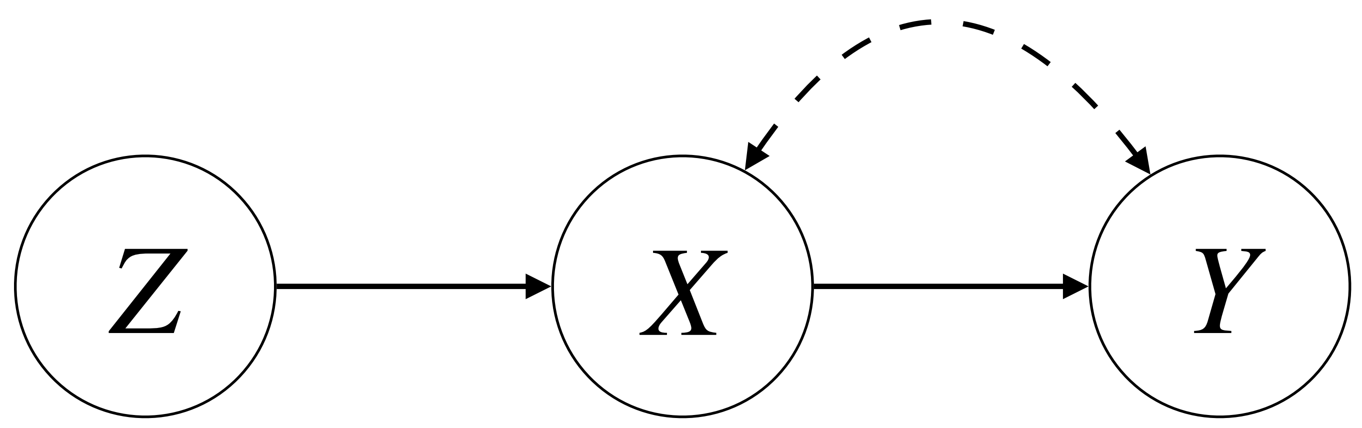
Counterfactual Identification Under Monotonicity Constraints

Aurghya Maiti, Drago Plecko, Elias Bareinboim
Causal Artificial Intelligence Lab
Columbia University

39th AAAI Conference on Artificial Intelligence
Philadelphia, 2025

A Puzzle!

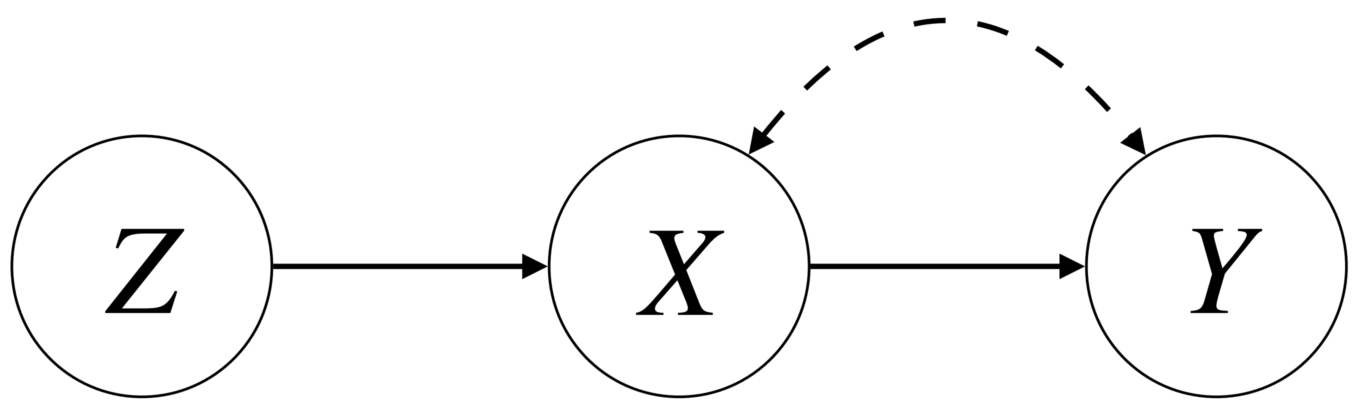
Instrument	Z	Invitation to the training program
Treatment	X	Participation in the training program
Outcome	Y	Performance



IV Graph

A Puzzle!

