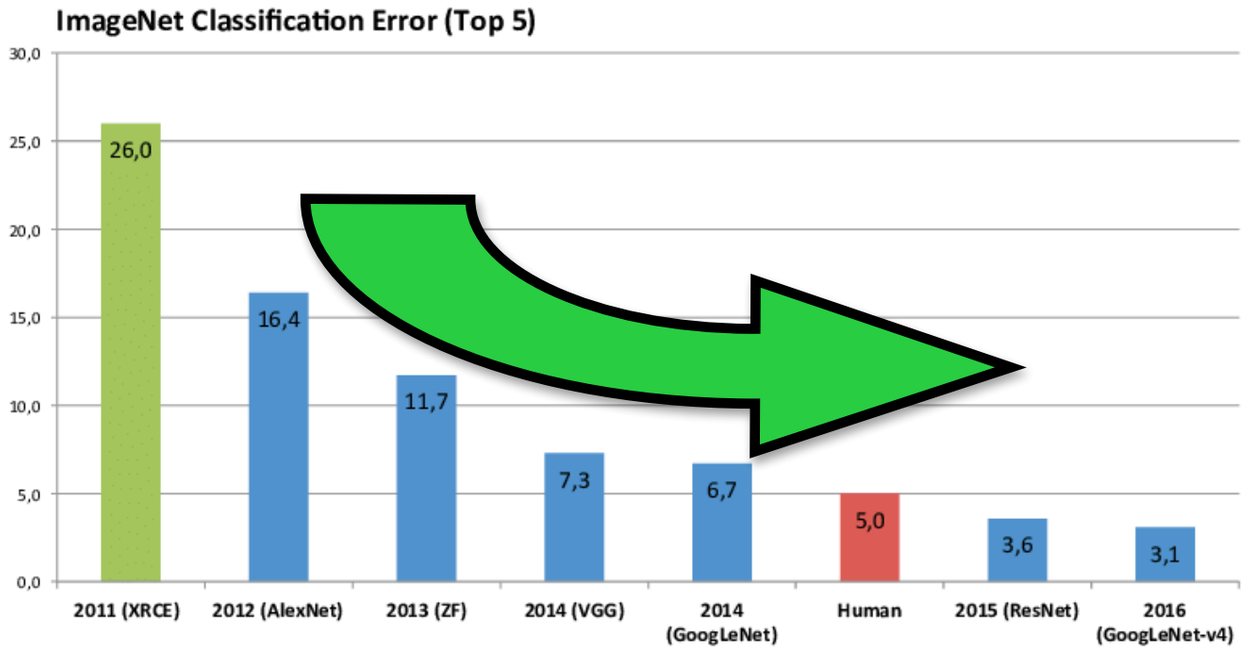
Algorithmic Fairness: Algorithms and Theory

Machine Learning Summer School, Okinawa

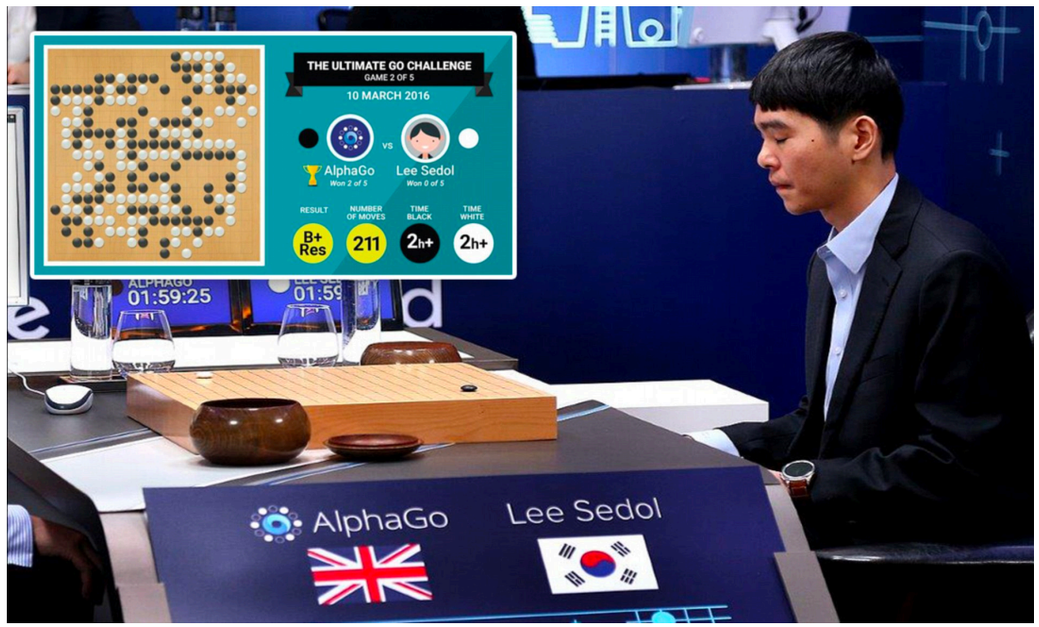
Mar. 7th, 2024

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

Success of Machine Learning

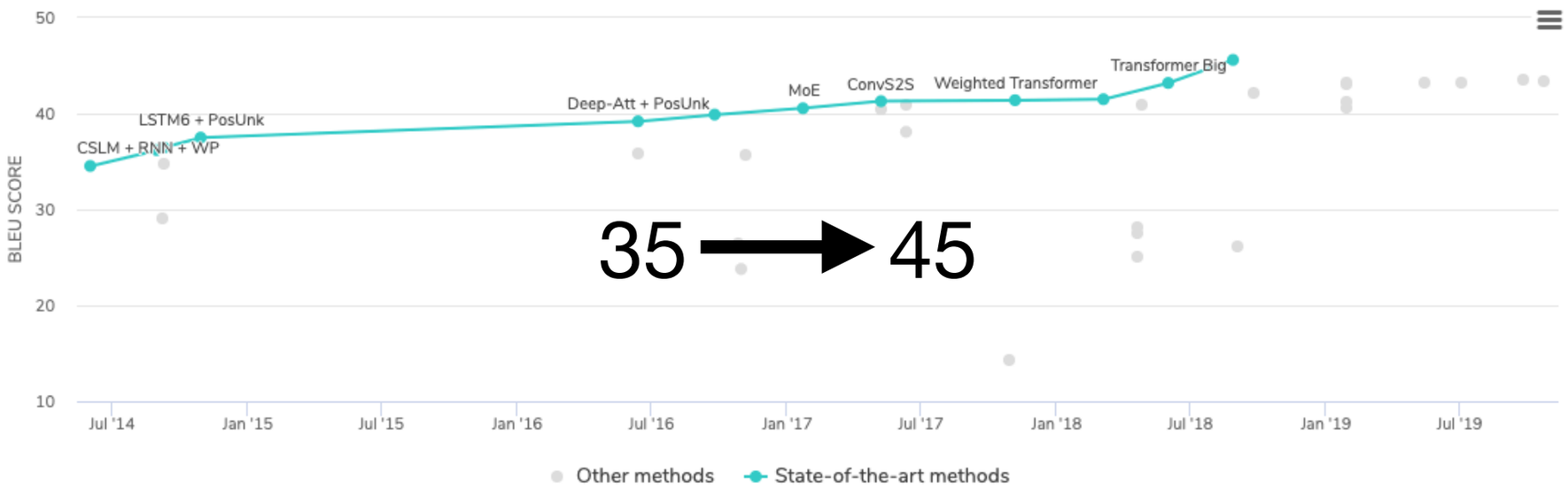


ImageNet: ~1M images, ~1K classes [Deng et al. 09]

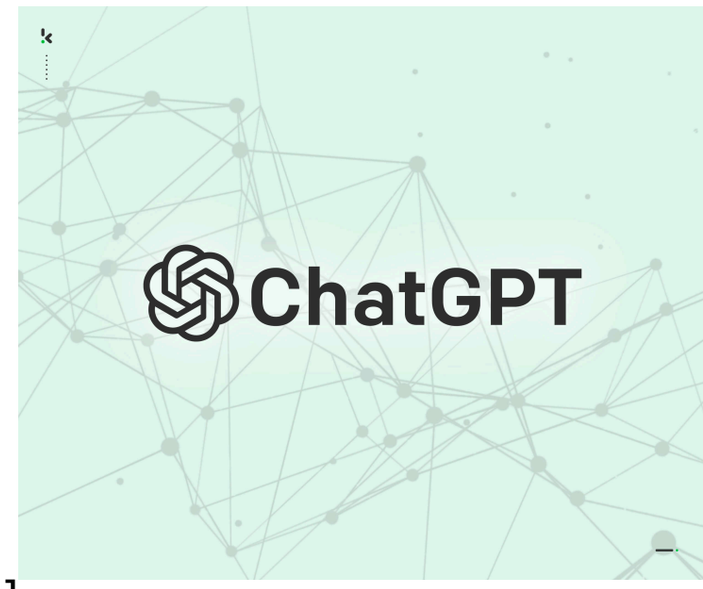


AlphaGo vs Lee Sedol [Silver et al.'16]

Machine Translation on WMT2014 English-French



Machine Translation, ~3M parallel sentences [Cho et al.'14; Devlin et al.'14]

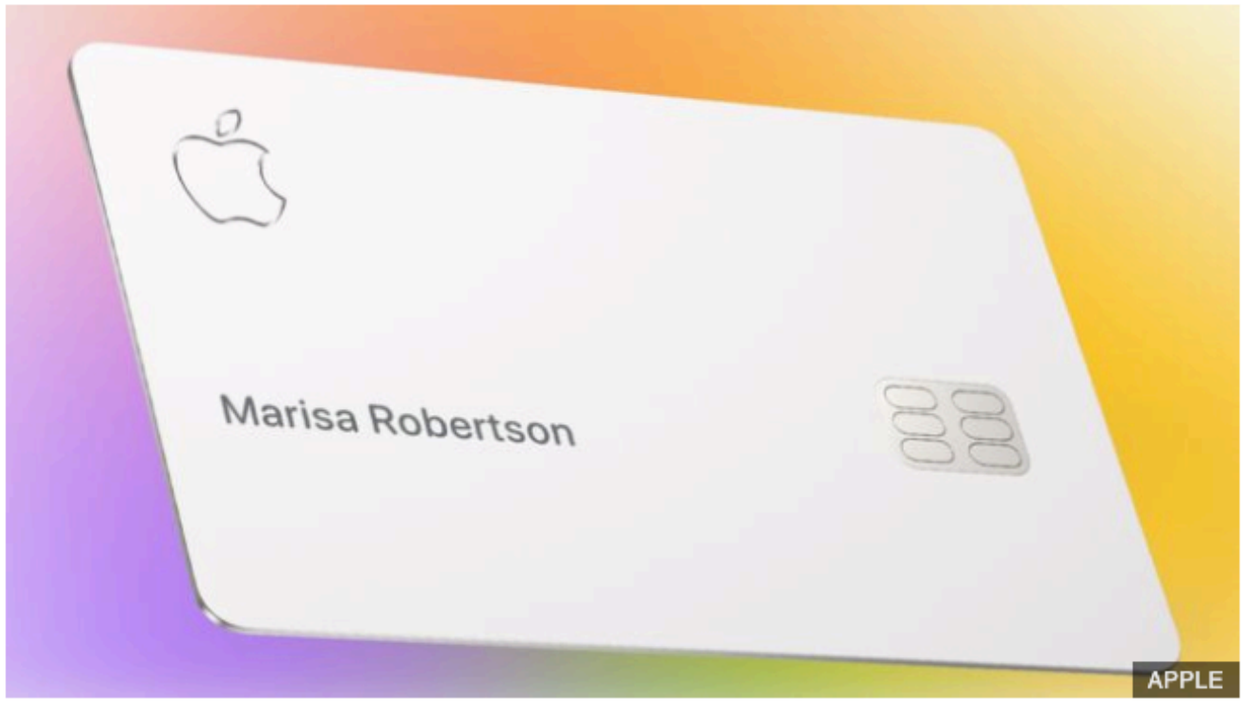


ChatGPT [OpenAI' 22]

Potential Bias of Data in High-stakes Domains

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

11 November 2019 Share



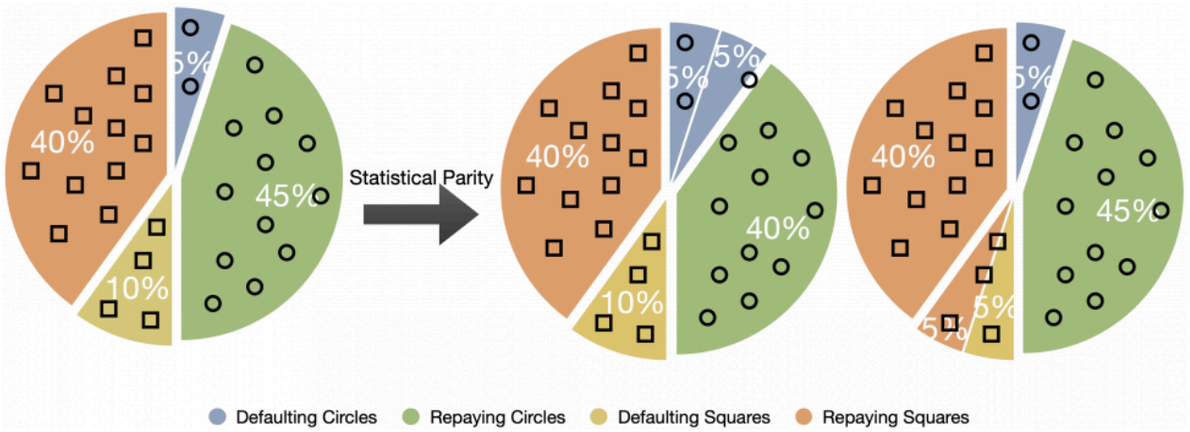
A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

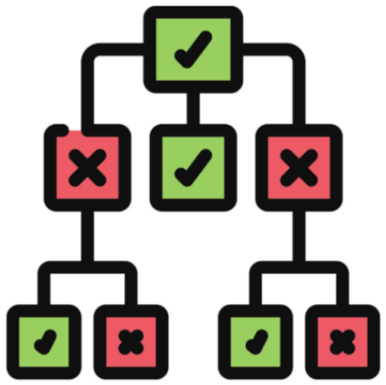
TECHNOLOGY NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

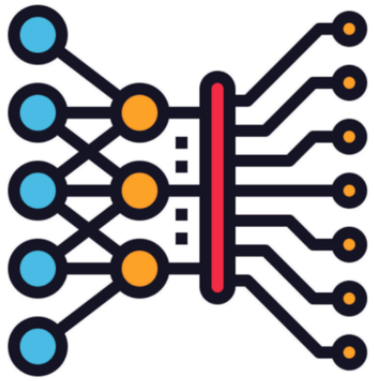
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

Trustworthy Machine Learning







Efficiency & Accuracy

Four aspects:

- Algorithmic Fairness
- (Distributional/Adversarial) Robustness
- Differential Privacy
- Model/Feature Interpretability

Robustness

Fairness

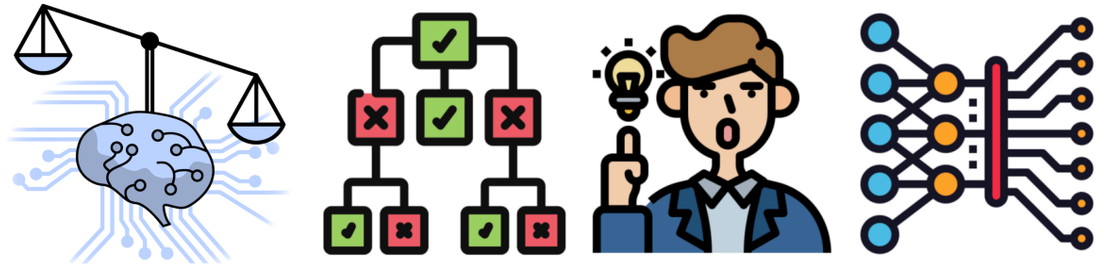
Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Trustworthy Machine Learning



Efficiency & Accuracy

Interpretability

Privacy

Robustness:

- Domain adaptation / generalization / Out-of-distribution generalization / transfer learning
- Invariant representation learning / invariant causal predictors

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Learning Fair Representations
Tradeoff between fairness & accuracy

Fairness

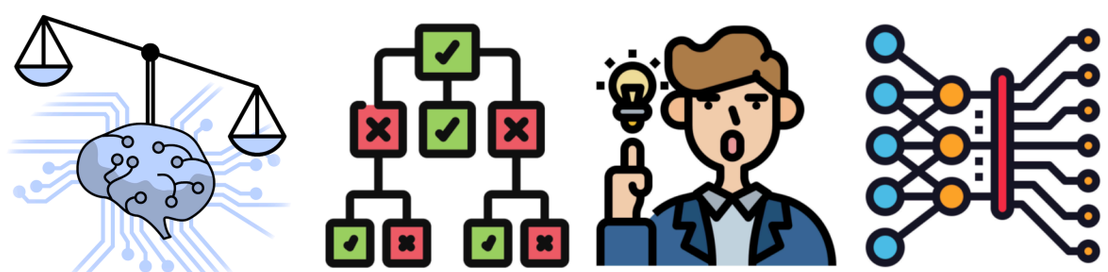
Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Efficiency & Accuracy

Interpretability

Privacy

Fairness:

- Learning fair presentations
- Trade-off between different notions of fairness and accuracy
- Interplay between robustness and fairness

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Learning Fair Representations
Tradeoff between fairness & accuracy

Fairness

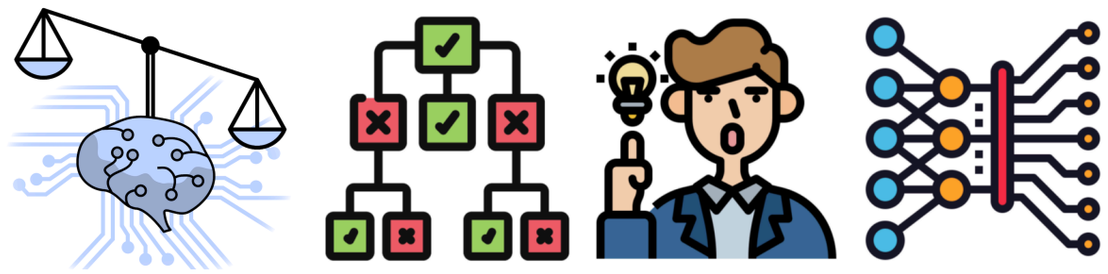
Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Efficiency & Accuracy

Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Interpretability

Privacy

Privacy:

- Fairness-Privacy-Accuracy tradeoff in classification/regression
- Machine unlearning
- Applications of DP in graph neural networks

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

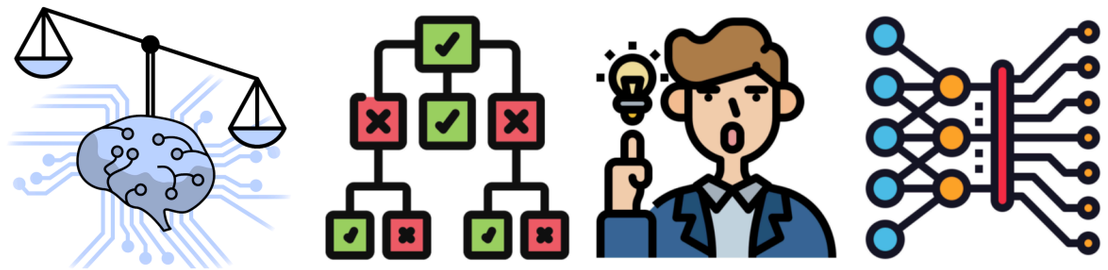
Fairness

Learning Fair Representations
Tradeoff between fairness & accuracy

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Efficiency & Accuracy

Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Maximum Influence Subset
Understand and select out the subset of data with maximum influence to the learned model

Structured Representations
Learning representations with class hierarchical information

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Interpretability

Privacy

Interpretability:

- Influence function / maximum influence subset identification, Shapley values
- Structured representations
- Understanding the internals of LLMs

How and why ML models could be *unfair*?

Case Study 1: Recidivism prediction

COMPAS (Northpointe):

Recidivism risk assessment tool used in a county in Florida

Goal: predict whether a defendant would commit a crime or not in 2 years if released?

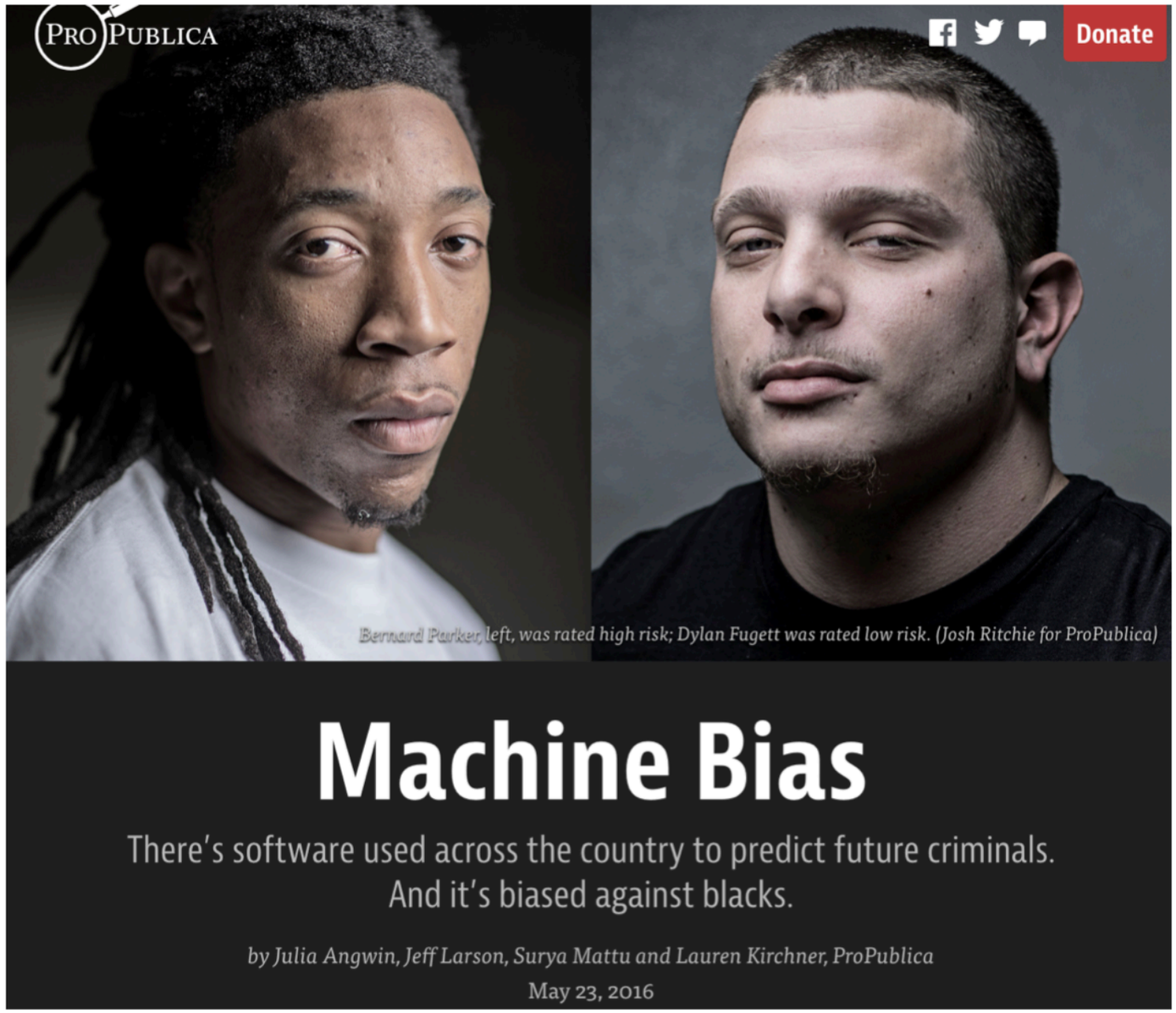
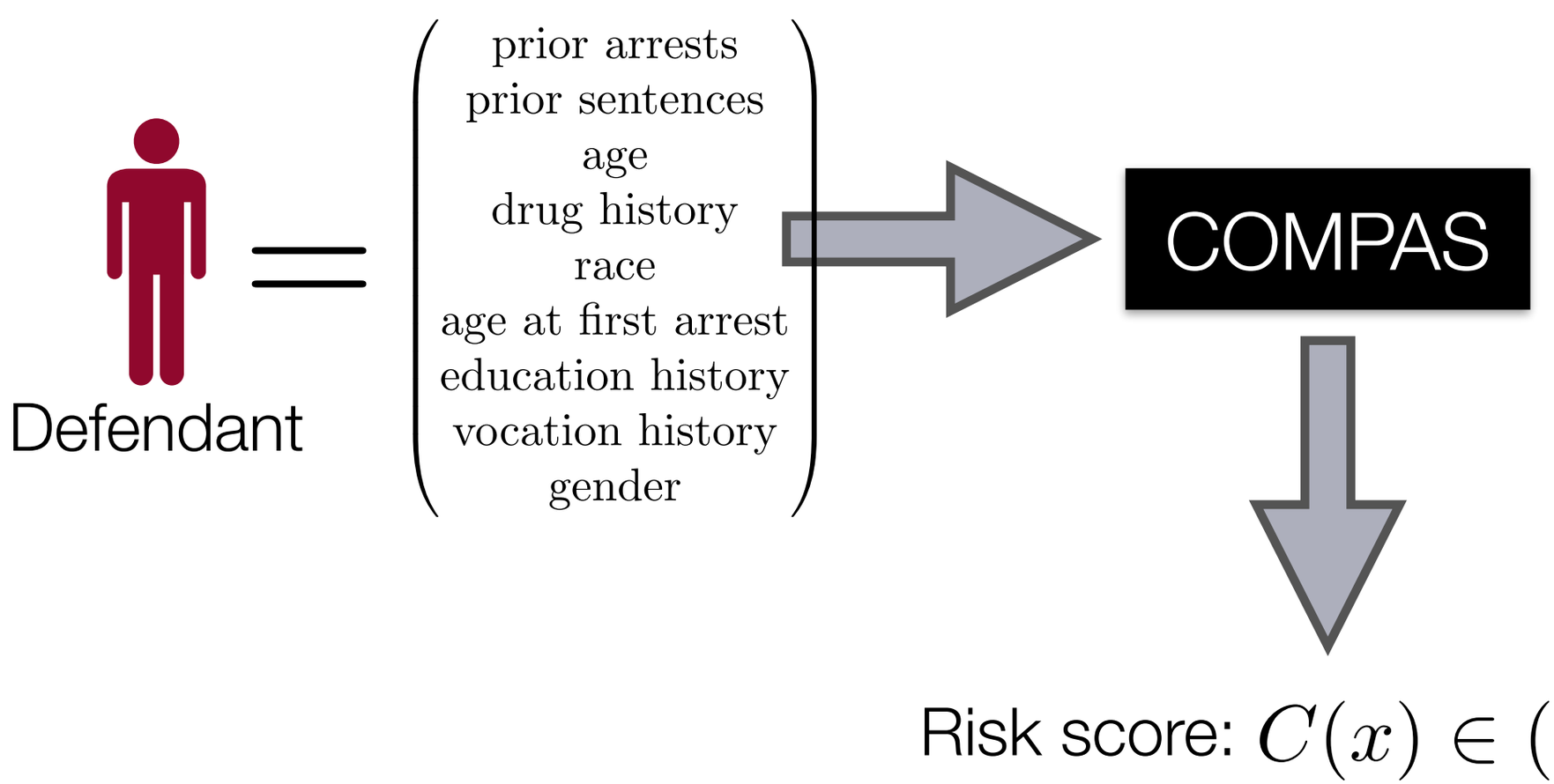


