

Doubly Robust Inference on Causal Derivative Effects for Continuous Treatments

Yikun Zhang

Joint work with *Professor Yen-Chi Chen*

*Department of Statistics,
University of Washington*

ENAR 2025 Spring Meeting
March 25, 2025

The Role of Derivatives in Causal Inference

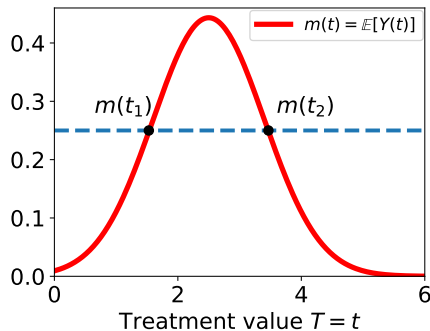
Goal: Study the causal effect of a treatment $T \in \mathcal{T}$ on an outcome of interest $Y \in \mathcal{Y}$.

- $\mathbb{E}[Y(t)]$ = mean potential outcome under a static intervention $T = t$.

The Role of Derivatives in Causal Inference

Goal: Study the causal effect of a treatment $T \in \mathcal{T}$ on an outcome of interest $Y \in \mathcal{Y}$.

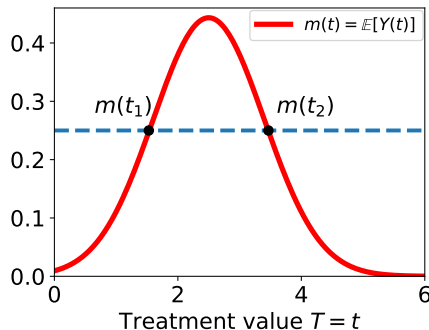
- $\mathbb{E}[Y(t)]$ = mean potential outcome under a static intervention $T = t$.
- When t varies in a continuous space, $t \mapsto \mathbb{E}[Y(t)] := m(t)$ is a curve!



The Role of Derivatives in Causal Inference

Goal: Study the causal effect of a treatment $T \in \mathcal{T}$ on an outcome of interest $Y \in \mathcal{Y}$.

- $\mathbb{E}[Y(t)]$ = mean potential outcome under a static intervention $T = t$.
- When t varies in a continuous space, $t \mapsto \mathbb{E}[Y(t)] := m(t)$ is a curve!



- While $m(t_1) = m(t_2)$, the derivative effects $m'(t_1)$, $m'(t_2)$ are distinct!
- The **derivative effect curve** $\theta(t) = m'(t) = \frac{d}{dt}\mathbb{E}[Y(t)]$ is a continuous generalization to the average treatment effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

Our causal estimand of interest is the **derivative effect curve**

$$t \mapsto \theta(t) = m'(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for } t \in \mathcal{T} \subset \mathbb{R}.$$

Our causal estimand of interest is the **derivative effect curve**

$$t \mapsto \theta(t) = m'(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for } t \in \mathcal{T} \subset \mathbb{R}.$$

Problem: $\theta(t)$ is non-regular and cannot be estimated in the rate $1/\sqrt{n}$.

Our causal estimand of interest is the **derivative effect curve**

$$t \mapsto \theta(t) = m'(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for } t \in \mathcal{T} \subset \mathbb{R}.$$

Problem: $\theta(t)$ is non-regular and cannot be estimated in the rate $1/\sqrt{n}$.

There are some closely related but distinct estimands:

- *Incremental Causal/Treatment Effect* ([Kennedy, 2019](#); [Rothenhäusler and Yu, 2019](#)):

$$\mathbb{E}[Y(T + \delta)] - \mathbb{E}[Y(T)] \quad \text{for some deterministic } \delta > 0.$$

Our causal estimand of interest is the **derivative effect curve**

$$t \mapsto \theta(t) = m'(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for } t \in \mathcal{T} \subset \mathbb{R}.$$

Problem: $\theta(t)$ is non-regular and cannot be estimated in the rate $1/\sqrt{n}$.

There are some closely related but distinct estimands:

- *Incremental Causal/Treatment Effect* (Kennedy, 2019; Rothenhäusler and Yu, 2019):

$$\mathbb{E}[Y(T + \delta)] - \mathbb{E}[Y(T)] \quad \text{for some deterministic } \delta > 0.$$

- *Average Derivative/Partial Effect* (Powell et al., 1989; Newey and Stoker, 1993):

$$\mathbb{E}[\theta(T)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}(Y|T, S) \right], \quad \text{where } S \in \mathcal{S} \subset \mathbb{R}^d \text{ is a covariate vector.}$$

Our causal estimand of interest is the **derivative effect curve**

$$t \mapsto \theta(t) = m'(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for } t \in \mathcal{T} \subset \mathbb{R}.$$

Problem: $\theta(t)$ is non-regular and cannot be estimated in the rate $1/\sqrt{n}$.

There are some closely related but distinct estimands:

- *Incremental Causal/Treatment Effect* (Kennedy, 2019; Rothenhäusler and Yu, 2019):

$$\mathbb{E}[Y(T + \delta)] - \mathbb{E}[Y(T)] \quad \text{for some deterministic } \delta > 0.$$

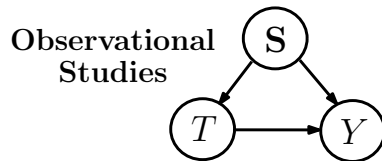
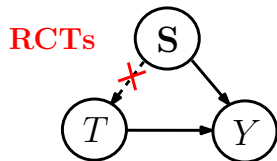
- *Average Derivative/Partial Effect* (Powell et al., 1989; Newey and Stoker, 1993):

$$\mathbb{E}[\theta(T)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}(Y|T, S) \right], \quad \text{where } S \in \mathcal{S} \subset \mathbb{R}^d \text{ is a covariate vector.}$$

Pros These estimands may have more realistic interpretations in the actual context.

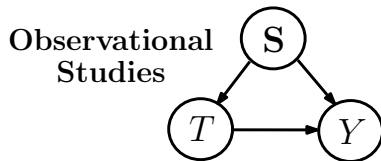
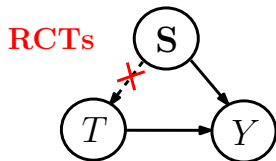
Cons They quantify only the overall causal effects, not those at a specific level of interest.

Identification Assumptions with Observational Data



¹Some mild interchangeability assumptions are needed; see Theorem 1.1 in [Shao \(2003\)](#).

Identification Assumptions with Observational Data

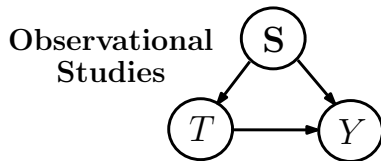
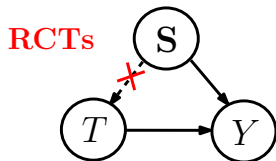


Assumption (Identification Conditions)

- 1 (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- 2 (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- 3 (**Positivity**) The conditional density satisfies $p_{T|S}(t|s) \geq p_{\min} > 0$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

¹Some mild interchangeability assumptions are needed; see Theorem 1.1 in [Shao \(2003\)](#).

Identification Assumptions with Observational Data



Assumption (Identification Conditions)

- 1 (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- 2 (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- 3 (**Positivity**) The conditional density satisfies $p_{T|S}(t|s) \geq p_{\min} > 0$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

$$\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \stackrel{(*)^1}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}(Y|T = t, S) \right].$$

- The positivity condition is required for $\frac{\partial}{\partial t} \mu(t, s) = \frac{\partial}{\partial t} \mathbb{E}(Y|T = t, S = s)$ to be well-defined on $\mathcal{T} \times \mathcal{S}$.

¹Some mild interchangeability assumptions are needed; see Theorem 1.1 in [Shao \(2003\)](#).

An Example of the Positivity Violation

Assumption (Positivity Condition)

There exists a constant $p_{\min} > 0$ such that $p_{T|S}(t|s) \geq p_{\min}$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

- Positivity is a very strong assumption with continuous treatments!