- For any combination of instrument Z and treatment X, there are four groups:
  - **Always-takers**: take the treatment even if they are assigned to control group, $X_{z_0} = 1,\ X_{z_1} = 1$ (or $f_X(z) = 1$)
  - **Never-takers:** do not take the treatment even if they are assigned to treatment group, $X_{z_0} = 0,\ X_{z_1} = 0$ (or $f_X(z) = 0$)
  - **Compliers:** take the treatment if and only if they are assigned to treatment group, $X_{z_0} = 0,\ X_{z_1} = 1$ (or $f_X(z) = z$)
  - **Defiers:** do the opposite of treatment assignment status, $X_{z_0} = 1.\ X_{z_1} = 0$ (or $f_X(z) = 1 - z$)

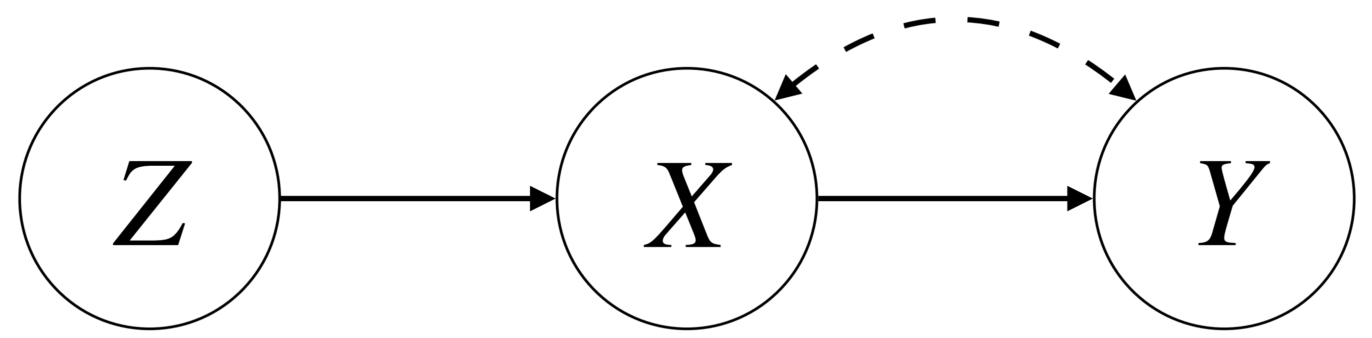


IV Graph

Local Average Treatment Effect

Question: *What is the effect of X on Y for the group of people who "comply" with the instrument Z?*

- *This quantity is not uniquely identifiable in general.*
- *Can it be identified when X is monotonically dependent on Z?*



