

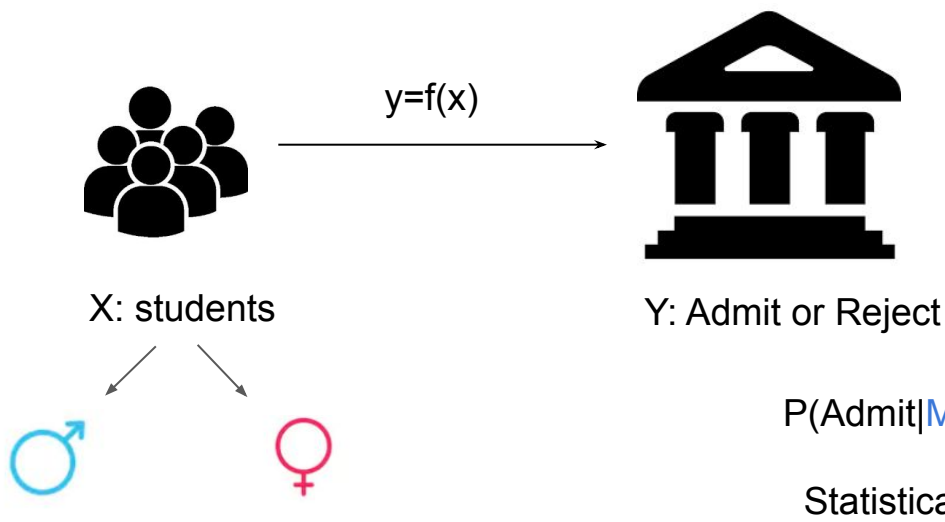
# Causality, Fairness, Representation Learning

Prince Zizhuang Wang

# Fair Classification

Reality

Dataset



$P(\text{Admit}|\text{Male}) >$

Unfair to F

# Group Fairness: Demographic Parity

Y: outcome; S: sensitive attribute

Demographic Parity (DP):  $E[Y|S=0] - E[Y|S=1]$

DP = 0 implies **statistical independence**

No discrimination against Groups

The Common Approach is to ensure **statistical independence** between Y and S  
Including:

Learning Fair Representations, Zemel, 2013

The Variational Fair Autoencoder, Louizos, 2016

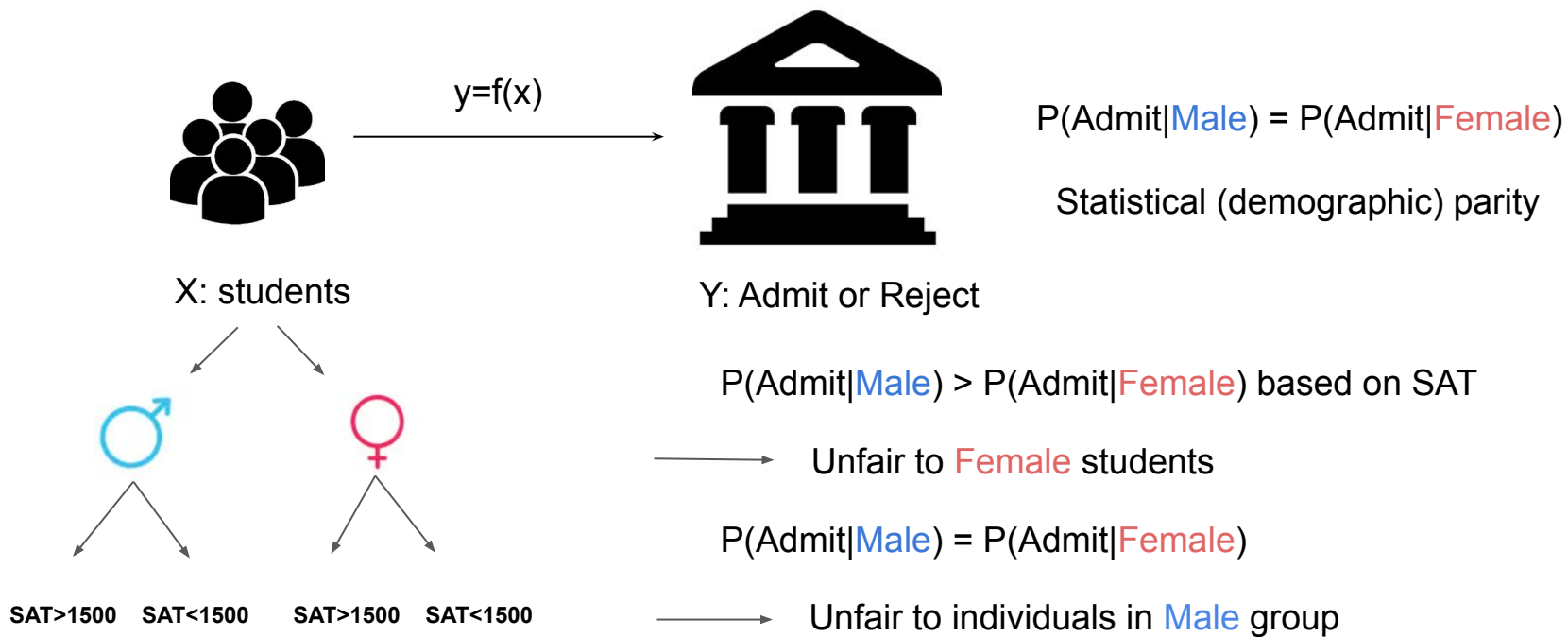
Learning Adversarially Fair and Transferrable Representations, Madras, 2018

