

# Factor Informed Double Deep Learning Estimation For Average Treatment Effect

**Jianqing Fan**

**Princeton** University

<https://fan.princeton.edu/>

with **Soham Jana**, **Sanjeev Kulkarni** and **Qishuo Yin**



# Outlines

- 1 Introduction
- 2 FAST-NN (FAN-Lasso)
- 3 FIDDLE
- 4 Theoretical Guarantee
- 5 Numerical Results
- 6 Conclusion Remarks



Soham Jana



Sanjeev Kulkarni



Qishuo Yin

# Outlines

- 1 Introduction
- 2 FAST-NN (FAN-Lasso)
- 3 FIDDLE
- 4 Theoretical Guarantee
- 5 Numerical Results
- 6 Conclusion Remarks



Soham Jana



Sanjeev Kulkarni



Qishuo Yin

# Outlines

- 1 Introduction
- 2 FAST-NN (FAN-Lasso)
- 3 FIDDLE
- 4 Theoretical Guarantee
- 5 Numerical Results
- 6 Conclusion Remarks



Soham Jana



Sanjeev Kulkarni

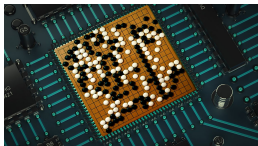


Qishuo Yin

# Introduction

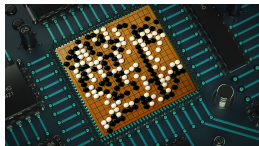
# Deep Learning

- **Games**: Atari, Go, Starcraft, Texas hold'em, ...
- **Control**: robotic hand, automatic driving, fleet management, ...
- **Prediction**: peotein folding, LLM, generative AI ...



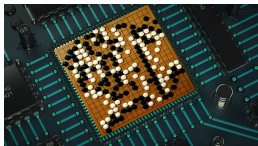
# Deep Learning

- **Games**: Atari, Go, Starcraft, Texas hold'em, ...
- **Control**: robotic hand, automatic driving, fleet management, ...
- **Prediction**: protein folding, LLM, generative AI ...



# Deep Learning

- **Games**: Atari, Go, Starcraft, Texas hold'em, ...
- **Control**: robotic hand, automatic driving, fleet management, ...
- **Prediction**: protein folding, LLM, generative AI ..





- ★ Nonparametric regression, density estimation (ImageGen, DALL-E)
- ★ Semiparametric, GLIM, Quantile regression, Survival Analysis
- ★ Transfer and generative learning
- ★ Causality Pursuit, Reinforcement Learning, and Fine Tuning

Adapt to unknown low-d structure

Causal Inference

- ★ Nonparametric regression, density estimation (ImageGen, DALL-E)
- ★ Semiparametric, GLIM, Quantile regression, Survival Analysis
- ★ Transfer and generative learning
- ★ Causality Pursuit, Reinforcement Learning, and Fine Tuning

Adapt to unknown low-d structure

Causal Inference

- ★ Nonparametric regression, density estimation (ImageGen, DALL-E)
- ★ Semiparametric, GLIM, Quantile regression, Survival Analysis
- ★ Transfer and generative learning
- ★ Causality Pursuit, Reinforcement Learning, and Fine Tuning

## Adapt to unknown low-d structure

## Causal Inference

- ★ Nonparametric regression, density estimation (ImageGen, DALL-E)
- ★ Semiparametric, GLIM, Quantile regression, Survival Analysis
- ★ Transfer and generative learning
- ★ Causality Pursuit, Reinforcement Learning, and Fine Tuning

**Adapt to unknown low-d structure**

**Causal Inference**

# Does aspirin relieves headaches?

- ★ **Task**: study whether taking aspirin will help relieve headaches;
- ★ **Unit**: the person;
- ★ **Treatment T**: Aspirin / no Aspirin;
- ★ **Outcome y**: headache / no headache;
- ★ **Causal (Treatment) Effect**: improvement of headache relief by Aspirin;
- ★ **Covariate x**: pretreatment information of the person.

# Does aspirin relieves headaches?

- ★ **Task**: study whether taking aspirin will help relieve headaches;
- ★ **Unit**: the person;
- ★ **Treatment T**: Aspirin / no Aspirin;
- ★ **Outcome y**: headache / no headache;
- ★ **Causal (Treatment) Effect**: improvement of headache relief by Aspirin;
- ★ **Covariate x**: pretreatment information of the person.

# Does aspirin relieves headaches?

- ★ **Task**: study whether taking aspirin will help relieve headaches;
- ★ **Unit**: the person;
- ★ **Treatment T**: Aspirin / no Aspirin;
- ★ **Outcome y**: headache / no headache;
- ★ **Causal (Treatment) Effect**: improvement of headache relief by Aspirin;
- ★ **Covariate x**: pretreatment information of the person.

# Potential outcome framework

Data  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} (y, T, \mathbf{x}),$

Rubin (1974)

★ Observed covariates:  $\mathbf{x} \in \mathbb{R}^p;$

★ Treatment:  $T = \begin{cases} 1, & \text{if treated} \\ 0, & \text{if control} \end{cases}$

★ Outcomes:  $y = \begin{cases} y(1), & \text{if } T = 1 \\ y(0), & \text{if } T = 0 \end{cases}$



# Potential outcome framework

Data  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} (y, T, \mathbf{x}),$

Rubin (1974)

★ Observed covariates:  $\mathbf{x} \in \mathbb{R}^p;$

★ Treatment:  $T = \begin{cases} 1, & \text{if treated} \\ 0, & \text{if control} \end{cases}$

★ Outcomes:  $y = \begin{cases} y(1), & \text{if } T = 1 \\ y(0), & \text{if } T = 0 \end{cases}$

# Potential outcome framework

Data  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} (y, T, \mathbf{x}),$

Rubin (1974)

★ Observed covariates:  $\mathbf{x} \in \mathbb{R}^p;$

★ Treatment:  $T = \begin{cases} 1, & \text{if treated} \\ 0, & \text{if control} \end{cases}$

★ Outcomes:  $y = \begin{cases} y(1), & \text{if } T = 1 \\ y(0), & \text{if } T = 0 \end{cases}$

# Average treatment effect Estimation

## ★ Propensity score:

$$\pi^*(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{P}(T = 1 | \mathbf{X} = \mathbf{x})$$

## ★ Outcome model:

$$y(t) \stackrel{\text{def}}{=} \mu_t^*(\mathbf{x}) + \varepsilon, \quad t \in \{0, 1\}$$

## ★ Average treatment effect (ATE):

$$\tau \stackrel{\text{def}}{=} \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x})]$$

Conditional ATE  $\tau(\mathbf{x}) = \mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x})$

# Average treatment effect Estimation

## ★ Propensity score:

$$\pi^*(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{P}(T = 1 | \mathbf{X} = \mathbf{x})$$

## ★ Outcome model:

$$y(t) \stackrel{\text{def}}{=} \mu_t^*(\mathbf{x}) + \varepsilon, \quad t \in \{0, 1\}$$

## ★ Average treatment effect (ATE):

$$\tau \stackrel{\text{def}}{=} \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x})]$$

Conditional ATE  $\tau(\mathbf{x}) = \mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x})$

**How to model propensity and outcome functions?**

Which variables affect propensity and outcome?

How to address dependence in high-d  $x$ ?

**How to model propensity and outcome functions?**

**Which variables affect propensity and outcome?**

How to address dependence in high-d  $x$ ?

**How to model propensity and outcome functions?**

**Which variables affect propensity and outcome?**

**How to address dependence in high-d  $x$ ?**

# About this work

## 1 Address dependence in high-d covariate by

- ★ Factor model  $\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$ ,
- learned via diversified projection

## 2 Model propensity and outcome functions by

- ★ Factor Augmented Sparse Throughput Neural network, FAST-NN model
- Select variables in **adaptive**, **nonparametric** and **algorithmic** manner

## 3 Ensure robust ATE estimator by

- ★ Augmented Inverse Propensity Weighted (AIPW) estimator
- Easily get semiparametric **efficiency**





# About this work

- 1 Address dependence in high-d covariate by

- ★ Factor model  $\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$ ,

- learned via diversified projection

- 2 Model propensity and outcome functions by

- ★ Factor Augmented Sparse Throughput Neural network, FAST-NN model

- Select variables in **adaptive**, **nonparametric** and **algorithmic** manner

- 3 Ensure robust ATE estimator by

- ★ Augmented Inverse Propensity Weighted (AIPW) estimator

- Easily get semiparametric **efficiency**



# About this work

- 1 Address dependence in high-d covariate by

- ★ Factor model  $\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$ ,

- learned via diversified projection

- 2 Model propensity and outcome functions by

- ★ Factor Augmented Sparse Throughput Neural network, FAST-NN model

- Select variables in **adaptive**, **nonparametric** and **algorithmic** manner

- 3 Ensure robust ATE estimator by

- ★ Augmented Inverse Propensity Weighted (AIPW) estimator

- Easily get semiparametric **efficiency**



# ATE estimation via factor decomposition

★ Model: For subsets  $\mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_2$  subset of covariates