$$E[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1]$$

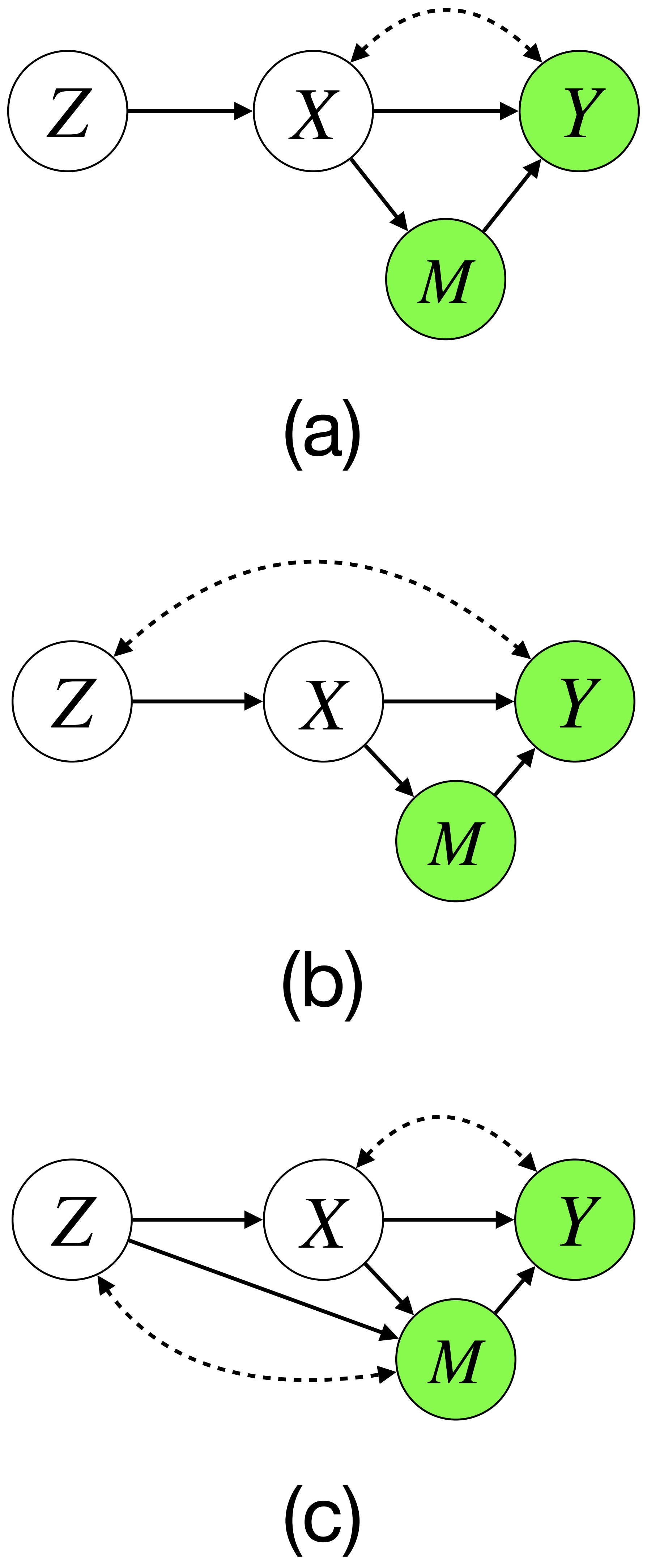
$$X_{z_1}(\mathbf{u}) \geq X_{z_0}(\mathbf{u}) \quad \forall \mathbf{u} \in D(\mathbf{U})$$

- In this example, employees cannot participate in the program without being invited. Hence, it satisfies the monotonicity constraint.

- The Nobel Prize in Economics in 2021 was awarded to David Card, Joshua Angrist, and Guido Imbens for their work on identifying LATE from observational data.

LATE

- The theory is still limited to special cases, such as instrumental variables (IVs).
- In reality, this entails strong assumptions about the underlying environment.
  - Graph (a) can be solved similarly to LATE.
  - What about models shown in (b) or (c)?
    - They violate the assumption that $Y_x \perp Z$.
      - Should we give up?



