Fair and Optimal Prediction via Post-Processing

Waterloo Artificial Intelligence Institute

Nov. 24th, 2023

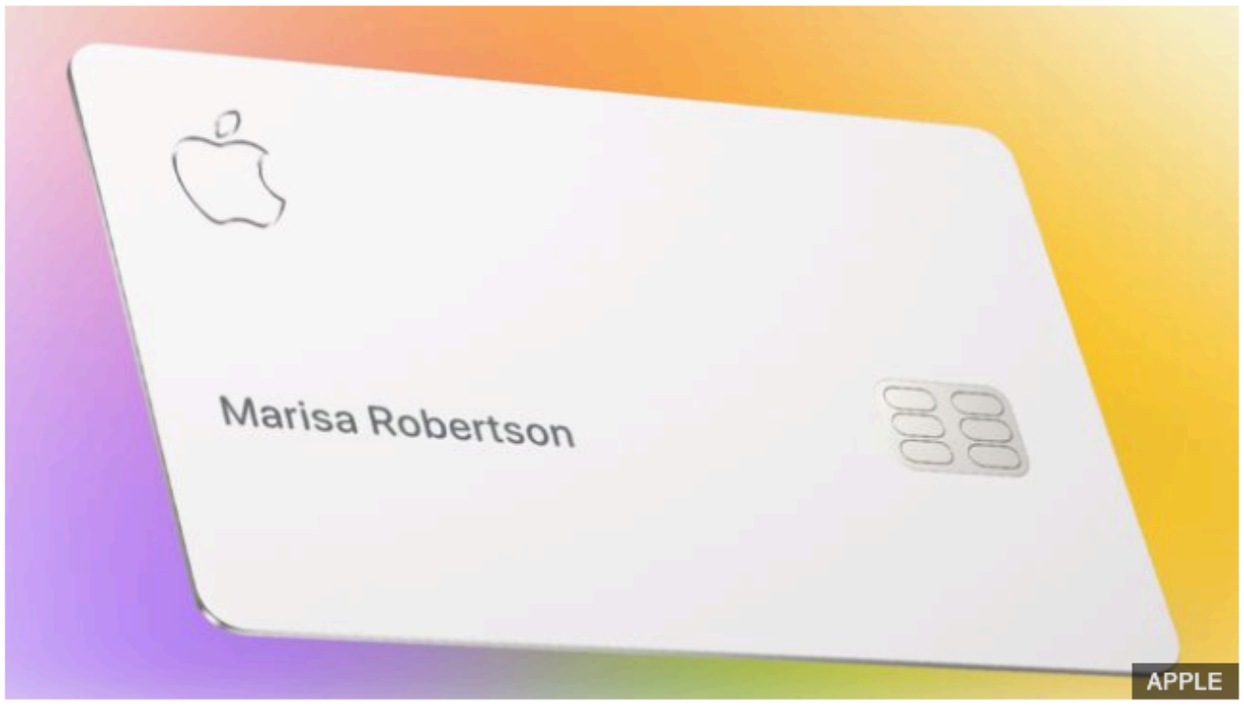
Han Zhao
hanzhao@illinois.edu
Amazon Visiting Academic
Assistant Professor
Department of Computer Science
University of Illinois Urbana-Champaign

ILLINOIS
Computer Science
GRAINGER COLLEGE OF ENGINEERING

Potential Bias of Data in High-stakes Domains

Apple's 'sexist' credit card investigated by US regulator

11 November 2019 Share



A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

TECHNOLOGY NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

Algorithmic Fairness: Statistical Parity

Statistical Parity (aka Demographic Parity)

$$\hat{Y} \perp A$$

ithm shouldn't take the

"Building classifiers with

CNN politics The Biden Presidency Facts First 2022 Midterms Audio Log In

LIVE UPDATES

Supreme Court guts affirmative action in college admissions

By Aditi Sangal, Adrienne Vogt, Sydney Kashiwagi, Matt Meyer and Tori B. Powell, CNN

Updated 2139 GMT (0539 HKT) June 29, 2023

Hear what happened inside the Supreme Court after historic ruling 05:22

What we're covering here

- The Supreme Court ruled colleges and universities can no longer take race into consideration as a specific basis in admissions — a landmark decision that overturns long-standing precedent that has benefited Black and Latino students in higher education.
- Chief Justice John Roberts, who wrote the opinion for the conservative majority, said Harvard and University of North Carolina admissions programs violated

All | Anal

46 Posts

17 min ago

Here's w landmarl

From CNN's Ariane d

The Supreme specific basis that has benef

Chief Justice

BBC NEWS | BBC

Affirmative action: US Supreme Court overturns race-based college admissions

Bernd Debusmann Jr - BBC News, Washington

Fri, June 30, 2023 at 5:00 a.m. GMT+9 · 5 min read

The US Supreme Court has ruled that race can no longer be considered as a factor in university admissions.

The landmark ruling upends decades-old US policies on so-called affirmative action, also known as positive discrimination.

It is one of the most contentious issues in US education.

Affirmative action first made its way into policy in the 1960s, and has been defended as a measure to increase diversity.

TRENDING

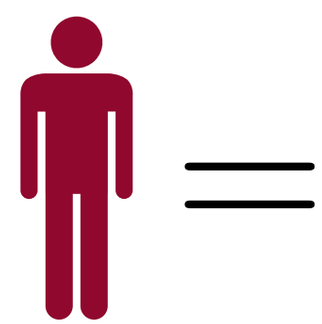
1. Toronto Argonauts have hit the ground running to open CFL season
2. 2023 NHL Draft: Final grades for all 32 teams
3. Blackhawks acquire Corey Perry from Lightning, adding more experience to Bedard-led rebuild
4. 2023 NHL Draft recap: Every pick made in Rounds 1-7
5. 'Pretty boring': Oilers ship Yamamoto to Wings, but little trade action at NHL draft

Fairness Through Blindness

Statistical Parity

Ignorance is bliss?! — Thomas Gray

$$C(X, \not{A}) \implies C(X)$$



Defendant

$=$

$$\begin{pmatrix} \text{prior arrests} \\ \text{prior sentences} \\ \text{age} \\ \text{drug history} \\ \cancel{\text{race}} \\ \text{age at first arrest} \\ \text{education history} \\ \text{vocation history} \\ \text{gender} \end{pmatrix}$$

