Causal Fairness Analysis

(Causal Inference II - Lecture 5)

Elias Bareinboim

Drago Plecko

Columbia University
Computer Science

Reference:

D. Plecko, E. Bareinboim.

Causal Fairness Analysis.

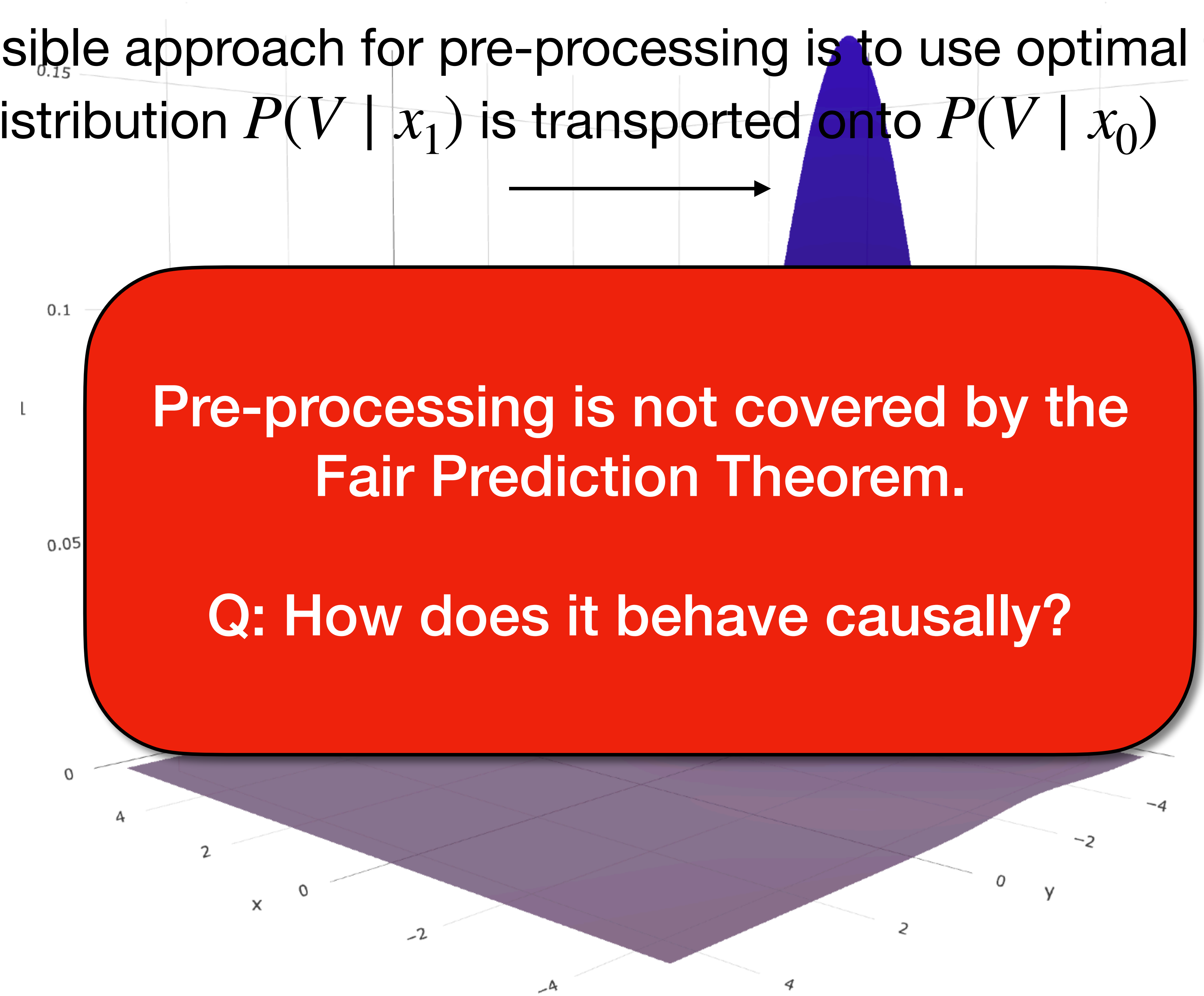
TR R-90, CausalAI Lab, Columbia University.

https://causalai.net/r90.pdf

Task 2. Fair Predictions (Continued)

Failure of Optimal Transport (in the Individual Fairness framework)

- A possible approach for pre-processing is to use optimal transport
- The distribution $P(V \mid x_1)$ is transported onto $P(V \mid x_0)$



Pre-processing is not covered by the Fair Prediction Theorem.

Q: How does it behave causally?

Failure of Optimal Transport (in the Individual Fairness framework)

- A common approach for pre-processing is to use optimal transport
- The distribution $P(V \mid x_1)$ is transported onto $P(V \mid x_0)$

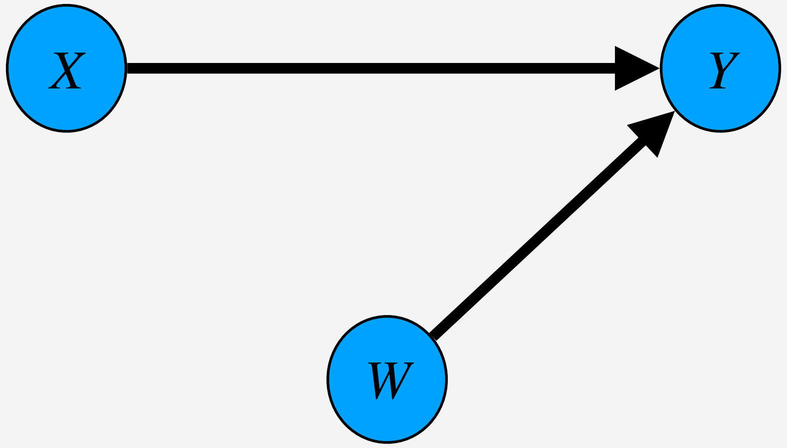
Example.

$$X \leftarrow U_X$$

$$W \leftarrow \epsilon(2U_W - 1)$$

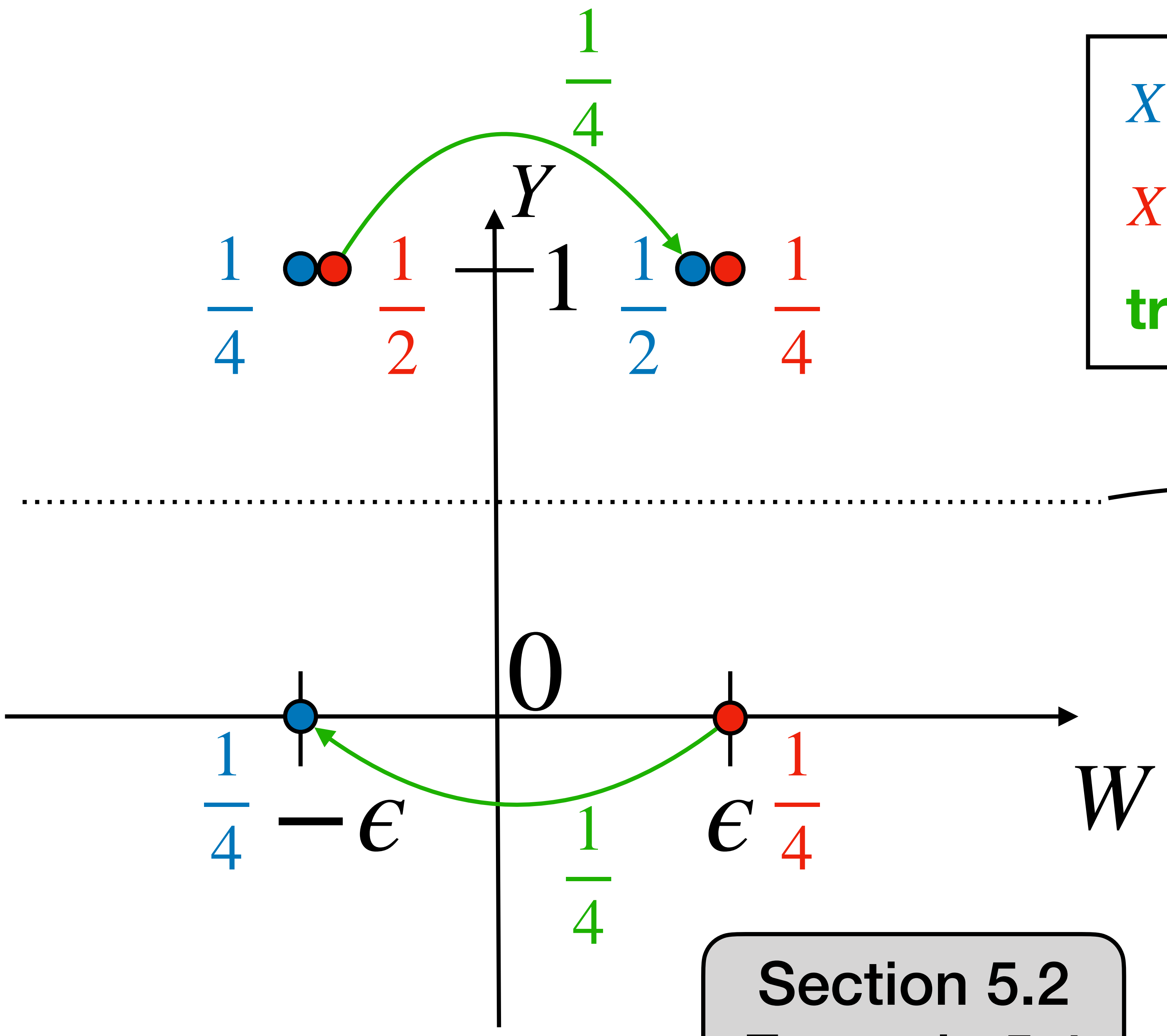
$$Y \leftarrow \begin{cases} U_Y \vee 1(W > 0) \textbf{ if } X = x_0 \\ U_Y \vee 1(W < 0) \textbf{ if } X = x_1 \end{cases}$$

U_X, U_W, U_Y **Bernoulli(0.5)**



- In the example, we wish to compute $\text{NIE}_{x_0,x_1}(\tilde{y}) = P(\tilde{y}_{x_0,\widetilde{W}_{x_1}}) - P(\tilde{y}_{x_0})$

Failure of Optimal Transport (in the Individual Fairness framework)



$X = x_0$ **distribution (blue)**

$X = x_1$ **distribution (red)**

transport map τ **(green)**

no transport across y-axis → $\widetilde{W}$ depends on Y → breaks the causal ordering

Section 5.2
Example 5.4

Failure of Optimal Transport (in the Individual Fairness framework)

$$P(\widetilde{y}_{x_0,\widetilde{W}_{x_1}}) = P(\widetilde{y}_{x_0,\epsilon}, \widetilde{W}_{x_1} = \epsilon) + P(\widetilde{y}_{x_0,-\epsilon}, \widetilde{W}_{x_1} = -\epsilon) \quad - \quad P(\widetilde{y}_{x_0}) = P(y_{x_0})$$

using the SCM

$\widetilde{y}_{x_0,\epsilon} = 1$ for any u

$\widetilde{W}_{x_1} = \epsilon$ for $U_W = 1$ w.p. $\frac{1}{2}$

$U_W = 0$ w.p. $\frac{1}{2}$ (1/4 for each U_Y)

$y_{x_0,-\epsilon} = U_Y$

for $U_Y = 1$, $\widetilde{W}_{x_1} = -\epsilon$

with prob. $\frac{1}{4}$ (0 for $U_W = 1$)

$y_{x_0} = U_Y \vee 1(W > 0)$

for $U_Y = 1$, $y_{x_0} = 1$

for $U_Y = 0$, $y_{x_0} = 1$

with prob. $\frac{1}{2}$

putting together

$$P(\widetilde{y}_{x_0,\widetilde{W}_{x_1}}) - P(\widetilde{y}_{x_0}) = \frac{1}{2} + \frac{1}{8} - \frac{3}{4} = -\frac{1}{8} \implies$$

Indirect Effect $\neq 0$!