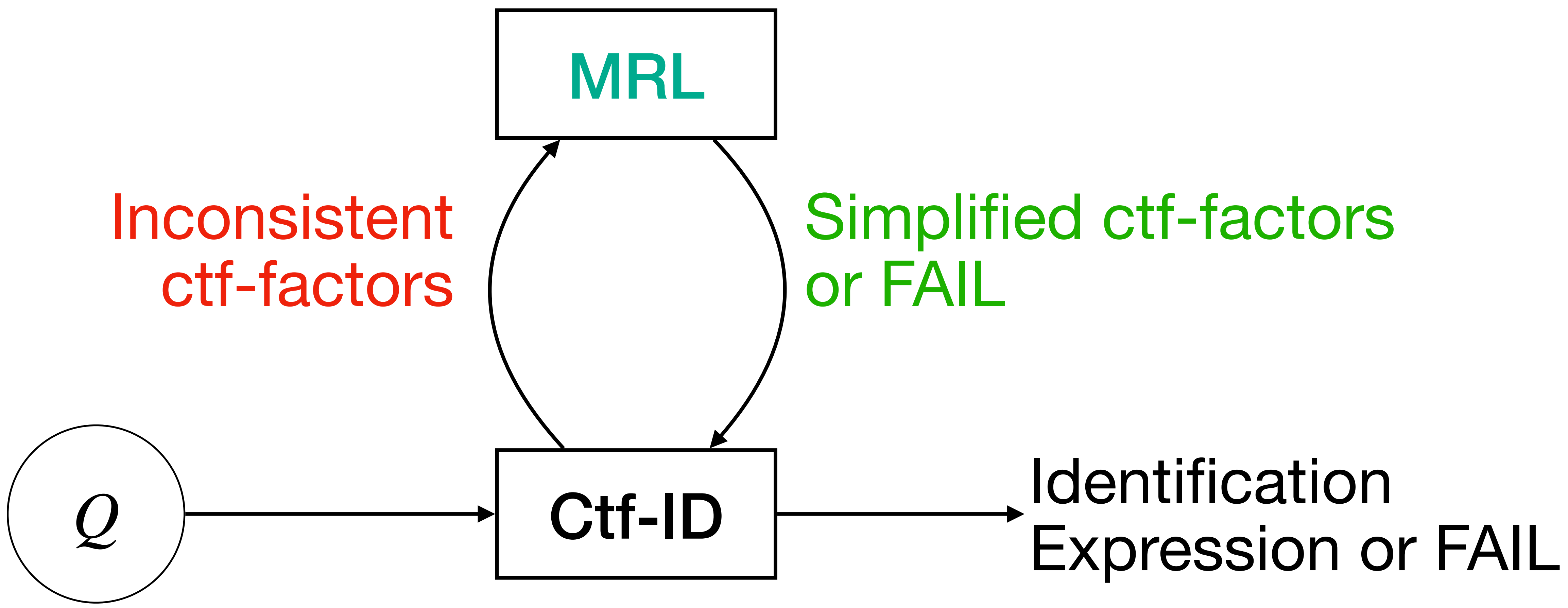
Contributions

- In this work, we generalize available machinery beyond IV settings, and develop the first **general algorithm** to identify LATE in an *arbitrary environment with monotonicity constraints*.
- This algorithm is also capable of evaluating other **counterfactual quantities**, such as direct and indirect effects, effects with post-treatment conditioning, thereby broadening the identification toolbox under popular parametric conditions.
- In doing so, we challenge the prior belief that "causal diagrams have difficulty encoding shape restrictions such as monotonicity" (Imbens 2020).

A General Algorithmic Approach

Monotonicity Reduction Lemma to simplify non-identifiable counterfactual (ctf) factors.

- W : a binary variable
- $\mathbf{T}$: set of monotonic parents of W
- $\mathbf{S}$: set of non-monotonic parents of W



Ctf-factors: $P(X_{1[pa_1]} = x_1, \ \dots \ X_{n[pa_n]} = x_n)$

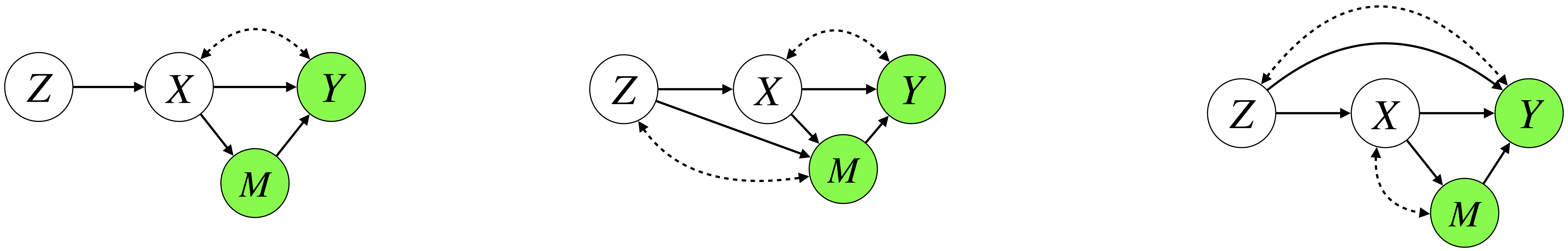
<u>Simplification Rule:</u> For $\mathbf{t} \leq \mathbf{t}'$,

