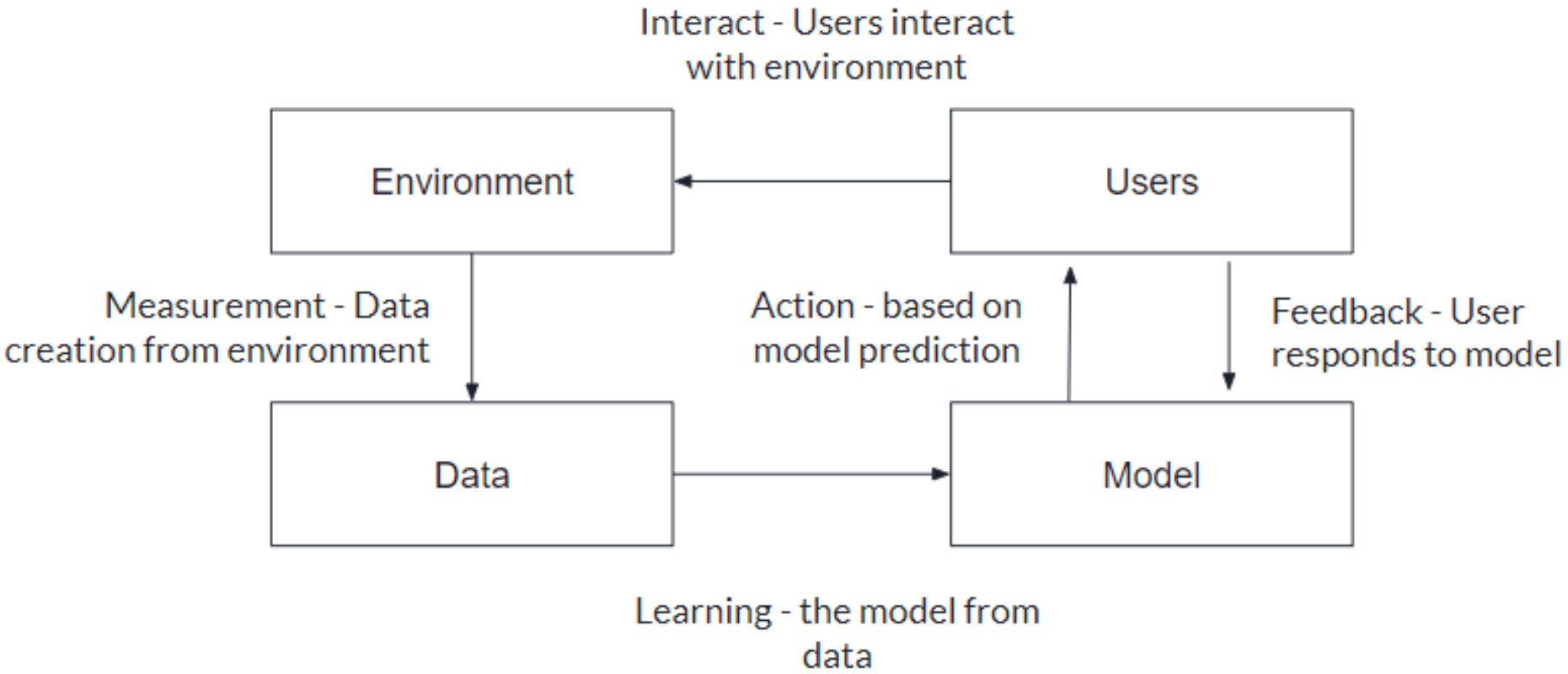
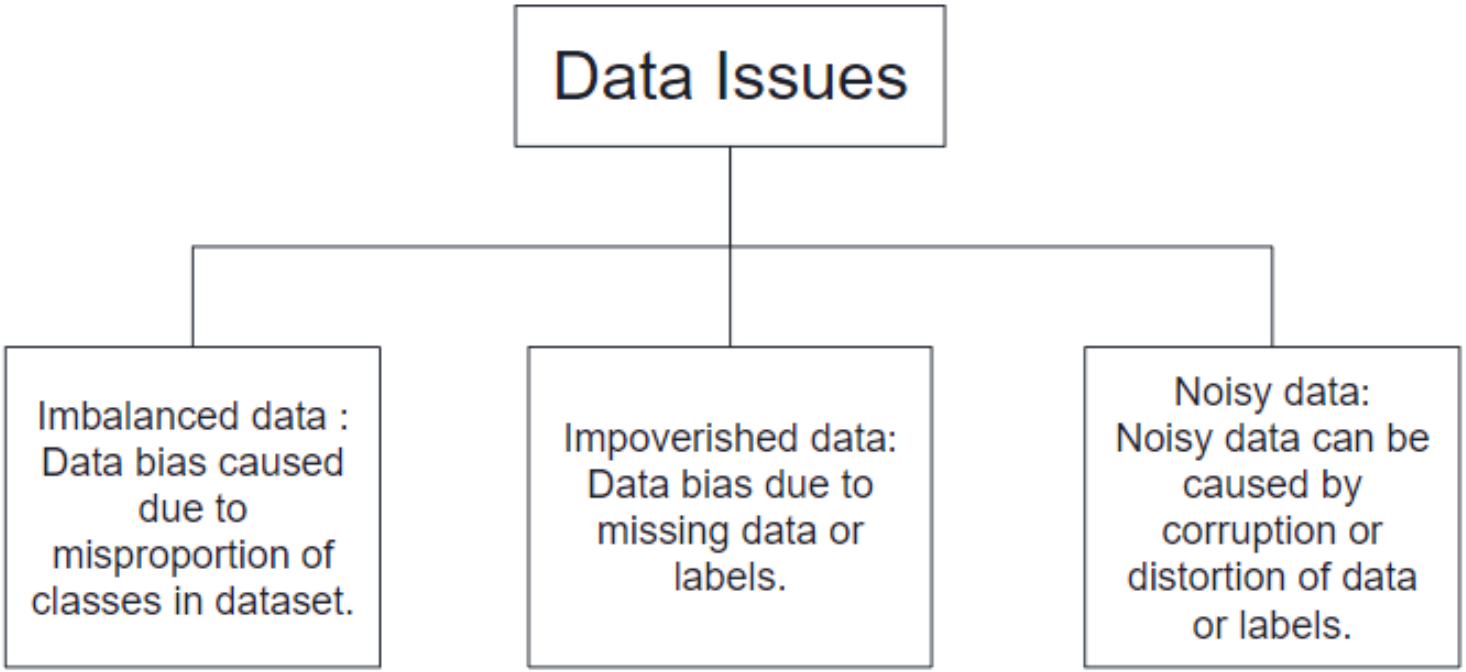
Fairness in AI/ML

Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

Stages of ML system



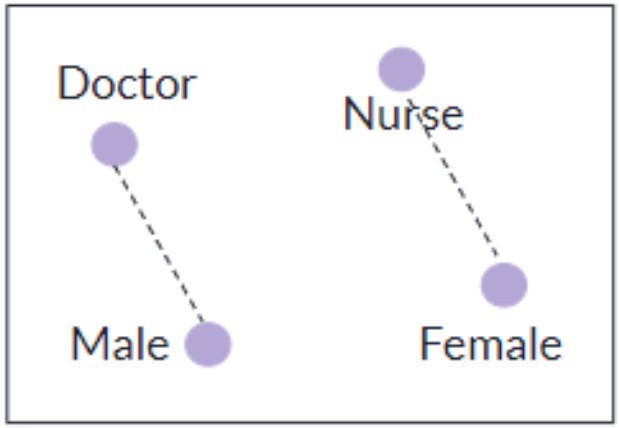
Unfairness due to data issues



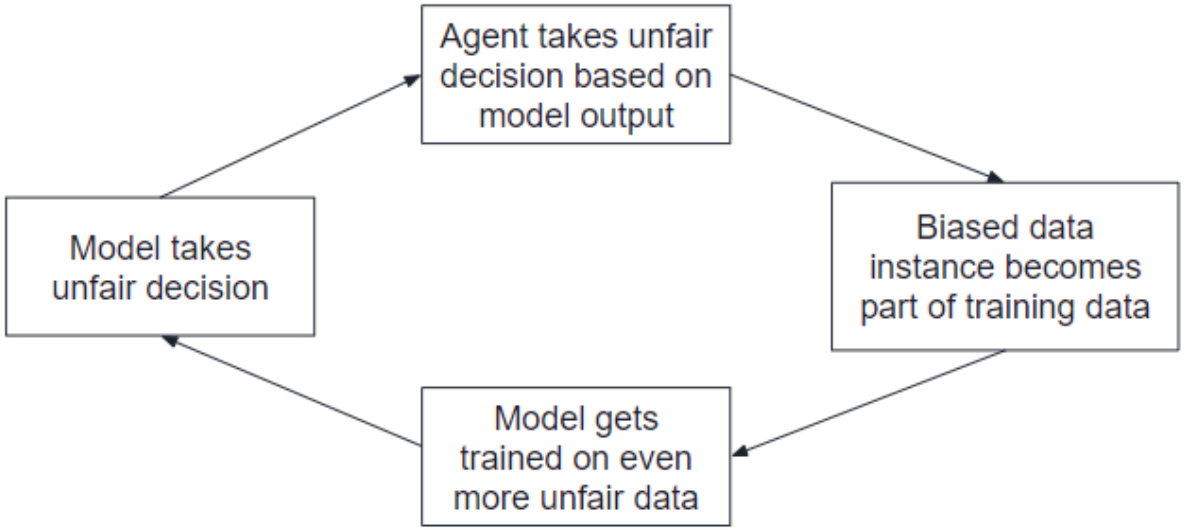
- Assumption that training and testing data are sampled from the same distribution may not always be true.
- ML models are trained based on the above assumption leading to unfairness issues.

Unfairness due to model issues

- Models can amplify disparities in data, including stereotypes. For example, embeddings trained on news data were shown to correlate gender with specific occupations (e.g., 'nurse' and 'female,' 'doctor' and 'male').
- Accuracy optimization often prioritizes the majority class, potentially worsening outcomes for smaller or underrepresented classes. Example: In a dataset with 95% 'majority class' and 5% 'minority class,' a model could achieve 95% overall accuracy by ignoring the minority class entirely, leaving its predictions unfair or unusable for that group.



Unfairness due to Feedback Loops



Loop extends to downstream tasks

- Order of search results - determines clicks.
- Predictive policing - more police to high-crime areas.
- Decision to detain a defendant affects probability of pleading guilty.