Flexibly Fair Representation Learning by Disentanglement, Creager, 2019











Learning Controllable Fair Representations, Song, 2019

And etc...

# Group or Individual



# Counterfactual Fairness: Individual Fairness?

Reality			Counterfactual Reality		
	GPA=3.8	(Admit)	→		GPA=3.8 (Reject) Unfair
	GPA=3.8	(Admit)	→		GPA=3.8 (Admit) Counterfactually fair
	GPA<3.5	(Reject)	→		GPA<3.5 (Admit) Unfair
	GPA<3.5	(Reject)	→		GPA<3.5 (Reject) Counterfactually fair
	GPA=3.8	(Admit)	→		GPA=3.8 (Reject) Unfair

Formal Definition: 
$$p(y_{s^+} | x, s^-) = p(y_{s^-} | x, s^-)$$

# Causality vs Correlation:



CRIME

## **When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?**

By JUSTIN PETERS

JULY 09, 2013 • 2:59 PM



# Causality:

Relation of Ice Cream Sales, Crime Rate, and Temperature

Joint distribution:  $P(\text{Ice Cream}, \text{Crime Rate})$



cause



$P(\text{Crime Rate}|\text{Ice Cream}) P(\text{Ice Cream})$



cause



$P(\text{Ice Cream}|\text{Crime Rate}) P(\text{Crime Rate})$

Temperature



$$\int P(\text{Ice Cream}|T)P(\text{Crime Rate}|T)P(T)dT$$

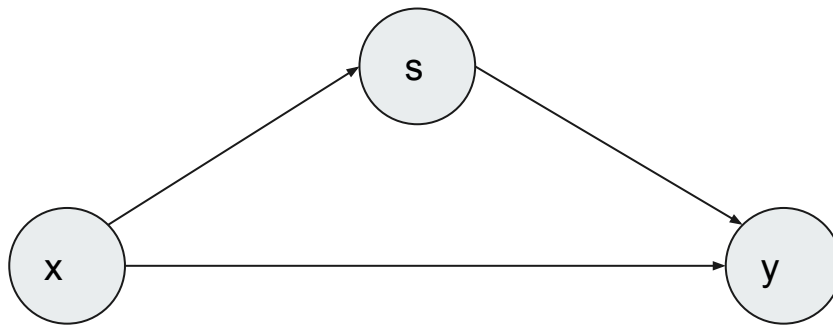
# Causality: intervention, do-calculus

## Kidney Stones

**S:** Treatment, +/-

**X:** Size of stone

**Y:** Recover



do notation:  $\text{do}(S=+)$  denotes the action of selecting + as the treatment

Counterfactual:  $y_{s+}$  denotes the potential recover rate when we intervene the treatment

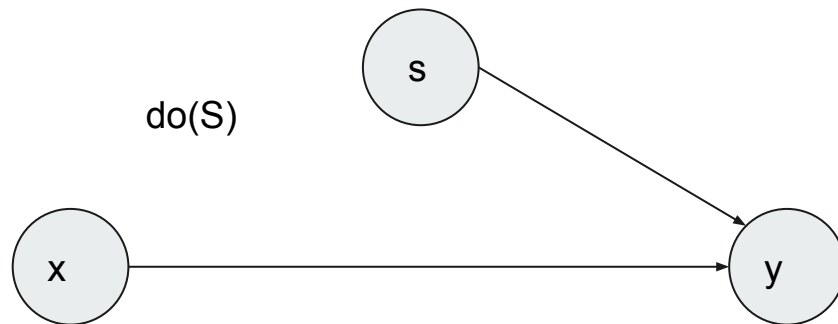
Interventional distribution:  $p(y|\text{do}(S = +))$