Figure credit: ProPublica, Larson et al., 2016

How and why ML models could be *unfair*?

Case Study 1: Recidivism prediction

COMPAS (high level):

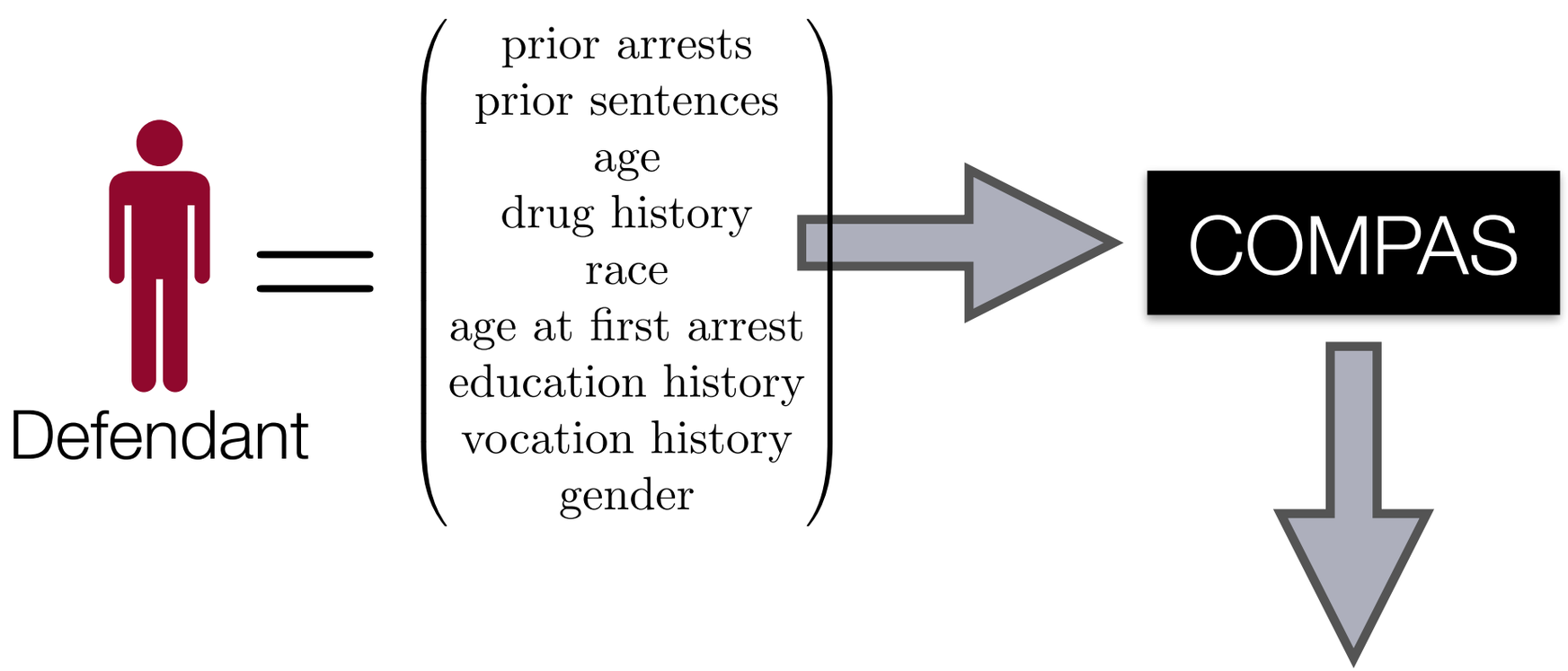


- Risk score ~ likelihood of defendant to recidivate
- Inputs have (noisy) true label: 0 (not recidivate) / 1 (will recidivate)
- The risk score + thresholding: 0 (low risk) / 1 (high risk)

How and why ML models could be *unfair*?

Case Study 1: Recidivism prediction

COMPAS (high level):



COMPAS is well calibrated (the prediction is correct on average):

$$\forall c \in (0,1), \;\; \Pr(Y = 1 \mid C(X) = c) = \mathbb{E}[Y \mid C(X) = c] = c$$

How and why ML models could be *unfair*?

Case Study 1: Recidivism prediction

ProPublica criticism:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

- Black defendants more likely than white to be **incorrectly** labeled "high risk"
- White defendants more likely than black to be **incorrectly** labeled "low risk"

Source: ProPublica, Larson et al., 2016

How and why ML models could be *unfair*?

Case Study 2: Face recognition

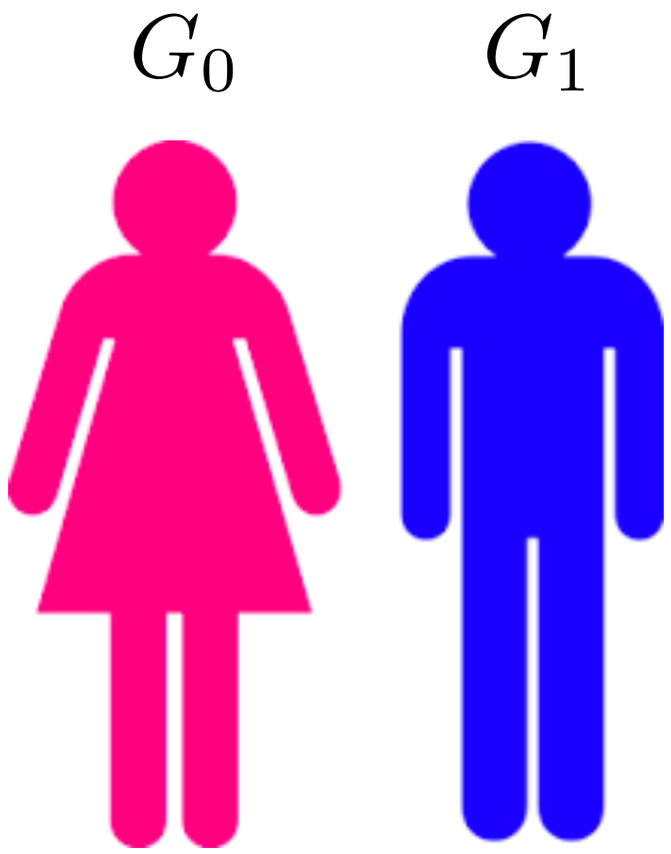
"Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Buolamwini and Gebru, FaccT 2018

Video link: https://www.youtube.com/watch?v=TWWsW1w-BVo

How and why ML models could be *unfair*?

Case Study 2: Face recognition

Some of the key findings from the paper:



- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Accuracy disparity:

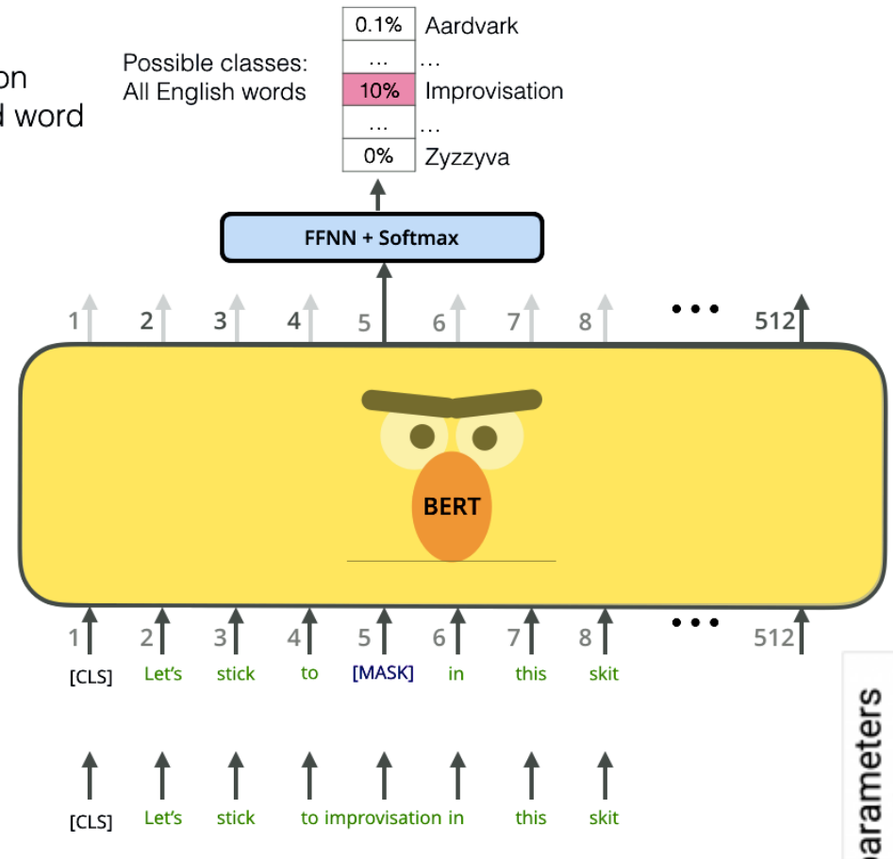
$$\Delta_{\mathrm{Err}}(h) := |\Pr(h(X) \neq Y \mid X \sim G_0) - \Pr(h(X) \neq Y \mid X \sim G_1)|$$

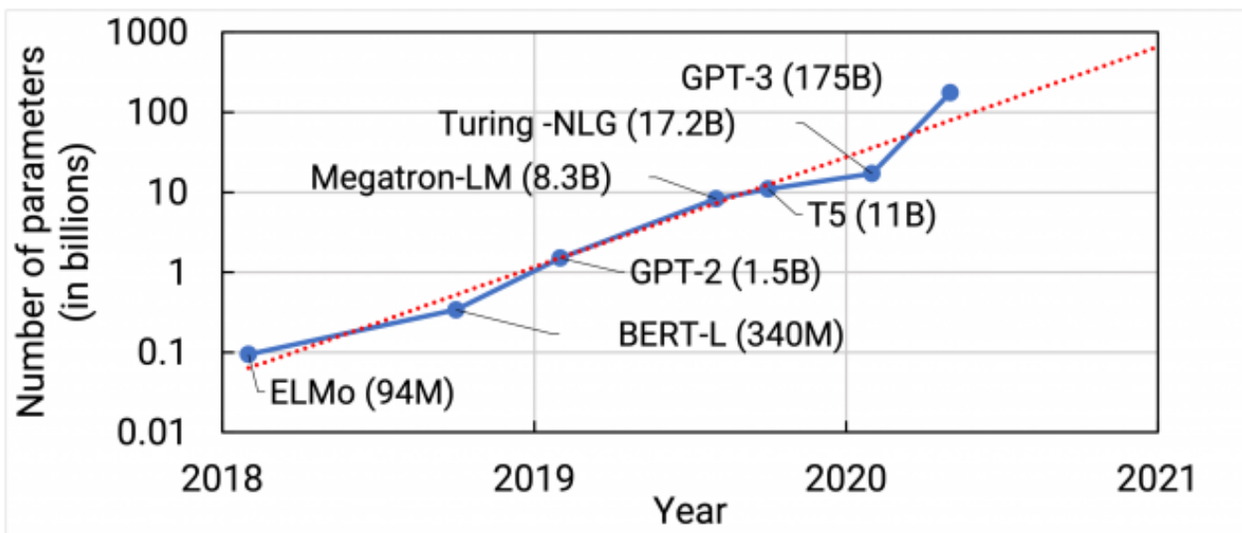
How and why ML models could be *unfair*?

Case Study 3: Contextual word-embeddings

Training large-scale language models

Use the output of the masked word's position to predict the masked word



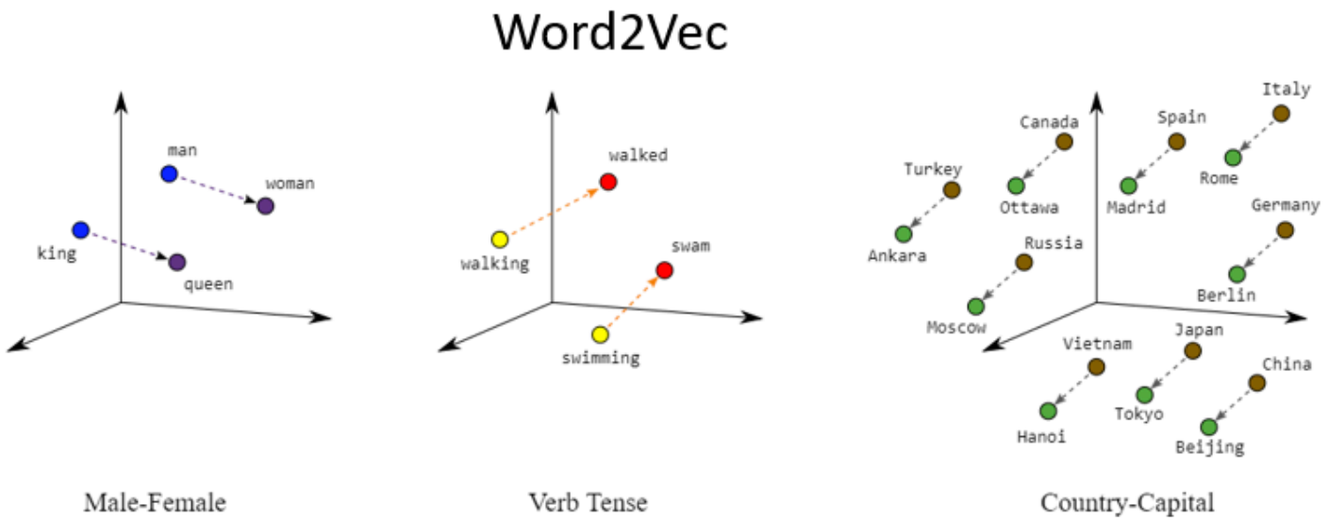


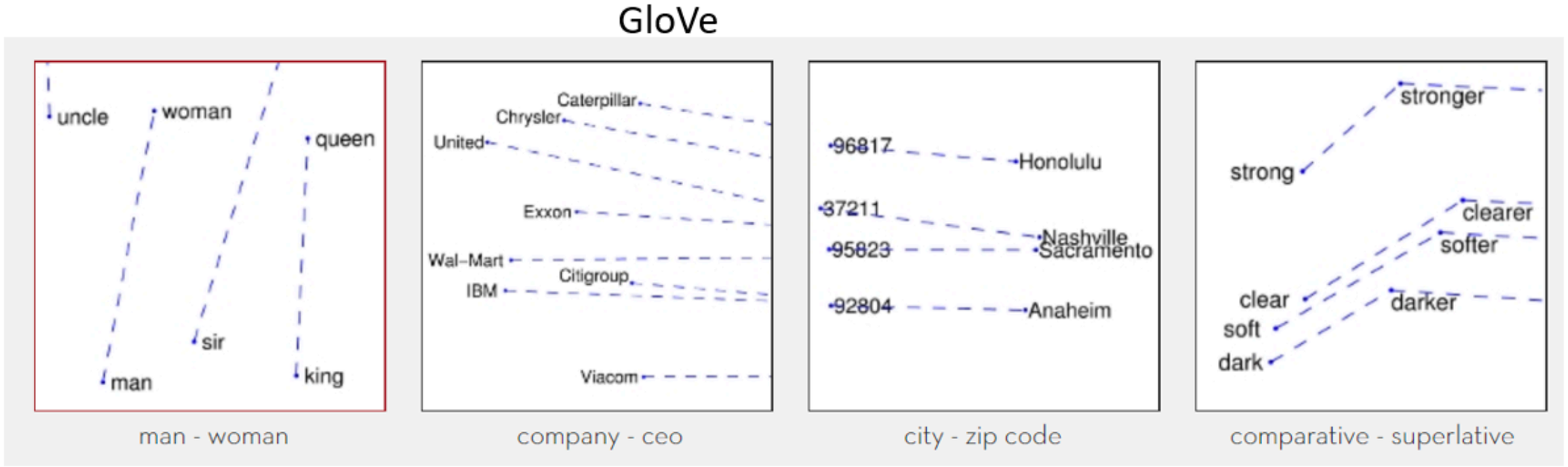
"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi, Chang, Zou, Saligrama, Kalai,

How and why ML models could be *unfair*?

Case Study 3: Contextual word-embeddings

Analogy relationship between the learned word-embeddings





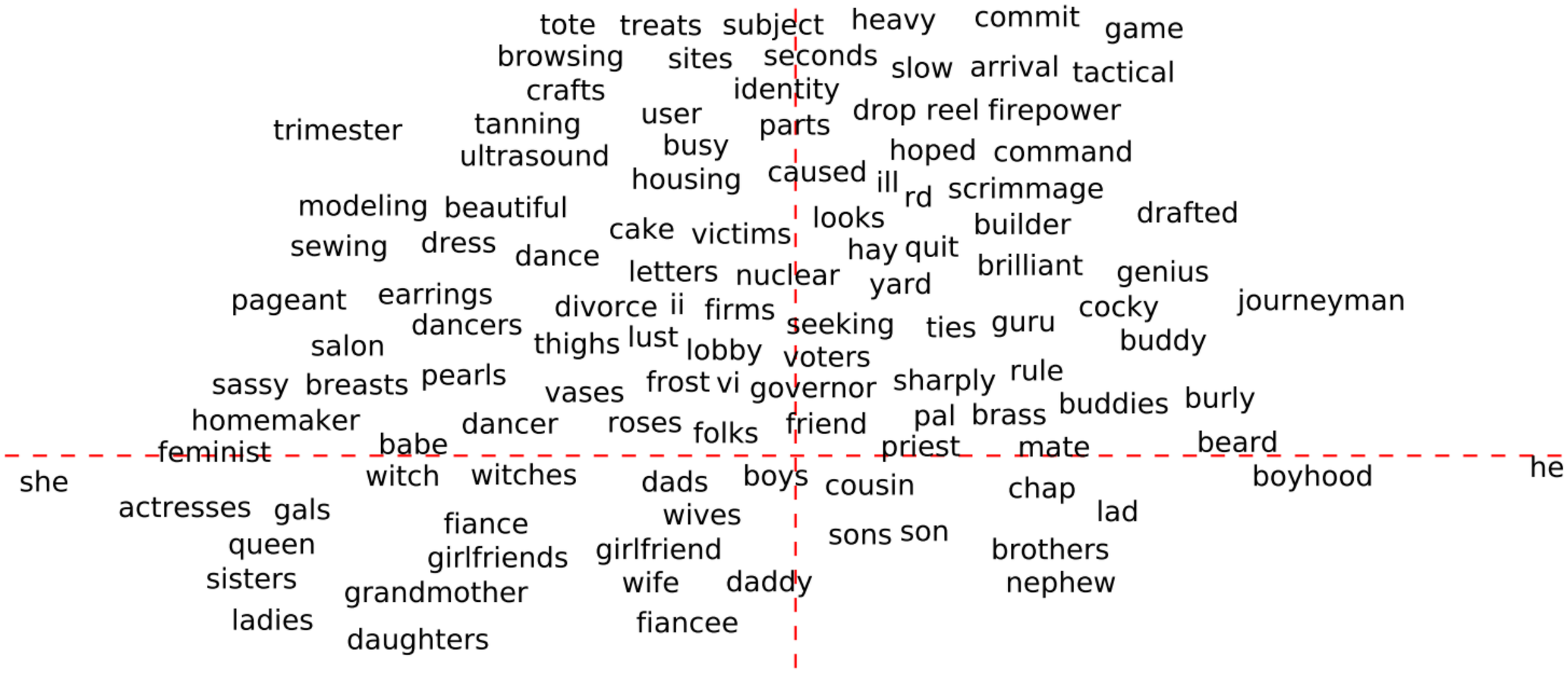
"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi, Chang, Zou, Saligrama, Kalai,

How and why ML models could be *unfair*?

Case Study 3: Contextual word-embeddings

However, biased analogy (stereotypes in the training corpus) also exists:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$



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

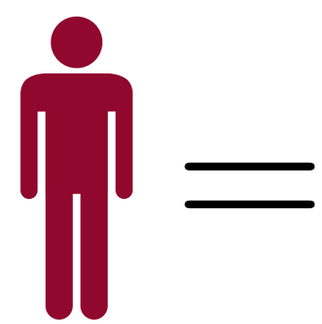
1. Toronto Argonauts have hit the ground running to open CFL season
2. 2023 NHL Draft: Final grades for all 32 teams
3. Blackhawks acquire Corey Perry from Lightning, adding more experience to Bedard-led rebuild
4. 2023 NHL Draft recap: Every pick made in Rounds 1-7
5. 'Pretty boring': Oilers ship Yamamoto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

Fairness Through Blindness

Statistical Parity

Ignorance is bliss?! — Thomas Gray

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



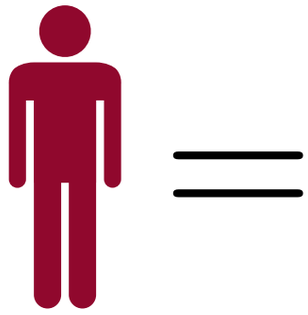
Defendant

$=$

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

Fairness Through Blindness

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



Defendant

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

Is this mechanism sufficient?