1. Initial Model Decision: model sends more police patrols to certain areas
2. Agent Action: Police patrol those areas frequently
3. Outcome: More crimes these areas, not cuz more crime, but more policing
4. Feedback Loop: newly recorded crime data reinforces model's unfair belief

How to do Fair Classification?

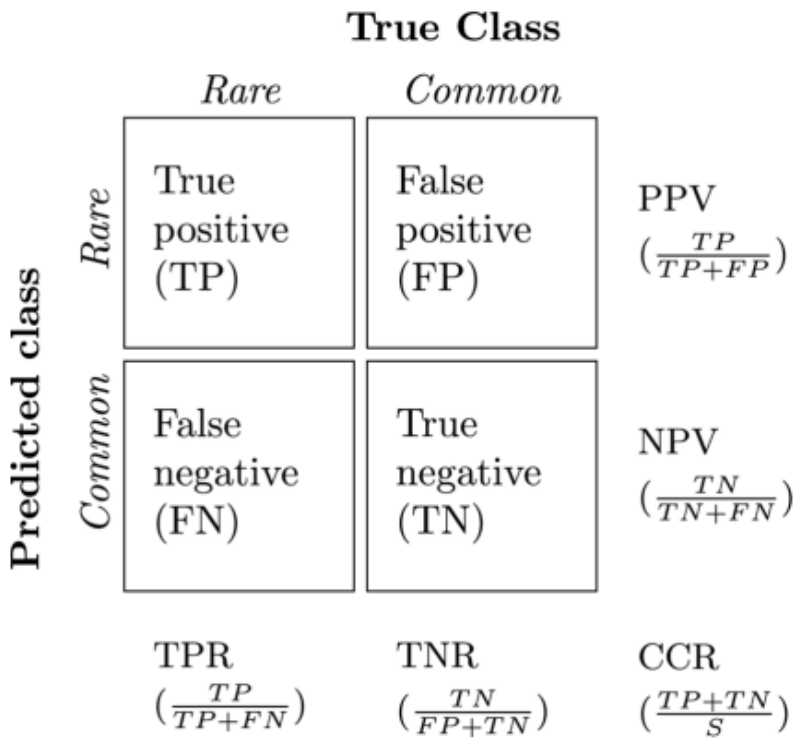
***S-blindness**: Removing or ignoring the "membership in A"*

This fails because membership in A may be encoded in other attributes.

***Awareness**: Assign each individual a representation by being aware of membership in group A.*

1. **Individual fairness**: If two individuals are close on the similarity metric, they should be close on the treatment metric.
2. **Group fairness**: It ensures some form of statistical parity (e.g. between positive outcomes, or errors) for members of different groups.

Confusion Matrix



Notations

Random Variables

- $X \in R^d$: Feature vector (Criminal Features like gender, age, race etc.)
- $Y \in 0, 1$: True labels (Recidivate or not in 2 years) - note we focus on binary classsificatiomn setting
- $\hat{Y} \in 0, 1$: Classifier's predicted labels (Recidivism prediction by Classifier)
- $A \in 0, 1$: Sensitive Attribute (Race = black,white)
- R : Classifier prediction scores
- (X, Y) : Data Distribution
- (X, Y, R, A, $\hat{Y}$) : Joint Distribution
- $B \perp C$: Two independent random variables

Predictive Rate Parity

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$$

This ensures that the predicted positive outcomes are equally likely to be true positives across groups, hence achieving parity in predictive rates.

Predictive Equality

$$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$$

This focuses on equalizing the rate of false positives (predicting positive when the true label is negative) across groups, hence the term predictive equality.

Fairness Definitions

Equal Opportunity

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b)$$

This ensures that true positive rates (opportunities for correct predictions) are equal across groups, leading to the name equal opportunity.

Equalized Odds

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b)$$

$$P(\hat{Y} = 1 | Y = 0, A = a) = P(\hat{Y} = 1 | Y = 0, A = b)$$

This achieves equality in both true positive rates and false positive rates, balancing the odds across groups.

Example 1: Overview

Scenario Description

- **Sensitive Attribute:** Gender (male, female). Female is the minority.
- **Outcome:** Selection for a job (after interview).
- **Goal:** Achieve Predictive Rate Parity while violating Equal Opportunity.

Key Fairness Definitions

Predictive Rate Parity: $P(Y = 1|\hat{Y} = 1, \text{male}) = P(Y = 1|\hat{Y} = 1, \text{female})1mm]\text{Equ}$

Example 1: Data for Male Group

Male Group Information

- **Total individuals:** 200
- **Actual positive outcomes (Y=1):** 50
- **Predicted positive outcomes (=1):** 50
- **True positives (correct predictions):** 40
- **False positives:** 10

Calculations (Male)

$$\text{PPV (Predictive Rate)} = P(Y = 1 | \hat{Y} = 1, \text{male}) = \frac{40}{50} = 0.81mm]\text{TPR (Equal Opp}$$

Example 1: Data for Female Group

Female Group Information

- **Total individuals:** 100
- **Actual positive outcomes (Y=1):** 40
- **Predicted positive outcomes (=1):** 25
- **True positives (correct predictions):** 20
- **False positives:** 5

Calculations (Female)

$$\text{PPV (Predictive Rate)} = P(Y = 1 | \hat{Y} = 1, \text{female}) = \frac{20}{25} = 0.81mm]\text{TPR (Equal O}$$

Example 1: Conclusion

Fairness Metric Comparison

- **Predictive Rate Parity (PPV):**

$$0.8 \text{ (male)} = 0.8 \text{ (female)}$$

- **Equal Opportunity (TPR):**

$$0.8 \text{ (male)} \neq 0.5 \text{ (female)}$$

Interpretation

Although the selected candidates (predicted positives) are equally likely to be truly qualified in both groups (PRP holds), the proportion of truly qualified individuals who are selected (TPR) is higher for males than for females. Thus, Equal Opportunity is violated.

