, \)

\[
\left\| \frac{1}{n} \sum_{i=1}^n T_{[j]}(Z_i) \mid X_n \right\|_{\psi_2} \leq \frac{4c_{n,j}(\Delta_{\pi,n})}{\sqrt{n}}, \text{ where}
\]

\[
c_{n,j}(\Delta_{\pi,n}) := \sigma_\varepsilon \|\Delta_{\pi,n}\|_{\infty,n} \|\bar{\pi}_n\|_{\infty,n} \left( \varphi_{[j]}^{(2)} \right)^{1/2}
\]

and all notations are as defined in (10.1)-(10.3). Using Lemma 13.2 again, it now follows that for any \( \varepsilon \geq 0, \)

\[
P \left\{ \left\| \frac{1}{n} \sum_{i=1}^n T_{[j]}(Z_i) \right\| > \varepsilon \right\} \leq 2 \exp \left\{ \frac{-n\varepsilon^2}{8c_{n,j}^2(\Delta_{\pi,n})} \right\} \forall 1 \leq j \leq d.
\]

The fundamental bound for \( \|T_{\pi,n}\|_{\infty}. \) Using (10.7), the union bound (u.b.) and the law of iterated expectations (l.i.e.), we then have: for any \( \varepsilon \geq 0, \)

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

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

\[
\leq \sum_{j=1}^d 2 \mathbb{E}_{X_n} \left[ \exp \left\{ \frac{-n\varepsilon^2}{8c_{n,j}^2(\Delta_{\pi,n})} \right\} \right] \text{ [using (10.7)].} \]
Next, we aim to control the behavior of the random variable $c^2_{n,j}(\mathcal{X}_n)$ appearing in the bound (10.8). Based on the definition of $c_{n,j}(\mathcal{X}_n)$ in (10.6), it suffices to separately control the variables $\|\Delta_{\pi,n}\|_{\infty,n}^2$, $\|\bar{\pi}_n\|_{\infty,n}^2$ 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 (10.1)-(10.2), we have: for any $\epsilon_1 \geq 0$,

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

(10.9) \hspace{1cm} \leq C_n \exp \left( -\frac{\epsilon_1^2}{v^2_{n,\pi}} \right) + nq_{n,\pi} \quad \text{[using (3.6)].} \quad \blacksquare

**Controlling $\|\bar{\pi}_n\|_{\infty,n}^2$.** 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),

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

\hspace{1cm} \leq \sum_{i=1}^{n} \mathbb{P} \left\{ \hat{\pi}(X_i) < \pi(X_i) - (\epsilon_2 + b_{n,\pi}) \right\} \quad \text{[using (1.1)],}

\hspace{1cm} \leq \sum_{i=1}^{n} \mathbb{P} \left\{ |\hat{\pi}(X_i) - \pi(X_i)| > \epsilon_2 + b_{n,\pi} \right\}
$$

(10.10) \hspace{1cm} \leq C_n \exp \left( -\frac{\epsilon_2^2}{v^2_{n,\pi}} \right) + nq_{n,\pi} \quad \text{[using (3.6)].} \quad \blacksquare

**Controlling $\varphi_{n[j]}^{(2)}$.** Finally, in order to control $\varphi_{n[j]}^{(2)}(\mathcal{X}_n)$ which is an average of the i.i.d. random variables $\{\varphi_{n[j]}^{(2)}(T_i, X_i)\}_{j=1}^{n}$, we first recall all notations from (10.3)-(10.4) and note that under Assumption 3.1 (a), $\|h_{n[j]}^{(2)}(X)\|_{\psi_1} \leq \sigma_{\pi}^2$ \forall $j \in \{1, \ldots, d\}$ owing to Lemma 13.1 (v). Further, $T^2/\pi^2(X) \leq \delta_{\pi}^{-2}$ a.s. \[\mathbb{P}\]. Hence, using Lemma 13.5 (b), we have: \forall $j \in \{1, \ldots, d\}$, and for some constants $\sigma_{\pi} \equiv \bar{\sigma}_{\varphi} := 2\sqrt{2}\sigma_{h}^2\delta_{\pi}^{-2}$ and $K_{\varphi} \equiv \bar{K}_{\varphi} := 2\sigma_{h}^2\delta_{\pi}^{-2}$,

$$
\varphi_{n[j]}^{(2)}(T, X) \equiv \frac{T^2}{\pi^2(X)}h_{n[j]}^{(2)}(X) \sim \text{BMC}(\bar{\sigma}_{\varphi}, \bar{K}_{\varphi}) \quad \text{and further,}
$$

(10.11)
(10.12) $\mu_h^{(2)} \equiv \mathbb{E} \left\{ \varphi_{[j]}^2(T, X) \right\} = \mathbb{E} \left\{ \frac{h_{[j]}^2(X)}{\pi(X)} \right\} \leq \frac{\mu_h^{(2)}}{\delta^\pi} \leq \left\| \mu^{(2)}_h \right\|_\infty,$

where $\left\| \mu^{(2)}_h \right\|_\infty := \max \{ \mu_h^{(2)} : j = 1, \ldots, d \} < \infty$ and $\mu^{(2)}_h$ is as in (10.4).

Using (10.11)-(10.12) along with Lemma 13.4, we then have: for any $\epsilon_3 > 0$ and for each $j \in \{1, \ldots, d\},$

$$P \left\{ \varphi^{(2)}_{n[j]} = \frac{1}{n} \sum_{i=1}^n \varphi_{[j]}^2(T_i, X_i) > \left\| \mu^{(2)}_h \right\|_{\infty} + \sqrt{2} \bar{\delta} \varphi \epsilon_3 + \bar{K} \varphi \epsilon_3^2 \right\}$$

$$
\leq P \left\{ \left\| \frac{1}{n} \sum_{i=1}^n \varphi_{[j]}^2(T_i, X_i) - \mu^{(2)}_h \right\| > \sqrt{2} \bar{\delta} \varphi \epsilon_3 + \bar{K} \varphi \epsilon_3^2 \right\}
$$

(10.13) \quad \leq 2 \exp \left(-n\epsilon_3^2\right).