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$$

$$\mathbb{E}[T | \mathbf{f}, \mathbf{u}] = \pi^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_2}) = \tilde{\pi}^*(\mathbf{f}, \mathbf{x}_{\mathcal{J}_2})$$

$$\mathbb{E}[y(t) | \mathbf{f}, \mathbf{u}] = \mu_t^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_t}), t \in \{0, 1\}$$

★ Data:  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} (y, T, \mathbf{x})$

★ Goal: estimate ATE

$$\tau = \mathbb{E} \{ \mathbb{E}[y(1) - y(0) | \mathbf{f}, \mathbf{u}] \} = \mathbb{E}[\mu_1^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_1}) - \mu_0^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_0})]$$

# ATE estimation via factor decomposition

★ Model: For subsets  $\mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_2$  subset of covariates

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$$

$$\mathbb{E}[T | \mathbf{f}, \mathbf{u}] = \pi^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_2}) = \tilde{\pi}^*(\mathbf{f}, \mathbf{x}_{\mathcal{J}_2})$$

$$\mathbb{E}[y(t) | \mathbf{f}, \mathbf{u}] = \mu_t^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_t}), t \in \{0, 1\}$$

★ Data:  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} (y, T, \mathbf{x})$

★ Goal: estimate ATE

$$\tau = \mathbb{E} \{ \mathbb{E}[y(1) - y(0) | \mathbf{f}, \mathbf{u}] \} = \mathbb{E}[\mu_1^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_1}) - \mu_0^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_0})]$$

# ATE estimation via factor decomposition

★ Model: For subsets  $\mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_2$  subset of covariates

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}$$

$$\mathbb{E}[T | \mathbf{f}, \mathbf{u}] = \pi^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_2}) = \tilde{\pi}^*(\mathbf{f}, \mathbf{x}_{\mathcal{J}_2})$$

$$\mathbb{E}[y(t) | \mathbf{f}, \mathbf{u}] = \mu_t^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_t}), t \in \{0, 1\}$$

★ Data:  $\{(y_i, T_i, \mathbf{x}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} (y, T, \mathbf{x})$

★ Goal: estimate ATE

$$\tau = \mathbb{E}\{\mathbb{E}[y(1) - y(0) | \mathbf{f}, \mathbf{u}]\} = \mathbb{E}[\mu_1^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_1}) - \mu_0^*(\mathbf{f}, \mathbf{u}_{\mathcal{J}_0})]$$

# Related works

- ★ **Factor model**: Forni et al. (2000); Stock & Watson (2002); Bai (2003); Bai & Ng (2006, 2008); Hallin & Liska, (2007); Fan et al. (2013, 2021, 2022); Fan & Liao (2022));
- ★ **Sparse additive models**: Ravikumar et al. (2009); Meier et al. (2009); Raskutti et al. (2012), Yuan & Zhou (2016);
- ★ **Neural network models**: Feng & Simon (2017), Chen et al. (2019); Gu et al. (2021); Schmidt-Hieber (2020); Kohler and Langer (2021); Shen et al. (2021), Fan, Gu, Zhou (2022), Ohn & Kim (2022), Fan et al. (2022, ; Fan & Gu (2024);
- ★ **ATE estimation**: Rosenbaum and Rubin (1983); Rosenbaum (1989); Abadie & Imbens (2006); Wager & Athey (2018); Chernozhukov et al. (2018); Yoon et al. (2018); Bang & Robins (2005); Farrell (2015), Farrell et al. (2021)

# Related works

- ★ **Factor model**: Forni et al. (2000); Stock & Watson (2002); Bai (2003); Bai & Ng (2006, 2008); Hallin & Liska, (2007); Fan et al. (2013, 2021, 2022); Fan & Liao (2022));
- ★ **Sparse additive models**: Ravikumar et al. (2009); Meier et al. (2009); Raskutti et al. (2012), Yuan & Zhou (2016);
- ★ **Neural network models**: Feng & Simon (2017), Chen et al. (2019); Gu et al. (2021); Schmidt-Hieber (2020); Kohler and Langer (2021); Shen et al. (2021), Fan, Gu, Zhou (2022), Ohn & Kim (2022), Fan et al. (2022, ; Fan & Gu (2024);
- ★ **ATE estimation**: Rosenbaum and Rubin (1983); Rosenbaum (1989); Abadie & Imbens (2006); Wager & Athey (2018); Chernozhukov et al. (2018); Yoon et al. (2018); Bang & Robins (2005); Farrell (2015), Farrell et al. (2021)

# Related works

- ★ **Factor model**: Forni et al. (2000); Stock & Watson (2002); Bai (2003); Bai & Ng (2006, 2008); Hallin & Liska, (2007); Fan et al. (2013, 2021, 2022); Fan & Liao (2022));
- ★ **Sparse additive models**: Ravikumar et al. (2009); Meier et al. (2009); Raskutti et al. (2012), Yuan & Zhou (2016);
- ★ **Neural network models**: Feng & Simon (2017), Chen et al. (2019); Gu et al. (2021); Schmidt-Hieber (2020); Kohler and Langer (2021); Shen et al. (2021), Fan, Gu, Zhou (2022), Ohn & Kim (2022), Fan et al. (2022, ; Fan & Gu (2024);
- ★ **ATE estimation**: Rosenbaum and Rubin (1983); Rosenbaum (1989); Abadie & Imbens (2006); Wager & Athey (2018); Chernozhukov et al. (2018); Yoon et al. (2018); Bang & Robins (2005); Farrell (2015), Farrell et al. (2021)



# Related works

- ★ [Factor model](#): Forni et al. (2000); Stock & Watson (2002); Bai (2003); Bai & Ng (2006, 2008); Hallin & Liska, (2007); Fan et al. (2013, 2021, 2022); Fan & Liao (2022));
- ★ [Sparse additive models](#): Ravikumar et al. (2009); Meier et al. (2009); Raskutti et al. (2012), Yuan & Zhou (2016);
- ★ [Neural network models](#): Feng & Simon (2017), Chen et al. (2019); Gu et al. (2021); Schmidt-Hieber (2020); Kohler and Langer (2021); Shen et al. (2021), Fan, Gu, Zhou (2022), Ohn & Kim (2022), Fan et al. (2022, ; Fan & Gu (2024);
- ★ [ATE estimation](#): Rosenbaum and Rubin (1983); Rosenbaum (1989); Abadie & Imbens (2006); Wager & Athey (2018); Chernozhukov et al. (2018); Yoon et al. (2018); Bang & Robins (2005); Farrell (2015), Farrell et al. (2021)

# FAST-NN

## Nonparametric Lasso

★ Fan, J. and Gu, Y. (2024). Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression. *JASA*, **119** (548), 2680-2694.

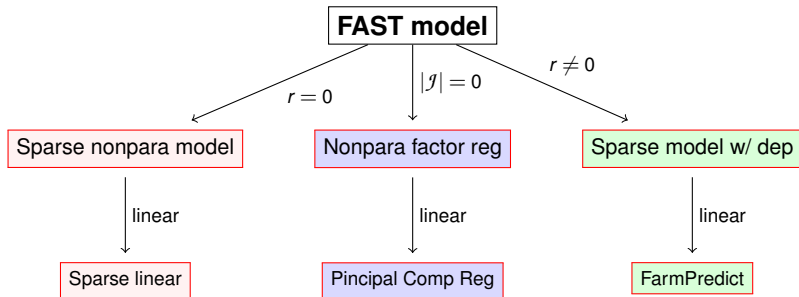
# FAST-NN

## ★ Factor-Adjusted Nonparametric Lasso (FAN-Lasso)

★ Fan, J. and Gu, Y. (2024). Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression. *JASA*, **119** (548), 2680-2694.

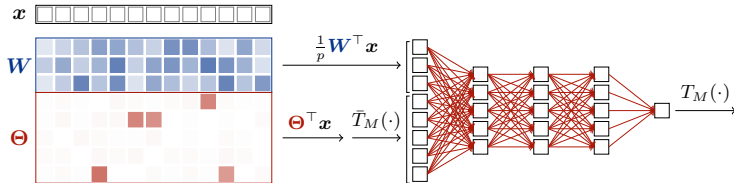
# Versality of FAST-model

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}, \quad \mathbb{E}[y|\mathbf{f}, \mathbf{u}] = m^*(\mathbf{f}, \mathbf{u}_g) = g^*(\mathbf{f}, \mathbf{x}_g),$$



# Archtechure of FAST-model

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \mathbf{u}, \quad \mathbb{E}[y|\mathbf{f}, \mathbf{u}] = m^*(\mathbf{f}, \mathbf{u}_j) = g^*(\mathbf{f}, \mathbf{x}_j),$$



## Joint training vs Pre-training?

# Diversified projections for factor learning

Challenges: Learning **simultaneously** factors and reg function via DNN overfit.

