Quantifying Discrimination in Contaminated Data with Automatic Causal Bounds

10 November 2021

Dean Knox **Penn**

Data is often fragmented

- Scientists often have incomplete information
    - Limited subsets of units, limited aspects of process

Data is often fragmented

- Scientists often have incomplete information
  - Limited subsets of units, limited aspects of process
- **Example:** discrimination in police-civilian interactions
  - Many police-civilian encounters not captured in data, inconsistent record-keeping in documented encounters

[1] Dean Knox. 2021. "Revealing racial bias: Causal inference can make sense of imperfect policing data." *Science*.

[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

[3] Dean Knox and Jonathan Mummolo. 2020. "Toward a General Causal Framework for the Study of Racial Bias in Policing." *JPIPE*.

Data is often fragmented

- Scientists often have incomplete information
  - Limited subsets of units, limited aspects of process

- **Example:** discrimination in police-civilian interactions
  - Many police-civilian encounters not captured in data, inconsistent record-keeping in documented encounters

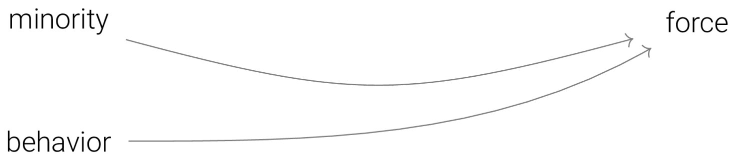
- Fragmented data leads to fragmented literatures
  - Proliferation of incompatible analytic approaches
  - Unstated, often contradictory modeling assumptions
  - Makes knowledge accumulation virtually impossible

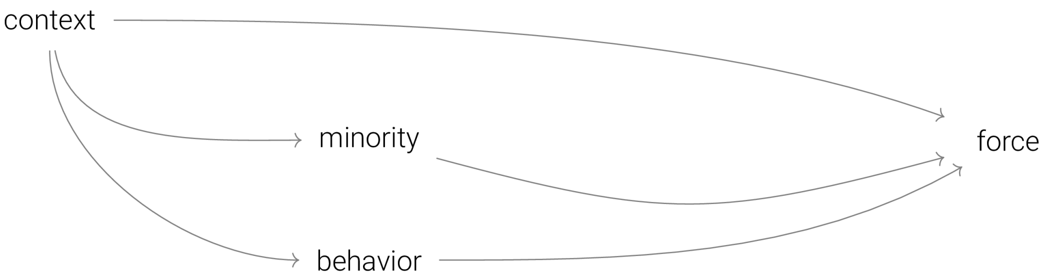
[1] Dean Knox. 2021. "Revealing racial bias: Causal inference can make sense of imperfect policing data." *Science*.

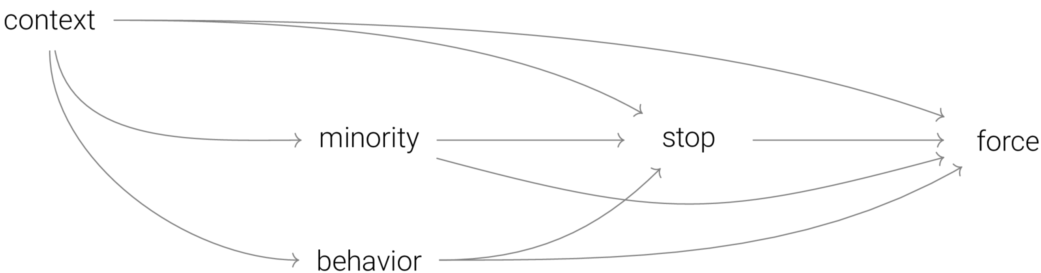
[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

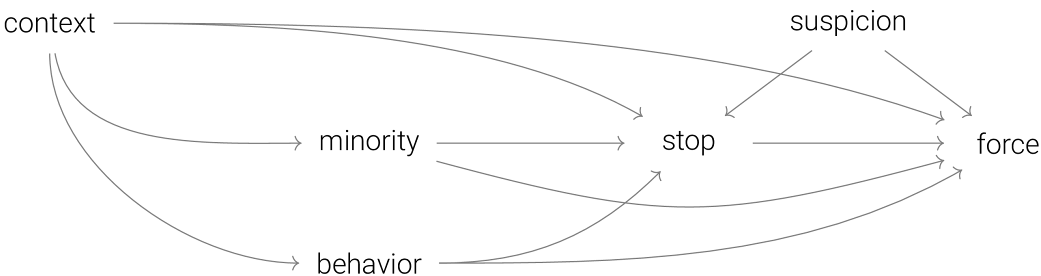
[3] Dean Knox and Jonathan Mummolo. 2020. "Toward a General Causal Framework for the Study of Racial Bias in Policing." *JPIPE*.

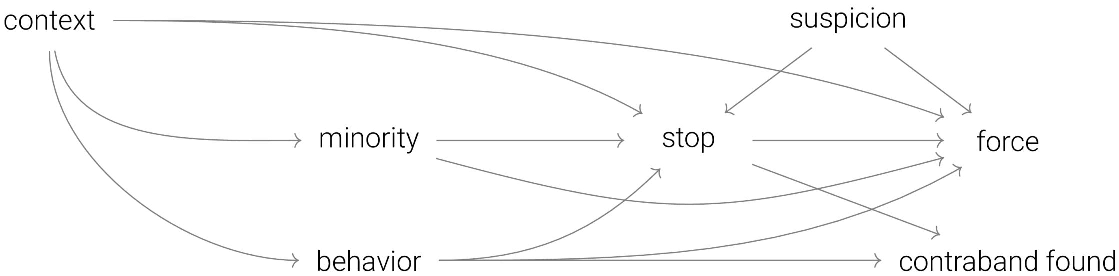
$$\begin{tikzcd} \text{minority} \arrow[r, bend right] & \text{force} \end{tikzcd}$$


minority
behavior
force


context
minority
behavior
force


context
minority
stop
force
behavior


context
suspicion
minority
stop
force
behavior



context
suspicion
minority
stop
force
behavior
contraband found

Bounds can reconcile claims

- Scientists often need to reconcile competing claims
    - Scholars, plaintiffs, external monitors, police departments
    - Different data access, modeling approaches, etc.

Bounds can reconcile claims

- Scientists often need to reconcile competing claims
  - Scholars, plaintiffs, external monitors, police departments
  - Different data access, modeling approaches, etc.

- **A promising approach:** nonparametric sharp bounds
  - Specify DAG, causal estimand, assumptions, available data

[1] Dean Knox, Teppei Yamamoto, Matthew Baum, and Adam Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *JASA*.

[2] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[3] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

Bounds can reconcile claims

- Scientists often need to reconcile competing claims
  - Scholars, plaintiffs, external monitors, police departments
  - Different data access, modeling approaches, etc.

- **A promising approach:** nonparametric sharp bounds
  - Specify DAG, causal estimand, assumptions, available data
  - Claims outside the bounds can be immediately rejected
  - Claims inside the bounds must explain where additional info. comes from (e.g. hidden assumptions, parametric models)

[1] Dean Knox, Teppei Yamamoto, Matthew Baum, and Adam Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *JASA*.

[2] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[3] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

Bounds can reconcile claims

- Scientists often need to reconcile competing claims
  - Scholars, plaintiffs, external monitors, police departments
  - Different data access, modeling approaches, etc.
- **A promising approach:** nonparametric sharp bounds
  - Specify DAG, causal estimand, assumptions, available data
  - Claims outside the bounds can be immediately rejected
  - Claims inside the bounds must explain where additional info. comes from (e.g. hidden assumptions, parametric models)
- **The challenge:** hard to derive analytically
  - Provide general-purpose bounding algorithm for any discrete causal system, any estimand & any information environment

[1] Dean Knox, Teppei Yamamoto, Matthew Baum, and Adam Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *JASA*.

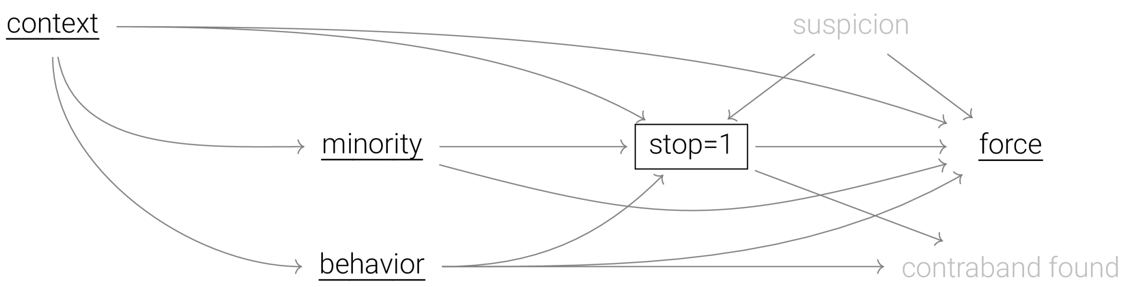
[2] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[3] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

Reconciling competing claims in the study of discrimination

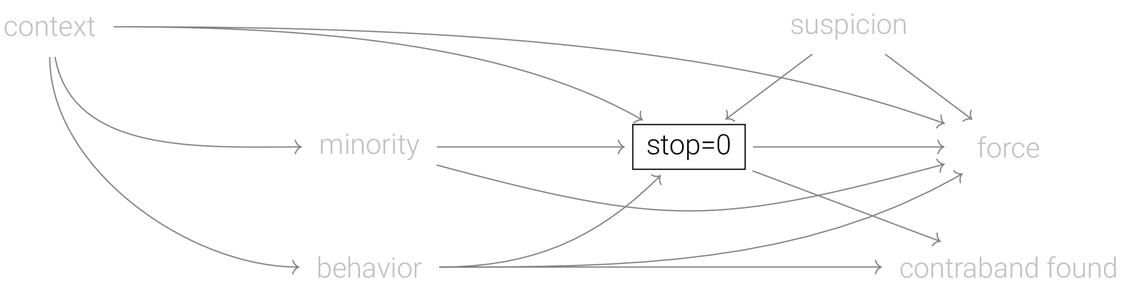
Fryer ('19), *JPE*

"... compelling case that there is no discrimination in officer-involved shootings" and reports surprisingly little discrimination in nonlethal force



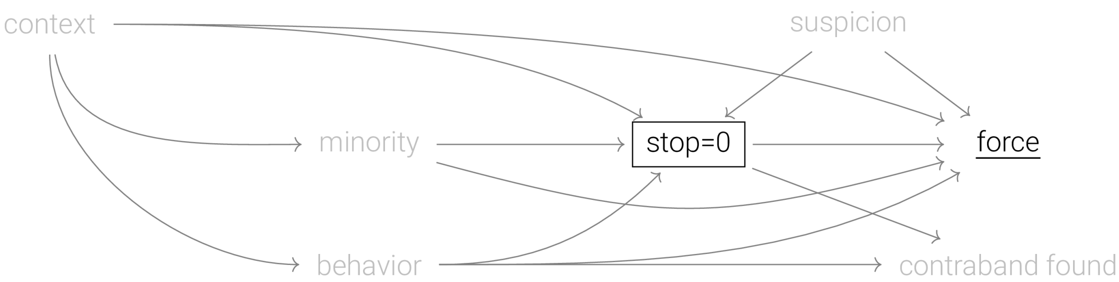
Fryer ('19), *JPE*

"... compelling case that there is no discrimination in officer-involved shootings" and reports surprisingly little discrimination in nonlethal force



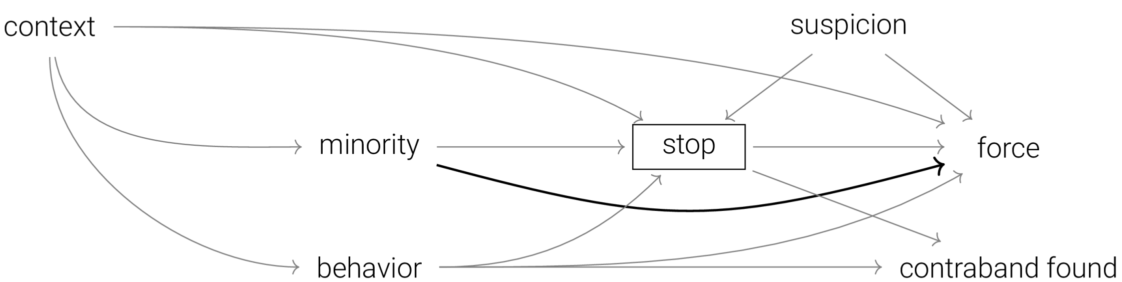
Fryer ('19), *JPE*

"... compelling case that there is no discrimination in officer-involved shootings" and reports surprisingly little discrimination in nonlethal force



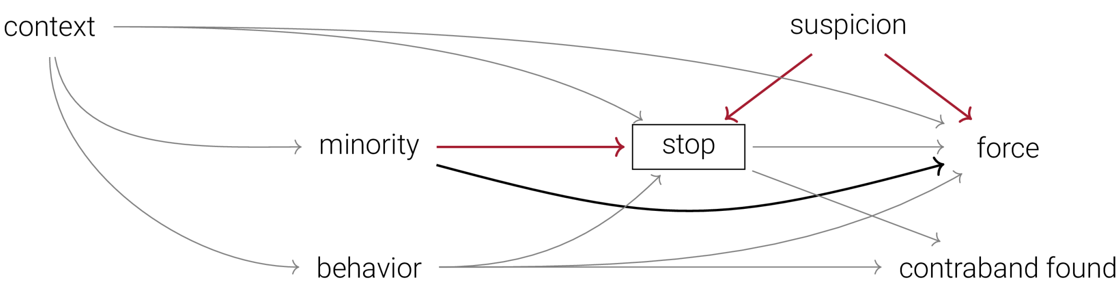
Fryer ('19), *JPE*

Conditioning on detainment records inherently introduces collider bias



Fryer ('19), *JPE*

Conditioning on detainment records inherently introduces collider bias



Fryer ('19), *JPE*

- Data environment:
  - **Observed:** Pr(minority, force | Stop=1), Pr(force | Stop=0)
  - **Unobserved:** Pr(stop), Pr(minority, force | Stop=0)
- Classic case of post-treatment selection
  - Minorities stopped for jaywalking, white civs. only for robbery

[1] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[2] Steven Durlauf and James Heckman. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment." *JPE*.

Fryer ('19), *JPE*

- Data environment:
  - **Observed:** Pr(minority, force | Stop=1), Pr(force | Stop=0)
  - **Unobserved:** Pr(stop), Pr(minority, force | Stop=0)
- Classic case of post-treatment selection
  - Minorities stopped for jaywalking, white civs. only for robbery
  - **Result:** analyzing stop records → comparing force rates used against min. jaywalkers & robbers, vs white robbers

[1] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[2] Steven Durlauf and James Heckman. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment." *JPE*.

Fryer ('19), *JPE*

- Data environment:
  - **Observed:** Pr(minority, force | Stop=1), Pr(force | Stop=0)
  - **Unobserved:** Pr(stop), Pr(minority, force | Stop=0)
- Classic case of post-treatment selection
  - Minorities stopped for jaywalking, white civs. only for robbery
  - **Result:** analyzing stop records → comparing force rates used against min. jaywalkers & robbers, vs white robbers
- Bounds say this design is fairly uninformative
  - Yet Fryer ('19) reports point estimates
  - **Hidden asm.:** E[Stop(minority=1) - Stop(minority=0)] = 0

[1] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *APSR*.

[2] Steven Durlauf and James Heckman. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment." *JPE*.

Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions
- Fryer ('19)
  - **Problem:** post-treatment conditioning (PTC)
  - **Hidden asm.:** no discrimination in stops

Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions

- Fryer ('19)
    - **Problem:** post-treatment conditioning (PTC)
    - **Hidden asm.:** no discrimination in stops

- Gaebler, Cai, Basse, Shroff, Goel & Hill ('20)
    - **Problem:** PTC + treatment confounding

Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions

- Fryer ('19)
  - **Problem:** post-treatment conditioning (PTC)
  - **Hidden asm.:** no discrimination in stops

- Gaebler, Cai, Basse, Shroff, Goel & Hill ('20)
  - **Problem:** PTC + treatment confounding
  - **Hidden asm.:** post-treatment bias = -omitted variable bias

Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions

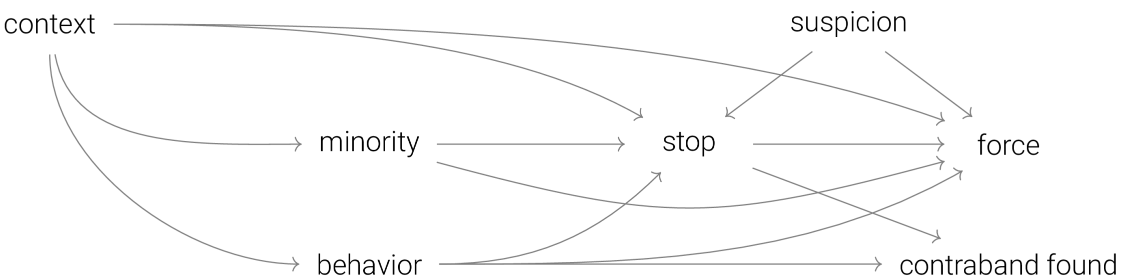
- Fryer ('19)
  - **Problem:** post-treatment conditioning (PTC)
  - **Hidden asm.:** no discrimination in stops

- Gaebler, Cai, Basse, Shroff, Goel & Hill ('20)
  - **Problem:** PTC + treatment confounding
  - **Hidden asm.:** post-treatment bias = -omitted variable bias

- Johnson, Tress, Burkel, Taylor & Cesario ('19)

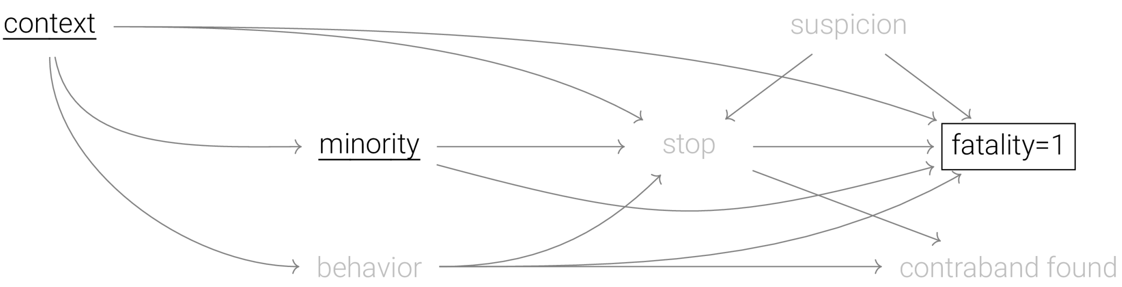
Johnson et al. ('19), *PNAS*

"We did not find evidence for anti-Black or anti-Hispanic disparity in police use of force... and, if anything, found anti-White disparities"



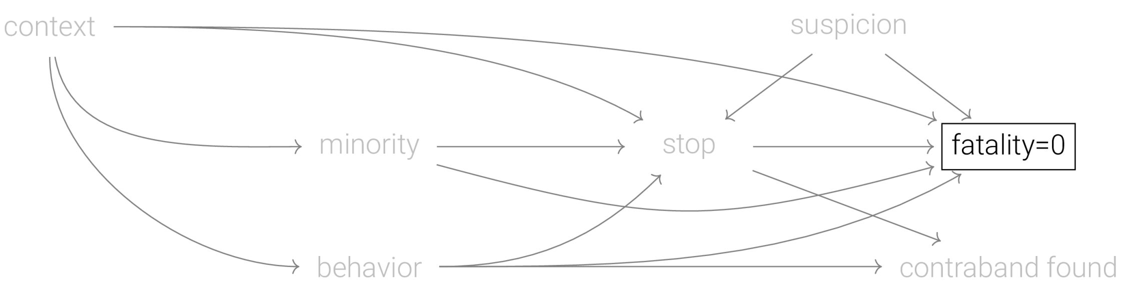
Johnson et al. ('19), *PNAS*

"We did not find evidence for anti-Black or anti-Hispanic disparity in police use of force... and, if anything, found anti-White disparities"



Johnson et al. ('19), *PNAS*

"We did not find evidence for anti-Black or anti-Hispanic disparity in police use of force... and, if anything, found anti-White disparities"



Johnson et al. ('19), *PNAS*

- Data environment:
  - **Observed:** Pr(minority | Fatality=1)
  - **Unobserved:** Pr(minority), Pr(fatality)

- Classic case of selection on dependent variable

[1] Dean Knox and Jonathan Mummolo. 2020. "Making inferences about racial disparities in police violence." *PNAS*.

[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

Johnson et al. ('19), *PNAS*

- Data environment:
  - **Observed:** Pr(minority | Fatality=1)
  - **Unobserved:** Pr(minority), Pr(fatality)
- Classic case of selection on dependent variable
- Bounds say this design is completely uninformative
  - $-1 \leq E[\text{Fatality}(\text{minority}=1) - \text{Fatality}(\text{minority}=0)] \leq 1$
  - Yet Johnson et al. ('19) reports point estimates

[1] Dean Knox and Jonathan Mummolo. 2020. "Making inferences about racial disparities in police violence." *PNAS*.

[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

Johnson et al. ('19), *PNAS*

- Data environment:
  - **Observed:** Pr(minority | Fatality=1)
  - **Unobserved:** Pr(minority), Pr(fatality)

- Classic case of selection on dependent variable

- Bounds say this design is completely uninformative
  - $-1 \leq E[\text{Fatality}(\text{minority}=1) - \text{Fatality}(\text{minority}=0)] \leq 1$
  - Yet Johnson et al. ('19) reports point estimates
  - **Hidden assumption:** Pr(Minority=1) = Pr(Minority=0) = ½

[1] Dean Knox and Jonathan Mummolo. 2020. "Making inferences about racial disparities in police violence." *PNAS*.

[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

Johnson et al. ('19), *PNAS*

- Data environment:
  - **Observed:** Pr(minority | Fatality=1)
  - **Unobserved:** Pr(minority), Pr(fatality)
- Classic case of selection on dependent variable
- Bounds say this design is completely uninformative
  - $-1 \leq E[\text{Fatality}(\text{minority}=1) - \text{Fatality}(\text{minority}=0)] \leq 1$
  - Yet Johnson et al. ('19) reports point estimates
  - **Hidden assumption:** Pr(Minority=1) = Pr(Minority=0) = ½
- Ultimately retracted after one year of harm

[1] Dean Knox and Jonathan Mummolo. 2020. "Making inferences about racial disparities in police violence." *PNAS*.

[2] Bocar Ba, Dean Knox, Jonathan Mummolo, and Roman Rivera. 2021. "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago." *Science*.

Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions

- Fryer ('19)
  - **Problem:** post-treatment conditioning (PTC)
  - **Hidden asm.:** no discrimination in stops

- Gaebler, Cai, Basse, Shroff, Goel & Hill ('20)
  - **Problem:** PTC + treatment confounding
  - **Hidden asm.:** post-treatment bias = -omitted variable bias

- Johnson, Tress, Burkel, Taylor & Cesario ('19)
  - **Problem:** selection on dependent variable
  - **Hidden asm.:** Pr(minority) = Pr(white) = ½

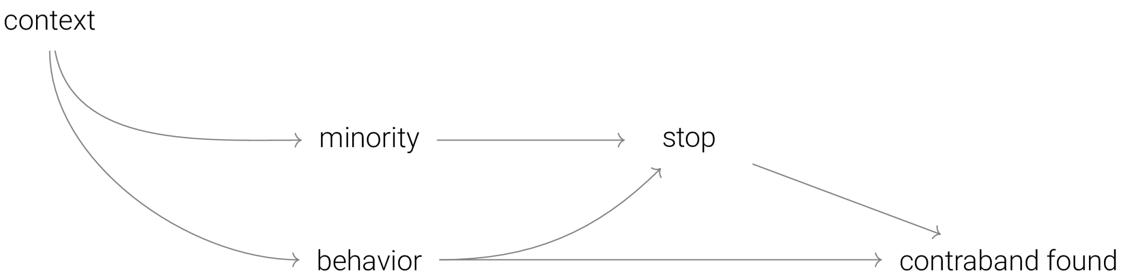
Reveal hidden assumptions

- Impossible point estimates are often reported; sharp bounds help reveal hidden assumptions
- Fryer ('19)
  - **Problem:** post-treatment conditioning (PTC)
  - **Hidden asm.:** no discrimination in stops
- Gaebler, Cai, Basse, Shroff, Goel & Hill ('20)
  - **Problem:** PTC + treatment confounding
  - **Hidden asm.:** post-treatment bias = -omitted variable bias
- Johnson, Tress, Burkel, Taylor & Cesario ('19)
  - **Problem:** selection on dependent variable
  - **Hidden asm.:** Pr(minority) = Pr(white) = ½
- Shoddy work on high-stakes policy has consequences

Sharp bounds can also build on existing methods

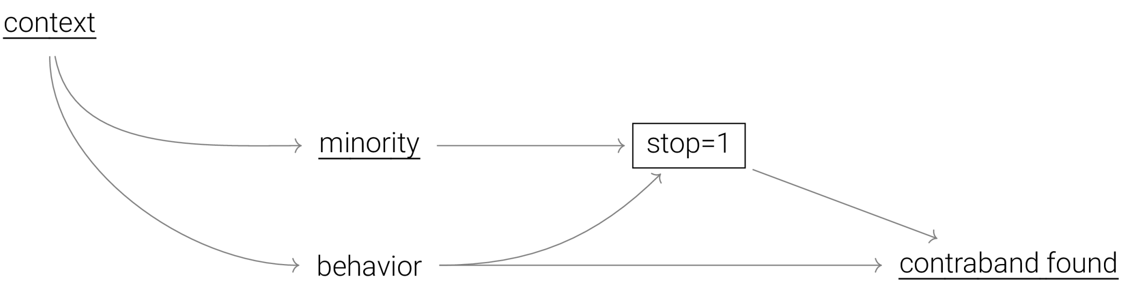
Sharp bounds can also build on existing methods

Outcome test exploits collider bias
to construct an indirect test of null discrimination.



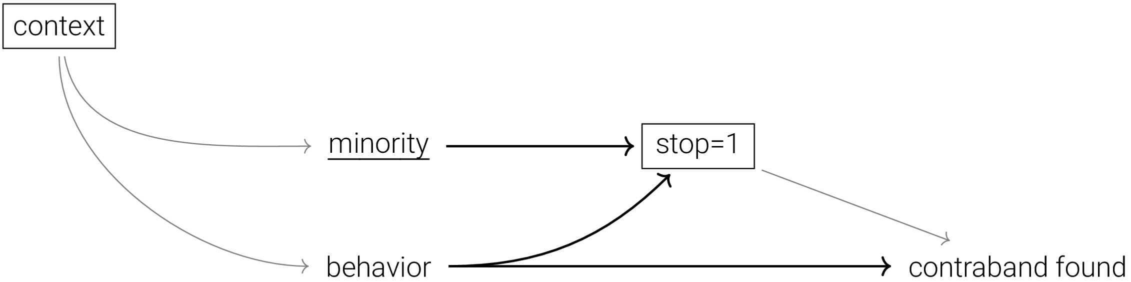
Sharp bounds can also build on existing methods

Outcome test exploits collider bias
to construct an indirect test of null discrimination.



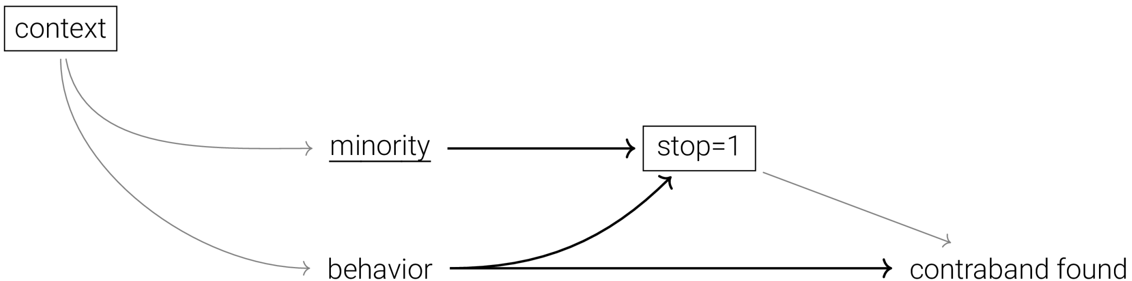
Sharp bounds can also build on existing methods

Outcome test exploits collider bias
to construct an indirect test of null discrimination.



Sharp bounds can also build on existing methods

Under reasonable assumptions, same data can be used to derive a nonparametric lower bound on discrimination.



Sharp bounds can also build on existing methods

Multiple datasets providing different, imperfect views can be optimally fused with single overarching approach.

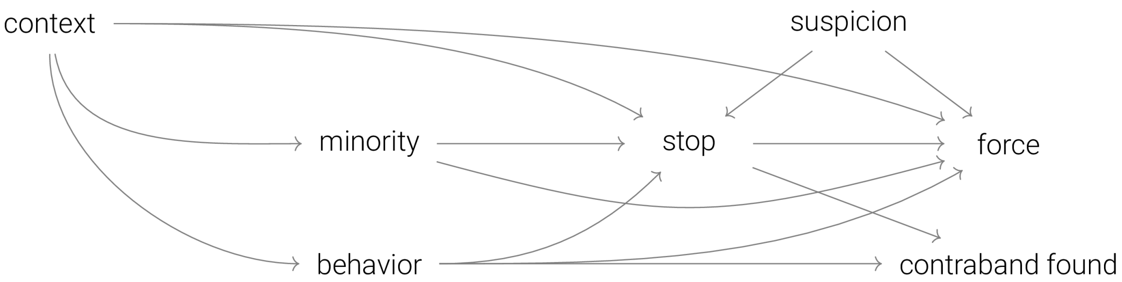
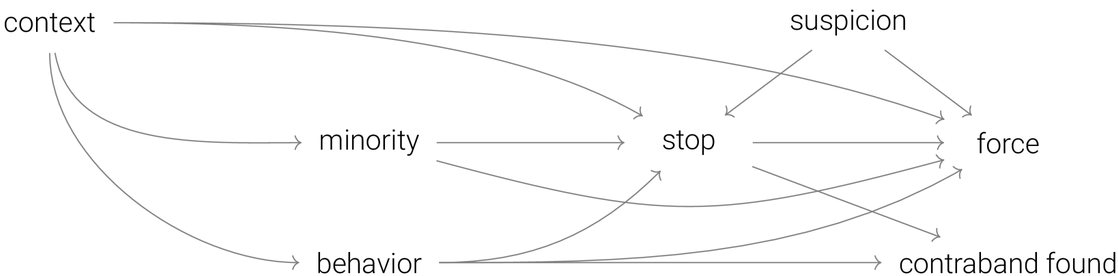


Illustration #1:
a simple bounding exercise


context
suspicion
minority
stop
force
behavior
contraband found

minority $\longrightarrow$ stop

minority $\longrightarrow$ stop

minority $\longrightarrow$ $\boxed{\text{stop=1}}$

minority $\longrightarrow$ $\boxed{\text{stop=0}}$

Observed data and all counterfactual manipulations are fully determined by joint distr. of response variables

Greenland & Robins ('86), Balke & Pearl ('94), Frangakis & Rubin ('02)

Observed data and all counterfactual manipulations are fully determined by joint distr. of response variables

Greenland & Robins ('86), Balke & Pearl ('94), Frangakis & Rubin ('02)

$$
\begin{tikzcd}
R_M \arrow[dr] & & & R_S \arrow[dl] \\
& \text{minority} \arrow[r] & \text{stop} &
\end{tikzcd}
$$

$$
\begin{array}{c}
\mathcal{R}_{\mathrm{M}} \\
\hline
r_{M,1} = 0 \\
r_{M,2} = 1
\end{array}
\qquad
\begin{array}{cl}
\mathcal{R}_{\mathrm{S}} & \\
\hline
r_{S,1}(m) = \begin{cases} 0 \text{ if } m=0 \\ 0 \text{ if } m=1 \end{cases} & \text{(never stop)} \\
r_{S,2}(m) = \begin{cases} 0 \text{ if } m=0 \\ 1 \text{ if } m=1 \end{cases} & \text{(anti-minority stop)} \\
r_{S,3}(m) = \begin{cases} 1 \text{ if } m=0 \\ 0 \text{ if } m=1 \end{cases} & \text{(anti-white stop)} \\
r_{S,4}(m) = \begin{cases} 1 \text{ if } m=0 \\ 1 \text{ if } m=1 \end{cases} & \text{(always stop)}
\end{array}
$$

Observed data and all counterfactual manipulations are fully determined by joint distr. of response variables

Greenland & Robins ('86), Balke & Pearl ('94), Frangakis & Rubin ('02)

$$\begin{tikzcd}
R_M \arrow[dr] & & & R_S \arrow[dl] \\
& \text{minority} \arrow[r] & \text{stop} &
\end{tikzcd}$$

$$\text{ATE} \quad = \quad E[\,\text{Stop}(\text{minority}) - \text{Stop}(\text{white})\,]$$

Observed data and all counterfactual manipulations are fully determined by joint distr. of response variables

Greenland & Robins ('86), Balke & Pearl ('94), Frangakis & Rubin ('02)

$$
\begin{tikzcd}
R_M \arrow[dr] & & & R_S \arrow[dl] \\
& \text{minority} \arrow[r] & \text{stop} &
\end{tikzcd}
$$

ATE = E[Stop(minority) - Stop(white)]

ATE = Pr(always stop) + Pr(anti-minority stop)

- Pr(always stop) - Pr(anti-white stop)

ATE = Pr(anti-minority stop) - Pr(anti-white stop)

Assumption:

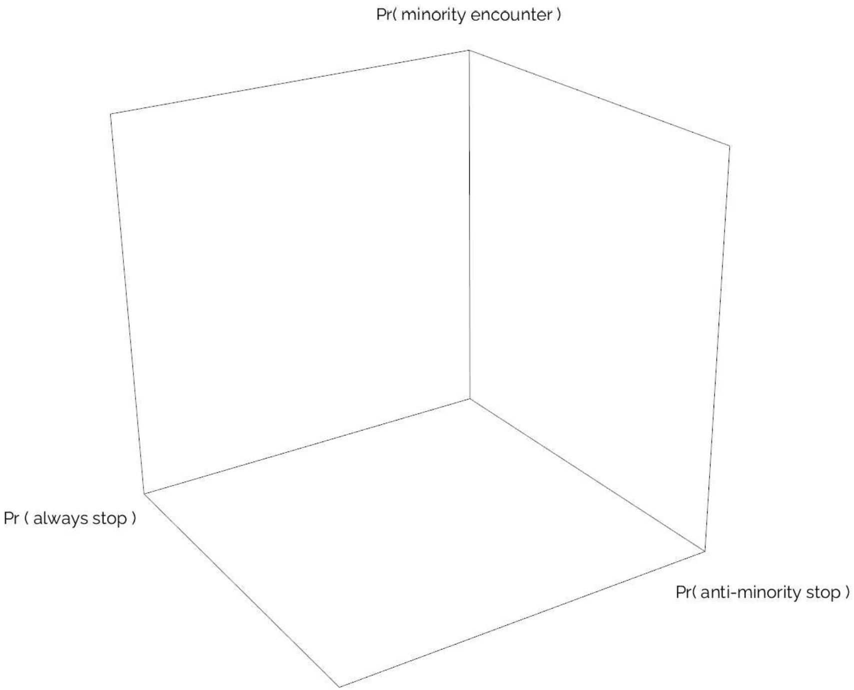
$p(\text{ anti-white stop }) = 0$

Assumption: p(anti-white stop) = 0

Laws of probability: $\sum_k p(r_{S,k}) = 1, \quad \sum_k p(r_{M,k}) = 1$

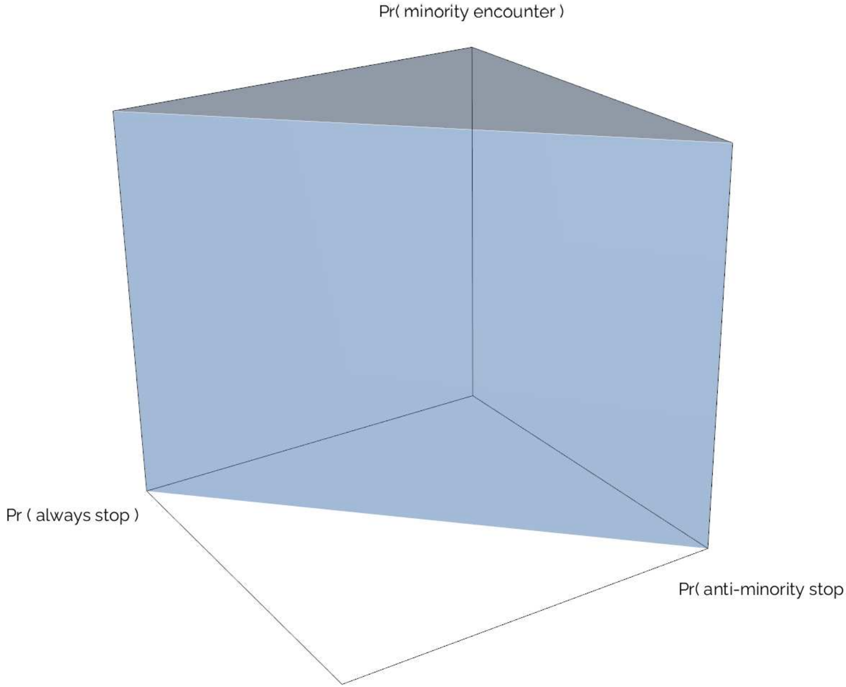
Assumption: p(anti-white stop) = 0

Laws of probability: $\sum_k p(r_{S,k}) = 1, \quad \sum_k p(r_{M,k}) = 1$



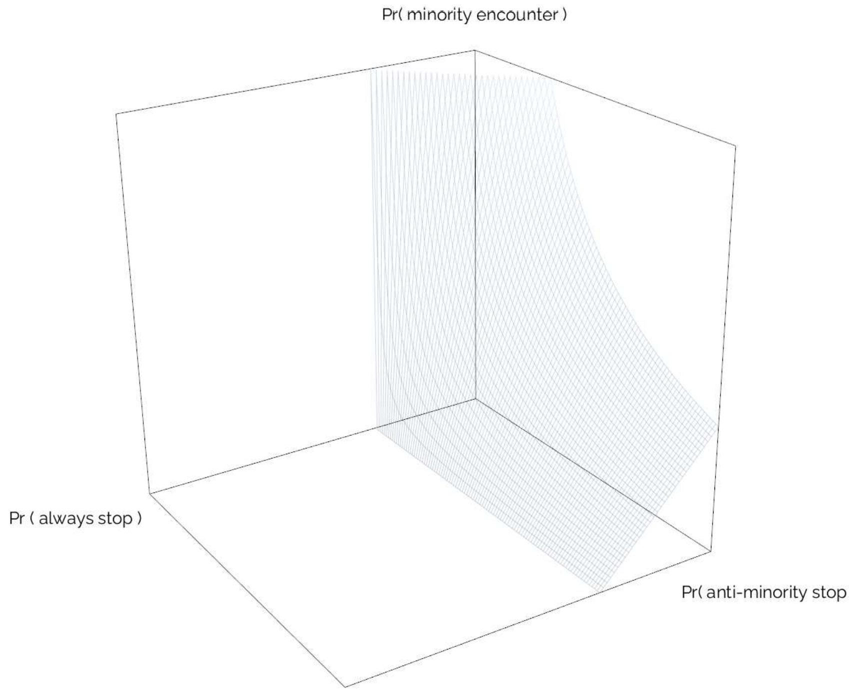
Assumption: $p(\text{ anti-white stop }) = 0$

Laws of probability: $\sum_k p(r_{S,k}) = 1, \quad \sum_k p(r_{M,k}) = 1$



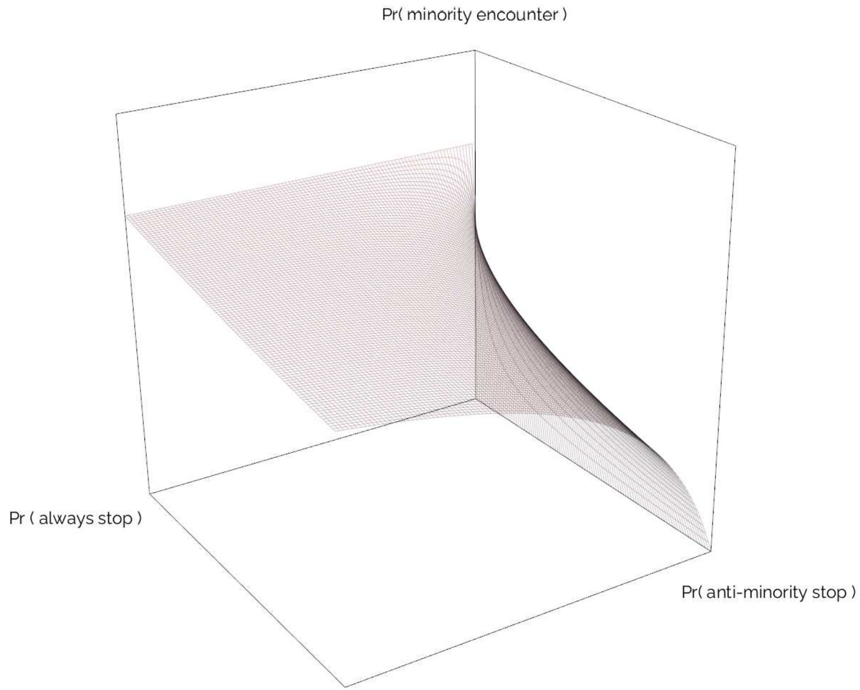
Observed stop margin:
p(stop)

p(min.) p(anti-min. stop)
+ p(min.) p(always stop)
+ [1 - p(min.)] p(always stop)

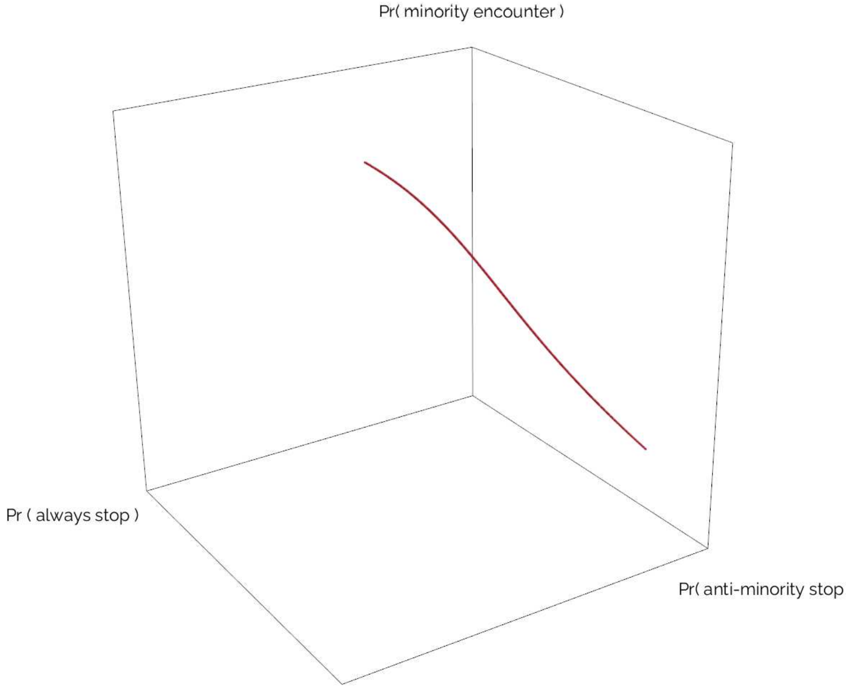


Observed stop records:
p(minority | Stop=1)

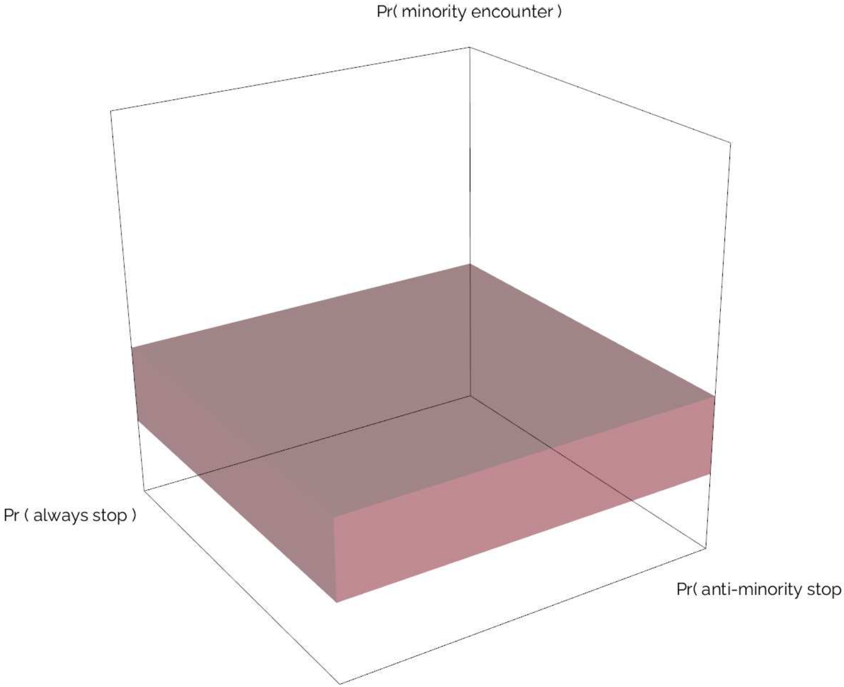
p(min.) p(anti-min. stop OR always stop)
/ [p(min.) p(anti-min. stop) + p(always stop)]



Remaining feasible region

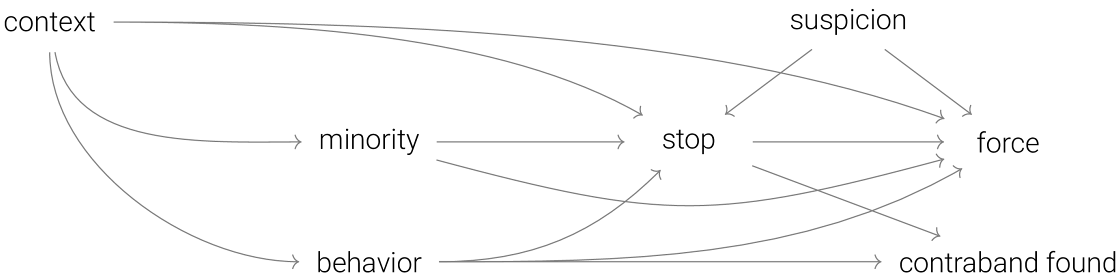


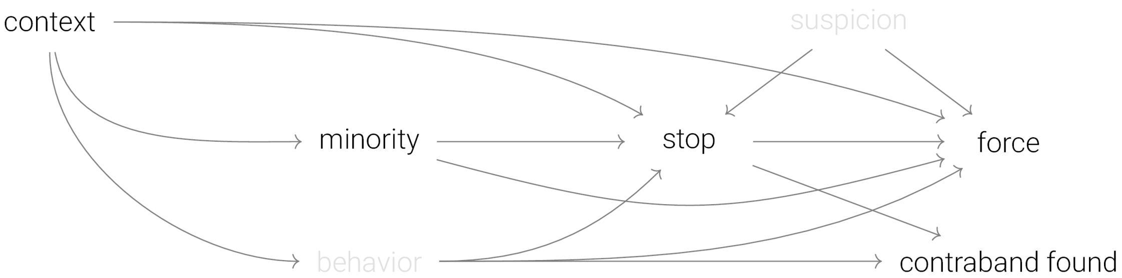
Population benchmarking: $\alpha \leq p(\text{min.}) \leq \beta$



Result:

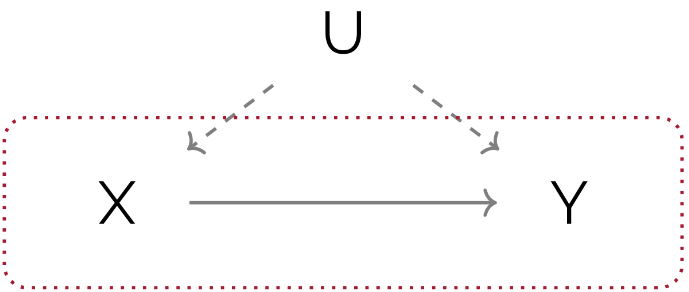
Discrete causal inference
is polynomial programming


context
suspicion
minority
stop
force
behavior
contraband found

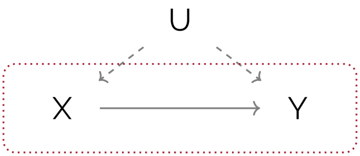

context
suspicion
minority
stop
force
behavior
contraband found

Implication:
Every question in imperfectly observed discrete causal systems can be solved automatically

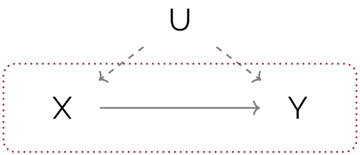
Applicable to essentially any research obstacle



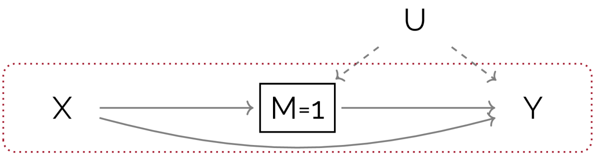
Confounding



Confounding



Mediator-based selection (Knox, Lowe, Mummolo, 2020)



A Detailed proofs

A.1 Bias for $\text{ATE}_{M=1}$

We first derive the bias of the local difference in means (that is, among encounters with $X_i = x$, $\hat{\Delta}_x = \overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}$), in estimating the local average treatment effect among stops, $\text{ATE}_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]$. The overall bias is then given by $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{ATE}_{M=1,x}\right)\Pr(X_i = x|M_i = 1)$.