For any $\epsilon_1, \epsilon_3 > 0$, and any $\epsilon_2 > 0$ such that $(\epsilon_2 + b_{n, \pi}) < \delta^\pi$, 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.

(10.14) \quad $A_{\pi, n, j}(\epsilon_1, \epsilon_2, \epsilon_3) := \{ 8c_{n,j}^2(\mathcal{X}_n) > d_n(\epsilon_1, \epsilon_2, \epsilon_3) \},$ 1 $\leq j \leq d$,

where $d_n(\epsilon_1, \epsilon_2, \epsilon_3) := \frac{8 \sigma^2(\epsilon_1 + b_{n, \pi})^2}{\left\| \delta^\pi - (\epsilon_2 + b_{n, \pi}) \right\|^2} \left( \left\| \mu^{(2)}_h \right\|_{\infty} + \sqrt{2} \bar{\delta} \varphi \epsilon_3 + \bar{K} \varphi \epsilon_3^2 \right).$

Then, recalling from (10.6) that $c_{n,j}^2(\mathcal{X}_n) \equiv \sigma^2 \left\| \Delta_{\pi, n} \right\|^2_{\infty, n} \left\| \tilde{\pi}_n \right\|^2_{\infty, n} \varphi^{(2)}_{n[j]}$ and using the bounds (10.9), (10.10) and (10.13) for $\left\| \Delta_{\pi, n} \right\|^2_{\infty, n}$, $\left\| \tilde{\pi}_n \right\|^2_{\infty, n}$ and $\varphi^{(2)}_{n[j]}$ respectively, along with the union bound, we have:

$$P \left( A_{\pi, n, j} \right) \equiv P_{\chi_n} \left( A_{\pi, n, j} \right) \equiv P_{\chi_n} \left\{ 8c_{n,j}^2(\mathcal{X}_n) > d_n(\epsilon_1, \epsilon_2, \epsilon_3) \right\}$$

(10.15) \quad \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) + 2nq_{n, \pi} + 2 \exp \left(-n\epsilon_3^2\right).$

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

$$E_{\chi_n} \left[ \exp \left\{ \frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)} \right\} \right] = E \left[ \exp \left\{ \frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)} \right\} \right| A^c_{\pi, n, j}] P( A^c_{\pi, n, j})$$

$$+ E \left[ \exp \left\{ \frac{-n\epsilon^2}{8c_{n,j}^2(\mathcal{X}_n)} \right\} \right| A_{\pi, n, j}] P( A_{\pi, n, j})$$

(10.16) \quad \leq \exp \left\{ \frac{-n\epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} \right\} + 2 \exp \left(-n\epsilon_3^2\right) + 2nq_{n, \pi}$$

$$+ Cn \exp \left( -\frac{-\epsilon_1^2}{v_{n, \pi}^2} \right) + Cn \exp \left( -\frac{-\epsilon_2^2}{v_{n, \pi}^2} \right) \text{ [using (10.14)-(10.15)].}$$
The final bound for $\|T_{\pi,n}\|_\infty$. Using (10.16) in the fundamental bound (10.8) for $\|T_{\pi,n}\|_\infty$, we finally have: for any $\epsilon \geq 0$,

$$P(\|T_{\pi,n}\|_\infty > \epsilon) \leq \sum_{j=1}^d 2 \mathbb{E} \chi_n \left( \exp \left( \frac{-n\epsilon^2}{8\sigma_{n,j}^2(X_n)} \right) \right)$$

$$\leq 2d \exp \left( \frac{-n\epsilon^2}{d_n(\epsilon_1, \epsilon_2, \epsilon_3)} \right) + 4d \exp \left( -n\epsilon_3^2 \right) + 4q_{n,\pi}(nd)$$

$$+ 2C(nd) \exp \left( \frac{-\epsilon_1^2}{v_{n,\pi}^2} \right) + 2C(nd) \exp \left( \frac{-\epsilon_2^2}{v_{n,\pi}^2} \right) \quad [\text{using (10.16)}],$$

(10.17) therefore establishes an explicit finite sample tail bound for $\|T_{\pi,n}\|_\infty$, as desired. This completes the proof of Theorem 10.1.

10.2. The Final Proof of Theorem 3.2. We next evaluate the general tail bound for $\|T_{\pi,n}\|_\infty$ in Theorem 10.1 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 $\|T_{\pi,n}\|_\infty$. To this end, let $(c_1, c_2, c_3) > 1$ be any universal constants, and set $\epsilon_1 = c_1v_{n,\pi}\sqrt{\log(nd)}$, $\epsilon_2 = c_2v_{n,\pi}\sqrt{\log(nd)}$ 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 10.1. Then,

$$\epsilon_1 + b_{n,\pi} \equiv c_1v_{n,\pi}\sqrt{\log(nd)} + b_{n,\pi} \leq c_1\{v_{n,\pi}\sqrt{\log(nd)} + b_{n,\pi} \} \quad \text{and}$$

$$\epsilon_2 + b_{n,\pi} \equiv c_2v_{n,\pi}\sqrt{\log(nd)} + b_{n,\pi} \leq \delta_\pi/2 \quad \text{(by choice),}$$

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

$$\frac{\|\mu_{h_{\epsilon_3}}^{(2)}\|_\infty}{\delta_\pi} + \sqrt{2}\sigma_\varphi\epsilon_3 + \bar{K}_\varphi\epsilon_3^2 \leq \frac{\|\mu_{h_{\epsilon_3}}^{(2)}\|_\infty}{\delta_\pi} + \left( \sqrt{2}\sigma_\varphi + \bar{K}_\varphi \right) c_3\sqrt{\frac{\log d}{n}}.$$
Given these choices of \( \{c_j\}_{j=1}^3 \), 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 10.1, we then have:

\[
\text{With probability } \geq 1 - \frac{2}{d^{2.5-1}} - \frac{4}{d^{3.5-1}} - \sum_{j=1}^{2} \frac{2C}{(nd)^{0.5-1}} - 4q_n,\pi(nd),
\]

\[
\|T_{\pi,n}\|_{\infty} \leq c \sqrt{\frac{\log d}{n}} \{\sqrt{\log(nd)} + b_n,\pi\}C_1 \left( \frac{\|\mu_1^{(2)}\|_{\delta_\pi}}{\delta_\pi} + C_2 \sqrt{\frac{\log d}{n}} \right)^{\frac{1}{2}},
\]

where \( C_1 := c_1(4\sqrt{2}\sigma_\pi/\sigma_\varphi) \) and \( C_2 := c_3C_\varphi \equiv c_3(\sqrt{2}\sigma_\varphi + K_\varphi) \), with \( \sigma_\varphi \) and \( K_\varphi \) being as in (10.11). This completes the proof of Theorem 3.2. \( \blacksquare \)

