High Dimensional $M$-Estimation & Inference from Observational Data with Incomplete Responses
A Semi-Parametric Doubly Robust Framework

Abhishek Chakrabortty\textsuperscript{1}

Department of Statistics
University of Pennsylvania

Group Meeting
April 24, 2019

\textsuperscript{1}Joint work with Jiarui Lu, T. Tony Cai and Hongzhe Li.
Current era of ‘big data’ and data science $\leadsto$ rapid influx of large and high dimensional data (easily available and computationally tractable).

Rich information on multitudes of variables at the same place $\leadsto$ many interesting scientific questions and also unique statistical challenges!
Current era of ‘big data’ and data science \(\leadsto\) rapid influx of large \textbf{and} high dimensional data (easily available and computationally tractable).

Rich information on multitudes of variables at the same place \(\leadsto\) many interesting scientific questions and also unique statistical challenges!

One frequently encountered challenge: incompleteness of the data and in particular, (partial) missingness of the response of interest.

Reasons could be ‘circumstantial’ (e.g. practical constraints such as logistics, time, cost issues etc.), or it could be ‘by design’ (e.g. due to the ‘treatment’ assignment/non-assignment mechanism).

The response corresponding to a ‘treatment’ of interest could not be observed for a person who is not ‘treated’ (and vice versa).
Current era of ‘big data’ and data science ⇔ rapid influx of large and high dimensional data (easily available and computationally tractable).

Rich information on multitudes of variables at the same place ⇔ many interesting scientific questions and also unique statistical challenges!

One frequently encountered challenge: incompleteness of the data and in particular, (partial) missingness of the response of interest.

Reasons could be ‘circumstantial’ (e.g. practical constraints such as logistics, time, cost issues etc.), or it could be ‘by design’ (e.g. due to the ‘treatment’ assignment/non-assignment mechanism).

The response corresponding to a ‘treatment’ of interest could not be observed for a person who is not ‘treated’ (and vice versa).

Another complication in both cases: observational nature of the data. The missingness mechanism could be informative (not randomized)!
Challenges of Incompleteness Contd. and Relevance in Modern Studies

- Observational data typically informative missingness (or treatment assignment) mechanism. Could depend on the person’s covariates.

- Often termed selection bias or treatment by indication or confounding (in causal inference) in observational studies. Has to be factored in!

- Need to account for the missingness in a proper principled way under minimal conditions to ensure valid, unbiased (and robust) inference.

Relevance: these issues occur in virtually any modern day large scale observational study arising in various scientific disciplines, including: Biomedical studies (e.g. electronic health records (EHR) data); and Integrative genomics (e.g. gene expression data and eQTL studies). Also econometrics (policy evaluation), computer science, finance etc.
Observational data $\sim$ typically informative missingness (or treatment assignment) mechanism. Could depend on the person's covariates.

Often termed selection bias or treatment by indication or confounding (in causal inference) in observational studies. Has to be factored in!

Need to account for the missingness in a proper principled way under minimal conditions to ensure valid, unbiased (and robust) inference.

Relevance: these issues occur in virtually any modern day large scale observational study arising in various scientific disciplines, including:

- Biomedical studies (e.g. electronic health records (EHR) data); and
- Integrative genomics (e.g. gene expression data and eQTL studies).

Also econometrics (policy evaluation), computer science, finance etc.
The Basic Framework and Set-Up

Variables of interest: outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and covariates $X \in \mathcal{X} \subseteq \mathbb{R}^p$ (possibly high dimensional, compared to the sample size).

The supports $\mathcal{Y}$ and $\mathcal{X}$ of $Y$ and $X$ need not be continuous.

Main issue: $Y$ may not always be observed. Let $T \in \{0, 1\}$ denote the indicator of the true $Y$ being observed.

The (partly) unobserved random vector $(T, Y, X)$ is assumed to be jointly defined on a common probability space with measure $\mathbb{P}(\cdot)$.
The Basic Framework and Set-Up

- **Variables of interest:** outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and covariates $X \in \mathcal{X} \subseteq \mathbb{R}^p$ (possibly high dimensional, compared to the sample size).

- The supports $\mathcal{Y}$ and $\mathcal{X}$ of $Y$ and $X$ need not be continuous.

- **Main issue:** $Y$ may not always be observed. Let $T \in \{0, 1\}$ denote the indicator of the true $Y$ being observed.

- The (partly) unobserved random vector $(T, Y, X)$ is assumed to be jointly defined on a common probability space with measure $\mathbb{P}(\cdot)$.

- **Observable data:** $\mathcal{D}_n := \{Z_i := (T_i, T_i Y_i, X_i) : i = 1, \ldots, n\} \overset{iid}{\sim} Z$, where $Z := (T, TY, X)$ whose distribution is defined via $\mathbb{P}(\cdot)$.

- **High dimensional setting:** $p$ can diverge with $n$ (including $p \gg n$).
Applicability of the Framework

- Generally applicable to **any** missing data setting - with missing outcomes $Y$ and (possibly) high dimensional covariates $X$.

- Causal inference problems (via ‘potential’ outcomes framework).
Applicability of the Framework

- Generally applicable to any missing data setting - with missing outcomes $Y$ and (possibly) high dimensional covariates $X$.

- Causal inference problems (via ‘potential’ outcomes framework).

- Here, $X$ is often called ‘confounders’ (for observational studies) or ‘adjustment’ variables/features (for randomized trials).

- Usual set-up: binary ‘treatment’ (a.k.a. exposure/intervention) assignment: $T \in \{0, 1\}$, and potential outcomes: $\{Y(0), Y(1)\}$.

- Observed outcome: $Y := Y(0)1(T = 0) + Y(1)1(T = 1)$, i.e. depending on $T$, we observe only one of $\{Y(0), Y(1)\}$.
Applicability of the Framework

- Generally applicable to **any** missing data setting - with missing outcomes $Y$ and (possibly) high dimensional covariates $X$.

- **Causal inference** problems (via ‘potential’ outcomes framework).

  - Here, $X$ is often called ‘confounders’ (for observational studies) or ‘adjustment’ variables/features (for randomized trials).

  - **Usual set-up:** binary ‘treatment’ (a.k.a. exposure/intervention) assignment: $T \in \{0, 1\}$, and potential outcomes: $\{Y_{(0)}, Y_{(1)}\}$.

  - **Observed outcome:** $\mathbb{Y} := Y_{(0)}1(T = 0) + Y_{(1)}1(T = 1)$, i.e. depending on $T$, we observe only one of $\{Y_{(0)}, Y_{(1)}\}$.

  - For each $j \in \{0, 1\}$, this set-up is included based on the ‘map’:

    $$(T, Y, X) \leftarrow (T_j, Y_{(j)}, X) \text{ with } T_j := 1(T = j) \quad \forall j \in \{0, 1\}.$$
Applicability of the Framework

- Generally applicable to any missing data setting - with missing outcomes \( Y \) and (possibly) high dimensional covariates \( X \).

- Causal inference problems (via ‘potential’ outcomes framework).

  - Here, \( X \) is often called ‘confounders’ (for observational studies) or ‘adjustment’ variables/features (for randomized trials).

  - Usual set-up: binary ‘treatment’ (a.k.a. exposure/intervention) assignment: \( T \in \{0, 1\} \), and potential outcomes: \( \{Y(0), Y(1)\} \).

  - Observed outcome: \( Y := Y(0)1(T = 0) + Y(1)1(T = 1) \), i.e. depending on \( T \), we observe only one of \( \{Y(0), Y(1)\} \).

  - For each \( j \in \{0, 1\} \), this set-up is included based on the ‘map’:

    \[
    (T, Y, X) \leftarrow (T_j, Y(j), X) \quad \text{with} \quad T_j := 1(T = j) \quad \forall \ j \in \{0, 1\}.
    \]

    The case of any multi-category treatment also similarly included.
The Two Standard (Fundamental) Assumptions

1. **Ignorability assumption:** $T \independent Y \mid X$.
   - A.k.a. ‘missing at random’ (MAR) in the missing data literature.
   - A.k.a. ‘no unmeasured confounding’ (NUC) in causal inference.
   - **Special case:** $T \independent (Y, X)$. A.k.a. missing completely at random (MCAR) in missing data literature, and complete randomization (e.g. randomized trials) in causal inference (CI) literature.
The Two Standard (Fundamental) Assumptions

1. **Ignorability assumption: \( T \perp \perp Y \mid X \).**
   - A.k.a. ‘missing at random’ (MAR) in the missing data literature.
   - A.k.a. ‘no unmeasured confounding’ (NUC) in causal inference.
   - Special case: \( T \perp \perp (Y, X) \). A.k.a. missing completely at random (MCAR) in missing data literature, and complete randomization (e.g. randomized trials) in causal inference (CI) literature.

2. **Positivity assumption** (a.k.a. ‘sufficient overlap’ in CI literature):
   - Let \( \pi(X) := \mathbb{P}(T = 1 \mid X) \) be the propensity score (PS), and let \( \pi_0 := \mathbb{P}(T = 1) \). Then, \( \pi(\cdot) \) is uniformly bounded away from 0:
     \[
     1 \geq \pi(x) \geq \delta_\pi > 0 \quad \forall x \in \mathcal{X}, \text{ for some constant } \delta_\pi > 0.
     \]
Relevance in Biomedical Studies: EHR Data

- Rich resources of data for discovery research; fast growing literature.

Using electronic health records to drive discovery in disease genomics

Isaac S. Kohane

Nature Reviews Genetics 12, 417–428 (2011)

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen, Lars J. Jensen & Søren Brunak

Rich resources of data for discovery research; fast growing literature.

Detailed clinical and phenotypic data collected electronically for large patient cohorts, as part of routine health care delivery.
- Rich resources of data for discovery research; fast growing literature.

- Detailed clinical and phenotypic data collected electronically for large patient cohorts, as part of routine health care delivery.

- Structured data: ICD codes, medications, lab tests, demographics etc.

- Unstructured text data (extracted from clinician notes via NLP): signs and symptoms, family history, social history, radiology reports etc.
Information on a **variety** of phenotypes (unlike usual cohort studies).

Opens up **unique opportunities for** novel **integrative analyses**.
EHR Data: The Promises and the Challenges

- Information on a **variety** of phenotypes (unlike usual cohort studies).
- Opens up **unique opportunities for novel integrative analyses**.

**EHR + Bio-repositories** $\leadsto$ genome-phenome association networks, PheWAS studies and genomic risk prediction of diseases.
EHR Data: The Promises and the Challenges

- Information on a **variety** of phenotypes (unlike usual cohort studies).
- Opens up **unique opportunities** for novel **integrative analyses**.

- **EHR + Bio-repositories** $\rightarrow$ genome-phenome association networks, **PheWAS studies** and genomic risk prediction of diseases.

- The key **challenges and bottlenecks** for EHR driven research:
  - Logistic difficulty in obtaining validated phenotype ($Y$) information.
  - Often **time/labor/cost intensive** (and the ICD codes are imprecise).
Some examples of missing $Y$ in EHRs and the reason for missingness:

1. Some (binary) disease phenotype (e.g. Rheumatoid Arthritis).
   - Requires manual chart review by physicians (logistic constraints).

2. Some biomarker (e.g. anti-CCP, an important RA biomarker).
   - Requires lab tests (cost constraints).
   - Similarly, any $Y$ requiring genomic measurements may also have cost/logistics constraints.

Verified phenotypes/treatment response/biomarkers/genomic vars ($Y$) available only for a subset. Clinical features ($X$) available for all.

Further issues: selection bias/treatment by indication/preferential labeling (e.g. sicker patients get labeled/treated/tested more often).

Causal inference problems (treatment effects estimation): EHRs also facilitate comparative effectiveness research on a large scale.

Many treatments/medications (and responses) being observed. All other clinical features ($X$) serve as potential confounders.
Some examples of missing $Y$ in EHRs and the reason for missingness:

1. $Y \rightsquigarrow$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).

2. $Y \rightsquigarrow$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any $Y$ requiring genomic measurements may also have cost/logistics constraints.
Some examples of missing $Y$ in EHRs and the reason for missingness:

1. $Y \sim$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).

2. $Y \sim$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any $Y$ requiring genomic measurements may also have cost/logistics constraints.

- Verified phenotypes/treatment response/biomarkers/genomic vars ($Y$) available only for a subset. Clinical features ($X$) available for all.

- Further issues: selection bias/treatment by indication/preferential labeling (e.g. sicker patients get labeled/treated/tested more often).
Some examples of missing $Y$ in EHRs and the reason for missingness:

1. $Y \rightsquigarrow$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).

2. $Y \rightsquigarrow$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any $Y$ requiring genomic measurements may also have cost/logistics constraints.

- Verified phenotypes/treatment response/biomarkers/genomic vars ($Y$) available only for a subset. Clinical features ($X$) available for all.

- Further issues: selection bias/treatment by indication/preferential labeling (e.g. sicker patients get labeled/treated/tested more often).

- Causal inference problems (treatment effects estimation): EHRs also facilitate comparative effectiveness research on a large scale.

