PC-Fairness: A Unified Framework for Measuring Causality-based Fairness

TL;DR

- We define a unified fairness notion that covers most previous causality-based fairness notions, namely **Path-specific Counterfactual fairness (PC fairness)**.
- We develop a constrained optimization-based approach for tightly bounding any unidentifiable PC fairness.
- The experiments show the creativeness and effectiveness of the proposed method under any unidentifiable situation or combinations.

Backgrounds for Causal Inference

Structural Causal Model (SEM) and Causal Graph

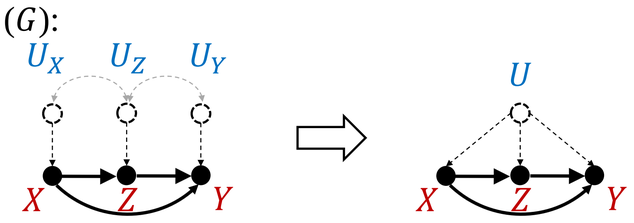
❖ Notations
- Exogenous Variables $\boldsymbol{U} = \{U_X, U_Y, U_Z\}$, and Distribution $P(\boldsymbol{U})$
- Endogenous Variables $\boldsymbol{V} = \{X, Y, Z\}$
- Functions $\boldsymbol{F} = \{f_X, f_Y, f_Z\}$

❖ Representations

SEM (M):

$$X = f_X(u_X)$$
$$Z = f_Z(x, u_Z)$$
$$Y = f_Y(x, z, u_Y)$$

Graph (G):



Causal Effects

❖ Causal effects are responses of the outcome of interest to an external intervention.

❖ Intervention is facilitated with do-operator.
- Intervention $do(x)$ is defined as the substitution of structural equation $X = f_X(u_X)$ with $X = x$.
- The response of Y after intervention $do(x)$ is denoted by Y_x.

❖ **Total Causal Effect** of $X: x_0 \to x_1$ on $Y = y$:

$$TCE(s^-, s^+) = P(y_{s^+}) - P(y_{s^-})$$

❖ **Path-specific Effect** of $X: x_0 \to x_1$ on $Y = y$ through a path set $\pi = \{X \to Z \to Y\}$:

$$PE(x_1, x_0) = P\left(y_{x_1|\pi, x_0|\bar{\pi}}\right) - P(y_{x_0})$$

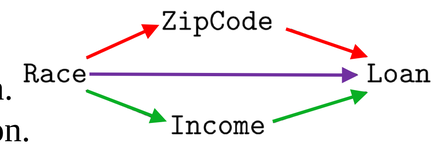
❖ **Counterfactual Effect** of $X: x_0 \to x_1$ on $Y = y$ for a condition $X = x', \boldsymbol{O} = \boldsymbol{o}'$:

$$CE(x_1, x_0) = P\left(y_{x_1}|x', \boldsymbol{o}'\right) - P(y_{x_0}|x', \boldsymbol{o}')$$

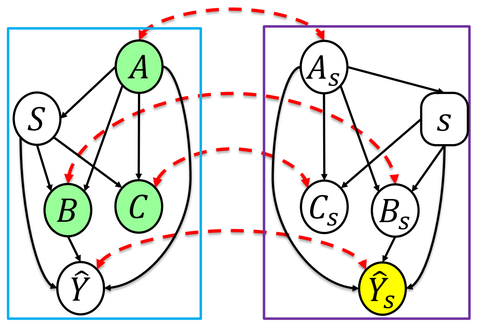
Backgrounds for Causal Fairness Learning

Path-specific Fairness and Counterfactual Fairness

❖ Three types of path-specific effects:
- The green-path effect is considered as explainable.
- The red-path is considered as indirect discrimination.
- The purple-path is considered as direct discrimination.



❖ Counterfactual Fairness: the total effect for a group specified by the factual observation.
- The blue box denotes the real/factual world.
- The purple box denotes the counterfactual world.
- The condition $\boldsymbol{O} = \{A, B, C\}$ (in green).
- The counterfactual quantity is $\mathrm{P}(\hat{Y}_s)$ (in yellow).



Path-specific Counterfactual Fairness

Path-specific Counterfactual Effect

❖ Path-specific Counterfactual Effect of $X: x_0 \to x_1$ on $Y = y$ for a given condition $\boldsymbol{O} = \boldsymbol{o}$ through a pathset π is

$$PCE_\pi(x_1, x_0|\boldsymbol{o}) = P\left(y_{x_1|\pi, x_0|\bar{\pi}}|\boldsymbol{o}\right) - P(y_{x_0}|\boldsymbol{o})$$

Path-specific Counterfactual Fairness (PC Fairness)

❖ Given a predictor $\hat{Y}$, a condition $\boldsymbol{o}$, a path set π,

1) $\hat{Y}$ achieves PC fairness if $PCE_\pi(s_1, s_0|\boldsymbol{o}) = 0$;

2) $\hat{Y}$ achieves τ-PC fairness if $|PCE_\pi(s_1, s_0|\boldsymbol{o})| \leq \tau$.

where s_0, s_1 are two values of the sensitive attribute.

❖ Thanks to the flexibility and capability of the path-specific counterfactual effect, the PC fairness generalize the previous causality-based fairness notions, as summarized in Table 1.

