Yale

# Algorithmic Fairness in ML

CPSC680: Trustworthy Deep Learning

Rex Ying

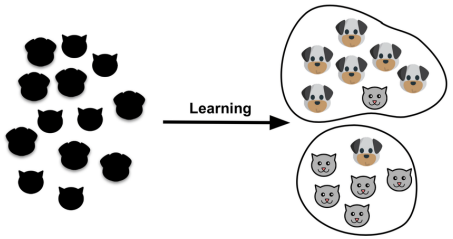
# Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

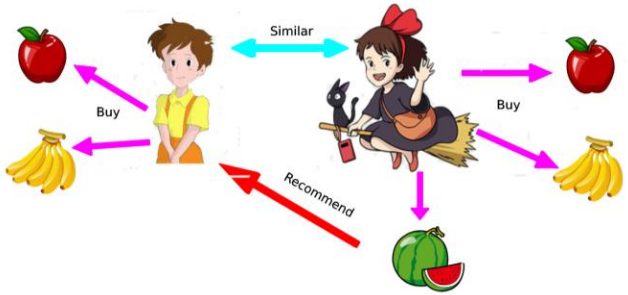
# Machine Learning in Consequential Domains

- ML models are increasingly applied in **consequential domains** like healthcare, law enforcement, and employment to automate decision-making.
- The application of ML models in those domains, even with many benefits (e.g., reducing labor costs), **raises many ethical concerns,** as a decision can significantly influence people's lives.

***In*consequential domains**




Dog-cat classification




Amazon recommendation

**Consequential domains**


Job hiring

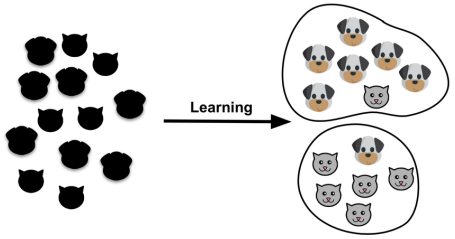
Criminal justice


Loan approval

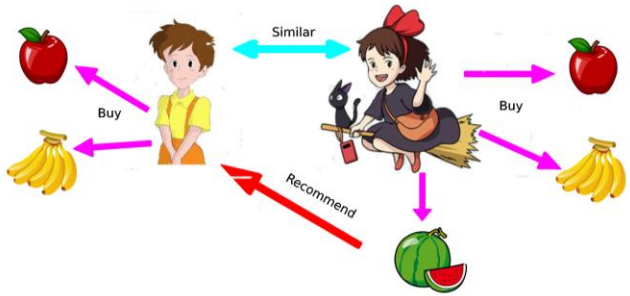
# Machine Learning in Consequential Domains

- In high-stakes domains, we do need to care more than just performance

**_In_consequential domains**




Dog-cat classification




Amazon Recommendation

**Performance, performance, performance!!!**

**Consequential domains**


Job hiring

Criminal justice


Loan approval

**Hmm!**

**Is the algorithm accurate?**

**Is the algorithm private?**

**Is the algorithm fair?**


**This lecture**

**....**

# COMPAS Case Study

- COMPAS is an algorithm developed by Equivant to **predict the chance that a criminal will commit another crime** in the future (recidivism).
- Recidivism scores impact criminal sentences: if a person is likely to commit another crime, shouldn't they get a longer sentence?
- Real systems that **have been used** in New York, Wisconsin, California, Florida,…
- The system is **claimed to be *fair*** as it did correctly predict recidivism for Black and White defendants at roughly the same rate.

# COMPAS Case Study

- The system is claimed to be *fair* as it did ***correctly*** predict recidivism for Black and White defendants at roughly the same rate.
- But, when it was wrong, it was wrong in different ways for Black and White

Black arrestees who would not be rearrested in a 2-year horizon scored as high risk at twice the rate of white arrestees not subsequently arrested


This is correct

But

PRO PUBLICA

Hmm! This is ***not*** *fair*

| **Prior Offense** | **Prior Offense** |
|---|---|
| 1 attempted burglary | 1 resisting arrest without violence |
| **Subsequent Offenses** | **Subsequent Offenses** |
| 3 drug possessions | None |