X	A	Y
0	1	1
1	0	1
1	0	0
0	1	1

X	$\neg X$	Y
0	1	1
1	0	1
1	0	0
0	1	1

Fairness Through Blindness

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

Defendant

$=$

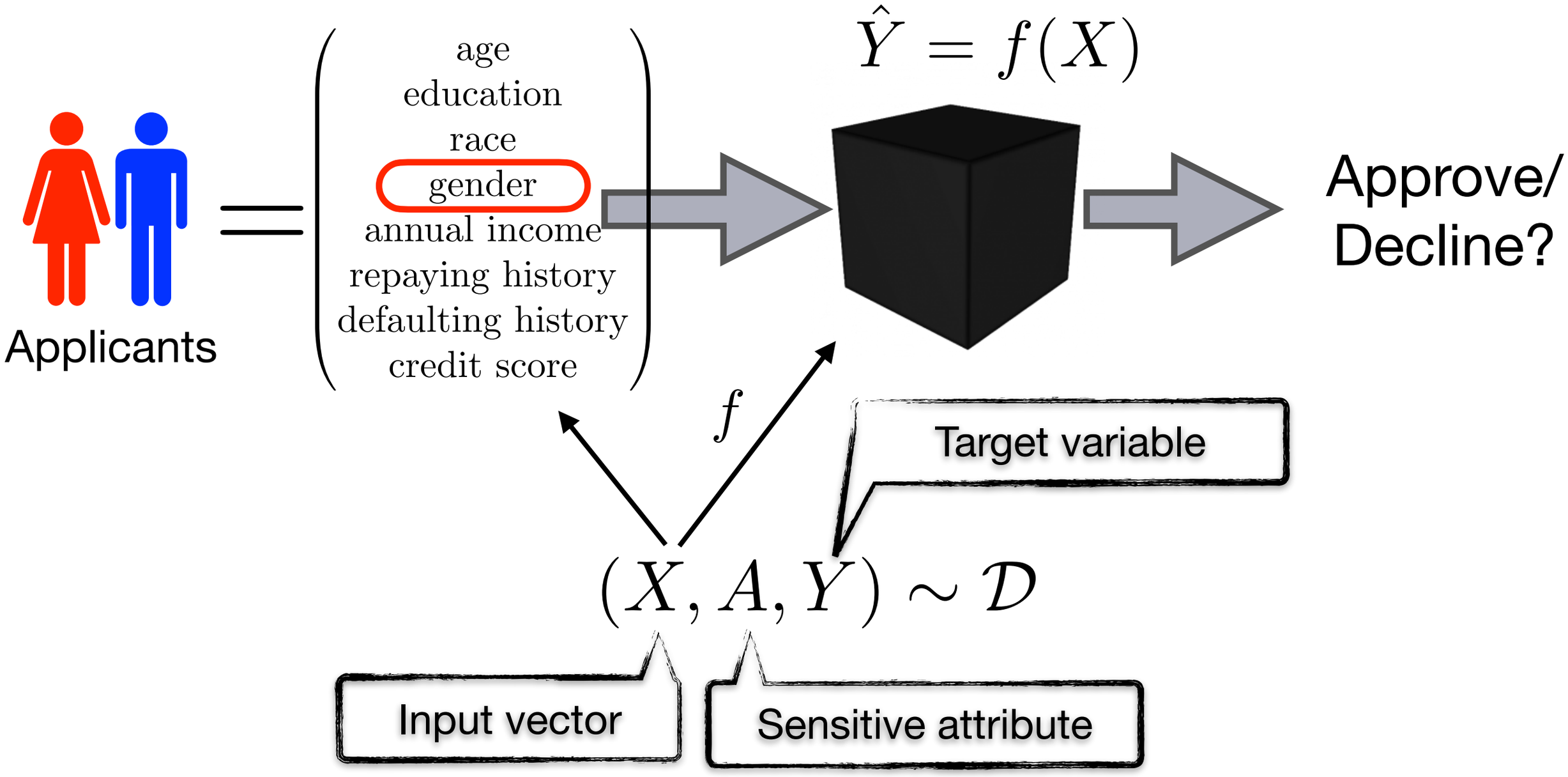
$$\begin{pmatrix} \text{prior arrests} \\ \text{prior sentences} \\ \text{age} \\ \text{drug history} \\ \cancel{\text{race}} \\ \text{age at first arrest} \\ \text{education history} \\ \text{vocation history} \\ \text{gender} \end{pmatrix}$$

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

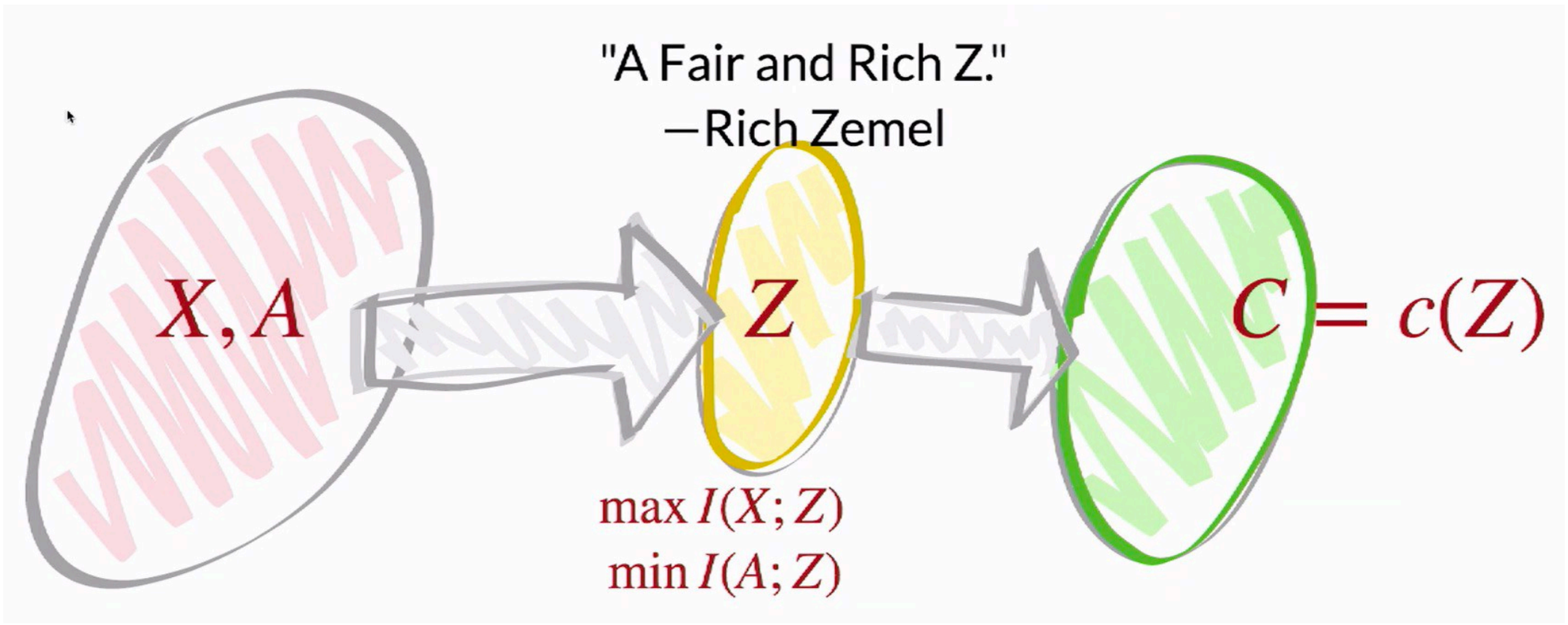
Pre-processing: Fair Representations

ICML 2013

Learning Fair Representations

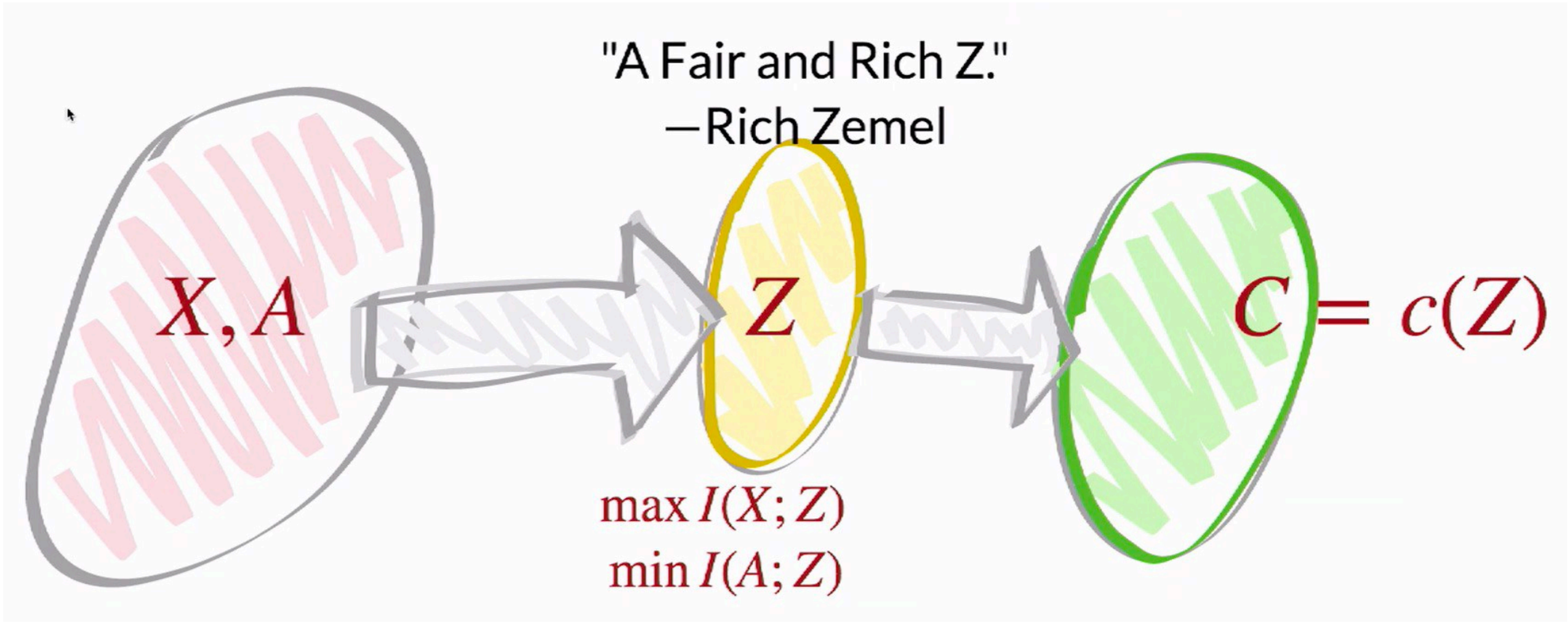
Richard Zemel ZEMEL@CS.TORONTO.EDU
Yu (Ledell) Wu WUYU@CS.TORONTO.EDU
Kevin Swersky KSWERSKY@CS.TORONTO.EDU
Toniann Pitassi TONI@CS.TORONTO.EDU
University of Toronto, 10 King's College Rd., Toronto, ON M6H 2T1 CANADA

Cynthia Dwork DWORK@MICROSOFT.COM
Microsoft Research, 1065 La Avenida Mountain View, CA. 94043 USA



Unsupervised learning, data pre-processing

Pre-processing: Fair Representations



Unsupervised learning, data pre-processing

How to utilize the target label Y?

$$\max I(Y;Z), \quad \min d_{\mathrm{TV}}(P,Q) \qquad P := \Pr_{A=0}(\cdot),\ Q := \Pr_{A=1}(\cdot)$$

$Z = g(X, A)$ is called the features/representations

Questions:

- How to learn the representations to ensure a small TV?
- Why is a small TV sufficient?

$$d_{\mathrm{TV}}(P,Q) := \sup_{Z \subseteq \mathcal{Z}} |P(Z) - Q(Z)|$$

Pre-processing: Fair Representations

Questions:

- Why is a small TV sufficient?

Hint: recall the data-processing inequality that we just learned

For any classifier $c : \mathcal{Z} \to \{0,1\}$ that acts on $\mathcal{Z}$, we know that

$$|\Pr_{A=0}(c(Z)=1) - \Pr_{A=1}(c(Z)=1)| \leq d_{\mathrm{TV}}(c_\sharp P, c_\sharp Q) \leq d_{\mathrm{TV}}(P,Q)$$

where $c_\sharp P$ is the induced distribution (pushforward) of P under c:

$$\forall E, c_\sharp P(E) := P(c^{-1}(E))$$

Implication: The gap of statistical parity could be bounded by the TV-distance between the representations

So, in order to ensure the outcome prediction to satisfy statistical parity, it is sufficient to minimize the TV distance

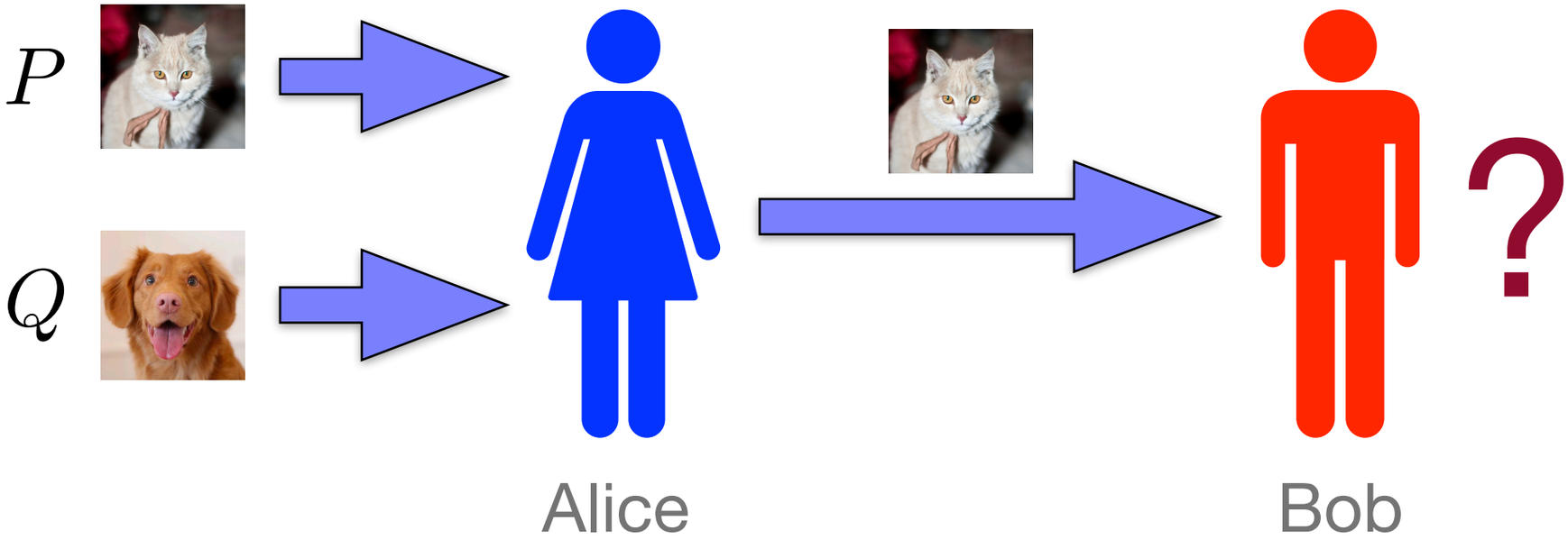
Pre-processing: Fair Representations

Questions:

- How to learn the representations to ensure a small TV?

Total Variation (TV) distance & the distinguishing game

TV distance: $d_{\mathrm{TV}}(P, Q) := \sup_{Z \subseteq \mathcal{Z}} |P(Z) - Q(Z)|$



Pre-processing: Fair Representations

Total Variation (TV) distance & the distinguishing game

TV distance: $d_{\mathrm{TV}}(P, Q) := \sup_{Z \subseteq \mathcal{Z}} |P(Z) - Q(Z)|$

Proposition:

Let P, Q be two probability distributions over the same space, then

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| =: \frac{1}{2} \|P - Q\|_1$$

Proof:

Let A^* be the event such that $d_{\mathrm{TV}}(P, Q) = P(A^*) - Q(A^*)$. Then

$$\forall x \in \mathcal{X} \backslash A^*, \; P(x) < Q(x)$$

Hence,

$$\begin{aligned}
\|P - Q\|_1 &= \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \\
&= \sum_{x \in A^*} (P(x) - Q(x)) + \sum_{x \in \mathcal{X} \backslash A^*} (Q(x) - P(x)) \\
&= (P(A^*) - Q(A*)) + (1 - Q(A^*) - (1 - P(A^*))) \\
&= 2\,(P(A^*) - Q(A*)) = 2 d_{\mathrm{TV}}(P, Q)
\end{aligned}$$

Pre-processing: Fair Representations

Interpretation of the TV-distance as a binary classification problem:

$$\eta(x) := \Pr(\text{my guess is } P \mid x)$$

Then, the error probability of using strategy $\eta(\cdot)$ when seeing x is:

$$\begin{aligned}\Pr(\text{error}, x) &= \Pr(\text{guess } P, x \sim Q) + \Pr(\text{guess } Q, x \sim P) \\ &= \frac{1}{2}Q(x)\eta(x) + \frac{1}{2}P(x)(1-\eta(x))\end{aligned}$$

Hence, the overall error probability is given by:

$$\begin{aligned}\Pr(\text{ error }) &= \sum_{x\in\mathcal{X}} \Pr(\text{ error }, x) \\ &= \sum_{x\in\mathcal{X}} \frac{1}{2}Q(x)\eta(x) + \frac{1}{2}P(x)(1-\eta(x)) \\ &= \frac{1}{2} + \frac{1}{2}\sum_{x\in\mathcal{X}} \eta(x)(Q(x)-P(x))\end{aligned}$$

Pre-processing: Fair Representations

Interpretation of the TV-distance as a binary classification problem:

$$\eta(x) := \Pr(\text{my guess is } P \mid x)$$

Hence, the overall error probability is given by:

$$\Pr(\text{error}) = \frac{1}{2} + \frac{1}{2} \sum_{x \in \mathcal{X}} \eta(x) \left(Q(x) - P(x)\right)$$

So, clearly, to minimize the overall guessing error, the best strategy is given by:

$$\eta(x) = \begin{cases} 1 & P(x) \geq Q(x) \\ 0 & P(x) < Q(x) \end{cases}$$

Under this optimal strategy, the optimal distinguishing error is:

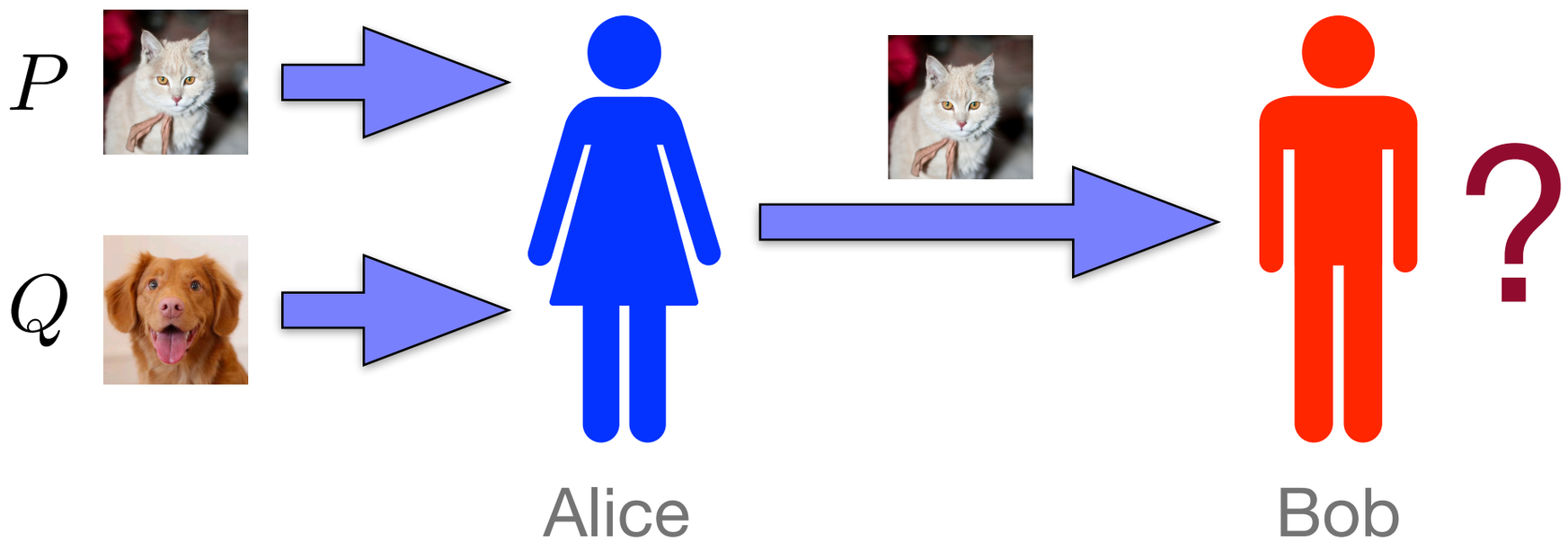
$$\begin{aligned} \Pr(\text{error}) &= \frac{1}{2} + \frac{1}{2} \sum_{x : P(x) \geq Q(x)} Q(x) - P(x) \\ &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(P, Q) \end{aligned}$$

Consider some extremal cases of this distinguishing game

Pre-processing: Fair Representations

Total Variation (TV) distance & the distinguishing game

TV distance: $d_{\mathrm{TV}}(P, Q) := \sup_{Z \subseteq \mathcal{Z}} |P(Z) - Q(Z)|$



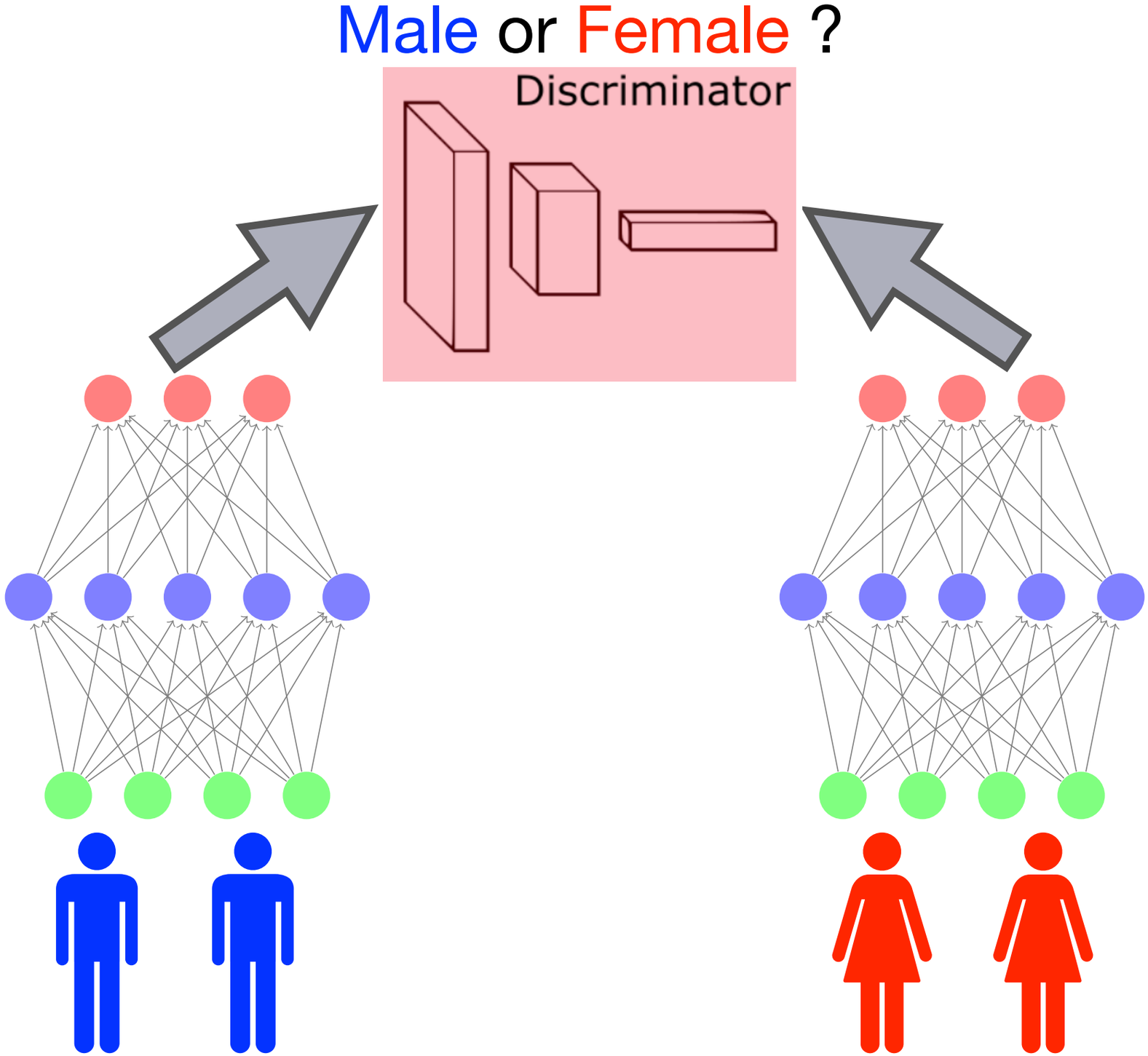
Bob's optimal error: $\Pr(\text{error}) = \frac{1}{2} + \frac{1}{2} \sum_{x: P(x) \geq Q(x)} Q(x) - P(x)$

$= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(P, Q)$

Want the TV to be small!

