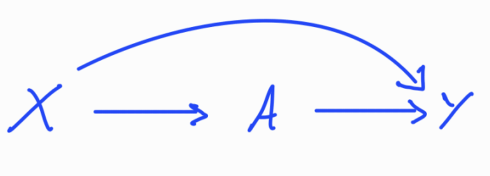
# Longitudinal trimmed and smooth trimmed effects with flip and S-flip interventions

**Alec McClean** and Iván Díaz

New York University Grossman School of Medicine

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0, 1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.
$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

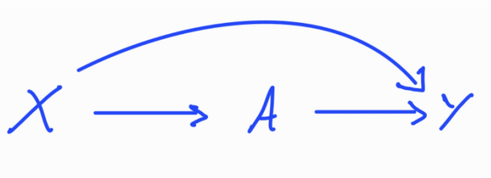
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\big\{Y(1) - Y(0)\big\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0,1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.
$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

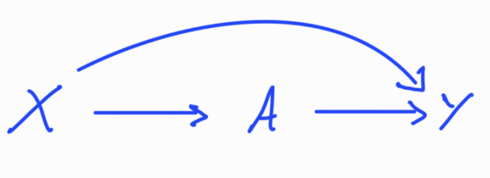
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0,1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.

$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

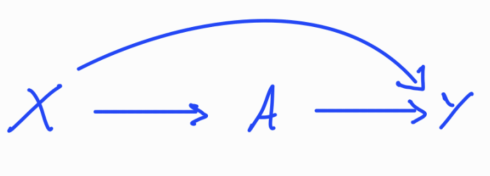
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0,1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.
$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

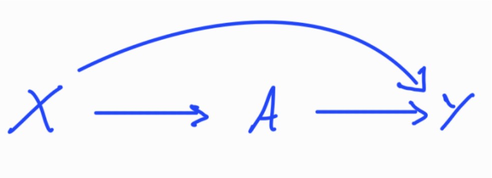
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0, 1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.
$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

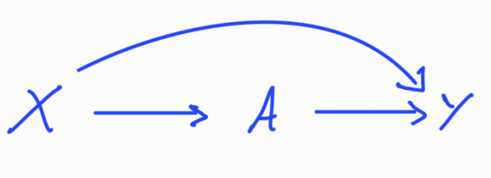
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

# Trimming with single timepoint data

$\{(X_i, A_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathcal{P}$ with $X_i$ covariates, $A_i \in \{0,1\}$ binary treatment and $Y_i \in \mathbb{R}$ outcome.
$Y(a)$ is potential outcome under treatment $a$.




**Positivity violations**: $\pi(X) = \mathbb{P}(A = 1 \mid X) \approx 0$ or $\approx 1$.
$\mathbb{E}\{Y(a)\}$ not identified ($\pi = 0$) or hard-to-estimate ($\pi \approx 0$)

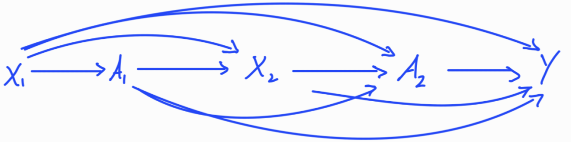
**Trimming:** Change population to satisfy positivity

$$\mathbb{E}\Big[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\Big] \text{ for } \varepsilon \in [0, 0.5)$$

(related to cond. trimming: $\mathbb{E}\{f(Z) \mid \mathcal{E}\} = \mathbb{E}\{f(Z)\mathbb{1}(\mathcal{E})\}/\mathbb{P}(\mathcal{E})$)

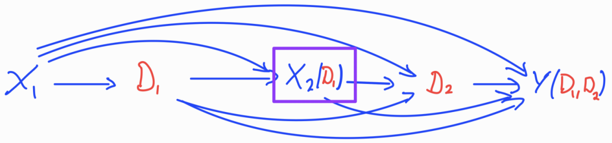
# Trimming on non-baseline covariates is challenging

Data: $(X_1, A_1, X_2, A_2, Y)$.




Can we trim on propensity scores in the second timepoint?
(where we often have positivity violations!)

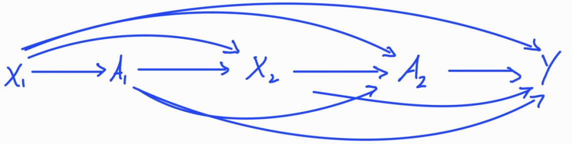
Traditional answer: no




*...the [post-treatment] covariates responsible for positivity violations may themselves be affected by past treatment. Trimming... amounts to conditioning on post-treatment covariates and can thus introduce new bias.* *[Petersen et al., 2012, Section 4.3]* (also, see, e.g., Jensen et al. [2024, Section 6.3])

# Trimming on non-baseline covariates is challenging

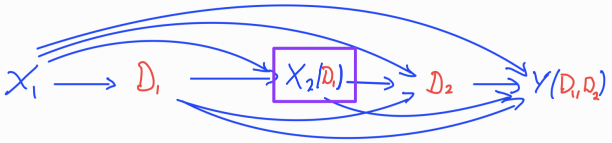
Data: $(X_1, A_1, X_2, A_2, Y)$.




Can we trim on propensity scores in the second timepoint?

(where we often have positivity violations!)

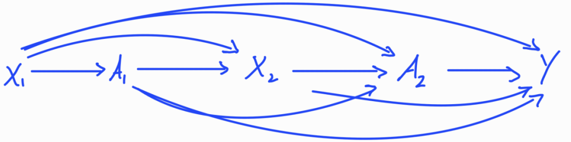
Traditional answer: no




*...the [post-treatment] covariates responsible for positivity violations may themselves be affected by past treatment. Trimming... amounts to conditioning on post-treatment covariates and can thus introduce new bias.* [Petersen et al., 2012, Section 4.3] (also, see, e.g., Jensen et al. [2024, Section 6.3])

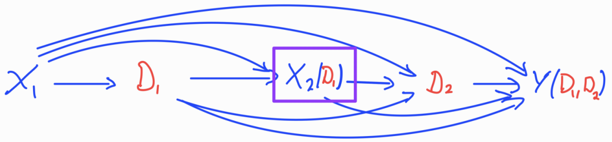
# Trimming on non-baseline covariates is challenging

Data: $(X_1, A_1, X_2, A_2, Y)$.