Fairness Through Blindness

$$C(X, \not{A}) \implies C(X)$$

Defendant

$=$

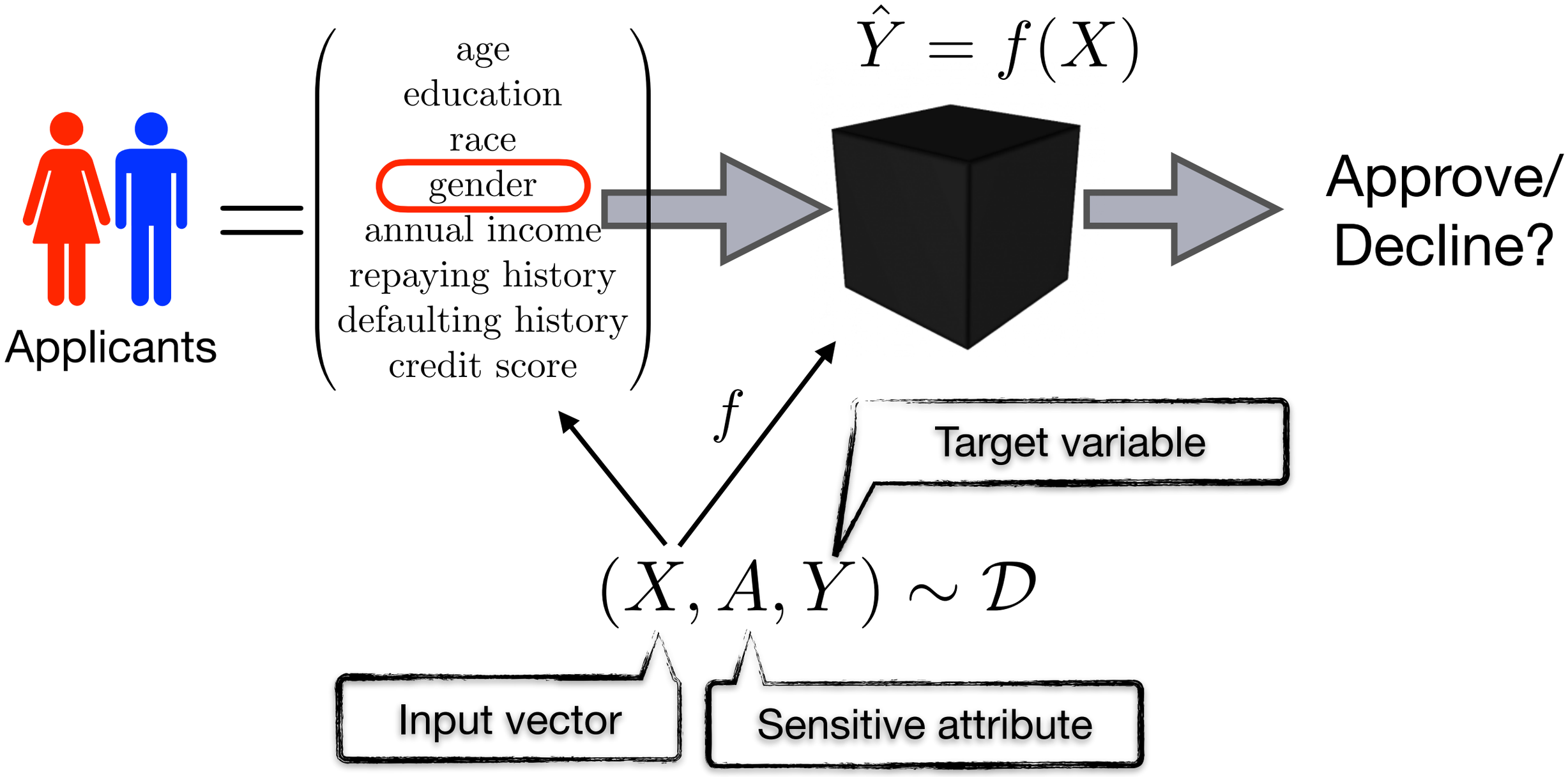
$$\begin{pmatrix} \text{prior arrests} \\ \text{prior sentences} \\ \text{age} \\ \text{drug history} \\ \cancel{\text{race}} \\ \text{age at first arrest} \\ \text{education history} \\ \text{vocation history} \\ \text{gender} \end{pmatrix}$$

Does this mechanism work?

- No, due to “redundant encoding”
- Other attributes in the inputs could be used to reconstruct the deleted sensitive attributes due to the potential correlations among them
  - Ethnicity vs hair color/last name
  - Race vs zipcode

Algorithmic Fairness: Statistical Parity

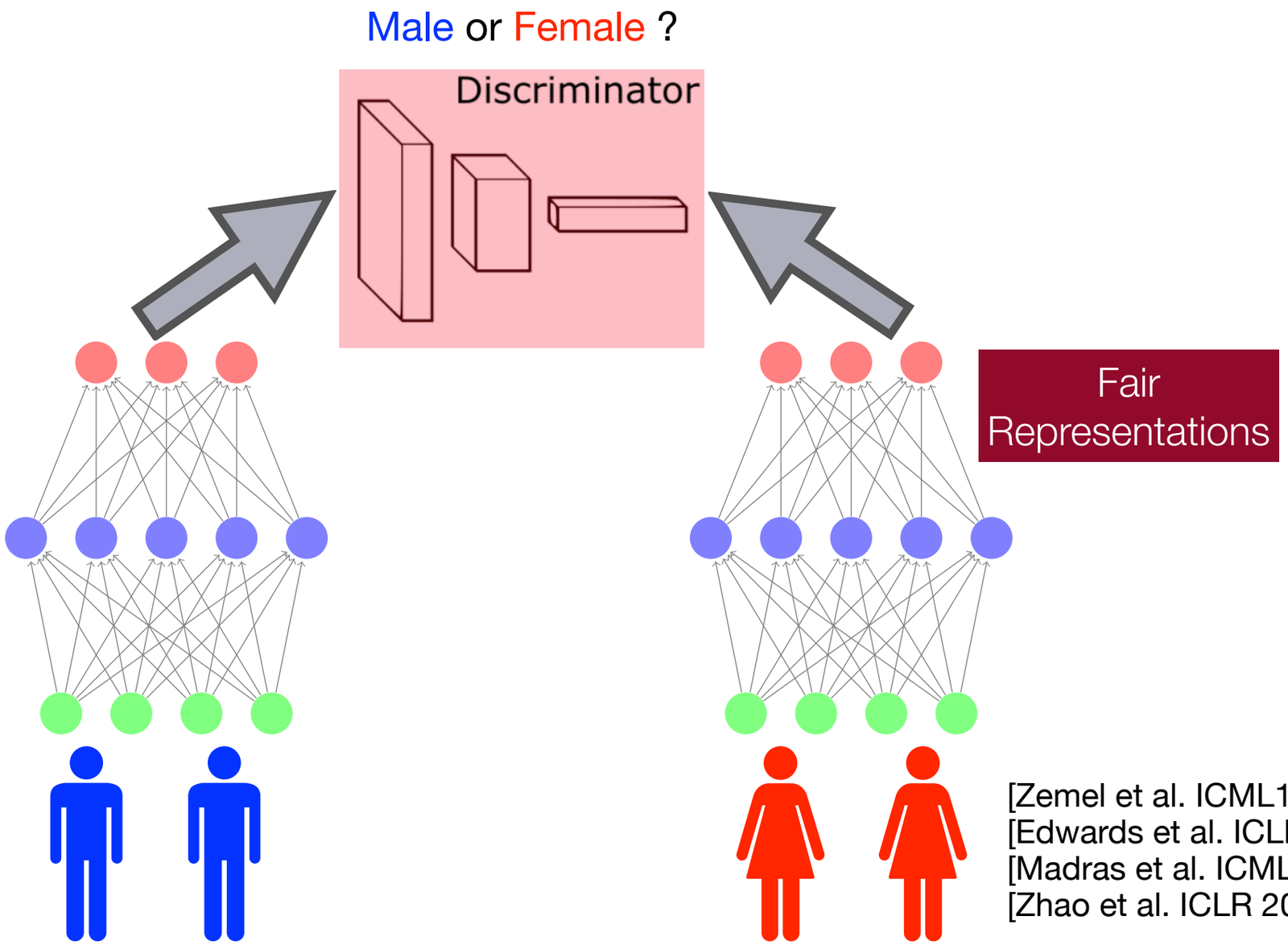
Example in loan application



Statistical parity: any fair algorithm cannot take information related to sensitive attribute during decision making