11. 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 11.1 (Tail bound for \( \|T_{m,n}\|_{\infty} \)).** Let Assumptions 1.1, 3.1 (a) and 3.3 hold, with the sequences \((v_{n,m}, b_{n,m}, q_{n,m})\), \( n \equiv n/2 \) and the constants \((\delta_\pi, C)\) being as defined therein. Then, for any \( \epsilon, \epsilon_1, \epsilon_2 \geq 0 \),

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

where

\[
t_\epsilon(\epsilon_1, \epsilon_2) := 8\bar{n}^2(\epsilon + b_{n,m})^2 \left( \|\mu_1^{(2)}\|_{\delta_\pi} + \sqrt{2}\sigma_m \epsilon_2 + K_m \epsilon_2^2 \right),
\]

\[
\|\mu_1^{(2)}\|_{\infty} := \max\{E(h_{[j]}^2(X)) : j = 1, \ldots, d\},
\]

and \( \delta_\pi, \sigma_m, K_m \geq 0 \) are constants with \( \delta_\pi \leq \delta_\pi^{-1} \) and \( \sigma_m := 2\sqrt{2}\sigma_h^2 \) and \( K_m := 2\sigma_h^2 \).

11.1. **Proof of Theorem 11.1.** Recalling \( T_{m,n} \) from the decomposition (3.1), we rewrite it as follows.

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

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

\[
(11.1) = \frac{1}{2} \sum_{k \neq k' = 1}^{2} T_{(k,k')}' m, \bar{n}, \text{ where } T_{(k,k')}' m, \bar{n} := \frac{1}{\bar{n}} \sum_{i \in I_k'} T_{(k,k)}'(X_i) \text{ and }
\]

\[
T_{(k)}'(Z) := \left\{ \frac{T}{\pi(X)} - 1 \right\} \left\{ \tilde{m}^{(k)}(X) - m(X) \right\} h(X) \ \forall \ k \neq k' \in \{1, 2\}.
\]
Define $\mathcal{X}_{n,k} := \{X_i : i \in I_k\} \forall k \in \{1, 2\}$, and let $\mathbb{E}_{\mathcal{X}_{n,k}}(\cdot)$ and $\mathbb{P}(\cdot \mid \mathcal{X}_{n,k})$ respectively denote expectation w.r.t. $\mathcal{X}_{n,k}$ and conditional probability given $\mathcal{X}_{n,k}$. Further, for each $k \neq k' \in \{1, 2\}$, let $\mathbb{E}_{D_n^{(k)} \mathcal{X}_{n,k}^*}(\cdot)$ and $\mathbb{P}(\cdot \mid D_n^{(k)} \mathcal{X}_{n,k}^*)$ respectively denote expectation w.r.t. $\{D_n^{(k)} \mathcal{X}_{n,k}^*\}$ and conditional probability given $\{D_n^{(k)} \mathcal{X}_{n,k}^*\}$. With $D_n^{(k)} \perp \mathcal{X}_{n,k'}^* \forall k \neq k' \in \{1, 2\}$, we note that $\mathbb{E}_{D_n^{(k)} \mathcal{X}_{n,k}^*}(\cdot) = \mathbb{E}_{\mathcal{X}_{n,k}^*}(\mathbb{E}_{D_n^{(k)}(\cdot)})$. Next, let us define: $\forall k \neq k' \in \{1, 2\}$,

$$\Delta_{m,\bar{a}}^{(k)}(X) := \bar{m}^{(k)}(X) - m(X), \quad \left\| \Delta_{m,\bar{a}}^{(k,k')} \right\|_{\infty,\bar{a}} := \max_{i \in I_k} \left\| \Delta_{m,\bar{a}}^{(k)}(X_i) \right\|,$$

(11.2) and let $\psi(T, X) := \frac{T}{\pi(X)} - 1$ and $h_{n,j}^{(2,k')} := \frac{1}{n} \sum_{i \in I_k} h_{n}^{2,j}(X_i)$.

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

Using the notations from (11.2) and (11.3), we have: for each $k \in \{1, 2\}$,

$$T_{m}^{(k)}(Z) = \left\{ \frac{T}{\pi(X)} - 1 \right\} \{\bar{m}^{(k)}(X) - m(X)\} h(X) = \psi(T, X) \Delta_{m,\bar{a}}^{(k)}(X) h(X).$$

Now, for each $k \in \{1, 2\}$ and $k' \neq k \in \{1, 2\}$, $D_n^{(k)} \perp \mathcal{X}_{n,k'}^*$, and we have:

$$\forall i \in I_{k'}, \quad \mathbb{E}\{\psi(T_i, X_i) \mid D_n^{(k)} \mathcal{X}_{n,k'}^*\} \equiv \mathbb{E}\{\psi(T_i, X_i) \mid X_i\} = 0 \quad \text{and} \quad \|\psi(T_i, X_i) \mid D_n^{(k)} \mathcal{X}_{n,k'}^*\|_{\psi_2} \equiv \|\psi(T_i, X_i) \mid X_i\|_{\psi_2} \leq \hat{\pi}^2(X_i) \leq \hat{\delta}_\pi^2,$$