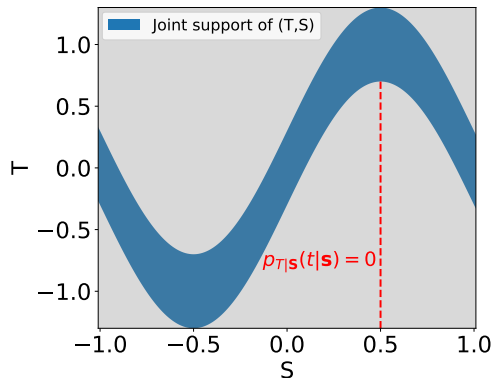
An Example of the Positivity Violation

Assumption (Positivity Condition)

There exists a constant $p_{\min} > 0$ such that $p_{T|S}(t|s) \geq p_{\min}$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

► Positivity is a very strong assumption with continuous treatments!

$$T = \sin(\pi S) + E, \quad E \sim \text{Uniform}[-0.3, 0.3], \quad S \sim \text{Uniform}[-1, 1], \quad \text{and} \quad E \perp\!\!\!\perp S.$$



Note that $p_{T|S}(t|s) = 0$ in the gray regions, and the positivity condition fails.

Highlights of Today's Talk

$$t \mapsto m(t) = \mathbb{E}[Y(t)] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for} \quad t \in \mathcal{T}.$$

Under the positivity condition:

- 1 Propose a doubly robust (DR) estimator of $\theta(t)$ via kernel smoothing.

Highlights of Today's Talk

$$t \mapsto m(t) = \mathbb{E}[Y(t)] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for} \quad t \in \mathcal{T}.$$

Under the positivity condition:

- 1 Propose a doubly robust (DR) estimator of $\theta(t)$ via kernel smoothing.

$$\text{Regression Adjustment (RA) + Inverse Probability Weighting (IPW)} \left\{ \begin{array}{l} \Rightarrow \\ \nRightarrow \end{array} \right. \text{DR}$$

$$t \mapsto m(t) = \mathbb{E}[Y(t)] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for} \quad t \in \mathcal{T}.$$

Under the positivity condition:

- 1 Propose a doubly robust (DR) estimator of $\theta(t)$ via kernel smoothing.

$$\text{Regression Adjustment (RA) + Inverse Probability Weighting (IPW)} \left\{ \begin{array}{l} \Rightarrow \\ \nRightarrow \end{array} \right. \text{DR}$$

Without the positivity condition:

- 2 $m(t)$ and $\theta(t)$ are identifiable with an additive structural assumption:

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon. \tag{1}$$

$$t \mapsto m(t) = \mathbb{E}[Y(t)] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for} \quad t \in \mathcal{T}.$$

Under the positivity condition:

- ① Propose a doubly robust (DR) estimator of $\theta(t)$ via kernel smoothing.

$$\text{Regression Adjustment (RA)} + \text{Inverse Probability Weighting (IPW)} \left\{ \begin{array}{l} \Rightarrow \\ \nRightarrow \end{array} \right. \text{DR}$$

Without the positivity condition:

- ② $m(t)$ and $\theta(t)$ are identifiable with an additive structural assumption:

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon. \tag{1}$$

- ③ The usual IPW estimators of $m(t)$ and $\theta(t)$ are still *biased* even under model (1).

Highlights of Today's Talk

$$t \mapsto m(t) = \mathbb{E}[Y(t)] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] \quad \text{for} \quad t \in \mathcal{T}.$$

Under the positivity condition:

- 1 Propose a doubly robust (DR) estimator of $\theta(t)$ via kernel smoothing.

$$\text{Regression Adjustment (RA)} + \text{Inverse Probability Weighting (IPW)} \left\{ \begin{array}{l} \Rightarrow \\ \nRightarrow \end{array} \right. \text{DR}$$

Without the positivity condition:

- 2 $m(t)$ and $\theta(t)$ are identifiable with an additive structural assumption:

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon. \tag{1}$$

- 3 The usual IPW estimators of $m(t)$ and $\theta(t)$ are still *biased* even under model (1).
- 4 Propose our bias-corrected IPW and DR estimators for $m(t)$ and $\theta(t)$.
 - Has a novel connection to nonparametric support and level set estimation problems.

Nonparametric Inference on $\theta(t)$ Under Positivity



Assumption (Identification Conditions)

- 1 (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- 2 (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- 3 (**Positivity**) The conditional density satisfies $p_{T|S}(t|s) \geq p_{\min} > 0$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

Given that $\mu(t, s) = \mathbb{E}(Y|T = t, S = s)$, we have

RA or G-computation:
$$\begin{cases} m(t) = \mathbb{E}[Y(t)] = \mathbb{E}[\mu(t, S)], \\ \theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \frac{d}{dt}\mathbb{E}[\mu(t, S)] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, S)\right]. \end{cases}$$

Assumption (Identification Conditions)

- ① (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- ② (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- ③ (**Positivity**) The conditional density satisfies $p_{T|S}(t|s) \geq p_{\min} > 0$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

Given that $\mu(t, s) = \mathbb{E}(Y|T = t, S = s)$, we have

RA or G-computation:
$$\begin{cases} m(t) = \mathbb{E}[Y(t)] = \mathbb{E}[\mu(t, S)], \\ \theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \frac{d}{dt}\mathbb{E}[\mu(t, S)] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, S)\right]. \end{cases}$$

IPW:
$$\begin{cases} m(t) = \mathbb{E}[Y(t)] = \lim_{h \rightarrow 0} \mathbb{E}\left[\frac{Y}{p_{T|S}(T|S)} \cdot \frac{1}{h}K\left(\frac{T-t}{h}\right)\right], \\ \theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = ??? \end{cases}$$

- $K : \mathbb{R} \rightarrow [0, \infty)$ is a kernel function, e.g., $K(u) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) & \text{(Gaussian),} \\ \frac{3}{4}(1 - u^2) \cdot \mathbb{1}_{\{|u| \leq 1\}} & \text{(Parabolic).} \end{cases}$
- $h > 0$ is a smoothing bandwidth parameter.

Dose-Response Curve Estimation Under Positivity

Given the observed data $\{(Y_i, T_i, S_i)\}_{i=1}^n$, there are three main strategies for estimating

$$m(t) = \mathbb{E}[Y(t)] = \mathbb{E}[\mu(t, S)] = \lim_{h \rightarrow 0} \mathbb{E} \left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p_{T|S}(T|S)} \right].$$

① **RA Estimator** (Robins, 1986; Gill and Robins, 2001):

$$\hat{m}_{\text{RA}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{\mu}(t, S_i).$$

② **IPW Estimator** (Hirano and Imbens, 2004; Imai and van Dyk, 2004):

$$\hat{m}_{\text{IPW}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)} \cdot Y_i.$$

③ **DR Estimator** (Kallus and Zhou, 2018; Colangelo and Lee, 2020):

$$\hat{m}_{\text{DR}}(t) = \frac{1}{nh} \sum_{i=1}^n \left\{ \frac{K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)} \cdot [Y_i - \hat{\mu}(t, S_i)] + h \cdot \hat{\mu}(t, S_i) \right\}.$$

To estimate $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\partial}{\partial t} \mu(t, S)\right]$ from $\{(Y_i, T_i, S_i)\}_{i=1}^n$, we could also have three strategies:

1 RA Estimator:

$$\hat{\theta}_{\text{RA}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i) \quad \text{with} \quad \beta(t, s) = \frac{\partial}{\partial t} \mu(t, s).$$

Question: How can we generalize the IPW form $m(t) = \lim_{h \rightarrow 0} \mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p_{T|S}(T|S)}\right]$ to identify and estimate $\theta(t)$?

To estimate $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\partial}{\partial t} \mu(t, S)\right]$ from $\{(Y_i, T_i, S_i)\}_{i=1}^n$, we could also have three strategies:

1 RA Estimator:

$$\hat{\theta}_{\text{RA}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i) \quad \text{with} \quad \beta(t, s) = \frac{\partial}{\partial t} \mu(t, s).$$

Question: How can we generalize the IPW form $m(t) = \lim_{h \rightarrow 0} \mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p_{T|S}(T|S)}\right]$ to identify and estimate $\theta(t)$?