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}_x] - \text{ATE}_{M=1,x}\\
&= (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}])\\
&\quad - (\mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x])\\
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x]\\
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x]\\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

under mediator monotonicity, $\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x) = 0$ and $\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x) = 1$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$
\begin{aligned}
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \cancel{\mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]}\\
&\qquad \cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)}\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \cancel{\mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]}\\
&\qquad \cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)}\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

adding and subtracting $\mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

substituting potential mediators based on principal strata,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

under mandatory reporting, $Y_i(d, 0) = 0$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

invoking assumption 4(b) (treatment ignorability),

$$
\begin{aligned}
&= \big(\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad - \mathbb{E}[Y_i(1, 1) - Y(0, 0)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \big)\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - (\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x])\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

which is Equation 6.

A.2 Bias for $\text{ATT}_{M=1}$

Next, we consider the bias that results when the local difference in means is used as an estimator for the local average racial effect among stopped minorities, $\text{ATT}_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x]$. Again, overall bias is found by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{ATT}_{M=1,x}\right)\Pr(X_i = x|D_i = 1, M_i = 1)$.

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}_x] - \text{ATT}_{M=1,x} &= (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}])\\
&\quad - (\mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x])\\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\\
&= -\mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\\
&= -\mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

by mediator monotonicity

$$
\begin{aligned}
&= -\mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

by mandatory reporting

$$
\begin{aligned}
&= -\mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

by treatment ignorability

$$
= -\mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|M_i(1) = 1, X_i = x)
$$

A.3 Bias for $\text{CDE}_{M=1}$

The $\text{CDE}_{M=1}$ is defined as

$$
\text{CDE}_{M=1} = \mathbb{E}[Y_i(1, 1)|M_i(D_i) = 1] - \mathbb{E}[Y_i(0, 1)|M_i(D_i) = 1] \tag{1}
$$

It is a somewhat contrived estimand because it necessarily involves counterfactuals that, for racial-stop encounters, could never realize even if researchers could somehow randomize civilian race in police-civilian encounters. For example, when a minority civilian is racially stopped for a "furtive movement" and reaches for their wallet, it makes little sense to consider an officer's potential use of force if the civilian suddenly became white at that moment: had police observed a white civilian from the onset, a stop would never have occurred. Moreover, the assumptions required for such counterfactuals are fundamentally unverifiable (Robins and Greenland, 1992), unless the experimentalist can somehow also manipulate officer stopping decisions without distorting outcomes. (We note that always-stop encounters are not subject to this issue, but analysts cannot hope to estimate the conditional CDE in this group because it is impossible to identify which minority encounters belong to this group.)

In the previous bias expression for the $\text{ATE}_{M=1}$, white individuals in the data—necessarily belonging to the always-stop group, $M_i(1) = M_i(0) = 1$—were used to estimate the $Y_i(0, M_i(0))$ potential outcomes of minority encounters in the data. Unavoidable bias arose as long as any minority individuals in the data belonged to the racial-stop group—had these individuals been white,

they would never have been stopped, and hence would not be subject to force. Changing the target estimand to the $\text{CDE}_{M=1}$ conceptually sidesteps this specific issue by considering a different counterfactual, $Y_i(0, 1)$ instead of $Y_i(0, M_i(0))$. But as we note above, for encounters in which only minority civilians would be stopped—that is, encounters with $M_i(1) = 1$ and $M_i(0) = 0$—this new counterfactual represents an impossible cross-world scenario. The $\text{CDE}_{M=1}$ asks whether force would have been used if officers were forced, against nature, to stop a white individual in this encounter as if they were a minority.

We now demonstrate that the local difference in means remains biased for the local controlled direct effect, $\text{CDE}_{M=1,x} = \mathbb{E}[Y_i(1, 1)|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, 1)|M_i = 1, X_i = x]$, unless officers are as violent toward minorities in always-stop encounters (where they are forced to intervene) as they are in racially discriminatory stops (where they are free to exercise discretion). In other words, for the naïve estimator to recover the $\text{CDE}_{M=1}$, Assumption 5 must hold. The derivation is almost identical to that of the $\text{ATE}_{M=1,x}$, differing only in that all individuals are held at $M_i = 1$ instead of allowed stops to vary with civilian race, $M_i(D_i)$. Bias for $\text{CDE}_{M=1}$ is then given by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{CDE}_{M=1,x}\right)\Pr(X_i = x|M_i = 1)$.

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}_x] - \text{CDE}_{M=1,x} &= (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}])\\
&\quad - (\mathbb{E}[Y_i(1, 1)|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, 1)|M_i = 1, X_i = x])\\
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, 1)|M_i(D_i) = 1, X_i = x]\\
&= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

under assumption 4(b),

$$
\begin{aligned}
&= \big(\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$
\begin{aligned}
&\qquad \big)\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - (\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x] - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x])\\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

which reproduces Equation 7.

Finally, we demonstrate that this bias is weakly negative. Rearranging terms yields

$$
\begin{aligned}
&\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&- \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\quad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&- \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&+ \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\quad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\\
&= -\mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x))\\
&\quad + \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x))
\end{aligned}
$$

The sum of the first and second terms is weakly negative by Assumption 3, because the magnitude of the first term is greater than that of the second, and similarly the sum of the third and fourth terms is also weakly negative. Therefore, bias is weakly negative when the naïve estimator is used to estimate the $\text{CDE}_{M=1}$.

A.4 Nonparametric sharp bounds for $\text{ATE}_{M=1}$

In this section, we derive nonparametric sharp bounds for the $\text{ATE}_{M=1,x}$. We begin with the case when the proportion of racially discriminatory stops among reported minority encounters, $\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$, is known or can be assumed. Rearrangement of Equa-

tion 6 (within levels of X) yields

$$
\begin{aligned}
\text{ATE}_{M=1,x} &= \mathbb{E}[\hat{\Delta}_x]\\
&\quad + \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x))\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x))\\
&\quad + \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&= \mathbb{E}[\hat{\Delta}_x]\\
&\quad + \frac{\mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \frac{\mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)^2 \qquad (2)\\
&\qquad \Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)\\
&= \mathbb{E}[\hat{\Delta}_x]\\
&\quad + \frac{\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i = 1, X_i = x)\\
&\quad - \frac{\mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)^2 \qquad (3)\\
&\qquad \Pr(D_i = 0|M_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i = 1, X_i = x)
\end{aligned} \tag{4}
$$

We then construct bounds on $\mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$ based on Fréchet inequalities for the joint distribution, $\Pr(Y_i(1, 1) = 1, M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)$, which incorporate marginal information about $Y_i(1, 1)$ and $M_i(0)$.

$$
\begin{aligned}
&\frac{\max\{0, \Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x) + \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - 1\}}{\Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)}\\
&\qquad \le \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \le \qquad (5)\\
&\frac{\min\{\Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x), \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]\}}{\Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)}
\end{aligned}
$$

These bounds are sharp given only marginal information, $\Pr(Y_i(1, 1) = 1|D_i = 1, M_i(1) = 1, X_i = x)$

and $\Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)$. However, the upper bound can be tightened further under Assumption 3, which implies $\mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \le \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]$; this is at least as small as the upper Fréchet bound.

Finally, note that the reported data contain no information that can be used to constrain the proportion of racially discriminatory minority stops, $\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$. If this proportion were zero, then the distribution of civilian race in police reports would reflect that of all police encounters (within levels of X). The reported data cannot distinguish between this possibility and an alternative population in which $\rho_x = \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$ is large, but white encounters are also larger by the proportion $1/(1 - \rho_x)$. Without side information about the total number of encounters, this proportion can take on any value in $[0, 1)$. Therefore, sharp bounds on $\text{ATE}_{M=1}$ alone are obtained by substituting Equation 5 into Equation 4 and setting the proportion of racial stops to unity. The bivariate bounds define the region in which $(\text{ATE}_{M=1}, \rho_x)$ pairs are consistent with the observed data. When ρ_x is set to zero or one, these respectively recover the difference in reported means and the marginal upper bounds on $\text{ATE}_{M=1}$. For $\rho_x \in (0, 1)$,

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}_X] + \rho_X\, \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x](1 - \Pr(D_i = 0|M_i = 1, X_i = x))\\
&\qquad \le \text{ATE}_{M=1,x} \le\\
&\qquad \mathbb{E}[\hat{\Delta}_x]\\
&+ \frac{\rho_X}{1 - \rho_X}\left(\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \max\left\{0, 1 + \frac{1}{\rho_X}\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \frac{1}{\rho_X}\right\}\right)\Pr(D_i = 0|M_i = 1, X_i = x)\\
&\qquad + \rho_X\, \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\,(1 - \Pr(D_i = 0|M_i = 1, X_i = x)),
\end{aligned}
$$

which reduces to Proposition 1 in the no-covariate case. Otherwise, bounds on $\text{ATE}_{M=1}$ are given by $\sum_x \underline{\text{ATE}}_{M=1,x}\Pr(X_i = x|M_i = 1) \le \text{ATE}_{M=1} \le \sum_x \overline{\text{ATE}}_{M=1,x}\Pr(X_i = x|M_i = 1)$, where $\underline{\text{ATE}}_{M=1,x}$ ($\overline{\text{ATE}}_{M=1,x}$) denote the lower (upper) bounds on the local average treatment effect.

Finally, we note that per Equation 3, the $\text{ATT}_{M=1,x}$ can be written

$$
\begin{aligned}
\text{ATT}_{M=1,x} &= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\,\Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\,\Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x)\\
&= \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\,\Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\,\Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x)
\end{aligned}
$$

under Assumption 1,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, 1) - |D_i = 1, M_i = 1, X_i = x]\\
&\quad - \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\,\Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x)
\end{aligned}
$$

and under Assumption 4,

$$
= \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\,(1 - \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x))
$$

which can be estimated from observed data if the proportion of racial stops is known. It then follows that

$$
\begin{aligned}
\text{ATT}_{M=1} &= \sum_X (\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x] + \rho_x\mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x])\\
&= \sum_X (\mathbb{E}[\Delta_X] + \rho_X\mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]).
\end{aligned}
$$

A.5 Uncertainty of bounds

Here, we describe our approach for constructing confidence intervals for the bounds on these causal quantities. We take X_i, D_i and M_i as fixed, so that uncertainty in the bounds arises strictly from the estimation of the conditional expectations, $\mathbb{E}[Y_i|D_i = d, M_i = 1, X_i = x]$. The asymptotic distribution of the estimated lower and upper bounds endpoints, $(\hat{\underline{\text{ATE}}}_{M=1}, \hat{\overline{\text{ATE}}}_{M=1})$, then follows directly from the asymptotic joint distribution of $\hat{\mathbb{E}}[Y_i|D_i = d, M_i = 1, X_i = x]$ for all d and x. We approximate this through a Monte Carlo simulation in which parameters of the logistic regression models described in Section 5 are sampled from a multivariate normal distribution centered on the parameter estimates and with the estimated covariance matrix. For each parameter sample θ^*, the corresponding bounds endpoint pair $(\underline{\text{ATE}}^*_{M=1}, \overline{\text{ATE}}^*_{M=1})$ is computed deterministically; after drawing a sufficient number of such samples, we numerically obtain the shortest range that fully contains 95% of all simulated bounds intervals. Closely related alternatives to this approach are the bootstrap-based method of Horowitz and Manski (2000) and the fully Bayesian approach taken in Knox et al. (2019). For the analysis in Section 5, we follow Fryer (2019) in using a cluster-robust covariance estimator, clustering on precinct, and 5,000 samples were drawn for each force threshold and model specification.

A.6 Point identification of ATE

First, we note that strata sizes are identified with information on the total count of encounters by race (both reported and unreported).

$$
\begin{aligned}
\Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) &= \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\\
&= \Pr(M_i = 1|D_i = 0, X_i = x)\\
\Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) &= \Pr(M_i(1) = 1, X_i = x)\\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x)\Pr(M_i = 1|D_i = 1, X_i = x)\\
&\quad - \Pr(M_i = 1|D_i = 0, X_i = x)\\
\Pr(M_i(1) = 0, M_i(0) = 1, X_i = x) &= 0\\
\Pr(M_i(1) = 0, M_i(0) = 0, X_i = x) &= 1 - \Pr(M_i(1) = 0, M_i(0) = 1, X_i = x)\\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 0, X_i = x)\\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x)\\
&= 1 - \Pr(M_i = 1|D_i = 1, X_i = x)
\end{aligned}
$$

We then reexpress the ATE in terms of strata-specific mean potential outcomes and simplify.

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$
\qquad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)
$$

under mandatory reporting

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)
\end{aligned}
$$

under mediator monotonicity

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\\
&\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x)\Pr(D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
\end{aligned}
$$

$$
\qquad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x)\Pr(D_i = 0, X_i = x)
$$

under treatment ignorability

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\,\Pr(M_i(1) = 1, M_i(0) = 1, X_i = x)\\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\,\Pr(M_i(1) = 1, M_i(0) = 0, X_i = x)\\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\,\Pr(M_i(1) = 1, M_i(0) = 1, X_i = x)
\end{aligned}
$$

which can be recovered from observed data

$$
\begin{aligned}
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x]\,\Pr(M_i = 1|D_i = 1, X_i = x)\\
&\quad - \mathbb{E}[Y_i|D_i = 0, M_i(D_i, X_i = x) = 1, X_i = x]\,\Pr(M_i = 1|D_i = 0, X_i = x)
\end{aligned}
$$

which reduces to Proposition 2 in the no-covariate case.

A.7 Derivation of outcome test bounds on ρ

Our paper focuses on the difficulty of estimating a race effect on post-stop police behavior such as the use of force. However, another popular approach, the outcome test, focuses on establishing whether there exists any bias in the decision to stop a civilian (Becker, 1971; Goel, Rao and Shroff, 2016; Engel, 2008; Knowles, Perisco and Todd, 2001; Ridgeway and MacDonald, 2010). Because the degree of the statistical bias we explore is a function of racial discrimination in stopping decisions, it is useful to clarify the assumptions undergirding outcome tests. In the process, we demonstrate that the principal stratification framework sheds light on the precise interpretation of outcome tests, and we prove that the outcome test can be used to establish a lower bound on the share of police stops of racial minorities that are racially discriminatory.

Outcome tests compare the rates of finding evidence of a crime—conditional on a suspect being stopped by police—across racial groups. The logic behind the test is that if the decision to stop a civilian is unbiased, the rate of discovering evidence of a crime ("hit rates") should be identical across groups. Proponents of outcome tests thus claim that differences in hit rates amount to evidence of racially biased policing. The empirical observation that hit rates are lower among minority stops can be written as $\mathbb{E}[Y_i|D_i = 0, M_i = 1] > \mathbb{E}[Y_i|D_i = 1, M_i = 1]$, where Y_i is an indicator, say, for finding contraband on a suspect. However, interpreting the above inequality as evidence of racial discrimination in fact requires assumptions that closely mirror those we describe above.

To see this, first observe that the overall hit rate among minority stops can be decomposed into

the weighted average of the hit rate among always-stop encounters and the hit rate among the (possibly nonexistent) set of racially discriminatory stops. In contrast, if we invoke Assumption 2 (which states that there are no white civilians stopped in circumstances where a minority civilian would be allowed to pass), then stops involving white civilians belong exclusively to the always-stop group.[1] In this case, the empirical difference in hit rates can be rewritten in the potential outcomes framework as

$$
\begin{aligned}
&\mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0]\\
&\quad -\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 1]\Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, M_i = 1)\\
&\quad - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, D_i = 1]\Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, M_i = 1) > 0
\end{aligned} \tag{6}
$$

A major critique of the outcome test is that observed racial disparities in hit rates alone do not constitute evidence of racially discriminatory stops because of the problem of "infra-marginality" (Ayres, 2002; Simoiu, Corbett-Davies and Goel, 2017). This critique suggests that the above inequality may hold simply because white civilians in always-stop encounters engage in more criminal conduct than minority suspects. In other words, the analyst might observe $\mathbb{E}[Y_i|D_i = 1, M_i = 1] < \mathbb{E}[Y_i|D_i = 0, M_i = 1]$ even if $\Pr(M_i(1) = 1, M_i(0) = 0) = 0$—that is, with no discrimination in stops—as long as $\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] < \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0]$. Some analysts employing the outcome test cast this scenario as unlikely, arguing that absent racial bias in stopping, "it would be difficult to explain why…whites for some reason had a systematically higher chance of possessing evidence of illegality" (Ayres, 2002) (137) and "there are not compelling reasons to suspect" this to be the case (138). Indeed, the validity of the outcome test hinges on the assumption that white and minority civilians in always-stop encounters commit crimes at the same rates, or that $\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] = \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0]$. This assumption closely parallels Assumption 4, which requires that treatment status is ignorable with respect to potential outcomes. (For simplicity, we suppose that this holds without conditioning on covariates, but the result also holds within levels of $X_i = x$.) In this case, the observed racial difference in hit rates can be rewritten as

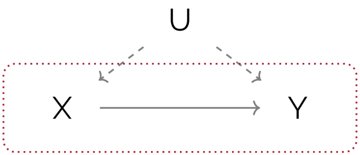
$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 0, M_i = 1] - \mathbb{E}[Y_i|D_i = 1, M_i = 1]\\
&\quad = \big(\mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0]\\
&\qquad -\mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, D_i = 1]\big)\Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, M_i = 1).
\end{aligned} \tag{7}
$$

This formulation makes clear that the observed evidence gap is due to the difference in hit

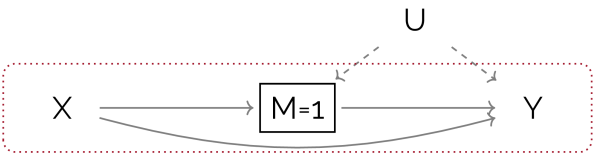
[1] If we do not assume mediator monotonicity, and allow for the presence of stops of white suspects that would not have occurred if the suspect was a racial minority, then the inequality used to estimate the outcome test becomes uninformative with respect to racial discrimination.