(11.4) where the bounds on the $\| \cdot \|_{\psi_2}$ norm follow from using Lemma 13.3 and Lemma 13.1 (i) (b) along with the definitions of $\hat{\pi}(\cdot)$ and $\hat{\delta}_\pi$ given earlier. Further, conditional on $D_n^{(k)}$ and $\mathcal{X}_{n,k'}^*$, $\{\Delta_{m,\bar{a}}^{(k,k')} (X_i)\} \in I_{k'}$, and $\{h_{n,j}(X_i)\} \in I_{k'}$, for each $j \in \{1, \ldots, d\}$, are all constants. Hence, using Lemma 13.2 and (11.4), along with (11.1)-(11.3), we have: $\forall k \neq k' \in \{1, 2\}$ and $j \in \{1, \ldots, d\}$,

$$\frac{1}{n} \sum_{i \in I_{k'}} T_{m}^{(k)}(Z_i) \|D_n^{(k)} \mathcal{X}_{n,k'}^*\|_{\psi_2} \leq \frac{4d_{\bar{a},j} \left( D_n^{(k)} \mathcal{X}_{n,k'}^* \right)}{\sqrt{n}},$$

where

$$d_{\bar{a},j} \left( D_n^{(k)} \mathcal{X}_{n,k'}^* \right) := \hat{\delta}_\pi \left\| \Delta_{m,\bar{a}}^{(k,k')} \right\|_{\infty,\bar{a}} \left( h_{n,j}^{(2,k')} \right)^{1/2}.$$
Using Lemma 13.2, we then have: \( \forall k \neq k' \in \{1, 2\}, 1 \leq j \leq d \) and \( \epsilon \geq 0 \),

\[
\mathbb{P}\left\{ \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} \mathbf{T}^{(k)}_{m[j]}(Z_i) \geq \epsilon \mid \mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}' \right\} \leq 2 \exp\left\{ \frac{-\bar{n} \epsilon^2}{8d_{n,j}^2 \left\langle \mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}' \right\rangle} \right\}.
\]

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

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

\[
= \sum_{j=1}^{d} \mathbb{E}_{\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}'} \left[ \mathbb{P}\left\{ \frac{1}{n} \sum_{i \in \mathcal{I}_{k'}} \mathbf{T}^{(k)}_{m[j]}(Z_i) \geq \epsilon \mid \mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}' \right\} \right] 
\]

\[
(11.6) \quad \leq 2 \sum_{j=1}^{d} \mathbb{E}_{\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}'} \left[ \exp\left\{ \frac{-\bar{n} \epsilon^2}{8d_{n,j}^2 \left\langle \mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}' \right\rangle} \right\} \right]. \quad \blacksquare
\]

Next, we aim to control the random variable \( d_{n,j}^2(\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}') \) appearing in (11.6). Based on the definition (11.5) of \( d_{n,j}^2(\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}') \), it suffices to separately control \( \|\Delta_{m,n}^{(k,k')}\|_\infty^{2} \) and \( \mathcal{H}_{n,j}^{(2,k')} \). To this end, let \( \mathbb{E}_{\mathcal{D}_{n}^{(k)}}(\cdot) \) and \( \mathbb{P}_{\mathcal{D}_{n}^{(k)}}(\cdot) \) denote expectation and probability w.r.t \( \mathcal{D}_{n}^{(k)} \forall k \in \{1, 2\} \).

With \( \mathcal{D}_{n}^{(k)} \perp \mathcal{X}_{n,k}' \) for each \( k \neq k' \in \{1, 2\} \), we note that for any event \( A \equiv A(\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}') \), \( \mathbb{P}(A) \equiv \mathbb{P}_{\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}'}(A) = \mathbb{E}_{\mathcal{X}_{n,k}'}[\mathbb{E}_{\mathcal{D}_{n}^{(k)}}\{1(A) \mid \mathcal{X}_{n,k}'\}] \equiv \mathbb{E}_{\mathcal{X}_{n,k}'}[\mathbb{E}_{\mathcal{D}_{n}^{(k)}}\{A(\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}') \mid \mathcal{X}_{n,k}'\}] = \mathbb{E}_{\mathcal{X}_{n,k}'}[\mathbb{E}_{\mathcal{D}_{n}^{(k)}}\{A(\mathcal{D}_{n}^{(k)}, \mathcal{X}_{n,k}')\}] \), where the final step holds since \( \mathbb{P}_{\mathcal{D}_{n}^{(k)}}(\cdot) \mid \mathcal{X}_{n,k}' = \mathbb{P}_{\mathcal{D}_{n}^{(k)}}(\cdot) \) as \( \mathcal{D}_{n}^{(k)} \perp \mathcal{X}_{n,k}' \).
Controlling $\| \Delta_{m,\bar{n}}^{(k,k')} \|^2_{\infty,\bar{n}}$. Using Assumption 3.3 along with the notations and facts discussed above, we have: for any $k \neq k' \in \{1, 2\}$ and any $\epsilon_1 \geq 0$,

$$\mathbb{P} \left[ \frac{\| \Delta_{m,\bar{n}}^{(k,k')} \|^2_{\infty,\bar{n}}}{\epsilon_1} \equiv \left\{ \max_{i \in I_{k'}} \| \Delta_{m,\bar{n}}^{(k)}(X_i) \| \right\}^2 > (\epsilon_1 + b_{\bar{n},m})^2 \right]$$

$$\leq \sum_{i \in I_{k'}} \mathbb{P} \left( \| \Delta_{m,\bar{n}}^{(k)}(X_i) \| > \epsilon_1 + b_{\bar{n},m} \right) \quad \text{[using the u.b.,]}$$

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

(11.7) $$\equiv C\bar{n} \exp \left( \frac{-\epsilon_1^2}{v_{\bar{n},m}^2} \right) + \bar{n}q_{\bar{n},m},$$

where we also used that $D_{\bar{n}}^{(k)} \perp \chi_{n,k'}$ which ensures $\mathbb{P}_{D_{\bar{n}}^{(k)}}(\cdot | \chi_{n,k}) = \mathbb{P}_{D_{\bar{n}}^{(k)}}(\cdot)$ and makes (3.8) in Assumption 4.3 applicable conditional on $\chi_{n,k'}$. \hfill \Box