Algorithmic Fairness: Statistical Parity

Pre-processing Methods: Feature Learning

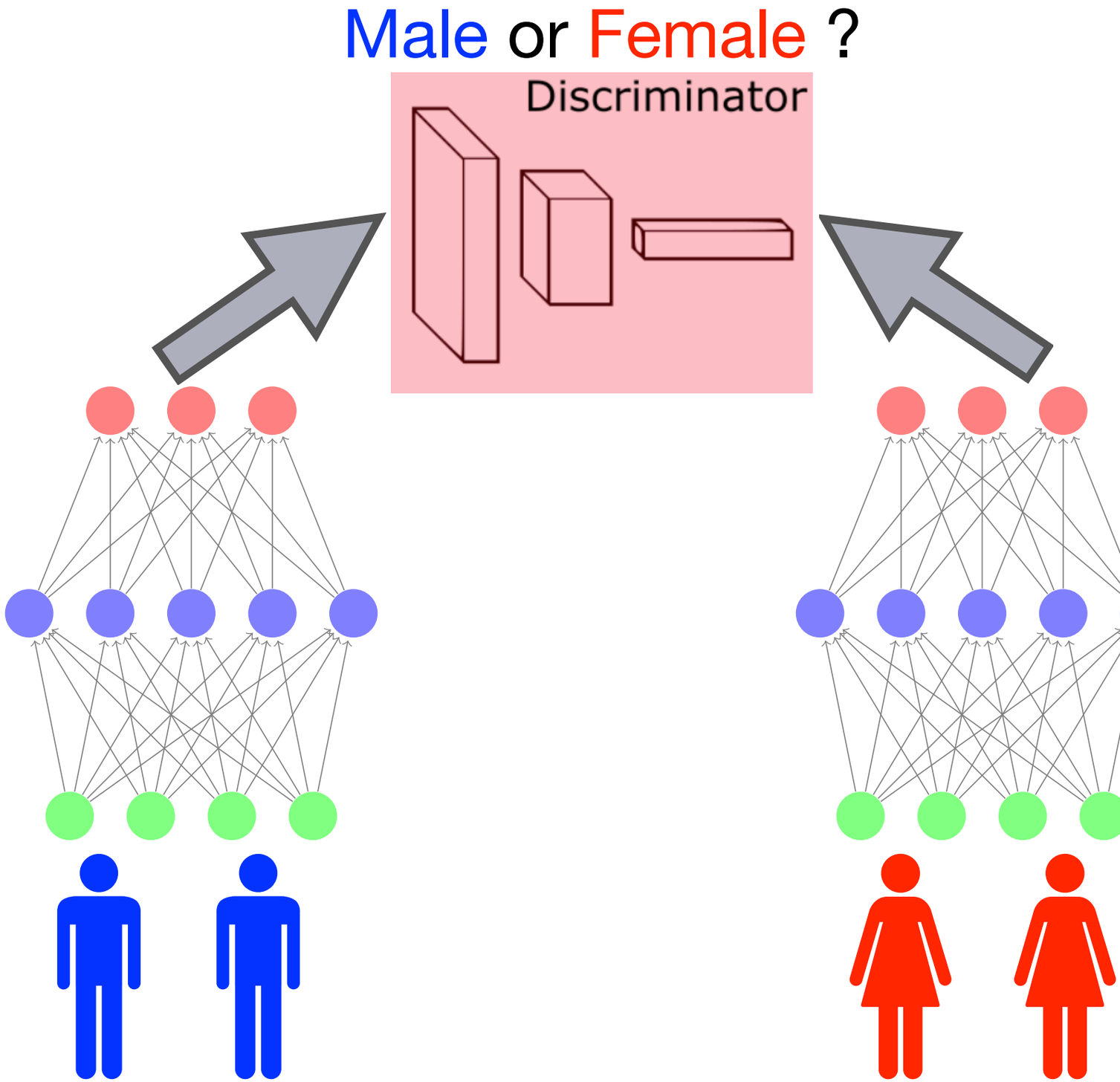


[Zemel et al. ICML13]
[Edwards et al. ICLR 15]
[Madras et al. ICML 18]
[Zhao et al. ICLR 20]

- Needs full access to (X, A, Y)
- In practice: minimax optimization can be unstable and hard for neural networks

Algorithmic Fairness: Statistical Parity

Adversarial Training



Fair Representations

[Zemel et al. ICML13]
[Edwards et al. ICLR 15]
[Madras et al. ICML 18]

Algorithmic Fairness: Statistical Parity

In-processing Methods: Constrained Optimization

$$\min_{\theta} \quad \mathbb{E}[\ell(f_\theta(x), y)]$$

$$\text{subject to} \quad |\Pr(f_\theta(x) = 1 \mid A = 0) - \Pr(f_\theta(x) = 1 \mid A = 1)| \leq \epsilon$$