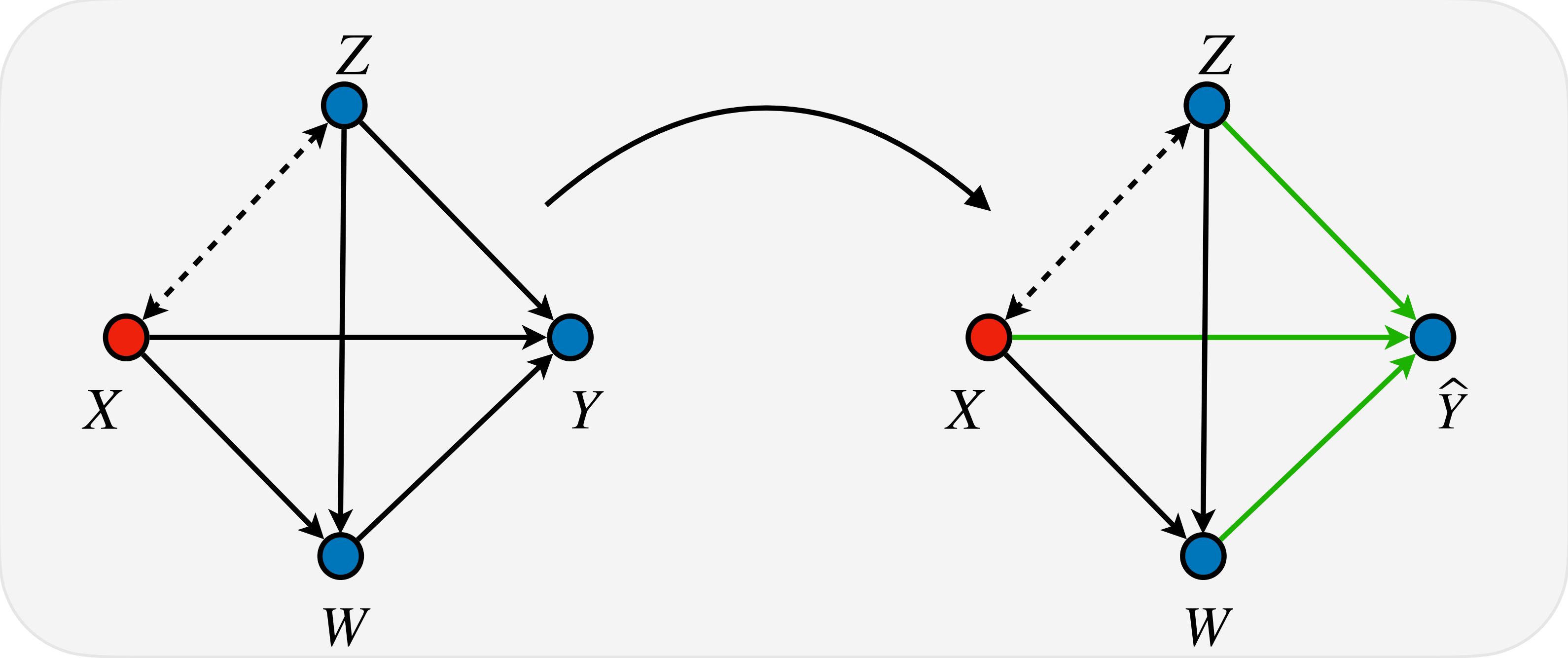
Towards the solution

- how can we construct “causal” fair predictions?

(i) causal structure of the SFM is preserved for the predictor $\widehat{Y}$

Section 5.2.8

(ii) identification expressions for x-DE, x-SE, and x-IE equal 0 for the predictor $\widehat{Y}$



$$x\text{-}\mathbf{DE}^{ID}_{x_0,x_1}(\hat{y}) = \sum_{z,w} [P(\hat{y} \mid x_1, z, w) - P(\hat{y} \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x_0) = 0$$

$$x\text{-}\mathbf{IE}^{ID}_{x_0,x_1}(\hat{y}) = \sum_{z,w} P(\hat{y} \mid x_1, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x) = 0$$

$$x\text{-}\mathbf{SE}^{ID}_{x_1,x_0}(\hat{y}) = \sum_{z} P(\hat{y} \mid x_1, z)[P(z \mid x_1) - P(z \mid x_0)] = 0.$$

In-processing solution

Theorem. Let $\widehat{Y}$ be the solution to the following optimization problem:

$$\widehat{Y} = \textbf{argmin}_f \quad E[Y - f(X, Z, W)]^2$$

$$\textbf{subject to} \quad x\textbf{-DE}^{\textbf{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0$$

$$x\textbf{-DE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0$$

$$x\textbf{-IE}^{\textbf{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0$$

$$x\textbf{-IE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0$$

$$x\textbf{-SE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y}) = 0$$

Section 5.2.9
Theorem 5.2

Then $\widehat{Y}$ satisfies

$$x\text{-DE}_{x_0,x_1}(\widehat{y} \mid x_0) = x\text{-IE}_{x_1,x_0}(\widehat{y} \mid x_0) = x\text{-SE}_{x_1,x_0}(\widehat{y}) = 0.$$

Pre-processing solution (Causal IF)

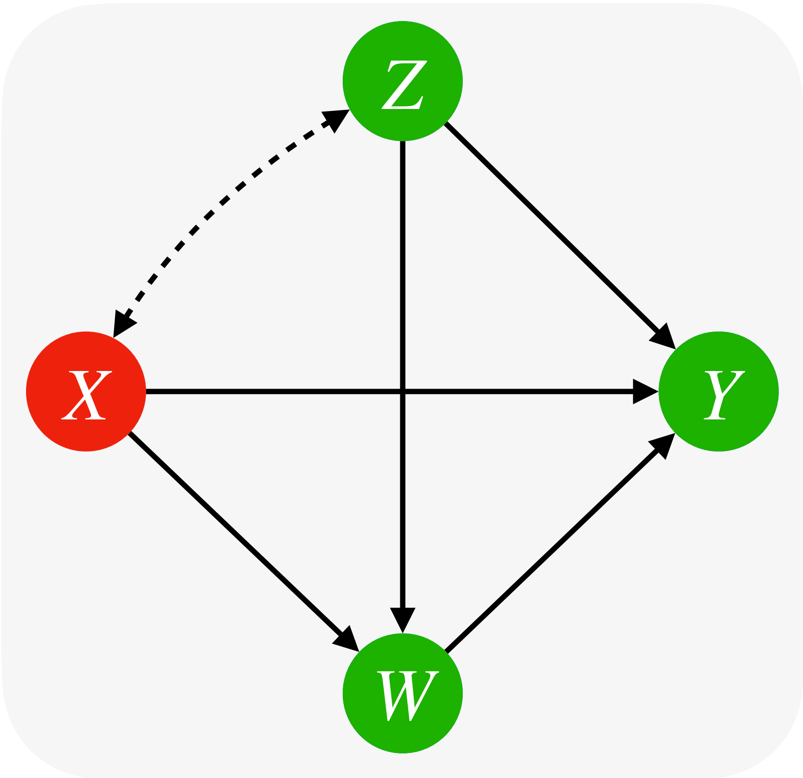
Definition. The Causal Individual Fairness (Causal IF, for short) algorithm is performed on a data coming from an SCM $\mathscr{M}$ compatible with the standard fairness model (SFM), in the following way:

1) if $Z \notin$ BN-set, transport
$Z \mid x_1 \mapsto Z \mid x_0$

2) if $W \notin$ BN-set, transport
$W \mid x_1, Z = z \mapsto W \mid x_0, Z = z$

3) transport
$Y \mid x_1, Z = z, W = w \mapsto Y \mid x_0, Z = z, W = w$

SFM



Data $\mathscr{D}$

X	Z	W	Y
$x^{(1)}$	$\tilde{z}^{(1)}$	$\tilde{w}^{(1)}$	$\tilde{y}^{(1)}$
$x^{(1)}$	$\tilde{z}^{(2)}$	$\tilde{w}^{(2)}$	$\tilde{y}^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x^{(n)}$	$\tilde{z}^{(n)}$	$\tilde{w}^{(n)}$	$\tilde{y}^{(n)}$

Section 5.2
Algorithm 2

Pre-processing solution (Causal IF)

Theorem. Let $\mathscr{M}$ be an SCM compatible with the SFM. Let τ be the optimal transport map obtained when applying Causal IF. Define a new, additional mechanism of the SCM $\mathscr{M}$ such that

$$\widetilde{Y} \leftarrow \tau^{Y}(Y; X, Z, W)\,.$$

Section 5.2
Theorem 5.3

For the transformed outcome $\widetilde{Y}$ we can then claim:

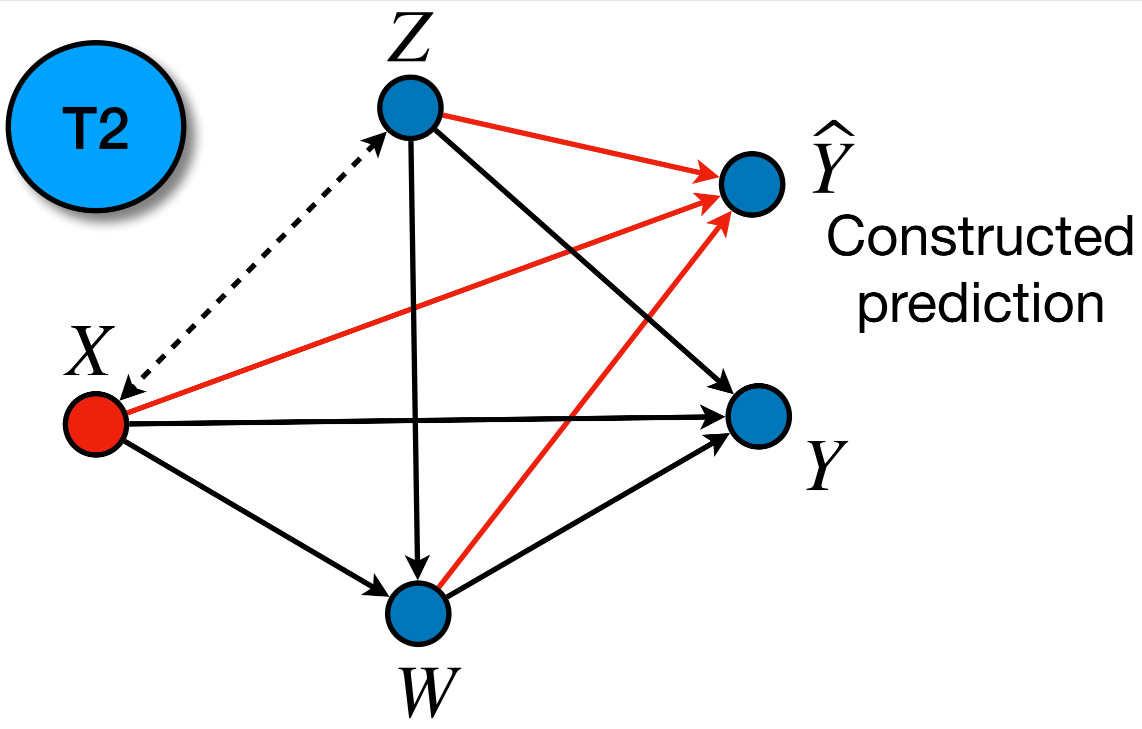
$$\text{if } Z \notin \text{BN-set} \implies x\text{-SE}_{x_1,x_0}(\widetilde{y}) = 0.$$

$$\text{if } W \notin \text{BN-set} \implies x\text{-IE}_{x_1,x_0}(\widetilde{y} \mid x_0) = 0.$$

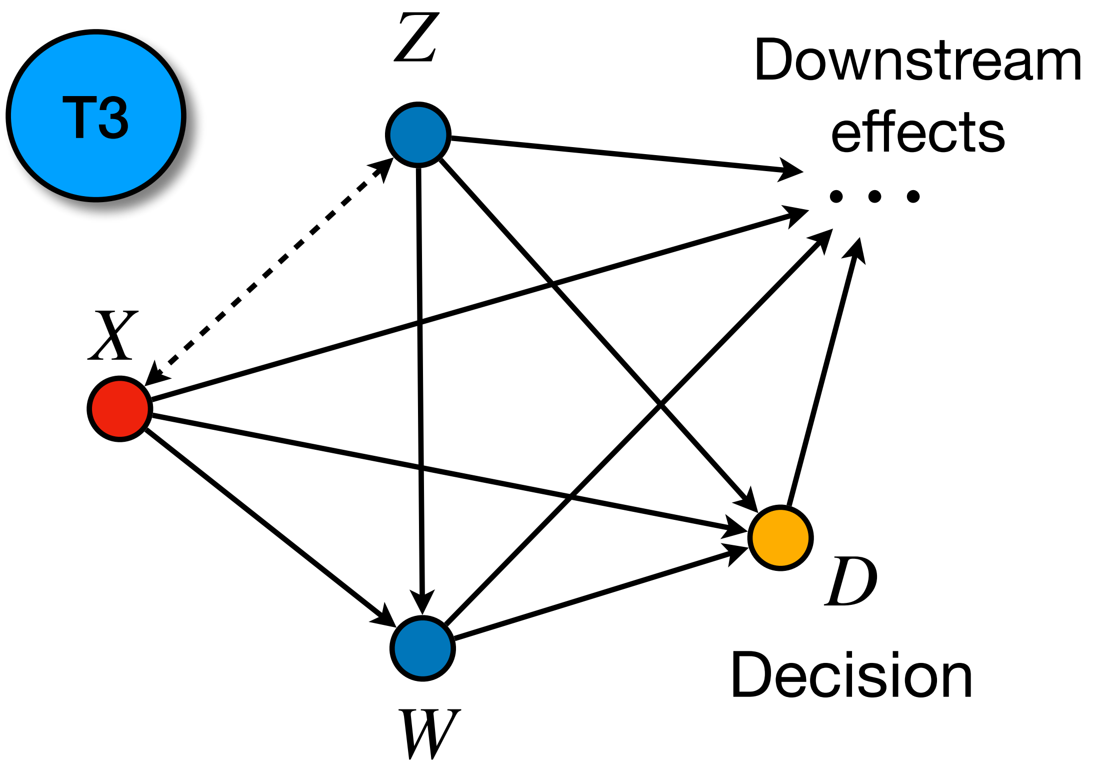
Furthermore, the transformed outcome $\widetilde{Y}$ also satisfies

$$x\text{-DE}_{x_0,x_1}(\widetilde{y} \mid x_0) = 0.$$

Task 3: Fair Decision-Making



Constructed prediction

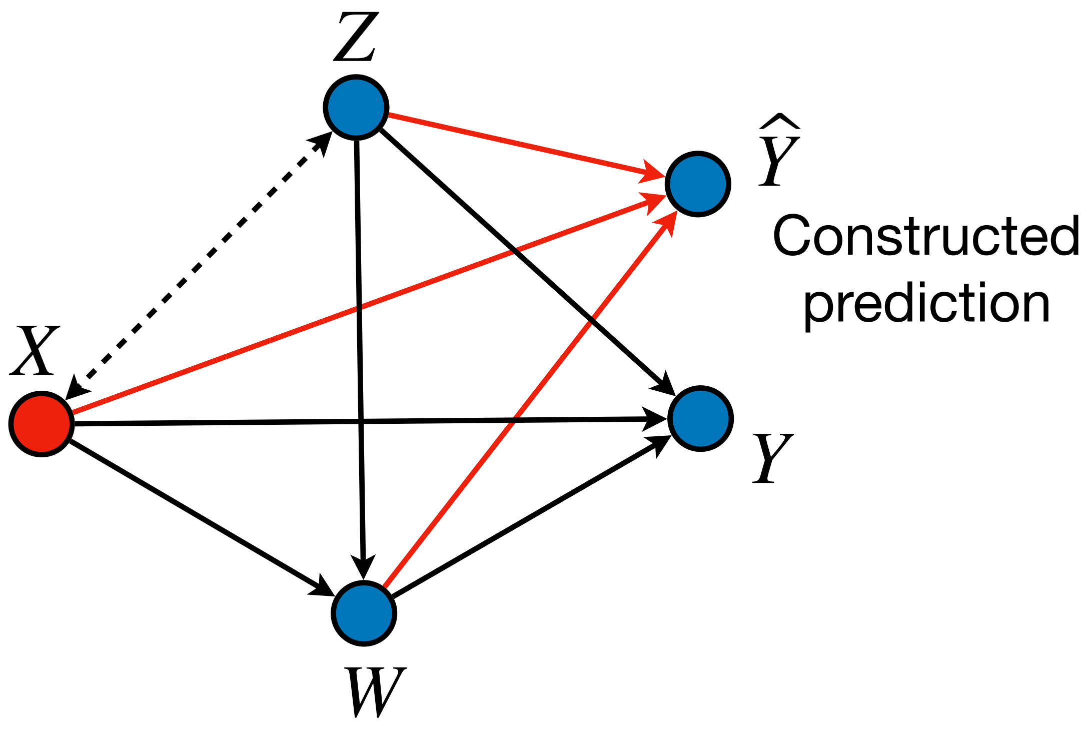


Downstream effects

Decision

College Admissions Example

Example (T2 vs. T3). A university is deciding on admissions of prospective applicants. The information available to the selection committee is the following —gender (X), the SAT scores (W), the predicted GPA score if enrolled ($\widehat{Y}$). Based on this information, the university needs to make a decision D on whether to admit the applicant.



T2

The aim of Task 2 is to produce predictions $\widehat{Y}$, which, for example, contain no direct effect of gender X, that is,

$$\text{NDE}_{x_0,x_1}(\widehat{y}) = \text{NDE}_{x_1,x_0}(\widehat{y}) = 0.$$

T3

The aim of Task 3 is to decide which of the applicants get admitted based on some utility function U. Different types of utility can be considered, e.g., the total expected income of the university coming from tuition fees; university reputation; minority representation.

Fair Decisions from Fair Predictions?

- A first possible idea might be to leverage fair predictions to construct fair decisions.

Proposition. Let μ be a fairness measure defined by a contrast C of the form (C_1, C_0, E_1, E_0). Suppose that a predictor $\widehat{Y}$ is fair w.r.t. μ, that is, $\mu(\widehat{y}) = 0$. Suppose that a decision policy D is constructed simply as a transformation of $\widehat{Y}$, i.e.,

$$D := f_D(\widehat{Y}) \, .$$

Then, we can say that D is fair with respect to μ if:

(a) function f_D is linear,
(b) measure μ is a unit level measure.

Chaining does not work in general.

- This suggests that transforming $\widehat{Y}$ does not *always* work in practice.

Failure of Thresholding Policies

Example (Thresholding). A university is deciding on admissions of prospective applicants. The information available to the selection committee is the following. Let X denote race (x_0 for minority groups, $x_1$for majority group), W_1 the SAT score, and W_2 the student's score on the admission exam. The predicted GPA of the student $\widehat{Y}$ is a function of W_2, and the final admission decision is denoted by D. Suppose the following (unknown) SCM describes this setting situation:

$$X \leftarrow U_X$$

$$W_1 \leftarrow U_W + X$$

$$W_2 \leftarrow \begin{cases} W_1 + 2(1 - X) \text{ if } W_1 > 0.5, \\ W_1 \text{ otherwise}. \end{cases}$$