Controlling $\bar{h}_{h_{[j]}}^{(2,k')}$. 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, $\forall k' \in \{1, 2\}$ and $j \in \{1, \ldots, d\}$, $\bar{h}_{h_{[j]}}^{(2,k')} \equiv \bar{h}_{h_{[j]}}^{(2)}$ is simply an average of the i.i.d. random variables $\{ h_{[j]}^2(X_i) \}_{i \in I_{k'}}$. Further, using Assumption 3.1 (a) and Lemma 13.5 (a), $h_{[j]}^2(X) \sim \text{BMC}(\bar{\sigma}_h, \bar{K}_h)$ for some constants $\sigma_m \equiv \bar{\sigma}_h := 2\sqrt{2\sigma_h^2}$ and $K_m \equiv \bar{K}_h := 2\sigma_h^2$. Hence, using Lemma 13.4, we have: for each $k' \in \{1, 2\}$ and $j \in \{1, \ldots, d\}$, and for any $\epsilon_2 \geq 0$,

(11.8) $$\mathbb{P} \left\{ \frac{1}{\bar{n}} \sum_{i \in I_{k'}} h_{[j]}^2(X_i) > \| \mu_{h}^{(2)} \|_{\infty} + \sqrt{2\bar{\sigma}_h} \epsilon_2 + \bar{K}_h \epsilon_2^2 \right\}$$

$$\leq \mathbb{P} \left\{ \frac{1}{\bar{n}} \sum_{i \in I_{k'}} h_{[j]}^2(X_i) - \mu_{h_{[j]}}^{(2)} \right\} \leq \sqrt{2\bar{\sigma}_h} \epsilon_2 + \bar{K}_h \epsilon_2^2 \right\} \leq 2 \exp(-\bar{n} \epsilon_2^2). \hfill \Box$$

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

(11.9) $$t_{\bar{n}}(\epsilon_1, \epsilon_2) := 8\bar{\sigma}_h^2 (\epsilon_1 + b_{\bar{n},m})^2 \left( \| \mu_{h}^{(2)} \|_{\infty} + \sqrt{2\bar{\sigma}_h} \epsilon_2 + \bar{K}_h \epsilon_2^2 \right).$$
Thus, using the bounds (11.7) and (11.8) in the definition of \(d_{n,j}^2(D_n^{(k)}, \mathcal{X}_{n,k'})\) in (11.5), we have: for each \(k \neq k' \in \{1, 2\}\), \(j \in \{1, \ldots, d\}\) and \(\epsilon_1, \epsilon_2 \geq 0\),

\[
P \left\{ 8d_{n,j}^2(D_n^{(k)}, \mathcal{X}_{n,k'}) > t_n(\epsilon_1, \epsilon_2) \right\}
\]

(11.10)

\[
\leq C n \exp \left( \frac{-\epsilon_1^2}{v_{n,m}^2} \right) + nq_{n,m} + 2 \exp(-\bar{\epsilon}_2^2).
\]

Using (11.10) in the fundamental bound (11.6) for \(\|T^{(k,k')}_{m,n}\|_\infty\), we then have: for each \(k \neq k' \in \{1, 2\}\) and for any \(\epsilon, \epsilon_1, \epsilon_2 \geq 0\),

\[
P \left\{ \|T^{(k,k')}_{m,n}\|_\infty > \epsilon \right\}
\]

\[
= 2 \sum_{j=1}^{d} \mathbb{E}_{D_n^{(k)}, \mathcal{X}_{n,k'}} \left[ \exp \left( \frac{-\bar{\epsilon}^2}{8d_{n,j}^2(D_n^{(k)}, \mathcal{X}_{n,k'})} \right) \right]
\]

(11.11)

\[
\leq 2d \left[ \exp \left( \frac{-\bar{\epsilon}^2}{t_n(\epsilon_1, \epsilon_2)} \right) \right] + P \left\{ 8d_{n,j}^2(D_n^{(k)}, \mathcal{X}_{n,k'}) > t_n(\epsilon_1, \epsilon_2) \right\}
\]

Thus, (11.11) establishes an explicit tail bound for \(\|T^{(k,k')}_{m,n}\|_\infty\).

The final bound for \(\|T_{m,n}\|_\infty\). A tail bound for \(\|T_{m,n}\|_\infty\) now follows easily using (11.1) and (11.11) along with the u.b. For any \(\epsilon, \epsilon_1, \epsilon_2 \geq 0\), we have:

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

(11.12)

\[
\leq 4d \exp \left( \frac{-\bar{\epsilon}^2}{t_n(\epsilon_1, \epsilon_2)} \right) + 4C nd \exp \left( \frac{-\epsilon_1^2}{v_{n,m}^2} \right) + 4nq_{n,m} + 8d \exp(-\bar{\epsilon}_2^2).
\]

(11.12) thus establishes an explicit finite sample tail bound for \(\|T_{m,n}\|_\infty\) as desired. This concludes the proof of Theorem 11.1.

11.2. The Final Proof of Theorem 3.3. Given the general tail bound for \(\|T_{m,n}\|_\infty\) in Theorem 11.1, 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 $\epsilon_1 = c_1 v_{\bar{m},m} \sqrt{\log(\bar{n}d)}$ and $\epsilon_2 = c_2 \sqrt{\log(d)}/\bar{n}$, where we further assume w.l.o.g. that $\epsilon_2 < 1$. Then, note that

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

and

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

Therefore, combining and using all the inequalities above in the definition (11.9) of $t_n(\epsilon_1, \epsilon_2)$, and letting $C_h := (\sqrt{2} \bar{\sigma}_h + \bar{K}_h)$, we have: $t_n(\epsilon_1, \epsilon_2)$

$$\leq 8c_1^2 \sigma_2 \{v_{\bar{m},m} \sqrt{\log(\bar{n}d)} + b_{\bar{m},m}\}^2 \left\{\|\mu^{(2)}_h\|_\infty + c_2 C_h \sqrt{\frac{\log d}{\bar{n}}}\right\}.$$  

Given these choices of $(\sigma_j^2)_{j=1}^d$, let us now set $\epsilon = c \sqrt{\{\log(\bar{n}d)/\bar{n}\} t_n(\epsilon_1, \epsilon_2)}$ for any $c > 1$. Using Theorem 11.1 and with $\bar{n} \equiv n/2 \leq n$, we then have:

$$\text{With probability } \geq 1 - \frac{4}{d^{k-1}} - \frac{8}{d^{k-1}} - \frac{4C}{(\bar{n}d)^{q-1}} - 4\bar{q}_{b,n}(\bar{n}d),$$

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

where $C_1^* := 4c_1 \bar{\delta}_\pi$ and $C_2^* := \sqrt{2} c_2 C_h \equiv \sqrt{2} c_2 (\sqrt{2} \bar{\sigma}_h + \bar{K}_h)$, with $\bar{\sigma}_h$ and $\bar{K}_h$ being as in (11.8). This completes the proof of Theorem 3.3.  