**It's not unfair; it satisfies a different notion of fairness!**

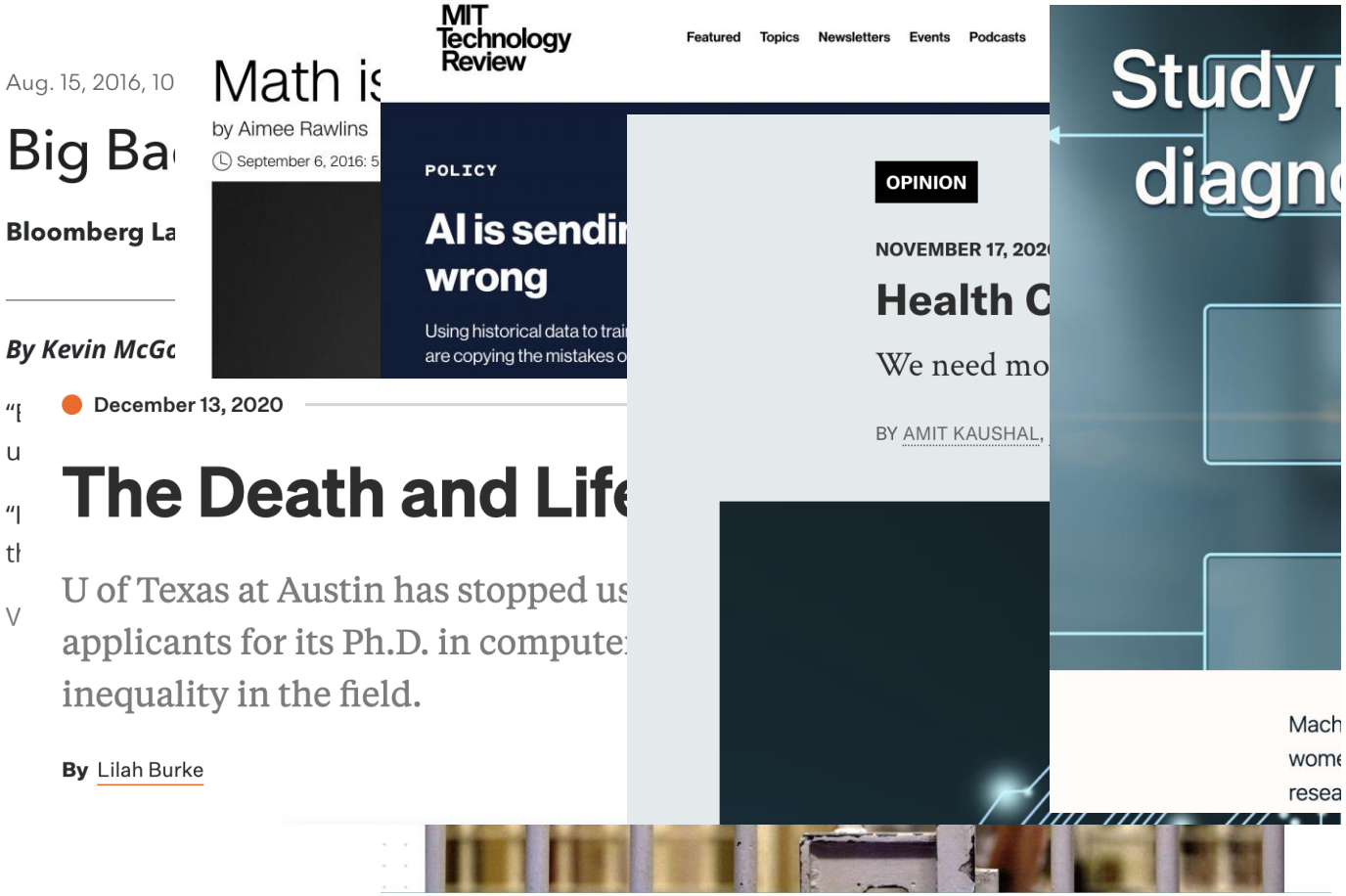
equivant

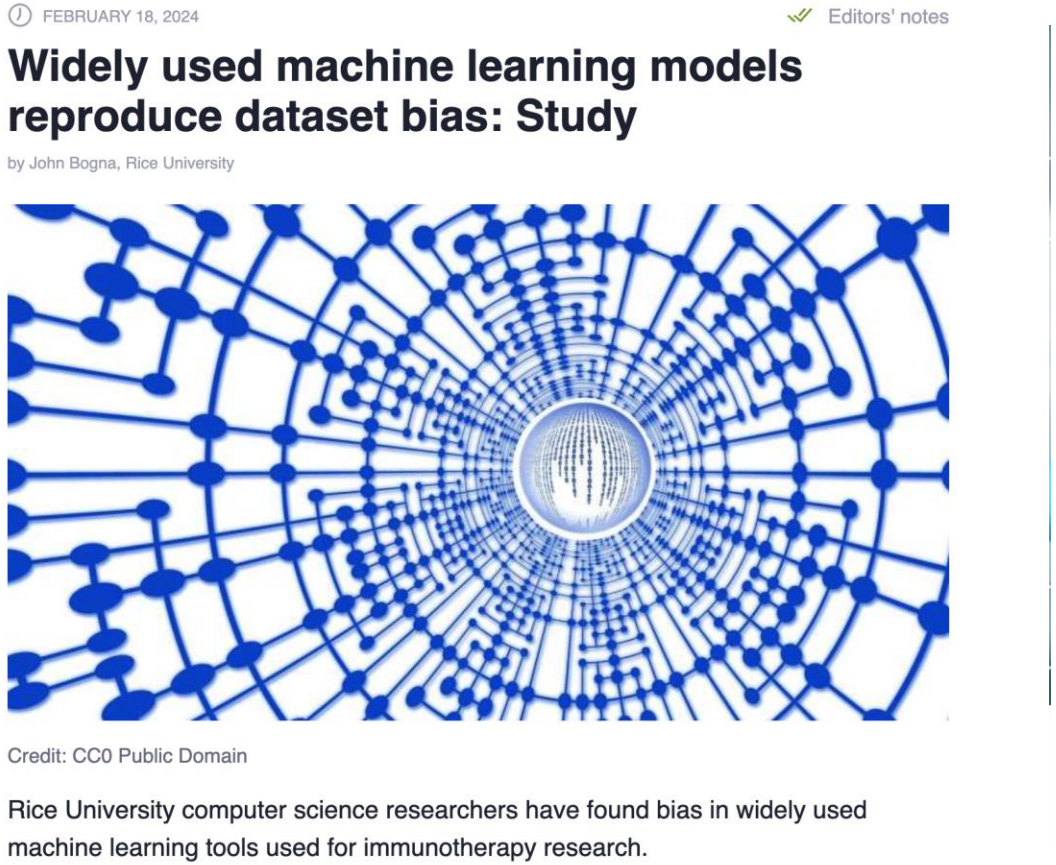
**There is no clear "wrong" or "right".**
**It turns out that these were incompatible definitions.**

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Algorithmic Bias

**It's more common than you thought!**



# Root Cause

- **Data, data, data!** If an ML model is **trained on data biased** by historical inequalities, human prejudices, or flawed collection practices, it will **likely learn and replicate these biases**.
- Biased data may come from
  - **Biased in data collection**: Data is gathered by humans who inherently possess biases.

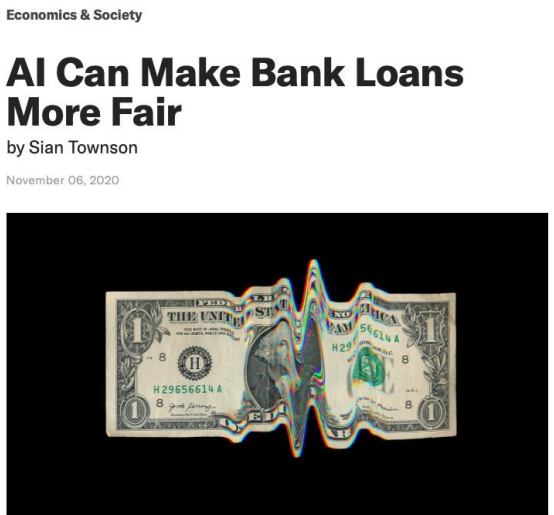
Study: People Associate 'Education' With Lighter Skin

Research participants remembered 'educated' black men as having a lighter skin tone.

By Brian Resnick and National Journal

# Root Cause

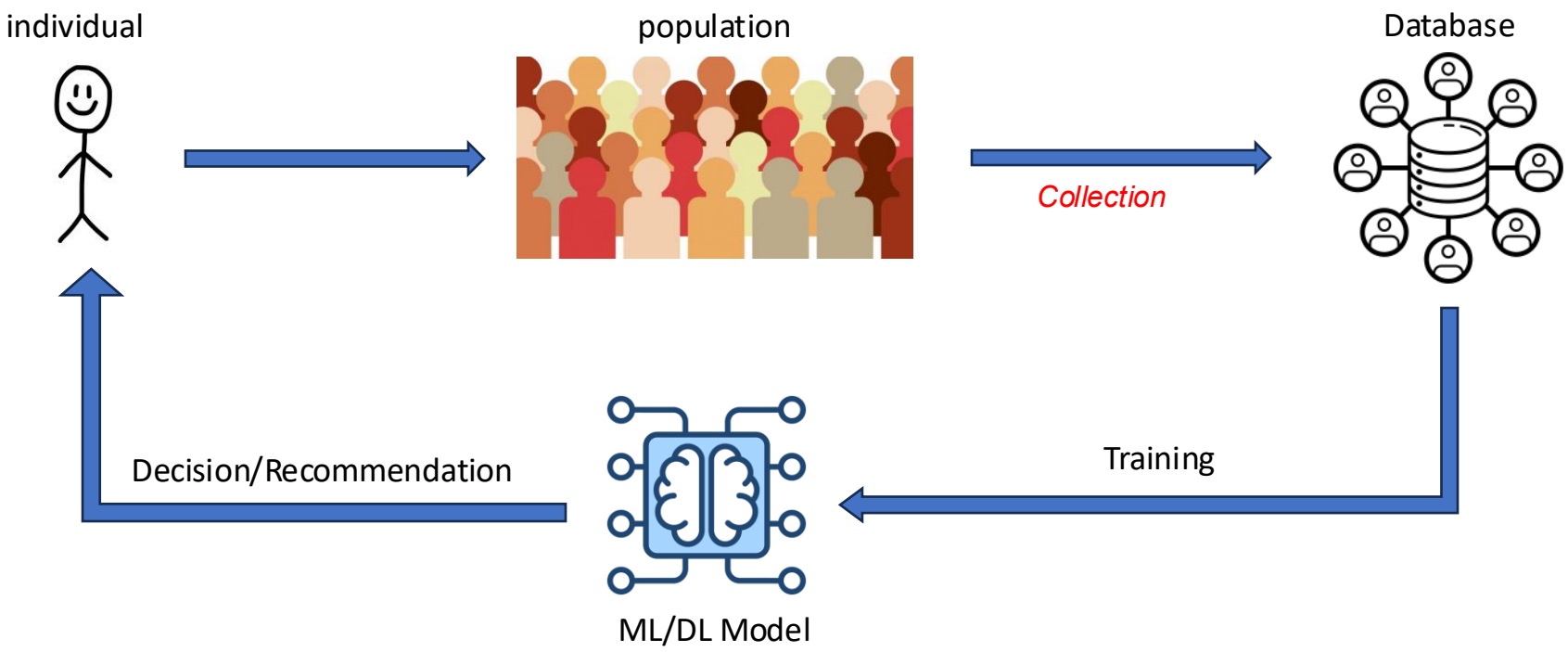
- If an ML model is **trained on data biased** by historical inequalities, human prejudices, or flawed collection practices, it will **likely learn and replicate these biases**.
- Biased data may come from
  - **Biased in data collection**: Data is gathered by humans who inherently possess biases.
  - **Imbalanced data**: ML models aim to optimize the performance on imbalanced data.




| | |
|---|---|
| Dataset: | Adult |
| Field of application: | Credit loan |
| Goal: | Predict income |
| Covariate: | Age, Occupation, Education, Gender... |
| Problem: | ~20% female, ~80% male |

# Feedback Loop

- **Amplifying feedback loop:** Systems that adapt based on their outputs can create a feedback loop that reinforces initial biases.



# Emerging Legislation

- **Emerging Legislation:** The use of AI models is getting more attention from regulators.

## NYC to Regulate Artificial Intelligence-Based Hiring Tools

Posted on December 15, 2021

POSTED IN U.S. STATE LAW, WORKPLACE PRIVACY

On November 10, 2021, the New York City Council passed a bill prohibiting employers and employment agencies from using automated employment decision tools to screen candidates or employees, unless a bias audit has been conducted prior to deploying the tool (the "Bill").

# Main Questions

- The fairness of AI-based systems is **unquestionably crucial**.
- However, this is still an open research direction. There are generally **more problems than solutions**.