Example 2: Overview

Scenario Description

- **Sensitive Attribute:** Gender (male, female). Female is the minority.
- **Outcome:** Selection for a job (after interview).
- **Goal:** Achieve Equal Opportunity while violating Predictive Rate Parity.

Key Fairness Definitions

Equal Opportunity: $P(\hat{Y} = 1 | Y = 1, \text{male}) = P(\hat{Y} = 1 | Y = 1, \text{female}) 1mm]\text{Predict}$

Example 2: Data for Male Group

Male Group Information

- **Total individuals:** 200
- **Actual positive outcomes (Y=1):** 50
- **Predicted positive outcomes (=1):** 80
- **True positives (correct predictions):** 40
- **False positives:** 40

Calculations (Male)

$$\text{TPR (Equal Opportunity)} = P(\hat{Y} = 1 | Y = 1, \text{male}) = \frac{40}{50} = 0.81mm]\text{PPV (Predicti}$$

Example 2: Data for Female Group

Female Group Information

- **Total individuals:** 100
- **Actual positive outcomes (Y=1):** 40
- **Predicted positive outcomes (=1):** 40
- **True positives (correct predictions):** 32
- **False positives:** 8

Calculations (Female)

$$\text{TPR (Equal Opportunity)} = P(\hat{Y} = 1 | Y = 1, \text{female}) = \frac{32}{40} = 0.81mm]\text{PPV (Predi}$$

Example 2: Conclusion

Fairness Metric Comparison

- **Equal Opportunity (TPR):**

$$0.8 \text{ (male)} = 0.8 \text{ (female)}$$

- **Predictive Rate Parity (PPV):**

$$0.5 \text{ (male)} \neq 0.8 \text{ (female)}$$

Interpretation

Although the proportion of truly qualified individuals selected is equal in both groups (TPR holds, hence Equal Opportunity is satisfied), the reliability of the predicted positives (PPV) differs significantly between males and females. Thus, Predictive Rate Parity is violated.

Fairness Definitions

Statistical Parity / Demographic Parity / Disparate Impact

- $P(\hat{Y}|A=a) = P(\hat{Y}|A=b)$ This ensures that the predicted outcomes are independent of the group membership, aiming for equal representation in predictions across groups, which justifies the terms statistical parity, demographic parity, or disparate impact.

Equalized Odds / Disparate Mistreatment (TPRs, FPRs are same)

- P($\hat{Y}$ = 1 | Y=1, A=a) = P($\hat{Y}$ = 1 | Y=1, A=b)
- P($\hat{Y}$ = 1 | Y=0, A=a) = P($\hat{Y}$ = 1 | Y=0, A=b)

Fairness Definitions

Predictive Equality: Equal FPR (False positive rate)

- $P(\hat{Y} = 1 | Y = 0, A = a) = P(\hat{Y} = 1 | Y = 0, A = b) \forall a, b \in A$

Equal Opportunity: Equal TPR (True positive rate)

- $P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = b) \forall a, b \in A$

Conditional Use Accuracy Equality (PPV / PRP and NPV are same)

- $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = a) \forall a, b \in A$
- $P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = a) \forall a, b \in A$

Overall Accuracy Equality

- $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \forall a, b \in A$

Fairness Definitions

Calibration

- $P(\hat{Y} = 1 | R = r, A = a) = P(\hat{Y} = 1 | R = r, A = b) \forall a, b \in A$

Well-Calibration

- $P(\hat{Y} = 1 | R = r, A = a) = P(\hat{Y} = 1 | R = r, A = b) = r \forall a, b \in A$

Balance for Positive Class

- $E(R | Y = 1, A = a) = E(R | Y = 1, A = b) \forall a, b \in A$

Balance for Negative Class

- $E(R | Y = 0, A = a) = E(R | Y = 0, A = b) \forall a, b \in A$

Statistical Definitions

Independence: Classifier scores R are independent of sensitive attribute A e.g. demographic parity, conditional demographic parity.

$$R \perp A$$

Separation: Classifier scores R are independent of the sensitive attribute given the labels e.g. equalized odds, equal opportunity.

$$R \perp A|Y$$

Sufficiency: Labels are independent of sensitive attribute given the classifier scores e.g. Predictive Rate Parity

$$Y \perp A|R$$

$$Y \perp A \tag{1}$$

No issues at all, as concept/study of fairness is required only when Y is dependent on A.

Statistical Definitions

Though there have been multiple metrics of fairness proposed, the three commonly used ones are presented below. (This does not imply that all fairness metrics can be grouped into these three categories; rather, these are among the most widely adopted metrics in fairness research.)

1. **Independence**: this ensures that the predictions are independent of the sensitive attribute, i.e. prediction probabilities are equal across all values of the sensitive attribute (e.g. Demographic Parity, Statistical Parity, etc.).

$$R \perp A \tag{2}$$

2. **Separation**

3. **Sufficiency**

Though there have been multiple metrics of fairness proposed, the three commonly used ones are presented below.