$$\widehat{Y} \leftarrow W_2 + 2$$

$$D \leftarrow 1(\widehat{Y} > 3.75)$$

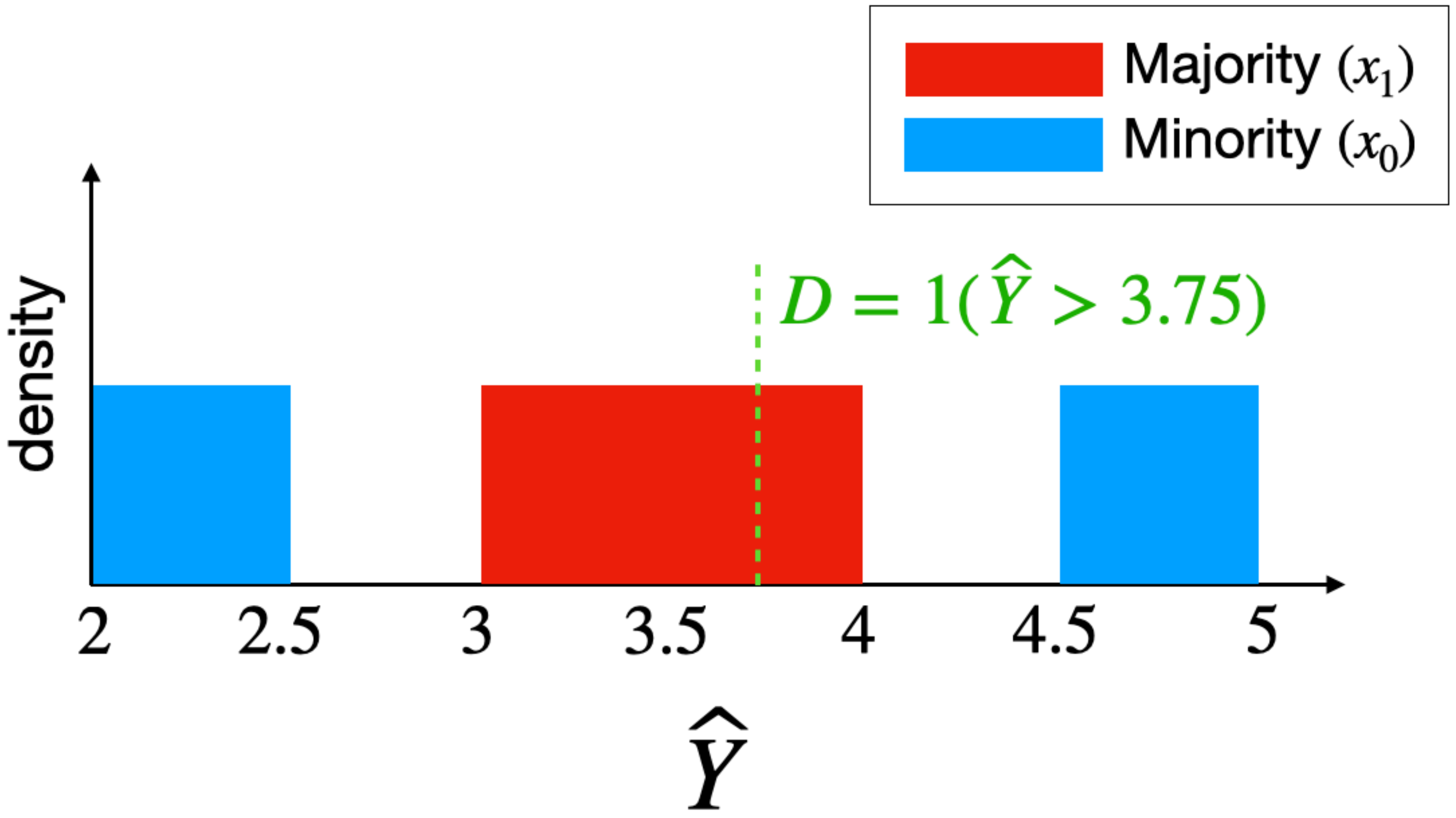
$$U_X \sim \text{Bernoulli}(0.8), U_W \sim \text{Unif}[0,1].$$

Can compute:

$$\text{NIE}_{x_1,x_0}(\widehat{y}) = 0.$$

$$\text{NIE}_{x_1,x_0}(d) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Failure of Thresholding



Utility Functions & Examples

- Different types of utility can be considered when performing Task 3 in practice

Utility Type	Mathematical Representation
University Reputation	$\lambda_1 W_1$, where W_1 is SAT score
Income From Tuition Fees	$\lambda_2 W_2$, where W_2 is SE status
Minority Representation	$\lambda_3 1(X = x_0)$, with x_0 protected

Decision-Making as in-processing

Theorem. Let the decision policy D be constructed as the optimal solution to

$$D = \textbf{argmin}_d \quad E[U(d; X, Z, W, Y)]$$

$$\begin{aligned} \textbf{subject to} \quad & x\text{-}\textbf{DE}^{\textbf{ID}}_{x_0,x_1}(d \mid x_0) = 0 \\ & x\text{-}\textbf{DE}^{\textbf{ID}}_{x_1,x_0}(d \mid x_0) = 0 \\ & x\text{-}\textbf{IE}^{\textbf{ID}}_{x_0,x_1}(d \mid x_0) = 0 \\ & x\text{-}\textbf{IE}^{\textbf{ID}}_{x_1,x_0}(d \mid x_0) = 0 \\ & x\text{-}\textbf{SE}^{\textbf{ID}}_{x_1,x_0}(d) = 0 \end{aligned}$$

Section 5.3
Theorem 5.4

Then D satisfies

$$x\text{-DE}^{\text{sym}}_x(d \mid x_0) = x\text{-IE}^{\text{sym}}_x(d \mid x_0) = x\text{-SE}_{x_1,x_0}(d) = 0.$$

Reward + Diversity Utility

- A popular class of utility functions takes the form:

$$U(D; X, Z, W, Y) = \boxed{R(D, Y)} + \boxed{\lambda 1(X = x_0)D}.$$

Reward term measuring individual's qualifications

Incentivising disadvantaged groups

- The λ parameter interpolates between the reward-only ($\lambda = 0$) and minority-representation only ($\lambda = \infty$) solutions
- (Nilforoshan et. al., 2022)'s result: the policy $d^{\lambda\text{-opt}}$ that is optimal for a specific λ value will with large probability does not satisfy any causal notion of fairness!
- Does the result sound familiar?

Causal Conceptions of Fairness as FPT Corollary

Theorem. Let the utility function be given as

$$U(D; X, Z, W, Y) = -[Y - D(X, Z, W)]^2 + \lambda 1(X = x_0)D \, .$$

Let Dmax denote the maximum utility policy that solves the problem

The result can be derived from the Fair Prediction Theorem!

$$x\text{-SE}_{x_0,x_1}(d) = 0.$$

Then, it is the case that

$$\exists \epsilon(n_Z, n_W) > 0 \text{ s.t. } P(U(D^{\text{max}}) - U(D^{\text{CF}}) > \epsilon(n_Z, n_W)) \geq \frac{3}{4} \, .$$

How to interpret Causal Conceptions?

Example (College Admissions: Who is who?). A university is deciding on admissions of prospective applicants. The information available to the selection committee is the following. Let X denote race (x_0 for minority groups, x_1 for majority group), W denotes the SAT score, Z denotes the socio-economic status of the family of the student ($Z = 0$ for poor, $Z = 1$ for rich). Let D be the decision whether to admit an applicant. Suppose that the following SCM describes the situation:

$$
\begin{aligned}
X &\leftarrow U_X \\
Z &\leftarrow U_Z \\
W &\leftarrow U_W - 5(1 - Z)(1 - X) \\
D &\leftarrow f_D(X, W),
\end{aligned}
$$

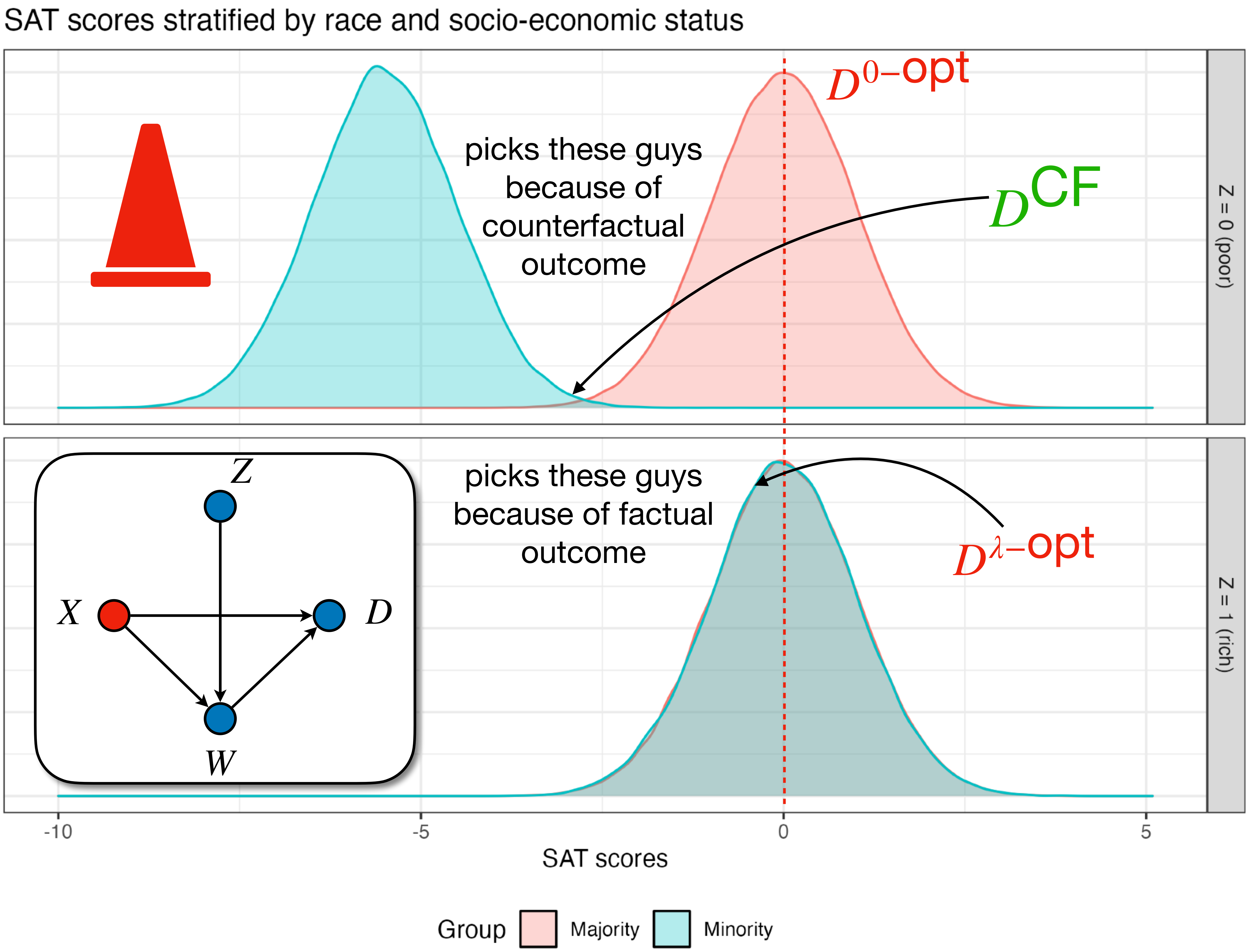
$$
\begin{aligned}
U_X &\sim \text{Bernoulli}(0.8), \\
U_Z &\sim \text{Bernoulli}(0.3), \\
U_W &\sim \text{Unif}[0,1] .
\end{aligned}
$$

How to interpret Causal Conceptions?

SCM Interpretation: 80% of applicants from the majority group, and 10% of the applicants come from privileged socio-economic background. Applicants from low income minority families have lower SAT scores on average, compared to their majority group counterparts. For the high-income families, there is no difference in the SAT scores between the majority and minority groups.