**1. Measure of Fairness**

**Q:** What is the correct measure of fairness?
**A:** It depends (maybe none)

**2. Model Training**

**Q:** How to train an ML model to satisfy a chosen fairness metric?
**A:** It depends (maybe impossible)

**3. Verification**

**Q:** How to verify that an ML model satisfies a chosen fairness metric?
**A:** It depends (maybe none)

# Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

# Types of Fairness

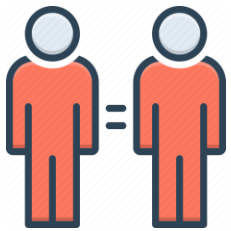
- Our idea of fairness is mainly based on two principles:

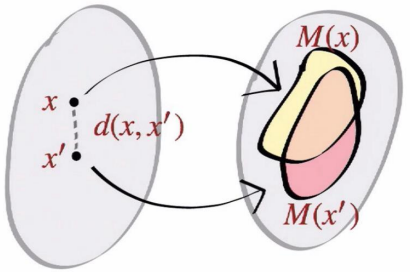
**Group Fairness**

"Models should ***perform equally well*** across different subgroups of a population"

**Individual Fairness**

"***Similar people*** should be ***treated similarly***"






- But…

**How to formally define these notions to measure fairness?**

# Group Measure of Fairness

- **Let's consider a binary classifier**
  - $Y = \{0, 1\}$ to be the ground truth label (e.g., recidivism)
  - $\hat{Y} = \{0, 1\}$ to be the model prediction
  - $G \in \{0, 1\}$ to be the sensitive attribute (e.g., race, gender)

| | | Predicted condition | |
|---|---|---|---|
| | Total population = P + N | **Predicted Positive (PP)** | **Predicted Negative (PN)** |
| **Actual condition** | **Positive (P)** [a] | **True positive** (TP), hit[b] | **False negative** (FN), miss, underestimation |
| | **Negative (N)**[d] | **False positive** (FP), false alarm, overestimation | **True negative** (TN), correct rejection[e] |

- Error rate = $\frac{FP+FN}{TN+FP+FN+TP}$ — Accuracy

Defendants care →

- False Positive Rate = $\frac{FP}{FP+TN}$ — Probability at which non-offenders being predicted to re-offend

Judges care →

- False Negative Rate = $\frac{FN}{FN+TP}$ — Probability at which offenders were predicted to not re-offend?



Source

# COMPAS Case Study

- We consider the COMPAS case study again…

| Black Defendants | Prediction: Low Risk | Prediction: High Risk |
|---|---|---|
| Outcome: No Recidivism | 990 (TN) | 805 (FP) |
| Outcome: Recidivated | 532 (FN) | 1369 (TP) |

| White Defendants | Prediction: Low Risk | Prediction: High Risk |
|---|---|---|
| Outcome: No Recidivism | 1139 (TN) | 349 (FP) |
| Outcome: Recidivated | 461 (FN) | 505 (TP) |

- Error rate $\approx 36.2\%$ $\approx$ Error rate $\approx 36.2\%$
- False Positive Rate $\approx 44.9\%$ $>$ False Positive Rate $\approx 23.5\%$
- False Negative Rate $\approx 28.0\%$ $<$ False Negative Rate $\approx 47.7\%$

**While having the the same error rate, Black defendants have 1.9x higher False Positive Rate!**



Source

# Group Measure of Fairness: Statistical Parity

- **Idea:** Different groups have the same probability of receiving favorable outcomes.

**Definition #1 (Statistical parity):** The classifier is said to satisfy statistical parity if the probability that the algorithm makes a positive prediction ($\hat{Y} = 1$) is the same across different groups

$$\mathbb{P}(\hat{Y} = 1 \mid G = 0) = \mathbb{P}(\hat{Y} = 1 \mid G = 1)$$

where $\mathbb{P}$ is the underlying data distribution.

- **Cons**: Does not take the ground truth label $Y$ into account (different groups have different underlying distributions for $Y$)

# Group Measure of Fairness: Equal Opportunity

- **Idea:** Different groups have the same **true positive rate (TPR)** across different groups

**Definition #2 (Equal Opportunity):** The classifier is said to satisfy equal opportunity if the classifier has the same true positive rate across different groups

$$TP(G = 0, \mathbb{P}) = TP(G = 1, \mathbb{P})$$
$$\Leftrightarrow \mathbb{P}(\hat{Y} = 1 \mid G = 0, \boldsymbol{Y = 1}) = \mathbb{P}(\hat{Y} = 1 \mid G = 1, \boldsymbol{Y = 1})$$

- **Cons**: Does not take error rate into account

| | | Predicted condition | |
|---|---|---|---|
| | Total population = P + N | Predicted Positive (PP) | Predicted Negative (PN) |
| Actual condition | Positive (P) [a] | True positive (TP), hit[b] | False negative (FN), miss, underestimation |
| | Negative (N)[d] | False positive (FP), false alarm, overestimation | True negative (TN), correct rejection[e] |

# Group Measure of Fairness: Equalized Odds

- **Idea:** Different groups have the same **true positive rate (TPR)** and **false positive rate (FPR)** across different groups

**Definition #3 (Equalized Odds):** The classifier is said to satisfy equal opportunity if the classifier has the same TPR and FPR across different groups

$$TP(G = 0, \mathbb{P}) = TP(G = 1, \mathbb{P})$$
$$\Leftrightarrow \mathbb{P}(\hat{Y} = 1 \mid G = 0, \boldsymbol{Y = 1}) = \mathbb{P}(\hat{Y} = 1 \mid G = 1, \boldsymbol{Y = 1})$$

and

$$FP(G = 0, ) = FP(G = 1, \mathbb{P})$$
$$\Leftrightarrow \mathbb{P}(\hat{Y} = 1 \mid G = 0, \boldsymbol{Y = 0}) = \mathbb{P}(\hat{Y} = 1 \mid G = 1, \boldsymbol{Y = 0})$$

| | | Predicted condition | |
|---|---|---|---|
| | Total population = P + N | Predicted Positive (PP) | Predicted Negative (PN) |
| Actual condition | Positive (P) [a] | True positive (TP), hit[b] | False negative (FN), miss, underestimation |
| | Negative (N)[d] | False positive (FP), false alarm, overestimation | True negative (TN), correct rejection[e] |

# Group Measure of Fairness: Calibration

- **Let $S$** be the predicted score (e.g., logits) of the model for a given input
- **Idea:** when two people from different groups get the same predicted score, they should have the same probability of belonging to the favorable class.

**Definition #4 (Calibration):** The classifier is said to satisfy calibration if the probability of $Y = 1$ across different groups is the same given predicted score

$$\mathbb{P}(Y = 1 \mid S = s, G = 0) = \mathbb{P}(Y = 1 \mid S = s, G = 1)$$

i.e., people from different groups with the same predicted score should have the same probability of belonging to $y = 1$

- **Pros**: Well-calibrated confidence score across different groups
- **Cons**: Not compatible with equalized odds

# Impossibility Theorem

- **Unfortunately**, we cannot consider all defined fairness criteria at the same time.

**Theorem (Impossibility):** No classifiers can satisfy three fairness (statistical parity, equalized odds, and calibration) criteria simultaneously.

- i.e., if an algorithm satisfies *statistical parity*, the algorithm ***cannot*** have *equal odds* and *calibration* at the same time.



Kleinberg et al. "Inherent trade-offs in the fair determination of risk scores." 2016

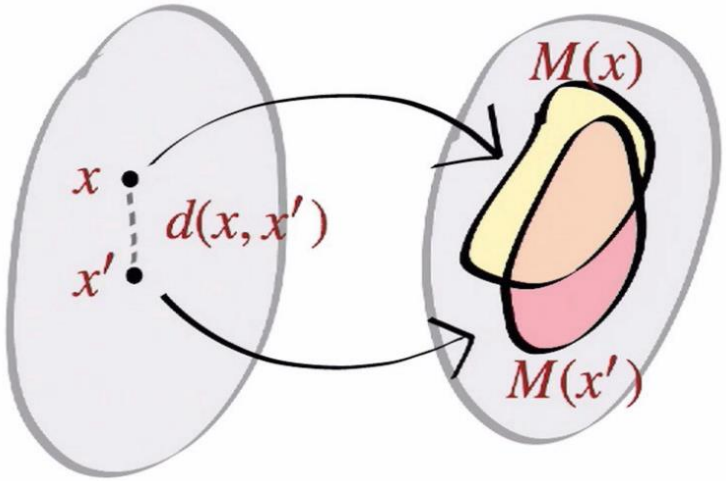
# Individual Fairness

- **Idea**: *Similar* individuals should be treated *similarly*
- **Let** $M(x)$ to be the algorithm output for the individual $x$