Factor learning: For a given  $\mathbf{W} \in \mathbb{R}^{p \times \bar{r}}$  ( $\bar{r} \geq r$ , Fan & Liao, 22),

+ no need  $n \rightarrow \infty$

$$\tilde{\mathbf{f}} = p^{-1} \mathbf{W}^\top \mathbf{x} = \underbrace{p^{-1} \mathbf{W}^\top \mathbf{B}}_{\mathbf{H}} + \underbrace{p^{-1} \mathbf{W}^\top \mathbf{u}}_{O_p(p^{-1/2})}$$

Requirements:  $\lambda_{\min}(\mathbf{H}) \gg p^{-1/2}$ ,  $\|\mathbf{W}\|_\infty \leq C$

Examples:  $\blacklozenge \mathbf{W}^{(1)} = \mathbf{B}$        $\blacklozenge \mathbf{W}^{(2)} = [\mathbf{B}, \mathbf{x}]$

pre-training:  $\mathbf{W}$  = top  $\bar{r}$  PCs using  $n_1$  data

Thm:  $n_1 \gtrsim r^2 \log p$  suffices.

# Diversified projections for factor learning

Challenges: Learning **simultaneously** factors and reg function via DNN overfit.

Factor learning: For a given  $\mathbf{W} \in \mathbb{R}^{p \times \bar{r}}$  ( $\bar{r} \geq r$ , Fan & Liao, 22),

+ no need  $n \rightarrow \infty$

$$\tilde{\mathbf{f}} = p^{-1} \mathbf{W}^\top \mathbf{x} = \underbrace{p^{-1} \mathbf{W}^\top \mathbf{B}}_{\mathbf{H}} + \underbrace{p^{-1} \mathbf{W}^\top \mathbf{u}}_{O_p(p^{-1/2})}$$

Requirements:  $v_{\min}(\mathbf{H}) \gg p^{-1/2}$ ,  $\|\mathbf{W}\|_\infty \leq C$

Examples:  $\diamond W^{(1)} = \mathbf{B}$   $\diamond W^{(2)} = [\mathbf{B}, \times]$

pre-training:  $\mathbf{W}$  = top  $\bar{r}$  PCs using  $n_1$  data

Thm:  $n_1 \gtrsim r^2 \log p$  suffices.

# Diversified projections for factor learning

Challenges: Learning **simultaneously** factors and reg function via DNN overfit.

Factor learning: For a given  $\mathbf{W} \in \mathbb{R}^{p \times \bar{r}}$  ( $\bar{r} \geq r$ , Fan & Liao, 22),

+ no need  $n \rightarrow \infty$

$$\tilde{\mathbf{f}} = p^{-1} \mathbf{W}^\top \mathbf{x} = \underbrace{p^{-1} \mathbf{W}^\top \mathbf{B}}_{\mathbf{H}} + \underbrace{p^{-1} \mathbf{W}^\top \mathbf{u}}_{O_p(p^{-1/2})}$$

Requirements:  $\lambda_{\min}(\mathbf{H}) \gg p^{-1/2}$ ,  $\|\mathbf{W}\|_\infty \leq C$

Examples:  $\blacklozenge \mathbf{W}^{(1)} = \mathbf{B}$   $\blacklozenge \mathbf{W}^{(2)} = [\mathbf{B}, \times]$

pre-training:  $\mathbf{W}$  = top  $\bar{r}$  PCs using  $n_1$  data

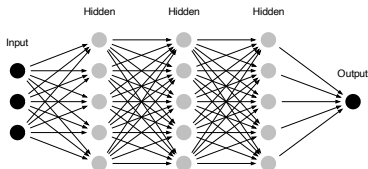
Thm:  $n_1 \gtrsim r^2 \log p$  suffices .



# Deep ReLU neural network

## Deep ReLU neural network with width $N$ and depth $L$

$$f: \mathbb{R}^d \rightarrow \mathbb{R}, \quad f(\mathbf{x}; \theta) = \mathcal{L}_{L+1} \circ \sigma \circ \mathcal{L}_L \circ \cdots \circ \sigma \circ \mathcal{L}_2 \circ \sigma \circ \mathcal{L}_1(\mathbf{x})$$



- $\mathcal{L}_i(\mathbf{x}) = \mathbf{W}_i \mathbf{x} + \mathbf{b}_i$  with weights  $\mathbf{W}_i, \mathbf{b}_i$ ,  $\sigma(t) = t_+ \text{ ReLU activation}$
- Dim:  $(d_0, d_1, \dots, d_L, d_{L+1}) = (d, N, \dots, N, 1)$
- **Parameter**  $\theta = \{\mathbf{W}_i, \mathbf{b}_i\}_{i=1}^L$  **# parameter**  $\asymp N^2 L$

**Bounded ReLU-DNN functions** :  $\mathcal{G}_n = \{T_M g(\mathbf{x}) : \sup_{\ell \in [L+1]} \|\mathbf{W}_\ell\|_{\max} \vee \|\mathbf{b}_\ell\|_\infty \leq B\}$

# Estimation for FAST model

★ Let diversified projection matrix  $\mathbf{W}$ , variable selection matrix  $\Theta \in \mathbb{R}^{p \times N}$ .

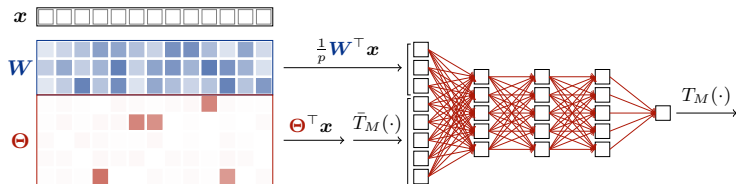
★ With depth  $L$  and width  $N$  ReLU  $\mathbf{g}_{NN}$ , output the model

$$m(\mathbf{x}) = \mathbf{g}_{NN} \left( \left[ p^{-1} \mathbf{W}^\top \mathbf{x}, \bar{T}_M(\Theta^\top \mathbf{x}) \right] \right)$$

★ Run penalized least squares with parameters  $\lambda$  and  $\tau$ :

FAST-NN

$$\hat{\mathbf{R}}_{\text{FAST}}(\Theta, \mathbf{g}_{NN}) = \frac{1}{n} \sum_{i=1}^n \{y_i - m(\mathbf{x}_i)\}^2 + \lambda \sum_{j,k} |\Theta_{j,k}| \wedge \tau$$



# Estimation for FAST model

★ Let diversified projection matrix  $\mathbf{W}$ , variable selection matrix  $\Theta \in \mathbb{R}^{p \times N}$ .

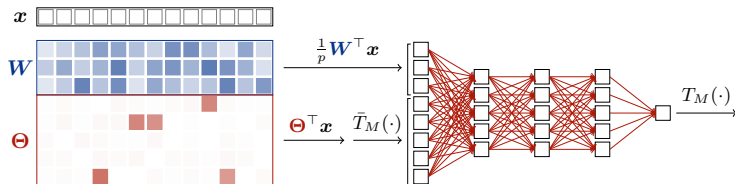
★ With depth  $L$  and width  $N$  ReLU  $\mathbf{g}_{NN}$ , output the model

$$m(\mathbf{x}) = \mathbf{g}_{NN} \left( \left[ p^{-1} \mathbf{W}^\top \mathbf{x}, \bar{T}_M(\Theta^\top \mathbf{x}) \right] \right)$$

★ Run penalized least squares with parameters  $\lambda$  and  $\tau$ :

FAST-NN

$$\hat{\mathbf{R}}_{\text{FAST}}(\Theta, \mathbf{g}_{NN}) = \frac{1}{n} \sum_{i=1}^n \{y_i - m(\mathbf{x}_i)\}^2 + \lambda \sum_{j,k} |\Theta_{j,k}| \wedge \tau$$



# Estimation for FAST model

★ Let diversified projection matrix  $\mathbf{W}$ , variable selection matrix  $\Theta \in \mathbb{R}^{p \times N}$ .