2 IPW Estimator: Inspired by the derivative estimator in [Mack and Müller \(1989\)](#), we propose

$$\hat{\theta}_{\text{IPW}}(t) = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{h^2 \cdot \kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} \quad \text{with} \quad \kappa_2 = \int u^2 \cdot K(u) du.$$

Doubly Robust Estimator for $\theta(t)$ Under Positivity

Recall that $\hat{m}_{\text{DR}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)} \cdot [Y_i - \hat{\mu}(t, S_i)] + \frac{1}{n} \sum_{i=1}^n \hat{\mu}(t, S_i).$

$$\hat{\theta}_{\text{RA}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i) \quad \text{"+"} \quad \hat{\theta}_{\text{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} \cdot Y_i \quad \Rightarrow$$

$$\hat{\theta}_{\text{DR}}(t) = \underbrace{\frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} \left[Y_i - \hat{\mu}(t, S_i) - (T_i - t) \cdot \hat{\beta}(t, S_i) \right]}_{\text{New IPW component}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i)}_{\text{RA component}}.$$

Doubly Robust Estimator for $\theta(t)$ Under Positivity

Recall that $\hat{m}_{\text{DR}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)} \cdot [Y_i - \hat{\mu}(t, S_i)] + \frac{1}{n} \sum_{i=1}^n \hat{\mu}(t, S_i)$.

$$\hat{\theta}_{\text{RA}}(t) = \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i) \quad \text{“+”} \quad \hat{\theta}_{\text{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} \cdot Y_i \quad \Rightarrow$$

$$\hat{\theta}_{\text{DR}}(t) = \underbrace{\frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} [Y_i - \hat{\mu}(t, S_i) - (T_i - t) \cdot \hat{\beta}(t, S_i)]}_{\text{New IPW component}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i)}_{\text{RA component}}.$$

The “New IPW component” leverages a local polynomial approximation to push the residual of the IPW component to (roughly) second order.

- Neyman orthogonality (Neyman, 1959; Chernozhukov et al., 2018) holds for this form of $\hat{\theta}_{\text{DR}}(t)$ as $h \rightarrow 0$.

Theorem (Theorem 1 in Zhang and Chen 2025)

Under some regularity assumptions and

- ① $\hat{\mu}, \hat{\beta}, \hat{p}_{T|S}$ are estimated on a dataset independent of $\{(Y_i, T_i, S_i)\}_{i=1}^n$;
- ② at least one of the model specification conditions hold:
 - $\hat{p}_{T|S}(t|s) \xrightarrow{P} \bar{p}_{T|S}(t|s) = p_{T|S}(t|s)$ (**conditional density model**),
 - $\hat{\mu}(t, s) \xrightarrow{P} \bar{\mu}(t, s) = \mu(t, s)$ and $\hat{\beta}(t, s) \xrightarrow{P} \bar{\beta}(t, s) = \beta(t, s)$ (**outcome model**);
- ③ $\sup_{|u-t| \leq h} \left\| \hat{p}_{T|S}(u|S) - p_{T|S}(u|S) \right\|_{L_2} \left[\left\| \hat{\mu}(t, S) - \mu(t, S) \right\|_{L_2} + h \left\| \hat{\beta}(t, S) - \beta(t, S) \right\|_{L_2} \right] = o_P \left(\frac{1}{\sqrt{nh}} \right),$

we prove that

Theorem (Theorem 1 in Zhang and Chen 2025)

Under some regularity assumptions and

- ① $\hat{\mu}, \hat{\beta}, \hat{p}_{T|S}$ are estimated on a dataset independent of $\{(Y_i, T_i, S_i)\}_{i=1}^n$;
- ② at least one of the model specification conditions hold:
 - $\hat{p}_{T|S}(t|s) \xrightarrow{P} \bar{p}_{T|S}(t|s) = p_{T|S}(t|s)$ (**conditional density model**),
 - $\hat{\mu}(t, s) \xrightarrow{P} \bar{\mu}(t, s) = \mu(t, s)$ and $\hat{\beta}(t, s) \xrightarrow{P} \bar{\beta}(t, s) = \beta(t, s)$ (**outcome model**);
- ③ $\sup_{|u-t| \leq h} \left\| \hat{p}_{T|S}(u|S) - p_{T|S}(u|S) \right\|_{L_2} \left[\left\| \hat{\mu}(t, S) - \mu(t, S) \right\|_{L_2} + h \left\| \hat{\beta}(t, S) - \beta(t, S) \right\|_{L_2} \right] = o_P \left(\frac{1}{\sqrt{nh}} \right),$

we prove that

- $\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{h,t} \left(Y_i, T_i, S_i; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S} \right) + o_P(1).$
- $\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N} \left(0, V_{\theta}(t) \right).$

An asymptotically valid inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$ can be conducted through

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

An asymptotically valid inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$ can be conducted through

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

① We estimate $V_{\theta}(t) = \mathbb{E} \left[\phi_{h,t}^2(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S}) \right]$ with

$$\phi_{h,t}(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S}) = \frac{\left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right)}{\sqrt{h} \cdot \kappa_2 \cdot \bar{p}_{T|S}(T|S)} \cdot [Y - \bar{\mu}(t, S) - (T-t) \cdot \bar{\beta}(t, S)]$$

$$\text{by } \hat{V}_{\theta}(t) = \frac{1}{n} \sum_{i=1}^n \phi_{h,t}^2(Y_i, T_i, S_i; \hat{\mu}, \hat{\beta}, \hat{p}_{T|S}).$$

Statistical Inference on $\theta(t)$

An asymptotically valid inference on $\theta(t) = \frac{d}{dt} \mathbb{E} [Y(t)]$ can be conducted through

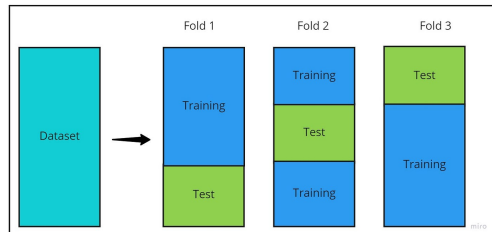
$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

- ① We estimate $V_{\theta}(t) = \mathbb{E} \left[\phi_{h,t}^2 \left(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S} \right) \right]$ with

$$\phi_{h,t} \left(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S} \right) = \frac{\left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right)}{\sqrt{h} \cdot \kappa_2 \cdot \bar{p}_{T|S}(T|S)} \cdot [Y - \bar{\mu}(t, S) - (T - t) \cdot \bar{\beta}(t, S)]$$

$$\text{by } \hat{V}_{\theta}(t) = \frac{1}{n} \sum_{i=1}^n \phi_{h,t}^2 \left(Y, T, S; \hat{\mu}, \hat{\beta}, \hat{p}_{T|S} \right).$$

- ② $\hat{\mu}, \hat{\beta}, \hat{p}_{T|S}$ can be estimated via sample-splitting or cross-fitting.



An asymptotically valid inference on $\theta(t) = \frac{d}{dt} \mathbb{E} [Y(t)]$ can be conducted through

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

- 1 We estimate $V_{\theta}(t) = \mathbb{E} \left[\phi_{h,t}^2 \left(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S} \right) \right]$ with

$$\phi_{h,t} \left(Y, T, S; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S} \right) = \frac{\left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right)}{\sqrt{h} \cdot \kappa_2 \cdot \bar{p}_{T|S}(T|S)} \cdot [Y - \bar{\mu}(t, S) - (T - t) \cdot \bar{\beta}(t, S)]$$

$$\text{by } \hat{V}_{\theta}(t) = \frac{1}{n} \sum_{i=1}^n \phi_{h,t}^2 \left(Y, T, S; \hat{\mu}, \hat{\beta}, \hat{p}_{T|S} \right).$$

- 2 $\hat{\mu}, \hat{\beta}, \hat{p}_{T|S}$ can be estimated via sample-splitting or cross-fitting.
- 3 The explicit form of $B_{\theta}(t)$ is complicated, but $h^2 B_{\theta}(t)$ is asymptotically negligible when $h = O \left(n^{-\frac{1}{5}} \right)$.
 - This order aligns with the outputs from usual bandwidth selection methods (Wand and Jones, 1994; Wasserman, 2006).

Question:² Do we have a nonparametric efficiency lower bound for $\hat{\theta}_{\text{DR}}(t)$?

²I acknowledge Ted Westling and Aaron Hudson for pointing out this direction.

Question:² Do we have a nonparametric efficiency lower bound for $\hat{\theta}_{\text{DR}}(t)$?

- $t \mapsto \theta(t) := \Psi(P_0)(t)$ is *not* pathwise differentiable (Bickel et al., 1998; Hirano and Porter, 2012; Luedtke and van der Laan, 2016):

$$\forall t \in \mathcal{T}, \quad \exists \{P_\epsilon : \epsilon \in \mathbb{R}\} \quad \text{s.t.} \quad \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_\epsilon)(t) - \Psi(P_0)(t)}{\epsilon} \quad \text{does not exist.}$$

²I acknowledge Ted Westling and Aaron Hudson for pointing out this direction.

Question:² Do we have a nonparametric efficiency lower bound for $\hat{\theta}_{\text{DR}}(t)$?