# How to estimate $p(y|do(S = +))$ : Back-door criterion

$$p(y|do(S = +)) \neq p(y|S = +)$$

Treatment +	Treatment -
276/350	289/350



Size	Treatment +	Treatment -
small	84/87 (0.96)	234/270 (0.87)
large	192/263 (0.73)	55/80 (0.68)

$$p(y|do(S = +)) = \sum_x p(y|S = +, x)p(x)$$

After adjustment,  
 $P(y|do(S=+)) = 0.85$   
 $P(y|do(S=-)) = 0.78$

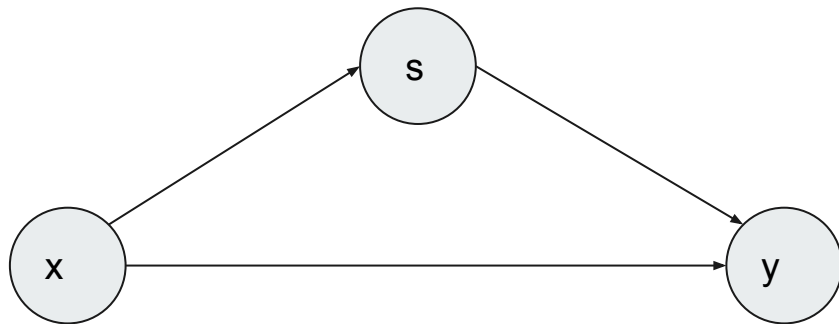
# Back-door criterion

Valid Adjustment set for S:

1, block all the back door from S to Y

2, not a child of Y

In this case, X is a valid adjustment for S

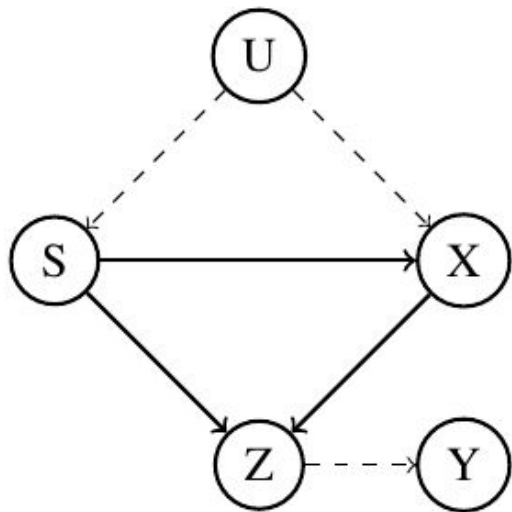


If X is a valid adjustment for S, then

$$p(y|do(S = +)) = \int p(y|do(S = +), x)p(x)dx$$
$$= \int p(y|S = +, x)p(x)dx$$

# Counterfactual Fair Representation

Unobserved confounder



Counterfactual Fairness:

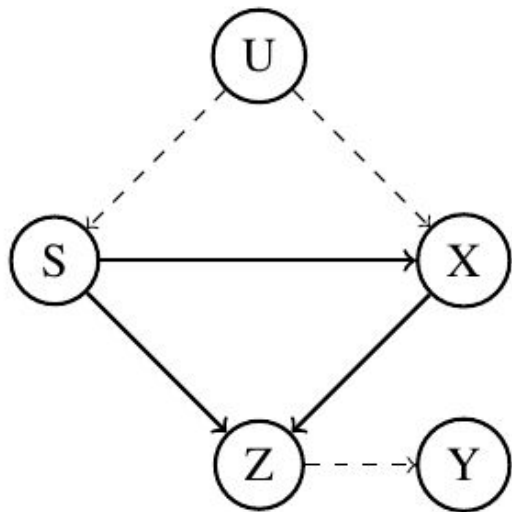
If  $z$  is a function of non-descendants of  $s$ , then

$$p(z|do(S = +)) = p(z|do(S = -))$$

This is infeasible.  $X$  may contain descendants of  $S$

# Counterfactual Fair Representation

Unobserved confounder



Our model is general because:  
Allow S and Y to be dependent  
Allow S to be cause of Y

$$p(z|do(S = s)) = \int p(z|S = s, x)p(x)dx$$

Define a metric (Distance or Divergence):

$$W(p_{\theta}(z|do(s = 0)) - p_{\theta}(z|do(s = 1)))$$

# SinkHorn: Lightspeed optimal transport cost

$$\begin{aligned}\mathcal{W}_{c,\epsilon}(\mu_\theta, \nu_\theta) = & \inf_{\pi \sim \Pi(\mu_\theta, \nu_\theta)} \mathbb{E}_{(z_0, z_1) \sim \pi} [c(z_0, z_1)] \\ & + \epsilon \int \log \frac{\pi(z_0, z_1)}{d\mu_\theta(x)d\nu_\theta(y)} d\pi(z_0, z_1) \quad \text{Entropy term}\end{aligned}$$

Sinkhorn Divergence

$$\begin{aligned}\mathcal{S}_{c,\epsilon}(\mu_\theta, \nu_\theta) = & 2\mathcal{W}_{c,\epsilon}(\mu_\theta, \nu_\theta) - \mathcal{W}_{c,\epsilon}(\mu_\theta, \mu_\theta) \\ & - \mathcal{W}_{c,\epsilon}(\nu_\theta, \nu_\theta)\end{aligned}$$