- $P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 0) = P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 0)$
- $P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) = P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1)$

<u>Difference Rule:</u> For $\mathbf{t} \leq \mathbf{t}'$,

$$
\begin{aligned}
&P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) \\
&= P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 1) \\
&= P(\mathbf{Y}_*, W_{\mathbf{t},\mathbf{s}} = 0) - P(\mathbf{Y}_*, W_{\mathbf{t}',\mathbf{s}} = 0)
\end{aligned}
$$

Local Natural Direct/Indirect Effect (LN{DE, IE})



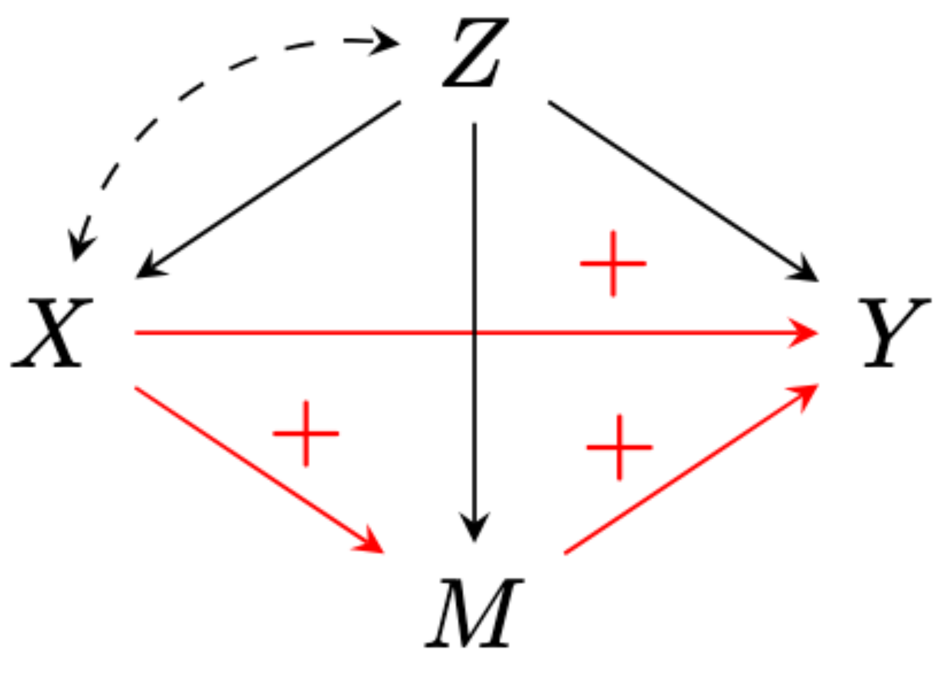
(a) (b) (c)

$$\mathrm{LNIE}(x) := \mathbb{E}[Y_{x,M_{x_1}} - Y_{x,M_{x_0}} \mid X_{z_0} = 0, X_{z_1} = 1]$$
$$\mathrm{LNDE}(x) := \mathbb{E}[Y_{x_1,M_x} - Y_{x_0,M_x} \mid X_{z_0} = 0, X_{z_1} = 1]$$

- $Y_x \perp Z$ is not satisfied in Graphs (b) and (c). However, LNDE, LNIE, and LATE are still computable from observational data.