**Definition #5 (Individual fairness):** The classifier is said to satisfy individual fairness, if for any $x, x'$

$$|M(x) - M(x')| \leq \delta \quad \text{if} \quad d(x, x') \leq \epsilon$$

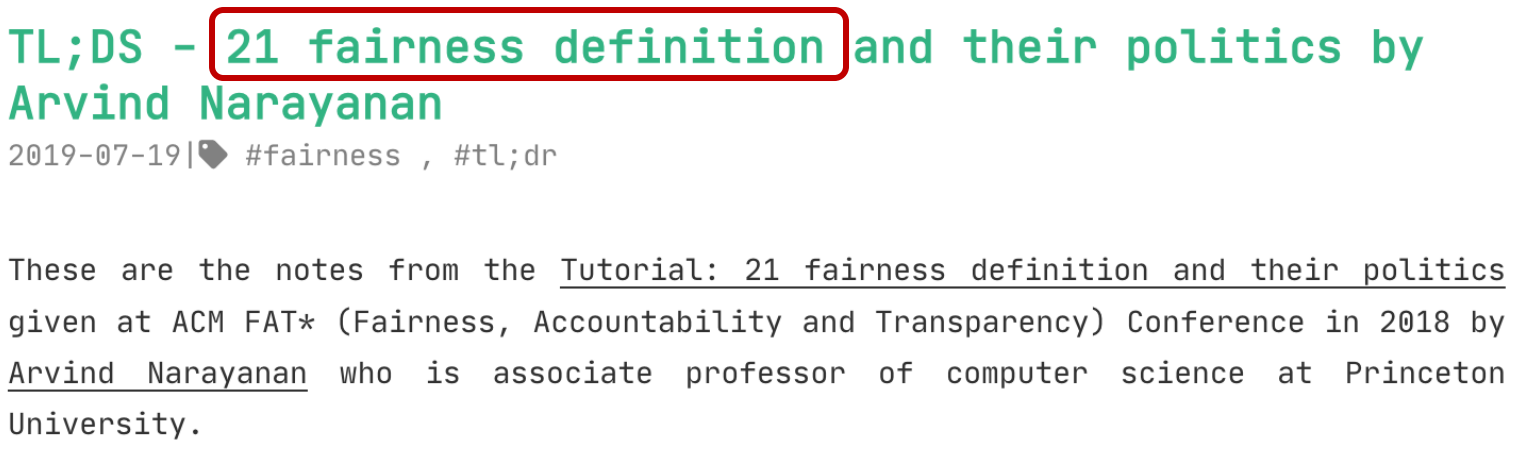
where $d$ is the distance function measure similarity between two individuals.




- **Pros**: can model heterogeneity within each group
- **Cons**: Notion of "similar" is hard to define mathematically, especially in high-dimensions.

# And many more…

- There are **many definitions** of fairness with different criteria.
  - Overall accuracy equality
  - Conditional use accuracy equality
  - Well-carlibration
  - Bayesian Fairness
  - Counterfactual Fairness
  - Generalized Entropy Index
  - Theil Index
  - …




- Fairness is ***not a purely technical issue***. We need to think about context and stake holders. ***Different notions of fairness matter to different stakeholders***.



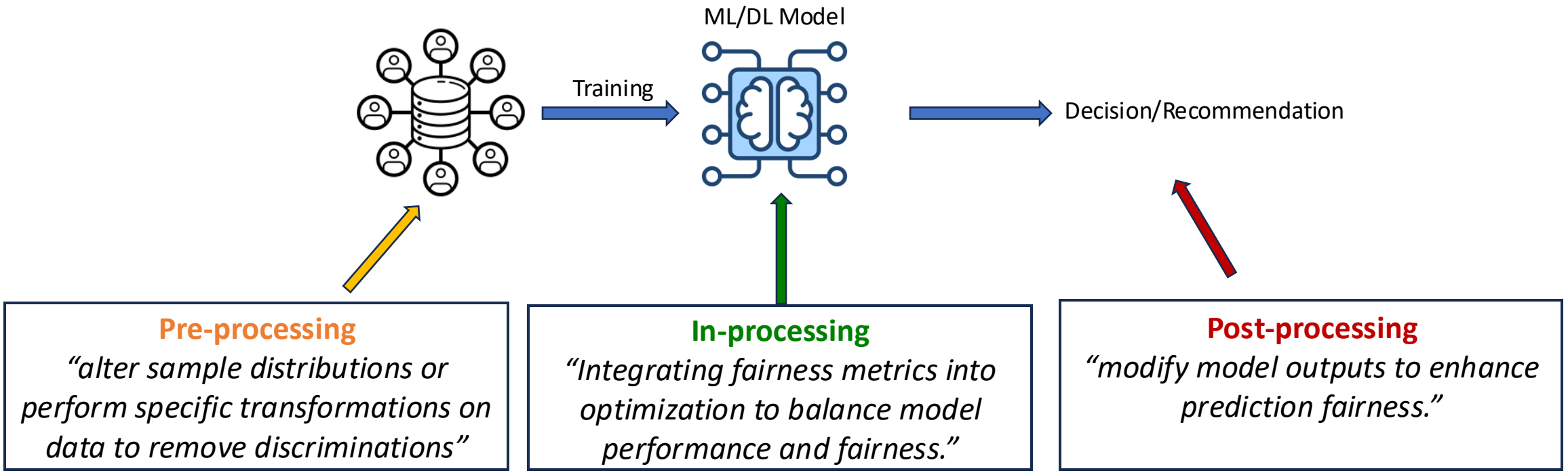
Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." ACM Computing Surveys (2020).

# Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

# Mitigating Algorithmic Bias

- Technical fairness interventions operate in different location in ML pipeline



# Pre-processing Approach

- **The goal** is to **make the classifier "immune" to one or more sensitive variables** (e.g., gender, race, etc.).

- **Naïve approach:** removing sensitive attribute?
  - COMPAS does not use race as an input to the algorithm but still gives very different outcomes for white vs black defendants!
  - **Reason**: other features (e.g. zip code) may correlate with the sensitive feature

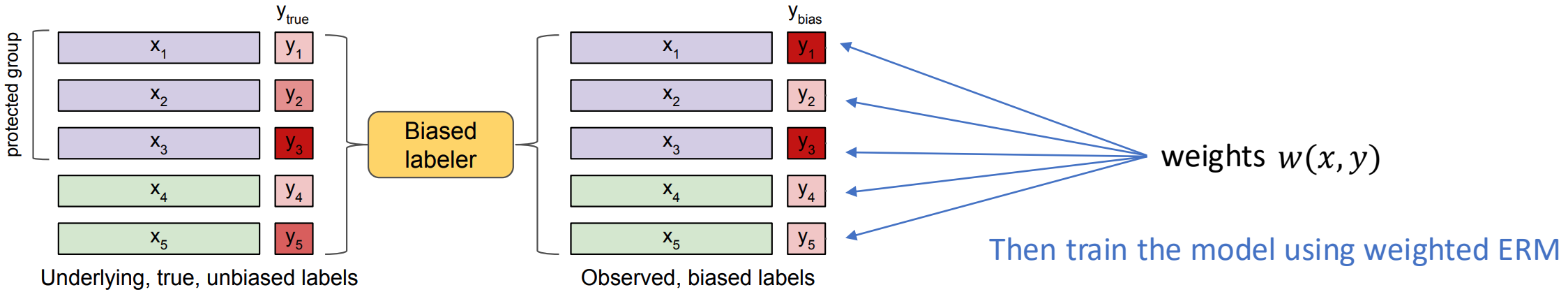
| Black Defendants | Prediction: Low Risk | Prediction: High Risk |
|---|---|---|
| Outcome: No Recidivism | 990 (TN) | 805 (FP) |
| Outcome: Recidivated | 532 (FN) | 1369 (TP) |

| White Defendants | Prediction: Low Risk | Prediction: High Risk |
|---|---|---|
| Outcome: No Recidivism | 1139 (TN) | 349 (FP) |
| Outcome: Recidivated | 461 (FN) | 505 (TP) |

- Error rate ≈ 36.2% ≈ Error rate ≈ 36.2%
- False Positive Rate ≈ 44.9% > False Positive Rate ≈ 23.5%
- False Negative Rate ≈ 28.0% < False Negative Rate ≈ 47.7%

# Pre-processing: Reweighting Training Data

- **Goal:** reweighting the training data to reverse the bias in the training data.
- **Settings**:
  - There exists an *unbiased* labeling function $y_{true}$: $\mathcal{X} \to [0,1], \mathbb{P}(y|x) = y_{true}(x)$
  - But we can only observe *biased labels* $y_{bias}$: $\mathcal{X} \to [0,1], \mathbb{P}(y|x) = y_{bias}(x)$ produced by annotators or collectors who are biased against a certain group (labels for other groups remain accurate).






Jiang, Heinrich, and Ofir Nachum. "Identifying and correcting label bias in machine learning." ICML, 2020.