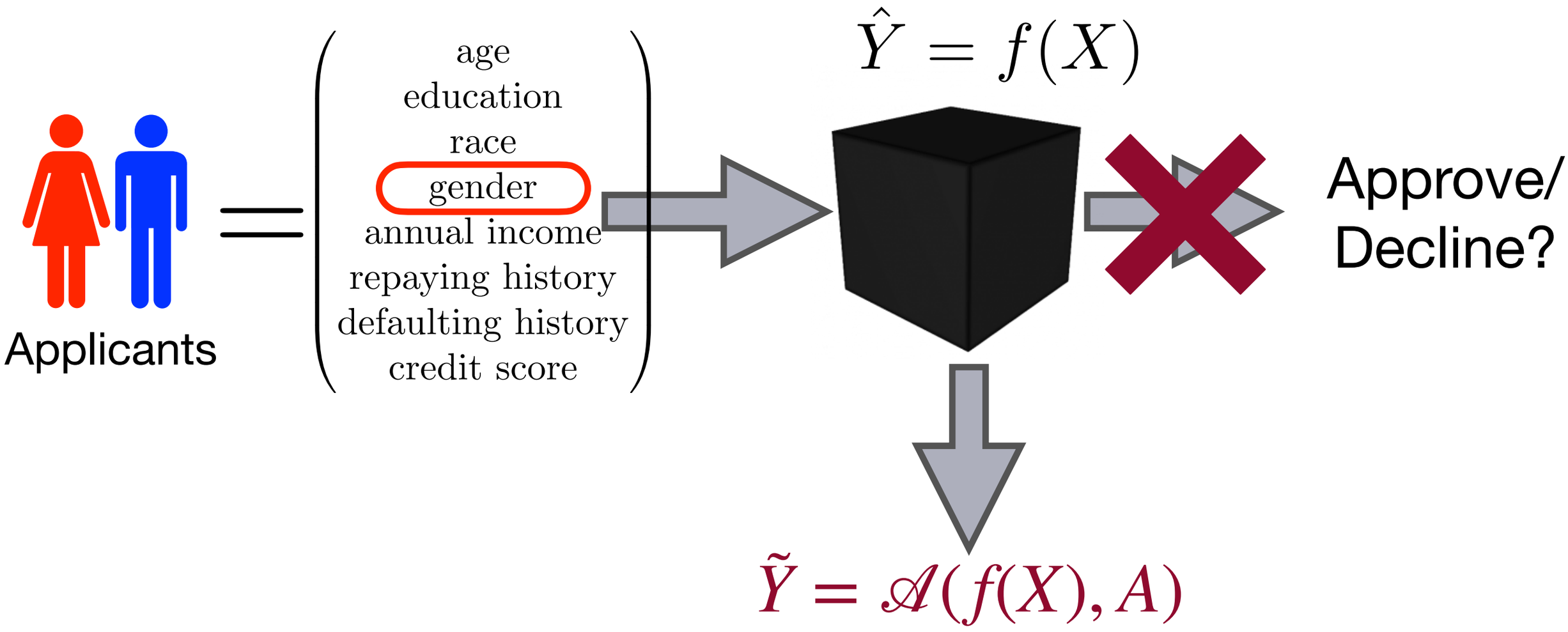
error minimization

constraint of approximate statistical parity

- Needs full access to (X, A, Y)
- Need to design dedicated optimization solvers for each different model $f_\theta(\cdot)$
- We may not be able to train the model from scratch due to limited computational resources, e.g., LLMs

Algorithmic Fairness: Statistical Parity

Post-processing Methods:



- No need to have full access to (X, A, Y)
- The given classifier $f(\cdot)$ can be treated as a black-box
- No need to re-train the model from scratch

Fairness-Accuracy Tradeoff

Before we talk about the algorithm:

- Is there any price we have to pay for fairness? If yes, what's the price (in terms of accuracy)?
- Is it possible to derive an algorithm that achieves the optimal accuracy under the constraint of fairness (statistical parity)?

Fairness-Accuracy Tradeoff

Statistical parity: enforcing statistical independence, will this lead to loss of accuracy?

Consider some extremal cases:

- What if $Y \perp A$ in the underlying distribution?
- What if $Y = A$ in the underlying distribution?

There should be a term that quantifies the dependency of these two random variables!

The tradeoff result should be inherent:

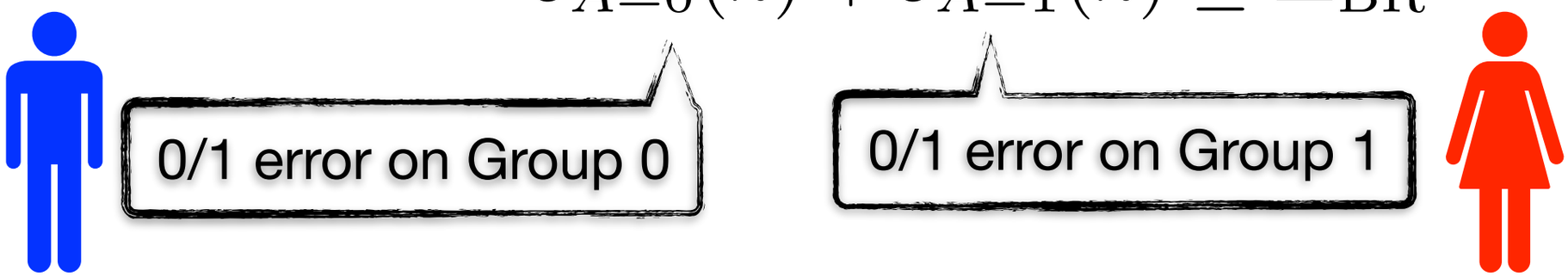
- Does not depend on the specific algorithm used to achieve statistical parity
- Does not depend on the computational resources available to the algorithm
- Does not depend on the sample size for training the predictor

Fairness-Accuracy Tradeoff

Statistical parity: $\hat{Y} \perp A$

Theorem [ZG, NeurIPS 19]: For any fair algorithm $\widehat{Y} = h(X)$ (in the sense of statistical parity), the following inequality holds:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$$



Key Message: when the base rates differ, any fair algorithm has to make a large error on at least one of the groups

(Improper) Analogy: a kind of uncertainty principle for fairness $\Delta p \cdot \Delta x \geq \frac{\hbar}{2}$

Difference of base rates:

$$\Delta_{\mathrm{BR}} := |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$$

Fairness-Accuracy Tradeoff

Theorem [ZG, NeurIPS 19]: For any fair algorithm $\widehat{Y} = h(X)$ (in the sense of statistical parity), the following inequality holds:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$$

Difference of base rates:

$$\Delta_{\mathrm{BR}} := \left| \Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1) \right|$$

- If $A = Y$, then $\Delta_{\mathrm{BR}} = 1$, meaning $\max\{\varepsilon_{A=0}(h), \varepsilon_{A=1}(h)\} \geq 0.5$
- If $A \perp Y$, then $\Delta_{\mathrm{BR}} = 0$, meaning no tension with utility

Δ_{BR} is a fundamental quantity to characterize the coupling between target and sensitive attribute

Fairness-Accuracy Tradeoff

But, why the specific form of this lower bound? Why not the joint error?

