# Learning treatment effects under covariate dependent left truncation and right censoring

Yuyao Wang[1], Andrew Ying and Ronghui Xu[12]

[1]Department of Mathematics, [2]Herbert Wertheim School of Public Health and Halicioglu Data Science Institute, UC San Diego

UC San Diego Mathematics

## Introduction: Selection Bias from Left Truncation

- Outcome of interest: time-to-event ($T^*$)
- $T^*$ is **left truncated** by the enrollment time ($Q^*$) if only subjects with $T^* > Q^*$ are included in the data. $\Rightarrow$ **Selection bias**.
  - Usually present in studies with *delayed entry*.
    e.g., aging studies, pregnancy studies, cancer survivorship studies.

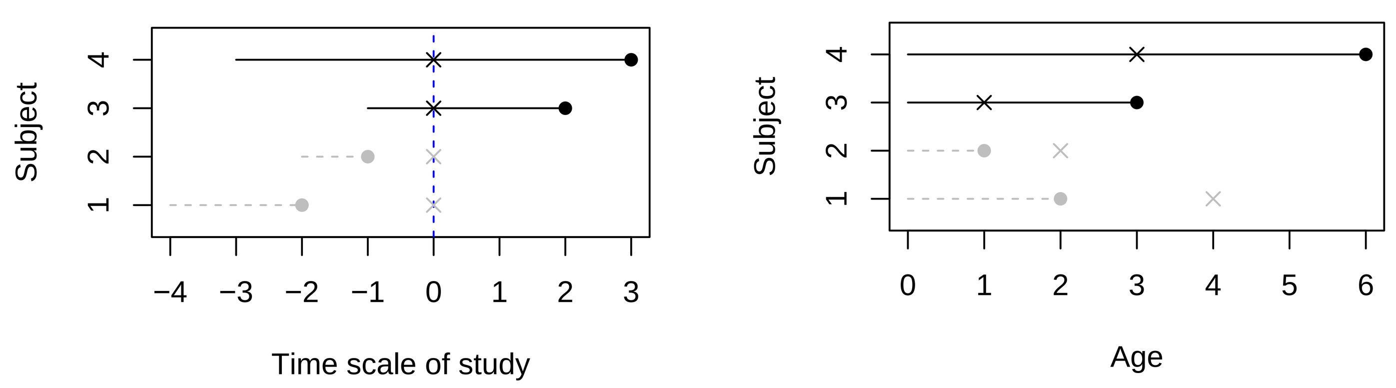



Figure: A toy example of aging study: people's lifespans on the time scale of study (left) and on the age scale (right). Solid dots: event times; '×': enrollment times.

- **Triple biases**: confounding, selection bias from left truncation, informative right censoring.
- **Estimand (CATE)**: $\tau(v) = \mathbb{E}[\nu\{T^*(1)\} - \nu\{T^*(0)\} \mid V^* = v]$.
- Notation: with '$*$' – truncation-free data; without '$*$' – truncated data.
- $C$: censoring time; $D = C - Q$: residual censoring time.
- Assumptions: $Q^* \perp\!\!\!\perp T^* \mid A^*, Z^*$; $D \perp\!\!\!\perp T \mid Q, A, Z$.

## Method

For any $\zeta = \zeta(T^*, A^*, Z^*)$ and $\varphi = \varphi(Q, T, A, Z)$,

- $\mathscr{V}_Q(\zeta; F, G) =$

$$\underbrace{\frac{\zeta(T,A,Z)}{G(T|A,Z)}}_{\text{IPW}} - \underbrace{\int_0^\infty m_\zeta(v, A, Z; F) \cdot \frac{F(v|A,Z)}{1 - F(v|A,Z)} \cdot \frac{d\bar{M}_Q(v;G)}{G(v|A,Z)}}_{\text{Augmentation}}$$

- $\mathscr{V}_C(\varphi; F, S_D) =$

$$\underbrace{\frac{\Delta\, \varphi(Q,X,A,Z)}{S_D(X-Q|Q,A,Z)}}_{\text{IPCW}} + \underbrace{\int_0^\infty \bar{m}_\varphi(u, Q, A, Z; F) \cdot \frac{dM_D(u;S_D)}{S_D(u|Q,A,Z)}}_{\text{Augmentation}}$$

  - $F$: conditional CDF of $T^* \mid A^*, Z^*$; $G$: conditional CDF of $Q^* \mid A^*, Z^*$.
  - $S_D$: conditional survival function of $D \mid Q, A, Z$.
  - $m_\zeta(v, a, z; \theta, F) = \mathbb{E}\{\zeta(T^*, A^*, Z^*; \theta) \mid T^* \le v, A^* = a, Z^* = z\}$.
  - $\bar{m}_\varphi(u, q, a, z; F) = \mathbb{E}\{\varphi(Q, T, A, Z) | T - Q \ge u, Q = q, A = a, Z = z\} = \int_{q+u}^\infty \varphi(q, t, a, z) dF(t|a,z)/\{1 - F(q+u|a,z)\}$.