# Pre-processing: Reweighting Training Data

- We can express notions of **fairness via linear constraints**. **Let**:
  - $h(y|x)$ be the probability of ML models labeling $x$ by $y$.
  - $P_X = \mathbb{E}_{\mathbb{P}}\, y_{true}(x)$ be the proportion of input $x$ having a positive label.
  - $\{G = b\}$ to be the protected group, probability of $x$ in the protected group is $g(x)$
  - $P_b = \mathbb{E}_{\mathbb{P}}\, y_{true}(x, x \in \{G = b\})$ be proportion of positive input $x$ in protected group.
- The classifier $h(\cdot)$ is said to satisfy ***equal opportunity*** if

$$\mathbb{P}\left(\hat{Y} = 1 \mid G = b, Y = 1\right) - \mathbb{P}\left(\hat{Y} = 1 \mid G = w, Y = 1\right) = 0$$

$$\Leftrightarrow \mathbb{E}_{\mathbb{P}}\left[\left(\frac{h(1|x)\, g(x)\, y_{true}(x)}{P_b} - \frac{h(1|x)\, y_{true}(x)}{P_X}\right)\right] = 0$$

$$\mathbb{E}_{\mathbb{P}}[(h(1|x)\; c(x, y))] = 0$$

**We could do similarly for other notions of fairness**



reference

# Pre-processing: Reweighting Training Data

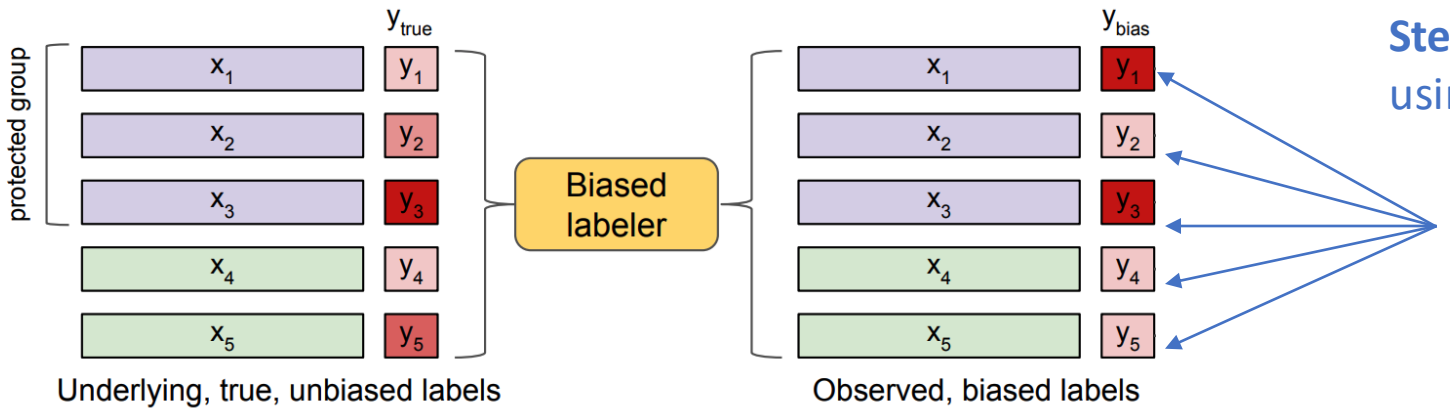
- Notice that $y_{bias}$ can also be seen as an ML model having bias score

$$\mathbb{E}_{\mathbb{P}}\left[\left(h(1|x)\; c(x,y)\right)\right] = \epsilon$$

- **Proposition**: Assume that $y_{bias}$ **is the closest labeling function** (defined via KL divergence) to the true labeling function $y_{true}$, then, $y_{bias}$ **has a closed form**

$$y_{bias}(y|x) \propto y_{true}(y|x)\exp\left(-\lambda_\epsilon c(x,y)\right) \quad \text{for some param } \lambda_\epsilon$$

- **Weighting scheme is then defined as**




**Step 1**: optimize $\lambda_\epsilon$ to satisfy fairness constraint.
**Step 2**: The model is then trained on the dataset using these weights

$$w(x,y) = \frac{\exp\left(\lambda_\epsilon c(x,y)\right)}{\exp\left(\lambda_\epsilon c(x,0)\right) + \exp\left(\lambda_\epsilon c(x,1)\right)}$$

**Guarantee?**

# Pre-processing: Reweighting Training Data

**Theorem (statistical consistency):** Let $h^\star$ minimize the re-weighted ERM with observed labels $y_{bias}$ over all Lipschiz classifiers, then the mean square loss between $h^\star$ and $y_{true}$ is bounded.

- **Empirical results:** smallest bias with acceptable utility trade-off

| | Unconst. | Post-fix | Lagrange | Ours |
|---|---|---|---|---|
| Adult error | 14.15 | 16.6 | 20.47 | 16.51 |
| Adult bias | 0.1173 | 0.0129 | 0.0198 | **0.0037** |
| Bank error | 9.41 | 9.7 | 10.46 | 9.63 |
| Bank bias | 0.0349 | 0.0068 | 0.0126 | **0.0056** |
| COMPAS error | 31.49 | 32.53 | 40.16 | 35.44 |
| COMPAS bias | 0.2045 | 0.0201 | 0.0495 | **0.0155** |
| Crime error | 11.62 | 32.06 | 28.46 | 30.06 |
| Crime bias | 0.4211 | 0.0653 | 0.1538 | **0.0107** |
| German error | 24.85 | 24.85 | 25.45 | 25.15 |
| German bias | 0.0766 | 0.0346 | 0.0410 | **0.0137** |

# In-Processing Approach

- **Goal**: Imposing fairness metrics into optimization to balance model performance and fairness.
- **Consider (again):** Binary classifier $h(y|x)$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$



paper

# In-Processing Approach

- **Goal**: Imposing fairness metrics into optimization to balance model performance and fairness.
- **Consider (again):** Binary classifier $h(y|x)$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$

- **Idea:**

$$\begin{aligned} \min_\theta \quad & \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \\ \text{s.t.} \quad & TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P}) \end{aligned}$$

Minimize expected loss (ERM)
Subject to **equal opportunity** constraints

**Step 1**: Propose a well-structured model that has potential



paper

# In-Processing Approach

- **Goal**: Imposing fairness metrics into optimization to balance model performance and fairness.
- **Consider (again):** Binary classifier $h(y|x)$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$

- **Idea:**

$$\begin{aligned} \min_\theta \quad & \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \\ \text{s.t.} \quad & TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P}) \end{aligned} \longrightarrow \begin{aligned} \min_\theta \quad & \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \\ \text{s.t.} \quad & |TP_\theta(b, \mathbb{P}) - TP_\theta(w, \mathbb{P})| \leq \epsilon \end{aligned}$$

difference in equalized opportunity (DEO)

**Step 1**: Propose a well-structured model that has potential

→

**Step 2**: Realize that it's difficult to solve. Perform relaxation…



paper

# In-Processing Approach

- **Goal**: Imposing fairness metrics into optimization to balance model performance and fairness.
- **Consider (again):** Binary classifier $h(y|x)$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$

- **Idea:**

$$\min_\theta \ \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \quad \text{s.t.} \quad TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P})$$

$$\min_\theta \ \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \quad \text{s.t.} \quad |TP_\theta(b, \mathbb{P}) - TP_\theta(w, \mathbb{P})| \leq \epsilon$$

$$\min_\theta \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] + \lambda\,|TP_\theta(b, \mathbb{P}) - TP_\theta(w, \mathbb{P})|$$

**Step 1**: Propose a well-structured model that has potential

**Step 2**: Realize that it's difficult to solve. Perform relaxation…

**Step 3**: Realize that it's *still* difficult to solve. Relax *further*…

Replace true dist by data (empirical)
Add regularization?
Add robustness?
Add kernel?
…



paper

# In-Processing Approach

- **Goal**: Imposing fairness metrics into optimization to balance model performance and fairness.
- **Consider (again):** Binary classifier $h(y|x)$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$

- **Idea:**

$$\begin{array}{ll} \min_\theta & \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \\ \text{s.t.} & TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P}) \end{array} \rightarrow \begin{array}{ll} \min_\theta & \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] \\ \text{s.t.} & |TP_\theta(b, \mathbb{P}) - TP_\theta(w, \mathbb{P})| \leq \epsilon \end{array} \rightarrow \min_\theta \mathbb{E}_{\mathbb{P}}[\ell(\theta, X, Y, G)] + \lambda |TP_\theta(b, \mathbb{P}) - TP_\theta(w, \mathbb{P})|$$

Relax *further*...

**Step 1**: Propose a well-structured model that has potential → **Step 2**: Realize that it's difficult to solve. Perform relaxation... → **Step 3**: Realize that it's *still* difficult to solve. Relax *further*... → **Step 4**: Define **Probabilistic Equal Opportunity**



paper paper

# Probabilistic Equal Opportunity

- Recall the definition of equal opportunity

$$\mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = b, Y = 1\right) = \mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = w, Y = 1\right)$$

$$\Leftrightarrow \mathbb{P}(\boldsymbol{h_\theta(X) \geq \tau} \mid G = b, Y = 1) = \mathbb{P}(\boldsymbol{h_\theta(X) \geq \tau} \mid G = w, Y = 1)$$

