A New and Efficient Debiased Estimation of General Treatment Models by Balanced Neural Networks Weighting

Zeqi Wu[1] Meilin Wang[1] Wei Huang[2] Zheng Zhang[1]

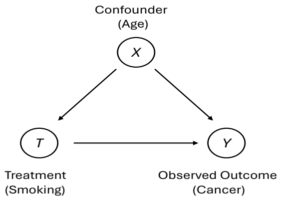
[1]Renmin University of China, China

[2]University of Melbourne, Australia

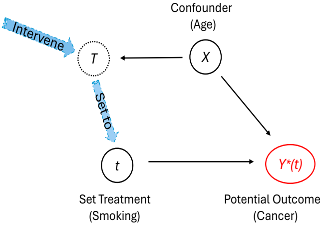
October 18th, 2025

Revisit some common treatment effect analysis

Observation



We want to infer



Assumption 1 (Unconfoundedness). $Y^*(t) \perp T \mid \boldsymbol{X}$ for all $t \in \mathcal{T} \subset \mathbb{R}$.

$\mathcal{T}$	Treatment effect	Identification
$\{0,1\}$	ATE: $\mathbb{E}[Y^*(1) - Y^*(0)]$	$\mathbb{E}\left[\frac{TY}{\mathbb{P}(T=1 \mid \boldsymbol{X})}\right] - \mathbb{E}\left[\frac{(1-T)Y}{1-\mathbb{P}(T=1 \mid \boldsymbol{X})}\right]$
General	ADRF: $\mathbb{E}[Y^*(t)]$	$\mathbb{E}\left[\frac{dF_T(T)}{dF_{T \mid X}(T \mid \boldsymbol{X})} Y \,\middle\vert\, T = t\right]$

Nuisance parameter (Propensity Scores): $\mathbb{P}(T = 1|\boldsymbol{X})$ or $dF_{T|X}(T|\boldsymbol{X})$ (Generalised).

Challenges:

- Unstable ratio-type estimation: Sensitive to small errors in estimating the propensity scores;
- $\boldsymbol{X} \in \mathbb{R}^d$, d is usually large:
  - Parametric models $\to$ Model misspecifications
  - Nonparametric estimation $\to$ Bias and/or curse of dimensionality

A unified framework

We consider a unified treatment model:

- $L(y, z) : \mathbb{R} \times \mathbb{R} \to [0, \infty)$: a general (possibly non-smooth) loss function; e.g., $(y - z)^2/2$;
- $g(t; \beta)$: a user-specified parametric dose-response function indexed by a p-dimensional parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$;
- For discrete $T \in \mathcal{T} = \{0, \ldots, J\}$, $g(t; \beta) = \sum_{j=0}^{J} \beta_j 1(t = j)$;
  - e.g., for binary T, $g(t; \beta) = \beta_0 \cdot (1 - t) + \beta_1 \cdot t$.

Then $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{p-1}^*)^\top$ is defined by:

$$\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \int_{\mathcal{T}} \mathbb{E}\left[L(Y^*(t), g(t; \beta))\right] dF_T(t).$$

A general causal framework

This framework covers a broad class of treatment effect models in the literature.

$L(y,z)$	$g(t;\beta^*)$
$(y-z)^2/2$	$\mathbb{E}\{Y^*(t)\}$
$(y-z)\cdot\{\tau - 1(y-z\le 0)\}$ for $\tau\in(0,1)$	the τ-th quantile of $Y^*(t)$
$-y\log z - (1-y)\log(1-z)$ for $y,z\in\{0,1\}$	$P\{Y^*(t)=1\}$

- Binary case ($g(t;\beta)=\beta_0\cdot(1-t)+\beta_1\cdot t$): $\beta_1^* - \beta_0^*$ is
  - row 1: Average Treatment Effect (ATE)
  - row 2: Quantile Treatment Effect (QTE)

A general causal framework

This framework covers a broad class of treatment effect models in the literature.