12. Proofs 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 12.1 (Tail bound for $\|R_{\pi,m,n}\|_\infty$).** Let Assumptions 1.1, 3.1, 3.2 and 3.3 hold, with the sequences $(v_{\pi,\pi}, b_{\pi,\pi}, q_{\pi,\pi})$, $(v_{\bar{m},m}, b_{\bar{m},m}, q_{\bar{m},m}, \bar{n})$ 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:

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

$$+ Cn \left\{\exp \left(\frac{-\epsilon_3^2}{v_{n,\pi}^2}\right) + \exp \left(\frac{-\epsilon_3^2}{v_{\bar{m},m}^2}\right) + \exp \left(\frac{-\epsilon_3^2}{v_{n,\pi}^2}\right)\right\} + 2n q_{n,\pi} + n q_{\bar{m},m},$$

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


12.1. Proof of Theorem 12.1. 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 hence, is expected to be small enough under reasonable assumptions on the behaviors of $\hat{\pi}(\cdot) - \pi(\cdot)$ and $\hat{m}(\cdot) - m(\cdot)$. Hence, one can attempt to control the behavior of this term by ‘naive’ techniques, as opposed to the more sophisticated analyses required for controlling $T_{\pi,n}$ and $T_{m,n}$. Recall from (3.1) that

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

Hence, with $\|\Delta_{\pi,n}\|_{\infty,n}$ and $\|\hat{\pi}_n\|_{\infty,n}$ as in (10.1) and (10.2) respectively, and with $\|\Delta^{(k,k')}\|_{\infty,n}$ as in (11.2) for any $k \neq k' \in \{1,2\}$, we have:

$$\|R_{\pi,m,n}\|_{\infty} \leq \|\hat{\pi}_n\|_{\infty,n} \|\Delta_{\pi,n}\|_{\infty,n} \|\Delta^*_m,\bar{n}\|_{\infty,n} \|\bar{\xi}_n\|_{\infty},$$

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

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

For most of the quantities appearing in the bound (12.2), we already have their explicit tail bounds. Specifically, using (10.9), we have: for any $\epsilon_1 \geq 0$,

$$\mathbb{P} \left\{ \|\Delta_{\pi,n}\|_{\infty,n} > r_{\pi,n}(\epsilon_1) \right\} \leq C n \exp \left( -\frac{\epsilon_1^2}{v_{n,\pi}} \right) + nq_{n,\pi},$$

where $r_{\pi,n}(\epsilon) := \epsilon + b_{n,\pi}$ for any $\epsilon \geq 0$, and using (10.10), for any $\epsilon_2 \geq 0$ small enough such that $r_{\pi,n}(\epsilon_2) < \delta_{\pi}$,

$$\mathbb{P} \left[ \|\hat{\pi}_n\|_{\infty,n} > \left\{ \delta_{\pi} - r_{\pi,n}(\epsilon_2) \right\}^{-1} \right] \leq C n \exp \left( -\frac{\epsilon_2^2}{v_{n,\pi}} \right) + nq_{n,\pi}.$$

Next, let $r_{m,\bar{n}}(\epsilon) := \epsilon + b_{\bar{n},m}$. Using (11.7), we have: for any $\epsilon_3 \geq 0$,

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

$$\leq 2C \bar{n} \exp \left( -\frac{\epsilon_3^2}{v_{\bar{n},m}^2} \right) + 2\bar{n} q_{\bar{n},m} \equiv C n \exp \left( -\frac{\epsilon_3^2}{v_{n,m}^2} \right) + nq_{n,m}.$$

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 13.1
(ii)(a), \( \|h_{ij}(X)\|_{\psi_1} = \|h_{ij}(X)\|_{\psi_1} \leq \sqrt{2}\|h_{ij}(X)\|_{\psi_2} \leq \sqrt{2}\sigma_h \forall 1 \leq j \leq d \).

Further, due to (1.1), \( T/\pi(X) \leq \delta_\pi^{-1} \) a.s. \([\mathbb{P}]\). Hence, using Lemma 13.5 (ii), we have: for constants \( \sigma_\pi, m \equiv \bar{\sigma}_\pi := 4\sigma_h\delta_\pi^{-1} \) and \( K_\pi, m \equiv \bar{K}_\pi := 2\sqrt{2}\sigma_h\delta_\pi^{-1} \),

\begin{equation}
(12.6) \quad \xi_{ij}(T, X) \equiv \frac{T}{\pi(X)}\|h_{ij}(X)\| \sim \text{BMC}(\bar{\sigma}_\pi, \bar{K}_\pi) \quad \forall j \in \{1, \ldots, d\}.
\end{equation}

Further, \( \mathbb{E}\{\xi_{ij}(T, X)\} = \mathbb{E}\{\|h_{ij}(X)\|\} = \mu_{h_{ij}} \) (say) \( \forall j \in \{1, \ldots, d\} \), and recall that \( \mu_{h_j} \) is the maximum value of \( \max\{\mu_{h_j} : j = 1, \ldots, d\} \). Using (12.6) and Lemma 13.4 along with the u.b., we then have: for any \( \epsilon_4 \geq 0 \),

\begin{align*}
\mathbb{P}\left\{ \|\bar{\xi}_n\| > r_\pi(\epsilon_4) \equiv \|\mu_{h_j}\|_{\infty} + \sqrt{2}\bar{\sigma}_\pi \epsilon_4 + \bar{K}_\pi \epsilon_4^2 \right\} \\
\leq \sum_{j=1}^{d} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \xi_{ij}(T_i, X_i) - \mu_{h_{ij}} \right| > \sqrt{2}\bar{\sigma}_\pi \epsilon_4 + \bar{K}_\pi \epsilon_4^2 \right\} \\
(12.7) \quad \leq 2d \exp(-n\epsilon_4^2) \equiv 2 \exp(-n\epsilon_4^2 + \log d).
\end{align*}