- **AIPW operator** for handling left truncation and right censoring (LTRC):

$$\mathscr{V}(F, G, S_D) = \mathscr{V}_C(F, S_D) \circ \mathscr{V}_Q(F, G).$$

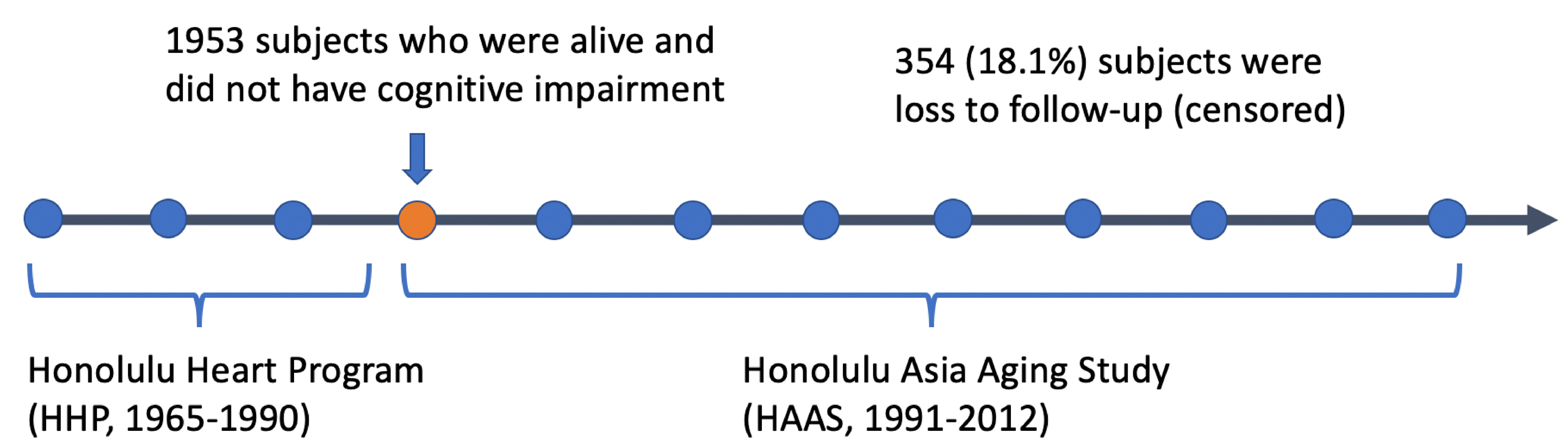
### Double Robustness and Neyman Orthogonality

- **Double robustness (DR)**: $\mathbb{E}\{\mathscr{V}(\zeta; F, G, S_D)\} = \beta^{-1}\mathbb{E}(\zeta)$ if either $F = F_0$ or $(G, S_D) = (G_0, S_{D0})$, where $\beta = \mathbb{P}(Q^* < T^*)$.
- **Neyman orthogonality**: $\mathscr{V}$ preserves Neyman orthogonality.

## Orthogonal and Doubly Robust Learners

- R-loss $\xrightarrow{\mathscr{V}}$ ltrcR-loss (Neyman orthogonal); DR-loss $\xrightarrow{\mathscr{V}}$ ltrcDR-loss (doubly robust).
- Two-stage algorithm with cross fitting: 1) Estimate nuisance parameters; 2) Empirical risk minimization with the estimated nuisance parameters plugged in.

## HHP-HAAS data




- Question: What is the impact of midlife alcohol consumption on late-life cognitive impairment?
- $T^*$ - age to moderate cognitive impairment or death; $Q^*$ - age at entry of HAAS.
- Baseline covariates: Education, ApoE genotype, systolic blood pressure (SBP), heart rate (HR).