1. **Independence**

2. **Separation**: this ensures that the accuracy of the model are independent of the sensitive attribute given the target label groups(e.g. Equalized Odds, Equal Opportunity etc.).

$$R \perp A \mid Y \tag{3}$$

3. **Sufficiency**

Statistical Definitions

Though there have been multiple metrics of fairness proposed, the three commonly used ones are presented below.

1. **Independence**

2. **Separation**

3. **Sufficiency**: this ensures that given the classifier scores, the labels are independent of sensitive attribute. In other words, given a score the probability of the true variable being 1 should be the same for each group(e.g. Predictive Rate Parity.).

$$Y \perp A \mid R \tag{4}$$

Impossibility Theorem

Theorem: The Impossibility Theorem states that no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias. In simple words, it states that any 2 of the 3 definition of fairness are mutually exclusive.

Proof of the Theorem

Lemma

Statistical Parity and Predictive Parity are mutually exclusive unless $A \perp Y$

Proof.

$$Given \;\; R \perp A \mid Y \Rightarrow \widehat{Y} \perp A \quad using(2) \tag{5}$$

$$Y \perp A \mid R \Rightarrow Y \perp A \mid \widehat{Y} \Rightarrow P(A \mid \widehat{Y}, Y) = P(A \mid \widehat{Y}) \tag{6}$$

$$P(A \mid \widehat{Y}) = P(A) \quad using(5)$$

$$To\,Prove: \;\; A \mid Y \Rightarrow P(A \mid Y) = P(A)$$

$$Solution: \;\; P(A \mid Y) = \sum_b P(\widehat{Y} = b) P(A \mid Y, \widehat{Y} = b)$$

$$P(A \mid Y) = \sum P(\widehat{Y} = b) P(A \mid \widehat{Y} = b) \quad using(8)$$

Proof of the Theorem

Lemma

Equalised Odds and Predictive Parity are mutually exclusive unless $A \perp Y$

Proof.

$$\textit{Given } R \perp A \mid Y \Rightarrow \widehat{Y} \perp A \mid Y \Rightarrow P(A \mid \widehat{Y}, Y) = P(A \mid \widehat{Y}) \tag{7}$$

$$Y \perp A \mid R \Rightarrow Y \perp A \mid \widehat{Y} \Rightarrow P(A \mid Y, \widehat{Y}) = P(A \mid Y) \tag{8}$$

$$\textit{To Prove}: \; Y \perp A$$

$$\textit{Solution}:$$

$$P(A \mid Y = 1) = P(A \mid \widehat{Y} = 1, Y = 1)P(\widehat{Y} = 1) + P(A \mid \widehat{Y} = 0, Y = 1)P(\widehat{Y} = 0)$$

$$P(A \mid Y = 1) = P(A \mid \widehat{Y} = 1)P(\widehat{Y} = 1) + P(A \mid \widehat{Y} = 0)P(\widehat{Y} = 0)$$

$$= P(A)$$

$$\Rightarrow Y \perp A$$

Proof of the Theorem

Lemma

Statistical Parity and Equalised Odds are mutually exclusive unless $A \perp Y$ ***or*** $Y \perp \widehat{Y}$

Proof.

$$Given \;\; R \perp A(\widehat{Y} \perp A) \Rightarrow P(A \mid \widehat{Y}) = P(A) \tag{9}$$

$$A \perp R \mid Y(A \perp \widehat{Y} \mid Y) \Rightarrow P(A \mid \widehat{Y}, Y) = P(A \mid Y) \tag{10}$$

$$Solution : P(\widehat{Y} = b) = P(\widehat{Y} = b \mid A = a)$$

$$P(\widehat{Y} = b) = \sum_{y} P(\widehat{Y} = b \mid A = a, Y = y) P(Y = y \mid A = a)$$

$$P(\widehat{Y} = b) = \sum P(\widehat{Y} = b \mid Y = y) P(Y = y \mid A = a)$$

$$P(\widehat{Y} = b) = \sum_{y} P(\widehat{Y} = b \mid Y = y) P(Y = y)$$

Proof of the Theorem

Lemma

Statistical Parity and Equalised Odds are mutually exclusive unless $A \perp Y$ ***or*** $Y \perp \widehat{Y}$

Proof.

$\Rightarrow$ Now let us define following f^n on y

$$p = P(Y = y)$$

$$p_a = P(Y = y \mid A = a)$$

$$b_0 = P(\widehat{Y} = b \mid y = 0)$$

$$b_1 = P(\widehat{Y} = b \mid y = 1)$$

$$pb_0 + (1-p)b_1 = p_a b_0 + (1-p_a)b_1$$

$$b_0(p - p_a) - b_1(p - p_a) = 0$$

$$(b_0 - b_1)(p - p_a) = 0$$

Metrics Considered While Building ML System

- **Accuracy:** the measure of how often the model is correct in its predictions.
- **Fairness:** the measure of how well the model treats different subgroups of the population, based on characteristics such as race, gender, or age.
- **Explainability/Simplicity:** the measure of how well the model can be understood and interpreted by humans, including how it arrived at its predictions.
- **Privacy:** the measure of how well the model protects sensitive information and maintains the privacy of individuals whose data is used to train or test the model.
- **Security:** the measure of how well the model is protected from attacks, such as adversarial examples or poisoning attacks.

