Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint

Yoichi Chikahara[1,3], Shinsaku Sakaue[2], Akinori Fujino[1], Hisashi Kashima[3]

[1] NTT, [2] The University of Tokyo, [3] Kyoto University

Motivation:
Machine learning (ML) for fair decision-making

- ML is increasingly used to make decisions for individuals

<u>**Application examples**</u>:
loan approval, job hiring, child abuse screening, and recidivism prediction

- Due to their huge societal impact on people's lives, these ML predictions should be **accurate** and **fair with respect to sensitive features** (e.g., gender, race, and sexual orientation)

Our approach:
Use causal graph to make accurate and fair predictions

Problem statement:
Learning fair binary classifier using causal graph

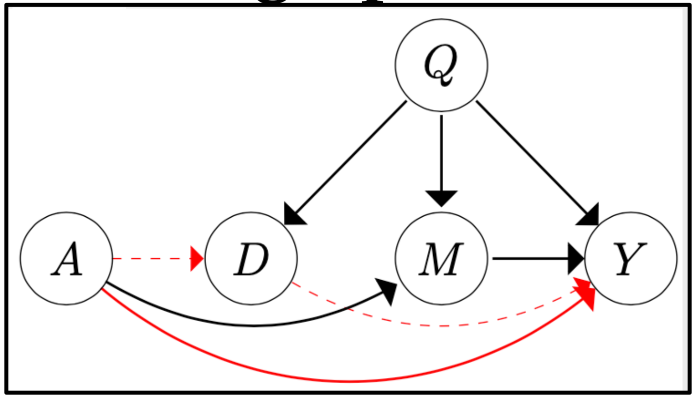
Input

Training data

A Sensitive	Q	D	M	Y
Female	B	0	B	Accept
Male	A	1	B	Reject
Female	C	0	D	Reject
Male	C	2	C	Reject

$\boldsymbol{X} = \{A, Q, D, M\}$: Features of each individual

Causal graph



(Given by experts or estimated from data)

Minimize
loss L_θ + penalty on unfairness G_θ

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

Output

Fair classifier

$h_{\hat{\theta}}(\boldsymbol{X})$

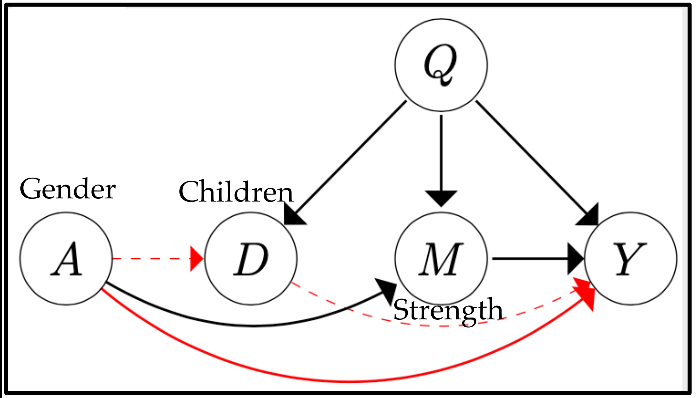
Causal graph allows us to design G_θ so that we can avoid imposing unnecessary fairness constraints.

Problem statement:
Using causal graph to express *what is unfair*

Causal graph can express our complex prior knowledge on discrimination in real-world scenarios

Motivating example

Hiring decisions for physically-demanding jobs



Following reasons for rejection is **unfair**:

1. female ($A \to Y$)
2. female, has no child ($A \to D \to Y$)

while following is **fair**:

3. female, has little physical strength ($A \to M \to Y$)

To formulate G_θ based on unfair pathways $\pi = \{A \to Y, A \to D \to Y\}$, we measure the unfairness as **path-specific causal effects (PSEs)**

Weaknesses of existing methods:
Needs strong assumptions or not individually fair

Existing methods **cannot achieve individual-level fairness** or **require restrictive functional assumptions on data**

Table 1: Comparison with existing methods

Method	Individually fair	Functional assumptions
Our method	Yes	Unnecessary
PSCF	Yes	Necessary
FIO	No	Unnecessary

A classifier achieves **(path-specific) individual-level fairness** if the following holds for any input feature value $\boldsymbol{x}$:

$$\mathbb{E}_{Y_{A \Leftarrow 0}, Y_{A \Leftarrow 1 \| \pi}} [\underline{Y_{A \Leftarrow 1 \| \pi} - Y_{A \Leftarrow 0}} | \boldsymbol{X} = \boldsymbol{x}] = 0$$ [Wu+; NeurIPS2019]

PSE [Avin+; IJCAI2005]: difference of two predictions (i.e., $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$), obtained by modifying feature attributes $\boldsymbol{x}$ to those of *counterfactual individuals*

How can we learn individually fair classifier without restrictive functional assumptions?

Proposed method:
Use upper bound on PIU for penalization

- To achieve individual-level fairness, we force probability of individual unfairness (PIU) to be zero, whose upper bound can be derived as

$$\underbrace{\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})}_{\text{PIU}} \leq 2\,\underbrace{\mathrm{P}^{I}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})}_{\text{upper bound on PIU}}$$

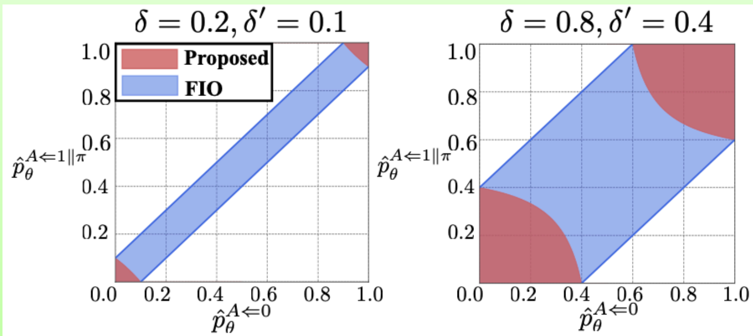
$\mathrm{P}^{I}(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = \mathrm{P}(Y_{A\Leftarrow 0})\,\mathrm{P}(Y_{A\Leftarrow 1\|\pi})$
is an *independent joint distribution*, which can be inferred **from data without any restrictive functional assumptions**

- To make the upper bound value close to zero, we use the estimator of $\mathrm{P}^{I}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$ as penalty; i.e.,

$$G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \hat{p}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A\Leftarrow 0}) + (1 - \hat{p}_\theta^{A\Leftarrow 1\|\pi})\hat{p}_\theta^{A\Leftarrow 0}$$

More details?
Check out our poster!

Why does penalty on upper bound guarantee individual-level fairness?



Can we deal with latent confounders?

$$G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \hat{u}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{l}_\theta^{A\Leftarrow 0}) + (1 - \hat{l}_\theta^{A\Leftarrow 1\|\pi})\hat{u}_\theta^{A\Leftarrow 0}$$

Experimental results?

Table 2: Test accuracy (%) on each dataset

Method	Synth	German	Adult
Proposed	80.0 ± 0.9	75.0	75.2
FIO	84.8 ± 0.6	78.0	81.2
PSCF	74.8 ± 1.6	76.0	73.4
Unconstrained	88.2 ± 0.9	81.0	83.2
Remove	76.9 ± 1.3	73.0	74.7