$L(y, z)$	$g(t; \beta^*)$
$(y-z)^2/2$	$\mathbb{E}\{Y^*(t)\}$
$(y-z) \cdot \{\tau - 1(y-z \leq 0)\}$ for $\tau \in (0,1)$	the τ-th quantile of $Y^*(t)$
$-y \log z - (1-y)\log(1-z)$ for $y, z \in \{0,1\}$	$P\{Y^*(t) = 1\}$

- Binary case $(g(t;\beta) = \beta_0 \cdot (1-t) + \beta_1 \cdot t)$: $\beta_1^* - \beta_0^*$ is
  - row 1: Average Treatment Effect (ATE)
  - row 2: Quantile Treatment Effect (QTE)

Identifying the general treatment effect parameter

- $Y^*(t)$ cannot be observed simultaneously for all t.
- Use the **stablised weight**:

$$\pi_0(T, \boldsymbol{X}) := \frac{dF_T(T)}{dF_{T|X}(T|\boldsymbol{X})} = \frac{dF_T(T)dF_{\boldsymbol{X}}(\boldsymbol{X})}{dF_{TX}(T, \boldsymbol{X})} .$$

- Identify: $\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \mathbb{E}[\pi_0(T, \boldsymbol{X}) L(Y, g(T; \boldsymbol{\beta}))]$, which can be estimated by

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{N} \widehat{\pi}(T_i, \boldsymbol{X}_i) L(Y_i, g(T_i; \boldsymbol{\beta})) ,$$

given an estimator $\widehat{\pi}$.

Double/Debiased Machine Learning (Chernozhukov et al. 2018, Kallus et al. 2024)

- $\widehat{\beta}_{\mathrm{DML}}$ solves the efficient influence function (EIF) score: $\sum_{i=1}^{N} \widehat{\psi}\{\beta; \widehat{\pi}, \widehat{\mu}, T_i, Y_i, \boldsymbol{X}_i\} = 0$, where $\widehat{\psi}$ estimates the EIF

$$\begin{aligned}\psi(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*, \pi_0, \mu_0) &:= \pi_0(T, \boldsymbol{X}) h(Y, T; \boldsymbol{\beta}^*) - \pi_0(T, \boldsymbol{X}) \mu_0(T, \boldsymbol{X}; \boldsymbol{\beta}^*) \\ &+ \mathbb{E}\left[\mu_0(T, \boldsymbol{X}; \boldsymbol{\beta}^*) \pi_0(T, \boldsymbol{X}) \mid T\right] + \mathbb{E}\left[\mu_0(T, \boldsymbol{X}; \boldsymbol{\beta}^*) \pi_0(T, \boldsymbol{X}) \mid \boldsymbol{X}\right],\end{aligned}$$

  with $\mu_0(t, \boldsymbol{x}; \boldsymbol{\beta}^*) = \mathbb{E}\left[L'(Y, g(T; \boldsymbol{\beta}^*)) \cdot \partial g(T; \boldsymbol{\beta}^*) / \partial \boldsymbol{\beta} | T = t, \boldsymbol{X} = \boldsymbol{x}\right]$.
- Under some regularity conditions, they show that

$$\widehat{\beta}_{\mathrm{DML}} - \beta^* \xrightarrow{d} N(0, V_{\mathrm{eff}}),$$

  given that $\|\widehat{\pi} - \pi_0\|_{P,2} \cdot \|\widehat{\mu} - \mu_0\|_{P,2} = o_p(N^{-1/2})$ – *rate-doubly robust.*

Double/Debiased Machine Learning (Chernozhukov et al. 2018, Kallus et al. 2024)

- Designed mainly for discrete treatment cases;
- They used a ratio-type estimator for π_0 – leads to an unstable estimator of β^* for N not large enough.
  - Automatic debiased machine learning (Chernozhukov, Newey, Singh, 2022) regards π_0 as a whole, but it mainly designed for $\beta = \mathbb{E}[m(W, \gamma)]$;
- The EIF can be computationally complicated in the general treatment models.

Our method

1. Directly estimate π_0 by $\widehat{\pi}_{\mathrm{DNN}}$, using deep neural networks;
2. Correct the bias in estimating β^* with $\widehat{\pi}_{\mathrm{DNN}}$, by reweighting $\widehat{\pi}_{\mathrm{DNN}}$ using the covariate balancing property.