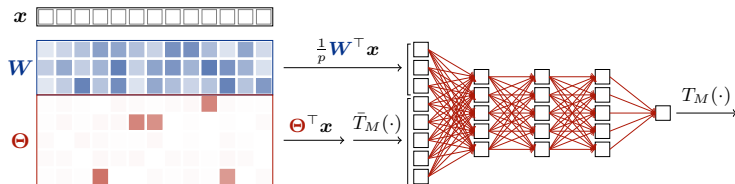
★ With depth  $L$  and width  $N$  ReLU  $\mathbf{g}_{NN}$ , output the model

$$m(\mathbf{x}) = \mathbf{g}_{NN} \left( \left[ p^{-1} \mathbf{W}^\top \mathbf{x}, \bar{T}_M(\Theta^\top \mathbf{x}) \right] \right)$$

★ Run penalized least squares with parameters  $\lambda$  and  $\tau$ :

FAST-NN

$$\hat{\mathbf{R}}_{\text{FAST}}(\Theta, \mathbf{g}_{NN}) = \frac{1}{n} \sum_{i=1}^n \{y_i - m(\mathbf{x}_i)\}^2 + \lambda \sum_{j,k} |\Theta_{j,k}| \wedge \tau$$



# Hierachial Composition Model

**Hierachial Composition Model:**  $\text{HCM}(\mathcal{P}) = \text{Finite compositions}$

$Q^* = Q_1 \circ \dots \circ Q_q$  of  $t$ -variate function with smoothness  $\beta$ , for  $(t, \beta) \in \mathcal{P}$ .

—  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$  is the hardest component.

(Kohler and Krzyzak, 2017)

★ e.g.  $Q^* = f_1(x_1) + \dots + f_p(x_p)$        $Q^* = f_1(x_1, f_2(x_2, x_3)) + f_3(x_2, x_5, x_9)$

★ Complexity determined by  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$ .

★ Adaptively learned by neural networks

# Hierachial Composition Model

**Hierachial Composition Model:**  $\text{HCM}(\mathcal{P}) = \text{Finite compositions}$

$Q^* = Q_1 \circ \dots \circ Q_q$  of  $t$ -variate function with smoothness  $\beta$ , for  $(t, \beta) \in \mathcal{P}$ .

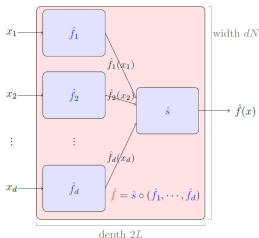
—  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$  is the hardest component.

(Kohler and Krzyzak, 2017)

★ e.g.  $Q^* = f_1(x_1) + \dots + f_p(x_p)$        $Q^* = f_1(x_1, f_2(x_2, x_3)) + f_3(x_2, x_5, x_9)$

★ Complexity determined by  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$ .

★ Adaptively learned by neural networks



# Hierachial Composition Model

**Hierachial Composition Model:**  $\text{HCM}(\mathcal{P}) = \text{Finite compositions}$

$Q^* = Q_1 \circ \dots \circ Q_q$  of  $t$ -variate function with smoothness  $\beta$ , for  $(t, \beta) \in \mathcal{P}$ .

—  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$  is the hardest component. (Kohler and Krzyzak, 2017)

★ e.g.  $Q^* = f_1(x_1) + \dots + f_p(x_p)$        $Q^* = f_1(x_1, f_2(x_2, x_3)) + f_3(x_2, x_5, x_9)$

★ Complexity determined by  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$ .

★ Adaptively learned by neural networks

# Hierachial Composition Model

**Hierachial Composition Model:**  $\text{HCM}(\mathcal{P}) = \text{Finite compositions}$

$Q^* = Q_1 \circ \dots \circ Q_q$  of  $t$ -variate function with smoothness  $\beta$ , for  $(t, \beta) \in \mathcal{P}$ .

—  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$  is the hardest component. (Kohler and Krzyzak, 2017)

★ e.g.  $Q^* = f_1(x_1) + \dots + f_p(x_p)$        $Q^* = f_1(x_1, f_2(x_2, x_3)) + f_3(x_2, x_5, x_9)$

★ Complexity determined by  $\gamma^* = \min_{(t, \beta) \in \mathcal{P}} \frac{\beta}{t}$ .

★ Adaptively learned by neural networks



# Factor Informed Double Deep Learning Estimator (FIDDLE)

# FIDDLE for ATE estimation

★ Outcome models  $\mu_t^*$  and propensity  $\pi^*$  estimated by FAST-NN:

$$\hat{\mu}_t(\mathbf{x}_i) = m_t^{\text{FAST}}(\mathbf{x}_i), \quad t = 0, 1$$

$$\hat{\pi}(\mathbf{x}_i) = m_2^{\text{FAST}}(\mathbf{x}_i)$$

★ FIDDLE: ATE estimation via AIPW (*Robins et al, 1994; Bang & Robins, 2005*)

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \left[ \hat{\mu}_1(\mathbf{x}_i) + \frac{T_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{\pi}(\mathbf{x}_i)} - \hat{\mu}_0(\mathbf{x}_i) - \frac{(1 - T_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1 - \hat{\pi}(\mathbf{x}_i)} \right]$$

# FIDDLE for ATE estimation

- ★ Outcome models  $\mu_t^*$  and propensity  $\pi^*$  estimated by FAST-NN:

$$\hat{\mu}_t(\mathbf{x}_i) = m_t^{\text{FAST}}(\mathbf{x}_i), \quad t = 0, 1$$

$$\hat{\pi}(\mathbf{x}_i) = m_2^{\text{FAST}}(\mathbf{x}_i)$$

- ★ FIDDLE: ATE estimation via AIPW (*Robins et al, 1994; Bang & Robins, 2005*)

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \left[ \hat{\mu}_1(\mathbf{x}_i) + \frac{T_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{\pi}(\mathbf{x}_i)} - \hat{\mu}_0(\mathbf{x}_i) - \frac{(1 - T_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1 - \hat{\pi}(\mathbf{x}_i)} \right]$$

# Augmented inverse propensity weighted estimator

★ Why Augmented Inverse Propensity Weighted (**AIPW**) Estimator?

$$\begin{aligned} \star \text{ 1}^{st}\text{-term} &\approx \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{T(Y - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] = \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{\pi^*(\mathbf{X})(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] \\ &= \mathbb{E} \left[ \mu_1^*(\mathbf{X}) + \frac{(\pi^*(\mathbf{X}) - \hat{\pi}(\mathbf{X}))(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] \quad \Longleftarrow \text{double-robustness} \end{aligned}$$

★ either propensity or outcome models correctly specified

$\implies$  consistent ATE estimation  $\implies$  error multiplied

# Augmented inverse propensity weighted estimator

★ Why Augmented Inverse Propensity Weighted (**AIPW**) Estimator?

★ 1<sup>st</sup>-term  $\approx \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{T(Y - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] = \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{\pi^*(\mathbf{X})(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right]$

$$= \mathbb{E} \left[ \mu_1^*(\mathbf{X}) + \frac{(\pi^*(\mathbf{X}) - \hat{\pi}(\mathbf{X}))(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] \quad \Longleftarrow \quad \text{double-robustness}$$

★ either propensity or outcome models correctly specified

$\implies$  consistent ATE estimation  $\implies$  error multiplied

# Augmented inverse propensity weighted estimator

★ Why Augmented Inverse Propensity Weighted (**AIPW**) Estimator?

$$\begin{aligned} \star \text{ 1}^{st}\text{-term} &\approx \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{T(Y - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] = \mathbb{E} \left[ \hat{\mu}_1(\mathbf{X}) + \frac{\pi^*(\mathbf{X})(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] \\ &= \mathbb{E} \left[ \mu_1^*(\mathbf{X}) + \frac{(\pi^*(\mathbf{X}) - \hat{\pi}(\mathbf{X}))(\mu_1^*(\mathbf{X}) - \hat{\mu}_1(\mathbf{X}))}{\hat{\pi}(\mathbf{X})} \right] \quad \Longleftarrow \text{double-robustness} \end{aligned}$$

★ **either** propensity **or** outcome models **correctly specified**

$\implies$  **consistent** ATE estimation  $\implies$  **error multiplied**

# Theory

# Technical assumptions

**Estimator:** Approximate ERM  $\hat{m}$  with  $\hat{R}(\hat{m}) \leq \inf_m \hat{R}(m) + \delta_{\text{opt}}$ .

★ Truncation level  $M$  is a large constant.

**Assump 1:**  $\|\mathbf{B}\|_{\max} \leq b$ ,  $\mathbf{f}$  and  $\mathbf{u}$  zero mean,  $\|\mathbf{f}\|_{\infty}, \|\mathbf{u}\|_{\infty} \leq b$ .

**Assump 2:** Weak dependence of  $\mathbf{u}$ :  $\sum_{j,k} |\mathbb{E}[u_j u_k]| \lesssim p$ .

**Assump 3:** Sub-Gaussian noise:  $\mathbb{P}[|\varepsilon| > t | \mathbf{f}, \mathbf{u}] \leq 2e^{-c_1 t^2}$  a.s.

**Assump 4:**  $\pi^*, \mu_0^*, \mu_1^*$  is  $C$ -Lipschitz and  $\|\pi^*\|_{\infty}, \|\mu_0^*\|_{\infty}, \|\mu_1^*\|_{\infty} \lesssim 1$ .