- A simple corollary regarding the joint error could be obtained:

$$\begin{aligned}\varepsilon(h) &= \Pr(A=0)\cdot\varepsilon_{A=0}(h) + \Pr(A=1)\cdot\varepsilon_{A=1}(h)\\ &\geq \min\{\Pr(A=0),\Pr(A=1)\}\cdot\left(\varepsilon_{A=0}(h)+\varepsilon_{A=1}(h)\right)\\ &\geq H_{01}(A)\cdot\Delta_{\mathrm{BR}}\end{aligned}$$

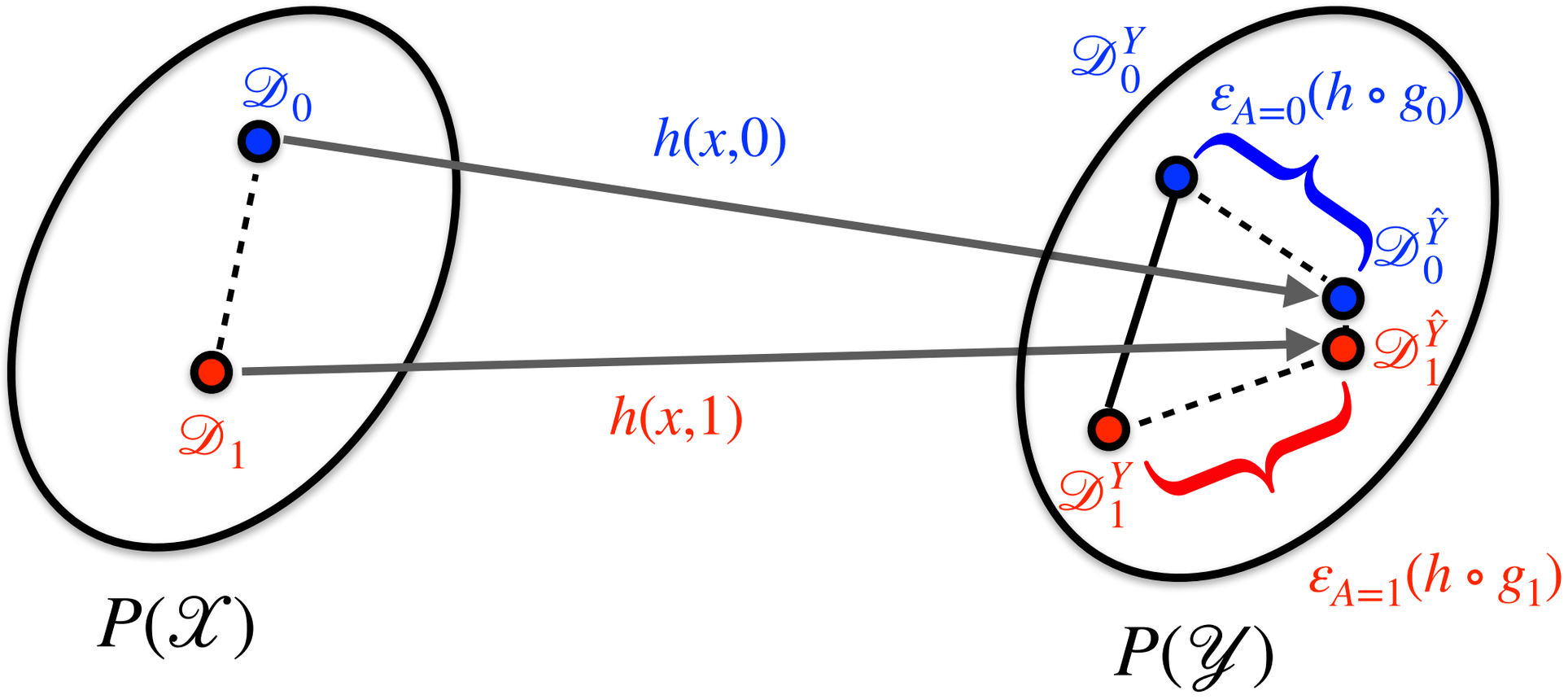
where $H_{01}(A) := 1 - \max_a \Pr(A=a)$ is called zero-one entropy of A

- Any lower bound for the joint error has to depend on the marginal distribution of the sensitive attribute A, which could bias towards the majority group
  - ✓ Instead, the lower bound in Theorem 1 treats both errors equally
  - ✓ In cases where the ratio between two groups is extremely imbalanced, the lower bound for the joint error could be 0, i.e., no price to pay in terms of the joint error

Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. Annals of statistics' 04

Proof Sketch

We provide a proof sketch for an attribute-aware classifier:



$\mathscr{D}_a$: input distributions over group $A = a$

$\mathscr{D}_a^{\hat{Y}}$: predicted label distributions over group $A = a$

$\mathscr{D}_a^{Y}$: ground-truth label distributions over group $A = a$

An Optimal Fair Classifier via Post-Processing

Is it possible to construct a classifier that verifies the lower bound?

Why should we care about this question?

- Can confirm the tightness of the inequality
- Can design optimal classifiers on the fairness-accuracy frontier

[Why?] However, this problem cannot be easier than learning the Bayes classifier without fairness constraint

An Optimal Fair Classifier via Post-Processing

This problem cannot be easier than learning the Bayes classifier

- Assume we have oracle access to the problem of learning optimal fair classifier
- Use this oracle access to learn the Bayes optimal classifier

Problem A:

Let μ' be a distribution over $\mathcal{X} \times \mathcal{Y}$. We want to learn $h'(\cdot)$, the Bayes optimal classifier over μ'

Problem B:

Let μ be a distribution over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. We want to learn $h(\cdot, \cdot)$, the optimal fair classifier over μ

To show that Problem A $\ll$ Problem B, suppose we have an algorithm to solve Problem B, we could use that algorithm to solve Problem A as well.

An Optimal Fair Classifier via Post-Processing

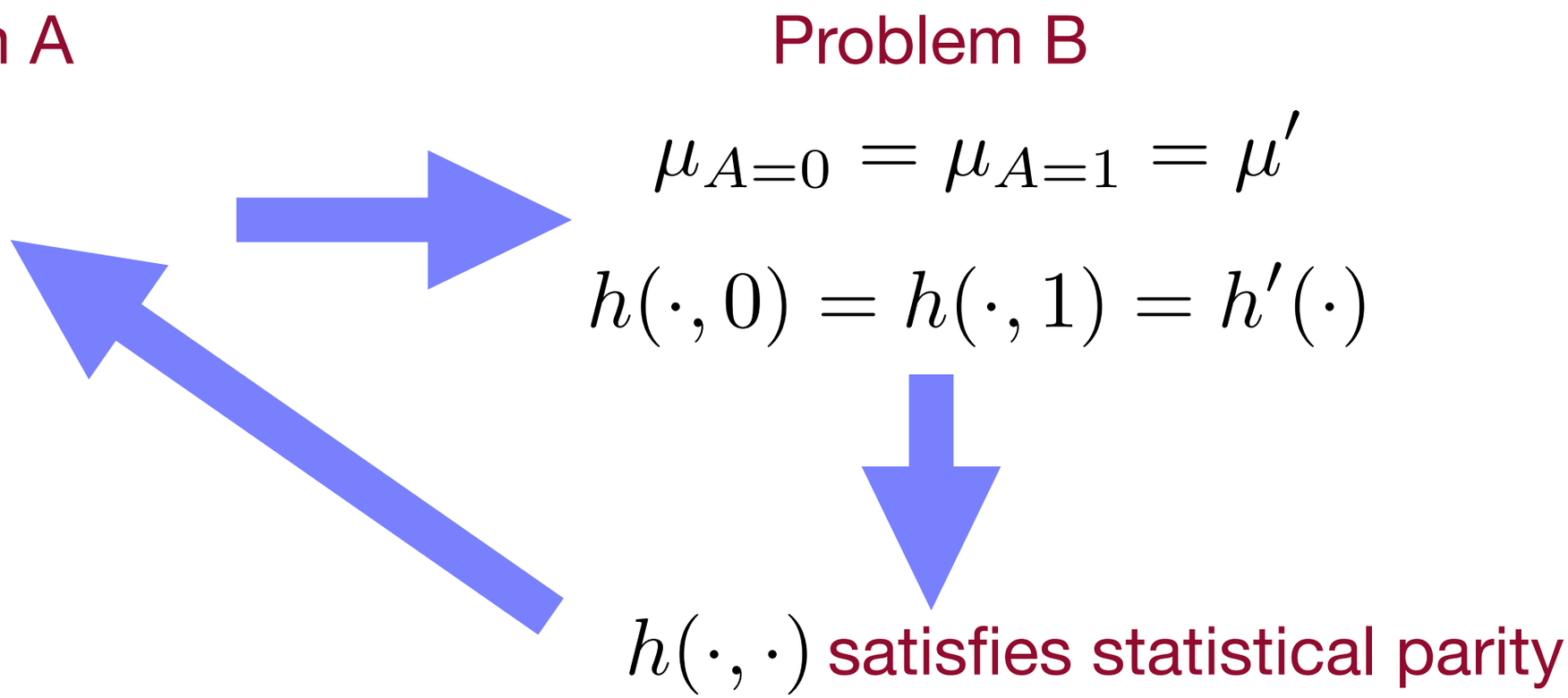
Problem A:

Let μ' be a distribution over $\mathcal{X} \times \mathcal{Y}$. We want to learn $h'(\cdot)$, the Bayes optimal classifier over μ'

Problem B:

Let μ be a distribution over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. We want to learn $h(\cdot, \cdot)$, the optimal fair classifier over μ such that the lower bound holds

Reduction:



[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

Bad news: We know that learning the Bayes optimal classifier is computationally hard in general, even for simple function classes like linear predictors

Instead, what we can aim for is:

Given oracle access to Bayes classifiers, could we construct an algorithm to learn the optimal fair classifier?

Algorithm 1 Optimal fair classifier

Input: Oracle access to h_0^* and h_1^*, the Bayes optimal classifiers over μ_0 and μ_1

Output: A randomized optimal fair classifier $h^*_{\text{Fair}} : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$

1: Compute $\alpha := \Pr_{\mu_0}(Y = 1)$ and $\beta := \Pr_{\mu_1}(Y = 1)$. Without loss of generality assume $\alpha \geq \beta$

2: For (x, a), randomly sample $s \sim U(0, 1)$, the uniform distribution between (0,1)

3: Construct $h^*_{\text{Fair}}(x, a)$ as

$$h^*_{\text{Fair}}(x,a) := \begin{cases} a = 0 : \begin{cases} 0 & \text{If } h_0^*(x) = 0 \text{ or } h_0^*(x) = 1 \text{ and } s > \frac{\alpha+\beta}{2\alpha} \\ 1 & \text{If } h_0^*(x) = 1 \text{ and } s \leq \frac{\alpha+\beta}{2\alpha} \end{cases} \\ a = 1 : \begin{cases} 0 & \text{If } h_1^*(x) = 0 \text{ and } s > \frac{\alpha-\beta}{2(1-\beta)} \\ 1 & \text{If } h_1^*(x) = 1 \text{ or } h_1^*(x) = 0 \text{ and } s \leq \frac{\alpha-\beta}{2(1-\beta)} \end{cases} \end{cases} \tag{4}$$

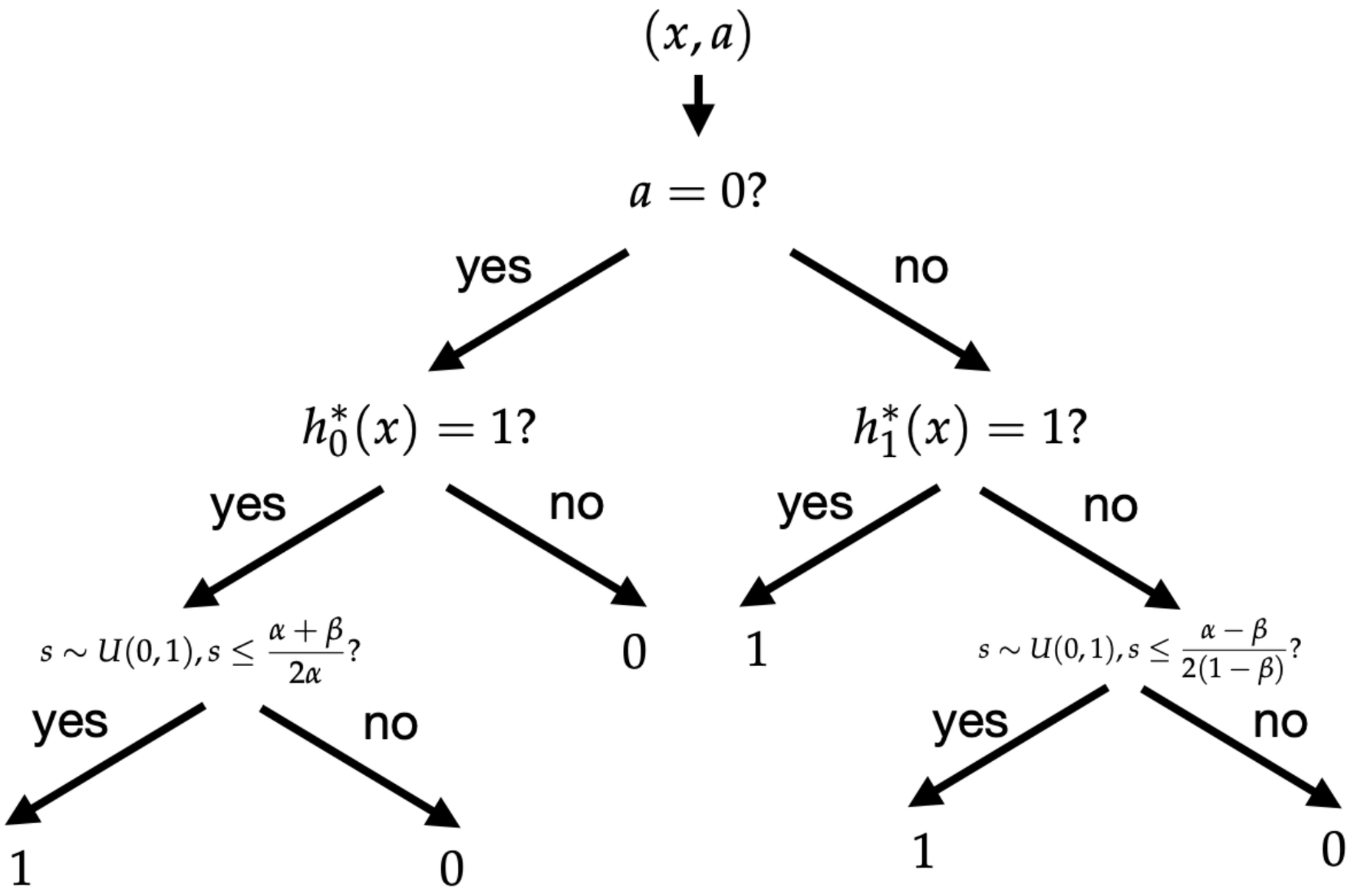
return h^*_{Fair}

[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

A constructive algorithm for the optimal fair classifier:

- It is a randomized classifier
- The classifier needs to have explicit access to the sensitive attribute



[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

Theorem (noiseless): For any distribution μ over (X, A, Y) such that $Y_{A=0} = h_0^*(X)$ and $Y_{A=1} = h_1^*(X)$, the classifier h^*_{Fair} constructed by the algorithm satisfies statistical parity and is optimal, i.e.,

$$\varepsilon_{A=0}(h^*_{\text{Fair}}) + \varepsilon_{A=1}(h^*_{\text{Fair}}) = \Delta_{\text{BR}}$$

Note:

- This theorem assumes 0 Bayes errors, so Δ_{BR} is purely due to the fairness constraint
- It shows that learning fair classifier is not much harder than learning the group-wise Bayes classifiers

[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

Extension to multi-groups under noiseless multi-class classification:

- Let $m \geq 2$ be the number of classes and $n \geq 2$ be the number of groups
- Let Δ_m be the standard $m-1$ dimensional probability simplex
- Let $p_i \in \Delta_m$ be the marginal label distribution of Y from group $i \in [n]$

$$\text{(TV-Barycenter)}: \quad \min_{q} \quad \frac{1}{2}\sum_{i=1}^{n} \|q - p_i\|_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} |(p_i)_j - q_j|$$

$$\text{subject to} \quad q \in \Delta_m : q \geq 0, \ \sum_{j=1}^{m} q_j = 1$$

Let $\mathrm{OPT}\left(\{p_i\}_{i=1}^{m}\right)$ be the optimal value of the above barycenter problem under the Total Variation (TV) distance, then,

$$\sum_{a=1}^{n} \varepsilon_{A=a}(h) \geq \mathrm{OPT}\left(\{p_i\}_{i=1}^{n}\right)$$

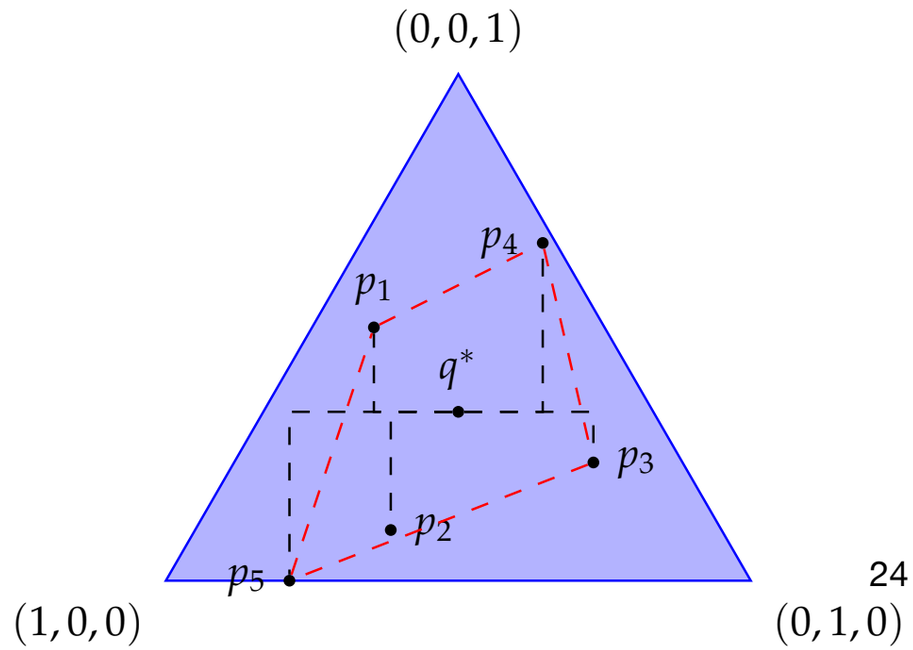
[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

Let $\mathrm{OPT}\left(\{p_i\}_{i=1}^m\right)$ be the optimal value of the above barycenter problem under the Total Variation (TV) distance, then,

$$\sum_{a=1}^{n} \varepsilon_{A=a}(h) \geq \mathrm{OPT}\left(\{p_i\}_{i=1}^n\right)$$

- We no longer have analytical lower bound but the optimal value can be computed efficiently via a linear program
- When $n = 2$, the OPT has a closed form via Δ_{BR}, which is essentially the TV distance between p_0 and p_1
- An extended version of the post-processing algorithm still works, by properly choosing the randomization configuration



[Zhao et al., JMLR' 22b]

An Optimal Fair Classifier via Post-Processing

What about multi-groups but **noisy** multi-class classification?

- Let $m \geq 2$ be the number of classes and $n \geq 2$ be the number of groups
- Let Δ_m be the standard $m-1$ dimensional probability simplex
- Let $f_i \in \Delta_m^{|\mathcal{X}|}$ be the Bayes score function of group $i \in [n]$, i.e., $f_i(x) \in \Delta_m$ with $f_i(x)(j) = \Pr(Y = j \mid X = x, A = i)$

Price of fairness:

(Wasserstein-Barycenter):

$$\begin{aligned} \min_{q} \quad & \frac{1}{2} \sum_{i=1}^{n} W_1 \left(f_i \sharp \mu_i^X, q \right) \\ \text{subjectto} \quad & q \in \Delta_m \end{aligned}$$

Note: $f_i \sharp \mu_i^X$ is the push-forward distribution over Δ_m given by the mapping f_i acting on the marginal distribution of X, i.e., μ_i^X. $W_1(\cdot, \cdot)$ is the 1-Wasserstein distance under the ℓ_1 metric

[Xian et al., ICML' 23]