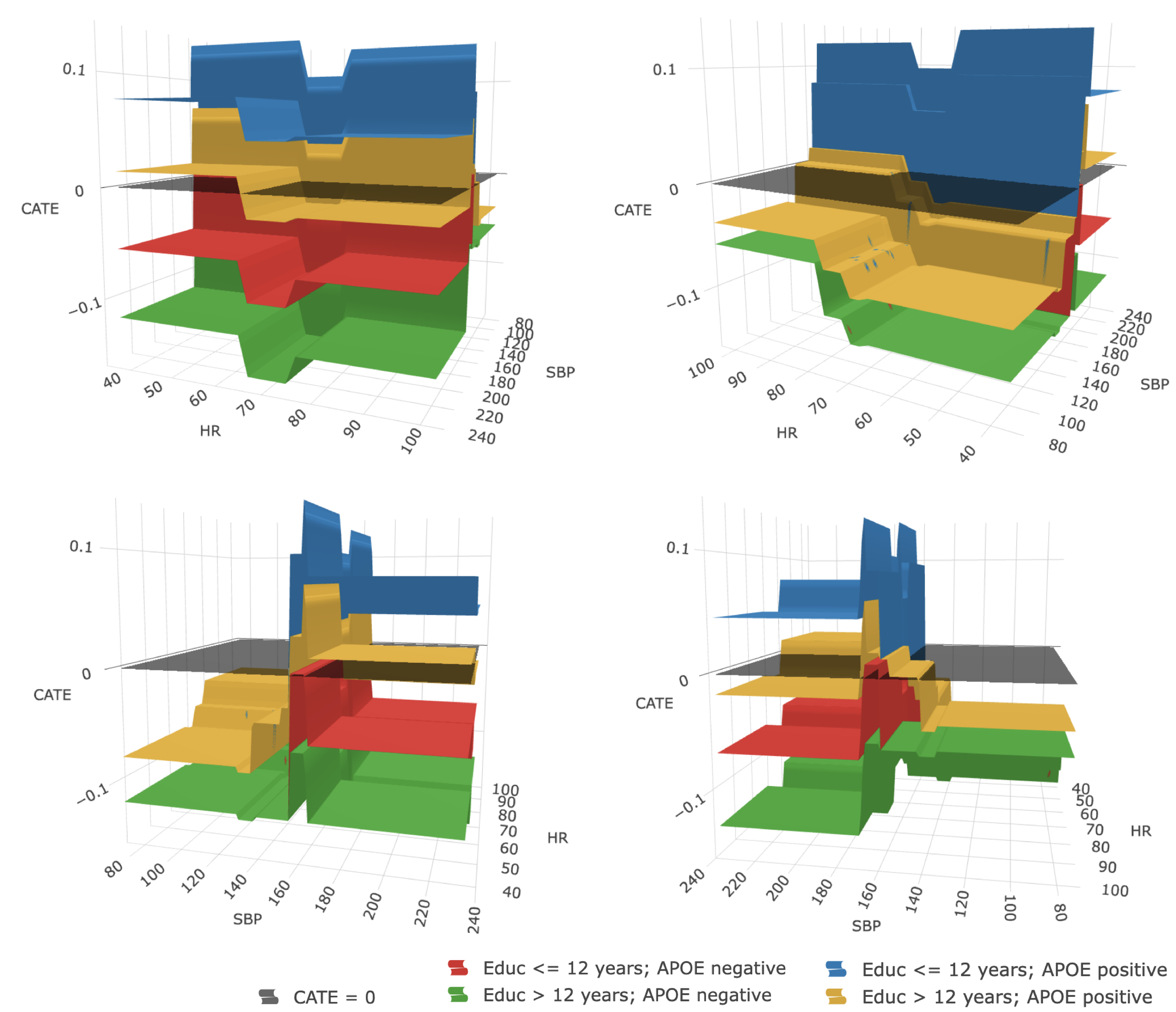



Figure: Estimated CATE surfaces from ltrcR-learner for cognitive-impairment-free survival at age 90 across the four education and ApoE genotype subgroups (views from four different angles); the estimated CATE surface for the two education groups overlaps for SBP < 158 mmHg. The spike of the CATE surfaces appear at SBP being 158 - 171 mmHg.

## Simulation Results

- Truncation rate: around 28%; censoring rate: around 50%; treatment rate: 50%.
- $\text{MSE} = \frac{1}{n}\sum_{i=1}^n \{\hat{\tau}(V_i) - \tau_0(V_i)\}^2$