Contributions:

- We do NOT need to estimate the EIF ψ.
- The reweighting technique improves finite sample performance.
- We show our estimator is rate-doubly robust, $\sqrt{N}$-consistent, asymptotically normal and semiparametric efficient.
- We propose a simple inference procedure based on the weighted bootstrap.

Our method

1. Directly estimate π_0 by $\widehat{\pi}_{\mathrm{DNN}}$, using deep neural networks;
2. Correct the bias in estimating β^* with $\widehat{\pi}_{\mathrm{DNN}}$, by reweighting $\widehat{\pi}_{\mathrm{DNN}}$ using the covariate balancing property.

Contributions:

- We do NOT need to estimate the EIF ψ.
- The reweighting technique improves finite sample performance.
- We show our estimator is rate-doubly robust, $\sqrt{N}$-consistent, asymptotically normal and semiparametric efficient.
- We propose a simple inference procedure based on the weighted bootstrap.

- Identify π_0:

$$\begin{aligned}\pi_0 &= \arg\min_{\pi} \mathbb{E}[\{\pi(T, \boldsymbol{X}) - \pi_0(T, \boldsymbol{X})\}^2] \\ &= \arg\min_{\pi} \left(\mathbb{E}[\{\pi(T, \boldsymbol{X})\}^2] - 2 \int \pi(t, \boldsymbol{x}) \cdot \frac{dF_T(t)}{dF_{T|X}(t|\boldsymbol{x})} \cdot dF_{T,X}(t, \boldsymbol{x}) \right) \\ &= \arg\min_{\pi} \left(\mathbb{E}[\{\pi(T, \boldsymbol{X})\}^2] - 2\mathbb{E}_T \mathbb{E}_X \{\pi(T, \boldsymbol{X})\} \right)\end{aligned}$$

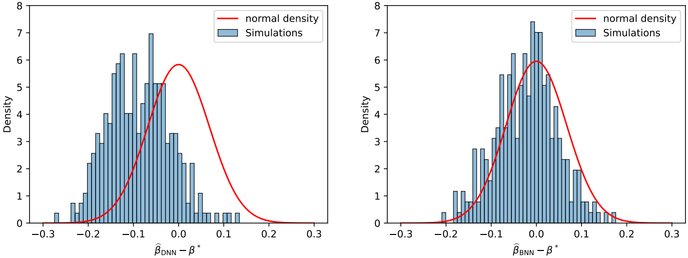
- We obtain the DNN estimator, $\widehat{\pi}_{\mathrm{DNN}}$, by minimising

$$\frac{1}{N} \sum_{i=1}^{N} \pi(T_i, \boldsymbol{X}_i)^2 - \frac{2}{N(N-1)} \sum_{j \neq i} \pi(T_i, \boldsymbol{X}_j)$$

over a class of deep neural networks (DNN) models for π.

Bias introduced by DNN

Under some regularity conditions, we show that the bias-variance trade-off L_2 rate of convergence of the DNN estimators, $\widehat{\pi}_{\mathrm{DNN}}$ of π_0 and $\widehat{\beta}_{\mathrm{DNNW}}$ of β^*, are both $N^{-s_\pi/(2s_\pi+d+1)} \cdot \log^5 N$, slower than $N^{-1/2}$.



Covariate Balancing

Covariate balancing property:

$$\mathbb{E}\{\pi_0(T, \boldsymbol{X})u(T, \boldsymbol{X})\} = \int\int u(t, \boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x})\, dF_T(t)\,,$$

for any integrable function u: T and $\boldsymbol{X}$ are weighted independent with π_0.

Balanced Neural Network

- $\widehat{\pi}_{\mathrm{DNN}}(T_i, \boldsymbol{X}_i)$, no covariate balancing property: for some u,

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_{\mathrm{DNN}}(T_i, \boldsymbol{X}_i)u(T_i, \boldsymbol{X}_i) \neq \frac{1}{N(N-1)}\sum_{j\neq i}u(T_i, \boldsymbol{X}_j)\,.$$