We can rewrite the TPR as the expectation of the indicator function

$$\Leftrightarrow \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = b, Y = 1\right) = \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = w, Y = 1\right)$$



paper paper

# Probabilistic Equal Opportunity

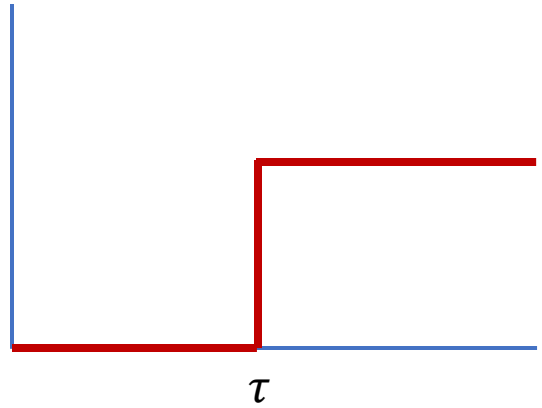
- Recall the definition of equal opportunity

$$\mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = b, Y = 1\right) = \mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = w, Y = 1\right)$$
$$\Leftrightarrow \mathbb{P}(\boldsymbol{h_\theta(X)} \geq \boldsymbol{\tau} \mid G = b, Y = 1) = \mathbb{P}(\boldsymbol{h_\theta(X)} \geq \boldsymbol{\tau} \mid G = w, Y = 1)$$

We can rewrite the TPR as the expectation of the indicator function

$$\Leftrightarrow \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = b, Y = 1\right) = \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = w, Y = 1\right)$$

- However, the indicator function is non-continuous, making it's very hard to solve the problem






paper paper

# Probabilistic Equal Opportunity

- Recall the definition of equal opportunity

$$\mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = b, Y = 1\right) = \mathbb{P}\left(\widehat{\boldsymbol{Y}} = \mathbf{1} \mid G = w, Y = 1\right)$$
$$\Leftrightarrow \mathbb{P}(\boldsymbol{h_\theta(X) \geq \tau} \mid G = b, Y = 1) = \mathbb{P}(\boldsymbol{h_\theta(X) \geq \tau} \mid G = w, Y = 1)$$

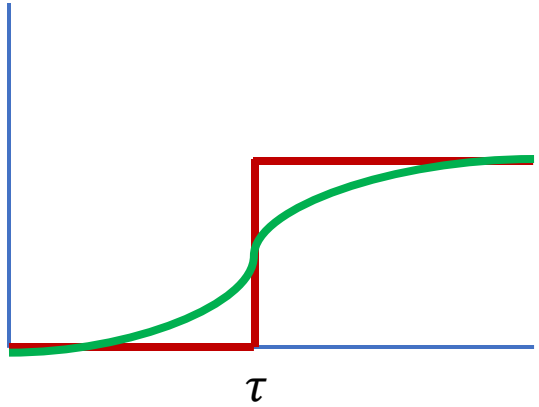
We can rewrite the TPR as the expectation of the indicator function

$$\Leftrightarrow \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = b, Y = 1\right) = \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{h_\theta(X) \geq \tau} \mid G = w, Y = 1\right)$$

- However, the indicator function is non-continuous, making it's very hard to solve the problem
- We thus relax to probabilistic equal opportunity

**Definition (Probabilistic Equal Opportunity):**

$$\mathbb{E}_{\mathbb{P}}\left[h_\theta(X) \mid G = b, Y = +\right] = \mathbb{E}_{\mathbb{P}}\left[h_\theta(X) \mid G = w, Y = +\right]$$





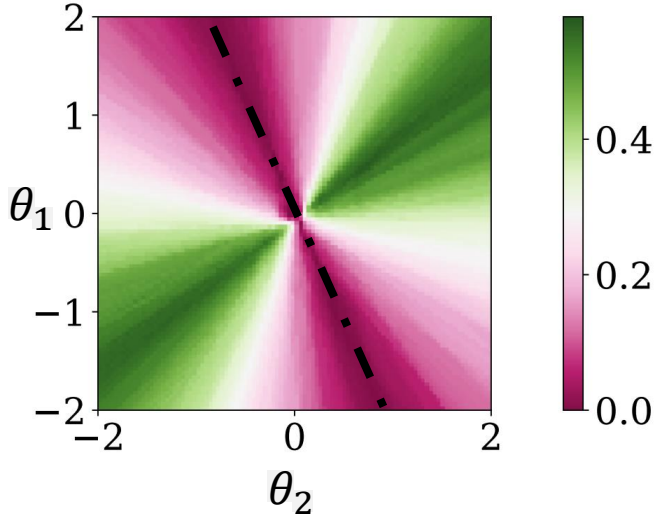

paper paper

# Probabilistic Equal Opportunity

- **Advantage:** the loss landscape of Probabilistic Equal Opportunity **is smoother** than Equal Opportunity while **having similar optimality** (black dash lines)

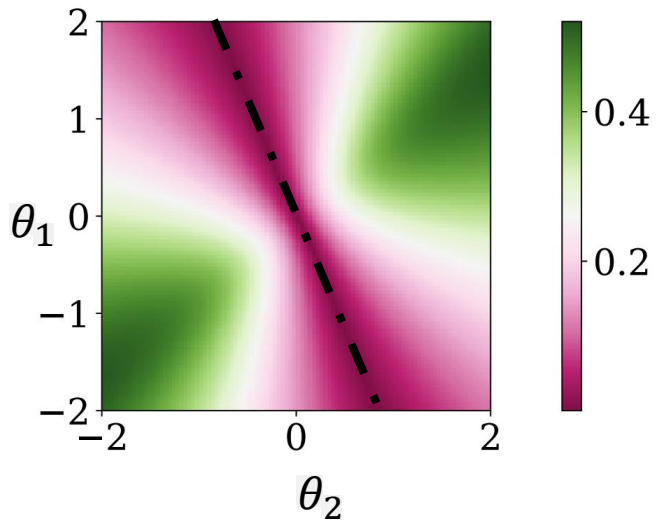
Difference of
Equal Opportunity

$$|\mathbb{P}[h_\theta(X) \geq \tau \mid G = b, Y = +] - \mathbb{P}[h_\theta(X) \geq \tau \mid G = w, Y = +]|$$




Difference of
Probabilistic Equal Opportunity

$$|\mathbb{E}_{\mathbb{P}}[h_\theta(X) \mid G = b, Y = +] - \mathbb{E}_{\mathbb{P}}[h_\theta(X) \mid G = w, Y = +]|$$




Taskesen et al. "A statistical test for probabilistic fairness." FaCCT. 2021.

# In-Processing Approach: FairERM

- **Empirical result:**

Better trade-off than vanilla SVM

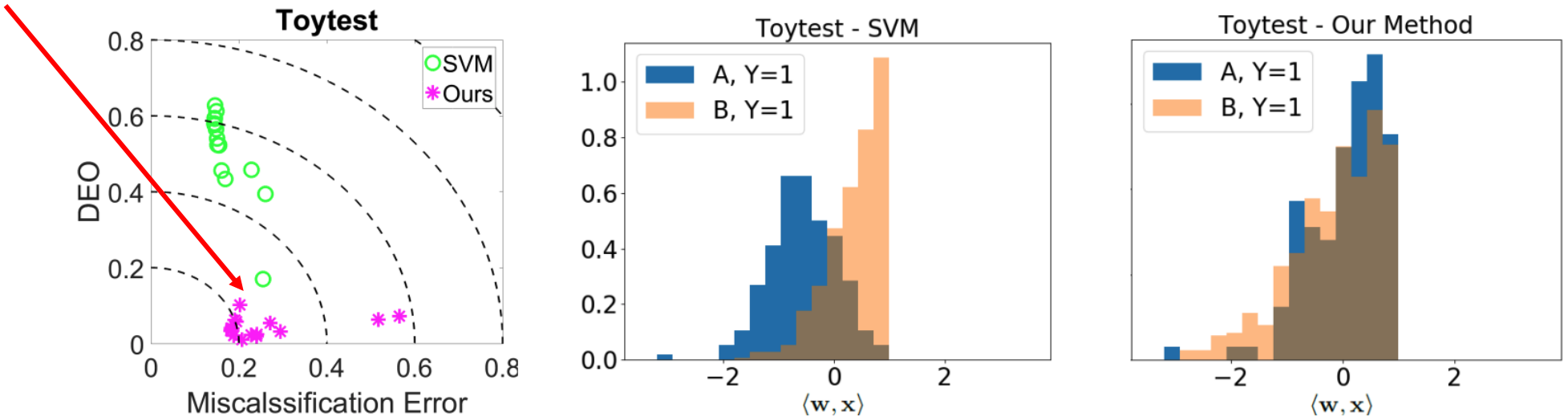



Figure 1: Left: Test classification error and DEO for different values of hyperparameter $C$ for standard linear SVM (green circles) and our modified linear SVM (magenta stars). Center and Rights: Histograms of the distribution of the values $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ for the two groups ($a$ in blue and $b$ in light orange) for test examples with label equals to $+1$. The results are collected by using the optimal validated model for the classical linear SVM (Center) and for our linear method (Right).



Donini, Michele, et al. "Empirical risk minimization under fairness constraints." NIPS, 2018.