Fairness Vs Accuracy

- An implication from the above impossibility theorem: If the sensitive attribute is predictive of the outcome (i.e., there is a correlation between the two), then it is impossible to satisfy both demographic parity and perfect accuracy.
- If the sensitive attribute is predictive of the outcome, then removing this information from the model may result in a loss of accuracy.

Fairness Vs Accuracy

To balance fairness and accuracy:

- one could use a model that achieves a certain level of accuracy while also satisfying some notion of fairness, such as demographic parity.
- one could use a model that achieves a high level of accuracy but adjusts the model's predictions based on the sensitive attribute to ensure that the model's predictions are not biased.

Fairness vs Accuracy Pareto Optimal Front:

- Let F be a fairness metric, such as statistical parity or equalized odds, and let E be an error metric, such as misclassification rate or cross-entropy loss. Then, the Pareto optimal front is defined as follows:

$$\mathcal{P} = (E, F) \in \mathbb{R}^2 : \nexists (E', F') \in \mathbb{R}^2 \text{ such that } E' < E \text{ and } F' \geq F$$

- A classifier is on the Pareto optimal front if there is no other classifier that achieves lower error and higher fairness, or equivalently, higher accuracy and lower fairness.

Wei, Susan, and Marc Niethammer. "The fairness-accuracy Pareto front." Statistical Analysis and Data Mining: The ASA Data Science Journal 15.3 (2022): 287-302.

Fairness Vs Accuracy

- Consider a binary classifier that assigns a label $\hat{y}$ to an input x based on some function $f(x)$:

$$\hat{y} = \begin{cases} 1, & \text{if } f(x) \geq t \\ 0, & \text{otherwise} \end{cases}$$

  where t is a threshold. Let $p(y = 1|a)$ denote the probability of observing a positive label given the protected attribute a.

- We say that the classifier satisfies minimum fairness if it achieves statistical parity, which is defined as follows:

$$p(\hat{y} = 1|a = 1) = p(\hat{y} = 1|a = 0)$$

Fairness Vs Accuracy

- Suppose there are n individuals, and let x_i denote the outcome for individual i. Then a solution x^* is Pareto optimal if there is no other feasible solution x' such that:
  - $x'_i \geq x_i$ for all i
  - There exists at least one j such that $x'_j > x_j$
- A solution is Pareto optimal if there is no other feasible solution that improves the outcome for at least one individual without making the outcome worse for any other individual.

Fairness Vs Accuracy

- The classifier satisfies the highest fairness if it achieves equalized odds, which is defined as follows: $p(\hat{y} = 1|y = 1, a = 1) = p(\hat{y} = 1|y = 1, a = 0)$ $p(\hat{y} = 1|y = 0, a = 1) = p(\hat{y} = 1|y = 0, a = 0)$
- Satisfying the highest fairness may require sacrificing accuracy. A perfect classifier that achieves maximum accuracy may not necessarily ensure fairness.

Simplicity - Notations

J. Kleinberg, S. Mullainathan, Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. ACM Conference on Economics and Computation, 2019

- A (productivity) score function **f** over the individuals is called **simple** if it is independent of certain features of individuals.
  - Simple means explainability
- Applicants are described by (Boolean) variables x.
- Set-Up: Assume that each data point is a k-dimensional (X) with (k+1)st dimension as A (sensitive attribute). All the dimensions are assumed to be binary (0/1).

$$x^{<1>}, x^{<2>}, x^{<3>}, \ldots, x^{<k>}$$

  - Assumption: Features are independent of each other.
- Function f describes productivity f(x) of applicant with features x.
- Plan: Sort by f-value high to low and admit the top r fraction

Simplicity - Notations

- Function f is independent of group f(x, A) = f(x, D) = f(x)
- Disadvantage: Group D has a worst distribution of feature vectors i.e they have more probability mass on feature vectors producing the lower values
- $\mu(x, \gamma)$ = fraction of population with features x in group γ.
- Function f is a conjunction of all features ('AND') here.

Case 1

- Applicants from A have $x^{<i>}$ = 1 with probability 2/3
- Applicants from D have $x^{<i>}$ = 1 with probability 1/3

$x^{(1)}$	$x^{(2)}$	γ	f	μ
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

Table 1.

Figure: No simplification / true values

Terms and Calculation

For the row-1 having $x^{<1>}$ = 1 , $x^{<2>}$ = 1 and belonging to the 'D' group

$$\mu\text{(fraction of people)} = \frac{1}{3} * \frac{1}{3} * \frac{1}{2} = 1/18$$

$$\mu = P(\gamma = D, x^{<1>} = 1) * P(\gamma = D, x^{<2>} = 1) * P(\gamma = D) \tag{11}$$

$$productivity(f) = x^{<1>} x^{<2>} \tag{12}$$

- At all admission rates $r \leq \frac{1}{18} + \frac{4}{18} = \frac{5}{18}$ (first 2 rows of the table), all admitted have f-value 1, with a $\frac{1}{5}$ fraction from group D
- Equity = ratio of Disadvantages to Advantageous in the accepted rate

$$equity = \frac{1}{4} \text{ (D = } \frac{1}{18}\text{, A = } \frac{4}{18}\text{)}$$

- average f = 1(all admitted are 1)

Reasons to simplify

- Collecting all data may be expensive
- Interepretability or cognitive complexity (for large instances)
- Out-of-sample generalization
- Removing a variable that confers some of the disadvantage

Case 2: using only $x^{<1>}$

- Now if we simplify f by using only x(1) the table would be
- First four rows of Table 1 compress to first row of Table 2 and similarly for the next four rows.