- Re-balance it by $\widehat{\pi}_{\mathrm{BNN}}(T_i, \boldsymbol{X}_i) := \widehat{w}_i\widehat{\pi}_{\mathrm{DNN}}(T_i, \boldsymbol{X}_i)$, where $\widehat{w}_i$ solves

$$\begin{cases}\min \sum_{i=1}^{N} D(w_i) \text{ subject to} \\ \frac{1}{N}\sum_{i=1}^{N} w_i\widehat{\pi}_{\mathrm{DNN}}(T_i, \boldsymbol{X}_i)\boldsymbol{\xi}(T_i, \boldsymbol{X}_i) = \frac{1}{N(N-1)}\sum_{j\neq i}\boldsymbol{\xi}(T_i, \boldsymbol{X}_j)\,,\end{cases}$$

  - $D(v)$ is a distance measure from v to 1
  - By choosing ξ appropriately, we can achieve a debiased estimator of β^*:

$$\widehat{\beta}_{\mathrm{BNNW}} := \arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{N}\widehat{\pi}_{\mathrm{BNN}}(T_i, \boldsymbol{X}_i)L(Y_i, g(T_i;\beta))\,.$$

Balanced Neural Network

- $\widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)$, no covariate balancing property: for some u,

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)u(T_i, \boldsymbol{X}_i) \neq \frac{1}{N(N-1)}\sum_{j\neq i} u(T_i, \boldsymbol{X}_j)\,.$$

- Re-balance it by $\widehat{\pi}_{\text{BNN}}(T_i, \boldsymbol{X}_i) := \widehat{w}_i\widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)$, where $\widehat{w}_i$ solves

$$\begin{cases} \min \sum_{i=1}^{N} D(w_i) \text{ subject to} \\ \frac{1}{N}\sum_{i=1}^{N} w_i\widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)\xi(T_i, \boldsymbol{X}_i) = \frac{1}{N(N-1)}\sum_{j\neq i}\xi(T_i, \boldsymbol{X}_j)\,, \end{cases}$$

  - $D(v)$ is a distance measure from v to 1
  - By choosing ξ appropriately, we can achieve a debiased estimator of β^*:

$$\widehat{\beta}_{\text{BNNW}} := \arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{N}\widehat{\pi}_{\text{BNN}}(T_i, \boldsymbol{X}_i)L(Y_i, g(T_i;\beta))\,.$$

Balanced Neural Network

- $\widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)$, no covariate balancing property: for some u,

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i) u(T_i, \boldsymbol{X}_i) \neq \frac{1}{N(N-1)} \sum_{j \neq i} u(T_i, \boldsymbol{X}_j)\,.$$

- Re-balance it by $\widehat{\pi}_{\text{BNN}}(T_i, \boldsymbol{X}_i) := \widehat{w}_i \widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)$, where $\widehat{w}_i$ solves

$$\begin{cases} \min \sum_{i=1}^{N} D(w_i) \text{ subject to} \\ \frac{1}{N}\sum_{i=1}^{N} w_i \widehat{\pi}_{\text{DNN}}(T_i, \boldsymbol{X}_i)\boldsymbol{\xi}(T_i, \boldsymbol{X}_i) = \frac{1}{N(N-1)} \sum_{j \neq i} \boldsymbol{\xi}(T_i, \boldsymbol{X}_j)\,, \end{cases}$$

  - $D(v)$ is a distance measure from v to 1
  - By choosing ξ appropriately, we can achieve a debiased estimator of $\boldsymbol{\beta}^*$:

$$\widehat{\boldsymbol{\beta}}_{\text{BNNW}} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{N} \widehat{\pi}_{\text{BNN}}(T_i, \boldsymbol{X}_i) L(Y_i, g(T_i; \boldsymbol{\beta}))\,.$$

Choices of ξ

- Let $L'(y,z) = \partial L(y,z)/\partial z$.
- We show: in order to eliminate the bias, it is desirable to take

$$\begin{aligned}\xi(t,\boldsymbol{x}) &= \mu_0(t,\boldsymbol{x};\boldsymbol{\beta}^*)\\ &:= \mathbb{E}\left[L'(Y,g(T;\boldsymbol{\beta}^*))\cdot\partial g(T;\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}|T=t,\boldsymbol{X}=\boldsymbol{x}\right].\end{aligned}$$