Table 1: Connection between previous fairness notions and PC fairness

Description	References	Relating to PC fairness
Total causal effect	[10, 7]	$\mathbf{O} = \emptyset$ and $\pi = \Pi$
(System) Direct discrimination	[10, 3, 7]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
(System) Indirect discrimination	[10, 3, 7]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_i \subset \Pi$
Individual direct discrimination	[8]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Path-specific Individual discrimination	[1]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi \in \Pi$
Group direct discrimination	[9]	$\mathbf{O} = \mathbf{Q} = \mathrm{PA}_Y \backslash \{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Counterfactual fairness	[2, 4, 5]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \Pi$
Counterfactual error rate	[6]	$\mathbf{O} = \{S, Y\}$ and $\pi = \pi_d$ or π_i

Note 1: please refer to our paper for the references.

Measuring Path-specific Counterfactual Fairness

Note 2: please refer to Section 4 for more technical details.

❖ The key barrier in PC fairness is the unidentifiable issues. Examples of typical unidentifiable situations are shown below

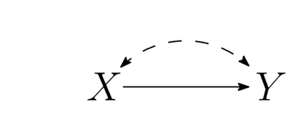


Figure 2: The "bow graph".

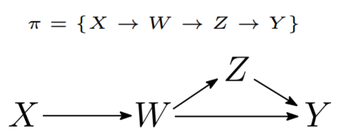


Figure 3: The "kite graph".

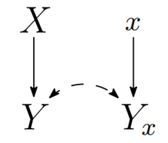
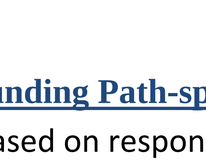


Figure 4: The "w graph".

❖ We formulate the bounding problem of PC fairness as a constrained problem by using "response-function variables".

Response-function Variables

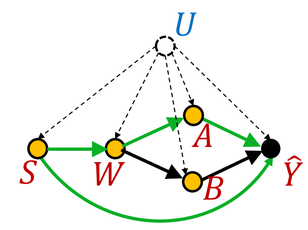
❖ Response-function variables $\boldsymbol{r}$ are used to parameterize the causal model.
- Categorize the unknown domain of $\boldsymbol{U}$ into limited number of equivalent regions, each of which is denoted by a value of a response-function variable.
- Represent unknown functions $\boldsymbol{F}$ by limited number of mappings from parents to dependent variables. Each corresponds to one equivalent region of $\boldsymbol{U}$.
- As a result, all uncertainties in the causal model parameterized by $P(\boldsymbol{r})$



$$r_Y = \begin{cases} 0 \text{ if } f_Y(x_0, u_Y) = y_0, f_Y(x_1, u_Y) = y_0 \\ 1 \text{ if } f_Y(x_0, u_Y) = y_0, f_Y(x_1, u_Y) = y_1 \\ 2 \text{ if } f_Y(x_0, u_Y) = y_1, f_Y(x_1, u_Y) = y_0 \\ 3 \text{ if } f_Y(x_0, u_Y) = y_1, f_Y(x_1, u_Y) = y_1 \end{cases}$$

$$r_X = \begin{cases} 0 \text{ if } f_X(u_X) = x_0 \\ 1 \text{ if } f_X(u_X) = x_1 \end{cases} \qquad P(x, y) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(x; r_X)\mathbb{I}(y; x, r_Y)$$

Expressing Path-specific Counterfactual Fairness



$$\pi = \{S \to W \to A \to \hat{Y}, S \to \hat{Y}\}$$
$$O = \{s_0, w', a', b'\}$$

❖ Response-function variables are given by $\boldsymbol{R} = \{R_S, R_W, R_A, R_B, R_{\hat{Y}}\}$.

❖ Applying response functions, we obtain:

$$P\left(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\boldsymbol{o}\right) = \sum_{\substack{a,b,w_1,w_0, \\ \boldsymbol{r} \in \boldsymbol{r_o}}} \frac{P(\boldsymbol{r})}{P(\boldsymbol{o})} \begin{bmatrix} \mathbb{I}(\hat{y}; a, b, s_1, r_{\hat{Y}}) \\ \mathbb{I}(a; w_1, r_A) \\ \mathbb{I}(b; w_0, r_B) \\ \mathbb{I}(w_1; s_1, r_W) \\ \mathbb{I}(w_0; s_0, r_w) \end{bmatrix}$$

$$P\left(\hat{y}_{s_0}|\boldsymbol{o}\right) = \sum_{\substack{a,b,w, \\ \boldsymbol{r} \in \boldsymbol{r_o}}} \frac{P(\boldsymbol{r})}{P(\boldsymbol{o})} \begin{bmatrix} \mathbb{I}(\hat{y}; a, b, s_0, r_{\hat{Y}}) \\ \mathbb{I}(a; w_0, r_A)\mathbb{I}(b; w_0, r_B) \\ \mathbb{I}(w_0; s_0, r_w) \end{bmatrix}$$

Bounding Path-specific Counterfactual Fairness

❖ Based on response-function variables, finding the bounds of PC fairness is formulated as follows

$$\begin{aligned} \min/\max \quad & P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o}), \\ \text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0, \end{aligned}$$

where the objective function is the path-specific counterfactual effect, $P(\boldsymbol{V})$ is the parameterized distribution, which agrees with the observational distribution $P(\boldsymbol{D})$.

Acknowledgement

This work was supported in part by NSF 1646654, 1920920, and 1940093.
Paper is available at NIPS Proceedings.
Code is available at http://tiny.cc/pc-fairness-code.

Email: {yw009, lz006, xintaowu}@uark.edu; htong@illinois.edu

2019 Conference on Neural Information Processing Systems (NeurIPS 2019)
Vancouver, Canada, Dec. 8-14, 2019