Can we trim on propensity scores in the second timepoint?
(where we often have positivity violations!)

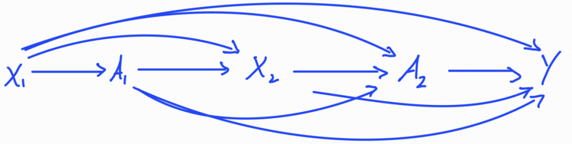
Traditional answer: no




*...the [post-treatment] covariates responsible for positivity violations may themselves be affected by past treatment. Trimming... amounts to conditioning on post-treatment covariates and can thus introduce new bias.* [Petersen et al., 2012, Section 4.3] (also, see, e.g., Jensen et al. [2024, Section 6.3])

# Trimming on non-baseline covariates is challenging

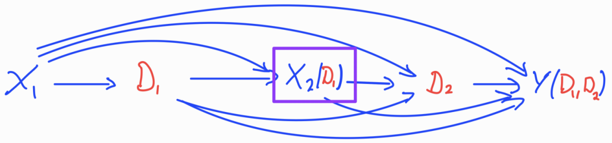
Data: $(X_1, A_1, X_2, A_2, Y)$.



Can we trim on propensity scores in the second timepoint?

(where we often have positivity violations!)

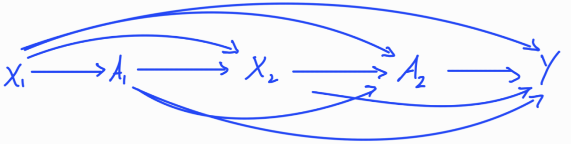
Traditional answer: no



*...the [post-treatment] covariates responsible for positivity violations may themselves be affected by past treatment. Trimming... amounts to conditioning on post-treatment covariates and can thus introduce new bias.* [Petersen et al., 2012, Section 4.3] (also, see, e.g., Jensen et al. [2024, Section 6.3])

# Trimming on non-baseline covariates is challenging

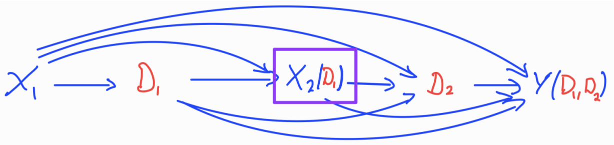
Data: $(X_1, A_1, X_2, A_2, Y)$.




Can we trim on propensity scores in the second timepoint?

(where we often have positivity violations!)

Traditional answer: no




*...the [post-treatment] covariates responsible for positivity violations may themselves be affected by past treatment. Trimming... amounts to conditioning on post-treatment covariates and can thus introduce new bias. [Petersen et al., 2012, Section 4.3]* (also, see, e.g., Jensen et al. [2024, Section 6.3])

# Trimming is a dynamic "flip" intervention

**Dynamic stochastic interventions:**
$D = d(A, X)$ adapts so $\mathbb{E}\{Y(D)\}$ IDable and estimable

- Modified treatment policies [Díaz et al., 2023]
- Incremental propensity score interventions [Kennedy, 2019]

Proposition

*Let* $\mathbb{1}(X) \equiv \mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}$ *and*

$$D(a) = \begin{cases} A & \text{if } A = a \\ \mathbb{1}(X)a + \{1 - \mathbb{1}(X)\}A & \text{if } A \neq a. \end{cases}$$

*Then, if* $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$,

$$\boxed{\mathbb{E}\left[Y\{D(1)\} - Y\{D(0)\}\right] = \mathbb{E}\left[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\right]}$$

$D(a)$ is a **flip intervention**:
If $A = a$, **do not intervene**; o/w, **flip** to $a$ **if** $\mathbb{1}(X) = 1$.

# Trimming is a dynamic "flip" intervention

**Dynamic stochastic interventions:**

$D = d(A, X)$ adapts so $\mathbb{E}\{Y(D)\}$ IDable and estimable

- Modified treatment policies [Díaz et al., 2023]
- Incremental propensity score interventions [Kennedy, 2019]

Proposition

*Let* $\mathbb{1}(X) \equiv \mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}$ *and*

$$D(a) = \begin{cases} A & \text{if } A = a \\ \mathbb{1}(X)a + \{1 - \mathbb{1}(X)\}A & \text{if } A \neq a. \end{cases}$$

*Then, if* $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$,

$$\boxed{\mathbb{E}\left[Y\{D(1)\} - Y\{D(0)\}\right] = \mathbb{E}\left[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\right]}$$

$D(a)$ is a **flip intervention**:
If $A = a$, **do not intervene**; o/w, **flip** to $a$ **if** $\mathbb{1}(X) = 1$.

# Trimming is a dynamic "flip" intervention

**Dynamic stochastic interventions:**
$D = d(A, X)$ adapts so $\mathbb{E}\{Y(D)\}$ IDable and estimable

- Modified treatment policies [Díaz et al., 2023]
- Incremental propensity score interventions [Kennedy, 2019]

Proposition

*Let* $\mathbb{1}(X) \equiv \mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}$ *and*

$$D(a) = \begin{cases} A & \text{if } A = a \\ \mathbb{1}(X)a + \{1 - \mathbb{1}(X)\}A & \text{if } A \neq a. \end{cases}$$

*Then, if* $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$,

$$\boxed{\mathbb{E}\left[Y\{D(1)\} - Y\{D(0)\}\right] = \mathbb{E}\left[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\right]}$$

$D(a)$ is a **flip intervention**:
If $A = a$, **do not intervene**; o/w, **flip** to $a$ **if** $\mathbb{1}(X) = 1$.