- $t \mapsto \theta(t) := \Psi(P_0)(t)$ is *not* pathwise differentiable (Bickel et al., 1998; Hirano and Porter, 2012; Luedtke and van der Laan, 2016):

$$\forall t \in \mathcal{T}, \quad \exists \{P_\epsilon : \epsilon \in \mathbb{R}\} \quad \text{s.t.} \quad \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_\epsilon)(t) - \Psi(P_0)(t)}{\epsilon} \quad \text{does not exist.}$$

- For a fixed $h > 0$, the smooth functional $\Phi(P_0)(t) := \mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S)} \right]$ is pathwise differentiable (van der Laan et al., 2018; Takatsu and Westling, 2024).

²I acknowledge Ted Westling and Aaron Hudson for pointing out this direction.

Question:² Do we have a nonparametric efficiency lower bound for $\hat{\theta}_{\text{DR}}(t)$?

- $t \mapsto \theta(t) := \Psi(P_0)(t)$ is *not* pathwise differentiable (Bickel et al., 1998; Hirano and Porter, 2012; Luedtke and van der Laan, 2016):

$$\forall t \in \mathcal{T}, \quad \exists \{P_\epsilon : \epsilon \in \mathbb{R}\} \quad \text{s.t.} \quad \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_\epsilon)(t) - \Psi(P_0)(t)}{\epsilon} \quad \text{does not exist.}$$

- For a fixed $h > 0$, the smooth functional $\Phi(P_0)(t) := \mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S)} \right]$ is pathwise differentiable (van der Laan et al., 2018; Takatsu and Westling, 2024).
- Up to a shrinking bias $O(h^2)$, the efficient influence function for $\Phi(P_0)(t)$ leads to

$$\hat{\theta}_{\text{EIF}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\kappa_2 \cdot \hat{p}_{T|S}(T_i|S_i)} [Y_i - \hat{\mu}(T_i, S_i)] + \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i).$$

- The asymptotic variances of $\hat{\theta}_{\text{DR}}(t)$ and $\hat{\theta}_{\text{EIF}}(t)$ are the same (or differing by $O(h^2)$)!

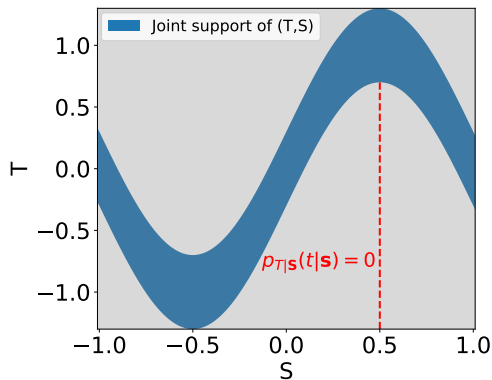
²I acknowledge Ted Westling and Aaron Hudson for pointing out this direction.

Nonparametric Inference on $\theta(t)$ Without Positivity



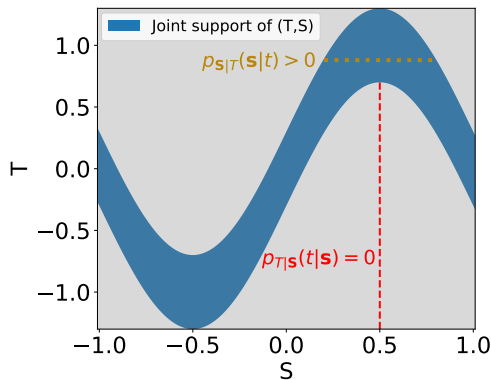
Assumption (Identification Conditions)

- 1 (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- 2 (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- 3 (Treatment Variation) $\text{Var}(T|S = s) > 0$ for all $s \in \mathcal{S}$.



Assumption (Identification Conditions)

- ① (Consistency) $Y = Y(t)$ whenever $T = t \in \mathcal{T}$.
- ② (Ignorability) $Y(t)$ is conditionally independent of T given S for all $t \in \mathcal{T}$.
- ③ (Treatment Variation) $\text{Var}(T|S = s) > 0$ for all $s \in \mathcal{S}$.



Assumption (Extrapolation; Zhang et al. 2024)

Assume $(t, s) \mapsto \mathbb{E}[Y(t)|S = s]$ to be differentiable w.r.to t for any $(t, s) \in \mathcal{T} \times \mathcal{S}$ with $p_{S|T}(s|t) > 0$ and

$$\begin{aligned} \theta(t) &= \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \right] \\ &\stackrel{*}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \mid T = t \right]. \end{aligned}$$

Additionally, it holds true that $\mathbb{E}(Y) = \mathbb{E}[m(T)]$.

Key Example: Additive Confounding Model

$$\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \right] \stackrel{*}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \mid T = t \right].$$

Key Example: Additive Confounding Model

$$\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \right] \stackrel{*}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \mid T = t \right].$$

Proposition 2 in [Zhang et al. \(2024\)](#) shows that the above equality holds under an additive structural assumption

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon.$$

- $\bar{m} : \mathcal{T} \rightarrow \mathbb{R}$ and $\eta : \mathcal{S} \rightarrow \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) > 0$.

Key Example: Additive Confounding Model

$$\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \right] \stackrel{*}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \mid T = t \right].$$

Proposition 2 in [Zhang et al. \(2024\)](#) shows that the above equality holds under an additive structural assumption

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon.$$

- $\bar{m} : \mathcal{T} \rightarrow \mathbb{R}$ and $\eta : \mathcal{S} \rightarrow \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) > 0$.
- **Identification:**

$$m(t) = \mathbb{E} \left[Y + \int_{u=T}^{u=t} \theta(u) du \right] \quad \text{and} \quad \theta(t) = \int \frac{\partial}{\partial t} \mu(t, s) dF_{S|T}(s|t).$$

Key Example: Additive Confounding Model

$$\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)] = \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \right] \stackrel{*}{=} \mathbb{E} \left[\frac{\partial}{\partial t} \mathbb{E}[Y(t)|S] \mid T = t \right].$$

Proposition 2 in [Zhang et al. \(2024\)](#) shows that the above equality holds under an additive structural assumption

$$Y(t) = \bar{m}(t) + \eta(S) + \epsilon.$$

- $\bar{m} : \mathcal{T} \rightarrow \mathbb{R}$ and $\eta : \mathcal{S} \rightarrow \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) > 0$.

- **Identification:**

$$m(t) = \mathbb{E} \left[Y + \int_{u=T}^{u=t} \theta(u) du \right] \quad \text{and} \quad \theta(t) = \int \frac{\partial}{\partial t} \mu(t, s) dF_{S|T}(s|t).$$

- **RA estimators without positivity ([Zhang et al., 2024](#)):**

$$\hat{m}_{\text{C,RA}}(t) = \frac{1}{n} \sum_{i=1}^n \left[Y_i + \int_{\tilde{t}=T_i}^{\tilde{t}=t} \hat{\theta}_{\text{C,RA}}(\tilde{t}) d\tilde{t} \right] \quad \text{and} \quad \hat{\theta}_{\text{C,RA}}(t) = \int \hat{\beta}(t, s) d\hat{F}_{S|T}(s|t).$$

Question: How about IPW and DR estimators for $\theta(t)$ without positivity?

- For identification, we assume $Y(t) = \bar{m}(t) + \eta(S) + \epsilon$.

Question: How about IPW and DR estimators for $\theta(t)$ without positivity?

- For identification, we assume $Y(t) = \bar{m}(t) + \eta(S) + \epsilon$.
- Recall the standard (oracle) IPW estimators for $m(t)$ and $\theta(t)$:

$$\tilde{m}_{\text{IPW}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{Y_i \cdot K\left(\frac{T_i - t}{h}\right)}{p_{T|S}(T_i | \mathbf{S}_i)} \quad \text{and} \quad \tilde{\theta}_{\text{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{\kappa_2 \cdot p_{T|S}(T_i | \mathbf{S}_i)}.$$

Question: How about IPW and DR estimators for $\theta(t)$ without positivity?

- For identification, we assume $Y(t) = \bar{m}(t) + \eta(S) + \epsilon$.
- Recall the standard (oracle) IPW estimators for $m(t)$ and $\theta(t)$:

$$\tilde{m}_{\text{IPW}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{Y_i \cdot K\left(\frac{T_i - t}{h}\right)}{p_{T|S}(T_i | S_i)} \quad \text{and} \quad \tilde{\theta}_{\text{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{\kappa_2 \cdot p_{T|S}(T_i | S_i)}.$$

Proposition (Proposition 2 in [Zhang and Chen 2025](#))

$$\lim_{h \rightarrow 0} \mathbb{E} [\tilde{m}_{\text{IPW}}(t)] = \bar{m}(t) \cdot \rho(t) + \omega(t) \neq m(t), \quad \text{with} \quad \rho(t) = \mathbb{P}(S \in \mathcal{S}(t)),$$

$$\lim_{h \rightarrow 0} \mathbb{E} [\tilde{\theta}_{\text{IPW}}(t)] = \begin{cases} \bar{m}'(t) \cdot \rho(t) \\ \infty \end{cases} \neq \theta(t), \quad \text{and} \quad \omega(t) = \mathbb{E} \left[\eta(S) \mathbb{1}_{\{S \in \mathcal{S}(t)\}} \right].$$

Question: How about IPW and DR estimators for $\theta(t)$ without positivity?