- Many treatments/medications (and responses) being observed. All other clinical features ($X$) serve as potential confounders.
Association studies for gene expression ($Y$) vs. genetic variants ($X$).
Another Example: eQTL Studies (Integrative Genomics)

- Association studies for **gene expression** (\(Y\)) vs. **genetic variants** (\(X\)).

- Popular tools in **integrative genomics** (genetic association studies + gene expression profiling) for understanding **gene regulatory networks**.
Another Example: eQTL Studies (Integrative Genomics)

- Association studies for gene expression ($Y$) vs. genetic variants ($X$).

- Popular tools in integrative genomics (genetic association studies + gene expression profiling) for understanding gene regulatory networks.

- Missing data issue: gene expression data often missing (loss of power), while genetic variants data often available for a much larger group.
Another Example: eQTL Studies (Integrative Genomics)

- Association studies for gene expression ($Y$) vs. genetic variants ($X$).

- Popular tools in integrative genomics (genetic association studies + gene expression profiling) for understanding gene regulatory networks.

- Missing data issue: gene expression data often missing (loss of power), while genetic variants data often available for a much larger group.

- Causal inference: estimate the causal effect of any one variant (the ‘treatment’) on $Y$ while all other variants are potential confounders.
Goal for $M$-estimation: estimation and inference, based on $D_n$, of $	heta_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

$$
\theta_0 \equiv \theta_0(\mathbb{P}) := \arg\min_{\theta \in \mathbb{R}^d} R(\theta), \quad \text{where } R(\theta) := \mathbb{E}\{L(Y, X, \theta)\}
$$

$L(\cdot) \in \mathbb{R}^+$ is any ‘loss’ function that is convex and differentiable in $\theta$. Existence of $\theta_0$ implicitly assumed (guaranteed for most usual probs).

- $d$ can diverge with $n$ (including $d \gg n$). Also, $\theta_0(\mathbb{P})$ is ‘model free’ (no restrictions on $\mathbb{P}$). In particular, no model assumptions on $Y | X$. 

The key challenges: the missingness via $T$ (if not accounted for, the estimator will be inconsistent!) and the high dimensional setting. Need suitable methods - involves estimation of nuisance functions and careful analyses (due to error terms with complex dependencies). 

Special (but low-$d$) case: $\theta_0 = \mathbb{E}(Y)$ and $L(Y, X, \theta) = (Y - \theta)^2$. Leads to the average treatment effect (ATE) estimation prob in CI.
Goal for $M$-estimation: estimation and inference, based on $D_n$, of $\theta_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

$$\theta_0 \equiv \theta_0(\mathbb{P}) := \arg\min_{\theta \in \mathbb{R}^d} R(\theta), \quad \text{where } R(\theta) := \mathbb{E}\{L(Y, X, \theta)\}$$

$L(\cdot) \in \mathbb{R}^+$ is any ‘loss’ function that is convex and differentiable in $\theta$. Existence of $\theta_0$ implicitly assumed (guaranteed for most usual probs).

$d$ can diverge with $n$ (including $d \gg n$). Also, $\theta_0(\mathbb{P})$ is ‘model free’ (no restrictions on $\mathbb{P}$). In particular, no model assumptions on $Y|X$.

The key challenges: the missingness via $T$ (if not accounted for, the estimator will be inconsistent!) and the high dimensional setting.

Need suitable methods - involves estimation of nuisance functions and careful analyses (due to error terms with complex dependencies).
Goal for $M$-estimation: estimation and inference, based on $D_n$, of $\theta_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

$$
\theta_0 \equiv \theta_0(\mathbb{P}) := \arg \min_{\theta \in \mathbb{R}^d} R(\theta), \text{ where } R(\theta) := \mathbb{E}\{L(Y, X, \theta)\} \quad \text{and}
$$

$L(\cdot) \in \mathbb{R}^+$ is any ‘loss’ function that is convex and differentiable in $\theta$. Existence of $\theta_0$ implicitly assumed (guaranteed for most usual probs).

$d$ can diverge with $n$ (including $d \gg n$). Also, $\theta_0(\mathbb{P})$ is ‘model free’ (no restrictions on $\mathbb{P}$). In particular, no model assumptions on $Y|X$.

The key challenges: the missingness via $T$ (if not accounted for, the estimator will be inconsistent!) and the high dimensional setting.

Need suitable methods - involves estimation of nuisance functions and careful analyses (due to error terms with complex dependencies).

Special (but low-$d$) case: $\theta_0 = \mathbb{E}(Y)$ and $L(Y, X, \theta) = (Y - \theta)^2$. Leads to the average treatment effect (ATE) estimation prob in Cl.
The framework includes a broad class of $M/Z$-estimation problems.


This work contributes to both literature above: $M$-estimation + missing data + high dimensional setting and parameter. (Also has applications in heterogeneous treatment effects estimation in CI).
The framework includes a broad class of $M/Z$-estimation problems.

- **$M$-estimation for fully observed data**: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.

- **Missing data/causal inference problems**: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
The framework includes a broad class of $M/Z$-estimation problems.

- **$M$-estimation for fully observed data:** well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.

- **Missing data/causal inference problems:** semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).

This work contributes to both literature above: $M$-estimation + missing data + high dimensional setting and parameter. (Also has applications in heterogeneous treatment effects estimation in CI).
The framework includes a broad class of $M/Z$-estimation problems.

- **$M$-estimation for fully observed data**: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.

- **Missing data/causal inference problems**: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).

- Much less attention when the parameter itself is high dimensional.
The framework includes a broad class of $M/Z$-estimation problems.

- **$M$-estimation for fully observed data**: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.

- **Missing data/causal inference problems**: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).

- Much less attention when the parameter itself is high dimensional.

This work contributes to both literature above: $M$-estimation + missing data + high dimensional setting and parameter. (Also has applications in heterogeneous treatment effects estimation in CI).
HD $M$-Estimation: A Few (Class of) Applications

1. All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.
All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.

- E.g. squared loss: \( L(Y, X, \theta) := (Y - X'\theta)^2 \rightarrow \) linear regression;
  logistic loss: \( L(Y, X, \theta) := \log\{1 + \exp(X'\theta)\} - Y(X'\theta) \rightarrow \) logistic regression (for binary \( Y \)), exponential loss (Poisson reg.) so on .

- Note: throughout, \textit{regardless} of any motivating ‘working model’ being true or not, the definition of \( \theta_0 \) is completely ‘model free’.

---

Series estimation problems (model free) with missing \( Y \) and HD basis functions (instead of \( X \) in Example 1 above). E.g. spline bases.

Use the same choices of \( L(\cdot) \) as in Example 1 above with \( X \) replaced by any set of \( d \) (possibly HD) basis functions \( \Psi(X) := \{\psi_j(X)\}_{j=1}^d \).

E.g. polynomial bases: \( \Psi(X) := \{1, x_1, x_2, \ldots, x_n\} \) with \( d_0 = 1 \rightarrow \) linear bases as in Example 1; \( d_0 = 3 \rightarrow \) cubic splines).
All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.

- E.g. squared loss: \( L(Y, X, \theta) := (Y - X'\theta)^2 \sim \) linear regression;
- logistic loss: \( L(Y, X, \theta) := \log\{1 + \exp(X'\theta)\} - Y(X'\theta) \sim \) logistic regression (for binary \( Y \)), exponential loss (Poisson reg.) so on . . . .

Note: throughout, regardless of any motivating ‘working model’ being true or not, the definition of \( \theta_0 \) is completely ‘model free’.

Series estimation problems (model free) with missing \( Y \) and HD basis functions (instead of \( X \) in Example 1 above). E.g. spline bases.

- Use the same choices of \( L(\cdot) \) as in Example 1 above with \( X \) replaced by any set of \( d \) (possibly HD) basis functions \( \Psi(X) := \{\psi_j(X)\}_{j=1}^d \).
- E.g. polynomial bases: \( \Psi(X) := \{1, x_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\} \) (\( d_0 = 1 \sim \) linear bases as in Example 1; \( d_0 = 3 \sim \) cubic splines).
Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. $X$ is Gaussian).

Let $Y = f(\beta_0'X, \epsilon)$ with $f : \mathbb{R}^2 \rightarrow \mathbb{Y}$ unknown (i.e. $\beta_0$ identifiable only upto scalar multiples) and $\epsilon \perp \perp X$ (i.e., $Y \perp \perp X \mid \beta_0'X$).
Another Application: HD Single Index Models (SIMs)

- Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. \( X \) is Gaussian).

- Let \( Y = f(\beta_0'X, \epsilon) \) with \( f : \mathbb{R}^2 \to \mathcal{Y} \) unknown (i.e. \( \beta_0 \) identifiable only upto scalar multiples) and \( \epsilon \perp \perp X \) (i.e., \( Y \perp \perp X \mid \beta_0'X \)).

- Consider any of the regression problems introduced in Example 1.

  - Let \( \theta_0 := \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E}\{L(Y, X'\theta)\} \) for any convex loss function \( L(\cdot) : \mathbb{R}^2 \to \mathbb{R} \) (convex in the second argument). Then, \( \theta_0 \propto \beta_0 \)!

  - A remarkable result due to Li and Duan (1989).
Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. \( X \) is Gaussian).

Let \( Y = f(\beta_0'X, \epsilon) \) with \( f : \mathbb{R}^2 \rightarrow \mathcal{Y} \) unknown (i.e. \( \beta_0 \) identifiable only up to scalar multiples) and \( \epsilon \perp \perp X \) (i.e., \( Y \perp \perp X \mid \beta_0'X \)).

Consider any of the regression problems introduced in Example 1.

Let \( \theta_0 := \arg \min_{\theta \in \mathbb{R}^p} \mathbb{E} \{ L(Y, X'\theta) \} \) for any convex loss function \( L(\cdot) : \mathbb{R}^2 \rightarrow \mathbb{R} \) (convex in the second argument). Then, \( \theta_0 \propto \beta_0 \).

A remarkable result due to Li and Duan (1989).

Classic example of a misspecified parametric model defining \( \theta_0 \), yet \( \theta_0 \) directly relates to an actual (interpretable) semi-parametric model!

The proportionality result also preserves any sparsity assumptions.
Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

1. Linear heterogeneous treatment effects estimation: application of the linear regression example (twice).

\[
Y(j) = X' \beta(j) + \epsilon(j), \quad \text{E}(\epsilon(j) | X) = 0 \quad \forall j = 0, 1,
\]

so that

\[
Y(1) - Y(0) = X' \beta^* + \epsilon^*, \quad \beta^* := \beta(1) - \beta(0), \quad \epsilon^* := \epsilon(1) - \epsilon(0).
\]

\(\beta^*\) denotes the (model free) linear projection of \(Y(1) - Y(0) | X\).

Of interest in HD settings when \(E(Y(1) - Y(0) | X)\) is difficult to model (Chernozhukov et al., 2017; Chernozhukov and Semenova, 2017).

2. Average conditional treatment effects (ACTE) estimation via series estimators: application of the series estimation example (twice).

3. Causal inference via SIMs (signal recovery, ACTE estimation and ATE estimation): application of the SIM example (twice).
Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

1. Linear heterogeneous treatment effects estimation: application of the linear regression example (twice). Write \( \{Y(0), Y(1)\} \) linearly as:

\[
Y(j) = X' \beta(j) + \epsilon(j), \quad E(\epsilon(j)X) = 0 \quad \forall \ j = 0, 1, \quad \text{so that}
\]

\[
Y(1) - Y(0) = X' \beta^* + \epsilon^*, \quad \beta^* := \beta(1) - \beta(0), \quad \epsilon^* := \epsilon(1) - \epsilon(0).
\]
Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

1. **Linear heterogeneous treatment effects estimation**: application of the linear regression example (twice). Write \( \{ Y(0), Y(1) \} \) linearly as:

\[
Y(j) = X' \beta(j) + \epsilon(j), \quad E(\epsilon(j)X) = 0 \quad \forall \ j = 0, 1,
\]

so that

\[
Y(1) - Y(0) = X' \beta^* + \epsilon^*, \quad \beta^* := \beta(1) - \beta(0), \quad \epsilon^* := \epsilon(1) - \epsilon(0).
\]

\( \beta^* \) denotes the (model free) linear projection of \( Y(1) - Y(0) | X \). Of interest in HD settings when \( E\{ Y(1) - Y(0) | X \} \) is difficult to model (Chernozhukov et al., 2017; Chernozhukov and Semenova, 2017).

2. **Average conditional treatment effects (ACTE) estimation via series estimators**: application of the series estimation example (twice).

3. **Causal inference via SIMs (signal recovery, ACTE estimation and ATE estimation)**: application of the SIM example (twice).
Applications of all these problems in causal inference (estimation of
treatment effects with useful applications in precision medicine):

1. Linear heterogeneous treatment effects estimation: application of
   the linear regression example (twice). Write \( \{Y_{(0)}, Y_{(1)}\} \) linearly as:

   \[
   Y_{(j)} = X'\beta_{(j)} + \epsilon_{(j)}, \quad \mathbb{E}(\epsilon_{(j)}|X) = 0 \quad \forall \ j = 0, 1, \quad \text{so that}
   \]

   \[
   Y_{(1)} - Y_{(0)} = X'\beta^* + \epsilon^*, \quad \beta^* := \beta_{(1)} - \beta_{(0)}, \quad \epsilon^* := \epsilon_{(1)} - \epsilon_{(0)}.
   \]

   \( \beta^* \) denotes the (model free) linear projection of \( Y_{(1)} - Y_{(0)}|X \). Of
   interest in HD settings when \( \mathbb{E}\{Y_{(1)} - Y_{(0)}|X\} \) is difficult to model
   (Chernozhukov et al., 2017; Chernozhukov and Semenova, 2017).