# Trimming is a dynamic "flip" intervention

**Dynamic stochastic interventions:**

$D = d(A, X)$ adapts so $\mathbb{E}\{Y(D)\}$ IDable and estimable

- Modified treatment policies [Díaz et al., 2023]
- Incremental propensity score interventions [Kennedy, 2019]

Proposition

*Let* $\mathbb{1}(X) \equiv \mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}$ *and*

$$D(a) = \begin{cases} A & \text{if } A = a \\ \mathbb{1}(X)a + \{1 - \mathbb{1}(X)\}A & \text{if } A \neq a. \end{cases}$$

*Then, if* $\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$,

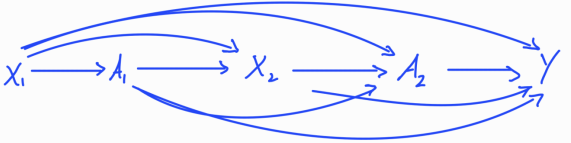
$$\boxed{\mathbb{E}\left[Y\{D(1)\} - Y\{D(0)\}\right] = \mathbb{E}\left[\{Y(1) - Y(0)\}\mathbb{1}\{\varepsilon < \pi(X) < 1 - \varepsilon\}\right]}$$

$D(a)$ is a **flip intervention**:
If $A = a$, **do not intervene**; o/w, **flip** to $a$ **if** $\mathbb{1}(X) = 1$.

# Longitudinal setup

# Longitudinal data

$\{Z_i\}_{i=1}^n \overset{iid}{\sim} \mathbb{P} \in \mathcal{P}$ where $Z = (X_1, A_1, X_2, A_2, \ldots, X_T, A_T, Y)$




$X_t \in \mathbb{R}^d$: time-varying covariates
$A_t \in \{0, 1\}$: time-varying binary treatment
$Y \in \mathbb{R}$: ultimate outcome of interest

History of $O_t$ at $t$: $\overline{O}_t = (O_1, \ldots, O_t)$
Future of $O_t$ from $t$: $\underline{O}_t = (O_t, \ldots, O_T)$
$H_t = (\overline{X}_t, \overline{A}_{t-1})$: covariate and treatment history at $t$

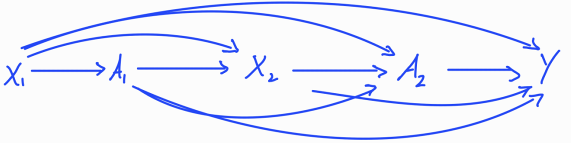
**NPSEM assumption:** there are $\{f_{X,t}, f_{A,t}\}_{t=1}^T$ and $f_Y$ such that

$$X_t = f_{X,t}(A_{t-1}, H_{t-1}, U_{X,t}),$$
$$A_t = f_{A,t}(H_t, U_{A,t}), \text{ and}$$
$$Y = f_Y(A_T, H_T, U_Y).$$

where $\{U_{X,t}, U_{A,t}, U_Y\}$ are exogeneous variables

# Longitudinal data

$\{Z_i\}_{i=1}^n \overset{iid}{\sim} \mathbb{P} \in \mathcal{P}$ where $Z = (X_1, A_1, X_2, A_2, \ldots, X_T, A_T, Y)$




$X_t \in \mathbb{R}^d$: time-varying covariates
$A_t \in \{0, 1\}$: time-varying binary treatment
$Y \in \mathbb{R}$: ultimate outcome of interest

History of $O_t$ at $t$: $\overline{O}_t = (O_1, \ldots, O_t)$
Future of $O_t$ from $t$: $\underline{O}_t = (O_t, \ldots, O_T)$
$H_t = (\overline{X}_t, \overline{A}_{t-1})$: covariate and treatment history at $t$

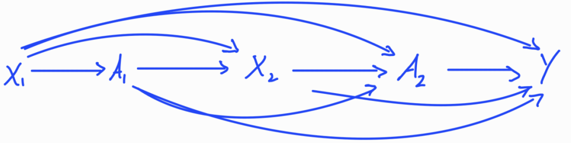
**NPSEM assumption:** there are $\{f_{X,t}, f_{A,t}\}_{t=1}^T$ and $f_Y$ such that

$$X_t = f_{X,t}(A_{t-1}, H_{t-1}, U_{X,t}),$$
$$A_t = f_{A,t}(H_t, U_{A,t}), \text{ and}$$
$$Y = f_Y(A_T, H_T, U_Y).$$

where $\{U_{X,t}, U_{A,t}, U_Y\}$ are exogeneous variables

# Longitudinal data

$\{Z_i\}_{i=1}^n \overset{iid}{\sim} \mathbb{P} \in \mathcal{P}$ where $Z = (X_1, A_1, X_2, A_2, \ldots, X_T, A_T, Y)$




$X_t \in \mathbb{R}^d$: time-varying covariates
$A_t \in \{0, 1\}$: time-varying binary treatment
$Y \in \mathbb{R}$: ultimate outcome of interest

History of $O_t$ at $t$: $\overline{O}_t = (O_1, \ldots, O_t)$
Future of $O_t$ from $t$: $\underline{O}_t = (O_t, \ldots, O_T)$
$H_t = (\overline{X}_t, \overline{A}_{t-1})$: covariate and treatment history at $t$

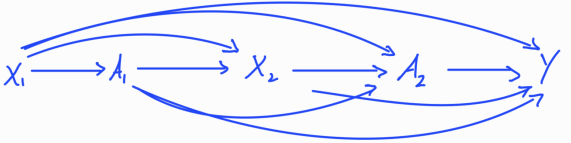
**NPSEM assumption:** there are $\{f_{X,t}, f_{A,t}\}_{t=1}^T$ and $f_Y$ such that

$$X_t = f_{X,t}(A_{t-1}, H_{t-1}, U_{X,t}),$$
$$A_t = f_{A,t}(H_t, U_{A,t}), \text{ and}$$
$$Y = f_Y(A_T, H_T, U_Y).$$

where $\{U_{X,t}, U_{A,t}, U_Y\}$ are exogeneous variables

# Longitudinal data

$$\{Z_i\}_{i=1}^n \overset{iid}{\sim} \mathbb{P} \in \mathcal{P} \text{ where } Z = (X_1, A_1, X_2, A_2, \ldots, X_T, A_T, Y)$$