Pre-processing: Fair Representations

Adversarial Training



Fair Representations

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

Pre-processing: Fair Representations

How to learn the fair representations for supervised learning?

- Goal 1: Learn discriminative features for the target task of interest
- Goal 2: Learn fair features to confuse the adversarial discriminator

$$\min_{\theta_f,\theta_y} \max_{\theta_d} L_y\left(G_y\left(G_f\left(X;\theta_f\right);\theta_y\right),Y\right) - \lambda \cdot L_d\left(G_d\left(G_f\left(X;\theta_f\right);\theta_d\right),A\right)$$

- $G_f(\cdot;\theta_f)$: feature transformation with tunable parameter θ_f
- $G_y(\cdot;\theta_y)$: classifier over feature space for the task of interest
- $G_d(\cdot;\theta_d)$: adversarial discriminator with tunable parameter θ_d

λ controls the relative importance of these two goals

Pre-processing: Fair Representations

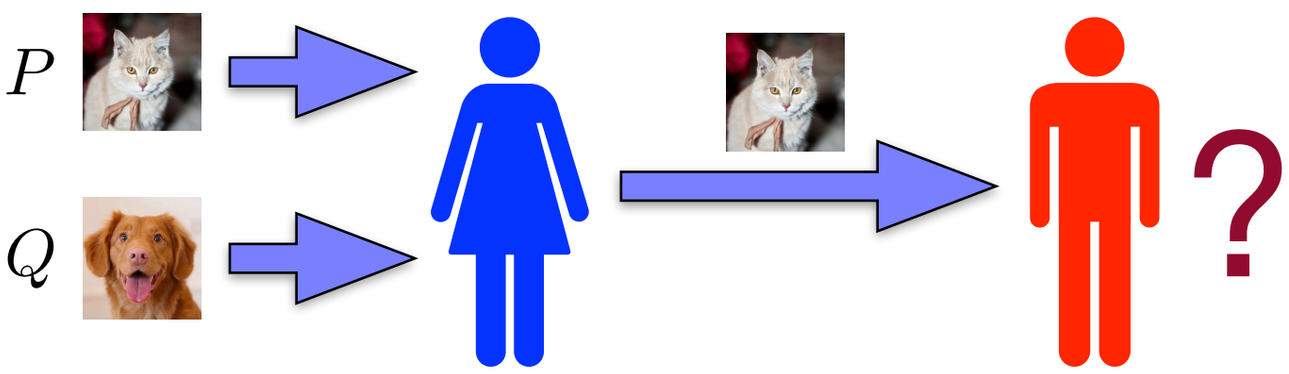
Objective function:

$$\min_{\theta_f,\theta_y} \max_{\theta_d} L_y\left(G_y\left(G_f\left(X;\theta_f\right);\theta_y\right),Y\right) - \lambda \cdot L_d\left(G_d\left(G_f\left(X;\theta_f\right);\theta_d\right),A\right)$$

Questions:

- Why the negative sign before λ?

$$L_d\left(G_d\left(G_f\left(X;\theta_f\right);\theta_d\right),A\right) \approx \mathbb{E}_{A=0}[G_d\left(G_f\left(X;\theta_f\right);\theta_d\right)=1]$$
$$+\mathbb{E}_{A=1}[G_d\left(G_f\left(X;\theta_f\right);\theta_d\right)=0]$$



$$\begin{aligned}\Pr(\text{error}) &= \frac{1}{2} + \frac{1}{2}\sum_{x:P(x)\geq Q(x)} Q(x) - P(x) \\ &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(P,Q)\end{aligned}$$

Pre-processing: Fair Representations

Stochastic Gradient Descent-Ascent algorithm:

Algorithm 1 Stochastic Gradient Descent Ascent

Require: Initial parameter of feature extractor $\theta_f^{(0)}$, parameter of target classifier $\theta_y^{(0)}$, parameter of adversarial discriminator $\theta_d^{(0)}$

for $t = 1, 2, \ldots$ until convergence **do**

  Sample a batch of data $\{x_j\}_{j=1}^B$ of size B from group $A = 0$

  Sample a batch of data $\{x'_j\}_{j=1}^B$ of size B from group $A = 1$

  Compute all the gradients:

  $\nabla_y^{(t)} \leftarrow \nabla_{\theta_y} L_y(\theta_f^{(t-1)}, \theta_y^{(t-1)}; \{x_j\}_{j=1}^B, \{x'_j\}_{j=1}^B)$

  $\nabla_d^{(t)} \leftarrow \nabla_{\theta_d} L_d(\theta_f^{(t-1)}, \theta_d^{(t-1)}; \{x_j\}_{j=1}^B, \{x'_j\}_{j=1}^B)$

  $\nabla_{f \leftarrow y}^{(t)} \leftarrow \nabla_{\theta_f} L_y(\theta_f^{(t-1)}, \theta_y^{(t-1)}; \{x_j\}_{j=1}^B, \{x'_j\}_{j=1}^B)$

  $\nabla_{f \leftarrow d}^{(t)} \leftarrow \nabla_{\theta_f} L_d(\theta_f^{(t-1)}, \theta_d^{(t-1)}; \{x_j\}_{j=1}^B, \{x'_j\}_{j=1}^B)$

  Gradient descent over θ_y and θ_d:

  $\theta_y^{(t)} \leftarrow \theta_y^{(t-1)} - \gamma \nabla_y^{(t)}$

  $\theta_d^{(t)} \leftarrow \theta_d^{(t-1)} - \lambda\gamma \nabla_d^{(t)}$

  Gradient descent and ascent over θ_f:

  $\theta_f^{(t)} \leftarrow \theta_f^{(t-1)} - \gamma \nabla_{f \leftarrow y}^{(t)} + \lambda\gamma \nabla_{f \leftarrow d}^{(t)}$

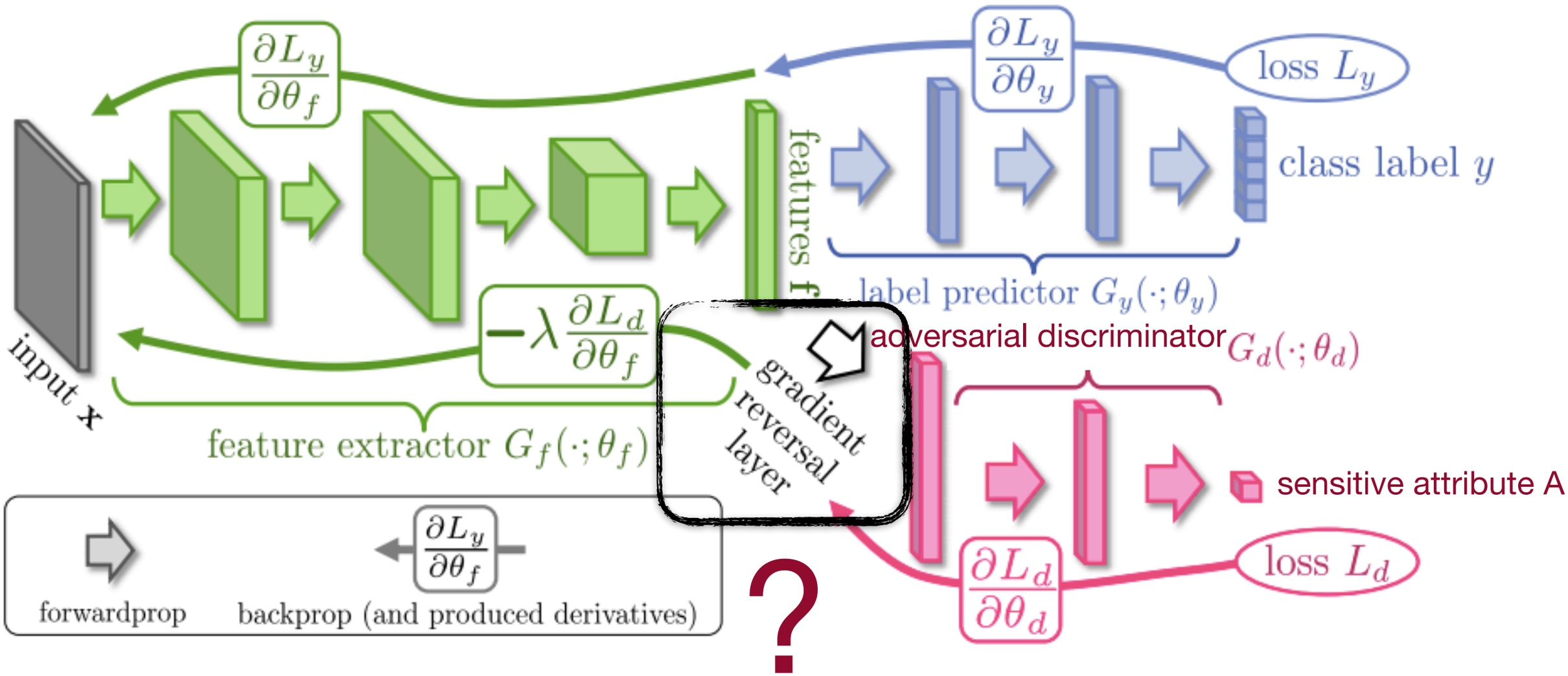
end for

return w^*

Pre-processing: Fair Representations

Objective function:

$$\min_{\theta_f, \theta_y} \max_{\theta_d} L_y\left(G_y\left(G_f\left(X;\theta_f\right);\theta_y\right), Y\right) - \lambda \cdot L_d\left(G_d\left(G_f\left(X;\theta_f\right);\theta_d\right), A\right)$$



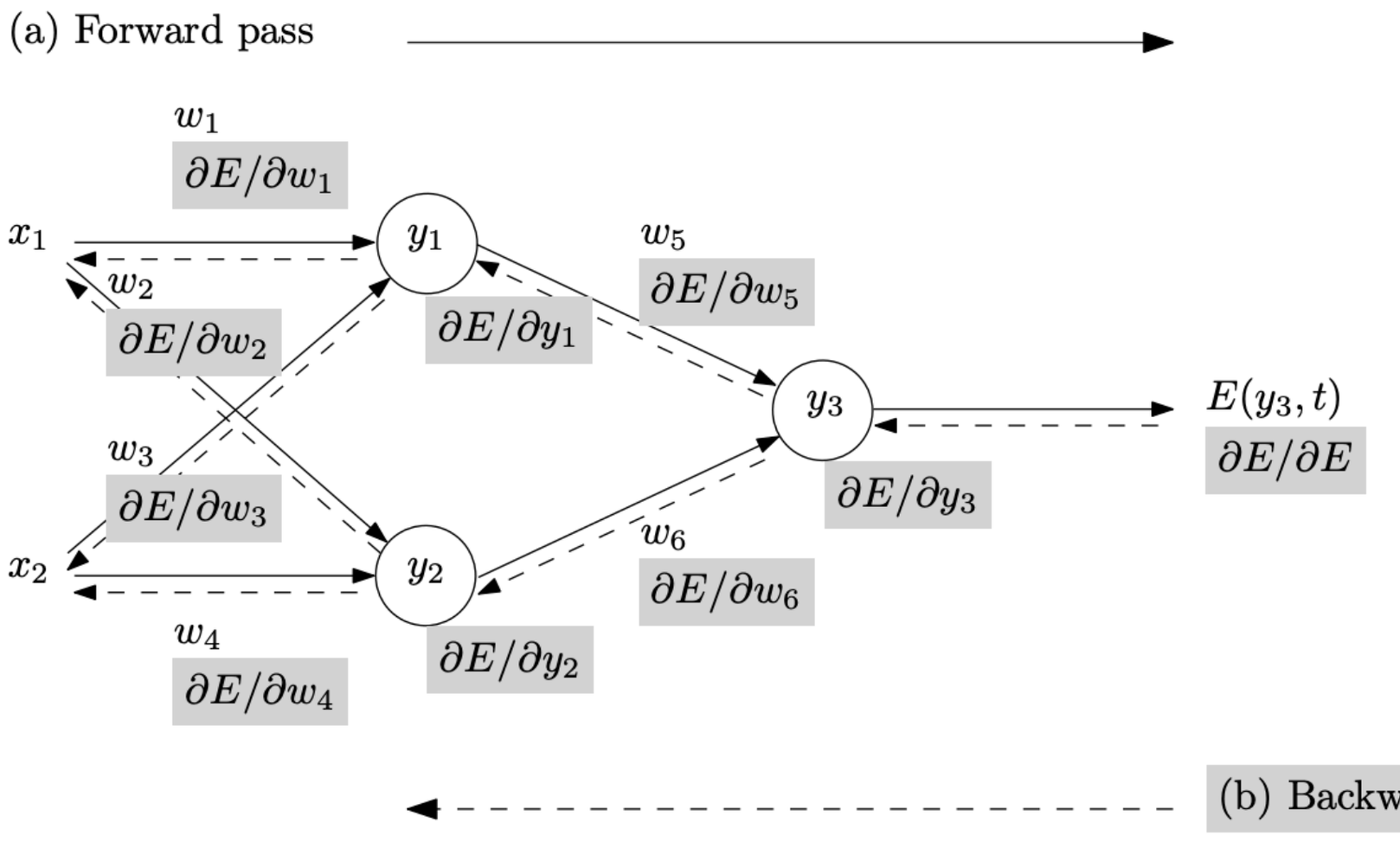
[Figure adapted from Ganin et al.' 16]

Pre-processing: Fair Representations

Autodiff:

TensorFlow PyTorch

Computational Circuits:



Pre-processing: Fair Representations

Basic module: computational node

```
class CompNode:
    def __init__(self, tape):
        # make sure that the gradient tape knows us
        tape.add(self)

    # perform the intended operation
    # and store the result in self.output
    def forward(self):
        pass

    # assume that self.gradient has all the information
    # from outgoing nodes prior to calling backward
    # -> perform the local gradient step with respect to inputs
    def backward(self):
        pass

    # needed to be initialized to 0
    def set_gradient(self, gradient):
        self.gradient = gradient

    # receive gradients from downstream nodes
    def add_gradient(self, gradient):
        self.gradient += gradient
```

Implement your logic of forward computation

Implement the gradient of the forward function

Pre-processing: Fair Representations

Gradient Reversal Layer

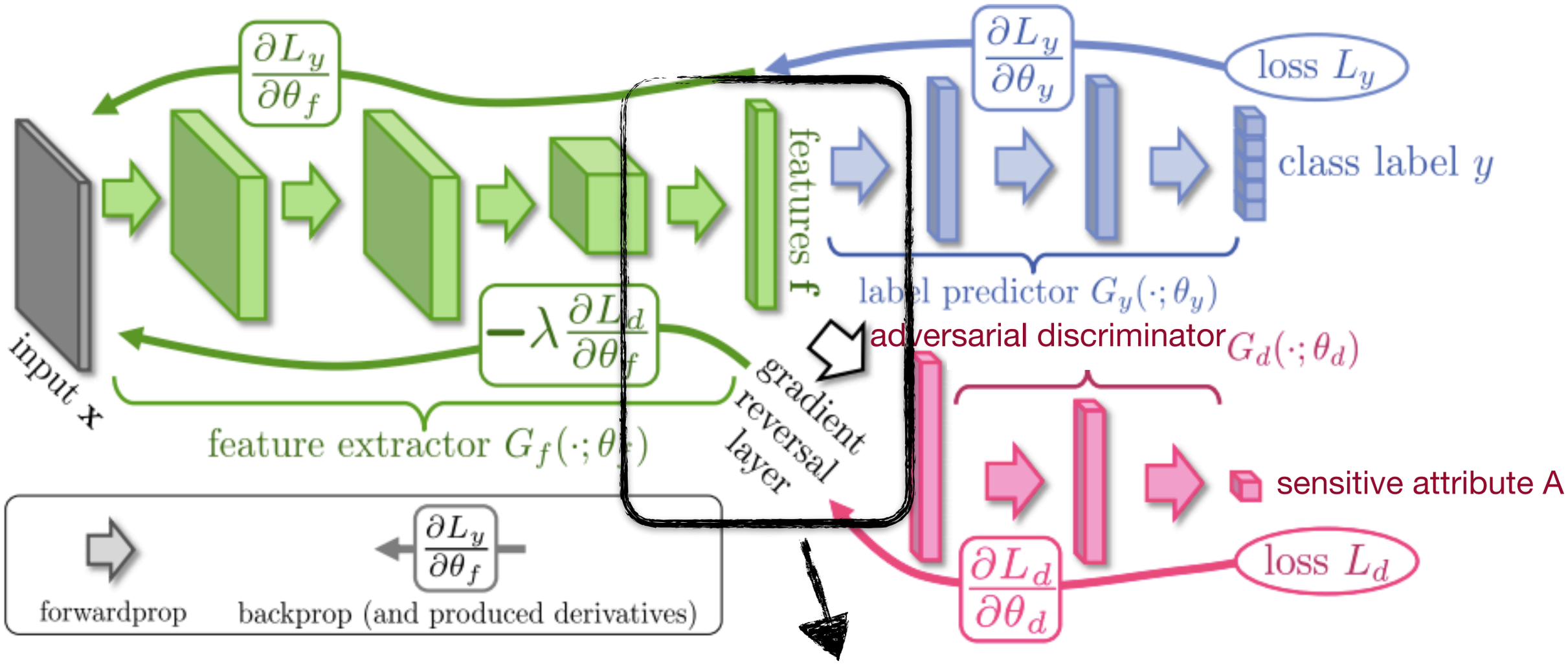
```
class GradientReversalLayer(torch.autograd.Function):
    """
    Implement the gradient reversal layer for the convenience of domain adaptation neural network.
    The forward part is the identity function while the backward part is the negative function.
    """
    def forward(self, inputs):
        return inputs

    def backward(self, grad_output):
        grad_input = grad_output.clone()
        grad_input = -grad_input
        return grad_input
```

- Forward: identity function
- Backward: -1 times identity

Pre-processing: Fair Representations

Gradient Reversal Layer

$$\min_{\theta_f,\theta_y} \max_{\theta_d} L_y\left(G_y\left(G_f\left(X;\theta_f\right);\theta_y\right),Y\right) - \lambda \cdot L_d\left(G_d\left(G_f\left(X;\theta_f\right);\theta_d\right),A\right)$$

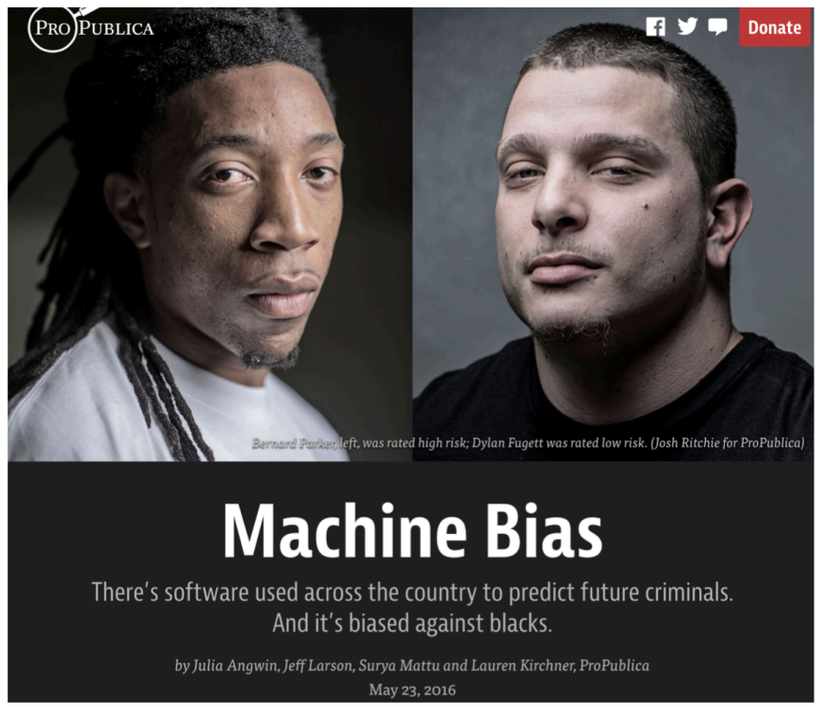


Gradient from the adversarial discriminator's loss to the feature extractor will be reversed.