# Post-Processing Approach: Thresholding

- **Goal**: calibrating the decision threshold ($\tau$) to enhance fairness
- **Consider (again):** score function $h_\theta$ and two groups $G = \{w, b\}$

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$

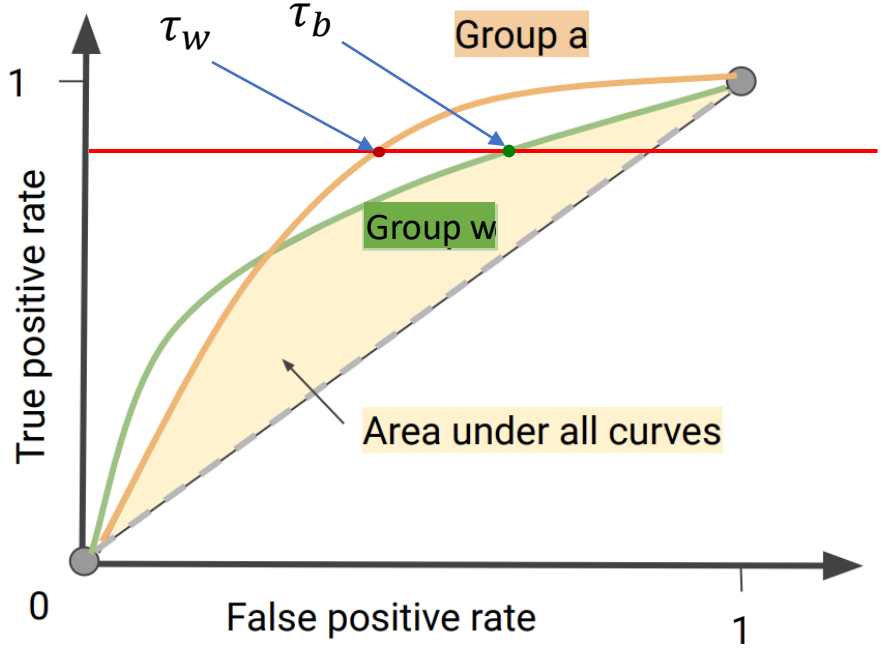
- Varying the threshold $\tau$ **changes the trade-off between** false positive rate (**FPR**) **and** false negative rate (**FNR**).
- In case the score function $h_\theta$ **satisfies the fairness constraints already**, we can just choose $\tau$ to balance FPR and FNR, thus minimizing the expected loss.
- In case the score function $h_\theta$ **does not satisfy the fairness constraints**, we might need to use **different threshold $\tau_G$ for different groups.**



Hardt et al. "Equality of opportunity in supervised learning." *NIPS*, 2016.

# Post-Processing Approach: Thresholding

- To choose the threshold $\tau_G$, we can **analyze the ROC (Receiver Operator Characteristic) curve of the score**, which captures the FPR and TPR at different thresholds.
- For the **Equal Opportunity** constraint, we can choose the thresholds for each group such that the **TPRs between groups match**.

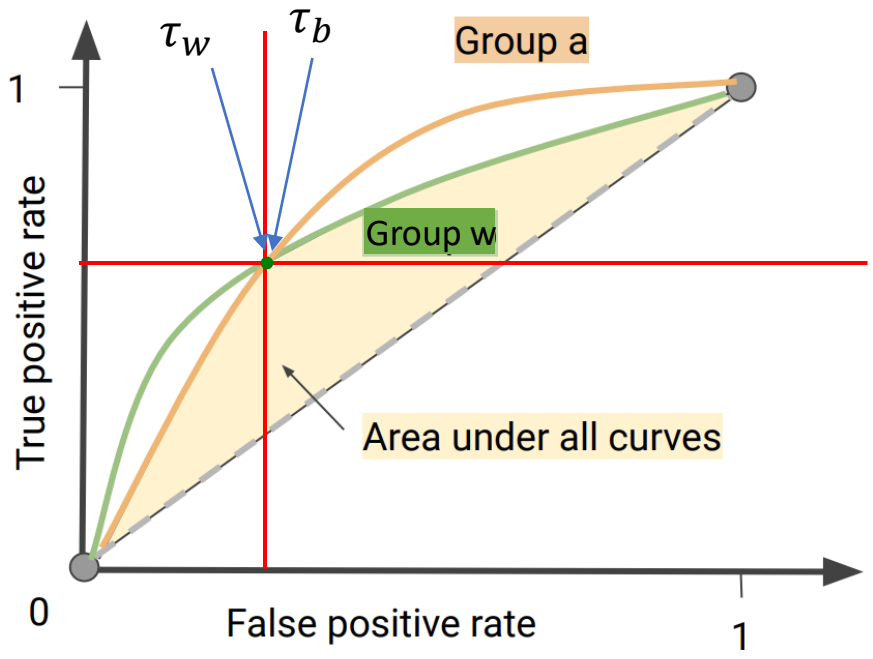





Hardt et al. "Equality of opportunity in supervised learning." *NIPS*, 2016.

# Post-Processing Approach: Thresholding

- To choose the threshold $\tau_G$, we can **analyze the ROC (Receiver Operator Characteristic) curve of the score**, which captures the FPR and TPR at different thresholds.
- For the **Equal Opportunity** constraint, we can choose the thresholds for each group such that the **TPRs between groups match**.
- Thresholds $\tau_w, \tau_b$ that makes the classifier satisfy **equalized odds** are only where **ROC curves of different groups intersect**.

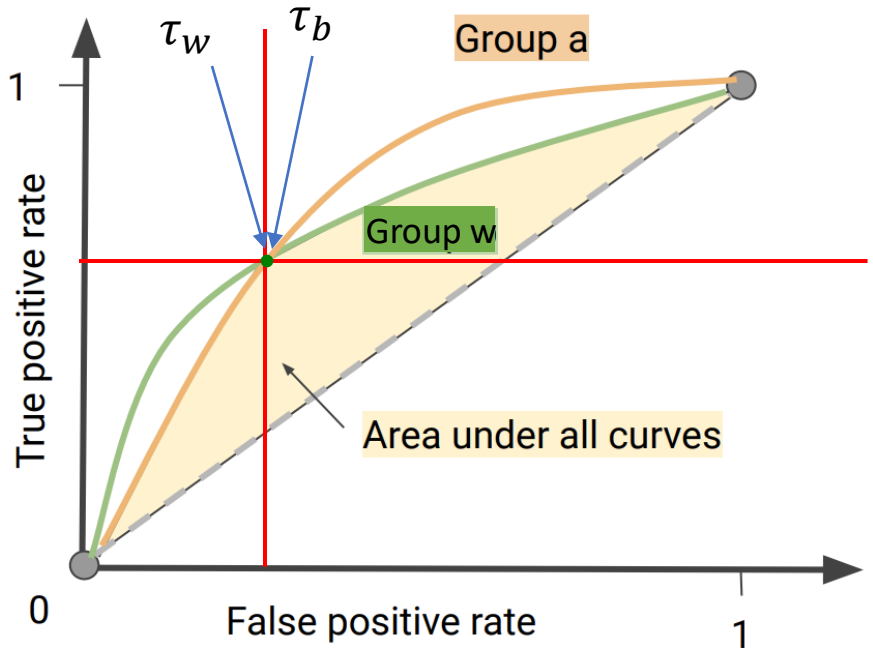





Hardt et al. "Equality of opportunity in supervised learning." *NIPS*, 2016.

# Post-Processing Approach: Thresholding

- To choose the threshold $\tau_G$, we can **analyze the ROC (Receiver Operator Characteristic) curve of the score**, which captures the FPR and TPR at different thresholds.
- For the **Equal Opportunity** constraint, we can choose the thresholds for each group such that the **TPRs between groups match**.
- Thresholds $\tau_w, \tau_b$ that makes the classifier satisfy **equalized odds** are only where **ROC curves of different groups intersect**.
  - What if the ROC curves of different groups do not intersect at all (except trivial endpoints)?
    → **Add randomization** to classifier's output to fill out the span of possible classifiers, allowing for intersections (area under all curves).
    (read more)






Hardt et al. "Equality of opportunity in supervised learning." *NIPS*, 2016.

# Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

# Fairness Verification

- **Goal**: An AI company claims that the blue-box is a fair classifier. Governments/Legislators want to verify the claim.
- **Consider (again):** binary classifier

$$\hat{Y} = \begin{cases} + & \text{if } h_\theta(X, G) \geq \tau \\ - & \text{otherwise} \end{cases}$$

$\mathbb{P}$ is the true distribution of $(X, Y, G)$


- **Testing fairness:**
  - **Step 1**: Agree on a notion of fairness (e.g., equal opportunity)
  - **Step 2**: Run hypothesis testing

Null Hypothesis $\mathcal{H}_0$: The classifier is fair $\quad TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P})$

Alterantive Hypothesis $\mathcal{H}_1$: The classifier is *not* fair $\quad TP_\theta(b, \mathbb{P}) \neq TP_\theta(w, \mathbb{P})$

- **Determine whether we accept the null hypothesis or not?**



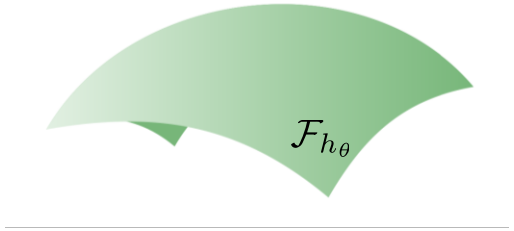
Paper 1

# Fairness Verification

- **Idea**: Look at the manifold of all data distributions with which our classifier is fair

$$\mathcal{F}_{h_\theta} = \{\mathbb{Q} \in \mathcal{P} : \text{ the classifier } h_\theta \text{ is fair relative to } \mathbb{Q}\}$$

  - If our data distribution $\mathbb{P} \in \boldsymbol{\mathcal{F}_{h_\theta}}$, we **accept** the null hypothesis $\mathcal{H}_0$
  - If our data distribution $\mathbb{P} \notin \boldsymbol{\mathcal{F}_{h_\theta}}$, we **reject** the null hypothesis $\mathcal{H}_0$



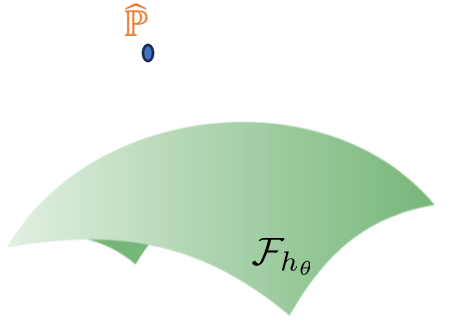



Paper 1

# Fairness Verification

- **Idea**: Look at the manifold of all data distributions with which our classifier is fair

$$\mathcal{F}_{h_\theta} = \{\mathbb{Q} \in \mathcal{P} : \text{ the classifier } h_\theta \text{ is fair relative to } \mathbb{Q}\}$$




  - If our data distribution $\mathbb{P} \in \boldsymbol{\mathcal{F}_{h_\theta}}$, we **accept** the null hypothesis $\mathcal{H}_0$
  - If our data distribution $\mathbb{P} \notin \boldsymbol{\mathcal{F}_{h_\theta}}$, we **reject** the null hypothesis $\mathcal{H}_0$

- **Questions**:
  - No access to the true data distribution $\mathbb{P}$ → use empirical data distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_i \text{Dirac}_{(x,g,y,)}$

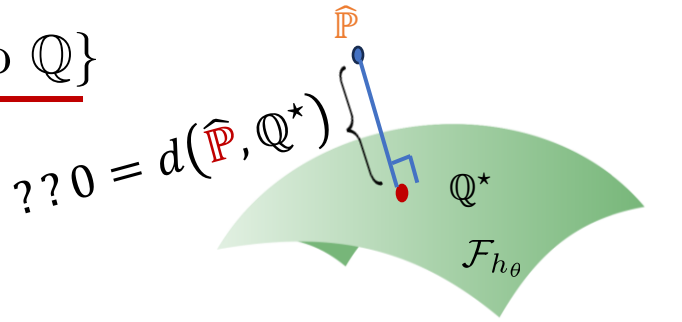


Paper 1

# Fairness Verification

- **Idea**: Look at the manifold of all data distributions with which our classifier is fair

$$\mathcal{F}_{h_\theta} = \{\mathbb{Q} \in \mathcal{P} : \text{ the classifier } h_\theta \text{ is fair relative to } \mathbb{Q}\}$$




  - If our data distribution $\mathbb{P} \in \boldsymbol{\mathcal{F}_{h_\theta}}$, we **accept** the null hypothesis $\mathcal{H}_0$
  - If our data distribution $\mathbb{P} \notin \boldsymbol{\mathcal{F}_{h_\theta}}$, we **reject** the null hypothesis $\mathcal{H}_0$

- **Questions**:
  - No access to the true data distribution $\mathbb{P}$ → use empirical data distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_i \text{Dirac}_{(x,g,y,)}$
  - Construction of $\boldsymbol{\mathcal{F}_{h_\theta}}$ → Use **probabilistic equal opportunity**

$$\mathcal{F}_h = \{\mathbb{Q} \in \mathcal{P} \text{ such that } \mathbb{E}_{\mathbb{Q}}(h(X)|G = b, Y = 1) = \mathbb{E}_{\mathbb{Q}}(h(X)|G = w, Y = 1)\}$$

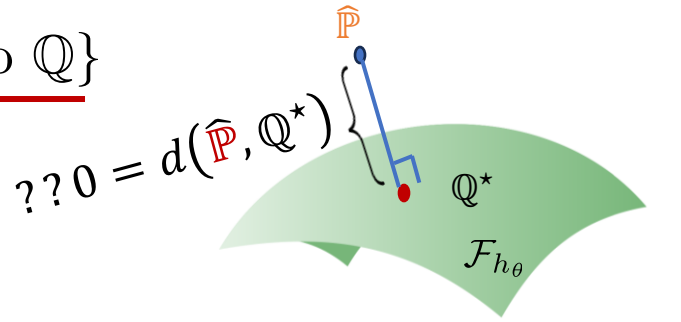


Paper 1

# Fairness Verification

- **Idea**: Look at the manifold of all data distributions with which our classifier is fair

$$\mathcal{F}_{h_\theta} = \{\mathbb{Q} \in \mathcal{P} : \text{ the classifier } h_\theta \text{ is fair relative to } \mathbb{Q}\}$$




  - If our data distribution $\mathbb{P} \in \boldsymbol{\mathcal{F}_{h_\theta}}$, we **accept** the null hypothesis $\mathcal{H}_0$
  - If our data distribution $\mathbb{P} \notin \boldsymbol{\mathcal{F}_{h_\theta}}$, we **reject** the null hypothesis $\mathcal{H}_0$

- **Questions**:
  - No access to the true data distribution $\mathbb{P}$ → use empirical data distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_i \text{Dirac}_{(x,g,y,)}$
  - Construction of $\boldsymbol{\mathcal{F}_{h_\theta}}$ → Use **probabilistic equal opportunity**

$$\mathcal{F}_h = \{\mathbb{Q} \in \mathcal{P} \text{ such that } \mathbb{E}_{\mathbb{Q}}(h(X)|G = b, Y = 1) = \mathbb{E}_{\mathbb{Q}}(h(X)|G = w, Y = 1)\}$$

  - Check if $\mathbb{P} \in \boldsymbol{\mathcal{F}_{h_\theta}}$ → Look at the projection distance of $\mathbb{P}$ to $\boldsymbol{\mathcal{F}_{h_\theta}}$

$$\mathcal{H}_0: \quad \inf_{\mathbb{Q} \in \boldsymbol{\mathcal{F}_{h_\theta}}} d(\hat{\mathbb{P}}, \mathbb{Q}) = 0$$

where $d(\hat{\mathbb{P}}, \mathbb{Q})$ is the distance between two distributions, e.g., Wasserstein distance.



Paper 1

# Fairness Verification

- **Testing fairness:**
  - **Step 1**: Agree on a notion of fairness (e.g., equal opportunity)
  - **Step 2**: Run hypothesis testing

Null Hypothesis $\mathcal{H}_0$: The classifier is fair $\quad TP_\theta(b, \mathbb{P}) = TP_\theta(w, \mathbb{P})$

Alterantive Hypothesis $\mathcal{H}_1$: The classifier is *not* fair $\quad TP_\theta(b, \mathbb{P}) \neq TP_\theta(w, \mathbb{P})$

- **Challenges**:
  - Multiple criteria (e.g., equalized odds)
  - Low sample size
  - Conditional probabilities in formula (e.g., $\mathbb{P}(\hat{Y} = 1 \mid G = 1, \boldsymbol{Y = 0})$) makes optimization more difficult


Paper 1 Paper 2

# Content

- Algorithmic Bias & Fairness
- Formal Definition
- Mitigating Algorithmic Bias
- Fairness Verification
- Unique Challenges

# Beyond Binary Classification

- **Most of fairness papers are about binary linear classifiers.**
  - which is understandable as many controversial applications in the domain involve binary decisions, such as hiring/not hiring, or offering a loan or not.
- Extensions to other ML problems are **more complex and under-explored**. The following is several notable extensions:
  - **Fair regression**: Aim to minimize the differences between the predicted and the ground truth value
    ($\hat{Y}$ and $Y$ are continuous value instead of binary or categorial).
  - **Recommender systems and ranking**: Group-based fairness in top-k ranking, individual fairness in ranking.
  - **Un/self-supervised learning**: fair clustering, fair representation learning, fair transfer learning, etc.

# Other Challenges

- Similar to other trustworthy aspects, imposing **fairness often leads to a trade-off in the model performance**; balancing the trade-off is crucial for practical use.
- **(Dis)agreement and incompatibility of "fairness".** There is little consensus in the literature about different notions of fairness. Unfortunately, we cannot combine different notions of fairness together (impossibility theorem).
  - Paper and paper even argue that some mathematical formulations of fairness do not match with the legal understanding of "fairness".
- **Lack of realistic data.** Current literature mostly relies on convenience datasets, often from the UCI repository, with limited context.
- Practical use of fairness algorithms **requires considerations of context, causality, and legal policy.**