$X_t \in \mathbb{R}^d$: time-varying covariates
$A_t \in \{0, 1\}$: time-varying binary treatment
$Y \in \mathbb{R}$: ultimate outcome of interest

History of $O_t$ at $t$: $\overline{O}_t = (O_1, \ldots, O_t)$
Future of $O_t$ from $t$: $\underline{O}_t = (O_t, \ldots, O_T)$
$H_t = (\overline{X}_t, \overline{A}_{t-1})$: covariate and treatment history at $t$

**NPSEM assumption:** there are $\{f_{X,t}, f_{A,t}\}_{t=1}^T$ and $f_Y$ such that

$$X_t = f_{X,t}(A_{t-1}, H_{t-1}, U_{X,t}),$$
$$A_t = f_{A,t}(H_t, U_{A,t}), \text{ and}$$
$$Y = f_Y(A_T, H_T, U_Y).$$

where $\{U_{X,t}, U_{A,t}, U_Y\}$ are exogeneous variables

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\, \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\, \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\, \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\ \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\, \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Causal assumptions

NPSEM embeds **consistency**

We will avoid **positivity** by trimming/flipping
(e.g. $\exists\ \delta > 0$ s.t. $\mathbb{P}\{\delta < \mathbb{P}(A_t = 1 \mid H_t) < 1 - \delta\} = 1$.)

**Sequential randomization:**

1. **Standard:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, U_Y) \mid H_t$ for all $t \leq T$
2. **Strong:** $U_{A,t} \perp\!\!\!\perp (\underline{U}_{X,t+1}, \underline{U}_{A,t+1}, U_Y) \mid H_t$

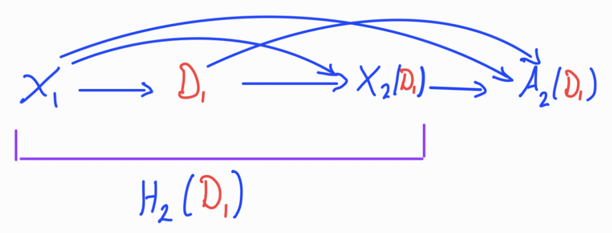
**Standard** SR: common causes of $A_t$ and future covariates and outcome measured. Typical in literature.

**Strong** SR: common causes $A_t$ and future covariates and outcome **and treatments** measured. Allows for ID when intervention depends on natural value of treatment.

# Counterfactual variables under $\overline{D}_{t-1}$

$\overline{D}_{t-1}$ generates **counterfactual variables at time $t$**:

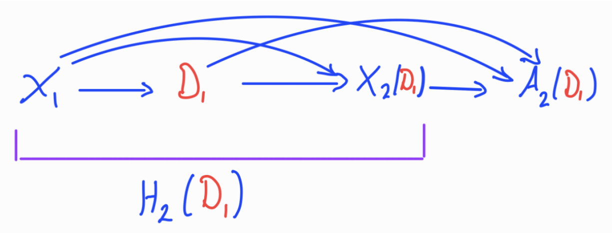
- $X_t(\overline{D}_{t-1}) = f_{X,t}(D_{t-1}, H_{t-1}(\overline{D}_{t-2}), U_{X,t})$
  *"Natural covariate value"*
- $H_t(\overline{D}_{t-1}) = (\overline{D}_{t-1}, \overline{X}_t(\overline{D}_{t-1}))$
  *"Natural covariate and treatment history"*
- $A_t(\overline{D}_{t-1}) = f_{A,t}(H_t(\overline{D}_{t-1}), U_{A,t})$
  *"Natural value of treatment"*



# Counterfactual variables under $\overline{D}_{t-1}$

$\overline{D}_{t-1}$ generates **counterfactual variables at time $t$**:

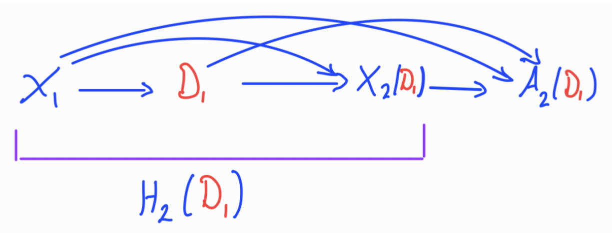
- $X_t(\overline{D}_{t-1}) = f_{X,t}(D_{t-1}, H_{t-1}(\overline{D}_{t-2}), U_{X,t})$
  *"Natural covariate value"*
- $H_t(\overline{D}_{t-1}) = (\overline{D}_{t-1}, \overline{X}_t(\overline{D}_{t-1}))$
  *"Natural covariate and treatment history"*
- $A_t(\overline{D}_{t-1}) = f_{A,t}(H_t(\overline{D}_{t-1}), U_{A,t})$
  *"Natural value of treatment"*



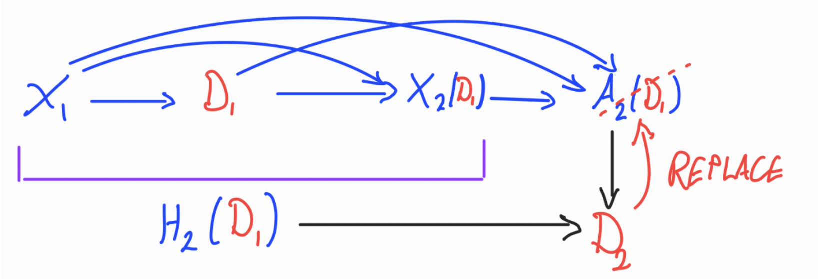
# Counterfactual variables under $\overline{D}_{t-1}$

$\overline{D}_{t-1}$ generates **counterfactual variables at time $t$**:

- $X_t(\overline{D}_{t-1}) = f_{X,t}(D_{t-1}, H_{t-1}(\overline{D}_{t-2}), U_{X,t})$
  *"Natural covariate value"*
- $H_t(\overline{D}_{t-1}) = (\overline{D}_{t-1}, \overline{X}_t(\overline{D}_{t-1}))$
  *"Natural covariate and treatment history"*
- $A_t(\overline{D}_{t-1}) = f_{A,t}(H_t(\overline{D}_{t-1}), U_{A,t})$
  *"Natural value of treatment"*