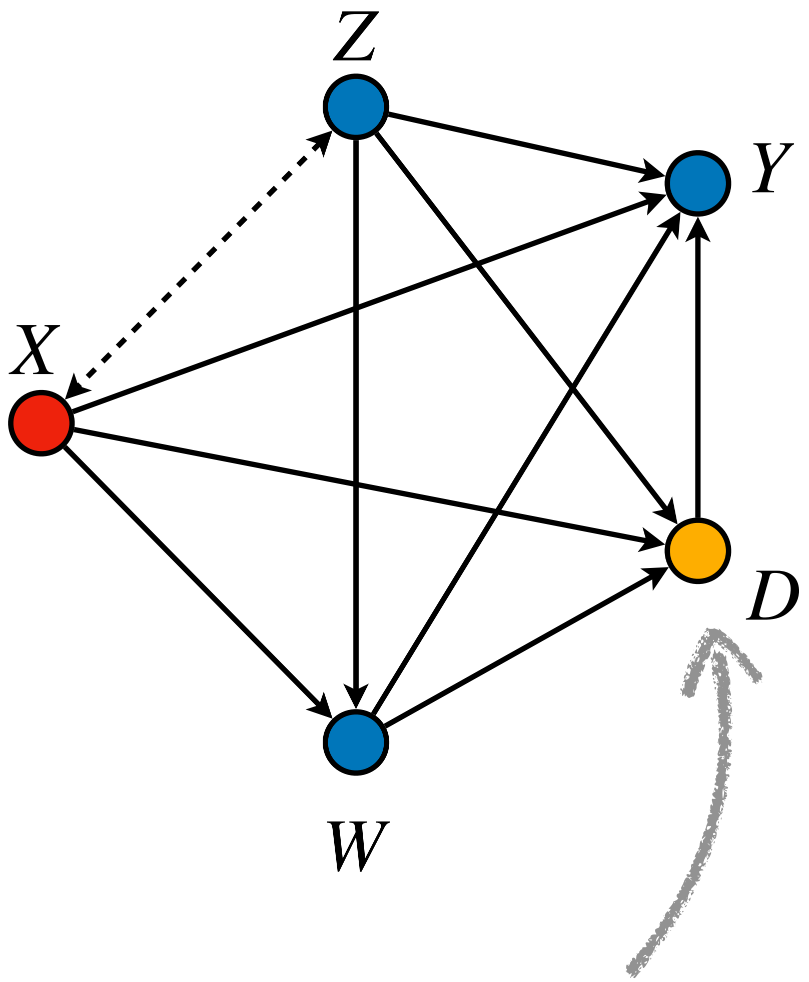
$$
\begin{aligned}
X &\leftarrow U_X \\
Z &\leftarrow U_Z \\
W &\leftarrow U_W - 5(1 - Z)(1 - X) \\
D &\leftarrow f_D(X, W),
\end{aligned}
$$

$$
\begin{aligned}
U_X &\sim \text{Bernoulli}(0.8), \\
U_Z &\sim \text{Bernoulli}(0.3), \\
U_W &\sim \text{Unif}[0,1] .
\end{aligned}
$$

SAT scores stratified by race and socio-economic status
D^{0}–opt
D^{CF}
picks these guys because of counterfactual outcome
picks these guys because of factual outcome
D^{λ}–opt
Z = 0 (poor)
Z = 1 (rich)
Z
X
D
W
-10
-5
0
5
SAT scores
Group
Majority
Minority

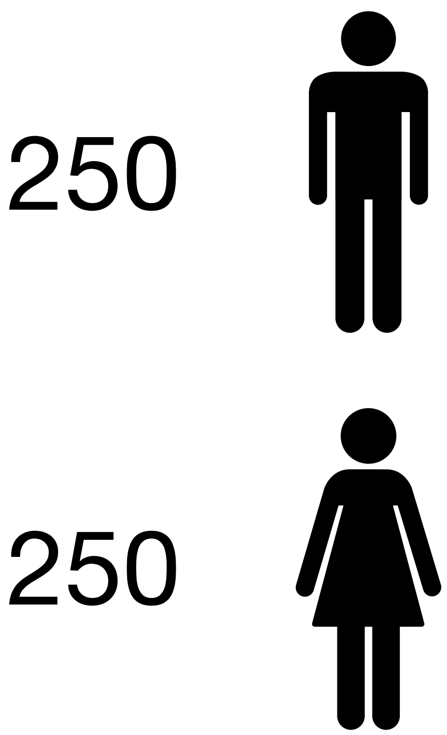
Task 3: Outcome Control

- There is an interesting specific setting of Task 3 (Decision-Making) in which the decision variable D possibly influences the outcome of interest Y we are trying to optimize.
- We refer to this setting as *Outcome Control.*
- Examples range from criminal justice (Y recidivism, D detention) to medical applications (Y survival, D surgery).
- How can we conceptualize fairness in such instances and leverage the previously developed tools?



Discretion of the decision-maker

Example (Judge and Oracle). A district court judge has to make decisions about whether to detain or release individuals that have been charged with a similar offense ($D = 0$ for detaining, $D = 1$ for releasing). The judge assesses a total of 500 individuals, half of whom are female, but has limited resources and can detain at most 100 of them. The judge's objective is to minimize the number of people who will re-offend ($Y = 0$ for re-offending, $Y = 1$ for not re-offending). However, the judge has access to an oracle that knows the potential outcomes Y_{d_0}, Y_{d_1} for every individual. Who does the judge detain?

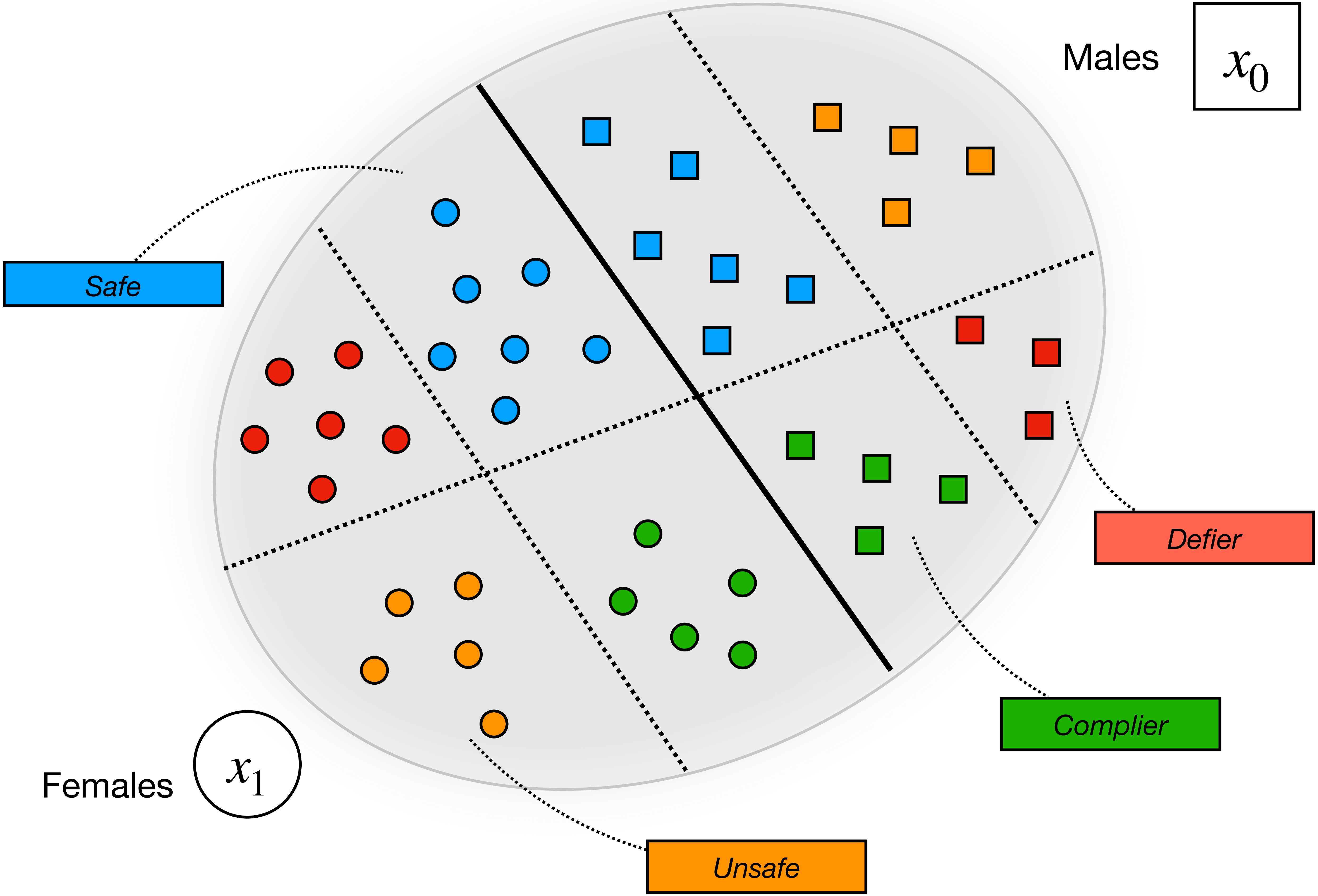


Potential outcomes (Y_{d_0}, Y_{d_1}) tell us:

- ✗ $(0{,}0)$ those who always re-offend,
- ✗ $(1{,}1)$ those who never re-offend,
- ✗ $(1{,}0)$ those who re-offend only if detained,
- ✓ $(0{,}1)$ those who re-offend only if released .

Pick 50 of each!

Units $\mathscr{U}$



Males x_0

Safe

Defier

Complier

Females x_1

Unsafe

Principal Fairness

- The appealing intuitive reasoning motivates the definition of *principal fairness (Imai & Jiang, 2020):*

$$P(d \mid y_{d_0}, y_{d_1}, x_1) = P(d \mid y_{d_0}, y_{d_1}, x_0)$$

Warning:
Joint distribution over counterfactual outcomes is notoriously difficult to get!

Imai & Jiang solution: Monotonicity

Assumption: $Y_{d_1}(u) \geq Y_{d_0}(u)$

Defiers!
(Think medical)

Decision-Maker’s Perspective

- So far, we considered perfect knowledge, which allows perfect utility!