$x^{(1)}$	$x^{(2)}$	γ	Average f	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

Table 2.

$$\text{average } f = \frac{\sum_i f_i . \mu_i}{\sum_i \mu_i} \tag{13}$$

Since r is 5/18, the admitted applicants are solely from row-1 average f = 5/9 (earlier this was 1)

For equity, row-1(x(1)=1) consists of (from table 1) D = 1/18 + 2/18 = 3/18 and A = 4/18 + 2/18 = 6/18

Equity is 1:2 (earlier was 1:4). We see a gain in equity but a fall in productivity or f

Case 2: using only $x^{<1>}$

$x^{(1)}$	$x^{(2)}$	γ	Average f	μ
1	any	any	5/9	1/2
0	any	any	0	1/2

Table 2.

However, the value obtained is when we choose randomly.

In many cases bias creeps in so when the x^2 is not shown, the selector may stereotype and choose all the selected candidates from the advantageous class. We show the behavior in the next slide.

Case 2: Bias Selection

- Now if we simplify f by using x(1) n group the table would be

$x^{(1)}$	$x^{(2)}$	γ	Average *f*	μ
1	any	A	2/3	1/3
1	any	D	1/3	1/6
0	any	A	0	1/6
0	any	D	0	1/3

Table 3.

Figure: Considering $x^{<1>}$ and group

As r is 5/18,
The applicants are admitted solely from row-1, hence only A is chosen
Average f = 2/3 [(1 * 4/18 + 0 * 2/18)/6/18],
Equity (D/A) is 0/1 We see that both f and equity are less than that of true values.

Case 3: Reservation

- Now if we simplify f by using x(1) n group the table would be

$x^{(1)}$	$x^{(2)}$	γ	Average *f*	μ
1	any	A	2/3	1/3
1	any	D	1/3	1/6
0	any	A	0	1/6
0	any	D	0	1/3

Table 3.

Figure: Considering $x^{<1>}$ + reservation

- If Equity is 1:4 (D:A), f = (2/3 * 4 + 1/3 * 1)/(4 + 1) = 3/5
- If Equity is 2:3 (D:A), f = (2/3 * 3 + 1/3 *2)/ (2 + 3) = 8/15

Case 4: Treating (1,1,D) separately from case-1

- As we find in the previous cases, the (1,1,D) row of D has a major disadvantage
- So we simplify the table as below

$x^{(1)}$	$x^{(2)}$	γ	Average f	μ
1	1	D	1	1/18
1	any	any	1/2	4/9
0	any	any	0	1/2

Table 4.

- **As r is 5/18, we chose 1/18 of row-1 and 4/18 of row-2**
- If the 4/18 is chosen uniformly at random from row-2

$$equity = \frac{2}{3}, \text{ Average } f = \frac{3}{5}$$

The average f is better than the reservation case with same equity.

Case 4: case of complete bias

- **As r is 5/18, we chose 1/18 of row-1 and 4/18 of row-2**
- If the 4/18 is chosen completely biased from row-2

$$equity = \frac{1}{4}, \text{ Average } f = \frac{11}{15}$$

f = (1 * 1/18 + 4/18 * 2/3)/(5/18) = 11/15

- The average f is better than Case 2 (Bias Selection), and also hugely better equity.
- Similarly, the f is better than reservation (Case 3) for same equity.

Case 4: case of 80% bias for A

- **As r is 5/18, we chose 1/18 of row-1 and 4/18 of row-2**
- If the 4/18 is chosen with a 80% biased from row-2
- let A1:D1 and f1 represent the fraction and avg. f-score of the row-2 respectively.
- A1 = 4/18 + 2/18 = 6/18 - ratio of 0.8
- D1 = 2/18 ratio of 0.2
- A1/D1 = $\frac{6/18*0.8}{2/18*0.2}$ = 12:1
- f1 = (12 * 2/3 + 1 * 0)/13
- f = (1*1/18 + f1*4/18)/(1/18 + 4/18)
- A:D = (12 * 4/18)/(13 * 1/18 + 1 * 4/18)
  This has a better f compared to reservation (Case 3) with an equity of 1:4. Also, for an equity of 2:3, f would be better in this case than Reservation
  Please find f when equity = 2:3

Result

For every Boolean function f with real-valued outputs satisfying the disadvantage condition and a genericity assumption, and for every simplification g of it (partitioning feature vectors into cells by fixing variables):

- There is always an f-approximator that strictly improves g in both efficiency and equity
- If g does not use group membership then adding group membership as a variable increases efficiency and reduces equity

Advantages: Easy to calculate, Can work with less data points
Disadvantages: Working with less data points may lead to stereotype answers, Productivity and equity may decrease.

Fair Classification Algorithms

A model can be made more fair, by attacking it at the following stages:

1. **Preprocessing**: We modify the data representation and make it bias free and apply a standard classifier like SVM, Logistic Regression, etc. on the new representation.