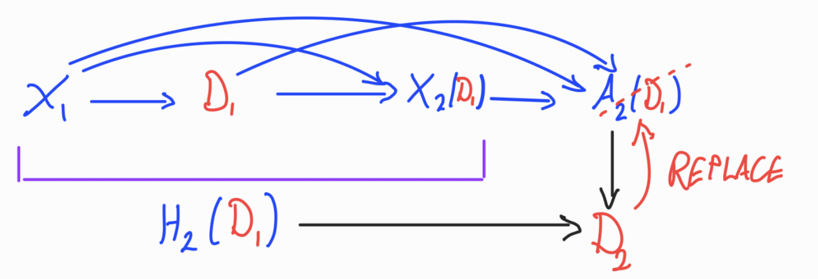
# Intervention $D_t$ and potential outcomes $\overline{D}_T$




$D_t$ can be a function of $A_t(\overline{D}_{t-1}), H_t(\overline{D}_{t-1})$.

Ultimately, replace $\overline{A}_T$ with $\overline{D}_T$:

- $Y(\overline{D}_T) = f_Y(D_T, H_T(\overline{D}_{T-1}), U_Y)$
  *"Counterfactual outcomes under $\overline{D}_T$"*




$D_t$ can be a function of $A_t(\overline{D}_{t-1}), H_t(\overline{D}_{t-1})$.

Ultimately, replace $\overline{A}_T$ with $\overline{D}_T$:

- $Y(\overline{D}_T) = f_Y(D_T, H_T(\overline{D}_{T-1}), U_Y)$
  *"Counterfactual outcomes under $\overline{D}_T$"*

# Longitudinal trimming with flip interventions

# Longitudinal flip interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $I_t^c \equiv I_t^c(a_t; h_t) = \mathbb{1}\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\} > \varepsilon\right]$ is trimming indicator

**Flip intervention** at time $t$ targeting $a_t$ is

$$\boxed{D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ I_t^c a_t + \left(1 - I_t^c\right) A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}}$$

At $t$, if the natural value of treatment is **already $a_t$, do nothing**; otherwise, "**flip**" the subject to $a_t$ if their **natural history lies in the trimmed set** (determined by $I_t^c$).

# Longitudinal flip interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $I_t^c \equiv I_t^c(a_t; h_t) = \mathbb{1}\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\} > \varepsilon\right]$ is trimming indicator

**Flip intervention** at time $t$ targeting $a_t$ is

$$D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ I_t^c a_t + (1 - I_t^c) A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}$$

At $t$, if the natural value of treatment is **already $a_t$, do nothing**; otherwise, "**flip**" the subject to $a_t$ if their **natural history lies in the trimmed set** (determined by $I_t^c$).

# Longitudinal flip interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $I_t^c \equiv I_t^c(a_t; h_t) = \mathbb{1}\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\} > \varepsilon\right]$ is trimming indicator

**Flip intervention** at time $t$ targeting $a_t$ is

$$D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ I_t^c a_t + \left(1 - I_t^c\right) A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}$$

At $t$, if the natural value of treatment is **already $a_t$, do nothing**; otherwise, "**flip**" the subject to $a_t$ if their **natural history lies in the trimmed set** (determined by $I_t^c$).

# Longitudinal flip interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $I_t^c \equiv I_t^c(a_t; h_t) = \mathbb{1}\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\} > \varepsilon\right]$ is trimming indicator

**Flip intervention** at time $t$ targeting $a_t$ is

$$\boxed{D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ I_t^c a_t + \left(1 - I_t^c\right) A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}}$$

At $t$, if the natural value of treatment is **already $a_t$, do nothing**; otherwise, "**flip**" the subject to $a_t$ if their **natural history lies in the trimmed set** (determined by $I_t^c$).

# Theorem: identification

Let $I_t(a_t; h_t) = \mathbb{1}\{\mathbb{P}(A_t = a_t \mid h_t) > \varepsilon\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **flip interventions** and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \overbrace{\sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T) \prod_{t=1}^{T} Q_t(b_t \mid \overline{b}_{t-1}, \overline{X}_t)\right\}}^{\textbf{extended g-formula}},$$

$$= \left.\mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right]\right\}\textbf{IPW},$$

where

$$Q_t(a_t \mid h_t) = I_t(a_t; h_t) + \{1 - I_t(a_t; h_t)\}\mathbb{P}(A_t = a_t \mid h_t)$$

and $Q_t(1 - a_t \mid h_t) = 1 - Q_t(a_t \mid h_t)$.

# Theorem: identification

Let $I_t(a_t; h_t) = \mathbb{1}\{\mathbb{P}(A_t = a_t \mid h_t) > \varepsilon\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **flip interventions** and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \overbrace{\sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T) \prod_{t=1}^{T} Q_t(b_t \mid \overline{b}_{t-1}, \overline{X}_t)\right\}}^{\text{extended g-formula}},$$

$$= \left.\mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right]\right\} \textbf{IPW},$$

where

$$Q_t(a_t \mid h_t) = I_t(a_t; h_t) + \{1 - I_t(a_t; h_t)\}\mathbb{P}(A_t = a_t \mid h_t)$$

and $Q_t(1 - a_t \mid h_t) = 1 - Q_t(a_t \mid h_t)$.

# Theorem: identification

Let $I_t(a_t; h_t) = \mathbb{1}\{\mathbb{P}(A_t = a_t \mid h_t) > \varepsilon\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **flip interventions** and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \overbrace{\sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T) \prod_{t=1}^{T} Q_t(b_t \mid \overline{b}_{t-1}, \overline{X}_t)\right\}}^{\text{extended g-formula}},$$

$$= \mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right] \Big\} \textbf{IPW},$$

where

$$Q_t(a_t \mid h_t) = I_t(a_t; h_t) + \{1 - I_t(a_t; h_t)\}\mathbb{P}(A_t = a_t \mid h_t)$$

and $Q_t(1 - a_t \mid h_t) = 1 - Q_t(a_t \mid h_t)$.