Post-Treatment Conditioning

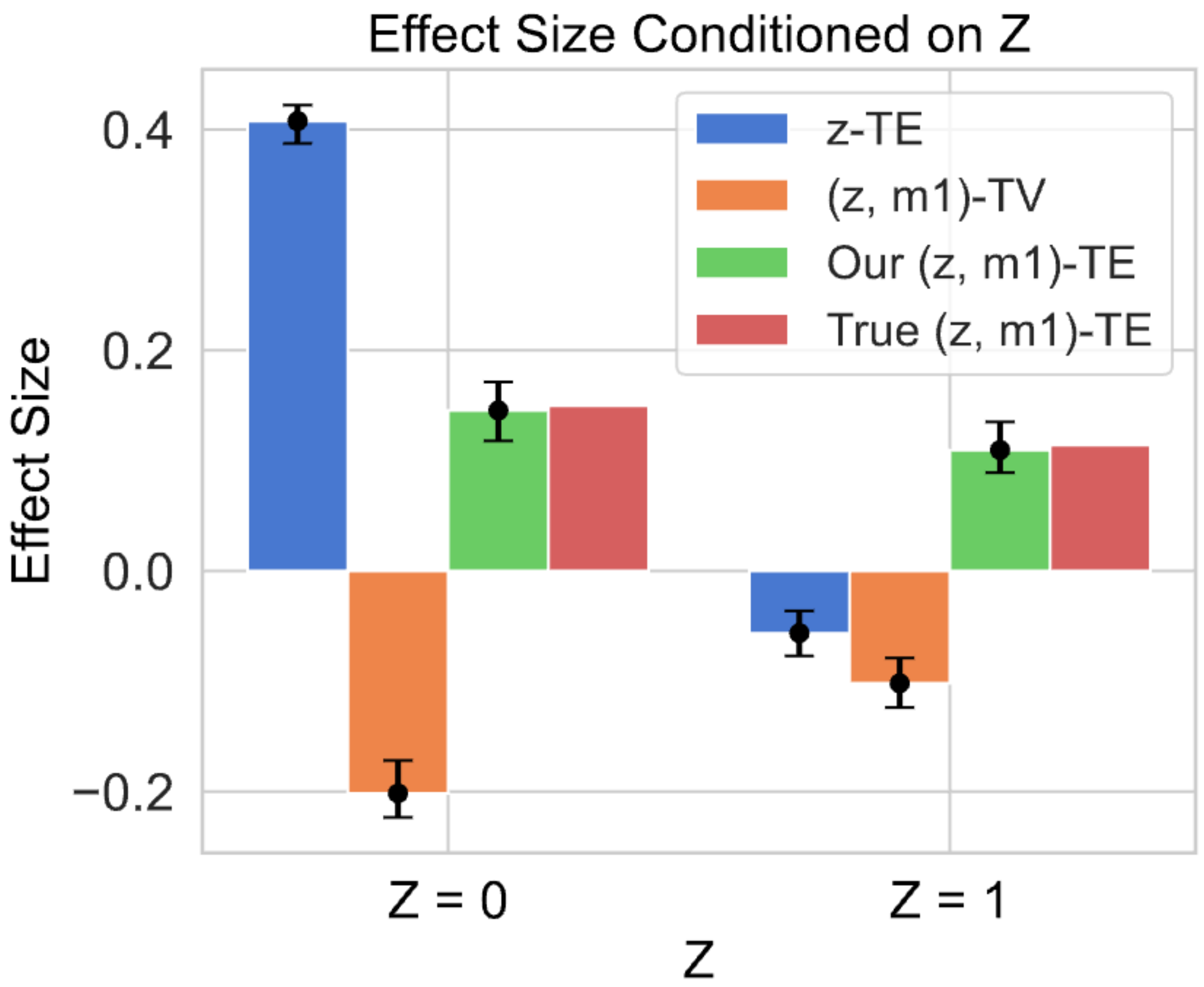
m-specific Total Effects



Target Quantity: $P(y_x \mid y', x')$

Effect of Interest: For a person of fixed age (z) and education level (m), how would their income change (y) if sex had been equal to male (x_1), compared to had it been equal to female (x_0)?