An Optimal Fair Classifier via Post-Processing

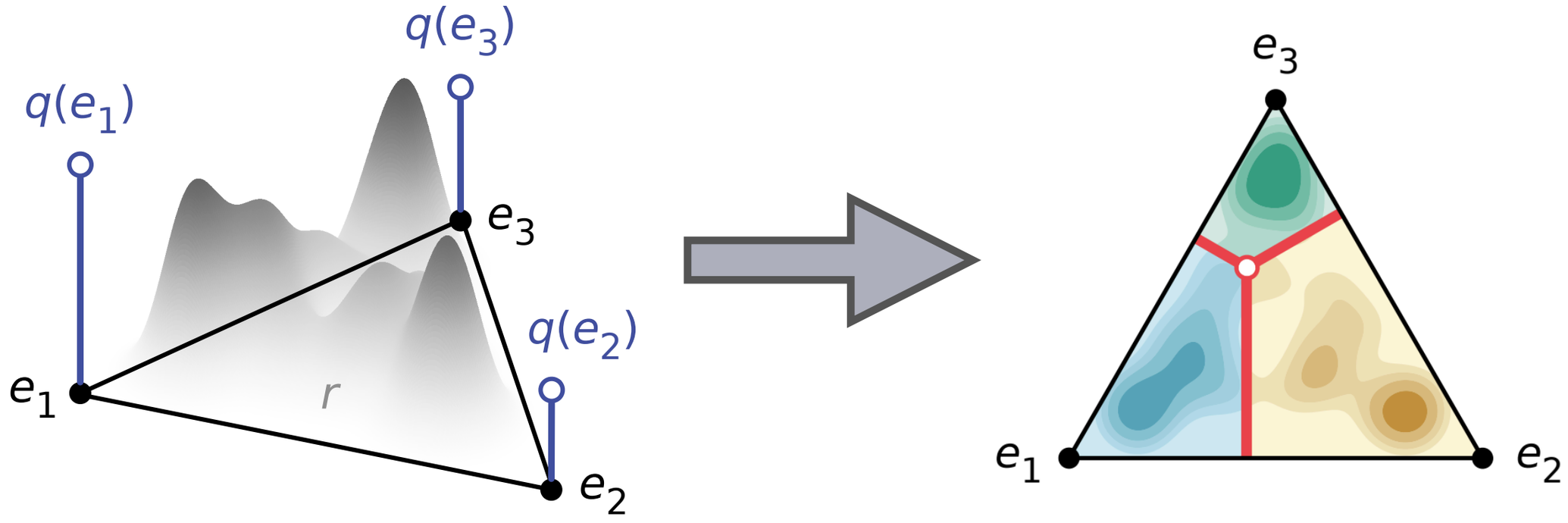
Price of fairness: (Wasserstein-Barycenter):

$$\min_{q} \quad \frac{1}{2}\sum_{i=1}^{n} W_1\left(f_i \sharp \mu_i^X, q\right)$$

$$\text{subjectto} \quad q \in \Delta_m$$

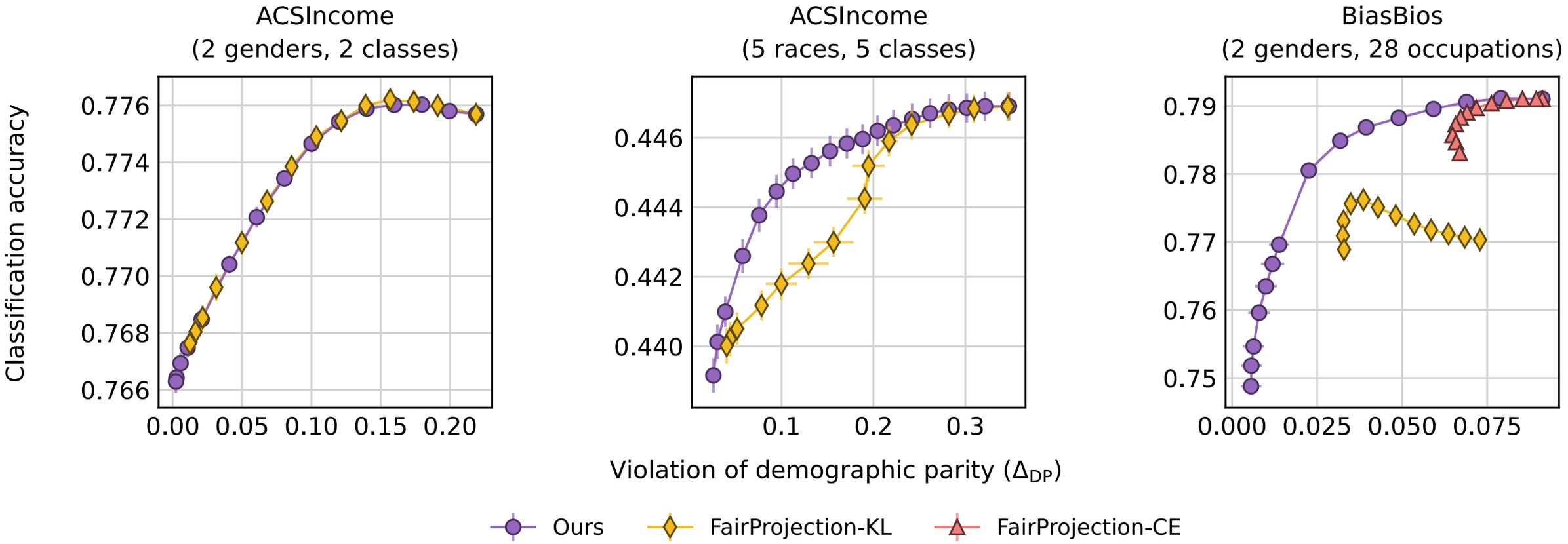
A two-step procedure:

1. Find the barycenter under the W_1 metric
2. Find the (randomized) transportation map to the barycenter as the post-processing map



[Xian et al., ICML' 23] Note: our algorithm also allows a relaxation of the exact fairness as well

Experiments



- FairProjection (Calmon et al. NeurIPS'22) is another post-processing method for fairness under the same setting
- FairProjection-KL works by minimizing the KL distance
- FairProjection-CE works by minimizing the reverse-KL distance
- Top-left points should be preferred (Pareto-optimal)

[Xian et al., ICML' 23]

Summary

Inherent tradeoffs by enforcing statistical parity under different settings:

Table 1: Characterizations of the inherent tradeoff of (strict) DP fairness.

Problem Setting	Minimum Risk Under DP
[Chzhen et al. NeurIPS' 20]: Regression	$\text{excess MSE} = \min_{q:\text{supp}(q)\subseteq\mathbb{R}} \sum_{a\in\mathcal{A}} w_a W_2^2(r_a^*, q)$ (1)
[Zhao et al. JMLR' 22]: Classification (Noiseless Setting)	$\text{excess} = \text{min. error} = \min_{q:\text{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \sum_{a\in\mathcal{A}} \frac{w_a}{2}\|p_a - q\|_1$ (2)
[Xian et al. ICML' 23]: Classification (General Setting)	$\text{minimum error} = \min_{q:\text{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \sum_{a\in\mathcal{A}} \frac{w_a}{2} W_1(r_a^*, q)$ (3)

- Attribute-aware post-processing is sufficient to achieve the optimal fair prediction, under both regression and classification settings
- Randomization is a powerful tool to enable the construction of the optimal fair predictor
- The prices of fairness are characterized by the barycenter problems under different metrics

Thanks

Q & A

Ruicheng Xian

Meta

DARPA

IBM Research

amazon

Code: https://github.com/rxian/fair-classification

Papers:

1. Inherent Tradeoffs in Learning Fair Representations, JMLR' 22b
2. Fundamental Limits and Tradeoffs in Invariant Representation Learning, JMLR' 22a
3. Fair and Optimal Classification via Post-Processing, ICML' 23