Using the bounds (12.3), (12.4), (12.5) and (12.7), along with the u.b., in the original bound (12.2) for \( \|R_{\pi,m,n}\|_{\infty} \), we then have: for any \( \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \geq 0 \),

\begin{align*}
(12.8) \quad \mathbb{P}\left\{ \|R_{\pi,m,n}\|_{\infty} > \frac{r_\pi,n(\epsilon_1)}{\delta_\pi - r_\pi, n(\epsilon_2)} r_{m,\bar{n}}(\epsilon_3) r_\pi(\epsilon_4) \right\} & \leq 2d \exp(-n\epsilon_4^2) \\
+ Cn \left\{ \exp\left( -\frac{\epsilon_1^2}{v_{\bar{n},\pi}} \right) + \exp\left( -\frac{\epsilon_2^2}{v_{\bar{n},\pi}} \right) + \exp\left( -\frac{\epsilon_3^2}{v_{\bar{n},\pi}} \right) \right\} + 2nq_{n,\pi} + nq_{\bar{n},m},
\end{align*}

where \( r_\pi,n(\cdot), r_{m,\bar{n}}(\cdot), r_\pi(\cdot) \geq 0 \) are as in (12.3), (12.5) and (12.7) respectively, and we assume that \( r_\pi,n(\epsilon_2) < \delta_\pi \). The proof of Theorem 12.1 is complete. 

\[ \tag{12.8} \]

12.2. The Final Proof of Theorem 3.4. Given the general tail bound for \( \|R_{\pi,m,n}\|_{\infty} \) in Theorem 12.1, we next evaluate it for a specific set of choices for \( \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 > 0 \) in order to understand its behavior and also establish the convergence rate of \( \|R_{\pi,m,n}\|_{\infty} \). To this end, let \( c_1, c_2, c_3, c_4 > 1 \) be universal constants, and set \( \epsilon_1 = c_1v_{n,\pi}\sqrt{\log n}, \epsilon_2 = c_2v_{n,\pi}\sqrt{\log n}, \epsilon_3 = c_3v_{n,m}\sqrt{\log n} \) and \( \epsilon_4 = c_4\sqrt{\log d}/n \), 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 12.1) and \( \epsilon_4 < 1 \). We then have:

\begin{align*}
r_\pi,n(\epsilon_1) & \leq c_1\{v_{n,\pi}\sqrt{\log n} + b_{n,\pi}\}, \quad r_\pi(\epsilon_4) \leq \|\mu_{h_j}\|_{\infty} + c_4C\xi \sqrt{\frac{\log d}{n}}, \\
r_{m,\bar{n}}(\epsilon_3) & \leq c_3\{v_{\bar{n},m}\sqrt{\log n} + b_{\bar{n},m}\}, \quad \text{and} \quad r_\pi,n(\epsilon_2) \leq \delta_\pi/2,
\end{align*}

\[ \tag{12.8} \]
where $C_\xi := \sqrt{2} \bar{\sigma}_\xi + \bar{K}_\xi$. Using Theorem 12.1, we then have:

$$\|R_{\pi,m,n}\|_\infty \leq \frac{2c_1c_3}{\delta} r_{\pi,n} r_{m,\bar{n}} \left( \|\mu_{[n]}\|_\infty + c_4 C_\xi \sqrt{\log n} \right),$$

where

$$r_{\pi,n} := v_{n,\pi} \sqrt{\log n} + b_{n,\pi}, \quad r_{m,\bar{n}} := v_{\bar{n},m} \sqrt{\log n} + b_{\bar{n},m}$$

and $C_\xi := \sqrt{2} \bar{\sigma}_\xi + \bar{K}_\xi$ with $\bar{\sigma}_\xi$ and $\bar{K}_\xi$ as in (12.6). This completes the proof of Theorem 3.4.

13. Technical Tools. In this section, we collect some key technical results, including definitions and supporting lemmas, that would be useful throughout in our technical analyses.

13.1. Orlicz Norms, Sub-Gaussians and Sub-Exponentials. We first introduce a few definitions and related results regarding concentration bounds which are a fundamental ingredient in the proofs of our results.

**Definition 13.1 (Orlicz norms).** For any $\alpha > 0$, let $\psi_\alpha(\cdot)$ denote the function given by: $\psi_\alpha(x) = \exp(x^\alpha) - 1 \forall \ x \geq 0$. Then, for any random variable $X$ and any $\alpha > 0$, the $\psi_\alpha$-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 $\alpha = \infty$. 
13.2. Properties of Orlicz Norms and Concentration Bounds. We next 
entail, 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 (pos-
sibly with slightly different constants) and proofs can be found in several 
relevant references, including Van der Vaart and Wellner (1996); Pollard 
(2015); Vershynin (2012, 2018); Wainwright (2017) and Rigollet and Hütter 
(2017), among others. The proofs are therefore skipped for brevity.

**Lemma 13.1** (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 $E(X) \in \mathbb{R}$.

(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 $\|X\|_{\psi_\alpha} = 0$ a.s., 
(b) $\|cX\|_{\psi_\alpha} = |c|\|X\|_{\psi_\alpha}$ for all $c \in \mathbb{R}$ and $\|X\|_{\psi_\alpha} = \|X\|_{\psi_\alpha}$, and (c) 
$\|X + Y\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}$.

(ii) (Monotonicities). (a) For any $0 < \alpha \leq \beta$, $(\log 2)^{1/\alpha}\|X\|_{\psi_\alpha} \leq (\log 2)^{1/\beta} 
\|X\|_{\psi_\beta}$. In particular, $\|X\|_{\psi_1} \leq (\log 2)^{-1/2}\|X\|_{\psi_2}$. (b) For any $\alpha > 0$, 
$\|X\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha}$. (c) If $|X| \leq |Y|$ a.s., then $\|X\|_{\psi_\alpha} \leq \|Y\|_{\psi_\alpha}$ for each $\alpha > 0$. (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 > 0, for some $(C,\sigma,\alpha) > 0$, then 
$\|X\|_{\psi_\alpha} \leq \sigma(1 + C/2)^{1/\alpha}$.