Example (Cancer Surgery). Clinicians at CUMC need to decide which cancer patients should undergo a cancer surgery in order to improve survival Y. They have information on sex X and illness severity W determined from tissue biopsy. The SCM is given by:

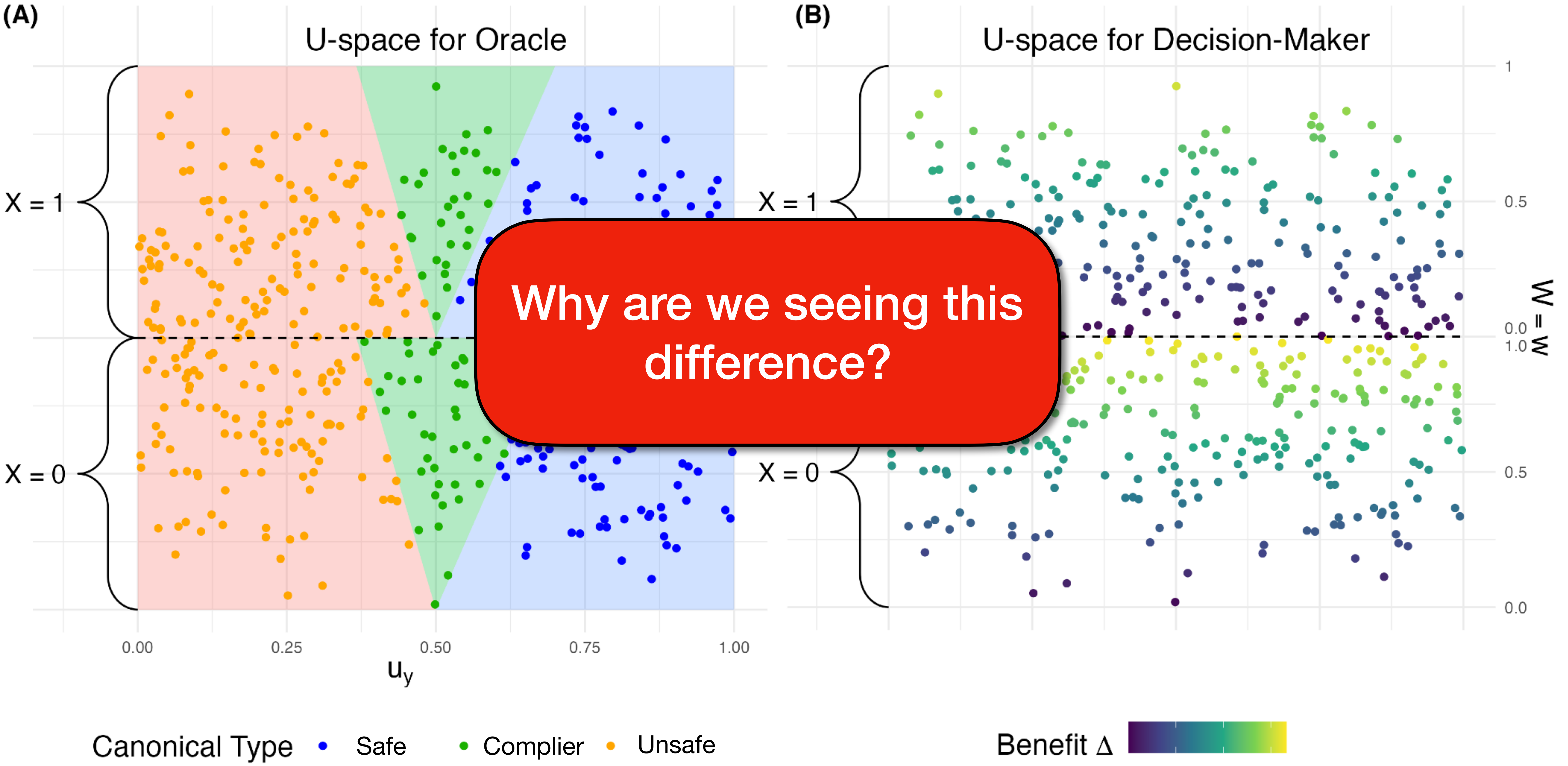
$$\begin{aligned} X &\leftarrow U_X \\ W &\leftarrow X + (-1)^X \sqrt{U_W} \\ D &\leftarrow f_D(X, W) \\ Y &\leftarrow \mathbb{1}(U_Y + \frac{1}{3}WD - \frac{1}{5}W > 0.5)\,. \end{aligned}$$

$$U_X \in \{0,1\}, P(U_X = 1) = 0.5,$$
$$U_W, U_Y \sim \text{Unif}[0,1],$$

The clinicians compute:

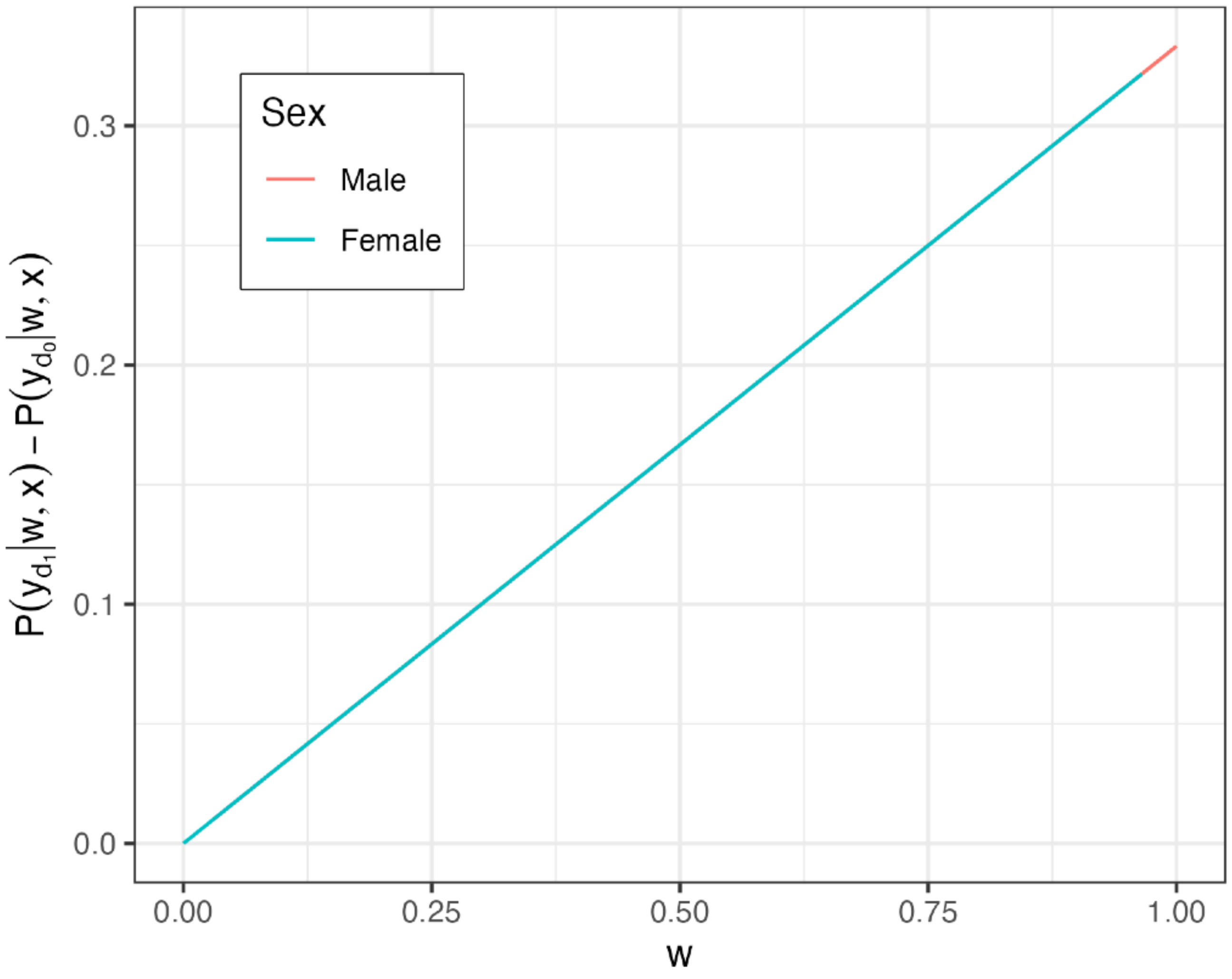
$$P(y_{d_1} = 1 \mid w, x_1) - P(y_{d_0} = 1 \mid w, x_1) = \frac{w}{3}$$
$$P(y_{d_1} = 1 \mid w, x_0) - P(y_{d_0} = 1 \mid w, x_0) = \frac{w}{3}\,.$$

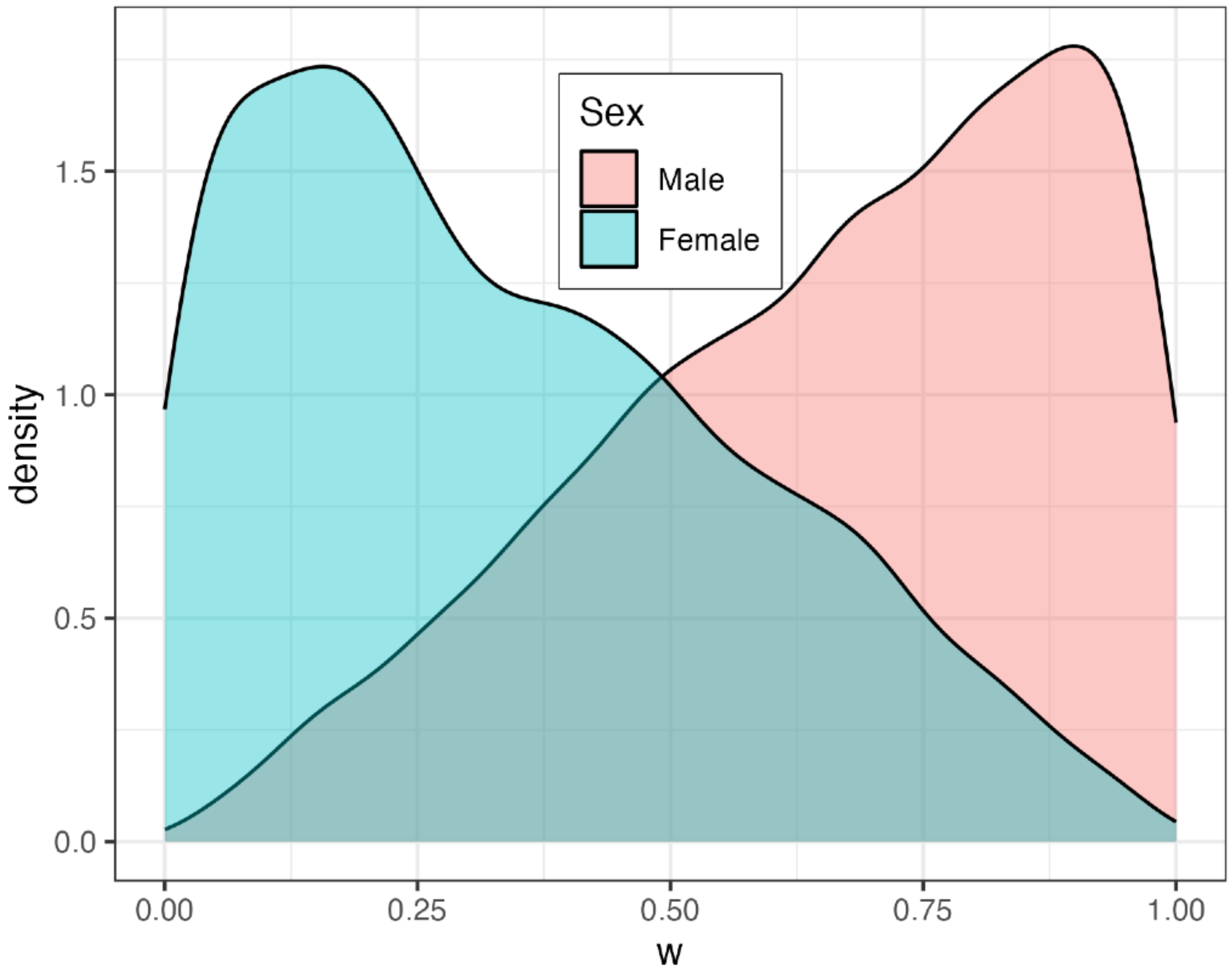
Oracle vs. Decision-Maker



Why are we seeing this difference?