# Theorem: identification

Let $I_t(a_t; h_t) = \mathbb{1}\{\mathbb{P}(A_t = a_t \mid h_t) > \varepsilon\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **flip interventions** and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \overbrace{\sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T) \prod_{t=1}^{T} Q_t(b_t \mid \overline{b}_{t-1}, \overline{X}_t)\right\}}^{\text{extended g-formula}},$$

$$= \left.\mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right]\right\}\textbf{IPW},$$

where

$$Q_t(a_t \mid h_t) = I_t(a_t; h_t) + \left\{1 - I_t(a_t; h_t)\right\}\mathbb{P}(A_t = a_t \mid h_t)$$

and $Q_t(1 - a_t \mid h_t) = 1 - Q_t(a_t \mid h_t)$.

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$

**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$ non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$

**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$ non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$
**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$ non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$

**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$ non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$
**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$
   non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# We are trimming, but we can't(/shouldn't) aim for "trimmed" effects

Not necessarily true that $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\}$ isolates

$$\psi = \mathbb{E}\Big[\big\{Y(\overline{a}_T) - Y(\overline{a}'_T)\big\}\mathbb{1}\big(\text{could take both } \overline{a}_T \text{ and } \overline{a}'_T\big)\Big].$$

**Why?** because $\overline{D}_t$ affects $X_{t+1}(\overline{D}_t), A_{t+1}(\overline{D}_t)$

**However,** $\psi$ is a bad target!

**Theorem:** Other flip interventions $\overline{D}_T$ and $\overline{D}'_T$ yield $\mathbb{E}\{Y(\overline{D}_T) - Y(\overline{D}'_T)\} = \psi$. However, they have two properties:

1. They depend on cross-world propensity scores: trim supposing non-zero propensity score under $D_1 = 1$ and $D_1 = 0$
   non-falsifiable
2. They depend on future propensity scores: trim at $t = 1$ supposing non-zero propensity scores at $t = 2$

(this justifies the caution typically advised)

# Smooth trimming with S-flip interventions

# Smooth trimming

Everything we've done so far can be generalized to smooth trimming!

**Why:** flip effects inherit $\widehat{\pi}$ convergence rate

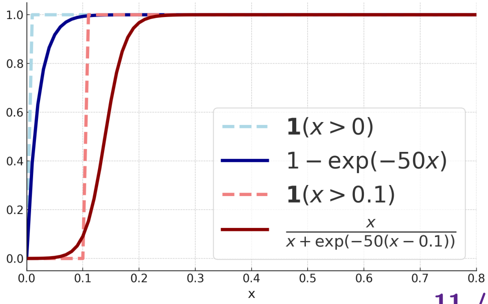
ML estimator $\implies$ slower-than-$\sqrt{n}$ convergence

Smooth trimmed effects: $\sqrt{n}$-convergence w/ ML estimators

**How:** replace $\mathbb{1}(\cdot)$ by $S(\cdot)$: $\mathbb{E}\Big[\{Y(1) - Y(0)\}S(X)\Big]$

**Examples:**

1. $\mathbb{1}(x > 0) \to$ $s(x; k) = 1 - \exp(-kx)$
2. $\mathbb{1}(x > \varepsilon) \to$ $s(x; k, \varepsilon) = \frac{x}{x + \exp\{-k(x-\varepsilon)\}}$



# Smooth trimming

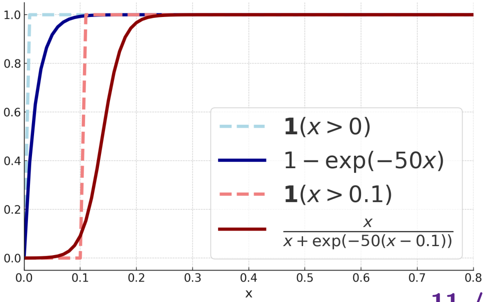
Everything we've done so far can be generalized to smooth trimming!

**Why:** flip effects inherit $\widehat{\pi}$ convergence rate
ML estimator $\Longrightarrow$ slower-than-$\sqrt{n}$ convergence
Smooth trimmed effects: $\sqrt{n}$-convergence w/ ML estimators

**How:** replace $\mathbb{1}(\cdot)$ by $S(\cdot)$: $\mathbb{E}\left[\{Y(1) - Y(0)\}S(X)\right]$

**Examples:**

1. $\mathbb{1}(x > 0) \rightarrow$
$s(x; k) = 1 - \exp(-kx)$
2. $\mathbb{1}(x > \varepsilon) \rightarrow$
$s(x; k, \varepsilon) = \frac{x}{x + \exp\{-k(x - \varepsilon)\}}$



# Smooth trimming

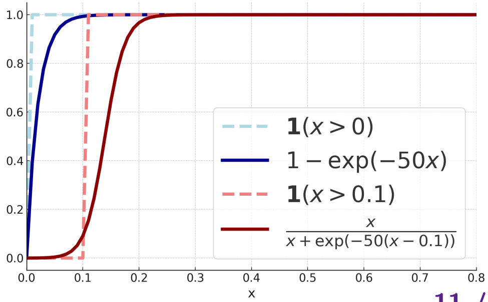
Everything we've done so far can be generalized to smooth trimming!

**Why:** flip effects inherit $\widehat{\pi}$ convergence rate
ML estimator $\implies$ slower-than-$\sqrt{n}$ convergence
Smooth trimmed effects: $\sqrt{n}$-convergence w/ ML estimators

**How:** replace $\mathbb{1}(\cdot)$ by $S(\cdot)$: $\mathbb{E}\Big[\{Y(1) - Y(0)\}S(X)\Big]$

**Examples:**

1. $\mathbb{1}(x > 0) \rightarrow$ $s(x; k) = 1 - \exp(-kx)$
2. $\mathbb{1}(x > \varepsilon) \rightarrow$ $s(x; k, \varepsilon) = \frac{x}{x + \exp\{-k(x-\varepsilon)\}}$



# Smooth trimming

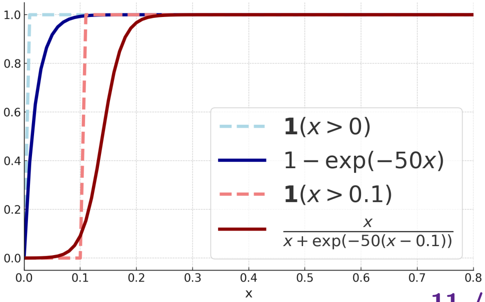
Everything we've done so far can be generalized to smooth trimming!