2. Average conditional treatment effects (ACTE) estimation via series
   estimators: application of the series estimation example (twice).

3. Causal inference via SIMs (signal recovery, ACTE estimation and
   ATE estimation): application of the SIM example (twice).
Some notations: \( m(X) := \mathbb{E}(Y|X) \) and \( \phi(X, \theta) := \mathbb{E}\{L(Y, X, \theta)|X\} \).
Before Getting Started: A Few Facts and Considerations

- Some notations: \( m(X) \) := \( E( Y | X ) \) and \( \phi(X, \theta) \) := \( E\{L(Y, X, \theta) | X\} \).

- It is generally necessary to ‘account’ for the missingness in \( Y \). The ‘complete case’ estimator of \( \theta_0 \) in general will be inconsistent!
Some notations: \( m(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X}) \) and \( \phi(\mathbf{X}, \theta) := \mathbb{E}\{L(Y, \mathbf{X}, \theta)|\mathbf{X}\} \).

It is generally necessary to ‘account’ for the missingness in \( Y \). The ‘complete case’ estimator of \( \theta_0 \) in general will be inconsistent!

That estimator may be consistent only if: (1) \( \nabla \phi(\mathbf{X}, \theta_0) = 0 \) a.s. for every \( \mathbf{X} \) (for regression problems, this indicates the ‘correct model’ case), and/or (2) \( T \perp \perp (Y, \mathbf{X}) \) (i.e. the MCAR case).

Illustration of (1) for sq. loss: \( \nabla \phi(\mathbf{X}, \theta_0) = \mathbb{E}\{\mathbf{X}(Y - \mathbf{X}'\theta_0)|\mathbf{X}\} = 0 \). Hence, \( \mathbb{E}(Y|\mathbf{X}) = \mathbf{X}'\theta_0 \) (i.e. a ‘linear model’ holds for \( Y|\mathbf{X} \)).
Before Getting Started: A Few Facts and Considerations

- Some notations: \( m(X) := \mathbb{E}(Y | X) \) and \( \phi(X, \theta) := \mathbb{E}\{L(Y, X, \theta) | X\} \).

- It is generally necessary to ‘account’ for the missingness in \( Y \). The ‘complete case’ estimator of \( \theta_0 \) in general will be inconsistent!

- That estimator may be consistent only if: (1) \( \nabla \phi(X, \theta_0) = 0 \) a.s. for every \( X \) (for regression problems, this indicates the ‘correct model’ case), and/or (2) \( T \perp \perp (Y, X) \) (i.e. the MCAR case).

- Illustration of (1) for sq. loss: \( \nabla \phi(X, \theta_0) = \mathbb{E}\{X(Y - X'\theta_0) | X\} = 0 \). Hence, \( \mathbb{E}(Y | X) = X'\theta_0 \) (i.e. a ‘linear model’ holds for \( Y | X \)).

- With \( \theta_0 \) (and \( X \)) being high dimensional (compared to \( n \)), we need some further structural constraints on \( \theta_0 \) to estimate it using \( \mathcal{D}_n \).

- We assume that \( \theta_0 \) is s-sparse: \( \|\theta_0\|_0 := s \) and \( s \leq \min(n, d) \).

- Note: the sparsity requirement has attractive (and fairly intuitive) geometric justification for all the examples we have given here.
Under MAR assmpn., \( R(\theta) := \mathbb{E}\{L(Y, X, \theta)\} \equiv \mathbb{E}_X\{\phi(X, \theta)\} \) admits the following debiased and doubly robust (DDR) representation:
Estimation of $\theta_0$: Getting Identifiable Representation(s) of $R(\theta)$

- Under MAR assmpn., $R(\theta) := \mathbb{E}\{L(Y, X, \theta)\} \equiv \mathbb{E}_X\{\phi(X, \theta)\}$ admits the following debiased and doubly robust (DDR) representation:

$$R(\theta) = \mathbb{E}_X\{\phi(X, \theta)\} + \mathbb{E}\left[\frac{T}{\pi(X)} \{L(Y, X, \theta) - \phi(X, \theta)\}\right].$$  \hspace{1cm} (1)

Purely non-parametric identification based on the observable $Z$ and the nuisance functions: $\pi(X)$ and $\phi(X, \theta)$ (unknown but estimable).
Estimation of $\theta_0$: Getting Identifiable Representation(s) of $R(\theta)$

- Under MAR assumption, $R(\theta) := \mathbb{E}\{L(Y, X, \theta)\} \equiv \mathbb{E}_X\{\phi(X, \theta)\}$ admits the following debiased and doubly robust (DDR) representation:

  $$R(\theta) = \mathbb{E}_X\{\phi(X, \theta)\} + \mathbb{E}\left[\frac{T}{\pi(X)} \{L(Y, X, \theta) - \phi(X, \theta)\}\right]. \tag{1}$$

Purely non-parametric identification based on the observable $Z$ and the nuisance functions: $\pi(X)$ and $\phi(X, \theta)$ (unknown but estimable).

- 2\textsuperscript{nd} term is simply 0, can be seen as a ‘debiasing’ term (of sorts).

- Plays a crucial role in analyzing the empirical version of (1). Ensures first order insensitivity to any estimation errors of $\pi(\cdot)$ and $\phi(\cdot)$. 
Estimation of $\theta_0$: Getting Identifiable Representation(s) of $R(\theta)$

Under MAR assmpn., $R(\theta) := \mathbb{E}\{L(Y, X, \theta)\} \equiv \mathbb{E}_X\{\phi(X, \theta)\}$ admits the following debiased and doubly robust (DDR) representation:

$$R(\theta) = \mathbb{E}_X\{\phi(X, \theta)\} + \mathbb{E}\left[\frac{T}{\pi(X)} \{L(Y, X, \theta) - \phi(X, \theta)\}\right].$$

Purely non-parametric identification based on the observable $Z$ and the nuisance functions: $\pi(X)$ and $\phi(X, \theta)$ (unknown but estimable).

2\textsuperscript{nd} term is simply 0, can be seen as a ‘debiasing’ term (of sorts).

Plays a crucial role in analyzing the empirical version of (1). Ensures first order insensitivity to any estimation errors of $\pi(\cdot)$ and $\phi(\cdot)$.

Double robustness (DR) aspect: replace $\{\phi(X, \theta), \pi(X)\}$ by any $\{\phi^*(X, \theta), \pi^*(X)\}$ and (1) continues to hold as long as one but not necessarily both of $\phi^*(\cdot) = \phi(\cdot)$ or $\pi^*(\cdot) = \pi(\cdot)$ hold.
Given any estimators \( \{\hat{\pi}(\cdot), \hat{\phi}(\cdot)\} \) be of the nuisance fns. \( \{\pi(\cdot), \phi(\cdot)\} \), we define our \( L_1 \)-penalized DDR estimator \( \hat{\theta}_{\text{DDR}} \) of \( \theta_0 \) as:

\[
\hat{\theta}_{\text{DDR}} \equiv \hat{\theta}_{\text{DDR}}(\lambda_n) := \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}^{\text{DDR}}_n(\theta) + \lambda_n \|\theta\|_1 \right\},
\]

where

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

\( \lambda_n \geq 0 \) is the tuning parameter and \( \{\hat{\pi}(\cdot), \hat{\phi}(\cdot)\} \) are arbitrary except for satisfying two basic conditions regarding their construction:
Given any estimators \( \{\hat{\pi}(\cdot), \hat{\phi}(\cdot)\} \) be of the nuisance fns. \( \{\pi(\cdot), \phi(\cdot)\} \), we define our \( L_1 \)-penalized DDR estimator \( \hat{\theta}_{\text{DDR}} \) of \( \theta_0 \) as:

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

where

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

\( \lambda_n \geq 0 \) is the tuning parameter and \( \{\hat{\pi}(\cdot), \hat{\phi}(\cdot)\} \) are arbitrary except for satisfying two basic conditions regarding their construction:

- \( \hat{\pi}(\cdot) \) obtained from the data \( T_n := \{T_i, X_i\}_{i=1}^{n} \) only; \( \{\hat{\phi}(X_i, \theta)\}_{i=1}^{n} \) obtained in a ‘cross-fitted’ manner (via sample splitting).

- Assume (temporarily) \( \{\hat{\pi}(\cdot), \hat{\phi}(\cdot)\} \) are both ‘correct’. DR properties (consistency) of \( \hat{\theta}_{\text{DDR}} \) under their misspecifications discussed later.
For simplicity, assume that the gradient $\nabla L(Y, X, \theta)$ of $L(\cdot)$ satisfies a ‘separable form’ as follows: for some $h(X) \in \mathbb{R}^d$ and $g(X, \theta) \in \mathbb{R}$,
For simplicity, assume that the gradient $\nabla L(Y, X, \theta)$ of $L(\cdot)$ satisfies a ‘separable form’ as follows: for some $h(X) \in \mathbb{R}^d$ and $g(X, \theta) \in \mathbb{R}$,

$$\nabla L(Y, X, \theta) = h(X)\{Y - g(X, \theta)\},$$

and hence,

$$\nabla \hat{\phi}(X, \theta) = h(X)\{\hat{m}(X) - g(X, \theta)\},$$

where $\hat{m}(X)$ denotes the corresponding (cross-fitted) estimator of $m(X)$. This simplifying assumption holds for all examples given before.

Assumed form $\Rightarrow$ only need to obtain $\hat{m}(X_i)$ and not $\hat{\phi}(X_i, \theta)$. 
For simplicity, assume that the gradient $\nabla L(Y, X, \theta)$ of $L(\cdot)$ satisfies a ‘separable form’ as follows: for some $h(X) \in \mathbb{R}^d$ and $g(X, \theta) \in \mathbb{R}$,

$$\nabla L(Y, X, \theta) = h(X)\{Y - g(X, \theta)\}, \quad \text{and hence,}$$

$$\nabla \hat{\phi}(X, \theta) = h(X)\{\hat{m}(X) - g(X, \theta)\}, \quad \text{where}$$

$\hat{m}(X)$ denotes the corresponding (cross-fitted) estimator of $m(X)$. This simplifying assumption holds for all examples given before.

Assumed form $\Rightarrow$ only need to obtain $\hat{m}(X_i)$ and not $\hat{\phi}(X_i, \theta)$.

**Implementation algorithm.** $\hat{\theta}_{\text{DDR}}$ can be obtained simply as:

$$\hat{\theta}_{\text{DDR}} \equiv \hat{\theta}_{\text{DDR}}(\lambda_n) := \arg \min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(\tilde{Y}_i, X_i, \theta) + \lambda_n \|\theta\|_1 \right\},$$

where $\tilde{Y}_i := \hat{m}(X_i) + \frac{T_i}{\hat{p}(X_i)}\{Y_i - \hat{m}(X_i)\}, \forall i$, is a ‘pseudo’ outcome.

Can use ‘glmnet’ in R. Pretend to have a ‘full’ data: $\{\tilde{Y}_i, X_i\}_{i=1}^n$. 

---

Abhishek Chakrabortty  High-Dim. $M$-Estimation with Missing Responses: A Semi-Parametric Framework  19/50
Assume $L(\cdot)$ is convex and differentiable in $\theta$ and $\mathcal{L}_{n}^{DDR}(\theta)$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\theta = \theta_0$. Then, for any choice of $\lambda_n \geq 2 \| \nabla \mathcal{L}_{n}^{DDR}(\theta_0) \|_\infty$, 

\[ \| \hat{\theta}_{DDR}(\lambda_n) - \theta_0 \|_2 \lesssim \lambda_n \sqrt{s}, \]

\[ \| \hat{\theta}_{DDR}(\lambda_n) - \theta_0 \|_1 \lesssim \lambda_n s. \]
Assume $L(\cdot)$ is convex and differentiable in $\theta$ and $\mathcal{L}_n^{DDR}(\theta)$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\theta = \theta_0$. Then, for any choice of $\lambda_n \geq 2 \|\nabla \mathcal{L}_n^{DDR}(\theta_0)\|_{\infty}$,

$$\left\| \hat{\theta}_{DDR}(\lambda_n) - \theta_0 \right\|_2 \lesssim \lambda_n \sqrt{s}, \text{ and } \left\| \hat{\theta}_{DDR}(\lambda_n) - \theta_0 \right\|_1 \lesssim \lambda_n s.$$ 

where $s := \|\theta_0\|_0$. This is a deterministic deviation bound. Holds for any choices of $\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}$ and for any realization of $\mathcal{D}_n$. 
Properties of $\hat{\theta}_{\text{DDR}}$: Deterministic Deviation Bounds

- Assume $L(\cdot)$ is convex and differentiable in $\theta$ and $L_n^{\text{DDR}}(\theta)$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\theta = \theta_0$. Then, for any choice of $\lambda_n \geq 2 \|\nabla L_n^{\text{DDR}}(\theta_0)\|_\infty$,

$$\|\hat{\theta}_{\text{DDR}}(\lambda_n) - \theta_0\|_2 \lesssim \lambda_n \sqrt{s}, \quad \text{and} \quad \|\hat{\theta}_{\text{DDR}}(\lambda_n) - \theta_0\|_1 \lesssim \lambda_n s.$$

where $s := \|\theta_0\|_0$. This is a deterministic deviation bound. Holds for any choices of $\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}$ and for any realization of $D_n$.