# Technical assumptions

**Estimator:** Approximate ERM  $\hat{m}$  with  $\hat{R}(\hat{m}) \leq \inf_m \hat{R}(m) + \delta_{\text{opt}}$ .

★ Truncation level  $M$  is a large constant.

**Assump 1:**  $\|\mathbf{B}\|_{\max} \leq b$ ,  $\mathbf{f}$  and  $\mathbf{u}$  zero mean,  $\|\mathbf{f}\|_{\infty}, \|\mathbf{u}\|_{\infty} \leq b$ .

**Assump 2:** Weak dependence of  $\mathbf{u}$ :  $\sum_{j,k} |\mathbb{E}[u_j u_k]| \lesssim p$ .

**Assump 3:** Sub-Gaussian noise:  $\mathbb{P}[|\varepsilon| > t | \mathbf{f}, \mathbf{u}] \leq 2e^{-c_1 t^2}$  a.s.

**Assump 4:**  $\pi^*, \mu_0^*, \mu_1^*$  is  $C$ -Lipschitz and  $\|\pi^*\|_{\infty}, \|\mu_0^*\|_{\infty}, \|\mu_1^*\|_{\infty} \lesssim 1$ .

# Technical assumptions

**Estimator:** Approximate ERM  $\hat{m}$  with  $\hat{R}(\hat{m}) \leq \inf_m \hat{R}(m) + \delta_{\text{opt}}$ .

★ Truncation level  $M$  is a large constant.

**Assump 1:**  $\|\mathbf{B}\|_{\max} \leq b$ ,  $\mathbf{f}$  and  $\mathbf{u}$  zero mean,  $\|\mathbf{f}\|_{\infty}, \|\mathbf{u}\|_{\infty} \leq b$ .

**Assump 2:** Weak dependence of  $\mathbf{u}$ :  $\sum_{j,k} |\mathbb{E}[u_j u_k]| \lesssim p$ .

**Assump 3:** Sub-Gaussian noise:  $\mathbb{P}[|\varepsilon| > t | \mathbf{f}, \mathbf{u}] \leq 2e^{-c_1 t^2}$  a.s.

**Assump 4:**  $\pi^*, \mu_0^*, \mu_1^*$  is  $C$ -Lipschitz and  $\|\pi^*\|_{\infty}, \|\mu_0^*\|_{\infty}, \|\mu_1^*\|_{\infty} \lesssim 1$ .

# A data-driven method for $W$

PCA for other  $n_1$  data  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$  with a large  $\bar{r}$ .

★ Apply PCA to sample covariance matrix  $\hat{\Sigma} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^\top$ .

★  $\tilde{W} = [\sqrt{\hat{\lambda}_1} \hat{\mathbf{v}}_1, \dots, \sqrt{\hat{\lambda}_{\bar{r}}} \hat{\mathbf{v}}_{\bar{r}}]$  for top- $\bar{r}$  eigenvalues and eigenvectors of  $\hat{\Sigma}$ .

↪ remove the incoherence assumption in Fan and Gu (24).

Theorem 1. Under assumptions 1-2,

if  $\lambda(\mathbf{B}^\top \mathbf{B}) \asymp p$  then w.h.p.,

$$v_{\min}(p^{-1} \tilde{W}^\top \mathbf{B}) \geq C \left( 1 - r \sqrt{\frac{\log p}{n_1}} + r^2 \sqrt{\frac{\log r}{n_1}} + \sqrt{\frac{1}{p}} \right).$$

↪ only requires  $n_1 \gg r^2 \log p$ .

★ negligible

# A data-driven method for $W$

PCA for other  $n_1$  data  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$  with a large  $\bar{r}$ .

★ Apply PCA to sample covariance matrix  $\hat{\Sigma} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^\top$ .

★  $\tilde{W} = [\sqrt{\hat{\lambda}_1} \hat{\mathbf{v}}_1, \dots, \sqrt{\hat{\lambda}_{\bar{r}}} \hat{\mathbf{v}}_{\bar{r}}]$  for top- $\bar{r}$  eigenvalues and eigenvectors of  $\hat{\Sigma}$ .

↪ remove the incoherence assumption in Fan and Gu (24).

**Theorem 1.** Under assumptions 1-2,

if  $\lambda(\mathbf{B}^\top \mathbf{B}) \asymp p$  then w.h.p.,

$$v_{\min}(p^{-1} \tilde{W}^\top \mathbf{B}) \geq C \left( 1 - r \sqrt{\frac{\log p}{n_1}} + r^2 \sqrt{\frac{\log r}{n_1}} + \sqrt{\frac{1}{p}} \right).$$

↪ only requires  $n_1 \gg r^2 \log p$ .

★ negligible

# A data-driven method for $W$

PCA for other  $n_1$  data  $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$  with a large  $\bar{r}$ .

★ Apply PCA to sample covariance matrix  $\hat{\Sigma} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^\top$ .

★  $\tilde{W} = [\sqrt{\hat{\lambda}_1} \hat{\mathbf{v}}_1, \dots, \sqrt{\hat{\lambda}_{\bar{r}}} \hat{\mathbf{v}}_{\bar{r}}]$  for top- $\bar{r}$  eigenvalues and eigenvectors of  $\hat{\Sigma}$ .

↪ remove the incoherence assumption in Fan and Gu (24).

**Theorem 1.** Under assumptions 1-2,

if  $\lambda(\mathbf{B}^\top \mathbf{B}) \asymp p$  then w.h.p.,

$$v_{\min}(p^{-1} \tilde{W}^\top \mathbf{B}) \geq C \left( 1 - r \sqrt{\frac{\log p}{n_1}} + r^2 \sqrt{\frac{\log r}{n_1}} + \sqrt{\frac{1}{p}} \right).$$

↪ only requires  $n_1 \gg r^2 \log p$ .

★ negligible

# Oracle-type inequality for FAST-NN estimator

Theorem 2. Under assumptions 1-4,

[Fan & Gu, 2024]

if  $N \geq 2(r + |\mathcal{J}|)$ ,  $B \geq c_1 r |\mathcal{J}|$  and  $\lambda \geq c_1 \frac{\log(p)}{n}$  and  $\eta^{-1} \geq c_2 n$ , then for each estimator  $\hat{g}$  (outcome and propensity functions), we have

$$\mathbb{P}\left(\|\hat{g}_j - g_j^*\|_2^2 \leq c_3 (\delta_{\text{opt}} + \delta_{j,a} + \delta_s + \delta_f + \frac{t}{n})\right) \geq 1 - 3e^{-t}$$

$$\text{approx. error: } \delta_{j,a} = \inf_{g_j \in \mathcal{G}_{NN}} \|m_j^{\text{FAST}}(\cdot) - g_j^*\|_\infty^2$$

$$\text{stochastic error: } \delta_s = \lambda + \frac{(LN)^2}{n}$$

$$\text{factor est error: } \delta_f = \frac{|\mathcal{J}| r \cdot \bar{r}}{v_{\min}^2(\mathbf{H}) \cdot p}, \quad \mathbf{H} = p^{-1} \tilde{\mathbf{W}}^\top \mathbf{B}$$

★ Need modifications of the proofs to random samples with  $T = 0$  and  $T = 1$ .

# Oracle-type inequality for FAST-NN estimator

Theorem 2. Under assumptions 1-4,

[Fan & Gu, 2024]

if  $N \geq 2(r + |\mathcal{J}|)$ ,  $B \geq c_1 r |\mathcal{J}|$  and  $\lambda \geq c_1 \frac{\log(p)}{n}$  and  $\eta^{-1} \geq c_2 n$ , then for each estimator  $\hat{g}$  (outcome and propensity functions), we have

$$\mathbb{P}\left(\|\hat{g}_j - g_j^*\|_2^2 \leq c_3 (\delta_{\text{opt}} + \delta_{j,a} + \delta_s + \delta_f + \frac{t}{n})\right) \geq 1 - 3e^{-t}$$

$$\text{approx. error: } \delta_{j,a} = \inf_{g_j \in \mathcal{G}_{NN}} \|m_j^{\text{FAST}}(\cdot) - g_j^*\|_\infty^2$$

$$\text{stochastic error: } \delta_s = \lambda + \frac{(LN)^2}{n}$$

$$\text{factor est error: } \delta_f = \frac{|\mathcal{J}| r \cdot \bar{r}}{v_{\min}^2(\mathbf{H}) \cdot p}, \quad \mathbf{H} = p^{-1} \tilde{\mathbf{W}}^\top \mathbf{B}$$

★ Need modifications of the proofs to random samples with  $T = 0$  and  $T = 1$ .