- By Taylor's expansion,

$$\mu_0(t,\boldsymbol{x};\boldsymbol{\beta}^*) \approx \mu_0(t,\boldsymbol{x};\widehat{\boldsymbol{\beta}}_{\mathrm{DNNW}}) + \partial_{\boldsymbol{\beta}}\mu_0(t,\boldsymbol{x};\widehat{\boldsymbol{\beta}}_{\mathrm{DNNW}})(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\mathrm{DNNW}}).$$

- Thus, we consider the following instrumental function:

$$\widehat{\xi}(t,\boldsymbol{x}) = \begin{pmatrix} \widehat{\mu}(t,\boldsymbol{x};\widehat{\boldsymbol{\beta}}_{\mathrm{DNNW}}) \\ \mathrm{vec}\left\{\widehat{\partial_{\boldsymbol{\beta}}\mu}(t,\boldsymbol{x};\widehat{\boldsymbol{\beta}}_{\mathrm{DNNW}})\right\} \end{pmatrix}.$$

Intuition

- Recall

$$\mu_0(t, \boldsymbol{x}; \boldsymbol{\beta}^*) := \mathbb{E}\left[L'(Y, g(T; \boldsymbol{\beta}^*)) \cdot \partial g(T; \boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} | T = t, \boldsymbol{X} = \boldsymbol{x}\right].$$

- By definition, $\boldsymbol{\beta}^*$ satisfies $\mathbb{E}\left[\pi_0(T, \boldsymbol{X})\mu_0(T, \boldsymbol{X}; \boldsymbol{\beta}^*)\right] = \mathbf{0}$.
- The estimating function for $\widehat{\boldsymbol{\beta}}_{\mathrm{BNNW}}$ is

$$\varphi_N(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N} \widehat{\pi}_{\mathrm{BNN}}(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i; \boldsymbol{\beta}))\partial g(T_i; \boldsymbol{\beta})/\partial \boldsymbol{\beta},$$

and $\varphi_N(\widehat{\boldsymbol{\beta}}_{\mathrm{BNNW}}) = \mathbf{0}$.
- The bias

$$\begin{aligned}\mathrm{Bias} &:= \left\|\mathbb{E}\left[\varphi_N(\boldsymbol{\beta}^*)\right]\right\| \\ &= \left\|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \widehat{w}_i \widehat{\pi}_{\mathrm{DNN}}(T_i, \boldsymbol{X}_i)\mu_0(T_i, \boldsymbol{X}_i; \boldsymbol{\beta}^*)\right]\right\|.\end{aligned}$$

- We have enforced that

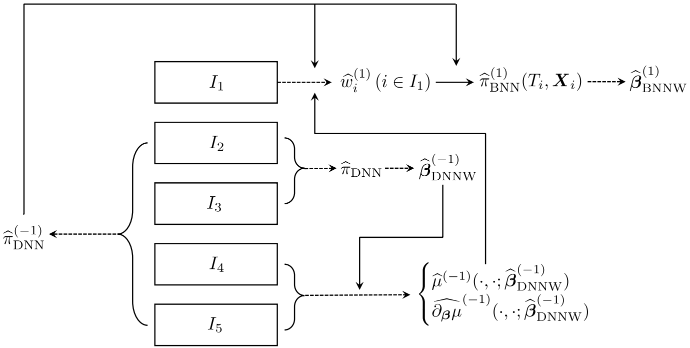
$$\frac{1}{N}\sum_{i=1}^{N}\widehat{w}_i\widehat{\pi}_{\mathrm{DNN}}(T_i,\boldsymbol{X}_i)\xi(T_i,\boldsymbol{X}_i)=\frac{1}{N(N-1)}\sum_{j=1}^{N}\sum_{l=1,l\neq j}^{N}\xi(T_j,\boldsymbol{X}_l).$$