But analytic results are difficult to derive even in small graphs

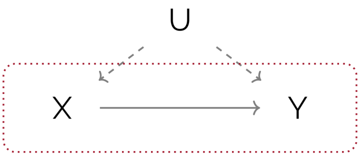
Confounding



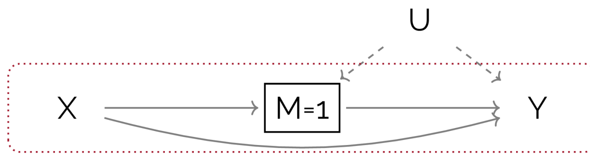
Mediator-based selection (Knox, Lowe, Mummolo, 2020)



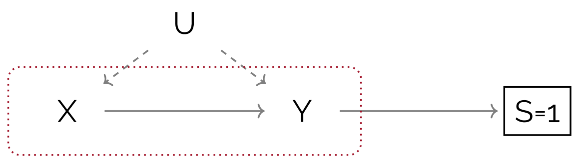
Confounding

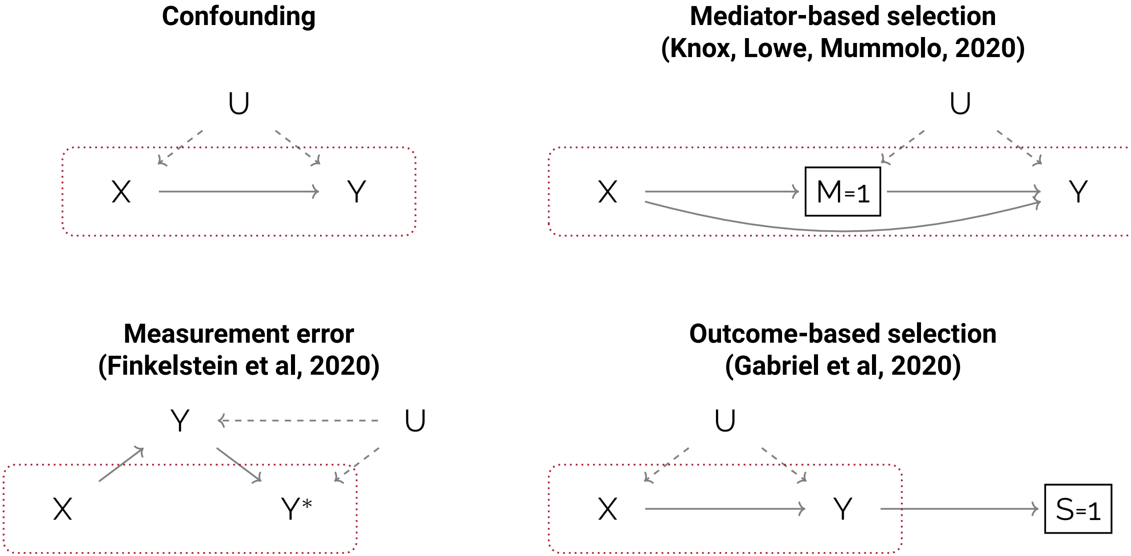


Mediator-based selection (Knox, Lowe, Mummolo, 2020)



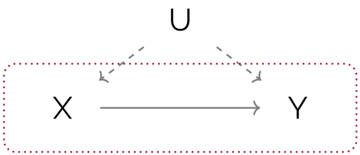
Outcome-based selection (Gabriel et al, 2020)



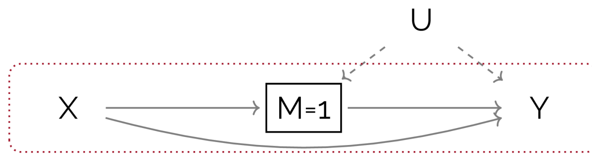


Confounding
U
X
Y
Mediator-based selection
(Knox, Lowe, Mummolo, 2020)
U
X
M=1
Y
Measurement error
(Finkelstein et al, 2020)
Y
U
X
Y*
Outcome-based selection
(Gabriel et al, 2020)
U
X
Y
S=1

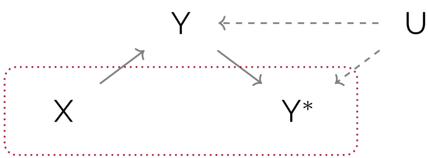
Confounding



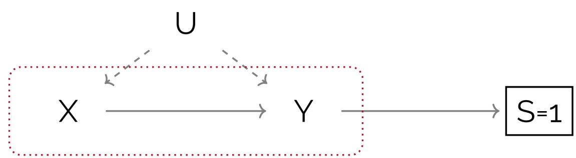
Mediator-based selection (Knox, Lowe, Mummolo, 2020)



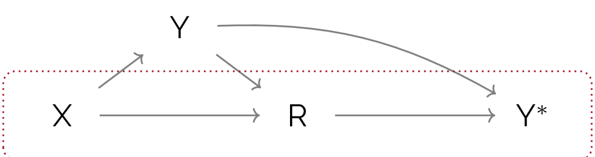
Measurement error (Finkelstein et al, 2020)

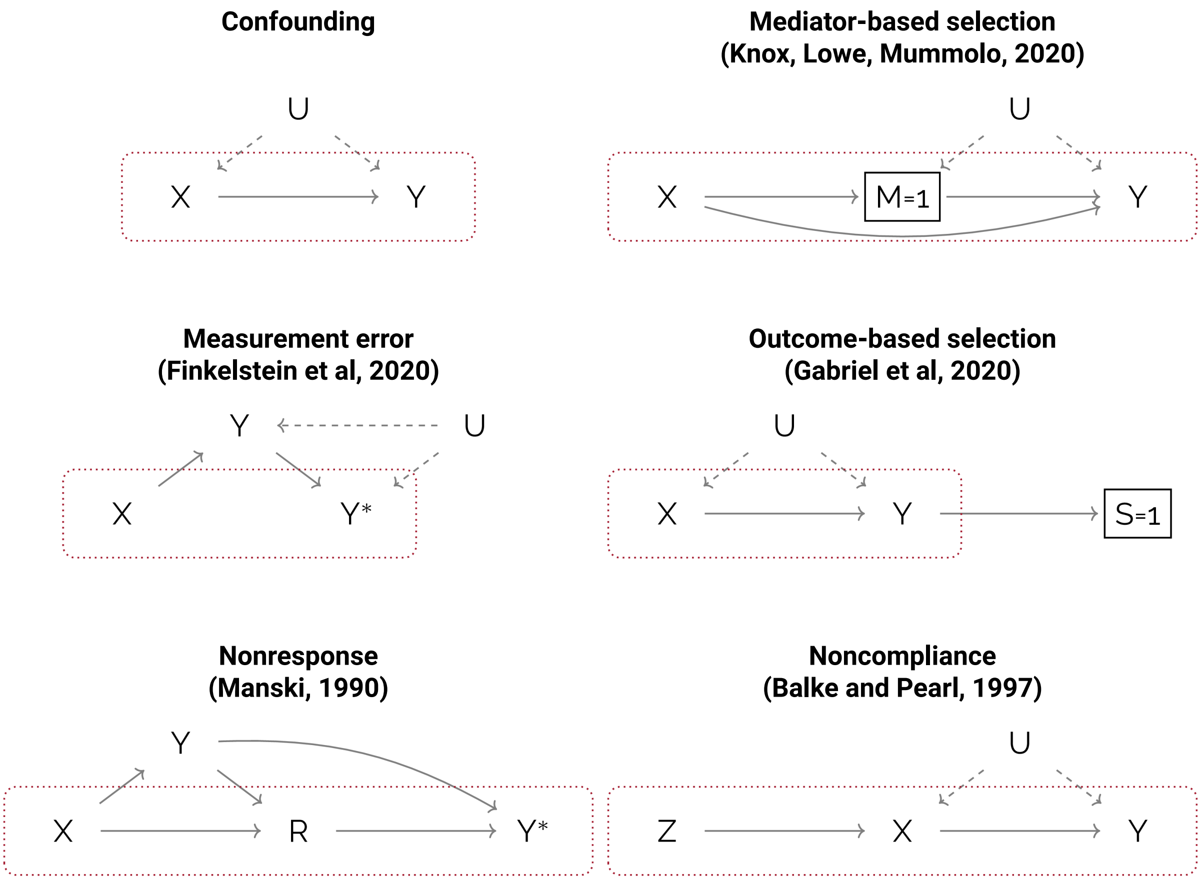


Outcome-based selection (Gabriel et al, 2020)



Nonresponse (Manski, 1990)




Confounding
U
X
Y
Mediator-based selection
(Knox, Lowe, Mummolo, 2020)
U
X
M=1
Y
Measurement error
(Finkelstein et al, 2020)
Y
U
X
Y*
Outcome-based selection
(Gabriel et al, 2020)
U
X
Y
S=1
Nonresponse
(Manski, 1990)
Y
X
R
Y*
Noncompliance
(Balke and Pearl, 1997)
U
Z
X
Y

A general solution: Automatic causal bounding

- Transform causal problems → optimization problems
  - Solves any partial identification problem
  - And also every point identification problem

[1] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

A general solution: Automatic causal bounding

- Transform causal problems → optimization problems
  - Solves any partial identification problem
  - And also every point identification problem

- Efficient spatial branch-and-bound procedure
  - Quickly produces valid non-sharp bounds
  - Iteratively refines bounds toward sharpness

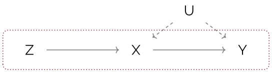
[1] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

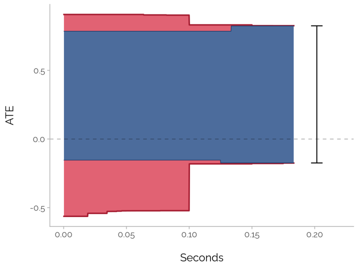
A general solution: Automatic causal bounding

- Transform causal problems → optimization problems
  - Solves any partial identification problem
  - And also every point identification problem
- Efficient spatial branch-and-bound procedure
  - Quickly produces valid non-sharp bounds
  - Iteratively refines bounds toward sharpness
- Strong theoretical guarantees
  - Guaranteed anytime validity of bounds
  - Guaranteed worst-case looseness of non-sharp bounds
  - Guaranteed eventual sharpness of bounds
  - Guaranteed conservative coverage of CI

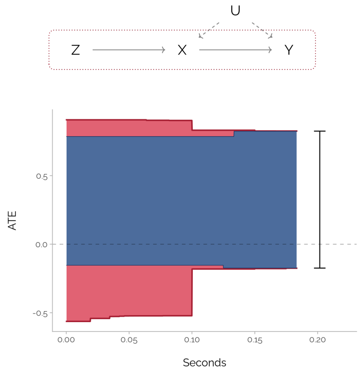
[1] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2021. "An Automated Approach to Causal Inference in Discrete Settings." Working paper.

Noncompliance (Balke & Pearl '94)

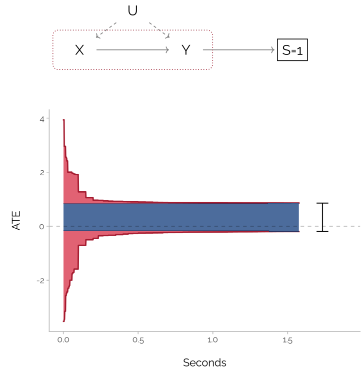




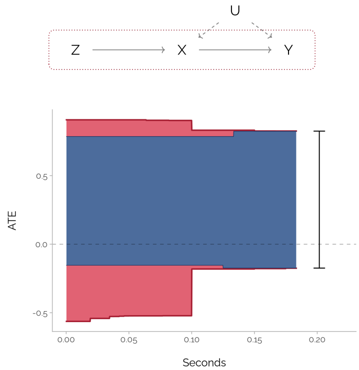
Noncompliance (Balke & Pearl '94)



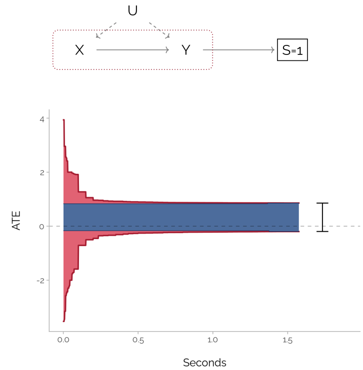
Outcome-based selection (Gabriel et al. '21)



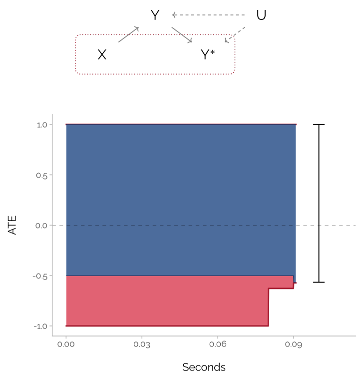
Noncompliance (Balke & Pearl '94)

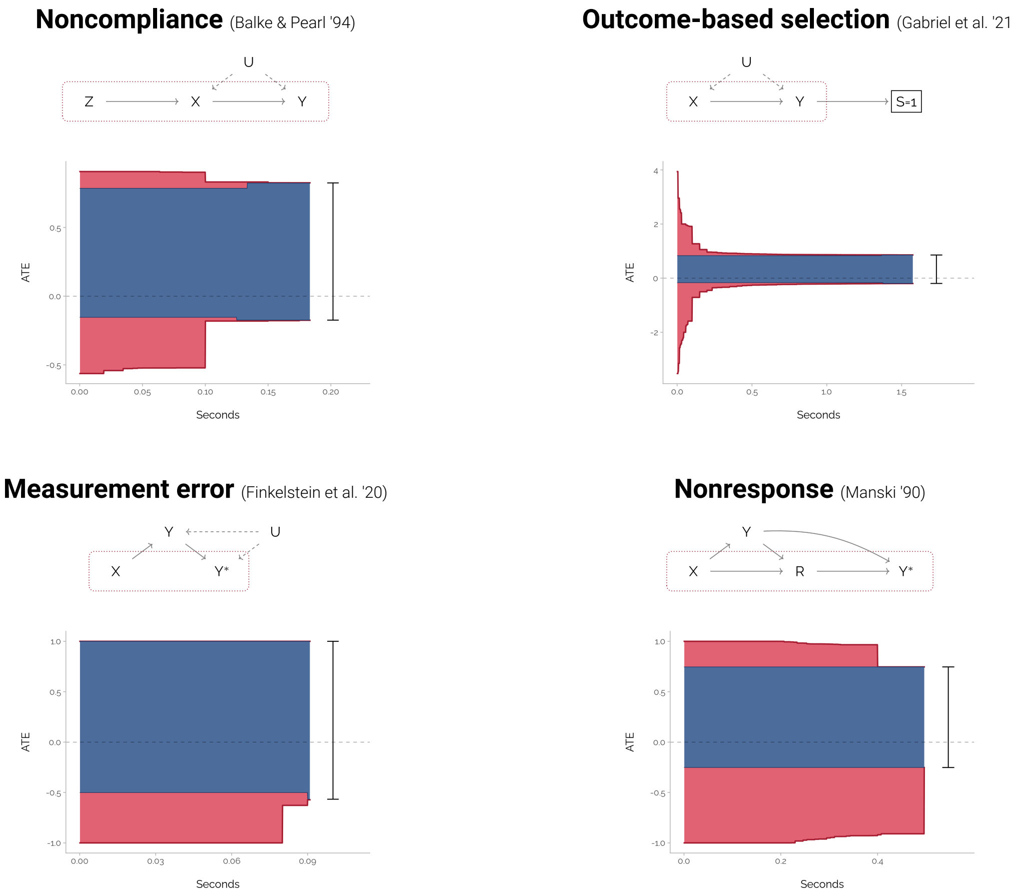


Outcome-based selection (Gabriel et al. '21)



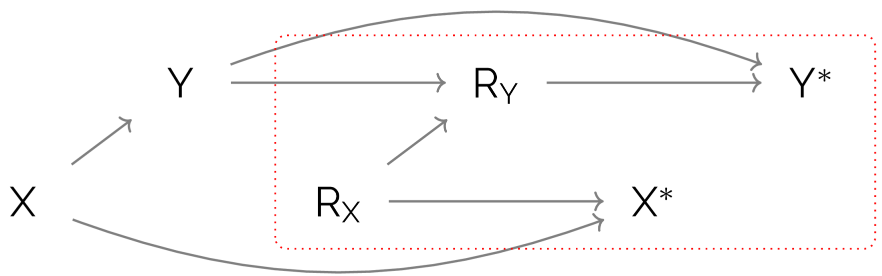
Measurement error (Finkelstein et al. '20)

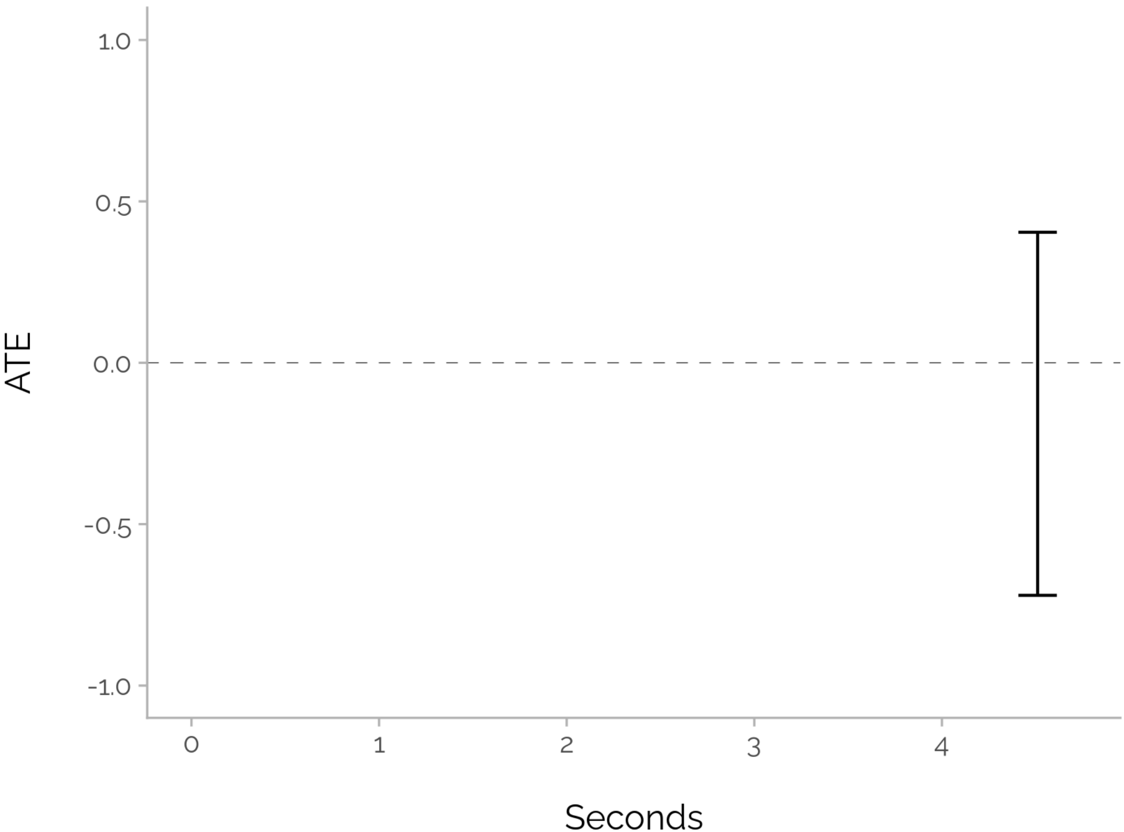




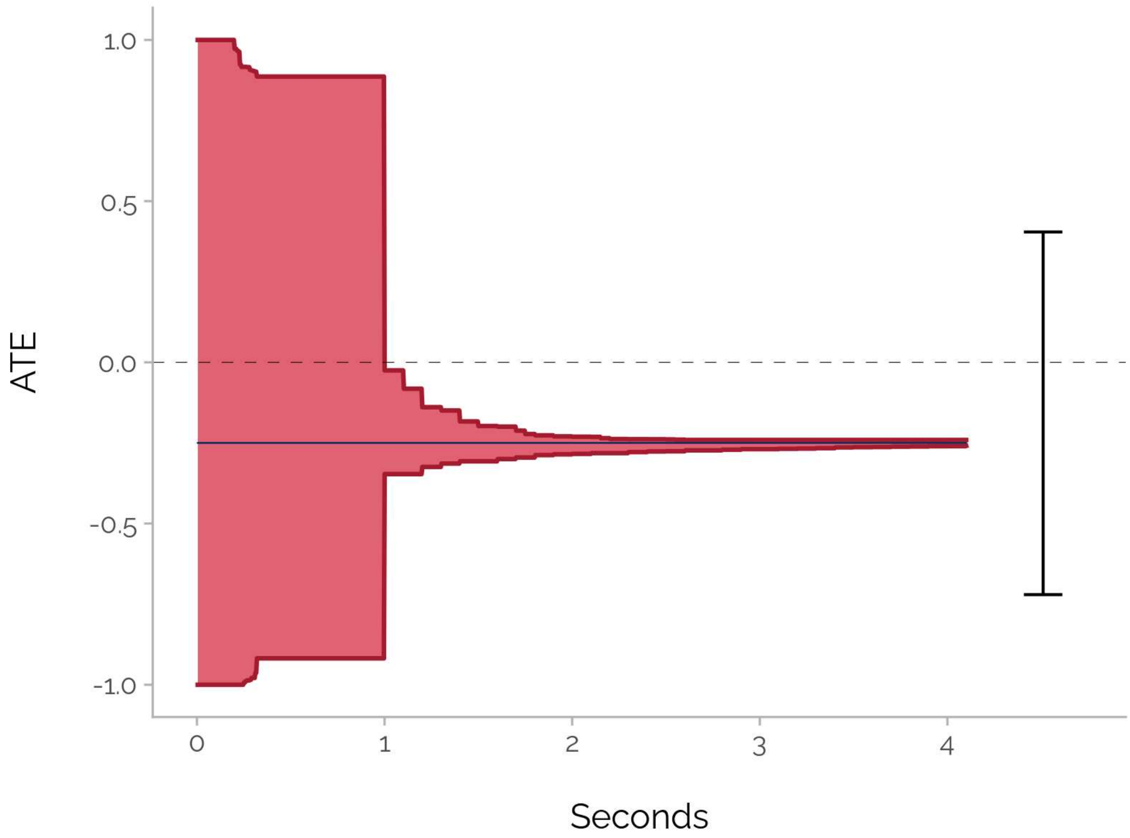
Noncompliance (Balke & Pearl '94)
U
Z
X
Y
ATE
0.5
0.0
-0.5
0.00
0.05
0.10
0.15
0.20
Seconds
Outcome-based selection (Gabriel et al. '21)
U
X
Y
S=1
ATE
4
2
0
-2
0.0
0.5
1.0
1.5
Seconds
Measurement error (Finkelstein et al. '20)
Y
U
X
Y*
ATE
1.0
0.5
0.0
-0.5
-1.0
0.00
0.03
0.06
0.09
Seconds
Nonresponse (Manski '90)
Y
X
R
Y*
ATE
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
Seconds

Dependent nonrandom missingness



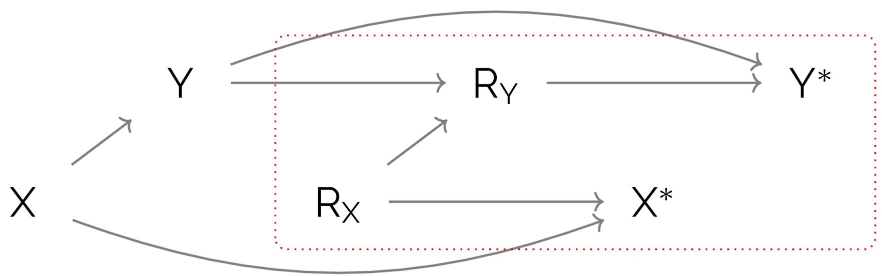


ATE
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
Seconds



ATE
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
Seconds

Dependent nonrandom missingness



Surprisingly, point identification with "shadow variables" demonstrated by Miao & Tchetgen Tchetgen ('16)

Takeaways

Takeaways

- Data imperfections are inevitable, often consequential

Takeaways

- Data imperfections are inevitable, often consequential
- Makes little sense to assume away discrimination in stage #1 of a process in order to estimate stage #2

Takeaways

- Data imperfections are inevitable, often consequential
- Makes little sense to assume away discrimination in stage #1 of a process in order to estimate stage #2
- Reporting best- and worst-case conclusions over all admissible worlds → maximally robust inferences

Takeaways

- Data imperfections are inevitable, often consequential
- Makes little sense to assume away discrimination in stage #1 of a process in order to estimate stage #2
- Reporting best- and worst-case conclusions over all admissible worlds → maximally robust inferences
- General-purpose algorithms for obtaining these bounds computationally are now available

Extra slides

State of the art

- **Point** identification has already been automated
  - *do*-calculus (Pearl '95; Huang & Valtorta '06; Shpitser & Pearl '06)
  - Fusing observational & experimental data (Lee, Correa, and Bareinboim, 2020; Lee and Shpitser, 2020)

State of the art

- **Point** identification has already been automated
  - *do*-calculus (Pearl '95; Huang & Valtorta '06; Shpitser & Pearl '06)
  - Fusing observational & experimental data (Lee, Correa, and Bareinboim, 2020; Lee and Shpitser, 2020)
- But **partial** identification (bounding) is much harder
  - Linear programming works for some cases (Balke & Pearl '97)
  - But many problems can't be represented as linear programs

State of the art

- **Point** identification has already been automated
  - *do*-calculus (Pearl '95; Huang & Valtorta '06; Shpitser & Pearl '06)
  - Fusing observational & experimental data (Lee, Correa, and Bareinboim, 2020; Lee and Shpitser, 2020)

- But **partial** identification (bounding) is much harder
  - Linear programming works for some cases (Balke & Pearl '97)
  - But many problems can't be represented as linear programs

- General solution to partial ID has been elusive
  - Symbolic solution is theoretically possible (Geiger & Meek '99) but in practice is wildly computationally infeasible
  - Problem instances are often nonconvex,

State of the art

- **Point** identification has already been automated
  - *do*-calculus (Pearl '95; Huang & Valtorta '06; Shpitser & Pearl '06)
  - Fusing observational & experimental data (Lee, Correa, and Bareinboim, 2020; Lee and Shpitser, 2020)
- But **partial** identification (bounding) is much harder
  - Linear programming works for some cases (Balke & Pearl '97)
  - But many problems can't be represented as linear programs
- General solution to partial ID has been elusive
  - Symbolic solution is theoretically possible (Geiger & Meek '99) but in practice is wildly computationally infeasible
  - Problem instances are often nonconvex, numeric approaches that fail to discover global extrema produce *invalid bounds*