(iv) (Moment bounds). If $\|X\|_{\psi_\alpha} \leq \sigma$ for some $(\alpha,\sigma) > 0$, \then \mathbb{E}(|X|^m) \leq C_\alpha^m \sigma^m m^{m/\alpha} \forall m \geq 1$, for some constant $C_\alpha$ depending only on $\alpha$. 
(A converse of this result also holds, although not explicitly presented here). For $\alpha = 1$ and 2 in particular, we have:

(a) If $\|X\|_{\psi_1} \leq \sigma$, then for each $m \geq 1$, \mathbb{E}(|X|^m) \leq \sigma^m m! \leq \sigma^m m^m$.
(b) If $\|X\|_{\psi_2} \leq \sigma$, then \mathbb{E}(|X|^m) \leq 2\sigma^m \Gamma(m/2 + 1)$ 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 (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 vari-
bles $X$ and $Y$ with $\|X\|_{\psi_\alpha} < \infty$ and $\|Y\|_{\psi_\beta} < \infty$, $\|XY\|_{\psi_\gamma} < \infty$ and 
$\|XY\|_{\psi_\gamma} \leq \|X\|_{\psi_\alpha}\|Y\|_{\psi_\beta}$. In particular, 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 $\|X\|_{\psi_1} < \infty$, then $\|XY\|_{\psi_a} \leq M \|X\|_{\psi_a}$.

(vi) (Orlicz norms and tail bounds for maximums). Let $\{X_i\}_{i=1}^n$ ($n \geq 1$) be random variables (possibly dependent) with $\max_{1 \leq i \leq n} \|X_i\|_{\psi_1} \leq \sigma$, for some $(\alpha, \sigma) > 0$, and let $Z_n := \max_{1 \leq i \leq n} |X_i|$. Then for any $n \geq 1$, $\|Z_n\|_{\psi_1} \leq \sigma (\log n + 2)^{1/\alpha} \leq \sigma \{3 \log(n + 1)\}^{1/\alpha}$, $\mathbb{P}(Z_n > \epsilon) \leq 2n \exp(-\epsilon^\alpha/\sigma^\alpha)$ $\forall \epsilon \geq 0$, and $\mathbb{P}\{Z_n > c \sigma (\log n)^{1/\alpha}\} \leq 2n^{-1}(\epsilon^{\alpha-1}) \forall c > 1$.

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

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

**Lemma 13.2** (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

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

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

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 $\sigma_i = \sigma$, $\|1/n \sum_{i=1}^n (X_i - \mu_i)\|_{\psi_2} \leq (4\sigma)/\sqrt{n}$.

**Lemma 13.3** (Sub-Gaussian properties of binary random variables). Let $Z \in \{0, 1\}$ be a binary random variable with $\mathbb{E}(Z) \equiv \mathbb{P}(Z = 1) = p \in [0, 1]$. 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 13.3 explicitly characterizes the sub-Gaussian properties of (centered) binary random variables and its proof can be found in Buldygin and Moskvichova (2013). The statement therein uses a MGF based definition.
of sub-Gaussians. The statement above is appropriately modified with the factor 2 multiplied in the $\| \cdot \|_\psi$ norm bound to adapt to our definition.

Next, we present a version of the well known Bernstein’s inequality. While the bounds in Lemma 13.2 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 13.5).

**Lemma 13.4 (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 \( \mathbb{E}(Z_i) = \mu_i \forall 1 \leq i \leq n \). Suppose \( \exists \) constants \( \sigma \geq 0 \) and \( K \geq 0 \), such that \( n^{-1} \sum_{i=1}^n \mathbb{E}(|Z_i - \mu_i|^m) \leq (m! / 2) \sigma^2 K^{m-2} \), for each positive integer \( m \geq 2 \). Then,

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

In particular, if \( \{Z_i\}_{i=1}^n \) are i.i.d. realizations of a sub-exponential variable \( Z \) with mean \( \mu \) and \( \| Z \|_\psi \leq \sigma_Z \) for some \( \sigma_Z \geq 0 \), then \( \| Z - \mu \|_\psi \leq 2 \sigma_Z \) 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 \( \sigma_Z \leq \| X \|_\psi \| Y \|_\psi \), and (b) \( Z = XY \) with \( X \) sub-exponential and \( |Y| \leq M \) a.s. for some \( M > 0 \), in which case \( \sigma_Z \leq M \| X \|_\psi \).

**Lemma 13.5 (The Bernstein moment conditions and its verification).** Consider the moment conditions required in Bernstein’s inequality in Lemma 13.4. Let us define any random variable \( Z \) with \( E(Z) = \mu \) to satisfy the Bernstein moment conditions (BMC) with parameters \( (\sigma, K) \), denoted as \( Z \sim \text{BMC}(\sigma, K) \), if for some constants \( \sigma, K \geq 0 \), \( \mathbb{E}(|Z - \mu|^m) \leq (m! / 2) \sigma^2 K^{m-2} \) for each \( m \geq 2 \). Then, if \( Z \) is sub-exponential with \( \| Z \|_\psi \leq \sigma_Z \) for some \( \sigma_Z \geq 0 \), \( Z \sim \text{BMC}(2 \sqrt{2} \sigma_Z, 2 \sigma_Z) \) and \( |Z| \sim \text{BMC}(2 \sqrt{2} \sigma_Z, 2 \sigma_Z) \). Further,

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

(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 \( \sigma_Z = M \| X \|_\psi \).

**Lemma 13.6 (Concentration bounds with variance in the leading term - adopted from Kuchibhotla and Chakrabortty (2018)).** Suppose \( \{X_i\}_{i=1}^n \) are independent mean zero random vectors in \( \mathbb{R}^p \), for any \( p \geq 1 \), such that
for some $\alpha > 0$ and some $K_n > 0$,

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

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

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_\infty \leq 7 \sqrt{\frac{\Gamma_n (t + \log p)}{n}} + \frac{C_\alpha K_n \{ \log(2n) \}^{1/\alpha} (t + \log p)^{1/\alpha^{*}}}{n},$$

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

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