- As a result,

$$\begin{aligned}
\mathbb{E}\left[\varphi_N(\beta^*)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\widehat{w}_i\widehat{\pi}_{\mathrm{DNN}}(T_i,\boldsymbol{X}_i)\mu_0(T_i,\boldsymbol{X}_i;\beta^*)\right]\\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\widehat{w}_i\widehat{\pi}_{\mathrm{DNN}}(T_i,\boldsymbol{X}_i)\left\{\mu_0(T_i,\boldsymbol{X}_i;\beta^*)-\mathbf{Q}\xi(T_i,\boldsymbol{X}_i)\right\}\right]\\
&\quad+\mathbb{E}\left[\frac{1}{N(N-1)}\sum_{j=1}^{N}\sum_{l=1,l\neq j}^{N}\mathbf{Q}\xi(T_j,\boldsymbol{X}_l)\right] \quad \text{Covariate Balancing of } \pi_0\\
&= \mathbb{E}\left[\widehat{\pi}_{\mathrm{BNN}}(T_i,\boldsymbol{X}_i)\left\{\mu_0(T_i,\boldsymbol{X}_i;\beta^*)-\mathbf{Q}\xi(T_i,\boldsymbol{X}_i)\right\}\right]+\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\mathbf{Q}\xi(T,\boldsymbol{X})\right]\\
&= \mathbb{E}\left[\left\{\widehat{\pi}_{\mathrm{BNN}}(T_i,\boldsymbol{X}_i)-\pi_0(T_i,\boldsymbol{X}_i)\right\}\left\{\mu_0(T_i,\boldsymbol{X}_i;\beta^*)-\mathbf{Q}\xi(T_i,\boldsymbol{X}_i)\right\}\right]
\end{aligned}$$

Cross-fitting

- Randomly partition the sample into equal parts $I_1, \ldots, I_K$.
- Take the final estimator to be $\widehat{\boldsymbol{\beta}}_{\text{BNNW}} = \sum_{k=1}^{K} \widehat{\boldsymbol{\beta}}_{\text{BNNW}}^{(k)} \Big/ K$.
- Cross-fitting procedure:



Rate doubly robust condition

- **Assumption 9.** Let $\zeta_\mu, \zeta_\pi, \zeta_u, \zeta_\beta, c^* \geq 0$ be some finite constants.
  - $\mathbb{E}\left[\sup_{\beta\in\Theta}\left\|\partial^2\mu_{0j}(T,\boldsymbol{X};\beta)/(\partial\beta\partial\beta^\top)\right\|^2\right]^{1/2} \leq c^*$.
  - Either one of the following conditions holds
    - $c^* = 0$, $\zeta_\pi + \zeta_\mu \geq 1/2$;
    - $c^* \neq 0$, $\zeta_\pi + \zeta_\mu \geq 1/2$, $\zeta_\pi + \zeta_\beta + \min\{\zeta_\beta, \zeta_u\} \geq 1/2$, $\zeta_\beta > 0$;
  - Suppose $\rho_0 > 0$ and $\rho_N = o((\log N)^{-1})$, it holds that,

$$\left\|\widehat{\mu}^{(-k)}(T,\boldsymbol{X};\widehat{\beta}^{(-k)}_{\mathrm{DNNW}}) - \mu_0(T,\boldsymbol{X};\widehat{\beta}^{(-k)}_{\mathrm{DNNW}})\right\|_{P,2} \leq \rho_N N^{-\zeta_\mu},$$

$$\left\|\widehat{\partial_\beta\mu}^{(-k)}(T,\boldsymbol{X};\widehat{\beta}^{(-k)}_{\mathrm{DNNW}}) - \partial_\beta\mu_0(T,\boldsymbol{X};\widehat{\beta}^{(-k)}_{\mathrm{DNNW}})\right\|_{P,2} \leq \rho_0 N^{-\zeta_u},$$

$$\left\|\widehat{\pi}^{(-k)}_{\mathrm{DNNW}}(T,\boldsymbol{X}) - \pi_0(T,\boldsymbol{X})\right\|_{P,2} \leq \rho_N N^{-\zeta_\pi}, \quad \left\|\widehat{\beta}^{(-k)}_{\mathrm{DNNW}} - \beta^*\right\| \leq \rho_0 N^{-\zeta_\beta}.$$

- For binary ATE, $c^* = 0$ and ζ_u can be arbitrarily large. Reduces to $\zeta_\pi + \zeta_\mu \geq 1/2$.
- For binary QTE, less restrictive than $\zeta_\pi + \zeta_\beta \geq 1/2$, required in Kallus et al. (2024, JMLR).