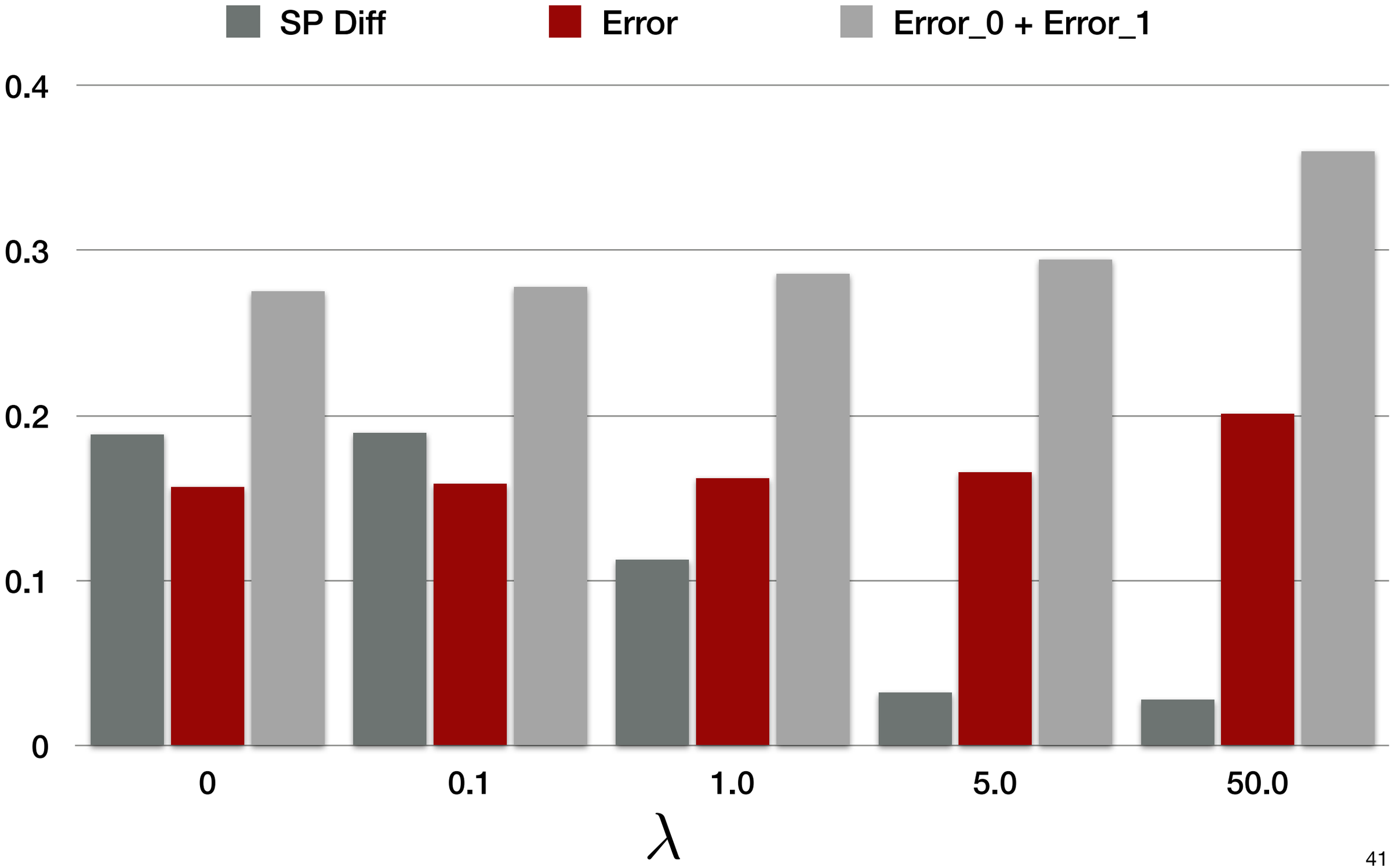
Experiment: Recidivism Prediction

COMPAS

- Train/Test: 4,320/1,852 instances from the Northpointe
- Target task: 0/1 classification (recidivism?)
- Sensitive attribute: race (Black/White)
- Other attributes: gender, education, prior arrest history, … (12 total)
- Difference of base rate: $\Delta_{\mathrm{BR}} = 0.129$



Experiment: Recidivism Prediction



Pre-processing: Fair Representations

Pros and Cons of fair representations:

Pros:

- A natural framework to separate data vendors vs data users (potentially malicious)
- Under the iid assumption (no distribution shift), guaranteed fairness for downstream tasks (due to the data-processing principle)
- Flexible — can control the tradeoff between data utility vs fairness by tuning λ

Cons:

- (Inevitably?) Hurts the accuracy of the downstream task
- Is not robust: under distribution shift (on the marginal distribution of X), fairness guarantee fails
- Computationally intractable: solving minimax opt is hard (especially for nonlinear feature maps)

"Impossibility Results for Fair Representations", Lechner, Ben-David, Agarwal, Anathakrishnan, 2021

In-processing: Constrained Optimization

Empirical risk minimization with approximate SP constraint:

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

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

error minimization

constraint of approximate statistical parity

- Model dependent — different choices of hypothesis class leads to different algorithms
- NP-hard under 0-1 loss for most hypothesis class
- Tractable under surrogate loss

"Certifying and removing disparate impact", Feldman, Friedler, Moeller, Scheidegger and Venkatasubramanian, KDD'15

"Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment", Zafar, Valera, Rodriguez, Gummadi, WWW' 17

"A Reductions Approach to Fair Classification", Agarwal, Beygelzimer, Dud´ık, Langford, Wallach, ICML'18

In-processing: Constrained Optimization

Logistic regression with approximate SP constraint:

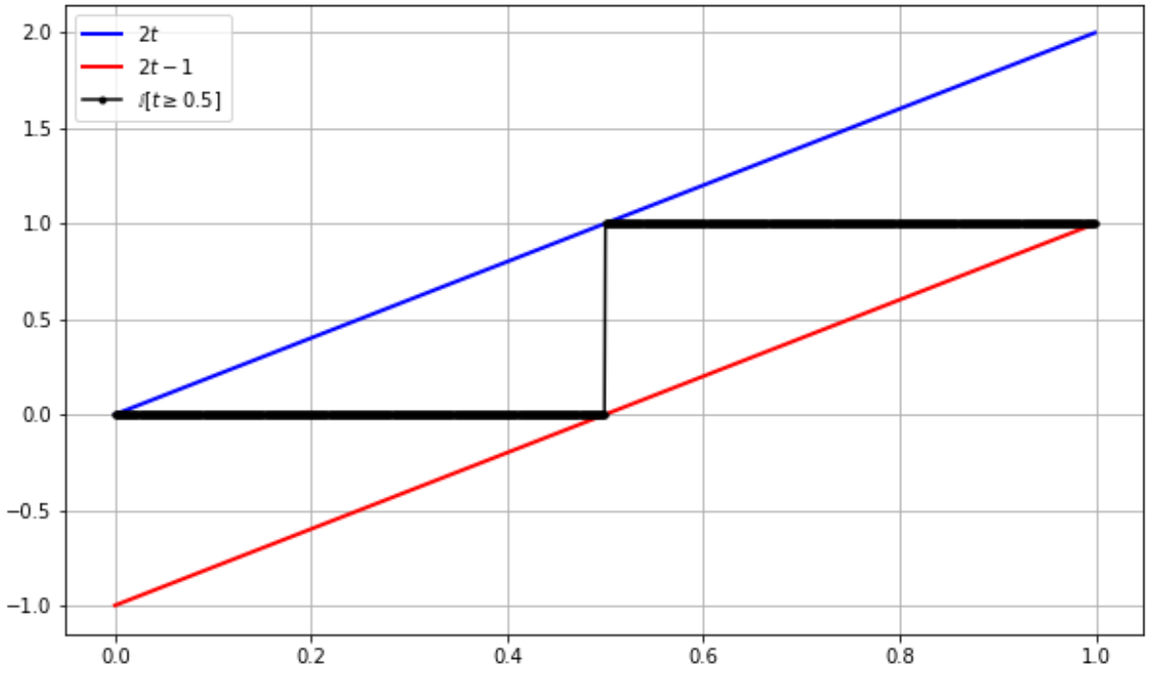
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} y_i \log \sigma(w^\top x_i) + (1-y_i)\log(1-\sigma(w^\top x_i))$$

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

For logistic regression, we know

$$f_w(x) = \begin{cases} 1 & \text{if } \sigma(w^\top x) \ge 1/2 \\ 0 & \text{o.w.} \end{cases}$$

where $\sigma(t) := 1/(1+\exp(-t))$ is the sigmoid function.



In-processing: Constrained Optimization

Logistic regression with approximate SP constraint:

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} y_i \log \sigma(w^\top x_i) + (1-y_i)\log(1-\sigma(w^\top x_i))$$

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

For logistic regression, we know

$$f_w(x) = \begin{cases} 1 & \text{if } \sigma(w^\top x) \ge 1/2 \\ 0 & \text{o.w.} \end{cases}$$

where $\sigma(t) := 1/(1+\exp(-t))$ is the sigmoid function.

Hence for $a \in \{0,1\}$ we have

$$2\mathbb{E}_{A=a}[\sigma(w^\top x)] - 1 \le \Pr(f_w(x)=1 \,|\, A=a) \le 2\mathbb{E}_{A=a}[\sigma(w^\top x)]$$

and we can relax the constraint as for $a \in \{0,1\}$:

$$\begin{aligned} \Pr(f_w(x)=1 \,|\, A=a) &- \Pr(f_w(x)=1 \,|\, A=1-a) \\ &\le 2\mathbb{E}_{A=a}[\sigma(w^\top x)] - 2\mathbb{E}_{A=1-a}[\sigma(w^\top x)] + 1 \\ &\le \epsilon \end{aligned}$$

In-processing: Constrained Optimization

Logistic regression with approximate SP constraint:

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} y_i \log \sigma(w^\top x_i) + (1-y_i)\log(1-\sigma(w^\top x_i))$$

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

Hence for $a \in \{0,1\}$ we have

$$2\mathbb{E}_{A=a}[\sigma(w^\top x)] - 1 \le \Pr(f_w(x)=1 \,|\, A=a) \le 2\mathbb{E}_{A=a}[\sigma(w^\top x)]$$

and we can relax the constraint as for $a \in \{0,1\}$:

$$\begin{aligned}\Pr(f_w(x)=1 \,|\, A=a) - \Pr(f_w(x)=1 \,|\, A=1-a) &\\ \le 2\mathbb{E}_{A=a}[\sigma(w^\top x)] - 2\mathbb{E}_{A=1-a}[\sigma(w^\top x)] + 1 &\\ \le \epsilon &\end{aligned}$$

Convert the constraint into an unconstrained penalized objective function and solve it.

$$|\Pr(f_w(x)=1 \mid A=0) - \Pr(f_w(x)=1 \mid A=1)| \le \epsilon$$

$$\Updownarrow$$

$$\Pr(f_w(x)=1 \mid A=0) - \Pr(f_w(x)=1 \mid A=1) \le \epsilon \wedge \Pr(f_w(x)=1 \mid A=1) - \Pr(f_w(x)=1 \mid A=0) \le \epsilon$$

Fairness-Accuracy Tradeoff

In order to satisfy statistical parity, no matter it's pre-processing or in-processing 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:

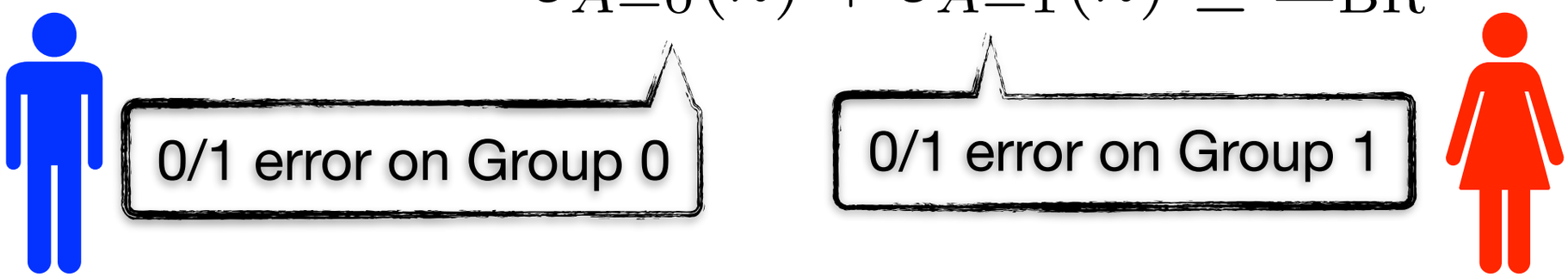
- Does not depend on the specific algorithm used to achieve statistical parity
- Does not depend on the computational resources available to the algorithm
- Does not depend on the sample size for training the predictor

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:

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

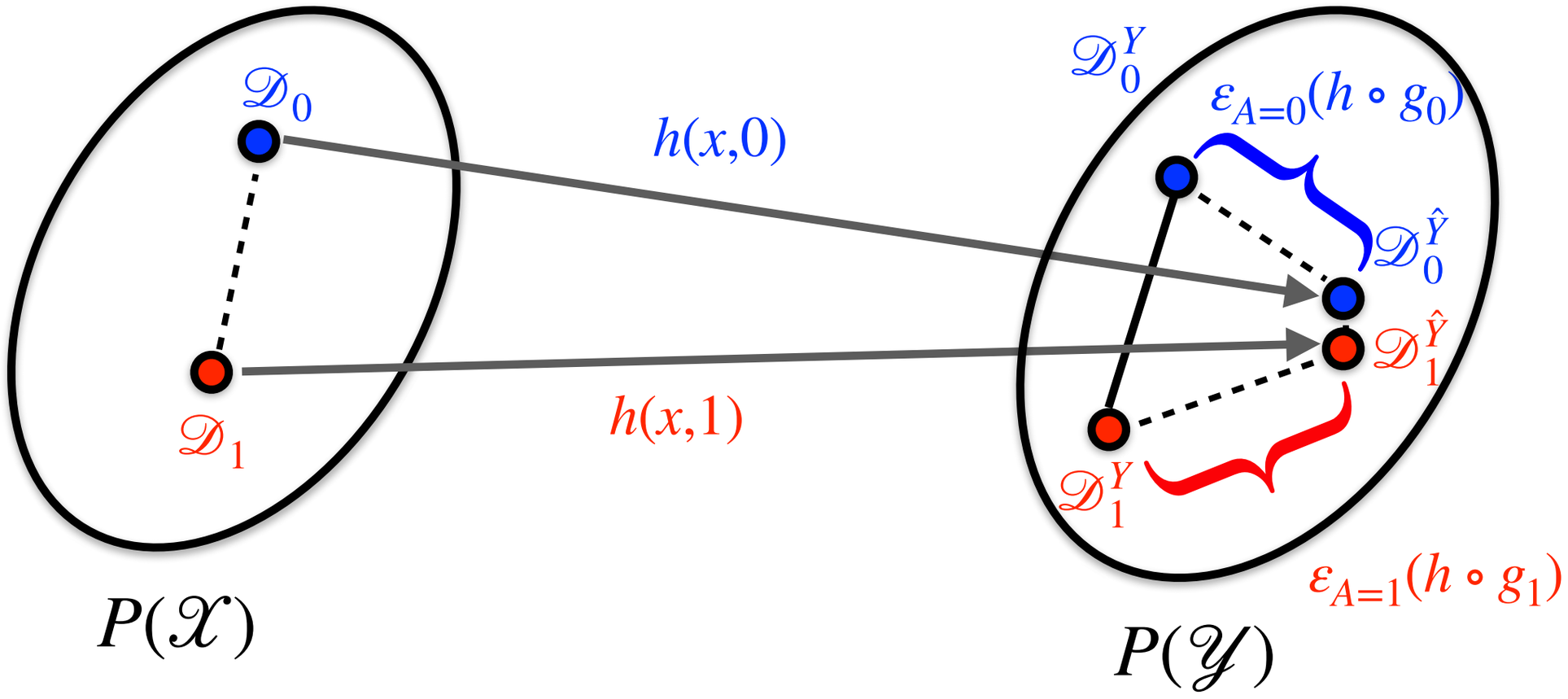
where $H_{01}(A) := 1 - \max_a \Pr(A=a)$ is called zero-one entropy of A

- 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 the Theorem 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

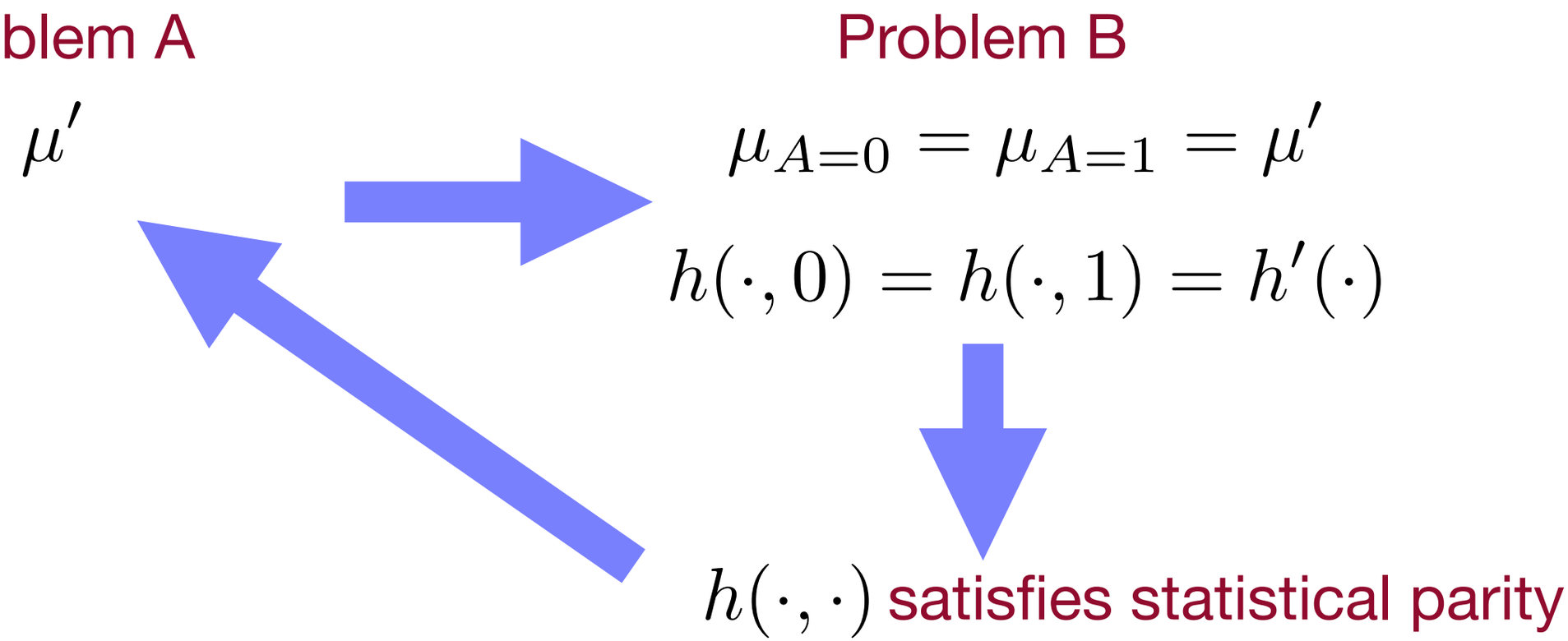
Problem A:

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:



"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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}^*

"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

An Optimal Fair Classifier via Post-Processing

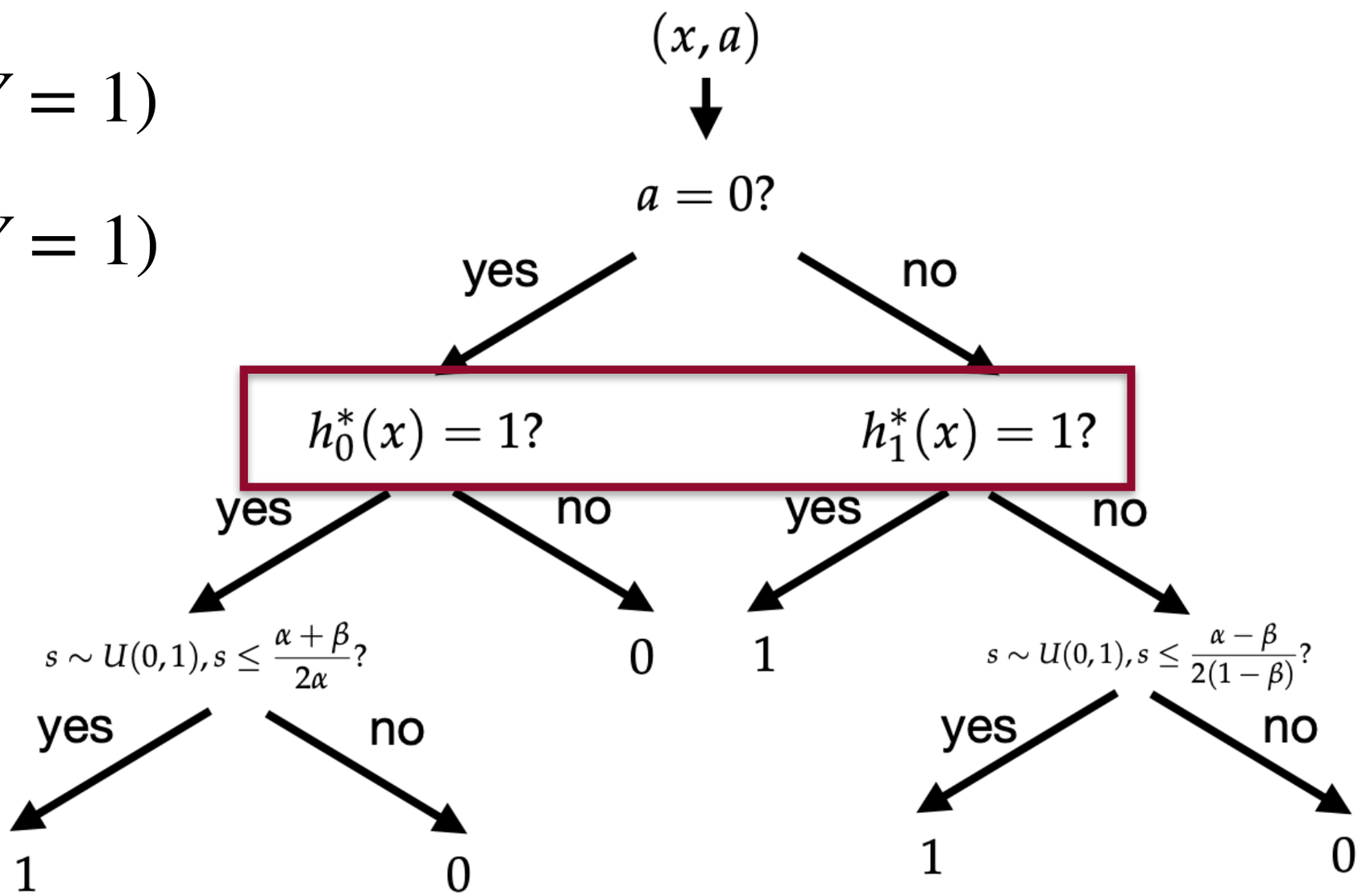
A constructive algorithm for the optimal fair classifier:

- It is a randomized classifier (the optimal classifier has to be randomized)
- The classifier needs to have explicit access to the sensitive attribute

$$\alpha = \Pr_{A=0}(Y=1)$$

$$\beta = \Pr_{A=1}(Y=1)$$

$$(\alpha \geq \beta)$$



"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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.,

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

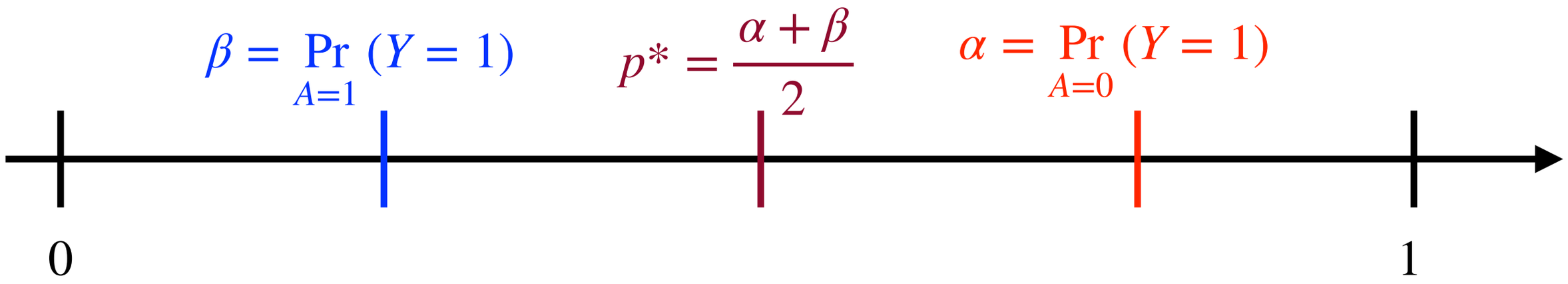
Note:

- 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

"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

An Optimal Fair Classifier via Post-Processing

Proof sketch:



Intuition:

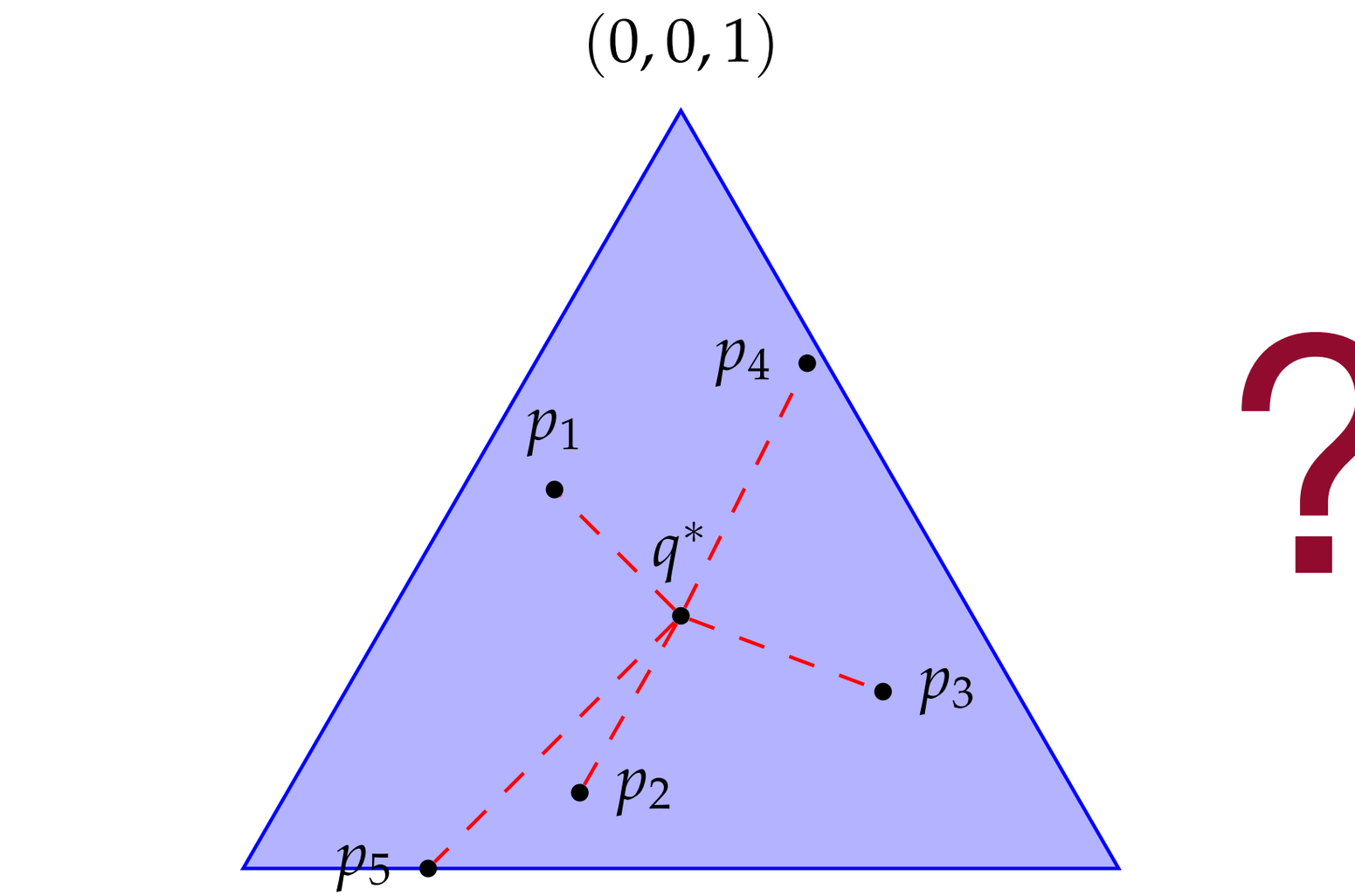
- Statistical parity enforces the same predictive probability
  $p = \Pr_{A=0}(\hat{Y}=1) = \Pr_{A=1}(\hat{Y}=1) \in [0,1]$
- The cost of fairness in $A=0$: $|\alpha - p|$
- The cost of fairness in $A=1$: $|\beta - p|$
- We should place $p^* \in [\beta, \alpha]$ and the algorithm chooses $p^* = \frac{\alpha+\beta}{2}$ (fairer since the price paid by both groups will be the same in this case)
- Randomization is used to achieve the post-process transportation

"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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 $i \in [n]$



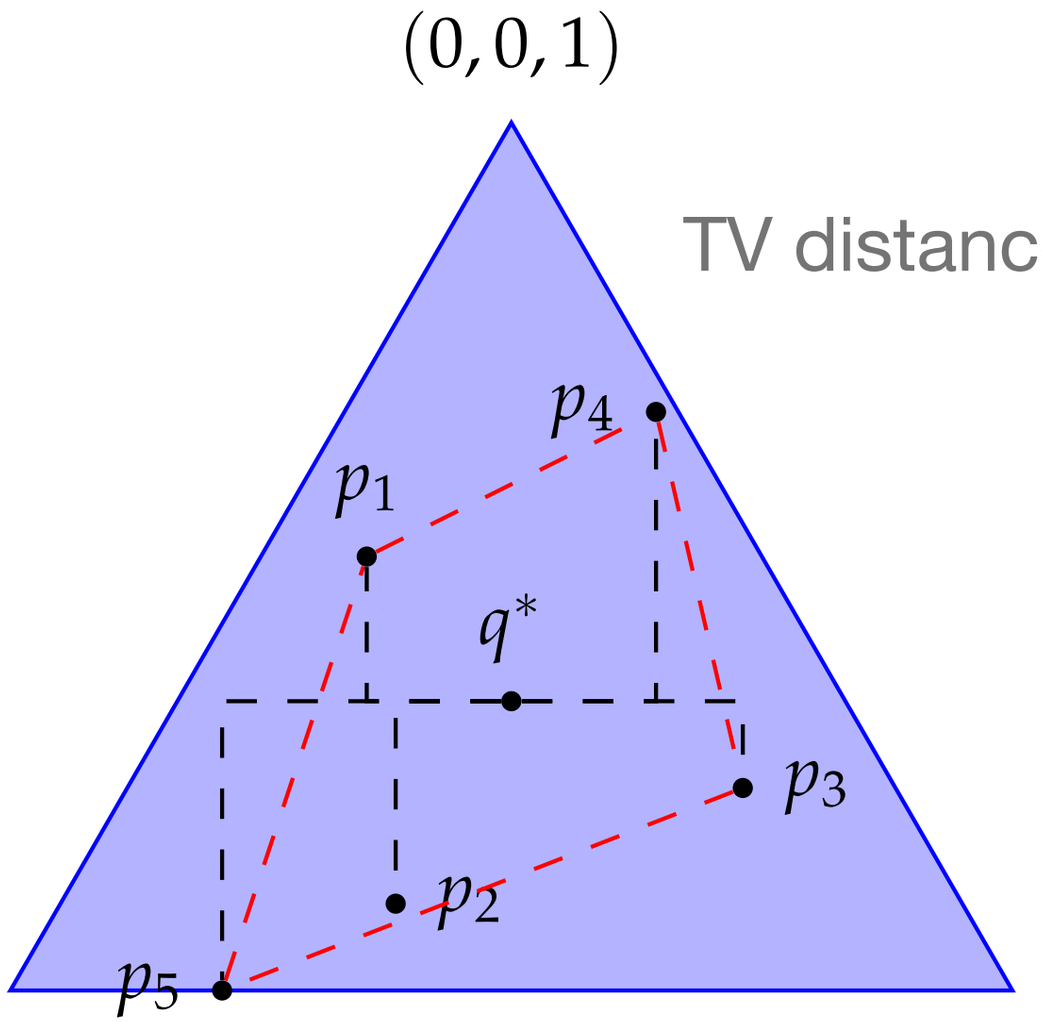
?

"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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 $i \in [n]$



TV distance: $d_{\mathrm{TV}}(p,q) := \sup_{E \subseteq [m]} |p(E) - q(E)|$

$$= \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\sum_{i=1}^{m} |p_i - q_i|$$

$$= \inf_{\gamma:\mathrm{Law}(Y)=p,\mathrm{Law}(Y')=q} \gamma \Pr(Y \neq Y')$$

"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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 for any fair classifier h

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

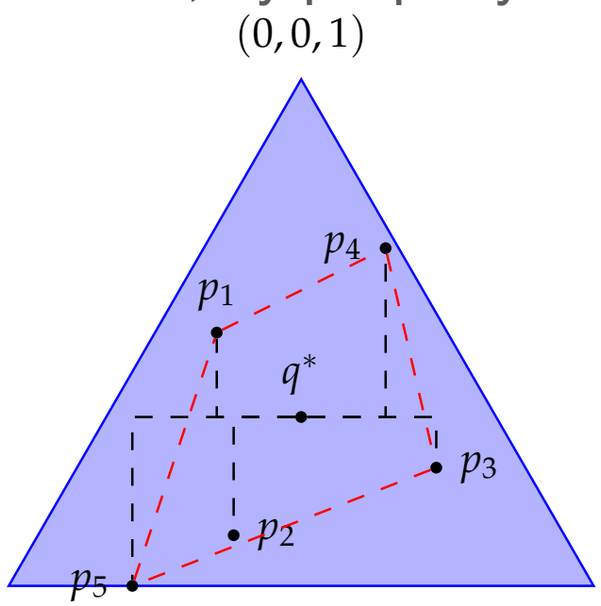
"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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 then for any fair classifier h:

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

- 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



"Inherent Tradeoffs in Learning Fair Representations", Zhao and Gordon, JMLR' 22

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 $f_i(x)(j) = \Pr(Y = j \mid X = x, A = i)$

Wasserstein distance/Optimal transport distance:

Let $d(\cdot, \cdot)$ be a distance over $\mathcal{Z} \times \mathcal{Z}$ and Z, Z' be two RVs over $\mathcal{Z}$ such that $\text{Law}(Z) = \mu, \text{Law}(Z') = \mu'$. Let γ be a coupling (joint distribution) over $Z \times Z'$ such that $\gamma_Z = \mu, \gamma_{Z'} = \mu'$. Then for $k \geq 1$, the $W_k(\mu, \mu')$ is defined to be:

$$W_k(\mu, \mu') := \inf_{\gamma} \left(\int d(Z, Z')^k \, d\gamma \right)^{1/k} = \inf_{\gamma} \mathbb{E}_{\gamma}^{1/k} \left[d(Z, Z')^k \right]$$

Example: if $\mathcal{Z} = [m]$ and $d(\cdot, \cdot) =$ Hamming distance, i.e., $d(Z, Z') = \mathbb{I}(Z \neq Z')$, then $W_1(\mu, \mu') = d_{\text{TV}}(\mu, \mu')$.

"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

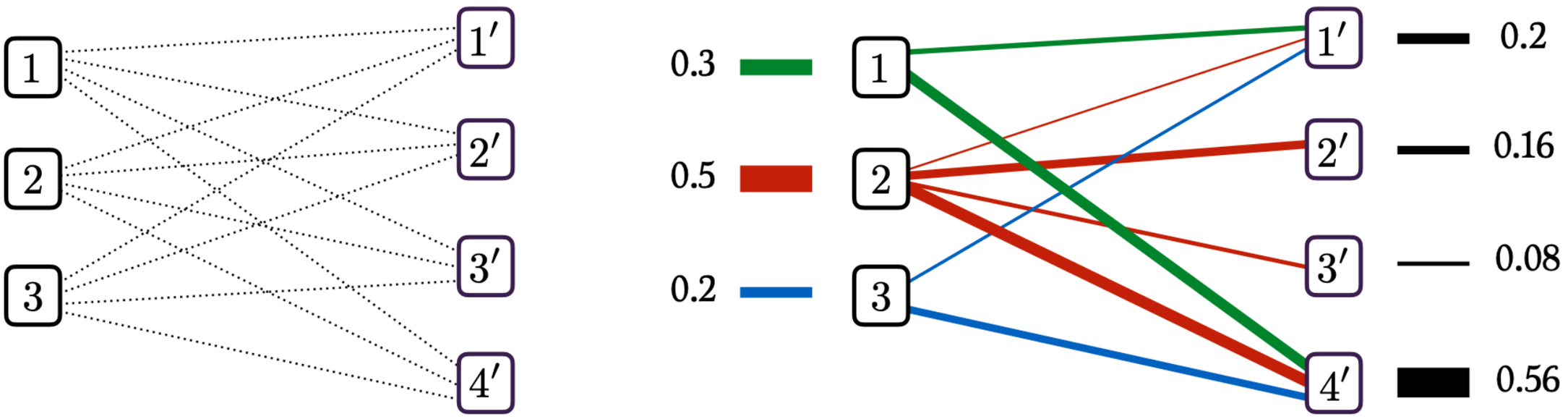
An Optimal Fair Classifier via Post-Processing

Wasserstein distance/Optimal transport distance:

Let $d(\cdot, \cdot)$ be a distance over $\mathcal{Z} \times \mathcal{Z}$ and Z, Z' be two RVs over $\mathcal{Z}$ such that $\text{Law}(Z) = \mu, \text{Law}(Z') = \mu'$. Let γ be a coupling (joint distribution) over $Z \times Z'$ such that $\gamma_Z = \mu, \gamma_{Z'} = \mu'$. Then for $k \geq 1$, the $W_k(\mu, \mu')$ is defined to be:

$$W_k(\mu, \mu') := \inf_{\gamma} \left(\int d(Z, Z')^k \, d\gamma \right)^{1/k} = \inf_{\gamma} \mathbb{E}_{\gamma}^{1/k} \left[d(Z, Z')^k \right]$$

Minimum cost network flow:



"Computational Optimal Transport", Peyré and Cuturi, 2018

An Optimal Fair Classifier via Post-Processing

Wasserstein distance/Optimal transport distance:

Let $d(\cdot, \cdot)$ be a distance over $\mathcal{Z} \times \mathcal{Z}$ and Z, Z' be two RVs over $\mathcal{Z}$ such that $\text{Law}(Z) = \mu, \text{Law}(Z') = \mu'$. Let γ be a coupling (joint distribution) over $Z \times Z'$ such that $\gamma_Z = \mu, \gamma_{Z'} = \mu'$. Then for $k \geq 1$, the $W_k(\mu, \mu')$ is defined to be:

$$W_k(\mu, \mu') := \inf_{\gamma} \left(\int d(Z, Z')^k \, d\gamma \right)^{1/k} = \inf_{\gamma} \mathbb{E}_{\gamma}^{1/k} \left[d(Z, Z')^k \right]$$

For empirical distributions induced from finite data:

- For any distance metric d for any $k \geq 1$, W_k can be computed via solving a linear program in $O(n^3)$, where n = # data points
- Faster approximations exist, e.g., Sinkhorn distance ($O(n^2)$), Sliced Wasserstein distance ($O(n)$), Tree-Wasserstein distance ($O(n)$)
- For $k = 1$, it is also known as the Earth-Mover distance in the computer science literature

"Computational Optimal Transport", Peyré and Cuturi, 2018

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):

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

$$\text{subject to} \quad q \in \Delta_m$$

Note: $f_i \sharp \mu_i^X$ is the induced distribution (push-forward) 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

"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

An Optimal Fair Classifier via Post-Processing

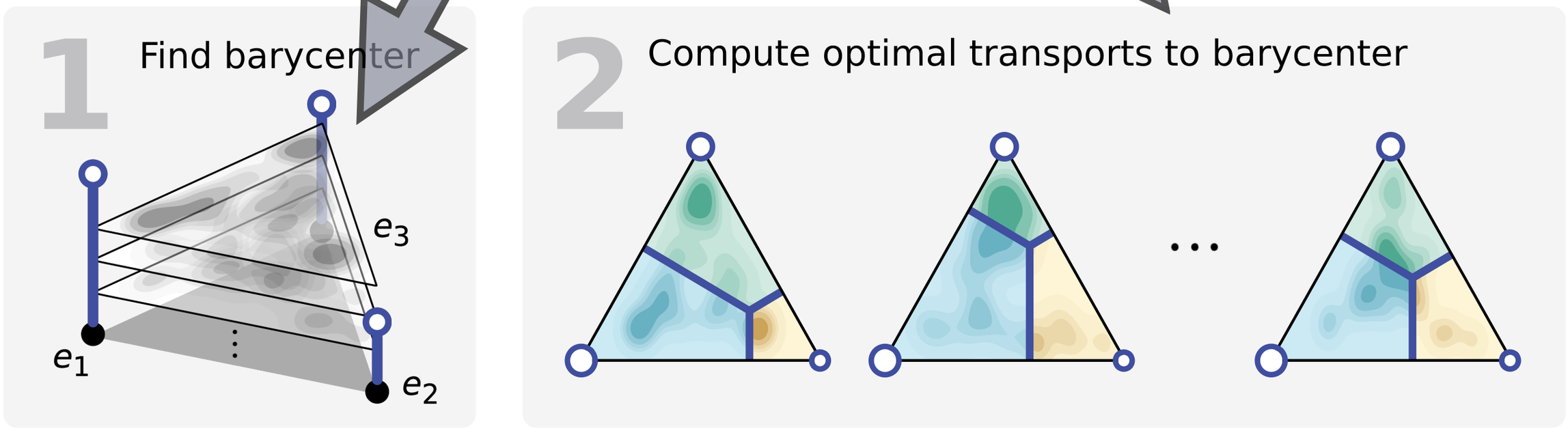
Price of fairness: (Wasserstein-Barycenter):

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

$$\text{subject to} \quad q \in \Delta_m$$

A two-step procedure:

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



"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

An Optimal Fair Classifier via Post-Processing

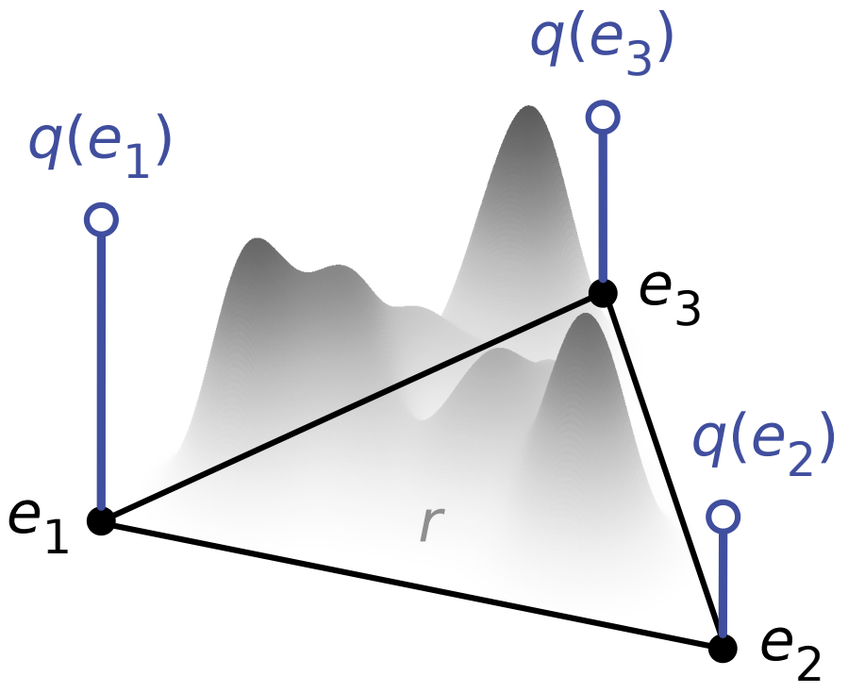
Price of fairness: (Wasserstein-Barycenter):

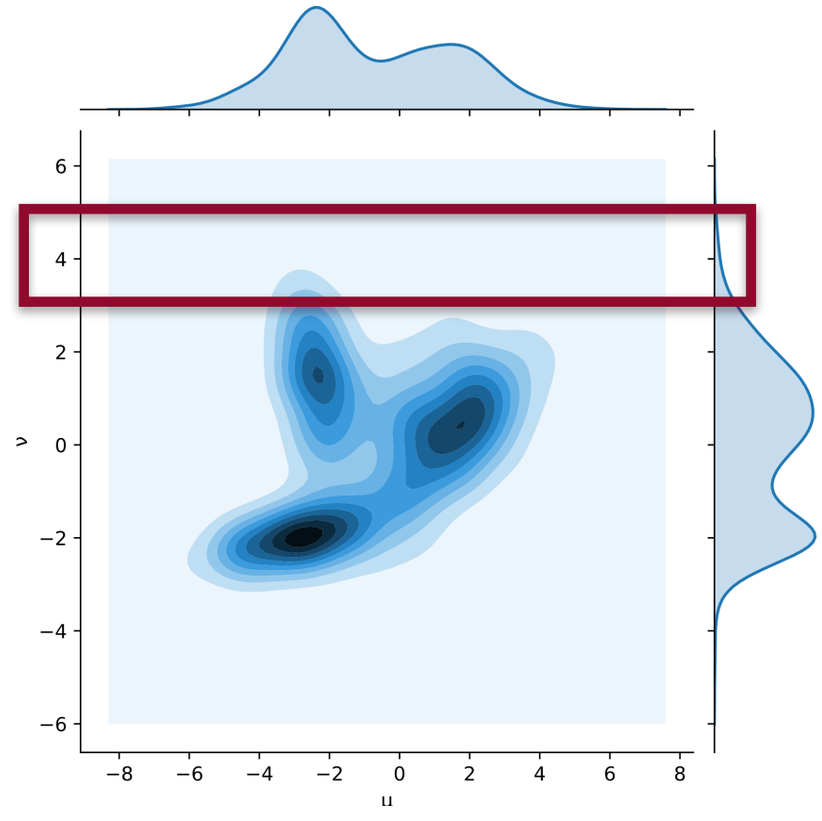
$$\min_{q} \quad \frac{1}{2}\sum_{i=1}^{n} W_1\left(f_i \sharp \mu_i^X, q\right)$$

$$\text{subject to} \quad q \in \Delta_m$$

A two-step procedure:

1. Find the barycenter under the W_1 metric (linear program)





"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

An Optimal Fair Classifier via Post-Processing

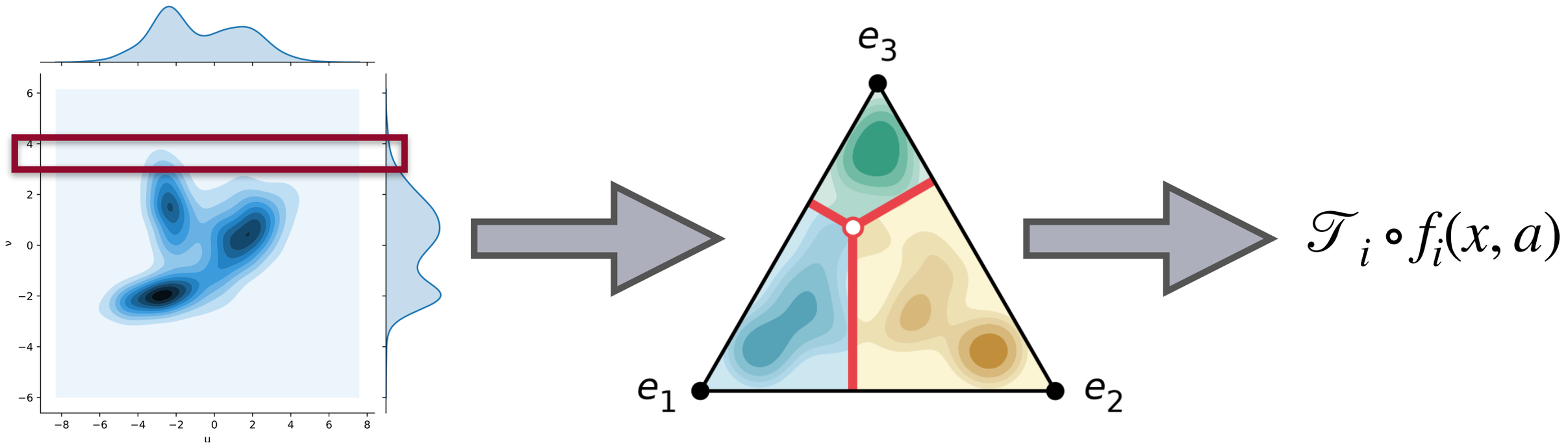
Price of fairness: (Wasserstein-Barycenter):

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

$$\text{subject to} \quad q \in \Delta_m$$

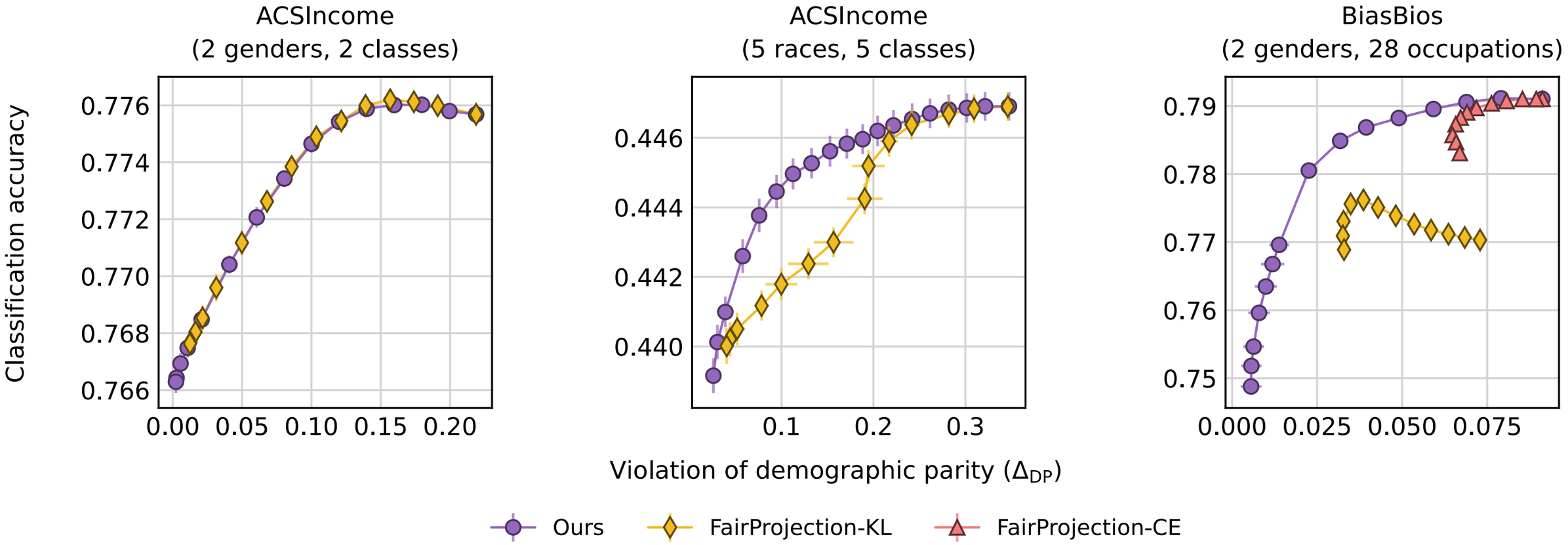
A two-step procedure:

2. Find the (randomized) transportation map to the barycenter as the post-processing map



"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

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)

Note: the algorithm also allows a relaxation of the exact fairness as well

"Fair and Optimal Classification via Post-Processing", Xian and Zhao, ICML' 23

An Optimal Fair Regression via Post-Processing

What about multi-groups regression under mean-squared error?