- For identification, we assume $Y(t) = \bar{m}(t) + \eta(S) + \epsilon$.
- Recall the standard (oracle) IPW estimators for $m(t)$ and $\theta(t)$:

$$\tilde{m}_{\text{IPW}}(t) = \frac{1}{nh} \sum_{i=1}^n \frac{Y_i \cdot K\left(\frac{T_i - t}{h}\right)}{p_{T|S}(T_i|S_i)} \quad \text{and} \quad \tilde{\theta}_{\text{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{\kappa_2 \cdot p_{T|S}(T_i|S_i)}.$$

Proposition (Proposition 2 in [Zhang and Chen 2025](#))

$$\lim_{h \rightarrow 0} \mathbb{E} [\tilde{m}_{\text{IPW}}(t)] = \bar{m}(t) \cdot \rho(t) + \omega(t) \neq m(t), \quad \text{with} \quad \rho(t) = \mathbb{P}(S \in \mathcal{S}(t)),$$

$$\lim_{h \rightarrow 0} \mathbb{E} [\tilde{\theta}_{\text{IPW}}(t)] = \begin{cases} \bar{m}'(t) \cdot \rho(t) \\ \infty \end{cases} \neq \theta(t), \quad \text{and} \quad \omega(t) = \mathbb{E} \left[\eta(S) \mathbb{1}_{\{S \in \mathcal{S}(t)\}} \right].$$

► **Key Issue:** The conditional support $\mathcal{S}(t)$ of $p_{S|T}(s|t)$ and the marginal support \mathcal{S} of $p_S(s)$ are different under the violations of positivity!!

$$\lim_{h \rightarrow 0} \mathbb{E} \left[\tilde{\theta}_{\text{IPW}}(t) \right] = \lim_{h \rightarrow 0} \mathbb{E} \left[\frac{Y \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S)} \right] = \begin{cases} \bar{m}'(t) \cdot \rho(t) \\ \infty \end{cases} \neq \theta(t),$$

where $\rho(t) = \mathbb{P}(S \in \mathcal{S}(t))$ and $\omega(t) = \mathbb{E} \left[\eta(S) \mathbb{1}_{\{S \in \mathcal{S}(t)\}} \right]$.

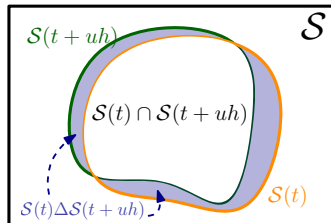
Bias-Corrected IPW Estimator for $\theta(t)$

$$\lim_{h \rightarrow 0} \mathbb{E} [\tilde{\theta}_{\text{IPW}}(t)] = \lim_{h \rightarrow 0} \mathbb{E} \left[\frac{Y \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S)} \right] = \begin{cases} \bar{m}'(t) \cdot \rho(t) & \neq \theta(t), \\ \infty & \end{cases}$$

where $\rho(t) = \mathbb{P}(S \in \mathcal{S}(t))$ and $\omega(t) = \mathbb{E} [\eta(S) \mathbb{1}_{\{S \in \mathcal{S}(t)\}}]$.

① We first want to disentangle $\theta(t) = \bar{m}'(t)$ from the bias term:

$$\begin{aligned} \mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right) \cdot p_{S|T}(S|t)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S) \cdot p_S(S)} \right] &= \bar{m}'(t) + O(h^2) \\ + \underbrace{\int_{\mathbb{R}} \mathbb{E} \left\{ [\bar{m}(t+uh) + \eta(S)] [\mathbb{1}_{\{S \in \mathcal{S}(t+uh) \setminus \mathcal{S}(t)\}} - \mathbb{1}_{\{S \in \mathcal{S}(t) \setminus \mathcal{S}(t+uh)\}}] \middle| T=t \right\} u \cdot K(u) du}_{\text{Non-vanishing Bias}}. \end{aligned}$$

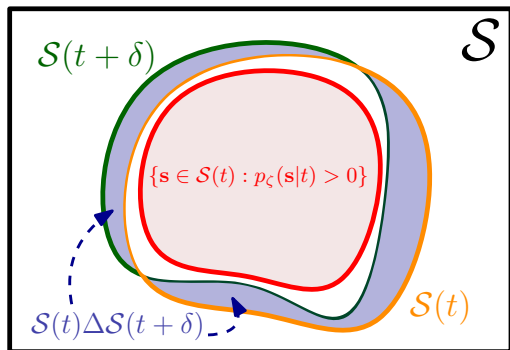


$$\mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right) p_{S|T}(S|t)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S) \cdot p_S(S)} \right] = \bar{m}'(t) + O(h^2) + \text{“Non-vanishing Bias”}.$$

$$\mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right) p_{S|T}(S|t)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S) \cdot p_S(S)} \right] = \bar{m}'(t) + O(h^2) + \text{"Non-vanishing Bias"}.$$

- 2 We replace $p_{S|T}(s|t)$ with a ζ -interior conditional density $p_\zeta(s|t)$ so that

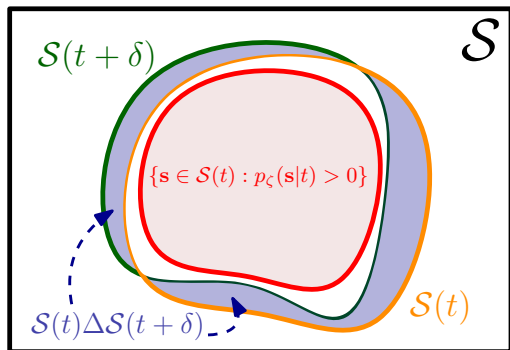
$$\{s \in \mathcal{S}(t) : p_\zeta(s|t) > 0\} \subset \mathcal{S}(t + \delta) \quad \text{for any } \delta \in [-h, h].$$



$$\mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right) p_{S|T}(S|t)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S) \cdot p_S(S)} \right] = \bar{m}'(t) + O(h^2) + \text{"Non-vanishing Bias"}.$$

- 2 We replace $p_{S|T}(s|t)$ with a ζ -interior conditional density $p_\zeta(s|t)$ so that

$$\{s \in \mathcal{S}(t) : p_\zeta(s|t) > 0\} \subset \mathcal{S}(t + \delta) \quad \text{for any } \delta \in [-h, h].$$



Now, we have that

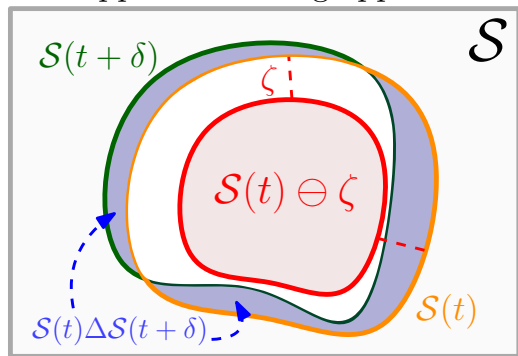
$$\mathbb{E} \left[\frac{Y \cdot \left(\frac{T-t}{h} \right) K \left(\frac{T-t}{h} \right) p_\zeta(S|t)}{h^2 \cdot \kappa_2 \cdot p_{T|S}(T|S) \cdot p_S(S)} \right] = \bar{m}'(t) + O(h^2).$$

Question: How can we find a ζ -interior conditional density $p_{\zeta}(s|t)$?

ζ -Interior Conditional Density

Question: How can we find a ζ -interior conditional density $p_\zeta(\mathbf{s}|t)$?