# Oracle-type inequality for FAST-NN estimator

Theorem 2. Under assumptions 1-4,

[Fan & Gu, 2024]

if  $N \geq 2(r + |\mathcal{J}|)$ ,  $B \geq c_1 r |\mathcal{J}|$  and  $\lambda \geq c_1 \frac{\log(p)}{n}$  and  $\eta^{-1} \geq c_2 n$ , then for each estimator  $\hat{g}$  (outcome and propensity functions), we have

$$\mathbb{P}\left(\|\hat{g}_j - g_j^*\|_2^2 \leq c_3 (\delta_{\text{opt}} + \delta_{j,a} + \delta_s + \delta_f + \frac{t}{n})\right) \geq 1 - 3e^{-t}$$

$$\text{approx. error: } \delta_{j,a} = \inf_{g_j \in \mathcal{G}_{NN}} \|m_j^{\text{FAST}}(\cdot) - g_j^*\|_\infty^2$$

$$\text{stochastic error: } \delta_s = \lambda + \frac{(LN)^2}{n}$$

$$\text{factor est error: } \delta_f = \frac{|\mathcal{J}| r \cdot \bar{r}}{v_{\min}^2(\mathbf{H}) \cdot p}, \quad \mathbf{H} = p^{-1} \tilde{\mathbf{W}}^\top \mathbf{B}$$

★ Need modifications of the proofs to random samples with  $T = 0$  and  $T = 1$ .



# Asymptotic normality

Theorem 3. Under assumptions 1-4,  $r + |\mathcal{J}| \leq C$ ,  $\|\mathbf{u}\|_\infty \leq b$ .

- $n^{c_1} < p < n^{100}$  for some constant  $c_1 > \frac{1}{2}$ ,
- $\mu_0^*, \mu_1^*, \pi^* \in \mathcal{H}(r + |\mathcal{J}|, l, \mathcal{P})$ ,  $\gamma^* > \frac{1}{2}$
- estimated by the FAST-NN estimators with  $\delta_{opt} < (n/\log n)^{-\frac{\gamma^*}{2\gamma^*+1}}$ .

★ asymptotic normality:

$$\sqrt{n}(\hat{\tau}^{\text{FIDDLE}} - \tau) \xrightarrow{d} N(0, \sigma^2),$$

★ semiparametric efficiency:

Hahn (1998)

$$\sigma^2 = \mathbb{E} \left[ (\mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x}) - \mu)^2 + \frac{\text{Var}[y(1)|\mathbf{x}]}{\pi^*(\mathbf{x})} + \frac{\text{Var}[y(0)|\mathbf{x}]}{1 - \pi^*(\mathbf{x})} \right]$$

★ Required  $\gamma_0^* \gamma_2^* > \frac{1}{4}$  and  $\gamma_1^* \gamma_2^* > \frac{1}{4}$ .

★ hold for  $r = 0$



# Asymptotic normality

Theorem 3. Under assumptions 1-4,  $r + |\mathcal{J}| \leq C$ ,  $\|\mathbf{u}\|_\infty \leq b$ .

- $n^{c_1} < p < n^{100}$  for some constant  $c_1 > \frac{1}{2}$ ,
- $\mu_0^*, \mu_1^*, \pi^* \in \mathcal{H}(r + |\mathcal{J}|, l, \mathcal{P})$ ,  $\gamma^* > \frac{1}{2}$
- estimated by the FAST-NN estimators with  $\delta_{opt} < (n/\log n)^{-\frac{\gamma^*}{2\gamma^*+1}}$ .

★ asymptotic normality:

$$\sqrt{n}(\hat{\tau}^{\text{FIDDLE}} - \tau) \xrightarrow{d} N(0, \sigma^2),$$

★ semiparametric efficiency:

Hahn (1998)

$$\sigma^2 = \mathbb{E} \left[ (\mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x}) - \mu)^2 + \frac{\text{Var}[y(1)|\mathbf{x}]}{\pi^*(\mathbf{x})} + \frac{\text{Var}[y(0)|\mathbf{x}]}{1 - \pi^*(\mathbf{x})} \right]$$

★ Required  $\gamma_0^* \gamma_2^* > \frac{1}{4}$  and  $\gamma_1^* \gamma_2^* > \frac{1}{4}$ .

★ hold for  $r = 0$

# Asymptotic normality

Theorem 3. Under assumptions 1-4,  $r + |\mathcal{J}| \leq C$ ,  $\|\mathbf{u}\|_\infty \leq b$ .

- $n^{c_1} < p < n^{100}$  for some constant  $c_1 > \frac{1}{2}$ ,
- $\mu_0^*, \mu_1^*, \pi^* \in \mathcal{H}(r + |\mathcal{J}|, l, \mathcal{P})$ ,  $\gamma^* > \frac{1}{2}$
- estimated by the FAST-NN estimators with  $\delta_{opt} < (n/\log n)^{-\frac{\gamma^*}{2\gamma^*+1}}$ .

★ asymptotic normality:

$$\sqrt{n}(\hat{\tau}^{\text{FIDDLE}} - \tau) \xrightarrow{d} N(0, \sigma^2),$$

★ semiparametric efficiency:

Hahn (1998)

$$\sigma^2 = \mathbb{E} \left[ (\mu_1^*(\mathbf{x}) - \mu_0^*(\mathbf{x}) - \mu)^2 + \frac{\text{Var}[y(1)|\mathbf{x}]}{\pi^*(\mathbf{x})} + \frac{\text{Var}[y(0)|\mathbf{x}]}{1 - \pi^*(\mathbf{x})} \right]$$

★ Required  $\gamma_0^* \gamma_2^* > \frac{1}{4}$  and  $\gamma_1^* \gamma_2^* > \frac{1}{4}$ .

★ hold for  $r = 0$

# Numerical Results

# Simulations: data generating process

★  $r = 4$ ,  $\mathcal{J} = \{1, \dots, 5\}$ ,  $n = 5000$ , varying  $p \in \{10, \dots, 10000\}$ ;

★ **Propensity model:**  $f_i, u_i \sim_{iid} U(-1, 1)$ ,  $b_{i,j} \sim_{iid} U(-\sqrt{3}, \sqrt{3})$ .

$$\pi^*(\mathbf{f}, \mathbf{u}) = 0.8\sigma\left(\sin(f_1) + \tan(f_2) + f_3 + f_4 + \sum_{j=1}^5 u_j\right) + 0.1;$$

★ **Outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_J) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\mathbf{f}, \mathbf{u}) = 10 + f_1 + f_2 f_3 + \sin(f_4) + \log(5 + u_1 + u_2 u_3) + \tan(u_4) + u_5;$$

$$\tau^*(\mathbf{f}, \mathbf{u}) = 5 + f_1 + f_2 + \sin(f_3) + \tan(f_4) + u_1 + u_2 + \sin(u_3 + u_4) + \tan(u_5).$$

★ ground truth  $\text{ATE}^* = \mathbb{E}[\tau^*(\mathbf{f}, \mathbf{u})] = 5$

$N_{sim} = 100$

# Simulations: data generating process

★  $r = 4$ ,  $\mathcal{J} = \{1, \dots, 5\}$ ,  $n = 5000$ , varying  $p \in \{10, \dots, 10000\}$ ;

★ **Propensity model:**  $f_i, u_i \sim_{iid} U(-1, 1)$ ,  $b_{i,j} \sim_{iid} U(-\sqrt{3}, \sqrt{3})$ .

$$\pi^*(\mathbf{f}, \mathbf{u}) = 0.8 \sigma \left( \sin(f_1) + \tan(f_2) + f_3 + f_4 + \sum_{j=1}^5 u_j \right) + 0.1;$$

★ **Outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T \tau^*(\mathbf{f}, \mathbf{u}_\mathcal{J}) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\mathbf{f}, \mathbf{u}) = 10 + f_1 + f_2 f_3 + \sin(f_4) + \log(5 + u_1 + u_2 u_3) + \tan(u_4) + u_5;$$

$$\tau^*(\mathbf{f}, \mathbf{u}) = 5 + f_1 + f_2 + \sin(f_3) + \tan(f_4) + u_1 + u_2 + \sin(u_3 + u_4) + \tan(u_5).$$

★ ground truth  $\text{ATE}^* = \mathbb{E}[\tau^*(\mathbf{f}, \mathbf{u})] = 5$

$N_{sim} = 100$

# Simulations: data generating process

★  $r = 4$ ,  $\mathcal{J} = \{1, \dots, 5\}$ ,  $n = 5000$ , varying  $p \in \{10, \dots, 10000\}$ ;

★ **Propensity model:**  $f_i, u_i \sim_{iid} U(-1, 1)$ ,  $b_{i,j} \sim_{iid} U(-\sqrt{3}, \sqrt{3})$ .

$$\pi^*(\mathbf{f}, \mathbf{u}) = 0.8\sigma\left(\sin(f_1) + \tan(f_2) + f_3 + f_4 + \sum_{j=1}^5 u_j\right) + 0.1;$$