- Let $n \geq 2$ be the number of groups
- Let f_i be the Bayes score function of group $i \in [n]$ under the mean-squared error, i.e., $f_i(x) = \mathbb{E}_{\mu_i}[Y \mid X = x]$

Price of fairness:

(Wasserstein-Barycenter):

$$\min_{\nu} \quad \sum_{i=1}^{n} W_2^2 \left(f_i \sharp \mu_i^X, \nu \right)$$

Note: $f_i \sharp \mu_i^X$ is the induced distribution (push-forward) over $\mathbb{R}$ given by the mapping f_i acting on the marginal distribution of X, i.e., μ_i^X. $W_2(\,\cdot\,,\cdot\,)$ is the 2-Wasserstein distance under the ℓ_2 metric

"Fair Regression with Wasserstein Barycenters", Chzhen, Denis, Hebiri, Oneto and Pontil, 2020
"Projection to Fairness in Statistical Learning", Le Gouic, Loubes and Rigollet, 2020

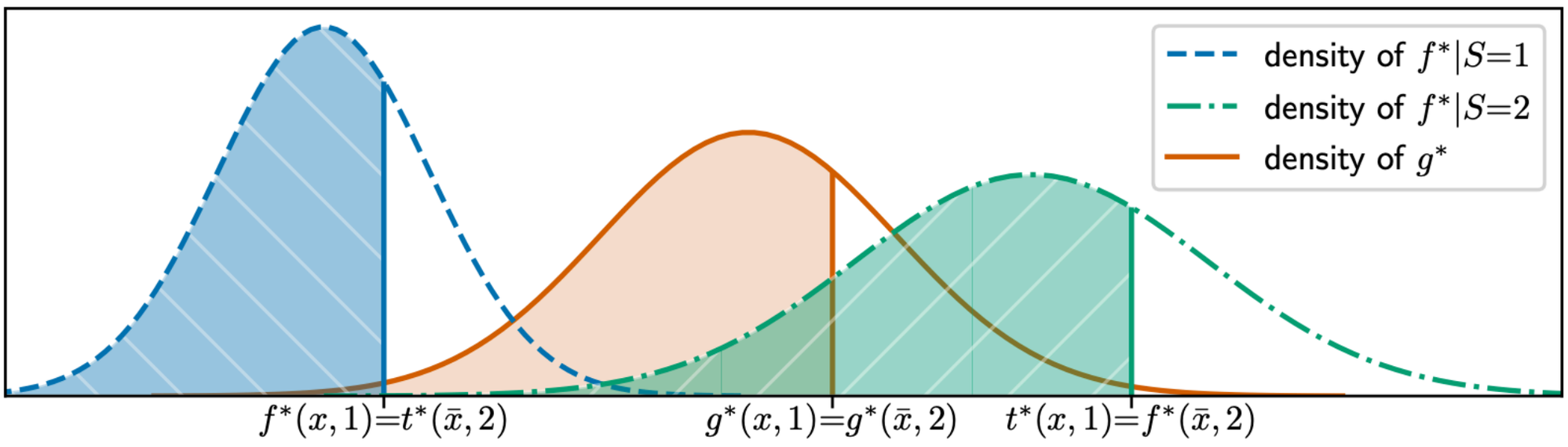
An Optimal Fair Regression via Post-Processing

What about multi-groups regression under mean-squared error?

- Let $n \geq 2$ be the number of groups
- Let f_i be the Bayes score function of group $i \in [n]$ under the mean-squared error, i.e., $f_i(x) = \mathbb{E}_{\mu_i}[Y \mid X = x]$

A two-step procedure:

1. Find the barycenter under the W_2 metric (linear program)
2. Find the transportation map to the barycenter as the post-processing map (ranking)



"Fair Regression with Wasserstein Barycenters", Chzhen, Denis, Hebiri, Oneto and Pontil, 2020

"Projection to Fairness in Statistical Learning", Le Gouic, Loubes and Rigollet, 2020

Post-processing via Wasserstein Barycenter

Inherent tradeoffs by enforcing statistical parity under different settings:

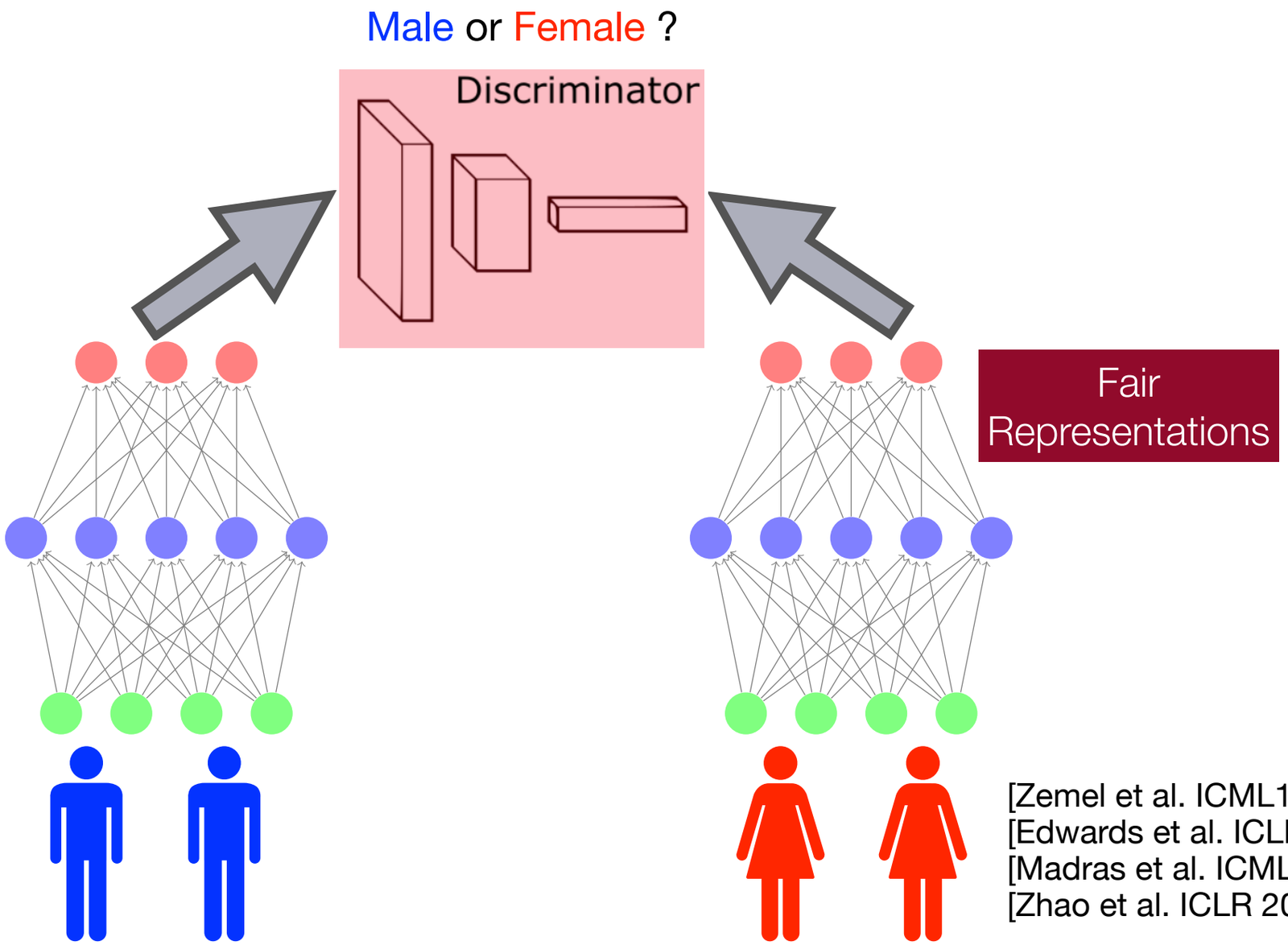
Table 1: Characterizations of the inherent tradeoff of (strict) DP fairness.

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

Summary

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

Summary

In-processing Methods: Constrained Optimization

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

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

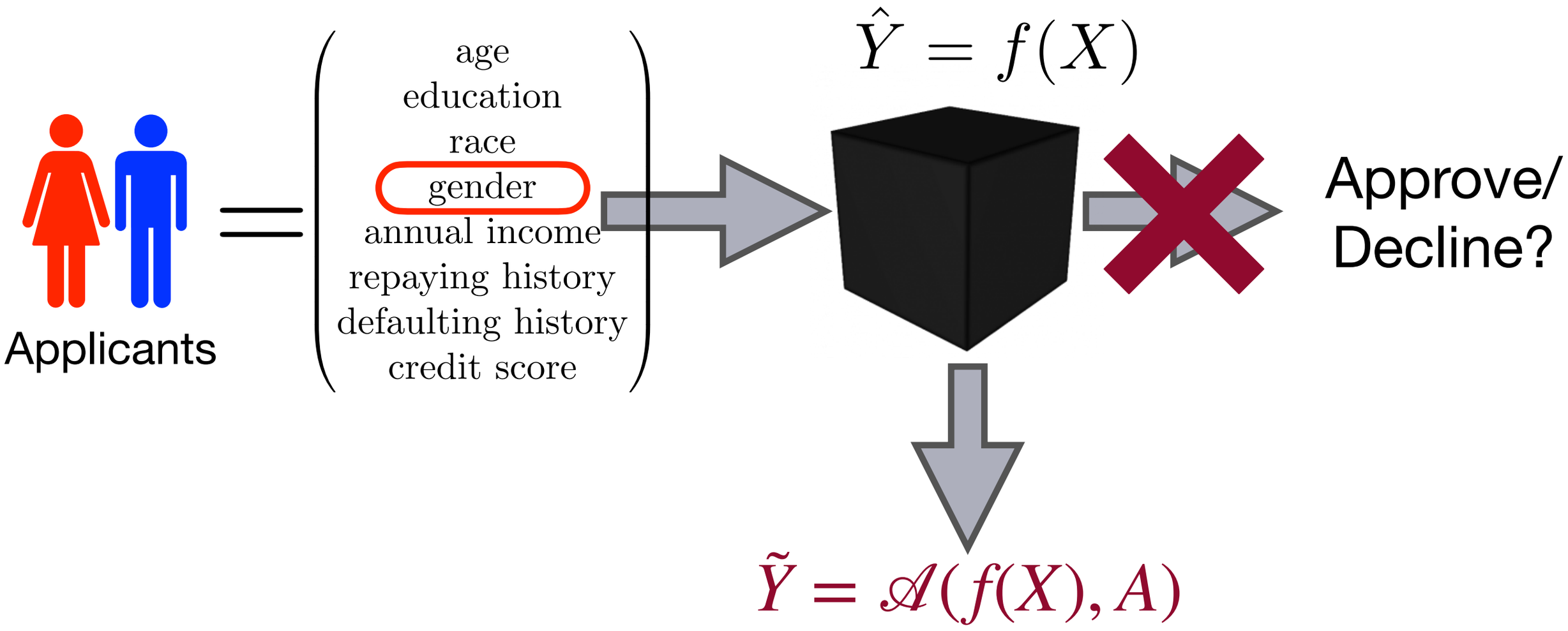
error minimization

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

Summary

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

Equalized Odds

What we have covered so far:

- Statistical parity, fair representations, optimal fair classifier
- Tradeoff between accuracy and statistical parity

Is statistical parity always a desirable notion of group fairness?

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

- ProPublica: Looking for equal false positive and negative rates
- Northpointe: Implemented a group-wise calibrated classifier

Equalized Odds

Is statistical parity always a desirable notion of group fairness?

- Rules out the perfect classifier if the base rates differ

Equalized Odds (Hardt et al.'16):

$$\widehat{Y} \perp A \mid Y$$

As a special case, when only the positive outcome (Y=1) is considered, this is also called equal opportunity in Hardt et al.'16.

Equal Opportunity (Hardt et al.'16):

$$\widehat{Y} \perp A \mid Y = 1$$

Intuitively, if a classifier $\widehat{Y}$ satisfies equalized odds, then

$$\Pr_{A=0}(\widehat{Y} = i \mid Y = j) = \Pr_{A=1}(\widehat{Y} = i \mid Y = j),\ \forall i, j, \in \{0, 1\}$$

Hence the classifier achieves equal FNR and FPR across the groups

"Equality of Opportunity in Supervised Learning", Hardt, Price, Srebro, NeurIPS' 16

Equalized Odds

Equalized Odds (Hardt et al.'16):

$$\widehat{Y} \perp A \mid Y$$

Is there also a similar tradeoff between accuracy and equalized odds in general?

Consider the noiseless setting, where there exists a perfect labeling function:

$$Y = f^*(X)$$

The classifier $\widehat{Y} = f^*(X)$ is optimal in accuracy, and satisfies EO

Note that for statistical parity, there is still a tradeoff in general even for the noiseless setting.

"Equality of Opportunity in Supervised Learning", Hardt, Price, Srebro, NeurIPS' 16

Equalized Odds

Equalized Odds (Hardt et al.'16):

$$\widehat{Y} \perp A \mid Y$$

Could we (approximately) ensure EO from a representation learning perspective?

Recall the fair representations approach for statistical parity

It suffices, if we could make sure

$$d_{\mathrm{TV}}(P_{|Y=0}, Q_{|Y=0}) \leq \epsilon,\ d_{\mathrm{TV}}(P_{|Y=1}, Q_{|Y=1}) \leq \epsilon$$

where

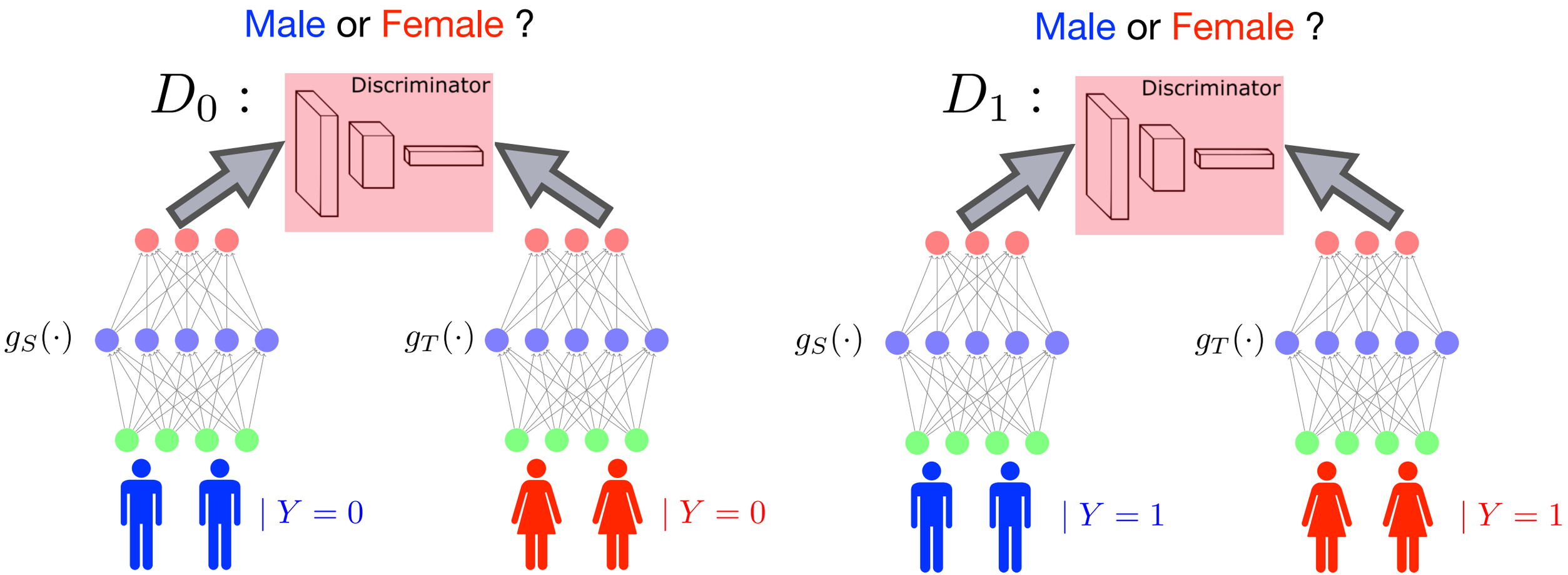
$$P_{|Y=i} := \Pr_{A=0}(\cdot \mid Y = i),\ Q_{|Y=i} := \Pr_{A=1}(\cdot \mid Y = i)$$

Idea: learn the features to simultaneously align both conditional distributions

Equalized Odds

Conditional Learning of Fair Representations

Training stage:



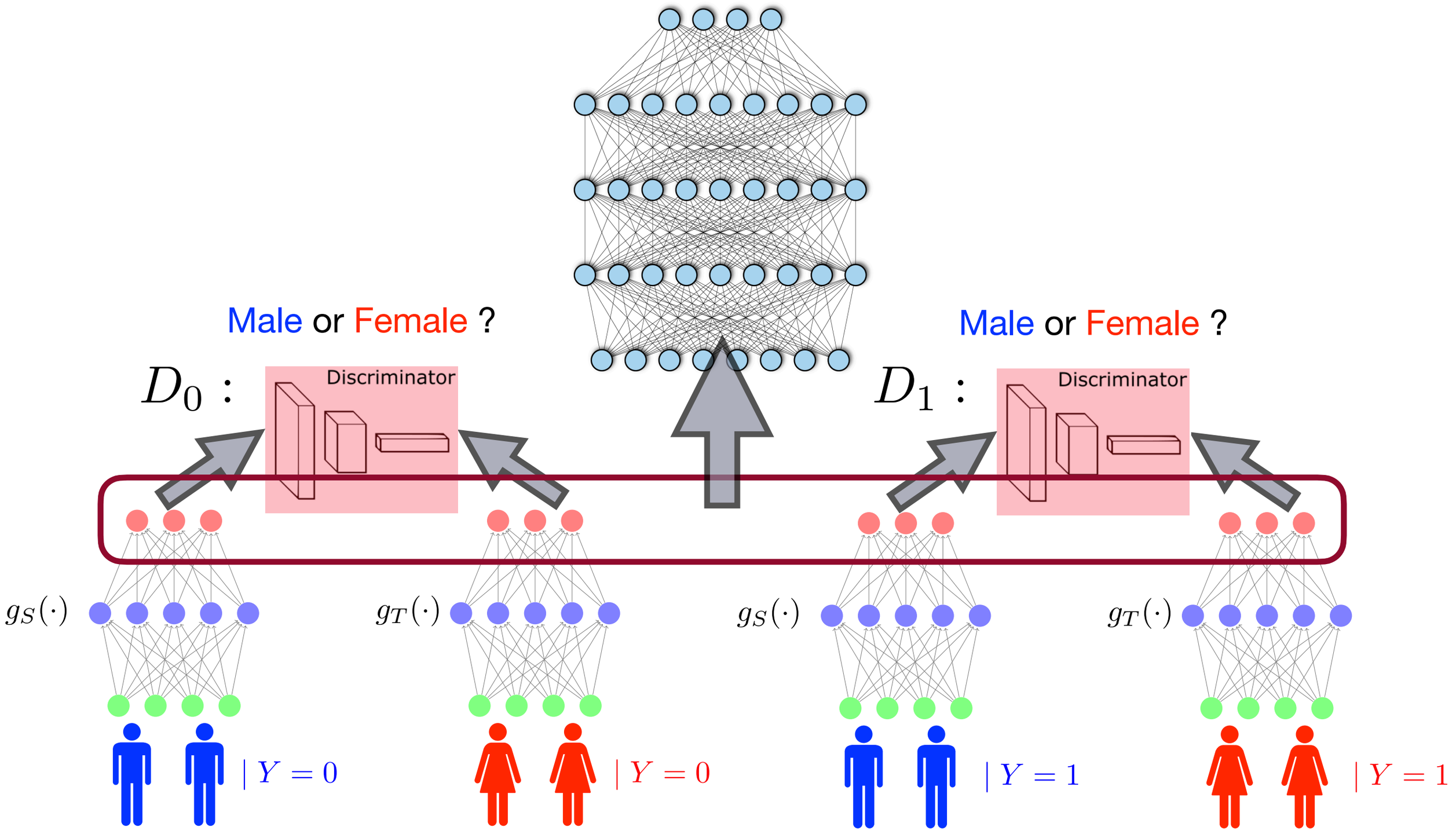
Note:

- Target labels needed only during the training stage
- Different feature extractors could be used for different groups

"Conditional Learning of Fair Representations", Zhao, Coston, Adel, and Gordon, ICLR' 20

Equalized Odds

Test stage:



Equalized Odds

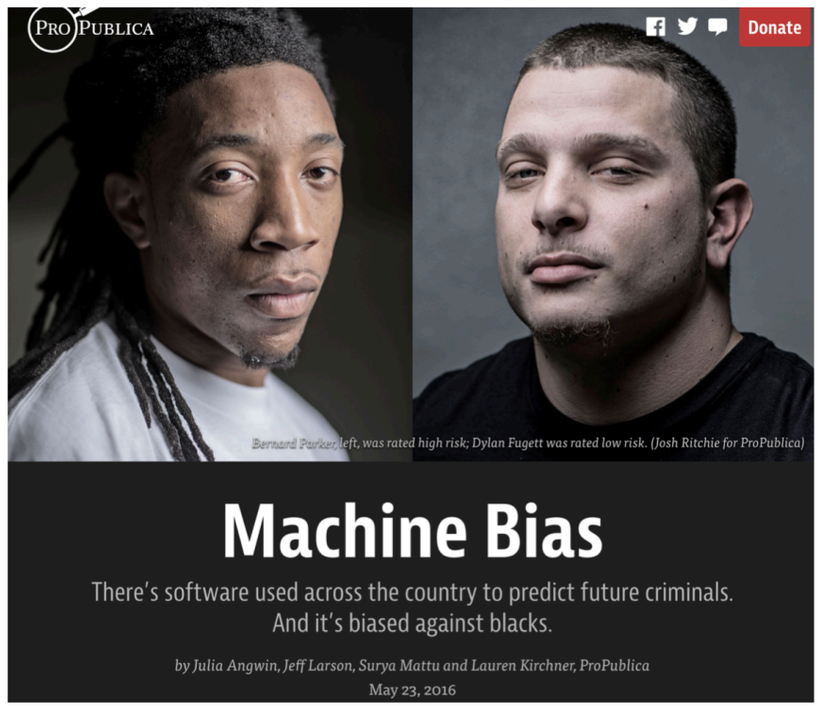
One more benefit of aligning the conditional distributions:

- We can show that the statistical disparity of any downstream classifier couldn’t be larger than the difference of base rates

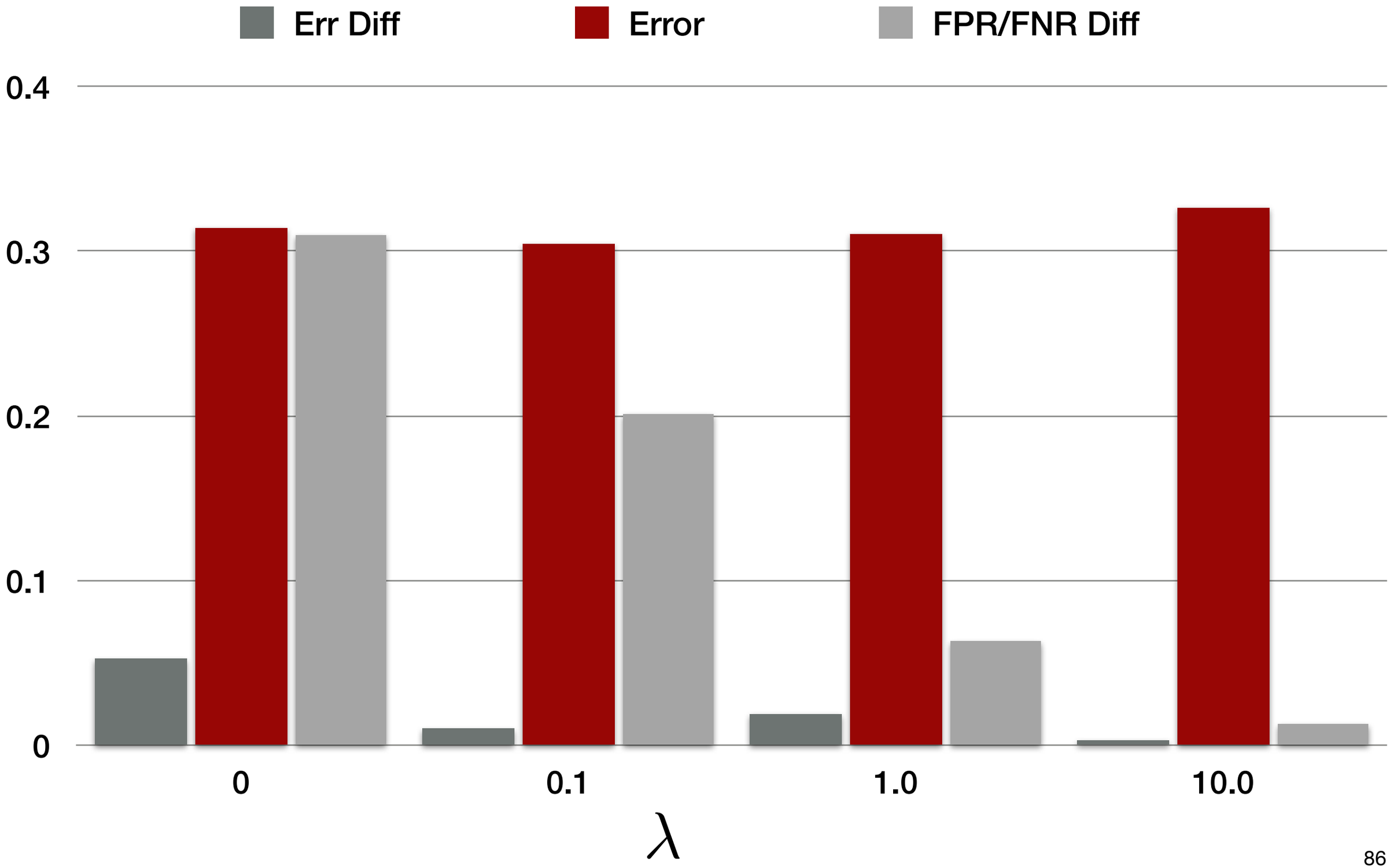
Experiment: Recidivism Prediction

COMPAS

- Train/Test: 4,320/1,852 instances from the Northpointe
- Target task: 0/1 classification (recidivism?)
- Sensitive attribute: race (Black/White)
- Other attributes: gender, education, prior arrest history, … (12 total)
- Difference of base rate: $\Delta_{\mathrm{BR}} = 0.129$



Experiment: Recidivism Prediction



Equalized Odds & Statistical Calibration

Recall in the COMPAS example, Northpointe's defense:

- The COMPAS tool is statistically calibrated by group, meaning that it is correct on average for each group

Statistical calibration by group:

$$\forall a \in \{0,1\}, c \in (0,1), \quad \Pr_{A=a}(Y=1 \mid C(X)=c) = c$$

Is it possible to further improve the COMPAS tool so that:

- It is still statistically calibrated by group
- It satisfies Equalized Odds

A weaker notion than statistical calibration by group:
predictive rate parity

$$\forall c \in \{0,1\}, \quad \Pr_{A=0}(Y=1 \mid C(X)=c) = \Pr_{A=1}(Y=1 \mid C(X)=c)$$

Equalized Odds & Statistical Calibration

Theorem (Chouldechova'17):

Assuming differing base rates and an imperfect classifier $\widehat{Y} \neq Y$.
Then, either

- Equalized odds fails, or
- Predictive rate parity fails.

Note: this also implies that, in general, statistical calibration by group and equalized odds cannot hold simultaneously (thus resolving the heated debate between Propublica and Northpointe)

Remark: this incompatibility result also holds for risk score functions, which is more general than binary classifier (proved by Kleinberg, Mullainathan, Raghavan'16)

"Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", Chouldechova' 17

"Inherent trade-offs in the fair determination of risk scores", Kleinberg, Mullainathan, Raghavan' 16

Equalized Odds & Statistical Calibration

Theorem (Chouldechova'17):

Assuming differing base rates and an imperfect classifier $\widehat{Y} \neq Y$.
Then, either

- Equalized odds fails, or
- Predictive rate parity fails.

Proof: We prove by contradiction. Assume both hold.

By the Bayes formula,

$$\Pr_{A=a}(Y = i \mid \widehat{Y} = j) = \frac{\Pr_{A=a}(\widehat{Y} = j \mid Y = i)\Pr_{A=a}(Y = i)}{\Pr_{A=a}(\widehat{Y} = j)}$$

Since $\widehat{Y} \neq Y$, there exists $i \neq j$, such that

$$\Pr_{A=a}(\widehat{Y} = j \mid Y = i) \neq 0$$

Equalized Odds & Statistical Calibration

Theorem (Chouldechova'17):

Assuming differing base rates and an imperfect classifier $\widehat{Y} \neq Y$.
Then, either

- Equalized odds fails, or
- Predictive rate parity fails.

Proof (cont'd):

By the predictive rate parity,

$$\frac{\Pr_{A=0}(\widehat{Y}=j \mid Y=i)\Pr_{A=0}(Y=i)}{\Pr_{A=0}(\widehat{Y}=j)} = \frac{\Pr_{A=1}(\widehat{Y}=j \mid Y=i)\Pr_{A=1}(Y=i)}{\Pr_{A=1}(\widehat{Y}=j)}$$

Since $\Pr_{A=a}(\widehat{Y}=j \mid Y=i) \neq 0$, by Equalized Odds,

$$\Pr_{A=0}(\widehat{Y}=j \mid Y=i) = \Pr_{A=1}(\widehat{Y}=j \mid Y=i) \neq 0$$

Equalized Odds & Statistical Calibration

Theorem (Chouldechova'17):

Assuming differing base rates and an imperfect classifier $\widehat{Y} \neq Y$.
Then, either

- Equalized odds fails, or
- Predictive rate parity fails.

Proof (cont'd):

Hence,

$$\frac{\Pr_{A=0}(Y=i)}{\Pr_{A=0}(\widehat{Y}=j)} = \frac{\Pr_{A=1}(Y=i)}{\Pr_{A=1}(\widehat{Y}=j)}$$

which means

$$\frac{\Pr_{A=0}(Y=i)}{\sum_j \Pr_{A=0}(\widehat{Y}=j)} = \frac{\Pr_{A=1}(Y=i)}{\sum_j \Pr_{A=1}(\widehat{Y}=j)}$$

yielding

$$A \perp Y$$

which contradicts with the assumption that the base rates differ across groups, finishing the proof.

Incompatibility Theorems

So far we have talked about three statistical notions of fairness:

- Statistical Parity
- Equalized Odds
- Predictive rate parity

All the three fairness concepts are statistical properties of the predictor. Broadly, these three are categorized under group fairness

In fact, except in certain degenerate cases, any two out of the three cannot hold simultaneously

Definitions of the three fairness concepts:

1. Statistical parity: $\widehat{Y} \perp A$
2. Equalized odds: $\widehat{Y} \perp A \mid Y$
3. Predictive rate parity: $Y \perp A \mid \widehat{Y}$

Incompatibility Theorems

We've proven the incompatibility between 2 & 3.

Definitions of the three fairness concepts:

1. Statistical parity: $\widehat{Y} \perp A$
2. Equalized odds: $\widehat{Y} \perp A \mid Y$
3. Predictive rate parity: $Y \perp A \mid \widehat{Y}$

1 & 3: statistical parity & predictive rate parity are mutually exclusive unless $A \perp Y$

Proof:

$$\widehat{Y} \perp A \implies I(A;\widehat{Y}) = 0$$
$$Y \perp A \mid \widehat{Y} \implies I(A;Y \mid \widehat{Y}) = 0$$

$$I(A;Y,\widehat{Y}) = I(A;\widehat{Y}) + I(A;Y \mid \widehat{Y}) = 0$$

$$0 \le I(A;Y) \le I(A;Y,\widehat{Y}) = 0 \implies A \perp Y$$

Incompatibility Theorems

We've proven the incompatibility between 2 & 3.

Definitions of the three fairness concepts:

1. Statistical parity: $\widehat{Y} \perp A$
2. Equalized odds: $\widehat{Y} \perp A \mid Y$
3. Predictive rate parity: $Y \perp A \mid \widehat{Y}$

1 & 2: For binary Y, statistical parity & equalized odds are mutually exclusive unless $A \perp Y$ or $\widehat{Y} \perp Y$

Incompatibility Theorems

1 & 2: For binary Y, statistical parity & equalized odds are mutually exclusive unless $A \perp Y$ or $\widehat{Y} \perp Y$

Proof:

Consider any event E, E' taken by $\widehat{Y}, A$, respectively.
On one hand, we have

$$\begin{aligned}\Pr(\widehat{Y} \in E) &= \Pr(\widehat{Y} \in E \mid A \in E') \\ &= \sum_y \Pr(\widehat{Y} \in E, y \mid A \in E') \\ &= \sum_y \Pr(\widehat{Y} \in E \mid A \in E', y) \cdot \Pr(y \mid A \in E') \\ &= \sum_y \Pr(\widehat{Y} \in E \mid y) \cdot \Pr(y \mid A \in E')\end{aligned}$$

On the other hand, we have

$$\Pr(\widehat{Y} \in E) = \sum_y \Pr(\widehat{Y} \in E \mid y) \cdot \Pr(y)$$

Incompatibility Theorems

1 & 2: For binary Y, statistical parity & equalized odds are mutually exclusive unless $A \perp Y$ or $\widehat{Y} \perp Y$

Proof (cont'd):

$$\sum_y \Pr(\widehat{Y} \in E \mid y) \cdot \Pr(y) = \sum_y \Pr(\widehat{Y} \in E \mid y) \cdot \Pr(y \mid A \in E')$$

$$\sum_y \Pr(\widehat{Y} \in E \mid y)(\Pr(y) - \Pr(y \mid A \in E')) = 0$$

Now that Y is binary, so we have $y \in \{0, 1\}$, hence

$$\begin{aligned} &\Pr(\widehat{Y} \in E \mid Y = 0)(\Pr(Y = 0) - \Pr(Y = 0 \mid A \in E')) \\ +&\Pr(\widehat{Y} \in E \mid Y = 1)(\Pr(Y = 1) - \Pr(Y = 1 \mid A \in E')) \\ &= 0 \end{aligned}$$

Furthermore,

$$\Pr(Y = 1) = 1 - \Pr(Y = 0), \qquad \Pr(Y = 1 \mid A \in E') = 1 - \Pr(Y = 0 \mid A \in E')$$

Incompatibility Theorems

1 & 2: For binary Y, statistical parity & equalized odds are mutually exclusive unless $A \perp Y$ or $\widehat{Y} \perp Y$

Proof (cont'd):

$$(\Pr(\widehat{Y} \in E \mid Y = 0) - \Pr(\widehat{Y} \in E \mid Y = 1)) \times$$
$$(\Pr(Y = 0) - \Pr(Y = 0 \mid A \in E')) = 0$$

$$\Pr(\widehat{Y} \in E \mid Y = 0) = \Pr(\widehat{Y} \in E \mid Y = 1) \quad \text{or}$$

$$\Pr(Y = 0) = \Pr(Y = 0 \mid A \in E')$$

Because both equalities hold for arbitrary events E, E', we must have

$$A \perp Y \quad \text{or} \quad \widehat{Y} \perp Y$$

Individual Fairness

Group fairness:

1. Statistical parity: $\widehat{Y} \perp A$
2. Equalized odds: $\widehat{Y} \perp A \mid Y$
3. Predictive rate parity: $Y \perp A \mid \widehat{Y}$

Except in certain degenerate cases, any two out of the three cannot hold simultaneously

What's missing from any statistical definition of fairness, i.e., group fairness?

Individual Fairness

Fairness notion defined between a pair of individuals

Group fairness only cares about the population level parity between two groups, e.g., predictive rate, FPR/FNR, etc.

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

November 30, 2011

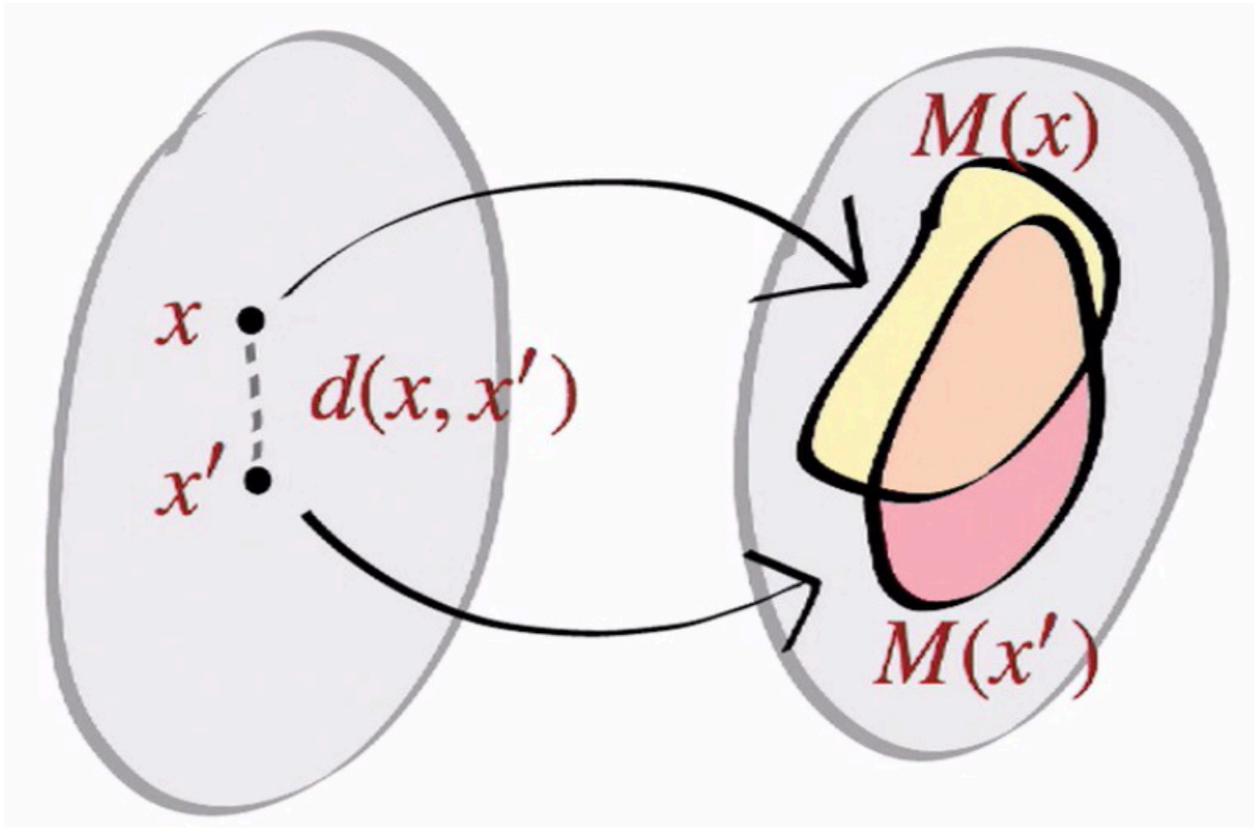
Similar individuals should be treated similarly

Abstract

We study *fairness in classification*, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand; (2) an algorithm for maximizing utility subject to the *fairness constraint*, that similar individuals are treated similarly. We also present an adaptation of our approach to achieve the complementary goal of "fair affirmative action," which guarantees *statistical parity* (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible. Finally, we discuss the relationship of fairness to privacy: when fairness implies privacy, and how tools developed in the context of differential privacy may be applied to fairness.

Individual Fairness

Fairness notion defined between a pair of individuals



Individual Fairness:

Definition 2.1 (Lipschitz mapping)**.** A mapping $M : V \to \Delta(A)$ satisfies the (D, d)-*Lipschitz* property if for every $x, y \in V$, we have

$$D(Mx, My) \leq d(x, y)\,. \tag{1}$$

When D and d are clear from the context we will refer to this simply as the *Lipschitz* property.

Individual Fairness

Fairness notion defined between a pair of individuals

Individual Fairness optimization formulation:

> Find a mapping from individuals to distributions over outcomes that minimizes expected loss subject to the Lipschitz condition.

Equivalent to solving a linear program under the transductive learning setting:

$$\begin{aligned} \mathrm{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \quad & \min_{\{\mu_x\}_{x \in V}} \quad \mathop{\mathbb{E}}_{x \sim V} \mathop{\mathbb{E}}_{a \sim \mu_x} L(x, a) \\ \text{subject to} \quad & \forall x, y \in V, : \quad D(\mu_x, \mu_y) \leq d(x, y) \\ & \forall x \in V : \quad \mu_x \in \Delta(A) \end{aligned}$$

Figure 1: The Fairness LP: Loss minimization subject to fairness constraint

Q: How to ensure individual fairness in inductive learning?

Q: What's the practical limitation of this fairness notion?

Individual Fairness

Fairness notion defined between a pair of individuals

Individual Fairness optimization formulation:

> Find a mapping from individuals to distributions over outcomes that minimizes expected loss subject to the Lipschitz condition.

Q: How to ensure individual fairness in inductive learning?

Goal: Let $f : \mathcal{X} \to \Delta_k$ be a classifier over k-output classes such that f satisfies individual fairness (Lipschitz condition).

Claim: If f is differentiable, then f is ρ-individual fairness under the Euclidean distance iff $\sup_{x \in \mathcal{X}} \|\nabla f(x)\|_2 \leq \rho$

Hence, given a similarity measure $d(\cdot, \cdot)$ between individuals, to achieve individual fairness, it suffices to learn Lipschitz continuous classifier.

Individual Fairness

Fairness notion defined between a pair of individuals

Individual Fairness optimization formulation:

> Find a mapping from individuals to distributions over outcomes that minimizes expected loss subject to the Lipschitz condition.

Q: What's the practical limitation of this fairness notion?

It's very hard, if not infeasible, to precisely quantify the similarity/ difference between individuals by using a single metric d

Roadmap of Trustworthy Machine Learning

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

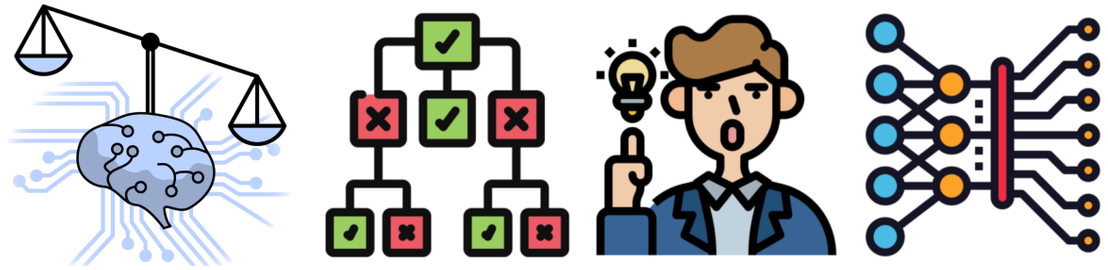
Fairness

Learning Fair Representations
Tradeoff between fairness & accuracy

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Efficiency & Accuracy

Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Maximum Influence Subset
Understand and select out the subset of data with maximum influence to the learned model

Structured Representations
Learning representations with class hierarchical information

Interpretability

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Privacy

Roadmap of Trustworthy Machine Learning

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

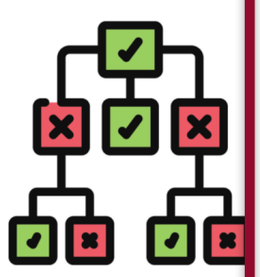
Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

Trustworthy Machine Learning



Efficiency & Accuracy

Fairness

Learning Fair Representations
Tradeoff between fairness & accuracy

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Maximum Influence Subset
Understand and select out the subset of data with maximum influence to the learned model

Structured Representations
Learning representations with class hierarchical information

Interpretability

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Privacy

Roadmap of Trustworthy Machine Learning

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Fairness

Learning Fair Representations
Tradeoff between fairness & accuracy

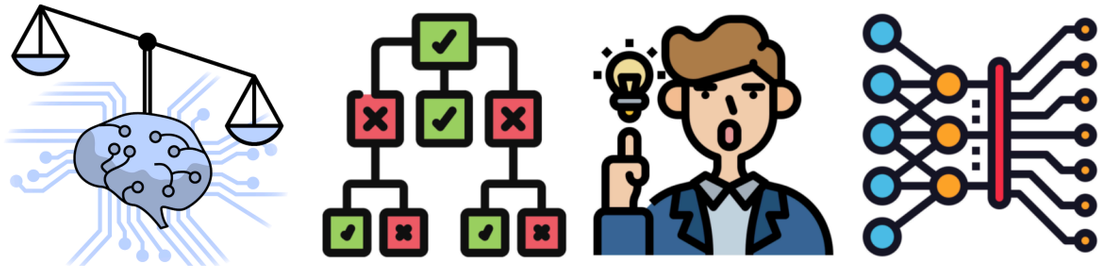
Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Efficiency & Accuracy

Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Maximum Influence Subset
Understand and select out the subset of data with maximum influence to the learned model

Structured Representations
Learning representations with class hierarchical information

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Interpretability

Privacy