Asymptotic Normal and Inference

Under some regularity conditions, and the rate-doubly robust assumption, we have

$$\widehat{\boldsymbol{\beta}}_{\mathrm{BNNW}} - \boldsymbol{\beta}^* = -\frac{1}{N}\sum_{i=1}^{N} \Sigma_0^{-1}\psi(Y_i, T_i, \boldsymbol{X}_i; \boldsymbol{\beta}^*) + o_P(N^{-1/2}),$$

and

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{BNNW}} - \boldsymbol{\beta}^*\right) \xrightarrow{d} N(\mathbf{0}, V_{\mathrm{eff}}),$$

where $\Sigma_0 = \partial\mathbb{E}\left[\pi_0(T, \boldsymbol{X})\mu_0(T, \boldsymbol{X}; \boldsymbol{\beta})\right] \Big/ \partial\boldsymbol{\beta}^\top\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is the Hessian matrix, $V_{\mathrm{eff}} := \Sigma_0^{-1}\mathbb{E}\left[\psi\psi^\top\right]\Sigma_0^{-1}$ attains the semiparametric efficiency bound.

- We propose a weighted bootstrap method for inference.
- The inference method is theoretically validated.

Binary Treatment

- DGP-B:

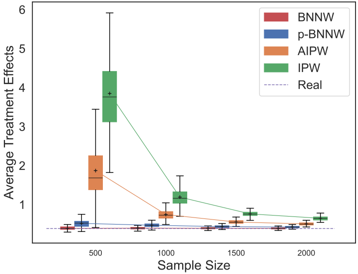
$$X_j = 0.3 + 0.4U_{xj},\ j = 1, \ldots, d \ \text{ and } \ T = 1(U_t < \boldsymbol{X}^\top \gamma_b);$$

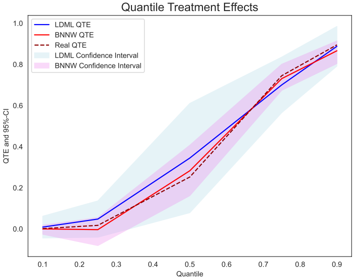
$$Y(0) = 1(U_{y_0} \leq \boldsymbol{X}^\top \gamma_b) \frac{U_{y_0}^2}{\boldsymbol{X}^\top \gamma_b} + 1(U_{y_0} > \boldsymbol{X}^\top \gamma_b) U_{y_0};$$

$$Y(1) = 1(U_{y_1} \leq 1 - \boldsymbol{X}^\top \gamma_b) \frac{2U_{y_1}^2}{1 - \boldsymbol{X}^\top \gamma_b} + 1(U_{y_0} > 1 - \boldsymbol{X}^\top \gamma_b) 2U_{y_0}.$$

Binary Treatment

- Left panel: Box plots of the estimated ATE using different methods over 100 replications of each sample size setting with $d = 50$. Right panel: Estimated QTE and the respective 95% confidence interval across different quantile levels on a simulated data set with $N = 2000$ and $d = 50$.





Binary Treatment (ATE)

Table: Bias, SE, RMSE of estimated ATE and coverage probability (CP), average width (AW) of the respective 95% confidence interval, using BNNW or AIPW estimators over 100 replications of simulated dataset with a binary treatment across different sample sizes.

	AIPW (Chernozhukov et al., 2018)					**BNNW**				
N	Bias	SE	RMSE	CP	AW	Bias	SE	RMSE	CP	AW
300	0.0781	0.070	0.0849	0.69	0.235	0.0022	0.061	0.0512	0.98	0.258
500	0.0670	0.048	0.0698	0.71	0.187	0.0040	0.052	0.0406	0.96	0.202
1000	0.0349	0.036	0.0419	0.82	0.140	0.0094	0.033	0.0278	0.95	0.144
5000	0.0097	0.017	0.0157	0.91	0.065	0.0049	0.017	0.0144	0.94	0.065

Binary Treatment (QTE)