Oracle vs. Decision-Maker: Intuition





Benefit Fairness

Definition. Define the degree of benefit Δ as:

$$\Delta(x, z, w) = P(y_{d_1} \mid x, z, w) - P(y_{d_0} \mid x, z, w)\,.$$

We say that the pair (Y, D) satisfies benefit fairness (BF, for short) if

$$P(d \mid \Delta = \delta, x_0) = P(d \mid \Delta = \delta, x_1) \quad \forall \delta\,.$$

$$P(\text{decision} \mid \Delta, \text{🚹}) = P(\text{decision} \mid \Delta, \text{🚺})$$

Canonical Types and Guarantees

- Principal Strata types “Safe”, “Unsafe”, “Complier”, “Defier” are sometimes referred to as canonical types.

- Can we provide results on the benefit Δ using the canonical types?

Proposition A. Let (s, d, c, u) denote the proportions of canonical types for a set of covariates (x, z, w). It then follows that

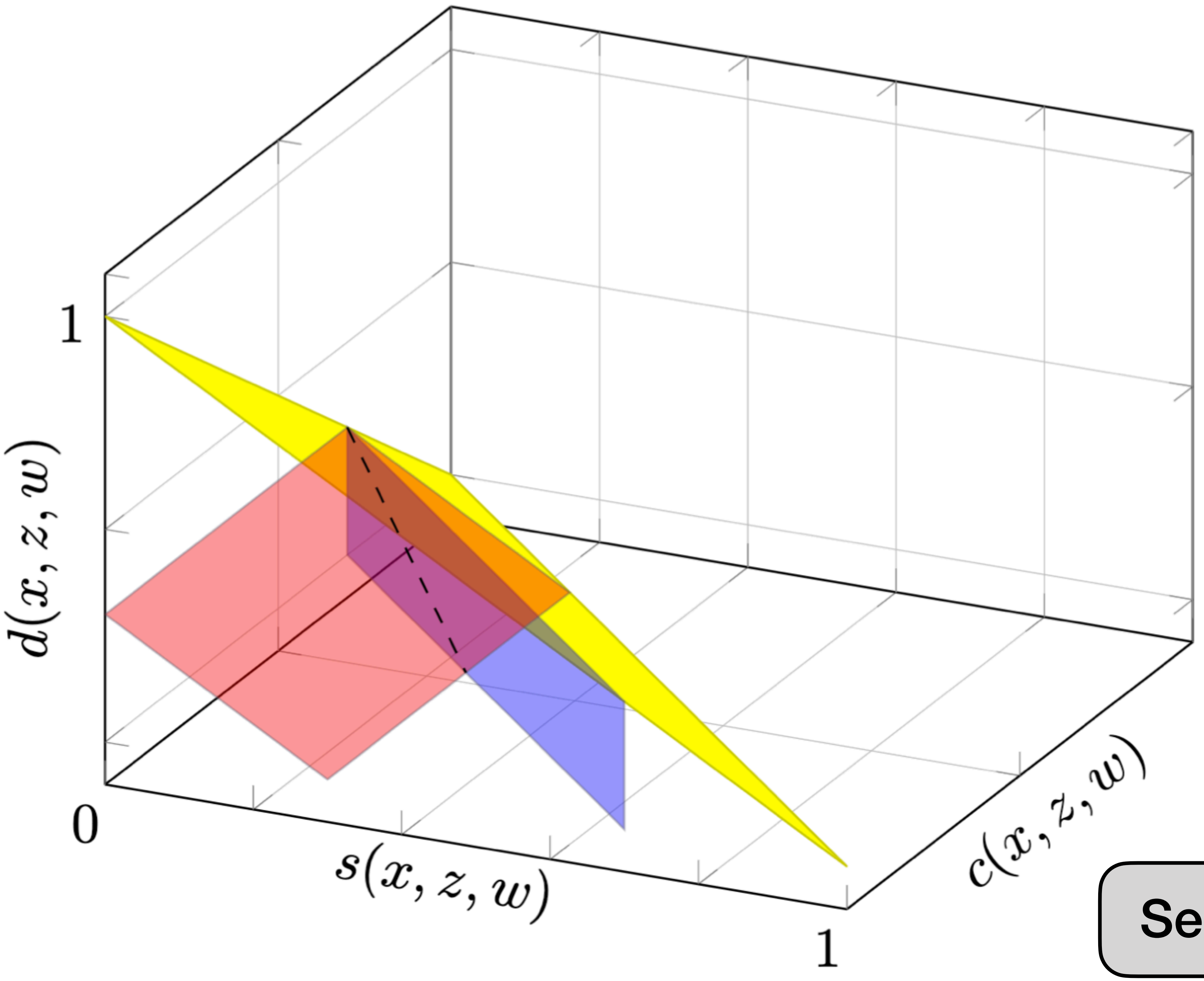
$$\begin{aligned}\Delta(x, z, w) &:= P(y_{d_1} \mid x, z, w) - P(y_{d_0} \mid x, z, w)\\ &= c(x, z, w) - d(x, z, w)\,.\end{aligned}$$

Proposition B. Let $m_1(x, z, w) = P(y_{d_1} \mid x, z, w)$ and $m_0(x, z, w) = P(y_{d_0} \mid x, z, w)$. It then follows that if $m_1 \geq m_0$ that

$$\begin{aligned}d &\in [0, \min(m_0, 1 - m_1)],\\ c &\in [m_1 - m_0, m_1]\,.\end{aligned}$$

In particular, the above bounds are tight.

Canonical Types and Guarantees



Section 5.3.3

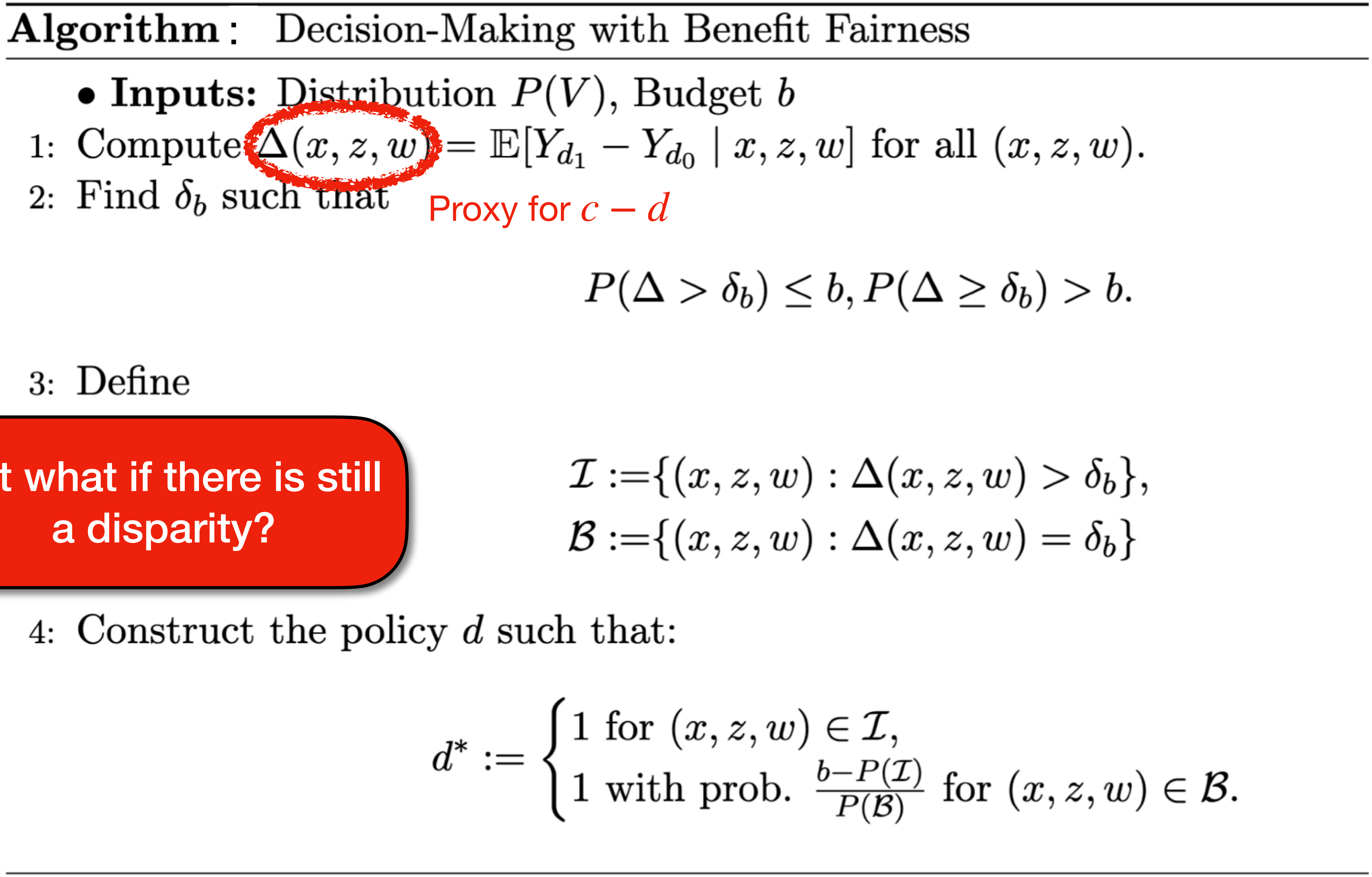
Decision-Making Problem

Definition. The optimal decision-making problem is defined as finding the solution to the following optimization problem, given a fixed budget b:

$$\begin{aligned} d^* = \text{argmax}_d \qquad & E[Y_d] \\ \text{subject to} \qquad & P(d) \leq b\,. \end{aligned}$$

$$\left.\begin{aligned} Y_d &= (1-D)Y_{d_0} + DY_{d_1} \\ Y_{d_1} &= 1(\text{safe}) + 1(\text{complier}) \\ Y_{d_0} &= 1(\text{safe}) + 1(\text{defier}) \end{aligned}\right\} \implies E[Y_d] = P(\text{safe}) + \boxed{E[D1(\text{complier})]}$$