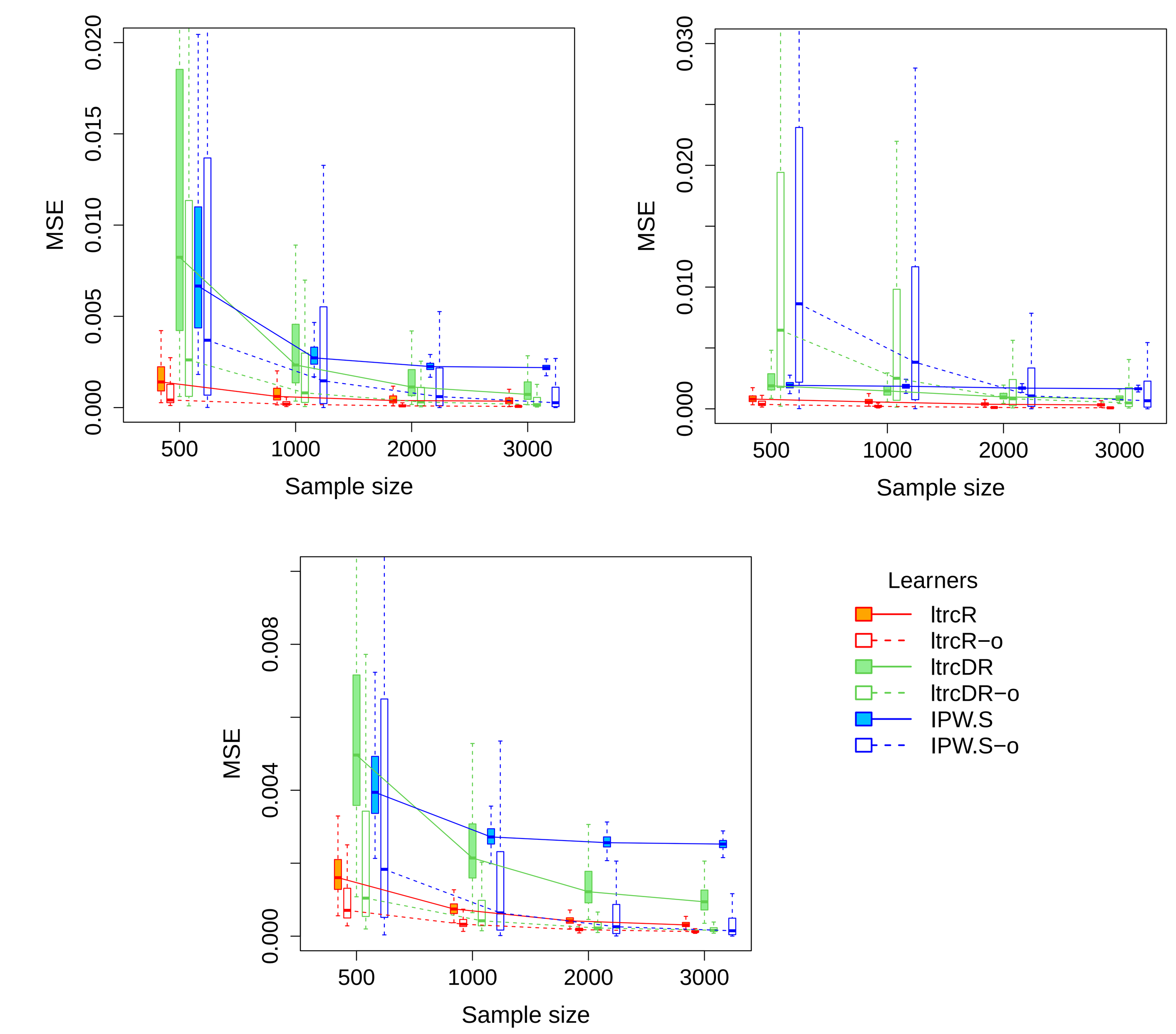



Figure: MSE's for different learners cross 500 simulated datasets under different scenarios; '-o' indicates the oracle learner with the true nuisance parameters.

### Oracle Rate Results

- **Error bound**: With probability at least $1 - \delta$,

$$\|\hat{\tau} - \tau_0\|_2^2 \le c_2(n) \cdot r_2(\mathscr{T}, \delta/2; \hat{\pi}, \hat{F}, \hat{G}, \hat{S}_D) + c_1(n) \cdot r_1(\mathscr{G}, \delta/2),$$

  where $r_1$: the error bound for the (integral) product estimation errors between $F$ and $(\pi, G, S_D)$;
  $r_2$: the excess risk bound of the second stage learning algorithm.
- **An oracle result**: $\|\hat{\tau} - \tau_0\|_2 = O_p(\delta_n^* + n^{-1/2} + a_n)$,
  where $\delta_n^*$: the critical radius of the second stage function class;
  $a_n$: the rate of the (integral) product estimation errors between $F$ and $(\pi, G, S_D)$.

## References and Contact Information

- Yuyao Wang, Andrew Ying, Ronghui Xu. (2024) Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika*, 111(3), 789-808.
- Yuyao Wang, Andrew Ying, Ronghui Xu. (2024) Learning treatment effects under covariate dependent left truncation and right censoring. *arXiv:2411.18879*.
- GitHub Repo: https://github.com/wangyuyao98/truncAC.
- R package: `truncAIPW`.

- Contact info: yuw079@ucsd.edu (Yuyao Wang).
  aying9339@gmail.com (Andrew Ying)
  rxu@ucsd.edu (Ronghui Xu)