**Why:** flip effects inherit $\widehat{\pi}$ convergence rate
ML estimator $\implies$ slower-than-$\sqrt{n}$ convergence
Smooth trimmed effects: $\sqrt{n}$-convergence w/ ML estimators

**How:** replace $\mathbb{1}(\cdot)$ by $S(\cdot)$: $\mathbb{E}\Big[\{Y(1) - Y(0)\}S(X)\Big]$

**Examples:**

1. $\mathbb{1}(x > 0) \to$
   $s(x; k) = 1 - \exp(-kx)$
2. $\mathbb{1}(x > \varepsilon) \to$
   $s(x; k, \varepsilon) = \frac{x}{x + \exp\{-k(x-\varepsilon)\}}$



# Smooth-flip (S-flip) interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $S_t^c \equiv S_t^c(a_t; h_t) = s\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\}, k, \varepsilon\right]$ is *smooth* trimming indicator

**S-flip intervention** at time $t$ targeting $a_t$ is

$$D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ \mathbb{1}(V_t > S_t^c)a_t + \mathbb{1}(V_t \leq S_t^c)A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}$$

and $V_1, \ldots, V_T \overset{iid}{\sim} \text{Uniform}(0,1)$ and $V_t \perp\!\!\!\perp Z$ for all $t \leq T$.

At time $t$, if the natural value of treatment is already $a_t$, **do nothing**; otherwise, "**flip**" the subject to $a_t$ **with probability $S_t^c$**.

$s(x) = 1 - \exp(-50x)$ approximates $\mathbb{1}(x > 0)$ :
$s(0) = 0, s(0.01) = 0.39, s(0.1) = 0.99$

# Smooth-flip (S-flip) interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $S_t^c \equiv S_t^c(a_t; h_t) = s\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\}, k, \varepsilon\right]$ is *smooth* trimming indicator

**S-flip intervention** at time $t$ targeting $a_t$ is

$$\boxed{D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ \mathbb{1}(V_t > S_t^c)a_t + \mathbb{1}(V_t \leq S_t^c)A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}}$$

and $V_1, \ldots, V_T \overset{iid}{\sim} \text{Uniform}(0,1)$ and $V_t \perp\!\!\!\perp Z$ for all $t \leq T$.

At time $t$, if the natural value of treatment is already $a_t$, **do nothing**; otherwise, "**flip**" the subject to $a_t$ **with probability** $S_t^c$.

$s(x) = 1 - \exp(-50x)$ approximates $\mathbb{1}(x > 0)$ :
$s(0) = 0, s(0.01) = 0.39, s(0.1) = 0.99$

# Smooth-flip (S-flip) interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $S_t^c \equiv S_t^c(a_t; h_t) = s\big[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\}, k, \varepsilon\big]$ is *smooth* trimming indicator

**S-flip intervention** at time $t$ targeting $a_t$ is

$$\boxed{D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ \mathbb{1}(V_t > S_t^c)a_t + \mathbb{1}(V_t \leq S_t^c)A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}}$$

and $V_1, \ldots, V_T \overset{iid}{\sim} \text{Uniform}(0,1)$ and $V_t \perp\!\!\!\perp Z$ for all $t \leq T$.

At time $t$, if the natural value of treatment is already $a_t$, **do nothing**; otherwise, "**flip**" the subject to $a_t$ **with probability $S_t^c$**.

$s(x) = 1 - \exp(-50x)$ approximates $\mathbb{1}(x > 0)$ :
$s(0) = 0, s(0.01) = 0.39, s(0.1) = 0.99$

# Smooth-flip (S-flip) interventions

- $\overline{a}_T \in \{0,1\}^T$ is the target regime
- $S_t^c \equiv S_t^c(a_t; h_t) = s\left[\mathbb{P}\{A_t(\overline{D}_{t-1}) = a_t \mid H_t(\overline{D}_{t-1}) = h_t\}, k, \varepsilon\right]$ is *smooth* trimming indicator

**S-flip intervention** at time $t$ targeting $a_t$ is

$$\boxed{D_t(a_t) = \begin{cases} A_t(\overline{D}_{t-1}), & \text{if } A_t(\overline{D}_{t-1}) = a_t, \\ \mathbb{1}(V_t > S_t^c)a_t + \mathbb{1}(V_t \le S_t^c)A_t(\overline{D}_{t-1}), & \text{otherwise,} \end{cases}}$$

and $V_1, \ldots, V_T \overset{iid}{\sim} \text{Uniform}(0,1)$ and $V_t \perp\!\!\!\perp Z$ for all $t \le T$.

At time $t$, if the natural value of treatment is already $a_t$, **do nothing**; otherwise, "**flip**" the subject to $a_t$ **with probability $S_t^c$**.