- The RSC (or ‘cone’) condition for $L_n^{\text{DDR}}(\theta)$ is exactly the same as the usual RSC condition required under a fully observed data! The fully observed data RSC condition’s validity is well studied.
Properties of $\hat{\theta}_{\text{DDR}}$: Deterministic Deviation Bounds

Assume $L(\cdot)$ is convex and differentiable in $\theta$ and $L_n^{\text{DDR}}(\theta)$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\theta = \theta_0$. Then, for any choice of $\lambda_n \geq 2 \| \nabla L_n^{\text{DDR}}(\theta_0) \|_{\infty}$,

$$
\| \hat{\theta}_{\text{DDR}}(\lambda_n) - \theta_0 \|_2 \lesssim \lambda_n \sqrt{s}, \ \text{and} \ \| \hat{\theta}_{\text{DDR}}(\lambda_n) - \theta_0 \|_1 \lesssim \lambda_n s.
$$

where $s := \| \theta_0 \|_0$. This is a deterministic deviation bound. Holds for any choices of $\{ \hat{\pi}(\cdot), \hat{m}(\cdot) \}$ and for any realization of $D_n$.

The RSC (or 'cone') condition for $L_n^{\text{DDR}}(\theta)$ is exactly the same as the usual RSC condition required under a fully observed data! The fully observed data RSC condition’s validity is well studied.

Key quantity of interest: the random lower bound $\| \nabla L_n^{\text{DDR}}(\theta_0) \|_{\infty}$ for $\lambda_n$. Need probabilistic bounds to determine convergence rate of $\hat{\theta}_{\text{DDR}}$. 
Bounds on $\|\nabla L_n^{\text{DDR}}(\theta_0)\|_\infty$ determines the rate of choice of $\lambda_n$ and hence the convergence rate of $\hat{\theta}_{\text{DDR}}$ (using the deviation bound).

**Probabilistic bounds for $\|\nabla L_n^{\text{DDR}}(\theta_0)\|_\infty$:**

The basic decomposition

$$\|\nabla L_n^{\text{DDR}}(\theta_0)\|_\infty \leq \|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty + \|R_{\pi,m,n}\|_\infty,$$
The Main Goal from Hereon: Probabilistic Bounds for $\| \nabla L_n^{DDR}(\theta_0) \|_\infty$

- Bounds on $\| \nabla L_n^{DDR}(\theta_0) \|_\infty$ determines the rate of choice of $\lambda_n$ and hence the convergence rate of $\hat{\theta}_{DDR}$ (using the deviation bound).

- **Probabilistic** bounds for $\| \nabla L_n^{DDR}(\theta_0) \|_\infty$: the basic decomposition

  $$\| \nabla L_n^{DDR}(\theta_0) \|_\infty \leq \| T_{0,n} \|_\infty + \| T_{\pi,n} \|_\infty + \| T_{m,n} \|_\infty + \| R_{\pi,m,n} \|_\infty,$$

  where $T_{0,n}$ is the ‘main’ term (a centered iid average), $T_{\pi,n}$ is the ‘$\pi$-error’ term involving $\hat{\pi}(\cdot) - \pi(\cdot)$ and $T_{m,n}$ is the ‘$m$-error’ term involving $\hat{m}(\cdot) - m(\cdot)$, while $R_{\pi,m,n}$ is the ‘($\pi$, $m$)-error’ term (usually lower order) involving the product of $\hat{\pi}(\cdot) - \pi(\cdot)$ and $\hat{m}(\cdot) - m(\cdot)$.

- Control each term separately. The analyses are all non-asymptotic and nuanced, especially in order to get sharp rates for $T_{\pi,n}$ and $T_{m,n}$.
The Main Goal from Hereon: Probabilistic Bounds for $\|\nabla L_n^{DDR}(\theta_0)\|_\infty$

- Bounds on $\|\nabla L_n^{DDR}(\theta_0)\|_\infty$ determines the rate of choice of $\lambda_n$ and hence the convergence rate of $\hat{\theta}_{DDR}$ (using the deviation bound).

- **Probabilistic** bounds for $\|\nabla L_n^{DDR}(\theta_0)\|_\infty$: the basic decomposition

  $$\|\nabla L_n^{DDR}(\theta_0)\|_\infty \leq \|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty + \|R_{\pi,m,n}\|_\infty,$$

  where $T_{0,n}$ is the ‘main’ term (a centered iid average), $T_{\pi,n}$ is the ‘$\pi$-error’ term involving $\hat{\pi}(\cdot) - \pi(\cdot)$ and $T_{m,n}$ is the ‘$m$-error’ term involving $\hat{m}(\cdot) - m(\cdot)$, while $R_{\pi,m,n}$ is the ‘($\pi$, $m$)-error’ term (usually lower order) involving the product of $\hat{\pi}(\cdot) - \pi(\cdot)$ and $\hat{m}(\cdot) - m(\cdot)$.

- Control each term separately. The analyses are all non-asymptotic and nuanced, especially in order to get sharp rates for $T_{\pi,n}$ and $T_{m,n}$.

- **We show:** $\|\nabla L_n^{DDR}(\theta_0)\|_\infty \lesssim \sqrt{(\log d)/n}$ with high probability, and hence $\|\hat{\theta}_{DDR} - \theta_0\|_2 \lesssim \sqrt{s(\log d)/n}$. So, clearly it is rate optimal.
Basic (high level) consistency conditions on \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \). Let \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \) be any general and ‘correct’ estimators of \( \{\pi(\cdot), m(\cdot)\} \), and assume they satisfy the following pointwise convergence rates:
Basic (high level) consistency conditions on \(\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}\). Let \(\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}\) be any general and ‘correct’ estimators of \(\{\pi(\cdot), m(\cdot)\}\), and assume they satisfy the following \textbf{pointwise} convergence rates:

\[
|\hat{\pi}(x) - \pi(x)| \lesssim_{\mathbb{P}} \delta_{n,\pi} \quad \text{and} \quad |\hat{m}(x) - m(x)| \lesssim_{\mathbb{P}} \xi_{n,m} \quad \forall \ x \in \mathcal{X}, \quad (2)
\]

for some sequences \(\delta_{n,\pi}, \xi_{n,m} \geq 0\) such that \((\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)\) and the product \(\delta_{n,\pi} \xi_{n,m}(\log n) = o\left(\sqrt{\log d}/n\right)\).
Basic (high level) consistency conditions on \{\hat{\pi}(\cdot), \hat{m}(\cdot)\}. Let \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} be any general and ‘correct’ estimators of \{\pi(\cdot), m(\cdot)\}, and assume they satisfy the following pointwise convergence rates:

\[ |\hat{\pi}(x) - \pi(x)| \lesssim_P \delta_{n,\pi} \quad \text{and} \quad |\hat{m}(x) - m(x)| \lesssim_P \xi_{n,m} \quad \forall \ x \in \mathcal{X}, \quad (2) \]

for some sequences \( \delta_{n,\pi}, \xi_{n,m} \geq 0 \) such that \((\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1) \) and the product \( \delta_{n,\pi}\xi_{n,m}(\log n) = o(\sqrt{\log d}/n) \).

Under condition (2), along with some more ‘suitable’ assumptions (sub-Gaussian tails etc.), we have: with high probability,

\[ \|T_{0,n}\|_{\infty} \lesssim \sqrt{\frac{\log d}{n}} \], \[ \|T_{\pi,n}\|_{\infty} \lesssim \sqrt{\frac{\log d}{n}} \left\{ \delta_{n,\pi}\sqrt{\log(nd)} \right\}, \quad \text{and} \]
Basic (high level) consistency conditions on \{\hat{\pi}(\cdot), \hat{m}(\cdot)\}. Let \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} be any general and ‘correct’ estimators of \{\pi(\cdot), m(\cdot)\}, and assume they satisfy the following pointwise convergence rates:

\[ |\hat{\pi}(x) - \pi(x)| \lesssim_P \delta_{n,\pi} \quad \text{and} \quad |\hat{m}(x) - m(x)| \lesssim_P \xi_{n,m} \quad \forall \; x \in \mathcal{X}, \quad (2) \]

for some sequences \(\delta_{n,\pi}, \xi_{n,m} \geq 0\) such that \((\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)\) and the product \(\delta_{n,\pi} \xi_{n,m}(\log n) = o\left(\sqrt{(\log d)/n}\right)\).

Under condition (2), along with some more ‘suitable’ assumptions (sub-Gaussian tails etc.), we have: with high probability,

\[ \|T_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}, \quad \|T_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ \delta_{n,\pi} \sqrt{\log(nd)} \right\}, \quad \text{and} \]

\[ \|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ \xi_{n,m} \sqrt{\log(nd)} \right\}, \quad \|R_{\pi,m,n}\|_\infty \lesssim \delta_{n,\pi} \xi_{n,m}(\log n). \]
Basic (high level) consistency conditions on \( \{ \hat{\pi}(\cdot), \hat{m}(\cdot) \} \). Let \( \{ \hat{\pi}(\cdot), \hat{m}(\cdot) \} \) be any general and ‘correct’ estimators of \( \{ \pi(\cdot), m(\cdot) \} \), and assume they satisfy the following pointwise convergence rates:

\[
|\hat{\pi}(x) - \pi(x)| \lesssim_P \delta_{n,\pi} \quad \text{and} \quad |\hat{m}(x) - m(x)| \lesssim_P \xi_{n,m} \quad \forall \ x \in \mathcal{X}, \quad (2)
\]

for some sequences \( \delta_{n,\pi}, \xi_{n,m} \geq 0 \) such that \( (\delta_{n,\pi} + \xi_{n,m}) \sqrt{\log(nd)} = o(1) \) and the product \( \delta_{n,\pi} \xi_{n,m} (\log n) = o(\sqrt{\log d/n}) \).

Under condition (2), along with some more ‘suitable’ assumptions (sub-Gaussian tails etc.), we have: with high probability,

\[
\|T_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}, \quad \|T_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ \delta_{n,\pi} \sqrt{\log(nd)} \right\}, \quad \text{and}
\]

\[
\|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ \xi_{n,m} \sqrt{\log(nd)} \right\}, \quad \|R_{\pi,m,n}\|_\infty \lesssim \delta_{n,\pi} \xi_{n,m} (\log n).
\]

Hence, \( \| \nabla L_n^{\text{DDR}}(\theta_0) \|_\infty \lesssim \sqrt{\frac{\log d}{n}} \{1 + o(1)\} \) with high probability.
Consider $\hat{\theta}_{\text{DDR}}$ for the squared loss: $L(Y, X, \theta) := \{ Y - \Psi(X)'\theta \}^2$, where $\Psi(X) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $X$.

Define $\Sigma := \mathbb{E}\{\Psi(X)\Psi(X)\}'$, $\Omega := \Sigma^{-1}$, and let $\hat{\Omega}$ be any reasonable estimator of $\Omega$ (and assume $\Omega$ is sparse if required).

We then define the desparsified DDR estimator $\tilde{\theta}_{\text{DDR}}$ as follows.
Consider $\hat{\theta}_{\text{DDR}}$ for the squared loss: $L(Y, X, \theta) := \{Y - \Psi(X)'\theta\}^2$, where $\Psi(X) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $X$.

Define $\Sigma := \mathbb{E}\{\Psi(X)\Psi(X)\}'$, $\Omega := \Sigma^{-1}$, and let $\hat{\Omega}$ be any reasonable estimator of $\Omega$ (and assume $\Omega$ is sparse if required).

We then define the desparsified DDR estimator $\tilde{\theta}_{\text{DDR}}$ as follows.

$$\tilde{\theta}_{\text{DDR}} := \hat{\theta}_{\text{DDR}} + \hat{\Omega} \frac{1}{n} \sum_{i=1}^{n} \{\tilde{Y}_i - \Psi(X_i)'\hat{\theta}_{\text{DDR}}\} \Psi(X_i),$$

where

Desparsification/Debiasing term

$$\tilde{Y}_i := \hat{m}(X_i) + \frac{T_i}{\hat{\pi}(X_i)} \{Y_i - \hat{m}(X_i)\}$$

are the pseudo outcomes.
Consider $\hat{\theta}_{\text{DDR}}$ for the squared loss: $L(Y, X, \theta) := \{Y - \Psi(X)\'\theta\}^2$, where $\Psi(X) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $X$.