Discrete causal inference is polynomial programming

Constraints are polynomials

$$p(x_1, \ldots, x_N) \quad = \prod_j p(x_j | \mathrm{pa}_{X_j})$$

Constraints are polynomials

$$
\begin{aligned}
p(x_1, \ldots, x_N) &= \prod_j p(x_j | \mathrm{pa}_{X_j}) \\
&= \prod_j \Big[\sum_k q(u_{X_j,k}) \, \mathbf{1}\{ X_j^{(u_{X_j,k})}(\mathrm{pa}_{X_j}) = x_j \} \Big]
\end{aligned}
$$

Constraints are polynomials

$$\begin{aligned} p(x_1, \ldots, x_N) &= \prod_j p(x_j | \mathrm{pa}_{X_j}) \\ &= \prod_j \Big[\sum_k q(u_{X_j,k}) \, \mathbf{1}\{X_j^{(u_{X_j,k})}(\mathrm{pa}_{X_j}) = x_j\} \Big] \\ p(x_1) &= \sum_{x_2,\ldots,x_N} \prod_j \Big[\sum_k q(u_{X_j,k}) \, \mathbf{1}\{X_j^{(u_{X_j,k})}(\mathrm{pa}_{X_j}) = x_j\} \Big] \end{aligned}$$

Constraints are polynomials

$$p(y_1, \ldots, y_N | x_1, \ldots, x_M)$$

$$= \frac{p(y_1, \ldots, y_N, x_1, \ldots, x_M)}{p(x_1, \ldots, x_M)}$$

Constraints are polynomials

$p(y_1, \ldots, y_N | x_1, \ldots, x_M)$

$$= \frac{p(y_1, \ldots, y_N, x_1, \ldots, x_M)}{p(x_1, \ldots, x_M)}$$

$$= \frac{\prod_i p(y_i | \mathrm{pa}_{Y_i}) \prod_j p(x_j | \mathrm{pa}_{X_j})}{\prod_j p(x_j | \mathrm{pa}_{X_j})}$$

Constraints are polynomials

$p(y_1, \ldots, y_N | x_1, \ldots, x_M)$

$$= \frac{p(y_1, \ldots, y_N, x_1, \ldots, x_M)}{p(x_1, \ldots, x_M)}$$

$$= \frac{\prod_i p(y_i | \mathrm{pa}_{Y_i}) \prod_j p(x_j | \mathrm{pa}_{X_j})}{\prod_j p(x_j | \mathrm{pa}_{X_j})}$$

$$= \frac{\prod_i \left[\sum_k q(u_{Y_i,k}) \, \mathbf{1}\{ Y_i^{(u_{Y_i,k})}(\mathrm{pa}_{Y_i}) = y_i \} \right] \prod_j \left[\sum_k q(u_{X_j,k}) \, \mathbf{1}\{ X_j^{(u_{X_j,k})}(\mathrm{pa}_{X_j}) = x_j \} \right]}{\prod_j \left[\sum_k q(u_{X_j,k}) \, \mathbf{1}\{ X_j^{(u_{X_j,k})}(\mathrm{pa}_{X_j}) = x_j \} \right]}$$

Discrete causal inference is polynomial programming

- Polynomial constraints:
  - Observed joint, marginal, conditional probabilities

Discrete causal inference is polynomial programming

- Polynomial constraints:
  - Observed joint, marginal, conditional probabilities
  - Monotonicity, disabling, elimination assumptions

Discrete causal inference is polynomial programming

- Polynomial constraints:
    - Observed joint, marginal, conditional probabilities
    - Monotonicity, disabling, elimination assumptions
- Polynomial objective functions:
    - Additive effects (joint, total, mediated)

Discrete causal inference is polynomial programming

- Polynomial constraints:
  - Observed joint, marginal, conditional probabilities
  - Monotonicity, disabling, elimination assumptions
- Polynomial objective functions:
  - Additive effects (joint, total, mediated)
  - Multiplicative effects

Discrete causal inference is polynomial programming

- Polynomial constraints:
  - Observed joint, marginal, conditional probabilities
  - Monotonicity, disabling, elimination assumptions
- Polynomial objective functions:
  - Additive effects (joint, total, mediated)
  - Multiplicative effects
  - Strictly monotonic transformations

$$
\begin{bmatrix}
q(r_{m,0}, r_{b,0}) \\
q(r_{m,0}, r_{b,1}) \\
q(r_{m,1}, r_{b,0}) \\
q(r_{m,1}, r_{b,1}) \\
q(r_{S,1}) \\
q(r_{S,2}) \\
q(r_{S,3}) \\
q(r_{S,4}) \\
q(r_{S,5}) \\
q(r_{S,6}) \\
q(r_{S,7}) \\
q(r_{S,8}) \\
q(r_{S,9}) \\
q(r_{S,10}) \\
q(r_{S,11}) \\
q(r_{S,12}) \\
q(r_{S,13}) \\
q(r_{S,14}) \\
q(r_{S,15}) \\
q(r_{S,16})
\end{bmatrix}
\geq \mathbf{0}
$$

$$\begin{bmatrix} q(r_{m,0}, r_{b,0}) \\ q(r_{m,0}, r_{b,1}) \\ q(r_{m,1}, r_{b,0}) \\ q(r_{m,1}, r_{b,1}) \\ q(r_{S,1}) \\ q(r_{S,2}) \\ q(r_{S,3}) \\ q(r_{S,4}) \\ q(r_{S,5}) \\ q(r_{S,6}) \\ q(r_{S,7}) \\ q(r_{S,8}) \\ q(r_{S,9}) \\ q(r_{S,10}) \\ q(r_{S,11}) \\ q(r_{S,12}) \\ q(r_{S,13}) \\ q(r_{S,14}) \\ q(r_{S,15}) \\ q(r_{S,16}) \end{bmatrix}^{\top} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = 1$$

$$\begin{bmatrix} q(r_{m,0}, r_{b,0}) \\ q(r_{m,0}, r_{b,1}) \\ q(r_{m,1}, r_{b,0}) \\ q(r_{m,1}, r_{b,1}) \\ q(r_{S,1}) \\ q(r_{S,2}) \\ q(r_{S,3}) \\ q(r_{S,4}) \\ q(r_{S,5}) \\ q(r_{S,6}) \\ q(r_{S,7}) \\ q(r_{S,8}) \\ q(r_{S,9}) \\ q(r_{S,10}) \\ q(r_{S,11}) \\ q(r_{S,12}) \\ q(r_{S,13}) \\ q(r_{S,14}) \\ q(r_{S,15}) \\ q(r_{S,16}) \end{bmatrix}^\top \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 1$$

$$\alpha \leq \begin{bmatrix} q(r_{m,0}, r_{b,0}) \\ q(r_{m,0}, r_{b,1}) \\ q(r_{m,1}, r_{b,0}) \\ q(r_{m,1}, r_{b,1}) \\ q(r_{S,1}) \\ q(r_{S,2}) \\ q(r_{S,3}) \\ q(r_{S,4}) \\ q(r_{S,5}) \\ q(r_{S,6}) \\ q(r_{S,7}) \\ q(r_{S,8}) \\ q(r_{S,9}) \\ q(r_{S,10}) \\ q(r_{S,11}) \\ q(r_{S,12}) \\ q(r_{S,13}) \\ q(r_{S,14}) \\ q(r_{S,15}) \\ q(r_{S,16}) \end{bmatrix}^{\top} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \leq \beta$$

$$
\begin{bmatrix} q(r_{m,0}, r_{b,0}) \\ q(r_{m,0}, r_{b,1}) \\ q(r_{m,1}, r_{b,0}) \\ q(r_{m,1}, r_{b,1}) \\ q(r_{S,1}) \\ q(r_{S,2}) \\ q(r_{S,3}) \\ q(r_{S,4}) \\ q(r_{S,5}) \\ q(r_{S,6}) \\ q(r_{S,7}) \\ q(r_{S,8}) \\ q(r_{S,9}) \\ q(r_{S,10}) \\ q(r_{S,11}) \\ q(r_{S,12}) \\ q(r_{S,13}) \\ q(r_{S,14}) \\ q(r_{S,15}) \\ q(r_{S,16}) \end{bmatrix}^{\top} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = p(b_1)
$$

$$
\begin{bmatrix}
q(r_{m,0}, r_{b,0}) \\
q(r_{m,0}, r_{b,1}) \\
q(r_{m,1}, r_{b,0}) \\
q(r_{m,1}, r_{b,1}) \\
q(r_{S,1}) \\
q(r_{S,2}) \\
q(r_{S,3}) \\
q(r_{S,4}) \\
q(r_{S,5}) \\
q(r_{S,6}) \\
q(r_{S,7}) \\
q(r_{S,8}) \\
q(r_{S,9}) \\
q(r_{S,10}) \\
q(r_{S,11}) \\
q(r_{S,12}) \\
q(r_{S,13}) \\
q(r_{S,14}) \\
q(r_{S,15}) \\
q(r_{S,16})
\end{bmatrix}^{\top}
\left[\begin{array}{cccccccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 1/2 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]
\begin{bmatrix}
q(r_{m,0}, r_{b,0}) \\
q(r_{m,0}, r_{b,1}) \\
q(r_{m,1}, r_{b,0}) \\
q(r_{m,1}, r_{b,1}) \\
q(r_{S,1}) \\
q(r_{S,2}) \\
q(r_{S,3}) \\
q(r_{S,4}) \\
q(r_{S,5}) \\
q(r_{S,6}) \\
q(r_{S,7}) \\
q(r_{S,8}) \\
q(r_{S,9}) \\
q(r_{S,10}) \\
q(r_{S,11}) \\
q(r_{S,12}) \\
q(r_{S,13}) \\
q(r_{S,14}) \\
q(r_{S,15}) \\
q(r_{S,16})
\end{bmatrix}
= p(s_1)
$$

We have a problem

We have a problem

$$
\begin{array}{lll}
\text{maximize/minimize} & \mathbf{q}^\top \mathbf{U}_{\text{obj}} \mathbf{q} & \\
\text{subject to} & \mathbf{q} & \geq \mathbf{0} \\
& \mathbf{q}^\top [\mathbf{1}_J^\top, \mathbf{0}_K^\top]^\top & = 1 \\
& \mathbf{q}^\top [\mathbf{0}_J^\top, \mathbf{1}_K^\top]^\top & = 1 \\
& \mathbf{q}^\top \mathbf{v}_M & \geq \alpha \\
& \mathbf{q}^\top \mathbf{v}_M & \leq \beta \\
& \mathbf{q}^\top \mathbf{v}_{\mathrm{B}} & = p(b_1) \\
& \mathbf{q}^\top \mathbf{U}_{\mathrm{S}} \mathbf{q} & = p(s_1)
\end{array}
$$

We have a problem

$$\begin{array}{lrcl}
\text{maximize/minimize} & \mathbf{q}^\top \mathbf{U}_{\text{obj}} \mathbf{q} & & \\
\text{subject to} & \mathbf{q} & \geq & \mathbf{0} \\
& \mathbf{q}^\top [\mathbf{1}_J^\top, \mathbf{0}_K^\top]^\top & = & 1 \\
& \mathbf{q}^\top [\mathbf{0}_J^\top, \mathbf{1}_K^\top]^\top & = & 1 \\
& \mathbf{q}^\top \mathbf{v}_M & \geq & \alpha \\
& \mathbf{q}^\top \mathbf{v}_M & \leq & \beta \\
& \mathbf{q}^\top \mathbf{v}_{\text{B}} & = & p(b_1) \\
& \mathbf{q}^\top \mathbf{U}_{\text{S}} \mathbf{q} & = & p(s_1)
\end{array}$$

- Generalization of LP bounds (run very fast) (Balke & Pearl '94)

We have a problem

$$
\begin{array}{lll}
\text{maximize/minimize} & \mathbf{q}^\top \mathbf{U}_{\text{obj}} \mathbf{q} & \\
\text{subject to} & \mathbf{q} & \geq \mathbf{0} \\
& \mathbf{q}^\top [\mathbf{1}_J^\top, \mathbf{0}_K^\top]^\top & = 1 \\
& \mathbf{q}^\top [\mathbf{0}_J^\top, \mathbf{1}_K^\top]^\top & = 1 \\
& \mathbf{q}^\top \mathbf{v}_M & \geq \alpha \\
& \mathbf{q}^\top \mathbf{v}_M & \leq \beta \\
& \mathbf{q}^\top \mathbf{v}_{\text{B}} & = p(b_1) \\
& \mathbf{q}^\top \mathbf{U}_{\text{S}} \mathbf{q} & = p(s_1)
\end{array}
$$

- Generalization of LP bounds (run very fast) (Balke & Pearl '94)
- Nonconvex QCQPs are known to be NP-hard

We have a problem

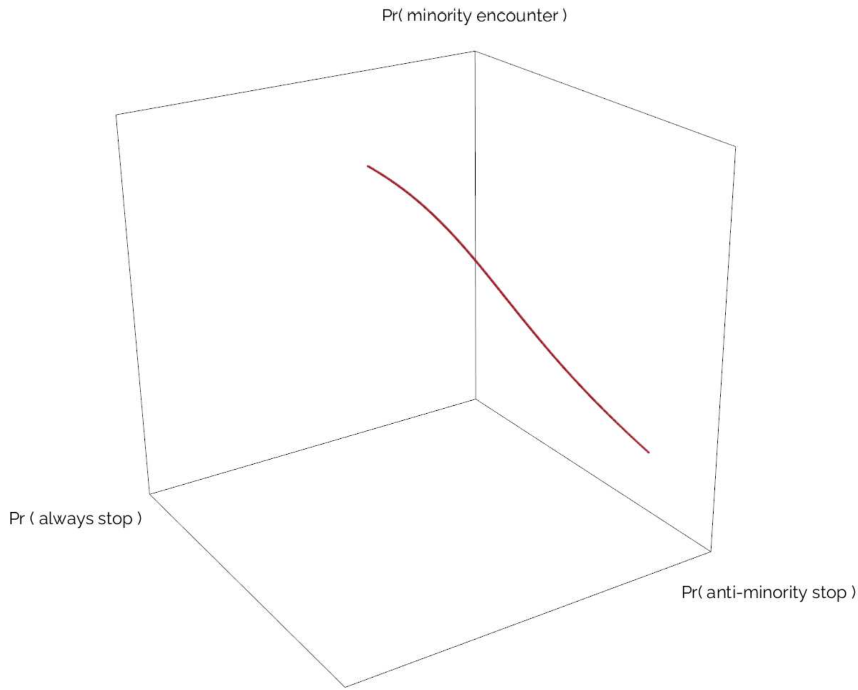
$$
\begin{array}{lllll}
\text{maximize/minimize} & \mathbf{q}^\top \mathbf{U}_{\text{obj}} \mathbf{q} & & \\
\text{subject to} & \mathbf{q} & \geq & \mathbf{0} \\
& \mathbf{q}^\top [\mathbf{1}_J^\top, \mathbf{0}_K^\top]^\top & = & 1 \\
& \mathbf{q}^\top [\mathbf{0}_J^\top, \mathbf{1}_K^\top]^\top & = & 1 \\
& \mathbf{q}^\top \mathbf{v}_M & \geq & \alpha \\
& \mathbf{q}^\top \mathbf{v}_M & \leq & \beta \\
& \mathbf{q}^\top \mathbf{v}_B & = & p(b_1) \\
& \mathbf{q}^\top \mathbf{U}_S \mathbf{q} & = & p(s_1)
\end{array}
$$

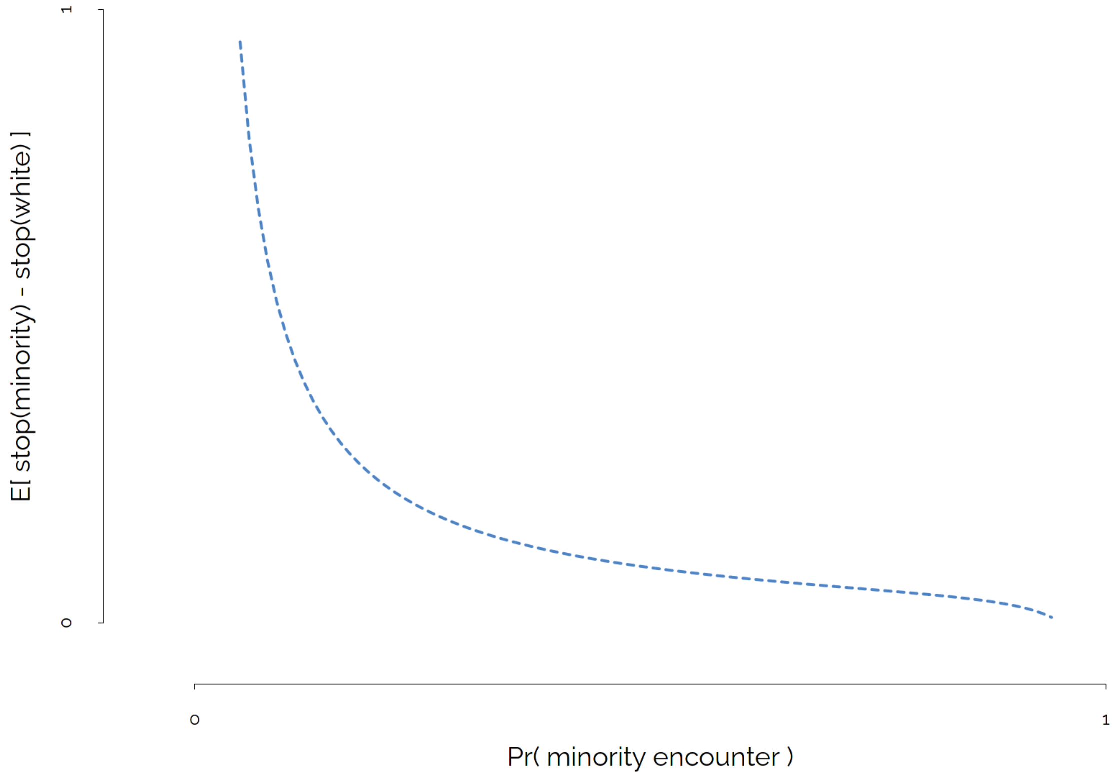
- Generalization of LP bounds (run very fast) (Balke & Pearl '94)
- Nonconvex QCQPs are known to be NP-hard
- Larger DAGs $\longrightarrow$ higher degree, greater difficulty

An efficient procedure for computing ε-sharp bounds

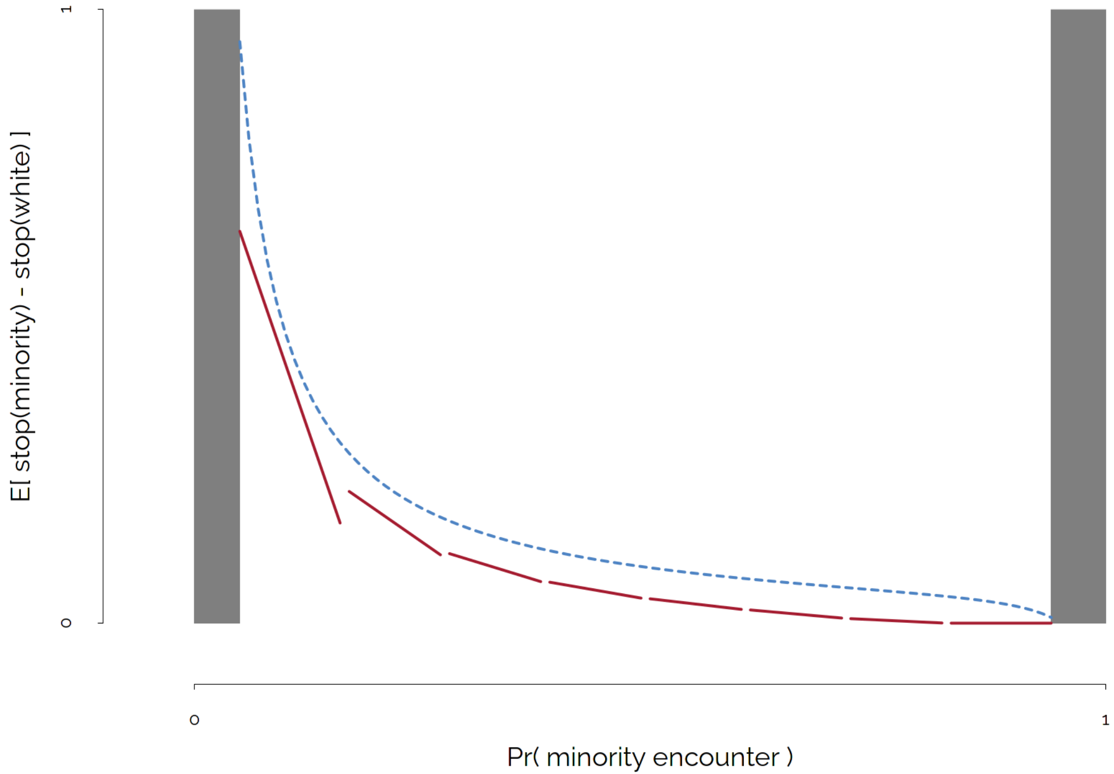
The procedure

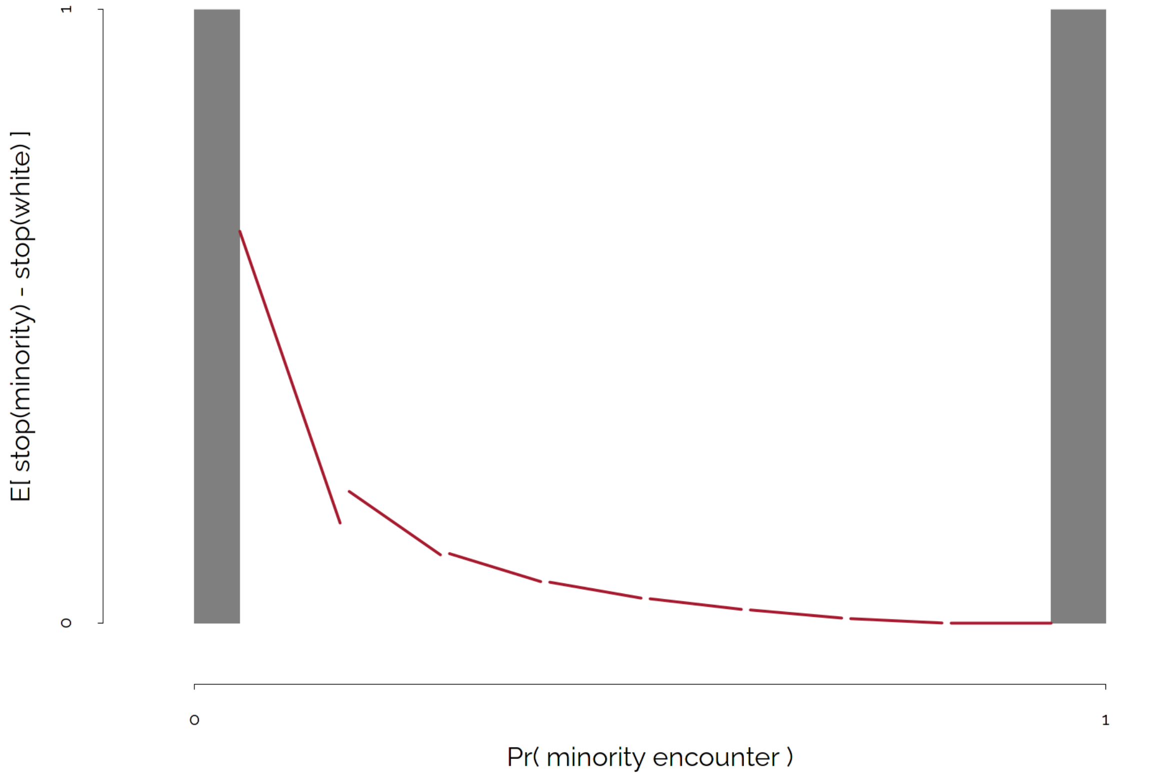
- Spatial branch & bound (Land & Doig '60)
  - Recursively divide model space into branches
  - Eliminate branches that cannot possibly be optimal
  - Much faster than brute-force enumeration