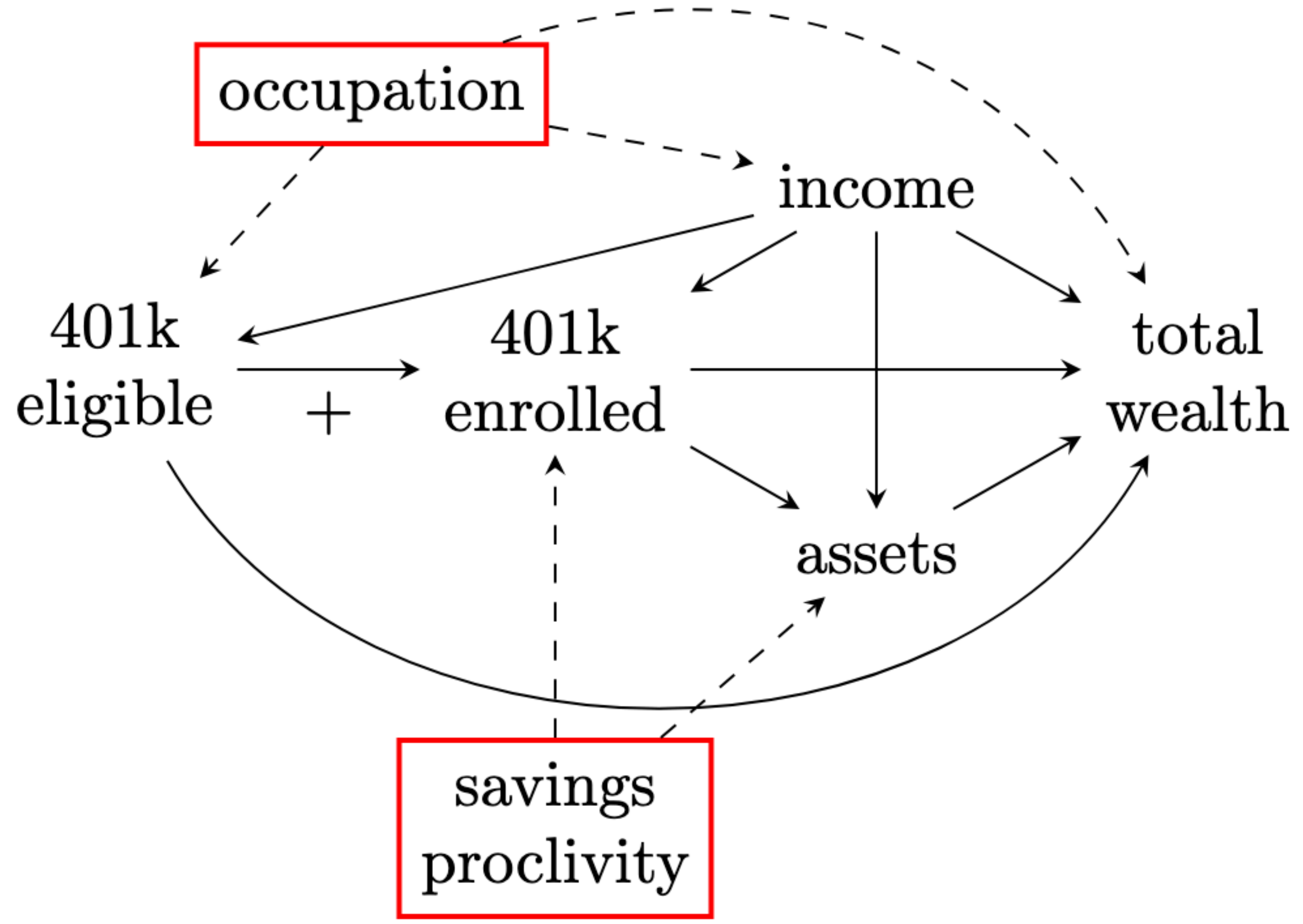
$$(z, m)\text{-}TE_{x_0,x_1}(y) = E[Y_{x_1} - Y_{x_0} \mid z, m]$$