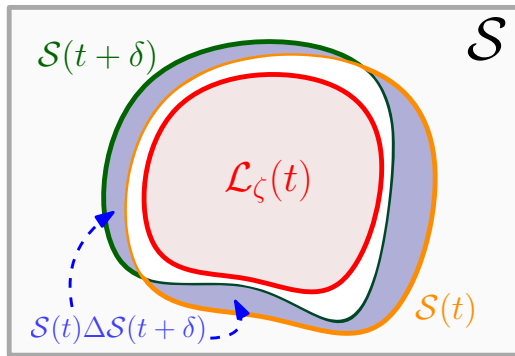
Support shrinking approach



$$\mathcal{S}(t) \ominus \zeta = \left\{ \mathbf{s} \in \mathcal{S}(t) : \inf_{\mathbf{x} \in \partial \mathcal{S}(t)} \|\mathbf{s} - \mathbf{x}\|_2 \geq \zeta \right\},$$

$$p_\zeta(\mathbf{s}|t) = \frac{p_{\mathcal{S}|T}(\mathbf{s}|t) \cdot \mathbb{1}_{\{\mathbf{s} \in \mathcal{S}(t) \ominus \zeta\}}}{\int_{\mathcal{S}(t) \ominus \zeta} p_{\mathcal{S}|T}(\mathbf{s}_1|t) d\mathbf{s}_1}.$$

Level set approach



$$\mathcal{L}_\zeta(t) = \left\{ \mathbf{s} \in \mathcal{S}(t) : p_{\mathcal{S}|T}(\mathbf{s}|t) \geq \zeta \right\},$$

$$p_\zeta(\mathbf{s}|t) = \frac{p_{\mathcal{S}|T}(\mathbf{s}|t) \cdot \mathbb{1}_{\{\mathbf{s} \in \mathcal{L}_\zeta(t)\}}}{\int_{\mathcal{L}_\zeta(t)} p_{\mathcal{S}|T}(\mathbf{s}_1|t) d\mathbf{s}_1}.$$

► **Bias-Corrected IPW Estimator Without Positivity:**

$$\hat{\theta}_{\text{C,IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right) \hat{p}_{\zeta}(\mathbf{S}_i|t)}{\kappa_2 \cdot \hat{p}(T_i, \mathbf{S}_i)},$$

- $\hat{p}(t, \mathbf{s}), \hat{p}_{\zeta}(\mathbf{s}|t)$ are estimators of $p(t, \mathbf{s}), p_{\zeta}(\mathbf{s}|t)$ and $\zeta = 0.5 \cdot \max \{\hat{p}_{\mathbf{S}|T}(\mathbf{S}_i|t) : i = 1, \dots, n\}$.

► Bias-Corrected IPW Estimator Without Positivity:

$$\hat{\theta}_{\text{C,IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right) \hat{p}_{\zeta}(S_i|t)}{\kappa_2 \cdot \hat{p}(T_i, S_i)},$$

- $\hat{p}(t, s), \hat{p}_{\zeta}(s|t)$ are estimators of $p(t, s), p_{\zeta}(s|t)$ and $\zeta = 0.5 \cdot \max \{\hat{p}_{S|T}(S_i|t) : i = 1, \dots, n\}$.

► Bias-Corrected DR Estimator Without Positivity:

$$\begin{aligned} \hat{\theta}_{\text{C,DR}}(t) = & \underbrace{\frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right) \hat{p}_{\zeta}(S_i|t)}{\kappa_2 \cdot \hat{p}(T_i, S_i)} \left[Y_i - \hat{\mu}(t, S_i) - (T_i - t) \cdot \hat{\beta}(t, S_i) \right]}_{\text{IPW component}} \\ & + \underbrace{\int \hat{\beta}(t, s) \cdot \hat{p}_{\zeta}(s|t) ds}_{\text{RA component}}. \end{aligned}$$

Theorem (Theorem 5 in Zhang and Chen 2025)

Under some regularity assumptions and

- ① $\hat{\mu}, \hat{\beta}, \hat{p}, \hat{p}_\zeta$ are estimated on a dataset independent of $\{(Y_i, T_i, S_i)\}_{i=1}^n$;
- ② $\sqrt{nh} \|\hat{p}_\zeta(S|t) - \bar{p}_\zeta(S|t)\|_{L_2} = o_P(1)$ with $\hat{p}_\zeta(s|t) \xrightarrow{P} \bar{p}_\zeta(s|t)$;
- ③ at least one of the model specification conditions hold:
 - $\hat{p}(t, s) \xrightarrow{P} \bar{p}(t, s) = p(t, s)$ (joint density model),
 - $\hat{\mu}(t, s) \xrightarrow{P} \bar{\mu}(t, s) = \mu(t, s)$ and $\hat{\beta}(t, s) \xrightarrow{P} \bar{\beta}(t, s) = \beta(t, s)$ (outcome model);
- ④ $\sup_{|u-t| \leq h} \|\hat{p}(u, S) - p(u, S)\|_{L_2} \left[\|\hat{\mu}(t, S) - \mu(t, S)\|_{L_2} + h \left\| \hat{\beta}(t, S) - \beta(t, S) \right\|_{L_2} \right] = o_P\left(\frac{1}{\sqrt{nh}}\right)$,

we prove that

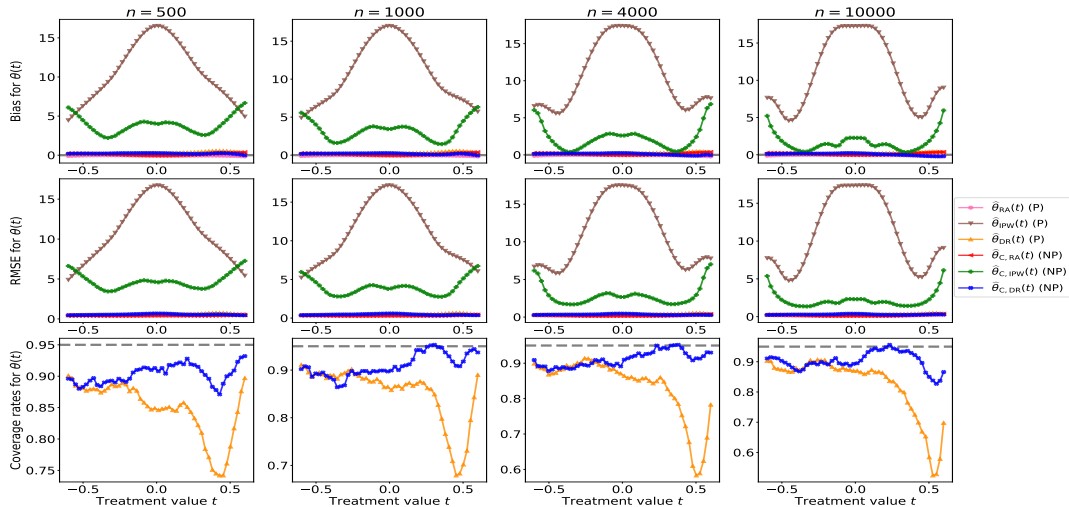
- $\sqrt{nh^3} [\hat{\theta}_{C,DR}(t) - \theta(t)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{C,h,t}(Y_i, T_i, S_i; \bar{\mu}, \bar{\beta}, \bar{p}_{T|S}) + o_P(1).$
- $\sqrt{nh^3} [\hat{\theta}_{C,DR}(t) - \theta(t) - h^2 \cdot B_{C,\theta}(t)] \xrightarrow{d} \mathcal{N}(0, V_{C,\theta}(t)).$

Experiments and Discussion



Simulations for $\hat{\theta}_{C,RA}(t)$, $\hat{\theta}_{C,IPW}(t)$, $\hat{\theta}_{C,DR}(t)$ Without Positivity

$$Y = T^3 + T^2 + 10S + \epsilon, \quad T = \sin(\pi S) + E, \quad S \sim \text{Unif}[-1, 1], \quad E \sim \text{Unif}[-0.3, 0.3].$$



Note: $\beta(t, s) = \frac{\partial}{\partial t} \mu(t, s)$ is estimated via automatic differentiation of a well-trained neural network (inspired by [Luedtke 2024](#)).

A Case Study Under Positivity

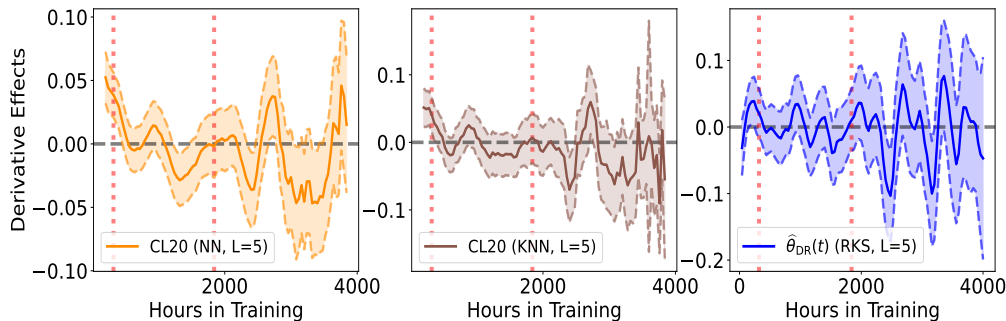
We compare our proposed DR estimator $\hat{\theta}_{\text{DR}}(t)$ under positivity with the finite-difference method (Colangelo and Lee 2020; CL20) on the U.S. Job Corps program (Schochet et al., 2001).