Algorithmic approach for Benefit Fairness

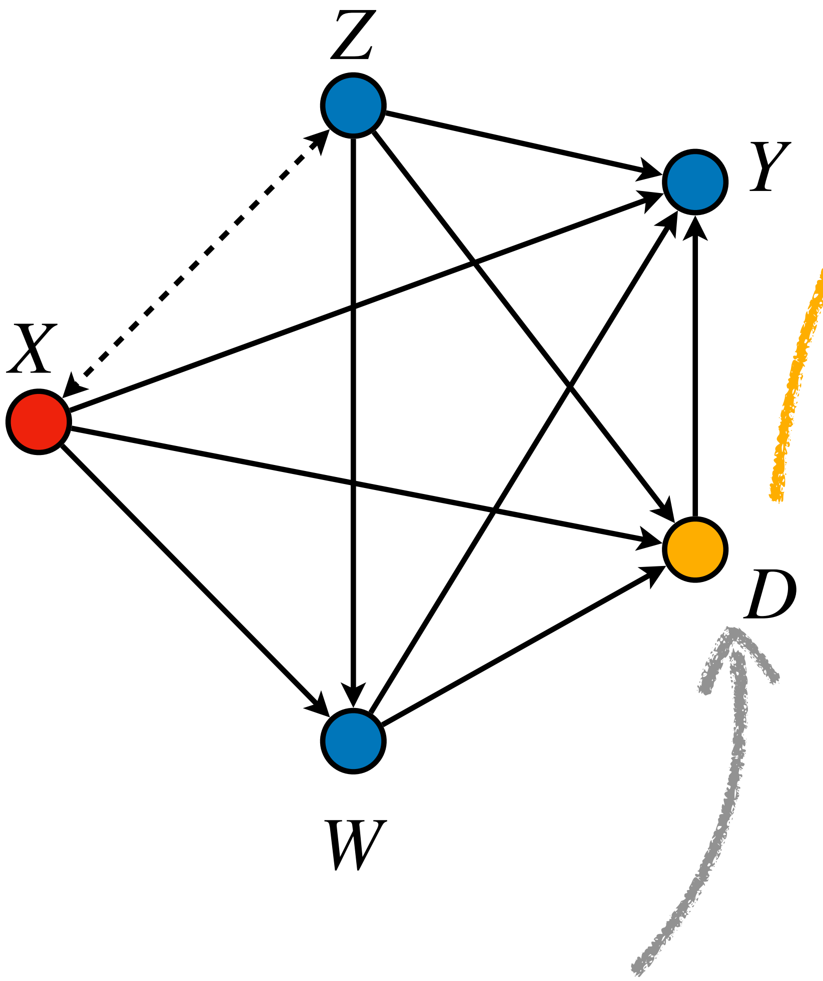


```
Algorithm: Decision-Making with Benefit Fairness
    • Inputs: Distribution $P(V)$, Budget $b$
1: Compute $\Delta(x, z, w) = \mathbb{E}[Y_{d_1} - Y_{d_0} \mid x, z, w]$ for all $(x, z, w)$.
2: Find $\delta_b$ such that
       $P(\Delta > \delta_b) \leq b, P(\Delta \geq \delta_b) > b.$
3: Define
       $\mathcal{I} := \{(x, z, w) : \Delta(x, z, w) > \delta_b\},$
       $\mathcal{B} := \{(x, z, w) : \Delta(x, z, w) = \delta_b\}$
4: Construct the policy $d$ such that:
       $d^* := \begin{cases} 1 \text{ for } (x, z, w) \in \mathcal{I}, \\ 1 \text{ with prob. } \frac{b - P(\mathcal{I})}{P(\mathcal{B})} \text{ for } (x, z, w) \in \mathcal{B}. \end{cases}$
```

Proxy for $c - d$

But what if there is still a disparity?

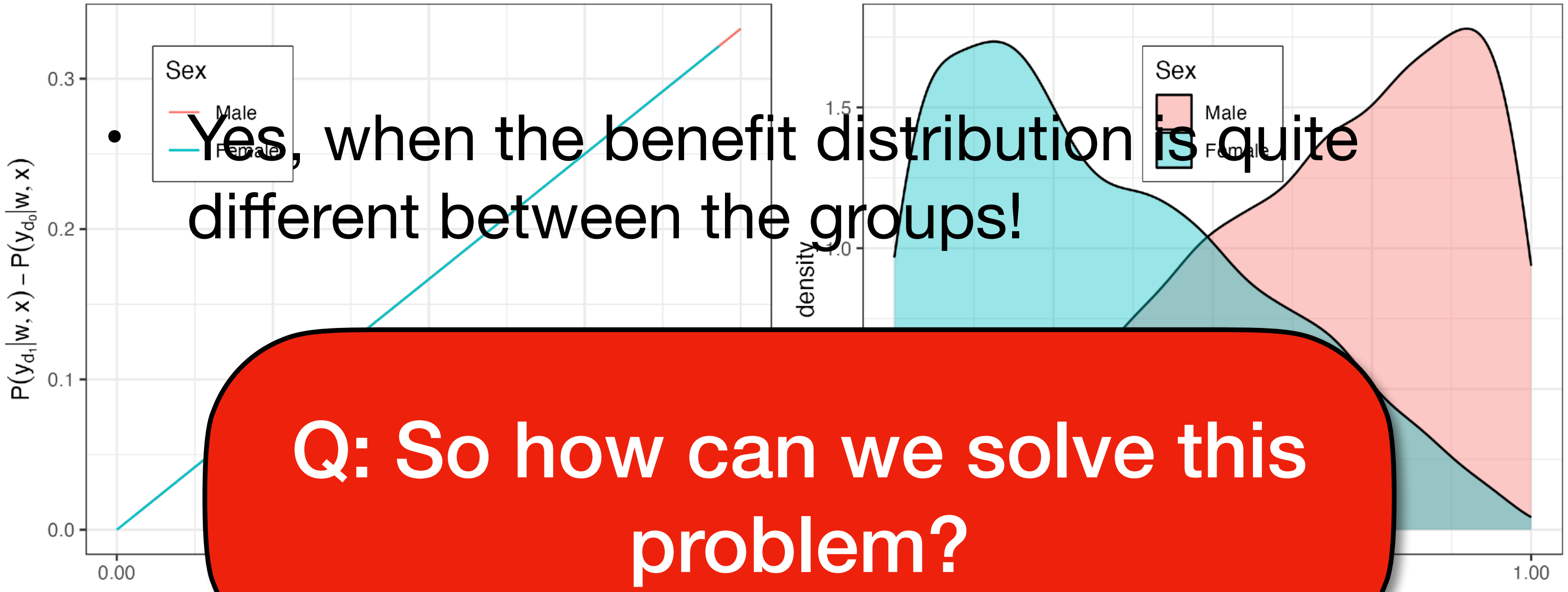
Is there still a gap?



Discretion of the decision-maker

- With benefit fairness, we make sure that both x_0 and x_1 groups are treated equally at any fixed level of the benefit Δ.

- Can this still result in a large gap between groups in terms of the allocation of resources?



- Yes, when the benefit distribution is quite different between the groups!

Q: So how can we solve this problem?

Controlling the Gap

1) Use the Decision-Making with Benefit Fairness Algorithm to construct a policy d

2) Decompose the total variation induced by the decision policy d based on the Fairness Map

$$P(d \mid x_1) - P(d \mid x_0) = \text{DE} + \text{IE} + \text{SE}\,.$$

3) Based on expert knowledge, determine if

3a) There is causal unfairness in the benefit through a contrast $C = (C_0, C_1)$

Use Causal Benefit Fairness:

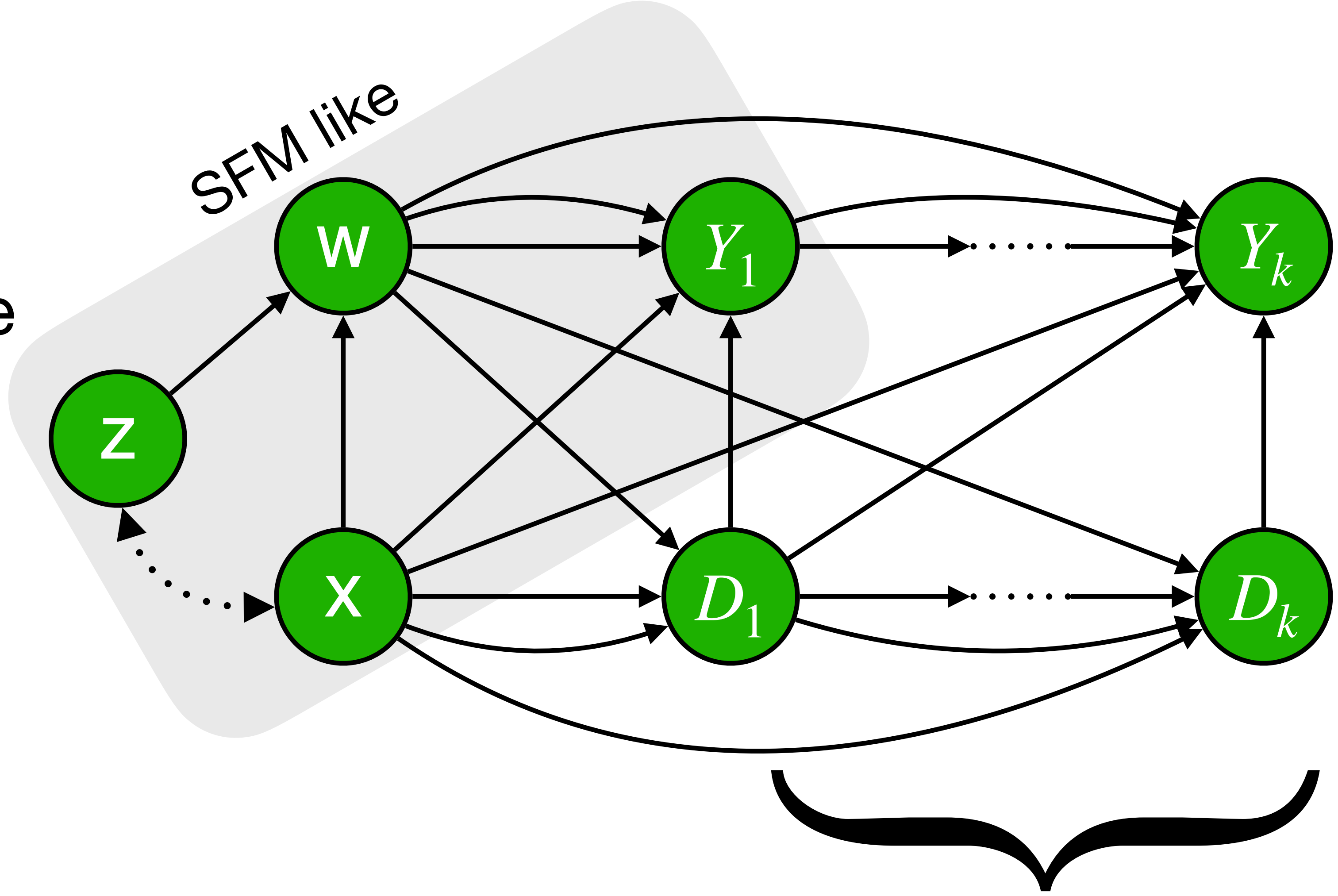
$$\begin{aligned} E(y_{C_1,d_1} \mid x, z, w) &= E(y_{C_0,d_1} \mid x, z, w)\ \forall x, z, w \\ E(y_{C_1,d_0} \mid x, z, w) &= E(y_{C_0,d_0} \mid x, z, w)\ \forall x, z, w \\ P(d \mid \Delta, x_0) &= P(d \mid \Delta, x_1)\,. \end{aligned}$$

3b) There is a “purely distributive” need for reducing the total variation

-> Repeat Step 1) by using different thresholds $\delta_b^{(x_0)}, \delta_b^{(x_1)}$ for the x_0, x_1 groups

Task 3: Long-term effects

- We discussed Task 3 with a specific utility function -> we call this the 0-step DM
- We discussed Task 3 with a specific outcome to be controlled -> we call this the 1-step DM
- Naturally, it is also possible to think about a k-step DM
- We leave this challenge for future work -> very exciting problem, related to RL!



decisions taken over time