Pr(minority encounter)
Pr (always stop)
Pr(anti-minority stop)

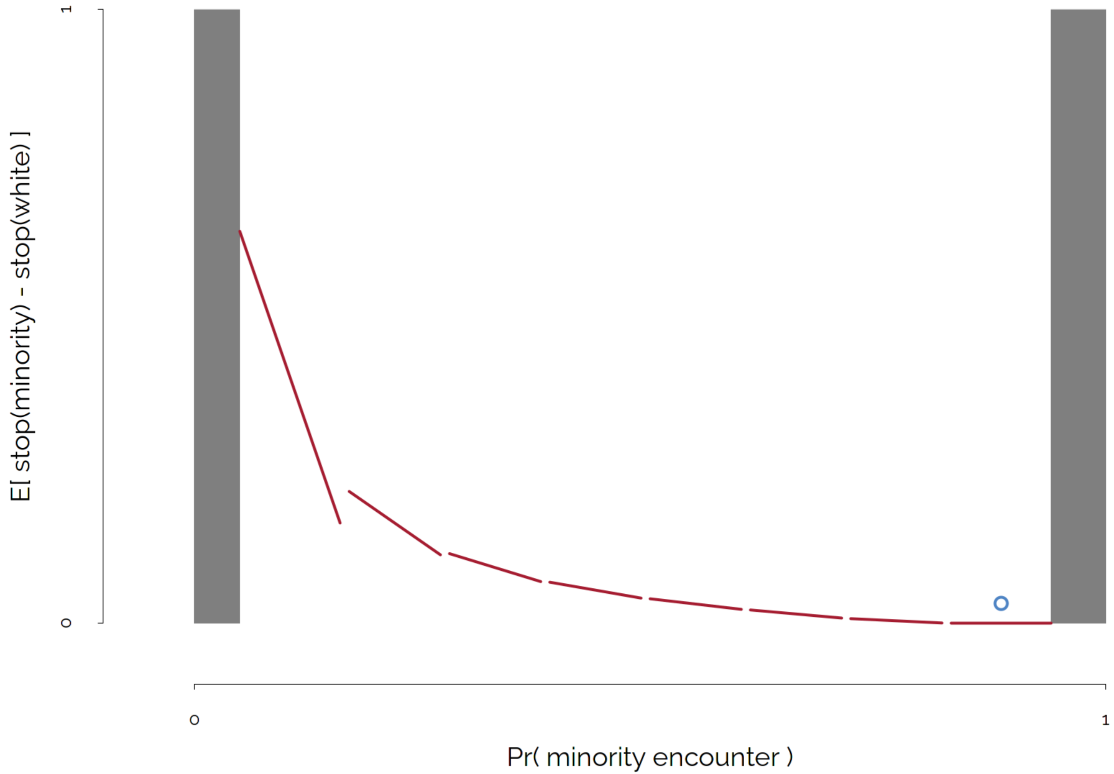


1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)

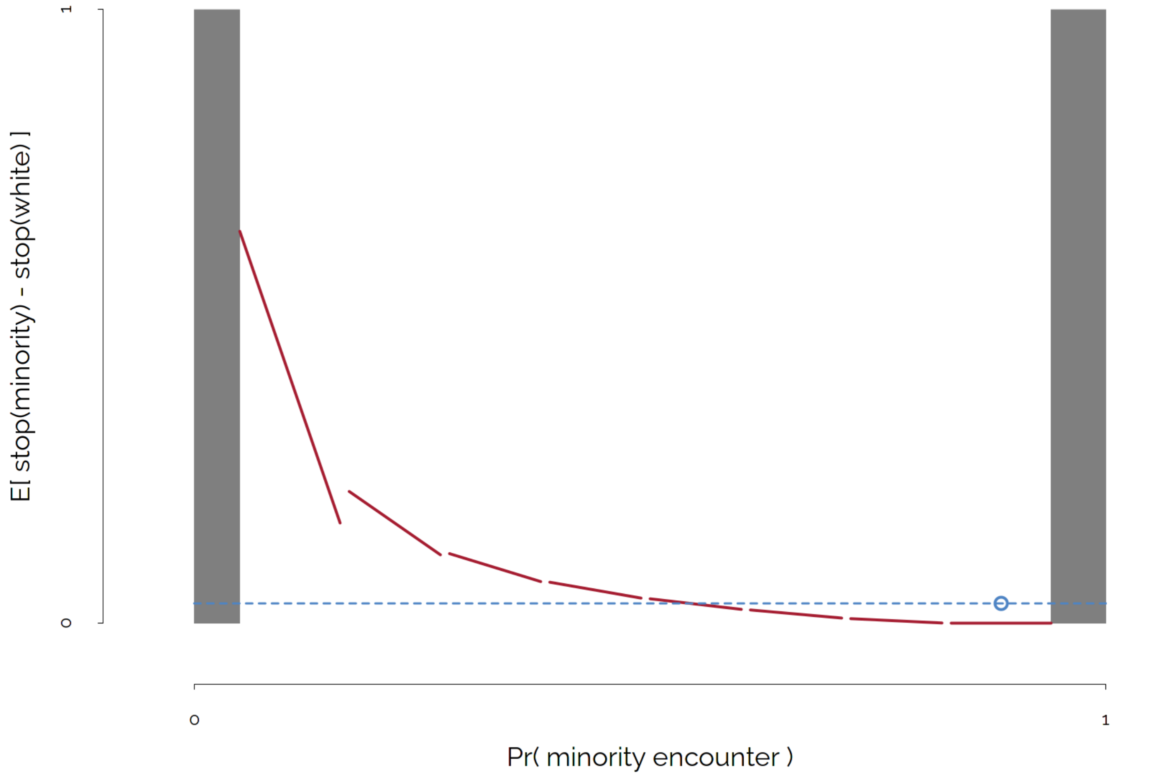

E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)

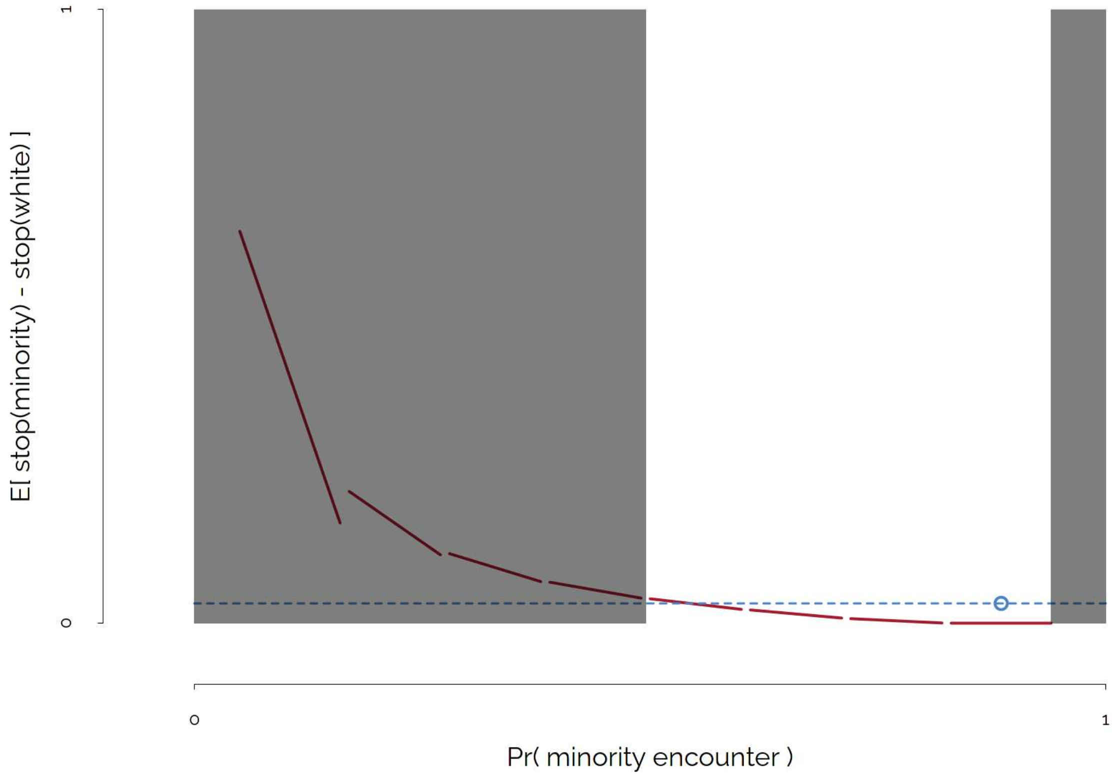


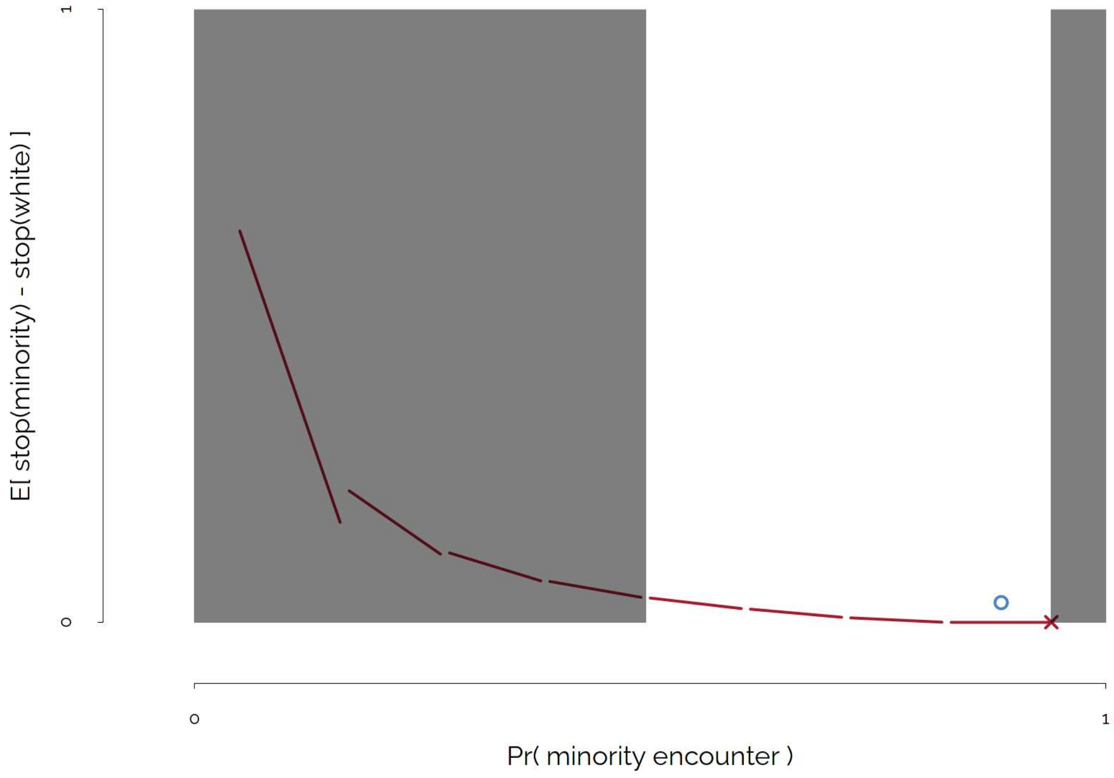
E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)



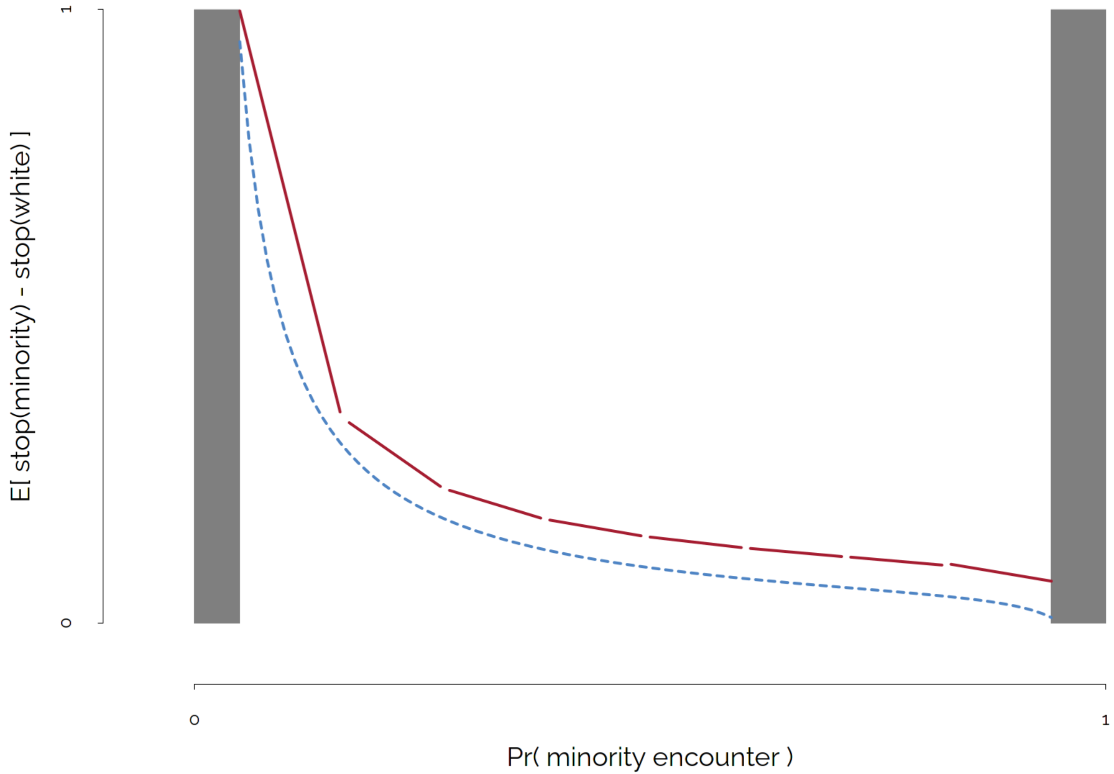
1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)

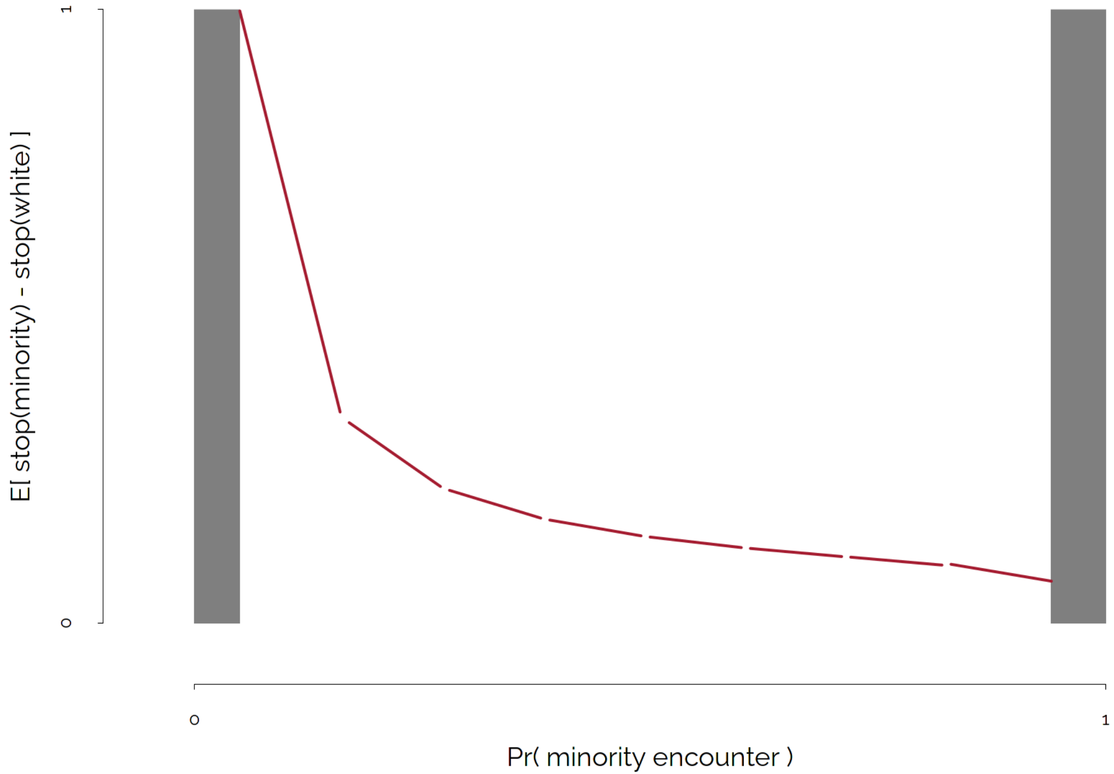

E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)

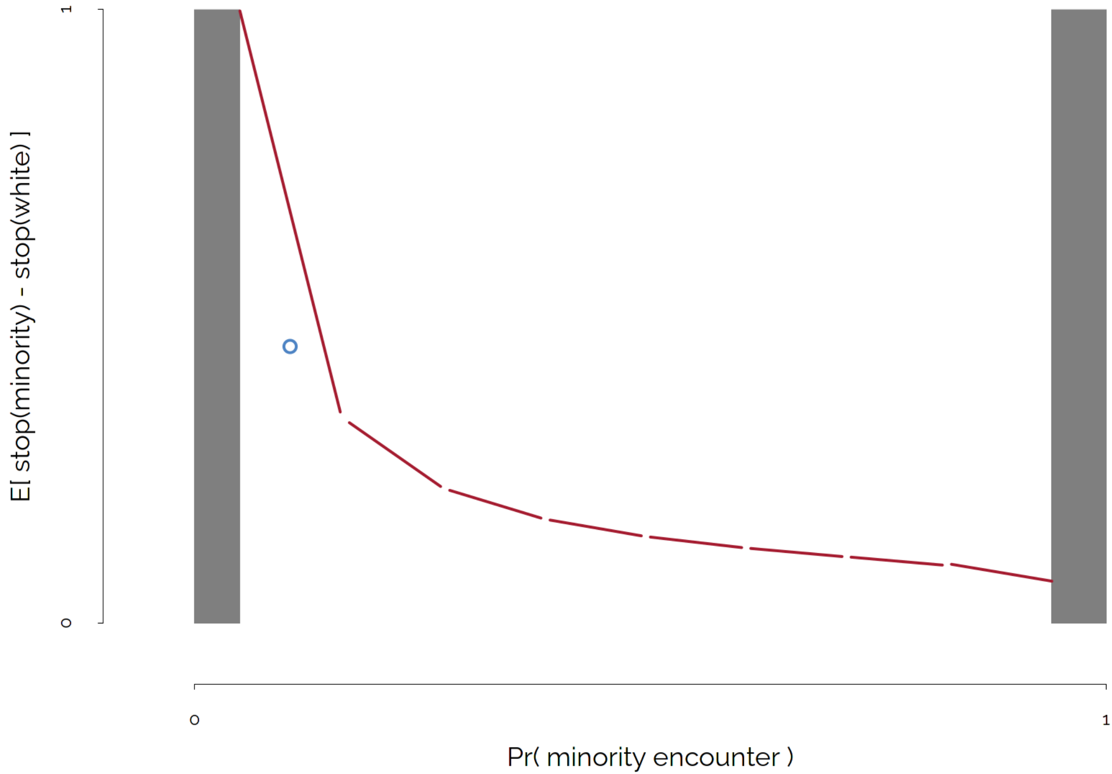

1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)



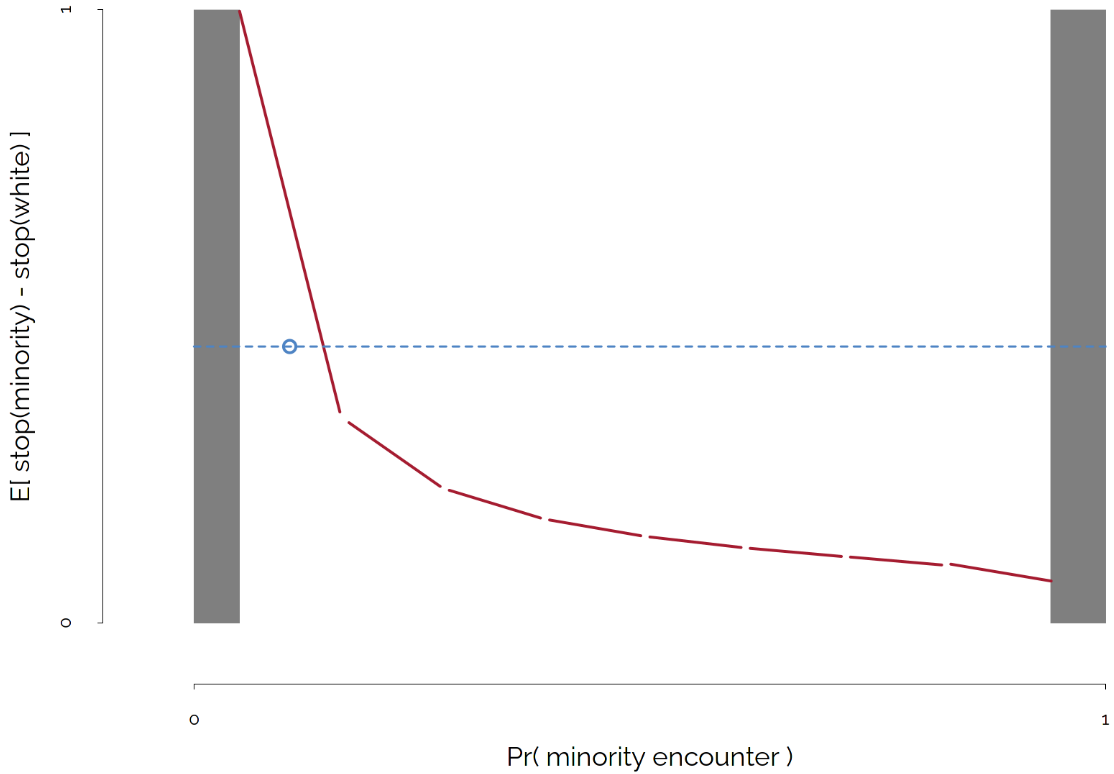
1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)


E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)

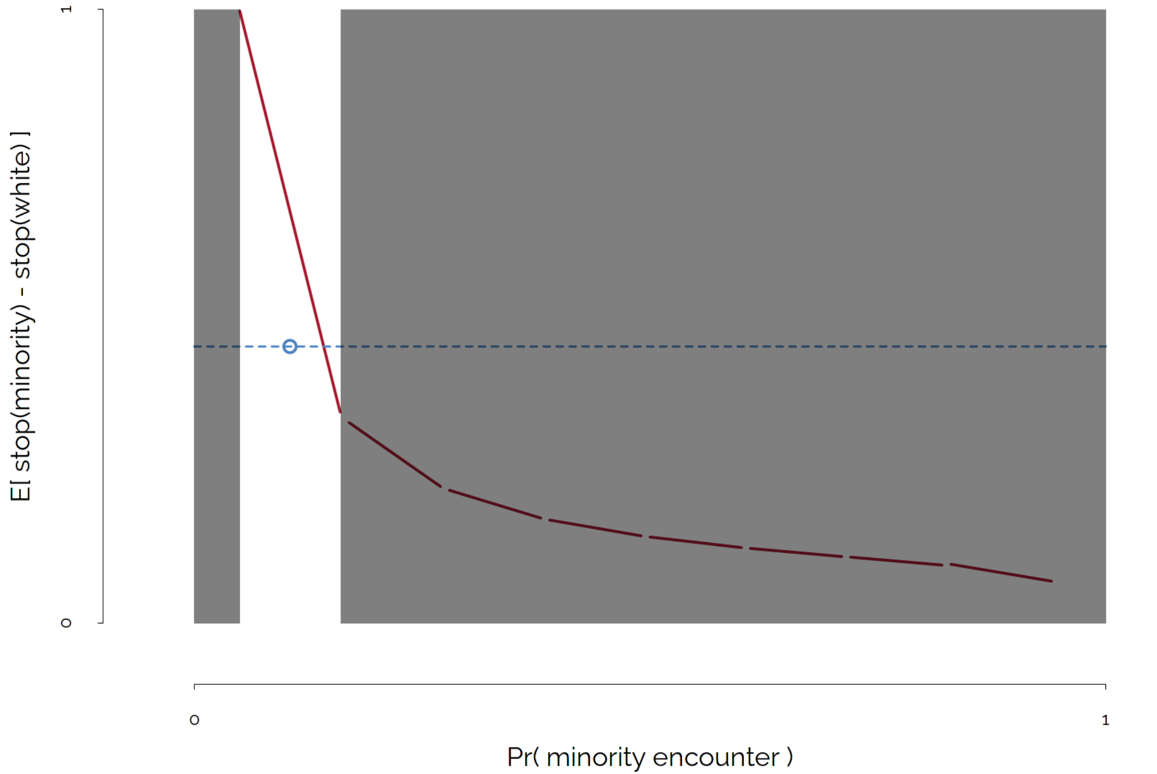

E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)