★ **Outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_g) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\mathbf{f}, \mathbf{u}) = 10 + f_1 + f_2 f_3 + \sin(f_4) + \log(5 + u_1 + u_2 u_3) + \tan(u_4) + u_5;$$

$$\tau^*(\mathbf{f}, \mathbf{u}) = 5 + f_1 + f_2 + \sin(f_3) + \tan(f_4) + u_1 + u_2 + \sin(u_3 + u_4) + \tan(u_5).$$

★ ground truth  $\text{ATE}^* = \mathbb{E}[\tau^*(\mathbf{f}, \mathbf{u})] = 5$

$N_{sim} = 100$

# Simulations: data generating process

★  $r = 4$ ,  $\mathcal{J} = \{1, \dots, 5\}$ ,  $n = 5000$ , varying  $p \in \{10, \dots, 10000\}$ ;

★ **Propensity model:**  $f_i, u_i \sim_{iid} U(-1, 1)$ ,  $b_{i,j} \sim_{iid} U(-\sqrt{3}, \sqrt{3})$ .

$$\pi^*(\mathbf{f}, \mathbf{u}) = 0.8\sigma\left(\sin(f_1) + \tan(f_2) + f_3 + f_4 + \sum_{j=1}^5 u_j\right) + 0.1;$$

★ **Outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_g) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\mathbf{f}, \mathbf{u}) = 10 + f_1 + f_2 f_3 + \sin(f_4) + \log(5 + u_1 + u_2 u_3) + \tan(u_4) + u_5;$$

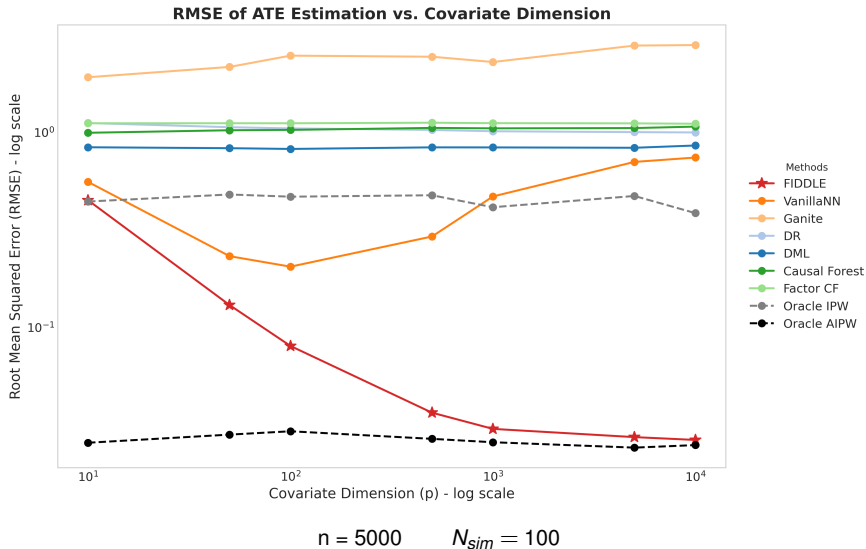
$$\tau^*(\mathbf{f}, \mathbf{u}) = 5 + f_1 + f_2 + \sin(f_3) + \tan(f_4) + u_1 + u_2 + \sin(u_3 + u_4) + \tan(u_5).$$

★ ground truth  $ATE^* = \mathbb{E}[\tau^*(\mathbf{f}, \mathbf{u})] = 5$

$N_{sim} = 100$



# Comparison of Methods

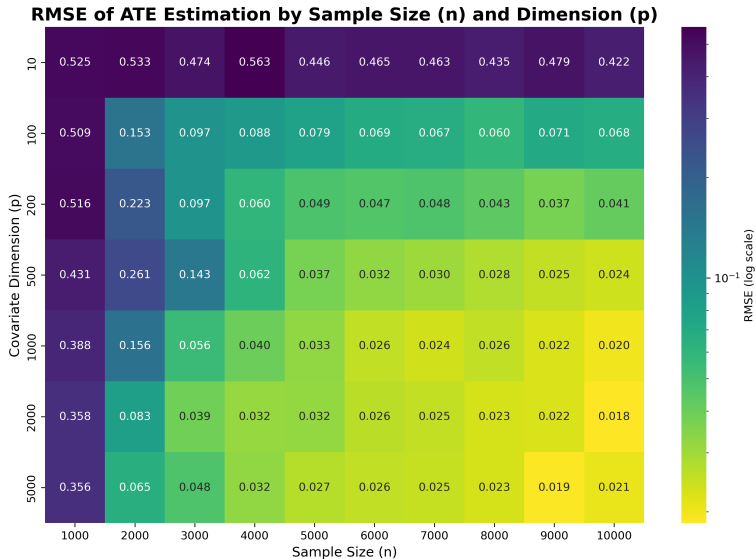


★ FIDDLE close to oracle

★ Vanilla-NN suffers COD.

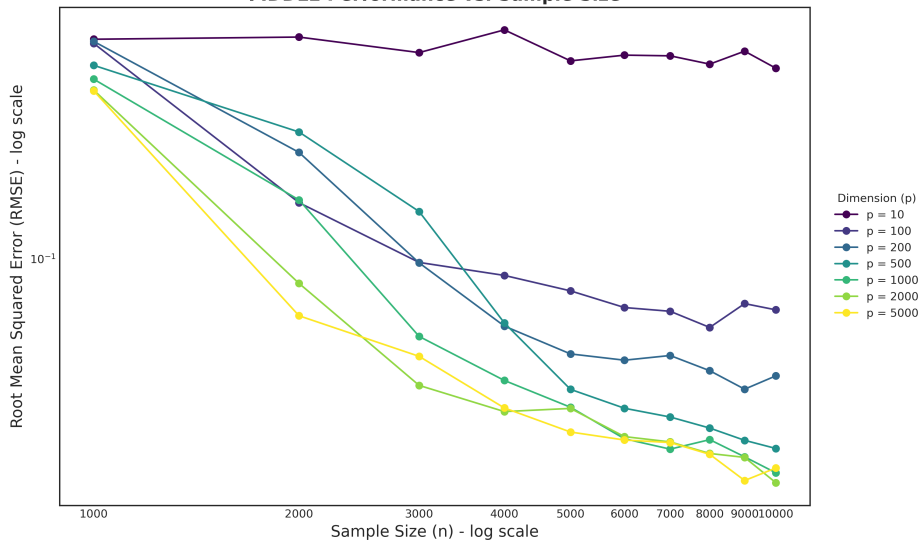
★ outperforms other methods

# Performance accross $n$ and $p$



# Rates of Convergence

FIDDLE Performance vs. Sample Size



# Semi-synthetic image data

★ CIFAR-10 dataset (Canadian Institute For Advanced Research)

★ benchmark in computer vision:  $n = 60,000$  and  $p = 3,072$   $32 \times 32 \times 3$

★ synthetic propensity model:  $\tilde{f}_{ij}, \tilde{u}_i$  taken from fitting factor model  $r = 4$

$$\pi^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 0.8\sigma(\sin(\tilde{f}_1) + \sum_{i=2}^4 \tilde{f}_i + \sin(\tilde{u}_1) + \sum_{j=2}^5 \tilde{u}_j) + 0.1$$

★ synthetic outcome model:  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_j) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 10 + \tilde{f}_1 + \sin(\tilde{f}_2) + \tilde{f}_3\tilde{f}_4 + \tilde{u}_1(\tilde{u}_2 + \sin(\tilde{u}_3)) + \tilde{u}_4 + \tilde{u}_5,$$

$$\tau^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = \tilde{f}_1(\tilde{f}_2 + 3) + \tilde{f}_3 + \sin(\tilde{f}_4) + \sin(\tilde{u}_1) + \tilde{u}_2 + \tilde{u}_3\tilde{u}_4\tilde{u}_5.$$

# Semi-synthetic image data

★ CIFAR-10 dataset (Canadian Institute For Advanced Research)

★ benchmark in computer vision:  $n = 60,000$  and  $p = 3,072$   $32 \times 32 \times 3$

★ **synthetic propensity model:**  $\tilde{f}_{ij}, \tilde{u}_i$  taken from fitting factor model  $r = 4$

$$\pi^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 0.8\sigma(\sin(\tilde{f}_1) + \sum_{i=2}^4 \tilde{f}_i + \sin(\tilde{u}_1) + \sum_{j=2}^5 \tilde{u}_j) + 0.1$$

★ **synthetic outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_j) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 10 + \tilde{f}_1 + \sin(\tilde{f}_2) + \tilde{f}_3\tilde{f}_4 + \tilde{u}_1(\tilde{u}_2 + \sin(\tilde{u}_3)) + \tilde{u}_4 + \tilde{u}_5,$$

$$\tau^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = \tilde{f}_1(\tilde{f}_2 + 3) + \tilde{f}_3 + \sin(\tilde{f}_4) + \sin(\tilde{u}_1) + \tilde{u}_2 + \tilde{u}_3\tilde{u}_4\tilde{u}_5.$$