Define $\Sigma := \mathbb{E}\{\Psi(X)\Psi(X)\}'$, $\Omega := \Sigma^{-1}$, and let $\hat{\Omega}$ be any reasonable estimator of $\Omega$ (and assume $\Omega$ is sparse if required).

We then define the desparsified DDR estimator $\tilde{\theta}_{\text{DDR}}$ as follows.

$$\tilde{\theta}_{\text{DDR}} := \hat{\theta}_{\text{DDR}} + \hat{\Omega} \frac{1}{n} \sum_{i=1}^{n} \{\tilde{Y}_i - \Psi(X_i)\'\hat{\theta}_{\text{DDR}}\} \Psi(X_i),$$

where

$$\tilde{Y}_i := \hat{m}(X_i) + \frac{T_i}{\hat{\pi}(X_i)} \{Y_i - \hat{m}(X_i)\}$$
are the pseudo outcomes.

Debiasing similar (in spirit) to van de Geer et al. (2014), except its the ‘right’ one for this problem (using pseudo outcomes in the full data).
Assume the basic convergence conditions (2) for \{\hat{\pi}(\cdot), \hat{m}(\cdot)\}, and that 
\[ \|\hat{\Omega} - \Omega\|_1 = O_P(a_n), \quad \|I - \hat{\Omega}\hat{\Sigma}\|_{\text{max}} = O_P(b_n) \text{ and } \|\Omega\|_1 = O(1), \]
with \(a_n\sqrt{\log d} = o(1)\) and \(b_n s \sqrt{\log d} = o(1)\), where \(s := \|\theta_0\|_0\).

Then, \(\tilde{\theta}_{\text{DDR}}\) satisfies the asymptotic linear expansion (ALE):
Assume the basic convergence conditions (2) for \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \), and that \( \|\hat{\Omega} - \Omega\|_1 = O_P(a_n) \), \( \|I - \hat{\Omega}\hat{\Sigma}\|_{\max} = O_P(b_n) \) and \( \|\Omega\|_1 = O(1) \), with \( a_n\sqrt{\log d} = o(1) \) and \( b_n s\sqrt{\log d} = o(1) \), where \( s := \|\theta_0\|_0 \).

Then, \( \tilde{\theta}_{\text{DDR}} \) satisfies the asymptotic linear expansion (ALE):

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

and \( \psi_0(Z) := \left[ \{m(X) - \Psi(X)'^T\theta_0\} + \frac{T}{\pi(X)} \{Y - m(X)\} \right] \Psi(X) \)

with \( \mathbb{E}\{\psi_0(Z)\} = 0 \). The ALE facilitates inference (e.g. confidence intervals etc.) for any low-d component of \( \theta_0 \) via Gaussian approx.
Assume the basic convergence conditions (2) for \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \), and that 
\[ \|\hat{\Omega} - \Omega\|_1 = O_P(a_n), \|I - \hat{\Omega}\hat{\Sigma}\|_{\text{max}} = O_P(b_n) \text{ and } \|\Omega\|_1 = O(1), \]
with \( a_n\sqrt{\log d} = o(1) \) and \( b_n s \sqrt{\log d} = o(1) \), where \( s := \|\theta_0\|_0 \).

Then, \( \tilde{\theta}_{\text{DDR}} \) satisfies the asymptotic linear expansion (ALE):

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

and \( \psi_0(Z) := \left[ \{m(X) - \Psi(X)'\theta_0\} + \frac{T}{\pi(X)} \{Y - m(X)\} \right] \Psi(X) \)

with \( \mathbb{E}\{\psi_0(Z)\} = 0 \). The ALE facilitates inference (e.g. confidence intervals etc.) for any low-d component of \( \theta_0 \) via Gaussian approx.

Further, the ALE is also ‘optimal’. The function \( \Omega \psi_0(Z) =: \Psi_{\text{eff}}(Z) \) is the ‘efficient’ influence function for \( \theta_0 \) (Robins et al., 1994). Thus, in classical settings, \( \tilde{\theta}_{\text{DDR}} \) achieves the semi-parametric efficiency bound.
Coordinate-wise asymptotic normality of $\tilde{\theta}_{\text{DDR}}$: $\forall 1 \leq j \leq d$,

$$\sqrt{n}(\tilde{\theta}_{\text{DDR}} - \theta_0)_j \overset{d}{\to} \mathcal{N}(0, \sigma^2_{0,j}),$$

where $\sigma^2_{0,j} := \text{Var}\{\Omega_j \cdot \psi_0(Z)\}$.

Further, $\max_{1 \leq j \leq d} |\hat{\sigma}_{0,j} - \sigma_{0,j}| = o_P(1)$, where $\hat{\sigma}_{0,j}$ is the plug-in estimator obtained by plugging in $\hat{\Omega}$, $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ in $\text{Var}\{\Omega_j \cdot \psi_0(Z)\}$.

Can choose $\hat{\Omega}$ to be any standard (sparse) precision matrix estimator, e.g. the node-wise Lasso estimator. Here, $a_n = s_\Omega \sqrt{(\log d)/n}$ and $b_n = \sqrt{(\log d)/n}$ under suitable conditions, with $s_\Omega := \max_{1 \leq j \leq d} \|\Omega_j\|_0$. 

Abhishek Chakraborty High-Dim. $M$-Estimation with Missing Responses: A Semi-Parametric Framework 25/50
Coordinate-wise asymptotic normality of $\tilde{\theta}_{\text{DDR}}$: $\forall 1 \leq j \leq d$,

$$\sqrt{n}(\tilde{\theta}_{\text{DDR}} - \theta_0)_j \xrightarrow{d} \mathcal{N}(0, \sigma^2_{0,j}),$$

where $\sigma^2_{0,j} := \text{Var}\{\Omega'_j \psi_0(Z)\}$.

Further, $\max_{1 \leq j \leq d} |\hat{\sigma}_{0,j} - \sigma_{0,j}| = o_P(1)$, where $\hat{\sigma}_{0,j}$ is the plug-in estimator obtained by plugging in $\hat{\Omega}$, $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ in $\text{Var}\{\Omega'_j \psi_0(Z)\}$.

Can choose $\hat{\Omega}$ to be any standard (sparse) precision matrix estimator, e.g. the node-wise Lasso estimator. Here, $a_n = s_{\Omega} \sqrt{(\log d) / n}$ and $b_n = \sqrt{(\log d) / n}$ under suitable conditions, with $s_{\Omega} := \max_{1 \leq j \leq d} \|\Omega_j\|_0$.

The error $\Delta_n$ can be decomposed as: $\Delta_n = \Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}$, where $\Delta_{n,1} := \frac{1}{n}(\hat{\Omega} - \Omega) \sum_{i=1}^{n} \psi_0(Z_i)$, $\Delta_{n,2} := (I_d - \hat{\Omega}\hat{\Sigma})(\hat{\theta}_{\text{DDR}} - \theta_0)$ and $\Delta_{n,3} := \hat{\Omega}(T_{\pi,n} + T_{m,n} + R_{\pi,m,n})$, with $\|\Delta_{n,3}\|_\infty \lesssim_P n^{-\frac{1}{2}}$ and

$$\|\Delta_{n,1}\|_\infty \lesssim a_n \sqrt{\frac{\log d}{n}} \quad \text{and} \quad \|\Delta_{n,2}\|_\infty \lesssim b_n s \sqrt{\frac{\log d}{n}}.$$
Finally, let \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\} \), with either \( \pi^*(\cdot) = \pi(\cdot) \) or \( m^*(\cdot) = m(\cdot) \) but not necessarily both. Assume the same pointwise convergence conditions and rates \( (\delta_{n,\pi}, \xi_{n,m}) \) for \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \) as in (2), but now with \( \{\pi(\cdot), m(\cdot)\} \) therein replaced by \( \{\pi^*(\cdot), m^*(\cdot)\} \).

Under some ‘suitable’ assumptions, we have: with high probability,
Finally, let \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \to \{\pi^*(\cdot), m^*(\cdot)\} \), with either \( \pi^*(\cdot) = \pi(\cdot) \) or \( m^*(\cdot) = m(\cdot) \) but not necessarily both. Assume the same pointwise convergence conditions and rates \((\delta_{n,\pi}, \xi_{n,m})\) for \(\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}\) as in (2), but now with \(\{\pi(\cdot), m(\cdot)\}\) therein replaced by \(\{\pi^*(\cdot), m^*(\cdot)\}\).

Under some ‘suitable’ assumptions, we have: with high probability,

\[
\|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ 1 + 1_{(\pi^*, m^*) \neq (\pi, m)} \right\}
\]
Finally, let \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\} \), with either \( \pi^*(\cdot) = \pi(\cdot) \) or \( m^*(\cdot) = m(\cdot) \) but not necessarily both. Assume the same pointwise convergence conditions and rates \( (\delta_n, \xi_n, \pi, m) \) for \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \) as in (2), but now with \( \{\pi(\cdot), m(\cdot)\} \) therein replaced by \( \{\pi^*(\cdot), m^*(\cdot)\} \).

Under some ‘suitable’ assumptions, we have: with high probability,

\[
\|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ 1 + 1_{(\pi^*, m^*) \neq (\pi, m)} \right\}
\]

and

\[
\|R_{\pi,m,n}\|_\infty \lesssim \left\{ \delta_{n,\pi} 1_{(m^* \neq m)} + \xi_{n,m} 1_{(\pi^* \neq \pi)} + \delta_{n,\pi} \xi_{n,m} \right\} (\log n).
\]
Finally, let $\{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \to \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ but not necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\hat{\pi}(\cdot), \hat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

Under some ‘suitable’ assumptions, we have: with high probability,

$$\|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty \lesssim \frac{\sqrt{\log d}}{n} \left\{ 1 + 1_{(\pi^*, m^*) \neq (\pi, m)} \right\}$$

and

$$\|R_{\pi,m,n}\|_\infty \lesssim \left\{ \delta_{n,\pi} 1_{(m^* \neq m)} + \xi_{n,m} 1_{(\pi^* \neq \pi)} + \delta_{n,\pi} \xi_{n,m} \right\} (\log n).$$

The $2^{nd}$ and/or $3^{rd}$ terms also contribute now to the rate $\sqrt{\log d}/n$. The $4^{th}$ term is $o(1)$ but no longer ignorable (and may be slower).
Finally, let \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \to \{\pi^*(\cdot), m^*(\cdot)\} \), with either \( \pi^*(\cdot) = \pi(\cdot) \) or \( m^*(\cdot) = m(\cdot) \) but not necessarily both. Assume the same pointwise convergence conditions and rates \( (\delta_{n,\pi}, \xi_{n,m}) \) for \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \) as in (2), but now with \( \{\pi(\cdot), m(\cdot)\} \) therein replaced by \( \{\pi^*(\cdot), m^*(\cdot)\} \).

Under some ‘suitable’ assumptions, we have: with high probability,

\[
\|T_{0,n}\|_\infty + \|T_{\pi,n}\|_\infty + \|T_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{ 1 + 1_{(\pi^*,m^*) \neq (\pi,m)} \right\}
\]

and

\[
\|R_{\pi,m,n}\|_\infty \lesssim \left\{ \delta_{n,\pi} 1_{(m^* \neq m)} + \xi_{n,m} 1_{(\pi^* \neq \pi)} + \delta_{n,\pi} \xi_{n,m} \right\} (\log n).
\]

The 2\textsuperscript{nd} and/or 3\textsuperscript{rd} terms also contribute now to the rate \( \sqrt{(\log d)/n} \).

The 4\textsuperscript{th} term is \( o(1) \) but no longer ignorable (and may be slower).

Regardless, this establishes general convergence rates and the DR property of \( \hat{\theta}_{\text{DDR}} \) under possible misspecification of \( \{\hat{\pi}(\cdot), \hat{m}(\cdot)\} \).

For the 4\textsuperscript{th} term, sharper rates need a case-by-case analysis.
Note: our theory holds generally for any choices of \( \hat{\pi}(\cdot) \) and \( \hat{m}(\cdot) \) under mild conditions (provided they are both ‘correct’ estimators).

Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates need case-by-case analyses. Even for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.
Note: our theory holds generally for any choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ under mild conditions (provided they are both ‘correct’ estimators).

Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates need case-by-case analyses.

Even for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.

Below we provide only some choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ that may be used to implement our theory & methods for $\hat{\theta}_{\text{DDR}}$. In general, one can use any reasonable method (including black box ML methods).

Choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$: we consider estimators from two families.
Choices of the Nuisance Component Estimators $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$

- **Note:** our theory holds generally for any choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ under mild conditions (provided they are both ‘correct’ estimators).

- Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates need case-by-case analyses.

- **Even** for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.

- Below we provide only some choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ that may be used to implement our theory & methods for $\hat{\theta}_{DDR}$. In general, one can use any reasonable method (including black box ML methods).

- Choices of $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$: we consider estimators from two families.
  
  - Parametric and ‘extended’ parametric families (series estimators).
  
  - Semi-parametric single index families.
If $\pi(\cdot)$ is known, we set $\hat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\hat{\pi}(\cdot)$ (each assumed to be ‘correct’).
Choices of $\hat{\pi}(\cdot)$: ‘Extended’ Parametric Families (Series Estimators)

- If $\pi(\cdot)$ is known, we set $\hat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\hat{\pi}(\cdot)$ (each assumed to be ‘correct’).