N	τ	LDML (Kallus et al., 2024)					**BNNW**				
		Bias	SE	RMSE	CP	AW	Bias	SE	RMSE	CP	AW
300	0.1	0.0064	0.073	0.0729	0.97	0.410	0.0219	0.038	0.0440	0.98	0.225
	0.25	0.0235	0.189	0.1905	0.91	0.646	0.0123	0.087	0.0874	0.97	0.410
	0.5	0.0204	0.302	0.3023	0.82	1.020	0.0413	0.137	0.1435	1.00	0.579
	0.75	0.0055	0.210	0.2102	0.95	0.775	0.0280	0.098	0.1022	0.97	0.464
	0.9	0.0167	0.154	0.1550	0.99	0.972	0.0391	0.063	0.0739	0.96	0.322
500	0.1	0.0115	0.043	0.0443	0.99	0.292	0.0033	0.022	0.0223	0.98	0.131
	0.25	0.0086	0.137	0.1377	0.88	0.458	0.0059	0.071	0.0712	0.96	0.297
	0.5	0.0152	0.242	0.2424	0.88	0.888	0.0239	0.119	0.1209	0.96	0.457
	0.75	0.0126	0.192	0.1925	0.89	0.616	0.0017	0.075	0.0750	0.98	0.342
	0.9	0.0006	0.117	0.1170	0.98	0.570	0.0012	0.045	0.0454	0.99	0.225
1000	0.1	0.0062	0.021	0.0219	1.00	0.142	0.0065	0.017	0.0183	0.90	0.073
	0.25	0.0122	0.076	0.0772	0.94	0.241	0.0119	0.043	0.0447	0.95	0.196
	0.5	0.0077	0.161	0.1608	0.88	0.554	0.0299	0.076	0.0813	0.97	0.332
	0.75	0.0153	0.077	0.0789	0.93	0.325	0.0030	0.047	0.0468	1.00	0.211
	0.9	0.0077	0.051	0.0519	0.95	0.235	0.0051	0.033	0.0331	0.96	0.147

Continuous Treatment

The generating process for the dosages, treatments and responses of IHDP-continuous dataset are,

$$
\begin{aligned}
\tilde{T} &= \frac{3X_1}{1+X_2} + \frac{3\max\{X_3, X_4, X_5\}}{0.2+\min\{X_3, X_4, X_5\}} + 3\tanh\left(\frac{5\sum_{j\in\mathcal{J}_1} X_j}{|\mathcal{J}_1|}\right) - 6 + \epsilon_1 \\
T &= (1+\exp(\tilde{T}))^{-1}, \\
Y^*(t) &= h(t, \boldsymbol{X}) + \epsilon_2, \\
h(t, \boldsymbol{X}) &= (-0.8 + 3.2t - 3.2t^2)\Bigg\{ \tanh\left(5\frac{\sum_{j\in\mathcal{J}_2} X_j}{|\mathcal{J}_2|}\right) + \\
&\qquad 3\exp\left(\frac{0.2(X_1 - X_5)}{0.1+\min\{X_2, X_3, X_4\}}\right)\Bigg\},
\end{aligned}
$$

where ϵ_1 and ϵ_2 are two independent standard normal random variables, $\boldsymbol{X} = (X_1, \ldots, X_{25})^\top$, $\mathcal{J}_1 = \{3, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$ and $\mathcal{J}_2 = \{15, 16, 17, 18, 19, 20, 21, 22, 23, 24\}$.

Continuous Treatment

Table: The ABias, ASE and ARMSE of the estimated average dosage responses and quantile dosage responses for various quantiles τ over 100 replications of the semi-synthetic IHDP dataset with a continuous treatment.

		ADRF	QDRF				
			$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
	ABias	0.215	0.391	0.198	0.203	0.198	0.212
BNNW	ASE	0.103	0.329	0.153	0.109	0.105	0.124
	ARMSE	0.247	0.524	0.256	0.236	0.231	0.255
	ABias	0.619	1.173	0.516	0.326	0.254	0.226
DNNW	ASE	0.078	0.251	0.131	0.098	0.094	0.119
	ARMSE	0.628	1.211	0.539	0.347	0.278	0.266
	ABias	0.605	1.405	0.643	0.295	0.253	0.206
GOE	ASE	0.398	0.924	0.591	0.500	0.545	0.725
(Sieve)	ARMSE	0.736	1.703	0.885	0.586	0.605	0.754

Thank You!