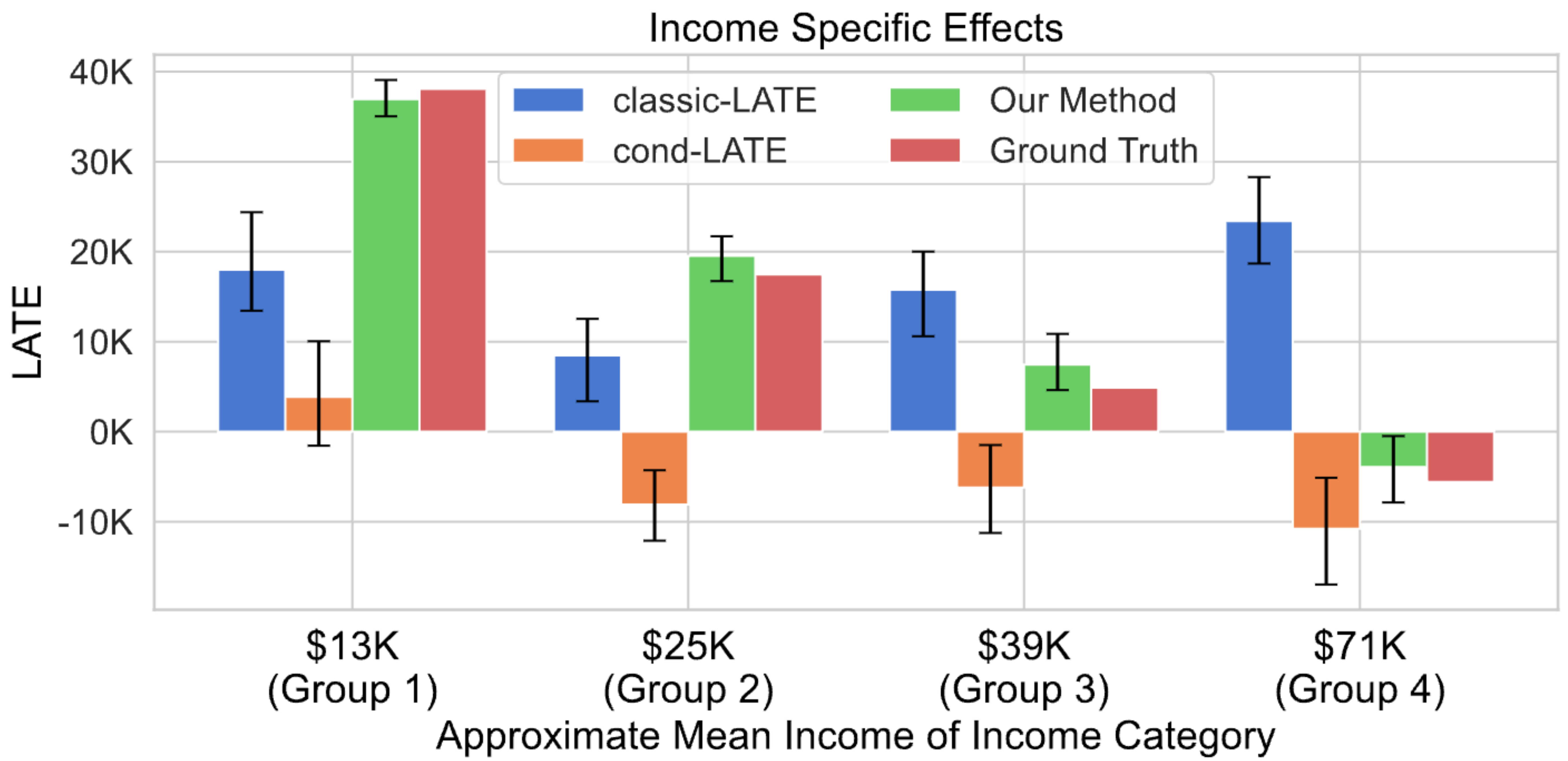


Identifying LATE in 401(k) Dataset

Effect of Interest: What is the LATE of 401(k) enrollment on total wealth for different income groups?

- Not uniquely identifiable from previous methods, as the scenario fails to satisfy some of their assumptions.
- Uniquely and correctly identifiable using our method





Thank You