- ‘Extended’ parametric family: $\pi(x) = g\{\alpha'\Psi(X)\}$, where $g(\cdot) \in [0, 1]$ is a known function [e.g. $g_{\text{expit}}(u) := \exp(u)/(1 + \exp(u))$], $\Psi(X) := \{\psi_k(X)\}_{k=1}^K$ is any set of $K$ basis functions (with $K \gg n$ possibly), and $\alpha \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
Choices of $\hat{\pi}(\cdot)$: ‘Extended’ Parametric Families (Series Estimators)

- If $\pi(\cdot)$ is known, we set $\hat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\hat{\pi}(\cdot)$ (each assumed to be ‘correct’).

- ‘Extended’ parametric family: $\pi(x) = g\{\alpha'\Psi(X)\}$, where $g(\cdot) \in [0,1]$ is a known function [e.g. $g_{\text{expit}}(u) := \exp(u)/(1 + \exp(u))$], $\Psi(X) := \{\psi_k(X)\}_{k=1}^K$ is any set of $K$ basis functions (with $K \gg n$ possibly), and $\alpha \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.

- Example: $\Psi(X)$ may correspond to the polynomial bases of $X$ upto any fixed degree $k$. Note: the special case of linear bases ($k = 1$) includes all standard parametric regression models. Further, the case of $\pi(\cdot) =$ constant (but unknown) i.e. MCAR is also included.
If \( \pi(\cdot) \) is known, we set \( \hat{\pi}(\cdot) := \pi(\cdot) \). Otherwise, we estimate \( \pi(\cdot) \) via two (class of) choices of \( \hat{\pi}(\cdot) \) (each assumed to be ‘correct’).

- ‘Extended’ parametric family: \( \pi(x) = g\{\alpha'\Psi(X)\} \), where \( g(\cdot) \in [0,1] \) is a known function [e.g. \( g_{\text{expit}}(u) := \exp(u)/(1 + \exp(u)) \)], \( \Psi(X) := \{\psi_k(X)\}_{k=1}^K \) is any set of \( K \) basis functions (with \( K \gg n \) possibly), and \( \alpha \in \mathbb{R}^K \) is an unknown (sparse) parameter vector.

  - Example: \( \Psi(X) \) may correspond to the polynomial bases of \( X \) upto any fixed degree \( k \). Note: the special case of linear bases (\( k = 1 \)) includes all standard parametric regression models. Further, the case of \( \pi(\cdot) = \text{constant} \) (but unknown) i.e. MCAR is also included.

- Estimator: we set \( \hat{\pi}(X) = g\{\hat{\alpha}'\Psi(X)\} \), where \( \hat{\alpha} \) denotes any suitable estimator (possibly penalized) of \( \alpha \) based on \( T_n := \{T_i, X_i\}_{i=1}^n \).
If \( \pi(\cdot) \) is known, we set \( \hat{\pi}(\cdot) := \pi(\cdot) \). Otherwise, we estimate \( \pi(\cdot) \) via two (class of) choices of \( \hat{\pi}(\cdot) \) (each assumed to be ‘correct’).

‘Extended’ parametric family: \( \pi(x) = g\{\alpha'\Psi(X)\} \), where \( g(\cdot) \in [0,1] \) is a known function [e.g. \( g_{\text{expit}}(u) := \exp(u)/(1 + \exp(u)) \)], \( \Psi(X) := \{\psi_k(X)\}_{k=1}^K \) is any set of \( K \) basis functions (with \( K \gg n \) possibly), and \( \alpha \in \mathbb{R}^K \) is an unknown (sparse) parameter vector.

Example: \( \Psi(X) \) may correspond to the polynomial bases of \( X \) upto any fixed degree \( k \). Note: the special case of linear bases (\( k = 1 \) includes all standard parametric regression models. Further, the case of \( \pi(\cdot) = \text{constant} \) (but unknown) i.e. MCAR is also included.

Estimator: we set \( \hat{\pi}(X) = g\{\hat{\alpha}'\Psi(X)\} \), where \( \hat{\alpha} \) denotes any suitable estimator (possibly penalized) of \( \alpha \) based on \( T_n := \{T_i, X_i\}_{i=1}^n \).

Example of \( \hat{\alpha} \): when \( g(\cdot) = g_{\text{expit}}(\cdot) \), \( \hat{\alpha} \) may be obtained based on a standard \( L_1 \)-penalized logistic regression of \( \{T_i \text{ vs. } \Psi(X_i)\}_{i=1}^n \).
Semi-parametric single index family: $\pi(\mathbf{X}) = g(\alpha' \mathbf{X})$, where $g(\cdot) \in (0, 1)$ is unknown and $\alpha \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable only upto scalar multiples, hence set $\|\alpha\|_2 = 1$ wlog).
Choices of \( \hat{\pi}(\cdot) \): Semi-Parametric Single Index Families

- Semi-parametric single index family: \( \pi(X) = g(\alpha'X) \), where \( g(\cdot) \in (0, 1) \) is unknown and \( \alpha \in \mathbb{R}^p \) is a (sparse) unknown parameter (identifiable only up to scalar multiples, hence set \( \|\alpha\|_2 = 1 \) wlog).

- Given an estimator \( \hat{\alpha} \) of \( \alpha \), we estimate \( \pi(X) \equiv \mathbb{E}(T | \alpha'X) \) as:

\[
\hat{\pi}(x) \equiv \hat{\pi}(\hat{\alpha}, x) := \frac{\frac{1}{nh} \sum_{i=1}^{n} T_i K \{ \hat{\alpha}'(X_i - x)/h \}}{\frac{1}{nh} \sum_{i=1}^{n} K \{ \hat{\alpha}'(X_i - x)/h \}},
\]

where \( K(\cdot) \) denotes any standard (2\textsuperscript{nd} order) kernel function and \( h = h_n > 0 \) denotes the bandwidth sequence with \( h = o(1) \).
Semi-parametric single index family: $\pi(X) = g(\alpha'X)$, where $g(\cdot) \in (0, 1)$ is unknown and $\alpha \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable only upto scalar multiples, hence set $\|\alpha\|_2 = 1$ wlog).

Given an estimator $\hat{\alpha}$ of $\alpha$, we estimate $\pi(X) \equiv \mathbb{E}(T | \alpha'X)$ as:

$$\hat{\pi}(x) \equiv \hat{\pi}(\hat{\alpha}, x) := \frac{\frac{1}{nh} \sum_{i=1}^{n} T_i K \left\{ \hat{\alpha}'(X_i - x)/h \right\}}{\frac{1}{nh} \sum_{i=1}^{n} K \left\{ \hat{\alpha}'(X_i - x)/h \right\}},$$

where $K(\cdot)$ denotes any standard ($2^{nd}$ order) kernel function and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.

Obtaining $\hat{\alpha}$: In general, any approach (if available) from (high dimensional) single index model literature can be used. But if $X$ is elliptically symmetric, then $\hat{\alpha}$ may be obtained as simply as a standard $L_1$-penalized logistic regression of $\{T_i \text{ vs. } X_i\}_{i=1}^{n}$. 

Abhishek Chakrabortty High-Dim. M-Estimation with Missing Responses: A Semi-Parametric Framework
‘Extended’ parametric family: $m(x) = g\{\gamma' \Psi(X)\}$, where $g(\cdot)$ is a known ‘link’ function [e.g. ‘canonical’ links: identity, expit or exp], $\Psi(X) := \{\psi_k(X)\}_{k=1}^K$ is any set of $K$ basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
‘Extended’ parametric family: \( m(x) = g\{\gamma' \Psi(X)\} \), where \( g(\cdot) \) is a known ‘link’ function [e.g. ‘canonical’ links: identity, expit or exp], \( \Psi(X) := \{\psi_k(X)\}_{k=1}^K \) is any set of \( K \) basis functions (with \( K \gg n \) possibly), and \( \gamma \in \mathbb{R}^K \) is an unknown (sparse) parameter vector.

Example: \( \Psi(X) \) may correspond to the polynomial bases of \( X \) upto any fixed degree \( k \). Note: the special case of linear bases \( (k = 1) \) includes all standard parametric regression models.
Choices of $\hat{m}(\cdot)$: ‘Extended’ Parametric Families

- ‘Extended’ parametric family: $m(x) = g\{\gamma'\Psi(X)\}$, where $g(\cdot)$ is a known ‘link’ function [e.g. ‘canonical’ links: identity, expit or exp], $\Psi(X) := \{\psi_k(X)\}_{k=1}^{K}$ is any set of $K$ basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.

- Example: $\Psi(X)$ may correspond to the polynomial bases of $X$ upto any fixed degree $k$. Note: the special case of linear bases ($k = 1$) includes all standard parametric regression models.

- Estimator: we set $\hat{m}(X) = g\{\hat{\gamma}'\Psi(X)\}$, where $\hat{\gamma}$ denotes any suitable estimator (possibly penalized) of $\gamma$ based on the data subset of ‘complete cases’: $D_n^{(c)} := \{(Y_i, X_i) \mid T_i = 1\}_{i=1}^{n}$.
Choices of $\hat{m}(\cdot)$: ‘Extended’ Parametric Families

- ‘Extended’ parametric family: $m(x) = g\{\gamma'\Psi(X)\}$, where $g(\cdot)$ is a known ‘link’ function [e.g. ‘canonical’ links: identity, expit or exp], $\Psi(X) := \{\psi_k(X)\}_{k=1}^{K}$ is any set of $K$ basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.

- Example: $\Psi(X)$ may correspond to the polynomial bases of $X$ upto any fixed degree $k$. Note: the special case of linear bases ($k = 1$) includes all standard parametric regression models.

- Estimator: we set $\hat{m}(X) = g\{\hat{\gamma}'\Psi(X)\}$, where $\hat{\gamma}$ denotes any suitable estimator (possibly penalized) of $\gamma$ based on the data subset of ‘complete cases’: $D_n^{(c)} := \{(Y_i, X_i) \mid T_i = 1\}_{i=1}^n$.

- Example of $\hat{\gamma}$: when $g(\cdot)$ := any ‘canonical’ link function, $\hat{\gamma}$ may be simply obtained based on the respective usual $L_1$-penalized ‘canonical’ link based regression (e.g. linear, logistic or poisson) of $\{(Y_i \ vs. X_i) \mid T_i = 1\}_{i=1}^n$ from the ‘complete case’ data $D_n^{(c)}$. 

Abhishek Chakrabortty  High-Dim. $M$-Estimation with Missing Responses: A Semi-Parametric Framework 30/50
Semi-parametric single index family: \( m(X) = g(\gamma'X) \), where \( g(\cdot) \) is an unknown ‘link’ and \( \gamma \in \mathbb{R}^p \) is a (sparse) unknown parameter (identifiable only upto scalar multiples, hence set \( \|\gamma\|_2 = 1 \) wlog).
Choices of $\hat{m}(\cdot)$: Semi-Parametric Single Index Families

- Semi-parametric single index family: $m(X) = g(\gamma'X)$, where $g(\cdot)$ is an unknown ‘link’ and $\gamma \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable only up to scalar multiples, hence set $\|\gamma\|_2 = 1$ wlog).

- Given an estimator $\hat{\gamma}$ of $\gamma$, we estimate $m(X) \equiv \mathbb{E}(Y | \gamma'X, T)$ as:

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

where $K(\cdot)$ denotes any standard (2\textsuperscript{nd} order) kernel function, and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.
Semi-parametric single index family: \( m(X) = g(\gamma'X) \), where \( g(\cdot) \) is an unknown ‘link’ and \( \gamma \in \mathbb{R}^p \) is a (sparse) unknown parameter (identifiable only up to scalar multiples, hence set \( ||\gamma||_2 = 1 \) wlog).

Given an estimator \( \hat{\gamma} \) of \( \gamma \), we estimate \( m(X) \equiv \mathbb{E}(Y | \gamma'X, T) \) as:

\[
\hat{m}(x) \equiv \hat{m}(\hat{\gamma}, x) := \frac{1}{nh} \sum_{i=1}^{n} T_i Y_i \frac{K\{\hat{\gamma}'(X_i - x)/h\}}{\frac{1}{nh} \sum_{i=1}^{n} T_i K\{\hat{\gamma}'(X_i - x)/h\}},
\]

where \( K(\cdot) \) denotes any standard (2nd order) kernel function, and \( h = h_n > 0 \) denotes the bandwidth sequence with \( h = o(1) \).

Obtaining \( \hat{\gamma} \): In general, any approach (if available) from HD SIM literature can be used on the complete case data subset \( D_n^{(c)} \).