- Y is the proportion of weeks employed in 2nd year after enrollment.
- T is the total hours of academic and vocational training received.
- S comprises 49 socioeconomic characteristics, and $n = 4024$.

A Case Study Under Positivity

We compare our proposed DR estimator $\hat{\theta}_{\text{DR}}(t)$ under positivity with the finite-difference method (Colangelo and Lee 2020; CL20) on the U.S. Job Corps program (Schochet et al., 2001).

- Y is the proportion of weeks employed in 2nd year after enrollment.
- T is the total hours of academic and vocational training received.
- S comprises 49 socioeconomic characteristics, and $n = 4024$.



Summary and Future Work

We study (nonparametric) doubly robust inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$, $t \in \mathcal{T} \subset \mathbb{R}$.

We study (nonparametric) doubly robust inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$, $t \in \mathcal{T} \subset \mathbb{R}$.

① Under the positivity condition:

- We propose $\hat{\theta}_{\text{DR}}(t)$ with standard nonparametric consistency and efficiency guarantee:

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

We study (nonparametric) doubly robust inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$, $t \in \mathcal{T} \subset \mathbb{R}$.

① Under the positivity condition:

- We propose $\hat{\theta}_{\text{DR}}(t)$ with standard nonparametric consistency and efficiency guarantee:

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

② Without the positivity condition:

- Our bias-corrected IPW and DR estimators $\hat{\theta}_{\text{C,IPW}}(t)$, $\hat{\theta}_{\text{C,DR}}(t)$ reveal interesting connections to nonparametric level set estimation problems (Bonvini et al., 2023):

Causal Inference \iff *Geometric Data Analysis* (<https://uwgeometry.github.io/>)!

We study (nonparametric) doubly robust inference on $\theta(t) = \frac{d}{dt} \mathbb{E}[Y(t)]$, $t \in \mathcal{T} \subset \mathbb{R}$.

1 Under the positivity condition:

- We propose $\hat{\theta}_{\text{DR}}(t)$ with standard nonparametric consistency and efficiency guarantee:

$$\sqrt{nh^3} \left[\hat{\theta}_{\text{DR}}(t) - \theta(t) - h^2 B_{\theta}(t) \right] \xrightarrow{d} \mathcal{N}(0, V_{\theta}(t)).$$

2 Without the positivity condition:

- Our bias-corrected IPW and DR estimators $\hat{\theta}_{\text{C,IPW}}(t)$, $\hat{\theta}_{\text{C,DR}}(t)$ reveal interesting connections to nonparametric level set estimation problems (Bonvini et al., 2023):

Causal Inference \iff *Geometric Data Analysis* (<https://uwgeometry.github.io/>)!

3 Future Works:

- Sensitivity analysis on unmeasured confounding (Chernozhukov et al., 2022).
- Generalize our derivative estimators to other causal estimands:
 - instantaneous causal effect $\frac{d}{dt} \mathbb{E}[Y(t)|S=s]$ (Stolzenberg, 1980);
 - direct and indirect effects in mediation analysis (Huber et al., 2020; Xu et al., 2021)?

Thank you!

More details can be found in

[1] Y. Zhang and Y.-C. Chen. Doubly Robust Inference on Causal Derivative Effects for Continuous Treatments. *arXiv preprint*, 2025. <https://arxiv.org/abs/2501.06969>.

All the code and data are available at
<https://github.com/zhangyk8/npDRDeriv>.

Python Package: [npDoseResponse](#).

- P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer New York, 1998.
- M. Bonvini, E. H. Kennedy, and L. J. Keele. Minimax optimal subgroup identification. *arXiv preprint arXiv:2306.17464*, 2023.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.
- V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.
- K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.
- R. D. Gill and J. M. Robins. Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, 29(6):1785–1811, 2001.
- K. Hirano and G. W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004.
- K. Hirano and J. R. Porter. Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790, 2012.
- M. Huber, Y.-C. Hsu, Y.-Y. Lee, and L. Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840, 2020.
- K. Imai and D. A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

- N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2018.
- E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.
- A. Luedtke. Simplifying debiased inference via automatic differentiation and probabilistic programming. *arXiv preprint arXiv:2405.08675*, 2024.
- A. R. Luedtke and M. J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713–742, 2016.
- Y. Mack and H.-G. Müller. Derivative estimation in nonparametric regression with random predictor variable. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 59–72, 1989.
- W. K. Newey and T. M. Stoker. Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223, 1993.
- J. Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and Statistics*, pages 213–234, 1959.
- J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.
- J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.
- D. Rothenhäusler and B. Yu. Incremental causal effects. *arXiv preprint arXiv:1907.13258*, 2019.

- P. Z. Schochet, J. Burghardt, and S. Glazerman. National job corps study: The impacts of job corps on participants' employment and related outcomes. Mathematica policy research reports, Mathematica Policy Research, 2001.
- J. Shao. *Mathematical Statistics*. Springer Science & Business Media, 2003.
- R. M. Stolzenberg. The measurement and decomposition of causal effects in nonlinear and nonadditive models. *Sociological Methodology*, 11:459–488, 1980.
- K. Takatsu and T. Westling. Debaised inference for a covariate-adjusted regression function. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae041, 2024.
- M. J. van der Laan, A. Bibaut, and A. R. Luedtke. Cv-tmlr for nonpathwise differentiable target parameters. In M. J. van der Laan and S. Rose, editors, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 455–481. Springer, 2018.
- M. P. Wand and M. C. Jones. *Kernel Smoothing*. CRC press, 1994.
- L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.
- Y. Xu, N. Sani, A. Ghassami, and I. Shpitser. Multiply robust causal mediation analysis with continuous treatments. *arXiv preprint arXiv:2105.09254*, 2021.
- Y. Zhang and Y.-C. Chen. Doubly robust inference on causal derivative effects for continuous treatments. *arXiv preprint arXiv:2501.*, 2025.
- Y. Zhang, Y.-C. Chen, and A. Giessing. Nonparametric inference on dose-response curves without the positivity condition. *arXiv preprint arXiv:2405.09003*, 2024.

Detailed Regularity Assumptions

Assumption (Differentiability of the conditional mean outcome function)

For any $(t, s) \in \mathcal{T} \times \mathcal{S}$ and $\mu(t, s) = \mathbb{E}(Y|T = t, S = s)$, it holds that

- ① $\mu(t, s)$ is at least four times continuously differentiable with respect to t .
- ② $\mu(t, s)$ and all of its partial derivatives are uniformly bounded on $\mathcal{T} \times \mathcal{S}$.

Let \mathcal{J} be the support of the joint density $p(t, s)$.

Assumption (Differentiability of the density functions)

For any $(t, s) \in \mathcal{J}$, it holds that

- ① The joint density $p(t, s)$ and the conditional density $p_{T|S}(t|s)$ are at least three times continuously differentiable with respect to t .
- ② $p(t, s)$, $p_{T|S}(t|s)$, $p_{S|T}(s|t)$, as well as all of the partial derivatives of $p(t, s)$ and $p_{T|S}(t|s)$ are bounded and continuous up to the boundary $\partial\mathcal{J}$.
- ③ The support \mathcal{T} of the marginal density $p_T(t)$ is compact and $p_T(t)$ is uniformly bounded away from 0 within \mathcal{T} .

Assumption (Regular kernel conditions)

A kernel function $K : \mathbb{R} \rightarrow [0, \infty)$ is bounded and compactly supported on $[-1, 1]$ with $\int_{\mathbb{R}} K(t) dt = 1$ and $K(t) = K(-t)$. In addition, it holds that

- ① $\kappa_j := \int_{\mathbb{R}} u^j K(u) du < \infty$ and $\nu_j := \int_{\mathbb{R}} u^j K^2(u) du < \infty$ for all $j = 1, 2, \dots$
- ② K is a second-order kernel, i.e., $\kappa_1 = 0$ and $\kappa_2 > 0$.
- ③ $\mathcal{K} = \left\{ t' \mapsto \left(\frac{t' - t}{h} \right)^{k_1} K \left(\frac{t' - t}{h} \right) : t \in \mathcal{T}, h > 0, k_1 = 0, 1 \right\}$ is a bounded VC-type class of measurable functions on \mathbb{R} .