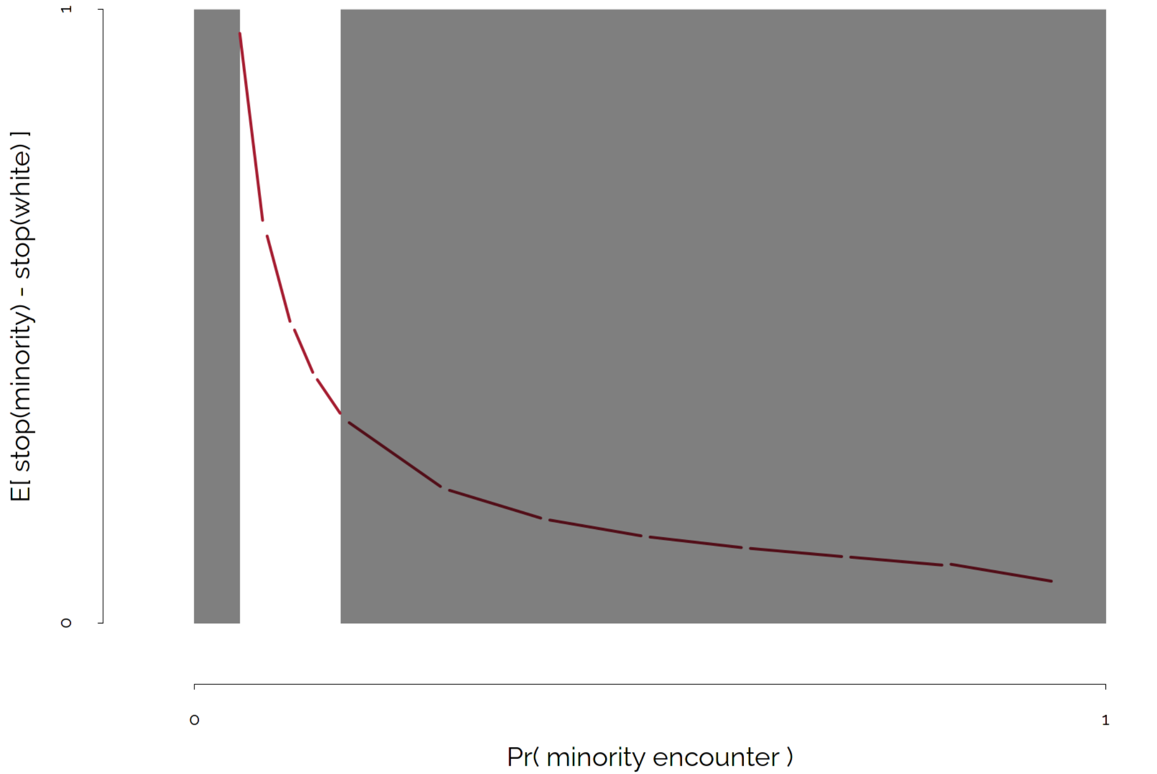
E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)

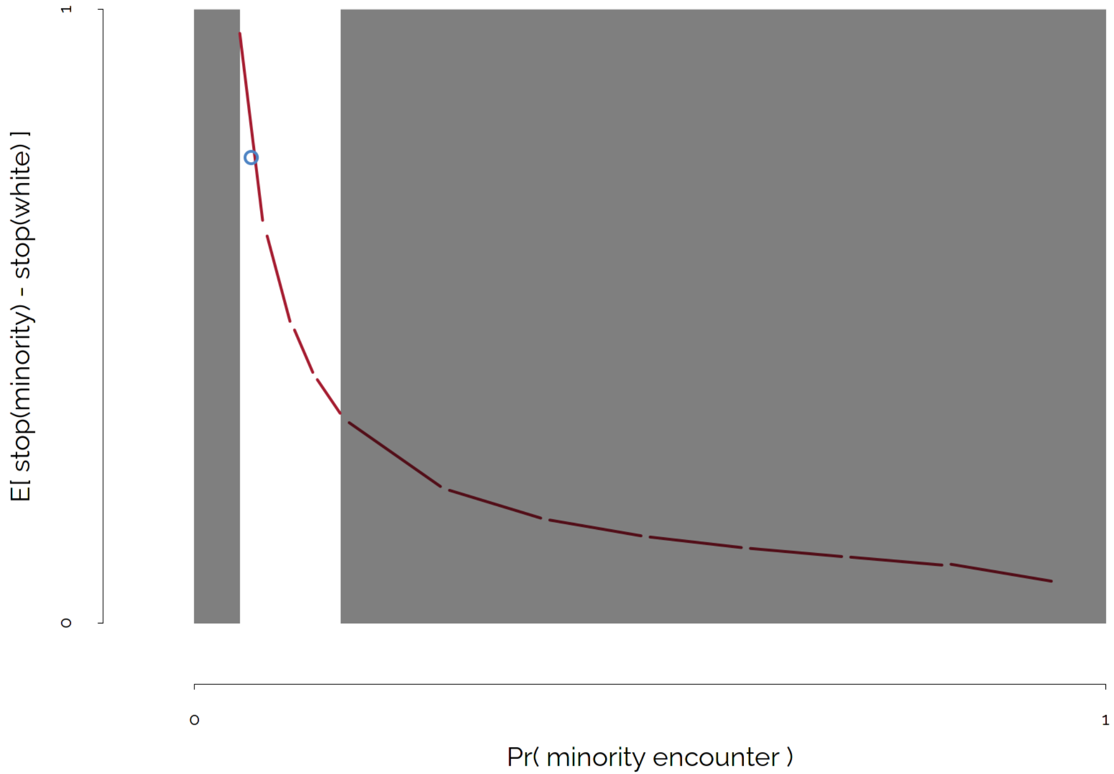


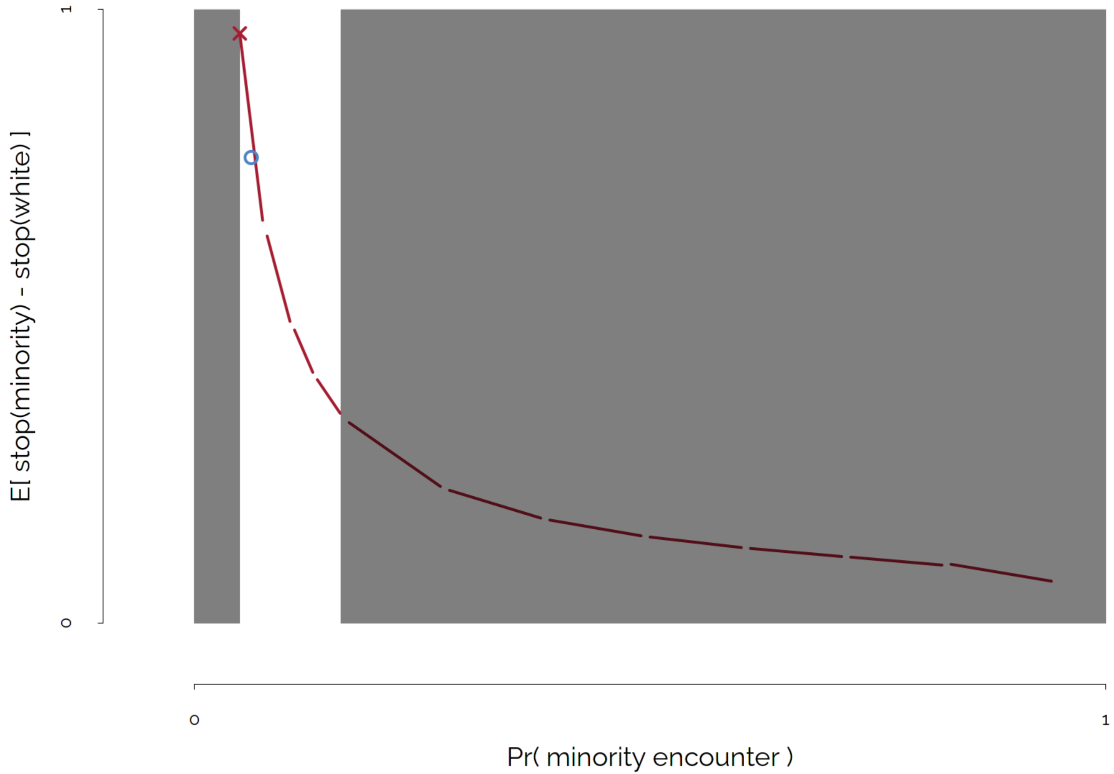
1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)



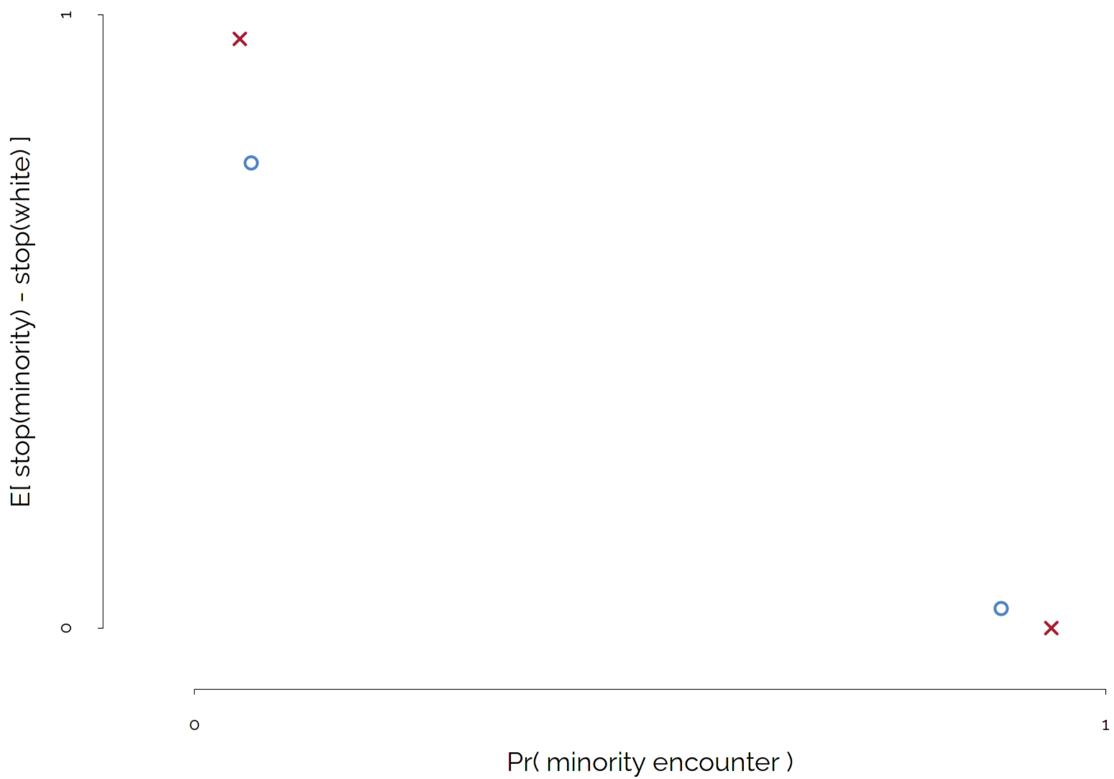
E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)


E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)


E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)

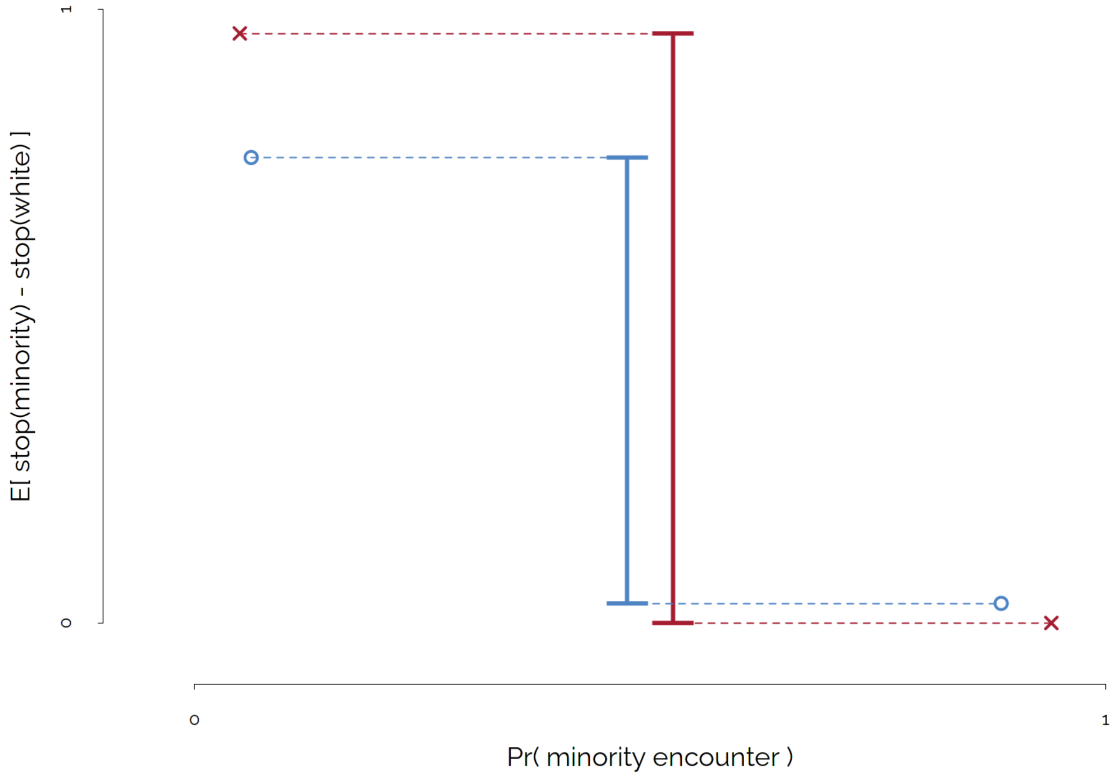


1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)


E[stop(minority) - stop(white)]
1
0
0
1
Pr(minority encounter)


max known feasible point
min known feasible point
E[stop(minority) - stop(white)]
Pr(minority encounter)
0
1
0
1


min guaranteed valid upper bound
max guaranteed valid lower bound
1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)



1
0
E[stop(minority) - stop(white)]
0
1
Pr(minority encounter)

The procedure

- Spatial branch & bound (Land & Doig '60)
  - Recursively divide model space into branches
  - Eliminate branches that cannot possibly be optimal
  - Much faster than brute-force enumeration

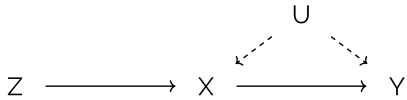
The procedure

- Spatial branch & bound (Land & Doig '60)
  - Recursively divide model space into branches
  - Eliminate branches that cannot possibly be optimal
  - Much faster than brute-force enumeration

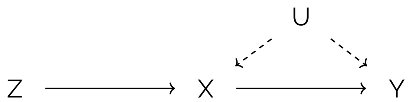
- Many algo. components (SCIP, Gamrath et al. '20; Vigerske & Gleixner '17)
  - Presolving (eliminate redundant variables, constraints)
  - Efficient branching strategies, primal exploration heuristics
  - Linear-programming relaxations for local dual problem