If \( X \) is elliptically symmetric and \( Y = f(\gamma'X; \epsilon) \) with \( f \) unknown and \( \epsilon \perp \perp (T, X) \), then \( \hat{\gamma} \) may be obtained as \( L_1 \)-penalized IPW estimator \( \hat{\theta}_{IPW} \) for any ‘canonical’ link based regression problem.
For either choices of $\hat{\pi}(\cdot)$, assume that the ingredient estimator $\hat{\alpha}$ satisfies:

$$\|\hat{\alpha} - \alpha\|_1 \lesssim P a_n$$

for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),
For either choices of $\hat{\pi}(\cdot)$, assume that the ingredient estimator $\hat{\alpha}$ satisfies: $\|\hat{\alpha} - \alpha\|_1 \lesssim_P a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$\left|\hat{\pi}(x) - \pi(x)\right| \lesssim a_n = o(1), \text{ for any fixed } x \in \mathcal{X}, \text{ (for method 1).}$$
For either choices of $\hat{\pi}(\cdot)$, assume that the ingredient estimator $\hat{\alpha}$ satisfies: $\|\hat{\alpha} - \alpha\|_1 \lesssim_P a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$|\hat{\pi}(x) - \pi(x)| \lesssim a_n = o(1), \text{ for any fixed } x \in \mathcal{X}, \text{ (for method 1)}.$$ 

For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $x \in \mathcal{X}$,

$$|\hat{\pi}(x) - \pi(x)| \lesssim \left(h^2 + \frac{1}{\sqrt{nh}}\right) + \left(a_n + \frac{\log(np)}{nh} + \frac{a_n^2}{h^2}\right) = o(1).$$
Convergence Rates Regarding The Choices of $\hat{\pi}(\cdot)$

- For either choices of $\hat{\pi}(\cdot)$, assume that the ingredient estimator $\hat{\alpha}$ satisfies: $\|\hat{\alpha} - \alpha\|_1 \lesssim_{\mathbb{P}} a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$|\hat{\pi}(x) - \pi(x)| \lesssim a_n = o(1), \text{ for any fixed } x \in \mathcal{X}, \text{ (for method 1)}.$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $x \in \mathcal{X}$,

$$|\hat{\pi}(x) - \pi(x)| \lesssim \left( h^2 + \frac{1}{\sqrt{nh}} \right) + \left( a_n + \frac{\log(np)}{nh} + \frac{a_n^2}{h^2} \right) = o(1).$$

- Usually, we expect the $L_1$ error rate of $\hat{\alpha}$ to be $a_n = s_\alpha \sqrt{(\log d_*)/n}$ where $s_\alpha := \|\alpha\|_0$ and $d_* = K$ or $p$ (depending on the method).
For either choices of $\hat{m}(\cdot)$, assume that the ingredient estimator $\hat{\gamma}$ satisfies: $\|\hat{\gamma} - \gamma\|_1 \lesssim_{\mathbb{P}} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: \textbf{with high probability},
For either choices of \( \hat{m}(\cdot) \), assume that the ingredient estimator \( \hat{\gamma} \) satisfies: \( \Vert \hat{\gamma} - \gamma \Vert_1 \lesssim_P b_n \) for some \( b_n = o(1) \). Then, under suitable smoothness and tail assumptions, we have: with high probability,

\[
|\hat{m}(x) - m(x)| \lesssim b_n = o(1) \quad \text{for any fixed } x \in \mathcal{X} \quad \text{(for method 1)}.
\]
Convergence Rates Regarding the Choices of $\hat{m}(\cdot)$

- For either choices of $\hat{m}(\cdot)$, assume that the ingredient estimator $\hat{\gamma}$ satisfies: $\|\hat{\gamma} - \gamma\|_1 \lesssim P b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

  $$|\hat{m}(x) - m(x)| \lesssim b_n = o(1) \quad \text{for any fixed } x \in \mathcal{X} \quad \text{(for method 1)}.$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $x \in \mathcal{X}$,

  $$|\hat{m}(x) - m(x)| \lesssim \left( h^2 + \frac{1}{\sqrt{nh}} \right) + \left( b_n + \frac{\log(np)}{nh} + \frac{b_n^2}{h^2} \right) = o(1).$$
For either choices of $\hat{m}(\cdot)$, assume that the ingredient estimator $\hat{\gamma}$ satisfies: $\|\hat{\gamma} - \gamma\|_1 \lesssim \mathbb{P} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

$$|\hat{m}(x) - m(x)| \lesssim b_n = o(1) \quad \text{for any fixed } x \in \mathcal{X} \quad \text{(for method 1)}.$$  

For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $x \in \mathcal{X}$,

$$|\hat{m}(x) - m(x)| \lesssim \left( h^2 + \frac{1}{\sqrt{nh}} \right) + \left( b_n + \frac{\log(np)}{nh} + \frac{b_n^2}{h^2} \right) = o(1).$$  

We typically expect the $L_1$ error rate of $\hat{\gamma}$ to be $b_n = s_\gamma \sqrt{(\log d_*)/n}$ where $s_\gamma := \|\alpha\|_0$ and $d_* = K$ or $p$ (depending on the method).
Basic parameters: \( n = 1000, \ p = 50 \text{ or } 500 \) and \( X \sim \mathcal{N}(0, \Sigma_p) \).

Three data generating processes (DGPs) for \( Y \mid X \) and \( T \mid X \) as follows:

1. **"Linear-Linear" DGP:**
   \[
   Y = \gamma_0 + \gamma' X + \epsilon, \quad \epsilon \mid X \sim \mathcal{N}(0, 1).
   \]

2. **"Quad-Quad" DGP:**
   \[
   Y = \gamma_0 + \gamma' X + p \sum_{j=1}^{p} \gamma_j^* X_j^2 + \epsilon, \quad \epsilon \mid X \sim \mathcal{N}(0, 1).
   \]

3. **"SIM-SIM" DGP:**
   \[
   Y = \gamma_0 + \gamma' X + c Y (\gamma' X)^2 + \epsilon, \quad \epsilon \mid X \sim \mathcal{N}(0, 1).
   \]
Simulation Studies: The Setup

- Basic parameters: \( n = 1000, \ p = 50 \) or \( 500 \) and \( \mathbf{X} \sim \mathcal{N}(0, \Sigma_p) \).

- Three data generating processes (DGPs) for \( Y|\mathbf{X} \) and \( T|\mathbf{X} \) as follows:
  1. “Linear-Linear” DGP:
     \[
     Y = \gamma_0 + \mathbf{\gamma}'\mathbf{X} + \epsilon, \quad \epsilon|\mathbf{X} \sim \mathcal{N}(0, 1).
     \]
     \[
     \text{logit}\{\pi(\mathbf{X})\} \equiv \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \alpha_0 + \mathbf{\alpha}'\mathbf{X}.
     \]
Simulation Studies: The Setup

- Basic parameters: \( n = 1000, p = 50 \) or \( 500 \) and \( \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \).

- Three data generating processes (DGPs) for \( Y|\mathbf{X} \) and \( T|\mathbf{X} \) as follows:

  1. “Linear-Linear” DGP:

     \[
     Y = \gamma_0 + \gamma' \mathbf{X} + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0, 1).
     \]

     \[
     \text{logit}\{\pi(\mathbf{X})\} \equiv \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \alpha_0 + \alpha' \mathbf{X}.
     \]

  2. “Quad-Quad” DGP:

     \[
     Y = \gamma_0 + \gamma' \mathbf{X} + \sum_{j=1}^{p} \gamma_j^* X_j^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0, 1).
     \]

     \[
     \text{logit}\{\pi(\mathbf{X})\} \equiv \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \alpha_0 + \alpha' \mathbf{X} + \sum_{j=1}^{p} \alpha_j^* X_{ij}^2.
     \]
Simulation Studies: The Setup

- Basic parameters: \( n = 1000, \ p = 50 \) or \( 500 \) and \( X \sim \mathcal{N}(0, \Sigma_p) \).

- Three data generating processes (DGPs) for \( Y \mid X \) and \( T \mid X \) as follows:

1. “Linear-Linear” DGP:
   \[
   Y = \gamma_0 + \gamma'X + \varepsilon, \quad \varepsilon \mid X \sim \mathcal{N}(0, 1).
   \]
   \[
   \logit\{\pi(X)\} \equiv \logit\{\mathbb{E}(T \mid X)\} = \alpha_0 + \alpha'X.
   \]

2. “Quad-Quad” DGP:
   \[
   Y = \gamma_0 + \gamma'X + \sum_{j=1}^{p} \gamma_j^* X_j^2 + \varepsilon, \quad \varepsilon \mid X \sim \mathcal{N}(0, 1).
   \]
   \[
   \logit\{\pi(X)\} \equiv \logit\{\mathbb{E}(T \mid X)\} = \alpha_0 + \alpha'X_i + \sum_{j=1}^{p} \alpha_j^* X_{ij}^2.
   \]

3. “SIM-SIM” DGP:
   \[
   Y = \gamma_0 + \gamma'X + c_Y (\gamma'X)^2 + \varepsilon, \quad \varepsilon \mid X \sim \mathcal{N}(0, 1).
   \]
   \[
   \logit\{\pi(X)\} \equiv \logit\{\mathbb{E}(T \mid X)\} = \alpha_0 + \alpha'X + c_T (\alpha'X)^2.
   \]
Choices of the parameters:

1. Covariance matrix $\Sigma_p$: $\Sigma_p = I_p$ (identity matrix).

2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.

3. When $p = 50$, $\alpha = \frac{1}{\sqrt{5}}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\alpha\|_0 = 5$, $\gamma = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\gamma\|_0 = 10$, $\alpha^* = (0.25, -0.25, 0, \cdots, 0)$ and $\gamma^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$. 

$K = 2$ fold cross-fitting used; all simulation settings replicated 500 times.

$\hat{\Omega}$ obtained as $\hat{\Sigma}^{-1}$ for $p = 50$ and using the nodewise Lasso for $p = 500$. 
Simulation Settings: Choice of Parameters

- **Choices of the parameters:**
  1. Covariance matrix $\Sigma_p$: $\Sigma_p = I_p$ (identity matrix).
  2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.
  3. When $p = 50$, $\alpha = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\alpha\|_0 = 5$, $\gamma = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\gamma\|_0 = 10$, $\alpha^* = (0.25, -0.25, 0, \cdots, 0)$ and $\gamma^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.
  4. When $p = 500$, $\|\alpha\|_0 = 10$ and $\alpha$ consists of three $1$s, two $-1$s, two $0.5$s and three $-0.5$s normalized by $1/\sqrt{10}$, while $\|\gamma\|_0 = 15$ and $\gamma$ consists of three $1$s, two $-1$s, five $0.5$s, five $-0.5$s, two $0.25$s and three $-0.25$s. Further, we set $\alpha^* = (0.25, 0.25, -0.25, -0.25, 0, \cdots, 0)$ and $\gamma^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$. 

$\hat{\Omega}$ obtained as $\hat{\Sigma}^{-1}$ for $p = 50$ and using the nodewise Lasso for $p = 500$. 

Abhishek Chakrabortty  
High-Dim. $M$-Estimation with Missing Responses: A Semi-Parametric Framework  
35/50
Simulation Settings: Choice of Parameters

- **Choices of the parameters:**
  1. Covariance matrix $\Sigma_p$: $\Sigma_p = I_p$ (identity matrix).
  2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.
  3. When $p = 50$, $\alpha = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\alpha\|_0 = 5$, $\gamma = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\gamma\|_0 = 10$, $\alpha^* = (0.25, -0.25, 0, \cdots, 0)$ and $\gamma^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.
  4. When $p = 500$, $\|\alpha\|_0 = 10$ and $\alpha$ consists of three 1s, two $-1$s, two 0.5s and three $-0.5$s normalized by $1/\sqrt{10}$, while $\|\gamma\|_0 = 15$ and $\gamma$ consists of three 1s, two $-1$s, five 0.5s, five $-0.5$s, two 0.25s and three $-0.25$s. Further, we set $\alpha^* = (0.25, 0.25, -0.25, -0.25, 0, \cdots, 0)$ and $\gamma^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.

- $K = 2$ fold cross-fitting used; all simulation settings replicated 500 times.

- $\hat{\Omega}$ obtained as $\hat{\Sigma}^{-1}$ for $p = 50$ and using the nodewise Lasso for $p = 500$. 
Obtain the DDR estimator $\hat{\theta}_{\text{DDR}}$ for linear regression: $\theta_0 = \Sigma^{-1}\mathbb{E}(XY)$. 

Two choices of the working nuisance models for $\pi(X)$ to obtain $\hat{\pi}(X)$:
1. Linear: $L_1$ penalized logistic-linear regression.
2. Quad: $L_1$ penalized logistic-linear regression with quadratic terms.

Three choices of the working nuisance models for $m(X)$ to obtain $\hat{m}(X)$:
1. Linear: $L_1$ penalized linear regression.
2. Quad: $L_1$ penalized linear regression with quadratic terms.
3. SIM: Single index model (with index parameter estimated via IPW Lasso)

Estimators used for comparison:
1. $\hat{\theta}_{\text{orac}}$ (Oracle): obtained assuming both $\pi(\cdot)$ and $m(\cdot)$ are known.
2. $\hat{\theta}_{\text{full}}$ (Super oracle): obtained assuming a full dataset is observed.

Criteria: $L_2$ errors for estimation and coverage probability for inference.
Obtain the DDR estimator $\hat{\theta}_{\text{DDR}}$ for linear regression: $\theta_0 = \Sigma^{-1}E(XY)$.