$s(x) = 1 - \exp(-50x)$ approximates $\mathbb{1}(x > 0)$ :
$s(0) = 0, s(0.01) = 0.39, s(0.1) = 0.99$

# Theorem: S-flip effect identification

Let $S_t(a_t; h_t) = s\{\mathbb{P}(A_t = a_t \mid h_t), \varepsilon, k\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **S-flip interventions**,
- $s(0) = 0$, and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T) \prod_{t=1}^{T} Q_t(b_t \mid \overline{b}_{t-1}, \overline{X}_t)\right\},$$

$$= \mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right],$$

where

$$\boxed{Q_t(a_t \mid h_t) = S_t(a_t; h_t) + \{1 - S_t(a_t; h_t)\}\mathbb{P}(A_t = a_t \mid h_t)}$$

$(S_t(a_t; h_t) \approx 1 \implies Q_t(a_t \mid h_t) \approx 1)$

# Theorem: S-flip effect identification

Let $S_t(a_t; h_t) = s\{\mathbb{P}(A_t = a_t \mid h_t), \varepsilon, k\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \ldots, D_T(a_T)\}$ are **S-flip interventions**,
- $s(0) = 0$, and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}\left(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T\right) \prod_{t=1}^{T} Q_t\left(b_t \mid \overline{b}_{t-1}, \overline{X}_t\right)\right\},$$

$$= \mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right],$$

where

$$\boxed{Q_t(a_t \mid h_t) = S_t(a_t; h_t) + \{1 - S_t(a_t; h_t)\}\mathbb{P}(A_t = a_t \mid h_t)}$$

$(S_t(a_t; h_t) \approx 1 \implies Q_t(a_t \mid h_t) \approx 1)$

# Theorem: S-flip effect identification

Let $S_t(a_t; h_t) = s\{\mathbb{P}(A_t = a_t \mid h_t), \varepsilon, k\}$. Suppose

- $\overline{D}_T = \{D_1(a_1), D_2(a_2), \dots, D_T(a_T)\}$ are **S-flip interventions**,
- $s(0) = 0$, and
- the **NPSEM and strong SR** hold.

Then,

$$\mathbb{E}\{Y(\overline{D}_T)\} = \sum_{\overline{b}_T \in \{0,1\}^T} \mathbb{E}\left\{\mathbb{E}\left(Y \mid \overline{A}_T = \overline{b}_T,\ \overline{X}_T\right) \prod_{t=1}^{T} Q_t\left(b_t \mid \overline{b}_{t-1}, \overline{X}_t\right)\right\},$$

$$= \mathbb{E}\left[Y \prod_{t=1}^{T} \frac{Q_t(A_t \mid H_t)}{\mathbb{P}(A_t \mid H_t)}\right],$$

where

$$\boxed{Q_t(a_t \mid h_t) = S_t(a_t; h_t) + \left\{1 - S_t(a_t; h_t)\right\}\mathbb{P}(A_t = a_t \mid h_t)}$$

$(S_t(a_t; h_t) \approx 1 \implies Q_t(a_t \mid h_t) \approx 1)$

# Estimation

(in two slides)

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\implies$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\implies$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\Longrightarrow$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\Longrightarrow$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\implies$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Estimation of S-flip effects

S-flip effects are pathwise differentiable $\implies$ **efficient influence function** (EIF)-based estimators

We derive the new EIF (plug-in plus weighted residuals)

Inspires two one-step estimators:

1. Multiply robust-style
2. Sequentially doubly robust-style (debias pseudo-outcome in multiply robust-style estimator)

We derive bias bounds, which imply weak convergence for S-flip estimator under nonparametric conditions on nuisance functions

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \left(\|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\|\right).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$–consistency & asymptotic normality

# Bias bounds

Two new results for multiply robust-style estimator:

1. Observe that bias is **minimum** of two errors
   Díaz et al. [2023] & Kennedy [2019]: unroll into future and past
   $\rightarrow$ Hence, $2(T+1)$ multiply robust-style bound
2. Tighten bias bound from IPSI-style result [Kennedy, 2019]

Sequentially DR-style result is **first-of-its-kind** where $Q_t$ depends on unknown propensity score

**Theorem, informal:** Let $\psi$ denote S-flip effect, $\pi_t$ prop. score at $t$, $\widetilde{m}_t$ denote seq. reg. at $t$ with estimated pseudo-outcome. Then,

$$\boxed{\left|\mathbb{E}\left(\widehat{\psi}_{sdr} - \psi\right)\right| \lesssim \sum_{t=1}^{T} \|\widehat{\pi}_t - \pi_t\| \Big( \|\widehat{m}_t - \widetilde{m}_t\| + \|\widehat{\pi}_t - \pi_t\| \Big).}$$

$\overbrace{\|\widehat{\pi}_t - \pi_t\| = o_{\mathbb{P}}(n^{-1/4}) = \|\widehat{m}_t - \widetilde{m}_t\|}^{\text{ML conv. rates}} \implies \sqrt{n}$−consistency & asymptotic normality

# Thank you!

Recap:

- $T = 1$:
  trimming $\equiv$ flip interventions
  smooth trimming $\equiv$ S-flip ints
- $T > 1$: not so simple.
  Flip/S-flip ints avoid positivity asmp and cross-world / future-dependence
- Efficient estimation of S-flip effects: (1) multiply robust and (2) sequentially doubly robust

Ongoing work:

- Data analysis (administrative censoring and censoring by death)

hadera01@nyu.edu

*Prelim draft:*

alecmcclean.github.io/files/LSTTEs-short.pdf

Iván Díaz, Nicholas Williams, Katherine L Hoffman, and Edward J Schenck. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857, 2023.

Aksel KG Jensen, Theis Lange, Olav L Schjørring, and Maya L Petersen. Identification and responses to positivity violations in longitudinal studies: an illustration based on invasively mechanically ventilated icu patients. *Biostatistics & Epidemiology*, 8(1):e2347709, 2024.

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Alexander W Levis, Edward H Kennedy, Alec McClean, Sivaraman Balakrishnan, and Larry Wasserman. Stochastic interventions, sensitivity analysis, and optimal transport. *arXiv preprint arXiv:2411.14285*, 2024.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

# Additional remarks

1. Robustness to positivity violations $\implies$ **S-flips are alternative to IPSIs that target a specific regime**
2. **Relax dependence on natural value of treatment:**

$$D_t(a_t) = \mathbb{1}\left[V_t \leq \mathbb{1}(a_t = 1)I_t^c + (1 - I_t^c)\mathbb{P}\{A_t(\overline{D}_{t-1}) = 1 \mid H_t(\overline{D}_{t-1})\}\right]$$

   - Can avoid practical issues — we may not observe natural value of treatment
   - Identification only requires standard sequential randomization
3. Flip interventions inspired by **maximally coupled policies** Levis et al. [2024]