Modularizing and testing assumptions

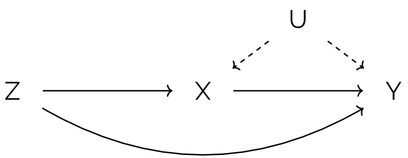
True data-generating process



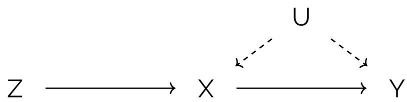
True data-generating process



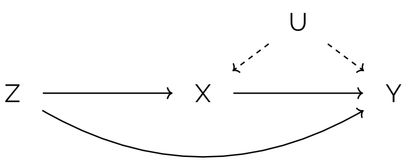
Overly cautious assumptions



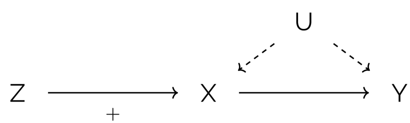
True data-generating process

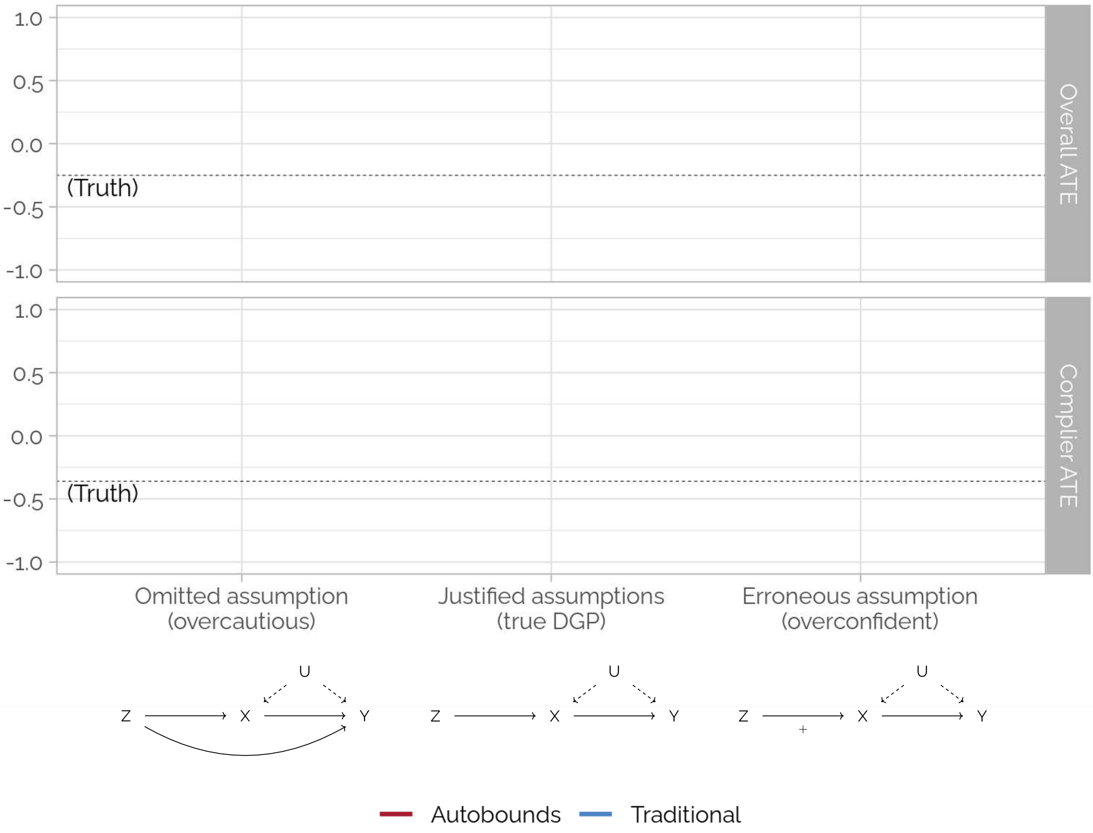


Overly cautious assumptions

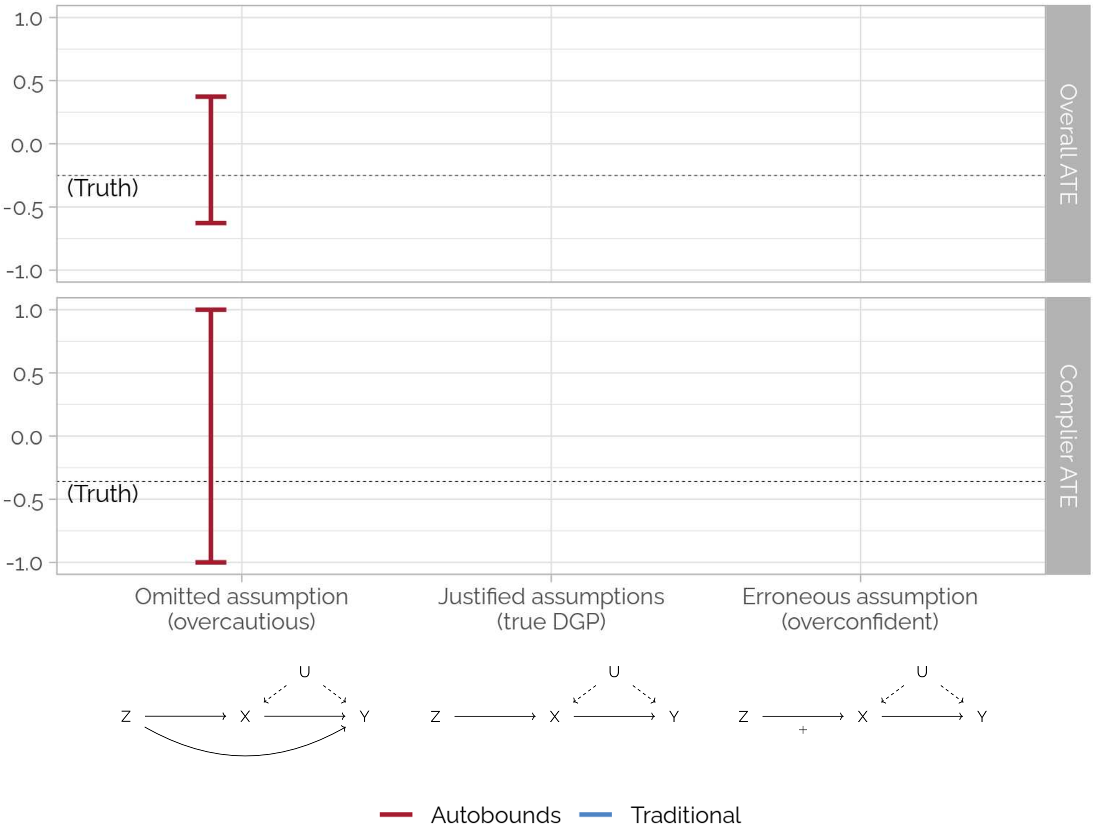


Overly confident assumptions

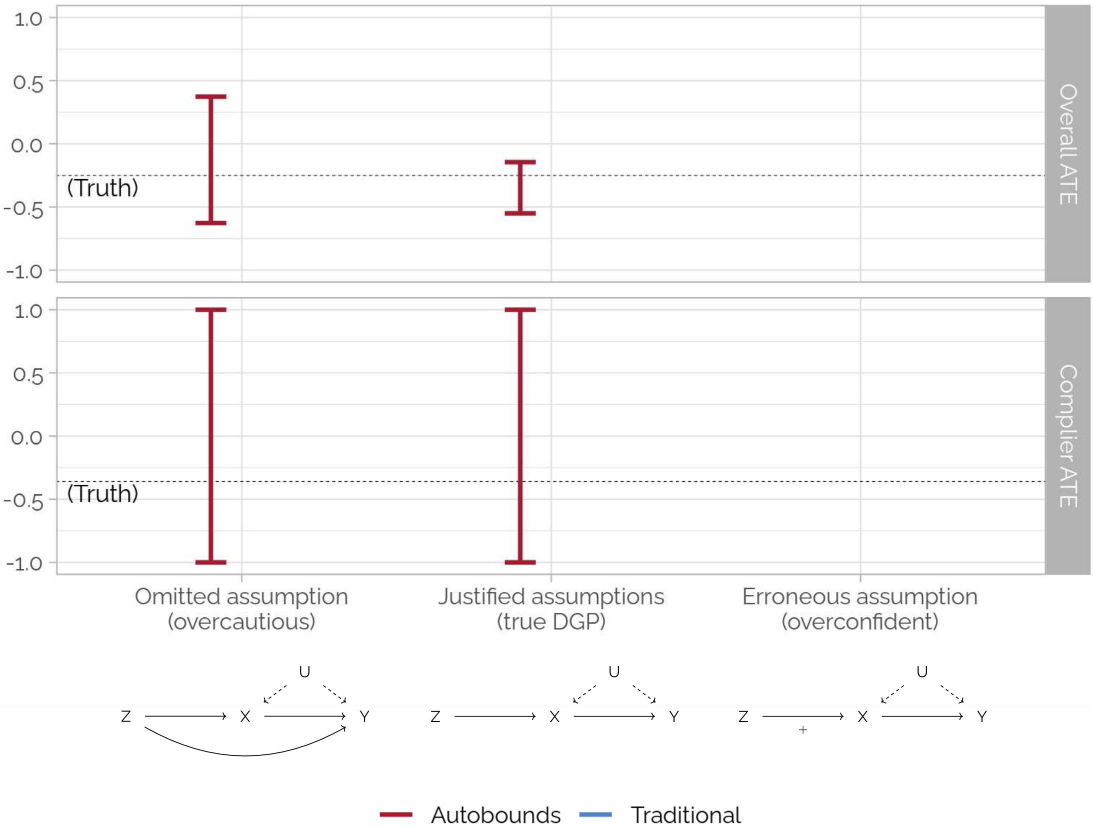




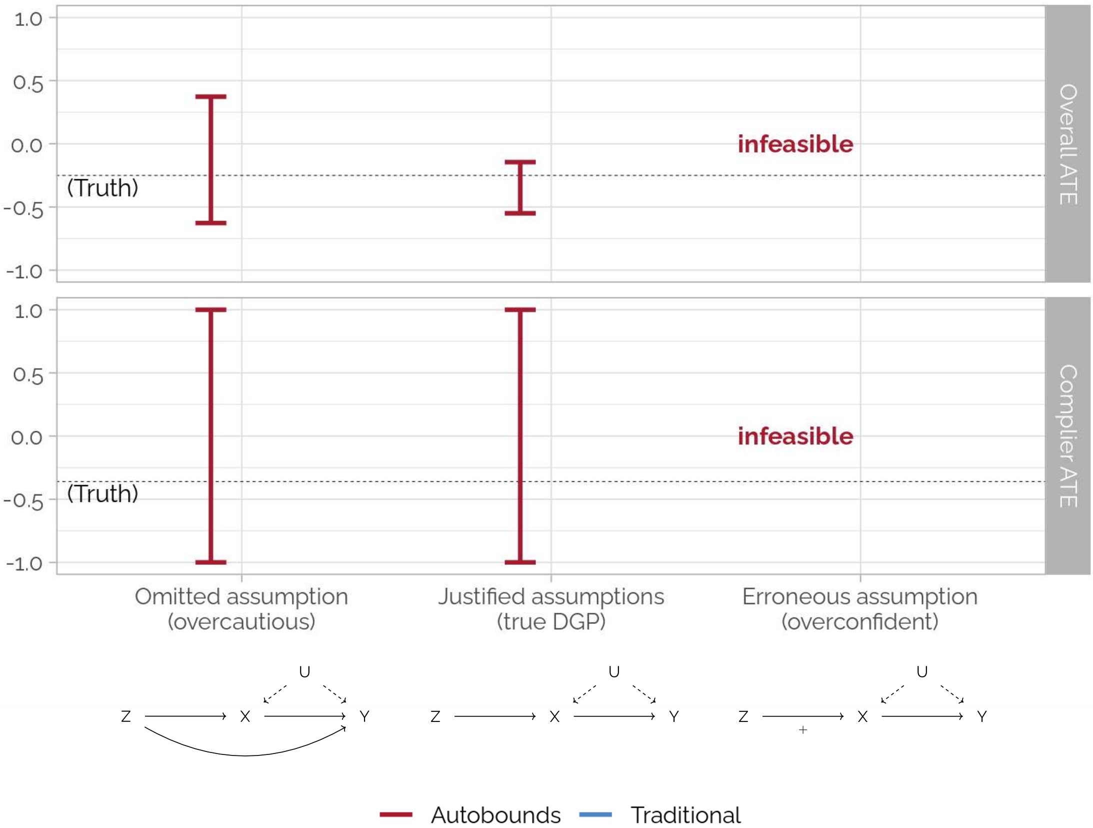
1.0
0.5
0.0
-0.5
-1.0
(Truth)
Overall ATE
Complier ATE
Omitted assumption (overcautious)
Justified assumptions (true DGP)
Erroneous assumption (overconfident)
U
Z
X
Y
+
Autobounds
Traditional



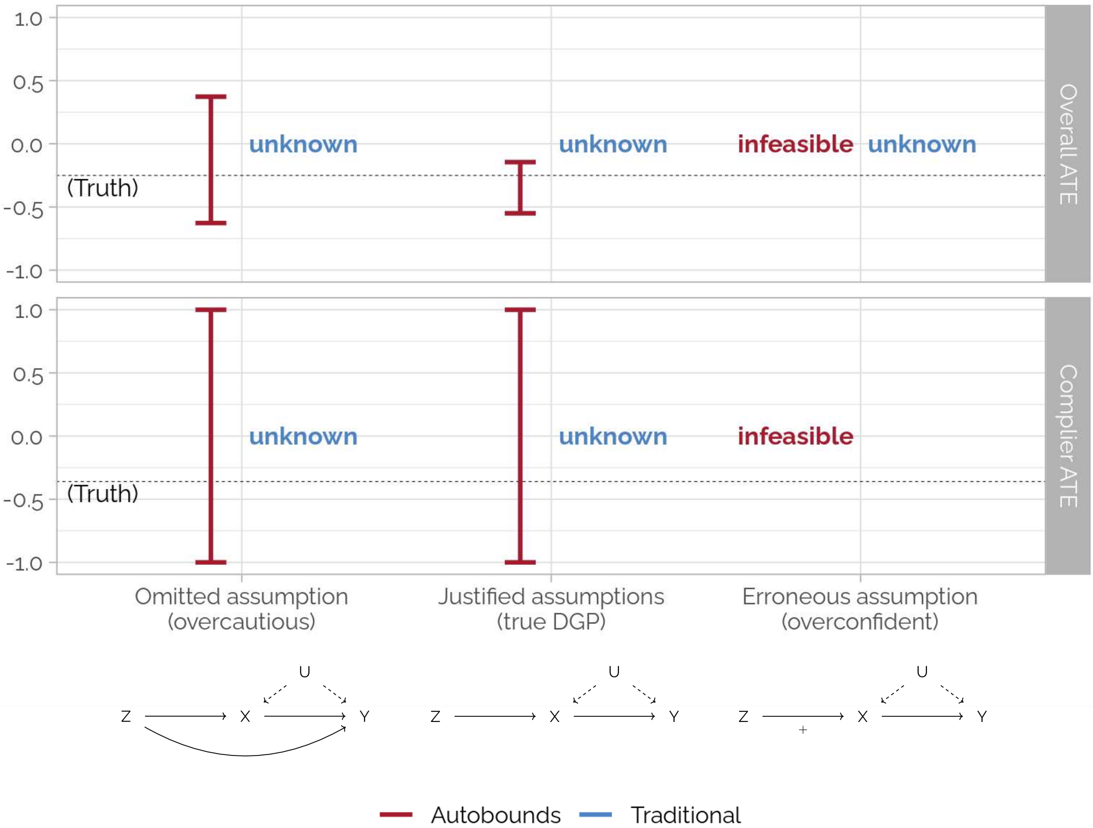
1.0
0.5
0.0
-0.5
-1.0
(Truth)
Overall ATE
Complier ATE
Omitted assumption
(overcautious)
Justified assumptions
(true DGP)
Erroneous assumption
(overconfident)
U
Z
X
Y
+
Autobounds
Traditional



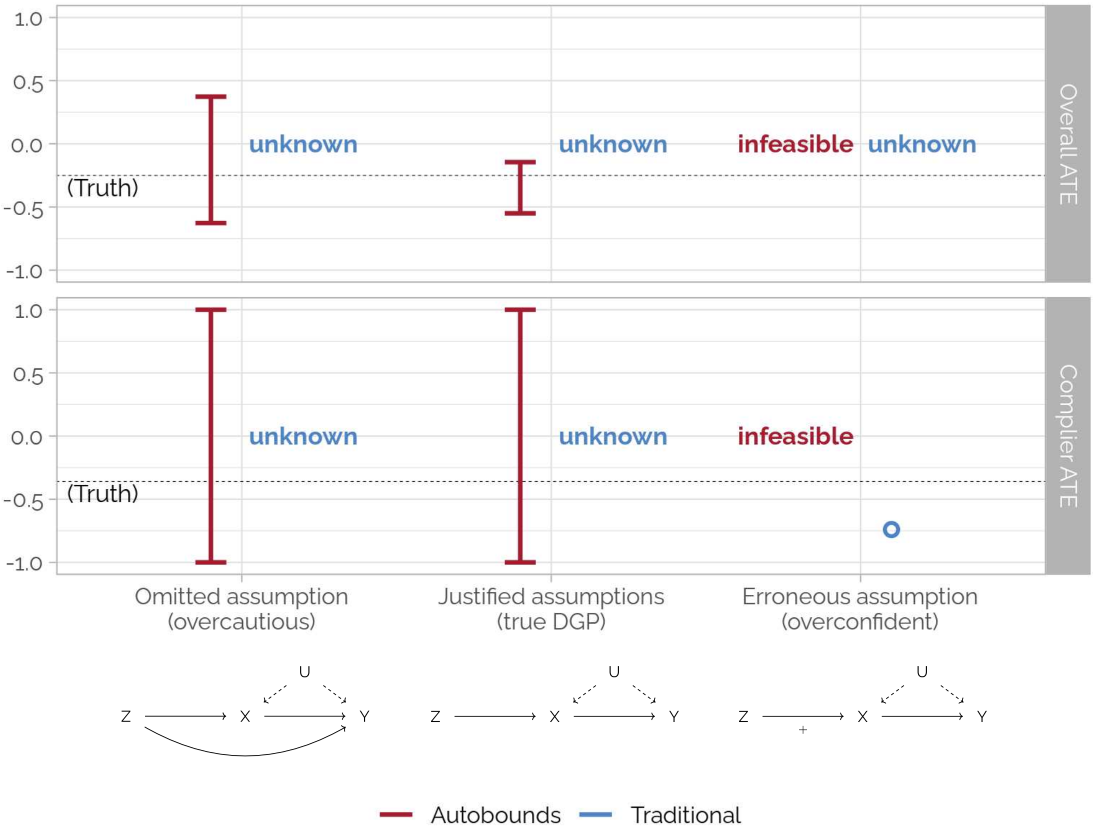
1.0
0.5
0.0
-0.5
-1.0
(Truth)
Overall ATE
Complier ATE
Omitted assumption
(overcautious)
Justified assumptions
(true DGP)
Erroneous assumption
(overconfident)
U
Z
X
Y
+
Autobounds
Traditional



1.0
0.5
0.0
-0.5
-1.0
(Truth)
infeasible
Overall ATE
Complier ATE
Omitted assumption
(overcautious)
Justified assumptions
(true DGP)
Erroneous assumption
(overconfident)
U
Z
X
Y
+
Autobounds
Traditional

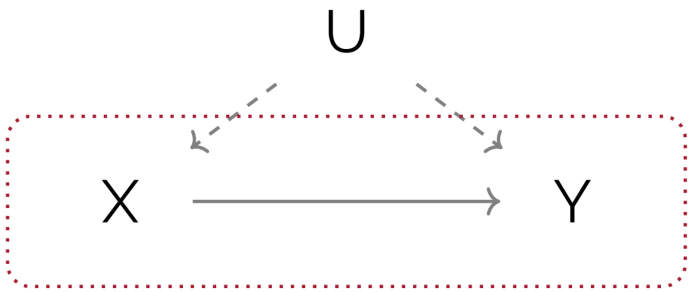


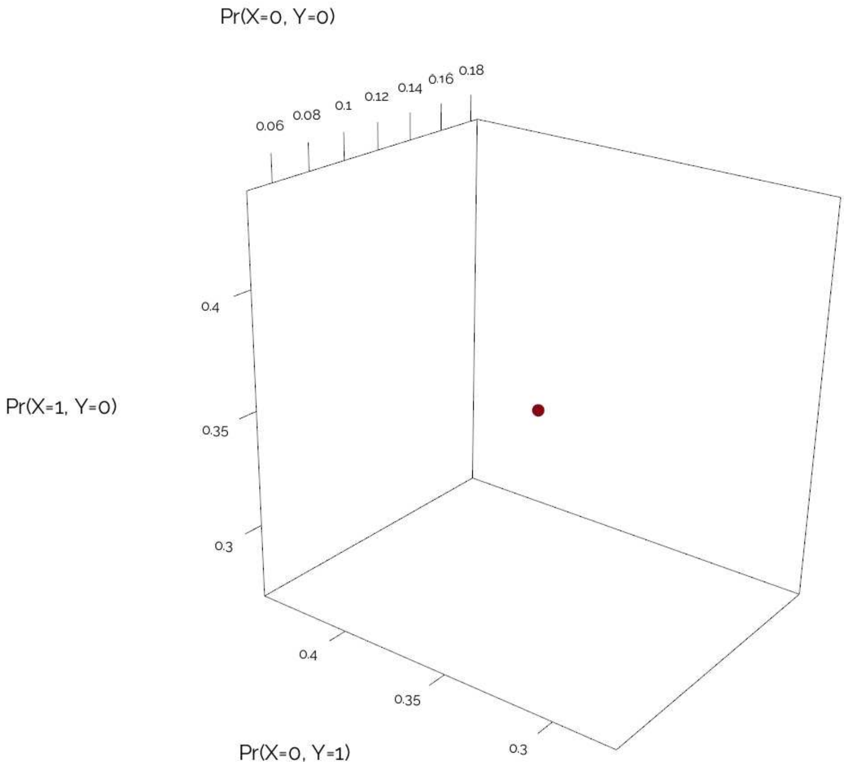
1.0
0.5
0.0
-0.5
-1.0
(Truth)
unknown
unknown
infeasible
unknown
Overall ATE
Complier ATE
Omitted assumption
(overcautious)
Justified assumptions
(true DGP)
Erroneous assumption
(overconfident)
U
Z
X
Y
+
Autobounds
Traditional



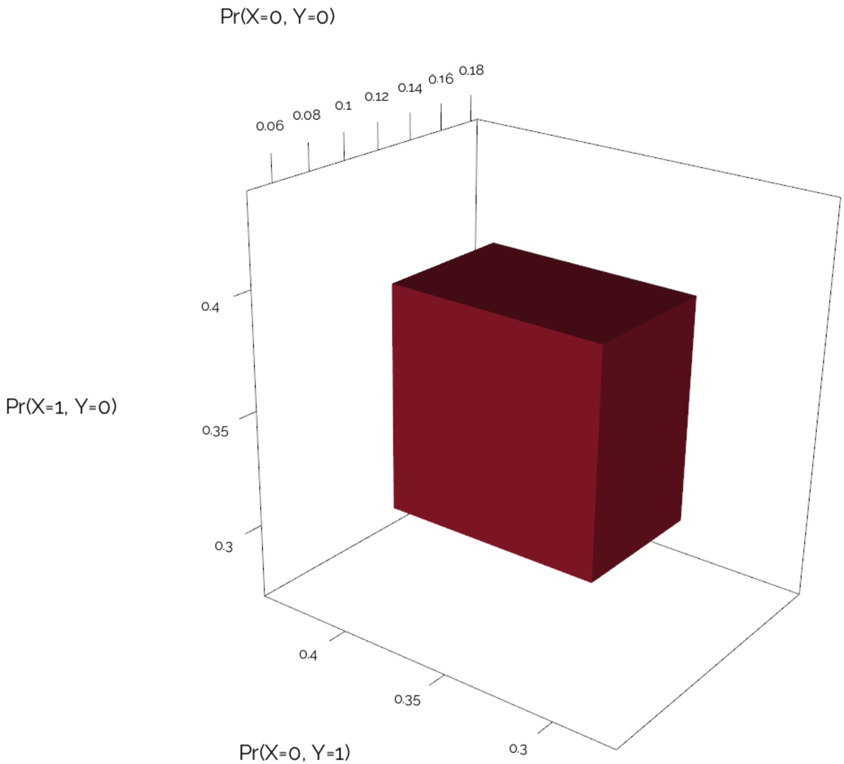
1.0
0.5
0.0
-0.5
-1.0
(Truth)
unknown
unknown
infeasible
unknown
Overall ATE
Complier ATE
Omitted assumption
(overcautious)
Justified assumptions
(true DGP)
Erroneous assumption
(overconfident)
U
Z
X
Y
+
Autobounds
Traditional

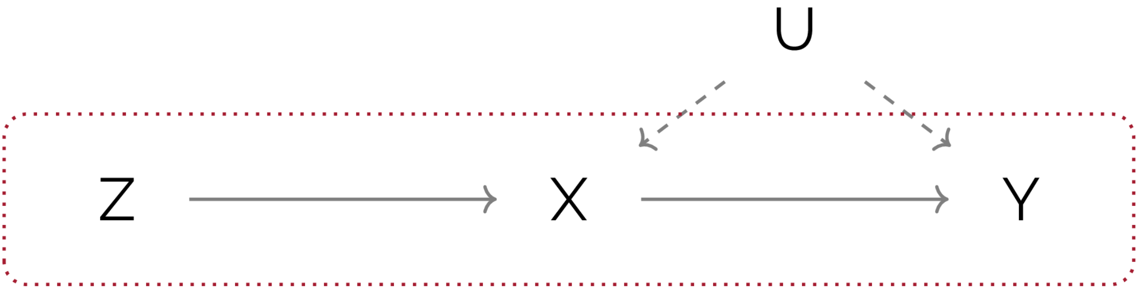
Inference on ε-sharp bounds

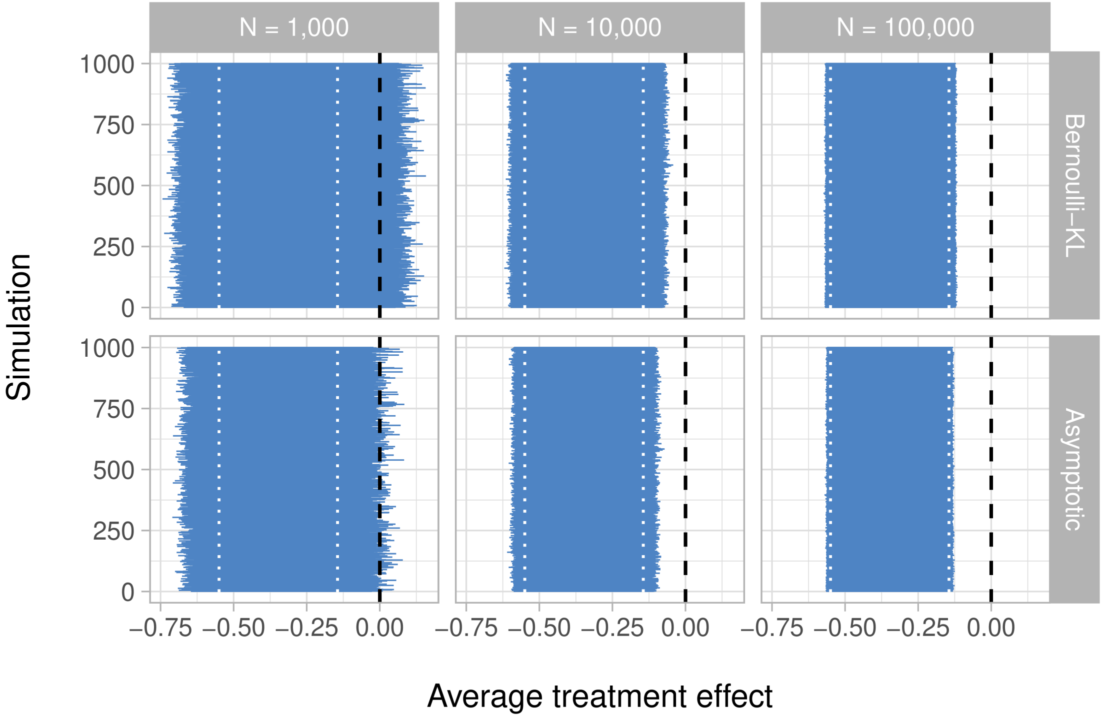

U
X
Y



Pr(X=0, Y=0)
0.06
0.08
0.1
0.12
0.14
0.16
0.18
Pr(X=1, Y=0)
0.4
0.35
0.3
Pr(X=0, Y=1)
0.4
0.35
0.3


Pr(X=0, Y=0)
0.06
0.08
0.1
0.12
0.14
0.16
0.18
Pr(X=1, Y=0)
0.4
0.35
0.3
Pr(X=0, Y=1)
0.4
0.35
0.3


U
Z
X
Y



N = 1,000
N = 10,000
N = 100,000
Bernoulli–KL
Asymptotic
Simulation
1000
750
500
250
0
−0.75
−0.50
−0.25
0.00
Average treatment effect

Potential critiques of autobounding

Potential critiques

- "The bounds may be too wide to be informative"

Potential critiques

- "The bounds may be too wide to be informative"
    - Yes.

Potential critiques

- "The bounds may be too wide to be informative"
  - Yes. This is a fact about the universe

Potential critiques

- "The bounds may be too wide to be informative"
  - Yes. This is a fact about the universe
  - Solution: collect more data or justify more assumptions

Potential critiques

- "The bounds may be too wide to be informative"
  - Yes. This is a fact about the universe
  - Solution: collect more data or justify more assumptions
- "The user must know the true causal model"
  - Causal inference without assumptions is impossible
  - Our approach allows them to be relaxed modularly

Potential critiques

- "The bounds may be too wide to be informative"
    - Yes. This is a fact about the universe
    - Solution: collect more data or justify more assumptions
- "The user must know the true causal model"
    - Causal inference without assumptions is impossible
    - Our approach allows them to be relaxed modularly
- "What about continuous variables?"
    - Discrete systems are a large part of applied work.
    - With more data, it is feasible to discretize variables.

Potential critiques

- "The bounds may be too wide to be informative"
  - Yes. This is a fact about the universe
  - Solution: collect more data or justify more assumptions
- "The user must know the true causal model"
  - Causal inference without assumptions is impossible
  - Our approach allows them to be relaxed modularly
- "What about continuous variables?"
  - Discrete systems are a large part of applied work.
  - With more data, it is feasible to discretize variables.
- "The bounds will take too long to compute"
  - Researchers can still narrow the range of answers
  - Our algorithm is "anytime," meaning bounds are always valid
  - More efficient methods are a key direction for future work