Two choices of the working nuisance models for $\pi(X)$ to obtain $\hat{\pi}(X)$:

1. Linear: $L_1$ penalized logistic-linear regression.
2. Quad: $L_1$ penalized logistic-linear regression with quadratic terms.

Estimators used for comparison:

1. $\hat{\theta}_{\text{orac}}$ (Oracle): obtained assuming both $\pi(\cdot)$ and $m(\cdot)$ are known.
2. $\hat{\theta}_{\text{full}}$ (Super oracle): obtained assuming a full dataset is observed.

Criteria: $L_2$ errors for estimation and coverage probability for inference.
Obtain the DDR estimator $\hat{\theta}_{\text{DDR}}$ for linear regression: $\theta_0 = \Sigma^{-1}\mathbb{E}(XY)$.

**Two choices of the working nuisance models for $\pi(X)$ to obtain $\hat{\pi}(X)$:**

1. **Linear**: $L_1$ penalized logistic-linear regression.
2. **Quad**: $L_1$ penalized logistic-linear regression with quadratic terms.

**Three choices of the working nuisance models for $m(X)$ to obtain $\hat{m}(X)$:**

1. **Linear**: $L_1$ penalized linear regression.
2. **Quad**: $L_1$ penalized linear regression with quadratic terms.
3. **SIM**: Single index model (with index parameter estimated via IPW Lasso)
Obtain the DDR estimator $\hat{\theta}_{\text{DDR}}$ for linear regression: $\theta_0 = \Sigma^{-1}E(\mathbf{X}Y)$.

Two choices of the working nuisance models for $\pi(\mathbf{X})$ to obtain $\hat{\pi}(\mathbf{X})$:
1. Linear: $L_1$ penalized logistic-linear regression.
2. Quad: $L_1$ penalized logistic-linear regression with quadratic terms.

Three choices of the working nuisance models for $m(\mathbf{X})$ to obtain $\hat{m}(\mathbf{X})$:
1. Linear: $L_1$ penalized linear regression.
2. Quad: $L_1$ penalized linear regression with quadratic terms.
3. SIM: Single index model (with index parameter estimated via IPW Lasso)

Estimators used for comparison:
1. $\hat{\theta}_{\text{orac}}$ (Oracle): obtained assuming both $\pi(\cdot)$ and $m(\cdot)$ are known.
2. $\hat{\theta}_{\text{full}}$ (Super oracle): obtained assuming a full dataset is observed.

Criteria: $L_2$ errors for estimation and coverage probability for inference.
Simulation Results: $L_2$ Error Comparison ($p = 50$) - I

$p = 50$, DGP: Linear-Linear.

![Bar chart showing error comparison for different estimators and models.](chart.png)

- $\hat{m}$: linear
- $\hat{m}$: quad
- $\hat{m}$: SIM
- $\hat{\pi}$: linear
- $\hat{\pi}$: quad
- $\hat{\pi}$: SIM

Legend:
- DDR
- Oracle
- Super Oracle
Simulation Results: $L_2$ Error Comparison ($p = 50$) - II

$p = 50$, DGP: Quad-Quad.
Simulation Results: $L_2$ Error Comparison ($p = 50$) - III

$p = 50$, DGP: SIM-SIM.
Simulation Results: $L_2$ Error Comparison ($p = 500$) - 1

$p = 500$, DGP: Linear-Linear.
Simulation Results: $L_2$ Error Comparison ($p = 500$) - II

$p = 500$, DGP: Quad-Quad.

![Graph showing comparison of different estimation methods for various $\hat{m}$ and $\hat{\pi}$ models.]
Simulation Results: $L_2$ Error Comparison ($p = 500$) - III

$p = 500$, **DGP: SIM-SIM.**

![Bar chart showing comparison of $\hat{m}$ and $\hat{\pi}$ for different methods: DDR, Oracle, Super Oracle. The chart compares linear, quadratic, and SIM models.]
Coverage probability of the DDR estimator:

**DGP: Linear-Linear.**
Coverage probability of the DDR estimator:

**DGP: Linear-Linear.**

1. **When \( p = 50 \):**

<table>
<thead>
<tr>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Coverage Probability</td>
<td>0.943 (0.011)</td>
<td>0.943 (0.012)</td>
<td>0.942 (0.012)</td>
<td>0.163 (0)</td>
</tr>
<tr>
<td>Average Length</td>
<td>0.163 (0)</td>
<td>0.164 (0)</td>
<td>0.163 (0)</td>
<td></td>
</tr>
</tbody>
</table>

2. **When \( p = 500 \):**

<table>
<thead>
<tr>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Coverage Probability</td>
<td>0.941 (0.025)</td>
<td>0.942 (0.027)</td>
<td>0.942 (0.025)</td>
<td>0.164 (0.001)</td>
</tr>
<tr>
<td>Average Length</td>
<td>0.164 (0.001)</td>
<td>0.168 (0.001)</td>
<td>0.162 (0.001)</td>
<td></td>
</tr>
</tbody>
</table>
Coverage probability of the DDR estimator:

\textbf{DGP: Quad-Quad.}
Coverage probability of the DDR estimator:

**DGP: Quad-Quad.**

1. **When \( p = 50 \):**

<table>
<thead>
<tr>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
<th>( \hat{\pi} ): SIM</th>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
<th>( \hat{\pi} ): SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m ): linear</td>
<td>0.916 ( 0.137 )</td>
<td>0.944 ( 0.009 )</td>
<td>0.919 ( 0.126 )</td>
<td>0.419 ( 0.039 )</td>
<td>0.344 ( 0.028 )</td>
</tr>
<tr>
<td>( m ): quad</td>
<td>0.923 ( 0.098 )</td>
<td>0.944 ( 0.009 )</td>
<td>0.926 ( 0.084 )</td>
<td>0.422 ( 0.036 )</td>
<td>0.343 ( 0.029 )</td>
</tr>
<tr>
<td><strong>Average Coverage Probability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Length</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. **When \( p = 500 \):**

<table>
<thead>
<tr>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
<th>( \hat{\pi} ): SIM</th>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
<th>( \hat{\pi} ): SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m ): linear</td>
<td>0.942 ( 0.024 )</td>
<td>0.944 ( 0.024 )</td>
<td>0.943 ( 0.024 )</td>
<td>0.44 ( 0.015 )</td>
<td>0.334 ( 0.01 )</td>
</tr>
<tr>
<td>( m ): quad</td>
<td>0.942 ( 0.025 )</td>
<td>0.943 ( 0.024 )</td>
<td>0.943 ( 0.024 )</td>
<td>0.434 ( 0.014 )</td>
<td>0.333 ( 0.01 )</td>
</tr>
<tr>
<td><strong>Average Coverage Probability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Length</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Coverage probability of the DDR estimator:

**DGP: SIM-SIM.**
Coverage probability of the DDR estimator:

**DGP: SIM-SIM.**

1. **When \( p = 50 \):**

<table>
<thead>
<tr>
<th></th>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi ): logit</td>
<td>0.944 (0.01)</td>
<td>0.944 (0.01)</td>
<td>0.945 (0.01)</td>
<td>0.468 (0.032)</td>
<td>0.467 (0.032)</td>
<td>0.478 (0.033)</td>
</tr>
<tr>
<td>( \pi ): quad</td>
<td>0.946 (0.01)</td>
<td>0.946 (0.01)</td>
<td>0.946 (0.011)</td>
<td>0.465 (0.031)</td>
<td>0.464 (0.031)</td>
<td>0.475 (0.032)</td>
</tr>
</tbody>
</table>

2. **When \( p = 500 \):**

<table>
<thead>
<tr>
<th></th>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi ): logit</td>
<td>0.941 (0.031)</td>
<td>0.943 (0.03)</td>
<td>0.949 (0.023)</td>
<td>0.534 (0.012)</td>
<td>0.538 (0.011)</td>
<td>0.504 (0.011)</td>
</tr>
<tr>
<td>( \pi ): quad</td>
<td>0.941 (0.033)</td>
<td>0.942 (0.031)</td>
<td>0.949 (0.023)</td>
<td>0.535 (0.012)</td>
<td>0.539 (0.011)</td>
<td>0.504 (0.011)</td>
</tr>
</tbody>
</table>
Consider $n = 50000$ and $p = 50$. In addition, also consider the complete case estimator $\hat{\theta}_{cc}$, obtained by using only the data with $T_i = 1$.

DGP: Linear-Linear
Consider \( n = 50000 \) and \( p = 50 \). In addition, also consider the complete case estimator \( \hat{\theta}_{cc} \), obtained by using only the data with \( T_i = 1 \).

**DGP: Linear-Linear**

**\( L_2 \) Error Comparison:**

<table>
<thead>
<tr>
<th>model</th>
<th>( \hat{\pi} ): logit</th>
<th>( \hat{\pi} ): quad</th>
<th>( \hat{\theta}_{DDR} )</th>
<th>( \hat{\theta}_{orac} )</th>
<th>( \theta_{full} )</th>
<th>( \hat{\theta}_{cc} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{m} ): linear</td>
<td>0.034 ( 0.005 )</td>
<td>0.033 ( 0.005 )</td>
<td>0.034 ( 0.005 )</td>
<td>0.034 ( 0.005 )</td>
<td>0.026 ( 0.004 )</td>
<td>0.032 ( 0.005 )</td>
</tr>
<tr>
<td>( \hat{m} ): quad</td>
<td>0.034 ( 0.005 )</td>
<td>0.033 ( 0.005 )</td>
<td>0.034 ( 0.005 )</td>
<td>0.033 ( 0.005 )</td>
<td>0.026 ( 0.004 )</td>
<td>0.032 ( 0.005 )</td>
</tr>
<tr>
<td>( \hat{m} ): SIM</td>
<td>0.034 ( 0.005 )</td>
<td>0.033 ( 0.005 )</td>
<td>0.034 ( 0.005 )</td>
<td>0.034 ( 0.005 )</td>
<td>0.026 ( 0.004 )</td>
<td>0.032 ( 0.005 )</td>
</tr>
</tbody>
</table>

**Inference:**

<table>
<thead>
<tr>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
<th>( \hat{m} ): linear</th>
<th>( \hat{m} ): quad</th>
<th>( \hat{m} ): SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\pi} ): logit</td>
<td>0.944 ( 0.012 )</td>
<td>0.944 ( 0.012 )</td>
<td>0.945 ( 0.015 )</td>
<td>0.023 ( 0 )</td>
<td>0.023 ( 0 )</td>
</tr>
<tr>
<td>( \hat{\pi} ): quad</td>
<td>0.944 ( 0.012 )</td>
<td>0.944 ( 0.012 )</td>
<td>0.944 ( 0.018 )</td>
<td>0.023 ( 0 )</td>
<td>0.023 ( 0 )</td>
</tr>
</tbody>
</table>
DGP: Quad-Quad
DGP: Quad-Quad

$L_2$ Error Comparison:

<table>
<thead>
<tr>
<th>model</th>
<th>$\hat{\theta}_{\text{DDR}}$</th>
<th>$\hat{\theta}_{\text{orac}}$</th>
<th>$\hat{\theta}_{\text{full}}$</th>
<th>$\hat{\theta}_{\text{cc}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{m}$: linear</td>
<td>$\hat{\pi}$: logit 0.46 ( 0.026 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
<tr>
<td></td>
<td>$\hat{\pi}$: quad 0.204 ( 0.137 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
<tr>
<td>$\hat{m}$: quad</td>
<td>$\hat{\pi}$: logit 0.071 ( 0.01 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
<tr>
<td></td>
<td>$\hat{\pi}$: quad 0.072 ( 0.011 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
<tr>
<td>$\hat{m}$: SIM</td>
<td>$\hat{\pi}$: logit 0.322 ( 0.019 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
<tr>
<td></td>
<td>$\hat{\pi}$: quad 0.172 ( 0.078 )</td>
<td>0.072 ( 0.011 )</td>
<td>0.069 ( 0.01 )</td>
<td>0.528 ( 0.021 )</td>
</tr>
</tbody>
</table>

Inference:

<table>
<thead>
<tr>
<th>$\hat{m}$: linear</th>
<th>$\hat{m}$: quad</th>
<th>$\hat{m}$: SIM</th>
<th>$\hat{m}$: linear</th>
<th>$\hat{m}$: quad</th>
<th>$\hat{m}$: SIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\pi}$: logit 0.889 ( 0.201 )</td>
<td>0.937 ( 0.031 )</td>
<td>0.912 ( 0.102 )</td>
<td>0.064 ( 0.008 )</td>
<td>0.048 ( 0.004 )</td>
<td>0.067 ( 0.007 )</td>
</tr>
<tr>
<td>$\hat{\pi}$: quad 0.959 ( 0.021 )</td>
<td>0.94 ( 0.03 )</td>
<td>0.951 ( 0.022 )</td>
<td>0.12 ( 0.037 )</td>
<td>0.049 ( 0.005 )</td>
<td>0.098 ( 0.025 )</td>
</tr>
</tbody>
</table>


Thank You!