# Semi-synthetic image data

★ CIFAR-10 dataset (Canadian Institute For Advanced Research)

★ benchmark in computer vision:  $n = 60,000$  and  $p = 3,072$   $32 \times 32 \times 3$

★ **synthetic propensity model:**  $\tilde{f}_{ij}, \tilde{u}_i$  taken from fitting factor model  $r = 4$

$$\pi^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 0.8\sigma(\sin(\tilde{f}_1) + \sum_{i=2}^4 \tilde{f}_i + \sin(\tilde{u}_1) + \sum_{j=2}^5 \tilde{u}_j) + 0.1$$

★ **synthetic outcome model:**  $y = \mu^*(\mathbf{f}, \mathbf{u}) + T\tau^*(\mathbf{f}, \mathbf{u}_g) + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, 1/4)$

$$\mu^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = 10 + \tilde{f}_1 + \sin(\tilde{f}_2) + \tilde{f}_3\tilde{f}_4 + \tilde{u}_1(\tilde{u}_2 + \sin(\tilde{u}_3)) + \tilde{u}_4 + \tilde{u}_5,$$

$$\tau^*(\tilde{\mathbf{f}}, \tilde{\mathbf{u}}) = \tilde{f}_1(\tilde{f}_2 + 3) + \tilde{f}_3 + \sin(\tilde{f}_4) + \sin(\tilde{u}_1) + \tilde{u}_2 + \tilde{u}_3\tilde{u}_4\tilde{u}_5.$$

# CIFAR-10 semi-synthetic dataset results

Method	RMSE	(SE)
Oracle AIPW	0.009	(0.001)
<b>FIDDLE</b>	<b>0.030</b>	<b>(0.003)</b>
Vanilla NN	0.282	(0.012)
GANITE	1.389	(0.032)
DR	1.664	(0.007)
DML	1.427	(0.007)
CF	1.878	(0.007)
Factor CF	1.990	(0.006)
Oracle IPW	0.448	(0.030)

★  $n = 5000$

# Application: bariatric surgery

- ★ MBSAQIP dataset (Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program): weight loss surgery
- ★ 174,013 records with 42 pretreatment covariates from year 2017
- ★ outcome: **30-day BMI reduction**; treatment: surgery type
  - ▶ Sleeve Gastrectomy: most widely performed as **control**;
  - ▶ RYGB (Roux-en-Y Gastric Bypass);
  - ▶ AGB (Adjustable Gastric Band);
  - ▶ BPD/DS (Biliopancreatic Diversion with Duodenal Switch);
  - ▶ SADI-S (Single Anastomosis Duodeno-Ileal Bypass with Sleeve Gastrectomy)



# Application: bariatric surgery

- ★ MBSAQIP dataset (Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program): weight loss surgery
- ★ 174,013 records with 42 pretreatment covariates from year 2017
- ★ outcome: **30-day BMI reduction**; treatment: surgery type
  - ▶ Sleeve Gastrectomy: most widely performed as **control**;
  - ▶ RYGB (Roux-en-Y Gastric Bypass);
  - ▶ AGB (Adjustable Gastric Band);
  - ▶ BPD/DS (Biliopancreatic Diversion with Duodenal Switch);
  - ▶ SADI-S (Single Anastomosis Duodeno-Ileal Bypass with Sleeve Gastrectomy)

# Bariatric surgery type comparison

Surgery	RYGB	AGB	BPD/DS	SADI-S
FIDDLE	-0.011 (-0.012, -0.011)	-1.039 (-1.052, -1.027)	0.336 (0.311, 0.362)	-0.502 (-0.519, -0.485)
Vanilla NN	-0.381 (-0.388, -0.375)	-1.681 (-1.691, -1.670)	-0.367 (-0.386, -0.348)	-1.544 (-1.557, -1.530)
GANITE	-0.451 (-0.516, -0.386)	-2.465 (-2.659, -2.271)	-1.502 (-1.655, -1.348)	-2.045 (-2.169, -1.922)
DR	-0.031 (-0.031, -0.031)	-0.734 (-0.741, -0.727)	0.031 (0.031, 0.031)	-0.512 (-0.518, -0.506)
DML	-0.015 (-0.016, -0.015)	-0.240 (-0.273, -0.208)	0.631 (0.606, 0.655)	-0.546 (-0.550, -0.542)
CF	-0.023 (-0.023, -0.022)	-1.080 (-1.082, -1.079)	0.214 (0.211, 0.216)	-0.647 (-0.649, -0.645)
Factor-CF	-0.057 (-0.057, -0.056)	-1.092 (-1.094, -1.091)	0.175 (0.173, 0.177)	-0.7773 (-0.779, -0.776)

# Bariatric surgery type comparison

- ★ FIDDLE yields robust results as the comparing methods
- ★ Sleeve - removes part of the stomach - as control
- ★ RYGB - create new stomach pouch - small negative ATE
- ★ AGB - restricts stomach with an adjustable band - large negative ATE
- ★ BPD/DS - bypass surgery after sleeve gastrectomy - positive ATE
- ★ SADI-S - simplified BPD/DS similar to Sleeve - moderate negative ATE

# Bariatric surgery type comparison

- ★ FIDDLE yields robust results as the comparing methods
- ★ Sleeve - removes part of the stomach - as control
- ★ RYGB - create new stomach pouch - small negative ATE
- ★ AGB - restricts stomach with an adjustable band - large negative ATE
- ★ BPD/DS - bypass surgery after sleeve gastrectomy - positive ATE
- ★ SADI-S - simplified BPD/DS similar to Sleeve - moderate negative ATE

# Bariatric surgery type comparison

- ★ FIDDLE yields robust results as the comparing methods
- ★ Sleeve - removes part of the stomach - as control
- ★ RYGB - create new stomach pouch - small negative ATE
- ★ AGB - restricts stomach with an adjustable band - large negative ATE
- ★ BPD/DS - bypass surgery after sleeve gastrectomy - positive ATE
- ★ SADI-S - simplified BPD/DS similar to Sleeve - moderate negative ATE

# Concluding Remarks

- 1 Introduce FIDDLE for ATE estimation with high-dimensional covariate.
- 2 Leverage •FAST-NN for non-param propensity and outcome modeling.
- 3 Encompass most statistical models, •factor models, •sparse models.
- 4 Learning attribution variables for propensity and outcome model
- 5 Guarantee doubly-robustness by •AIPW estimator
- 6 Demonstrate asymptotic normality and semiparametric efficiency
- 7 Verify results on both generated and semi-synthetic datasets
- 8 Apply to bariatric surgery data for surgical type-weight loss analysis

# Concluding Remarks

- 1 Introduce FIDDLE for ATE estimation with high-dimensional covariate.
- 2 Leverage •FAST-NN for non-param propensity and outcome modeling.
- 3 Encompass most statistical models, •factor models, •sparse models.
- 4 Learning attribution variables for propensity and outcome model
- 5 Guarantee doubly-robustness by •AIPW estimator
- 6 Demonstrate asymptotic normality and semiparametric efficiency
- 7 Verify results on both generated and semi-synthetic datasets
- 8 Apply to bariatric surgery data for surgical type-weight loss analysis

# Concluding Remarks

- 1 Introduce FIDDLE for ATE estimation with high-dimensional covariate.
- 2 Leverage •FAST-NN for non-param propensity and outcome modeling.
- 3 Encompass most statistical models, •factor models, •sparse models.
- 4 Learning attribution variables for propensity and outcome model
- 5 Guarantee doubly-robustness by •AIPW estimator
- 6 Demonstrate asymptotic normality and semiparametric efficiency
- 7 Verify results on both generated and semi-synthetic datasets
- 8 Apply to bariatric surgery data for surgical type-weight loss analysis



# Concluding Remarks

- 1 Introduce FIDDLE for ATE estimation with high-dimensional covariate.
- 2 Leverage •FAST-NN for non-param propensity and outcome modeling.
- 3 Encompass most statistical models, •factor models, •sparse models.
- 4 Learning attribution variables for propensity and outcome model
- 5 Guarantee doubly-robustness by •AIPW estimator
- 6 Demonstrate asymptotic normality and semiparametric efficiency
- 7 Verify results on both generated and semi-synthetic datasets
- 8 Apply to bariatric surgery data for surgical type-weight loss analysis

# The End

*Thank*



*You*

★ Fan, J., S. Jana, S. Kulkarni, and Q. Yin (2025). “Factor Informed Double Deep Learning For Average Treatment Effect Estimation.” arXiv preprint arXiv:2508.17136 .