Fast and Stable

# Implementation

method	Accuracy	Adult	
		DP	ACE $\times 10^{-2}$
3-layer NN	82.25	1.80	5.82
Fair VAE	84.82	1.62	0.71
IFCM	85.21	1.10	3.06

$$ACE = \mathbb{E}[y|do(s=0)] - \mathbb{E}[y|do(s=1)]$$

$$\Delta_{DP} = |\mathbb{E}[y|s=0] - \mathbb{E}[y|s=1]|$$

---

**Algorithm 1** Independent Fair Causal Mechanism Trained with SinkHorn Divergence.

---

- 1: **Input:** Binary sensitive attribute  $S$ , non-sensitive attributes  $X$ ,  $\theta$  (model parameters),  $\varphi$  (critic parameters), critic function  $f_\varphi$ . clipping limit  $c$
  - 2:  $\theta \leftarrow \theta_0, \varphi \leftarrow \varphi_0$
  - 3: **for**  $k = 1, 2, \dots$  **do**
  - 4:     **for**  $t = 1, 2, \dots, n_c$  **do**
  - 5:         Sample  $\{x_i\}_1^M$  from the empirical distribution.
  - 6:         compute  $\{z_0\}_1^M = f_\theta(\{x_i\}_1^M, s=0)$ .
  - 7:         compute  $\{z_1\}_1^M = f_\theta(\{x_i\}_1^M, s=1)$ .
  - 8:          $\varphi \leftarrow \varphi + \alpha \text{RMSProp}(\nabla_\varphi \mathcal{S}_{c,\epsilon})$
  - 9:     **end for**
  - 10:     Sample  $\{x_i\}_1^M$  and  $\{s_i\}_1^M$  from the empirical distribution.
  - 11:      $\{z\}_1^M = f_\theta(\{x_i\}_1^M, \{s_i\}_1^M)$
  - 12:      $\{\hat{y}_i\}_1^M = g_\theta(\{z\}_1^M)$
  - 13:      $L = \mathcal{S}_{c,\epsilon} + \text{CrossEntropy}(\{\hat{y}_i\}_1^M, \{y_i\}_1^M)$
  - 14:      $\theta \leftarrow \theta - \alpha \text{RMSProp}(\nabla_\theta L)$
  - 15: **end for**
-

# Optimal transport cost for Group Fairness?

Given a Lipschitz loss function  $f$ , define the expected risk as

$$R(f) = \int_{\mathcal{Z}} f(z, y) dD(z, y) = \mathbb{E}_{z, y \sim D} [f(z, y)]$$

Define  $R_0$  and  $R_1$  as the expected risk on the interventional distributions  $p(z|do(S=0))$   
 $p(z|do(S=1))$

Theorem:  $R_0 - R_1 \leq K * W(p(z|do(s=0)), p(z|do(s=1)))$

What does the Theorem tells us?

Intervention does not affect accuracy by much

Does it imply group fairness?

# Proof

$$\begin{aligned}R_0 - R_1 &= \mathbb{E}_{z \sim P_0} \mathbb{E}_{p(y|z)}[f(z, y)] - \mathbb{E}_{z \sim P_1} \mathbb{E}_{p(y|z)}[f(z, y)] \\&= \mathbb{E}_{p(y|z)}[\mathbb{E}_{z \sim P_0}[f(z, y)] - \mathbb{E}_{z \sim P_1}[f(z, y)]] \\&= \mathbb{E}_{p(y|z)}[\mathbb{E}_{z \sim P_0}[f_y(z)] - \mathbb{E}_{z \sim P_1}[f_y(z)]] \\&\leq K \cdot \mathbb{E}_{p(y|z)}\left[\frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{z \sim P_0}[f(z)]\right. \\&\qquad\qquad\qquad \left. - \mathbb{E}_{z \sim P_1}[f(z)]\right] \\&= K \cdot \mathbb{E}_{p(y|z)}[\mathcal{W}(P_0, P_1)] \\&= K \cdot W(P_0, P_1)\end{aligned}$$



# Causal Inference:

DAG-DNN: DAG Structure Learning with Graph Neural Network, 2019

Neural Attribution: A Causal Perspective, 2019

Group Invariance principles for Causal Generative Models, 2017

Counterfactuals Uncover the Modular Structure of Deep Generative Models, 2018

Identification of Conditional Causal Effects under Markov Equivalence, 2019

# Thanks!

# Questions ?

Github: [kingofspace0wzz/uai2020-fair](https://github.com/kingofspace0wzz/uai2020-fair)

Overleaf: Let me know