Assumption (Smoothness condition on $\mathcal{S}(t)$)

For any $\delta \in \mathbb{R}$ and $t \in \mathcal{T}$, there exists an absolute constant $A_0 > 0$ such that either (i) “ $\mathcal{S}(t) \ominus (A_0|\delta|) \subset \mathcal{S}(t + \delta)$ ” for the support shrinking approach or (ii) “ $\mathcal{L}_{A_0|\delta|}(t) \subset \mathcal{S}(t + \delta)$ ” for the level set approach.

The self-normalizing technique can reduce the instability of IPW and DR estimators (Kallus and Zhou, 2018):

1 Self-Normalized Estimators Under Positivity:

$$\hat{\theta}_{\text{IPW}}^{\text{norm}}(t) = \frac{\hat{\theta}_{\text{IPW}}(t)}{\frac{1}{nh} \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right)}{\hat{p}_{T|S}(T_j|S_j)}} = \frac{\sum_{i=1}^n \frac{Y_i\left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)}}{\kappa_2 h \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right)}{\hat{p}_{T|S}(T_j|S_j)}},$$

and

$$\hat{\theta}_{\text{DR}}^{\text{norm}}(t) = \frac{\sum_{i=1}^n \frac{[Y_i - \hat{\mu}(t, S_i) - (T_i - t) \cdot \hat{\beta}(t, S_i)] \left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right)}{\hat{p}_{T|S}(T_i|S_i)}}{\kappa_2 h \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right)}{\hat{p}_{T|S}(T_j|S_j)}} + \frac{1}{n} \sum_{i=1}^n \hat{\beta}(t, S_i).$$

2 Self-Normalized Estimators Without Positivity:

$$\hat{\theta}_{C,IPW}^{\text{norm}}(t) = \frac{\hat{\theta}_{C,IPW}(t)}{\frac{1}{nh} \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right) \cdot \hat{p}_{\zeta}(S_j|t)}{\hat{p}(T_j, S_j)}} = \frac{\sum_{i=1}^n \frac{Y_i \left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right) \cdot \hat{p}_{\zeta}(S_i|t)}{\hat{p}(T_i, S_i)}}{\kappa_2 h \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right) \cdot \hat{p}_{\zeta}(S_j|t)}{\hat{p}(T_j, S_j)}},$$

and

$$\begin{aligned} \hat{\theta}_{C,DR}^{\text{norm}}(t) = & \frac{\sum_{i=1}^n \frac{[Y_i - \hat{\mu}(t, S_i) - (T_i - t) \cdot \hat{\beta}(t, S_i)] \left(\frac{T_i-t}{h}\right) K\left(\frac{T_i-t}{h}\right) \cdot \hat{p}_{\zeta}(S_i|t)}{\hat{p}(T_i, S_i)}}{\kappa_2 h \sum_{j=1}^n \frac{K\left(\frac{T_j-t}{h}\right) \cdot \hat{p}_{\zeta}(S_j|t)}{\hat{p}(T_j, S_j)}} \\ & + \int \hat{\beta}(t, s) \cdot \hat{p}_{\zeta}(s|t) ds. \end{aligned}$$

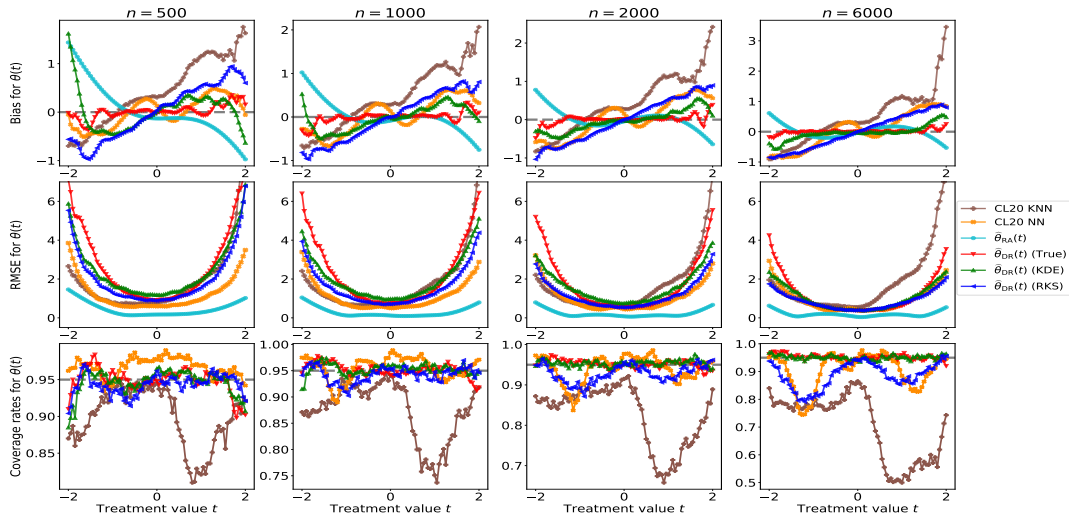
We generate i.i.d. observations $\{(Y_i, T_i, S_i)\}_{i=1}^n$ from the following data-generating model (Colangelo and Lee, 2020):

$$Y = 1.2T + T^2 + TS_1 + 1.2\boldsymbol{\xi}^T \mathbf{S} + \epsilon \sqrt{0.5 + F_{\mathcal{N}(0,1)}(S_1)}, \quad \epsilon \sim \mathcal{N}(0, 1),$$
$$T = F_{\mathcal{N}(0,1)}\left(3\boldsymbol{\xi}^T \mathbf{S}\right) - 0.5 + 0.75E, \quad \mathbf{S} = (S_1, \dots, S_d)^T \sim \mathcal{N}_d(\mathbf{0}, \Sigma), \quad E \sim \mathcal{N}(0, 1),$$

where

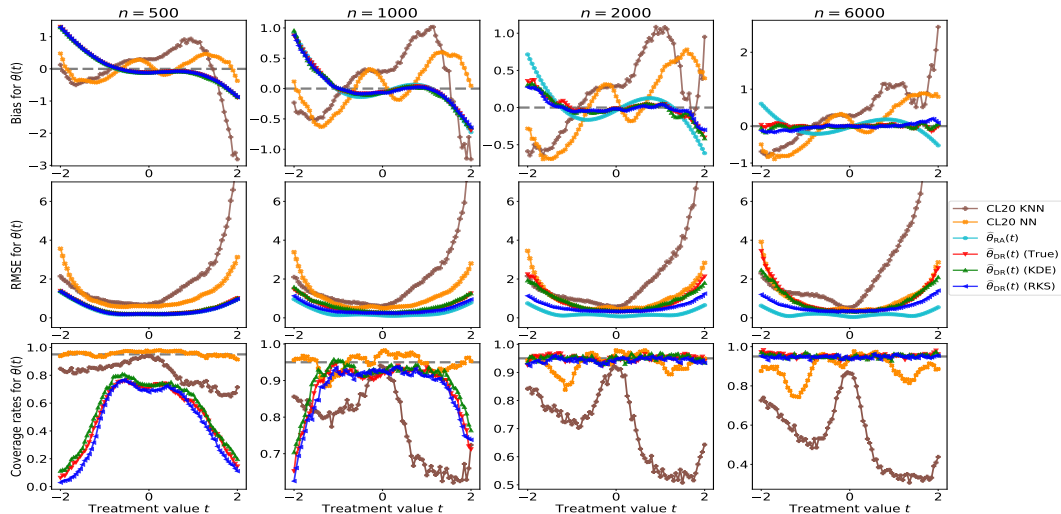
- $F_{\mathcal{N}(0,1)}$ is the CDF of $\mathcal{N}(0, 1)$ and $d = 20$.
- $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)^T \in \mathbb{R}^d$ has its entry $\xi_j = \frac{1}{j^2}$ for $j = 1, \dots, d$ and $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$ when $|i - j| = 1$, and $\Sigma_{ij} = 0$ when $|i - j| > 1$ for $i, j = 1, \dots, d$.
- The dose-response curve is given by $m(t) = 1.2t + t^2$, and our parameter of interest is the derivative effect curve $\theta(t) = 1.2 + 2t$.

Simulations for Estimating $\theta(t)$ Under Positivity



Comparisons between our proposed estimators and the finite-difference approaches by [Colangelo and Lee \(2020\)](#) (“CL20”) under positivity and **with 5-fold cross-fitting** across various sample sizes.

Simulations for Estimating $\theta(t)$ Under Positivity



Comparisons between our proposed estimators and the finite-difference approaches by Colangelo and Lee (2020) (“CL20”) under positivity and **without cross-fitting** across various sample sizes.