2. **Inprocessing**: We add fairness constraints during the training process of the classification algorithm to learn a less biased model. We work the constraint into the optimization process that constructs a classifier from training data.

3. **Postprocessing**: The output of the machine learning model is modified in order to reduce bias in the output.

Preprocessing

Certifying and removing disparate impact

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

July 2015

Fair Classification - Preprocessing

- We discuss **Feldman et al (2014)**, one of the early preprocessing algorithms in fair machine learning.
- The paper formalizes the connection between fairness and the predictability of protected class.
- The algorithm modifies each attribute so that the marginal distributions based on the subsets of that attribute with a given sensitive value are all equal. It does not modify the training labels.
- Aim is to ensure that the marginal distribution of an attribute is the same across the sensitive attribute.

Disparate Impact ("80% rule"): Given data set $D = (X, Y, C)$, with

- Protected Attribute X (e.g., race, sex, religion, etc.). e.g. gender $X \in \{M, F\}$
- Remaining attributes Y
- Binary class to be predicted C (e.g., "will hire").$C = 0$ for positive class and $C = 1$ for the negative class.

we will say that D has disparate impact if

$$\frac{Pr(C = YES|X = 0)}{Pr(C = YES|X = 1)} \leq \tau = 0.8 \tag{14}$$

That is protected class is positively classified less than τ times as often as the unprotected class. (legally, τ = 0.8 is common).

Preprocessing - Feldman et al

Can we verify that a classifier on Y will not have disparate impact with respect to X?

Big Idea: A classifier learned from Y will not have disparate impact if X cannot be predicted from Y

Computational Fairness: Alice uses algorithm A to decide who to hire. A takes data set D with protected attribute X and unprotected attributes Y and makes a binary decision C. It is Bob's job to verify that on the data D, Alice's algorithm A is not liable for a claim of disparate impact. Then the idea is that if Bob cannot predict X given Y, A is fair on D.

The **disparate impact certification problem** is to guarantee that, given D, any classification algorithm aiming to predict some C' (which is potentially different from the given C) from Y would not have disparate impact.

The **disparate impact removal problem** is to take some data set D and return a data set $\bar{D} = (X, \bar{Y}, C)$ that can be certified as not having disparate impact. The goal is to change only the remaining attributes Y, leaving C as in the original data set so that the ability to classify can be preserved as much as possible.

Following Table describes the confusion matrix for a classification with respect to the above attributes where each entry is the probability of that particular pair of outcomes for data sampled from the input distribution

Outcome	$X = 0$	$X = 1$
C = NO	a	b
C = YES	c	d

Table: A confusion matrix

Likelihood Ratio (Positive):

$$LR_+(C,X) = \frac{sensitivity}{1 - specificity} = \frac{d/(b+d)}{c/(a+c)} \tag{15}$$

Disparate Impact: A data set has disparate impact if

$$LR_+(C,X) > \frac{1}{\tau} = 1.25 \tag{16}$$

Balanced Error Rate (BER):

$$BER(f(Y),X) = \frac{Pr[f(Y)=0|X=1] + Pr[f(Y)=1|X=0]}{2} \tag{17}$$

Predictability: X is ε-predictable from Y if there exists a function f : Y $\rightarrow$ X such that $BER(f(Y),X) \leq \varepsilon$

Theorem: A data set is $(\frac{1}{2} - \frac{\beta}{8})$-predictable iff it admits disparate impact, where β is the fraction of elements in the minority class $(X = 0)$ that are selected $(C = 1)$.

Proof of Theorem:

- Disparate Impact $\rightarrow$ Predictability
- Predictability $\rightarrow$ Disparate Impact

Proof: Disparate Impact → Predictability

- Suppose there exists some function $g : Y \to C$ such that: $LR_+(g(Y), X) \geq \frac{1}{\tau}$
- Consider the confusion matrix associated with g :

Outcome	$X = 0$	$X = 1$
g(y) = NO	a	b
g(y) = YES	c	d

Table: Confusion matrix for g

- Set $\alpha = \frac{b}{b+d}$ and $\beta = \frac{c}{a+c}$
- $LR_+(g(Y), X) = \frac{1-\alpha}{\beta}$
- $DI(g) = \frac{\beta}{1-\alpha}$

Proof: Disparate Impact → Predictability

- We define the purely biased mapping $\psi : C \to X$ as $\psi(YES) = 1$ and $\psi(NO) = 0$
- Let $\phi : Y \to X = \psi o g$
- Consider the confusion matrix associated with ϕ :

Outcome	$X = 0$	$X = 1$
$\phi(y) = NO$	a	b
$\phi(y) = YES$	c	d

Table: Confusion matrix for ϕ

- Confusion matrix for ϕ is identical to the matrix for g and $BER(\phi) = \frac{\alpha+\beta}{2}$

Proof: Disparate Impact → Predictability

- We can now express contours of the DI and BER functions as curves in the unit square $[0,1]^2$
- Reparameterizing $\pi_1 = 1 - \alpha$ and $\pi_0 = \beta$, we can express the error measures:
  - $DI(g) = \frac{\pi_0}{\pi_1}$: Any classifier g with $DI(g) = \delta$ can be represented in the $[0,1]^2$ as the line $\pi_1 = \frac{\pi_0}{\delta}$
  - $BER(\phi) = \frac{(1+\pi_0-\pi_1)}{2}$: Any classifier ϕ with $BER(\phi) = \varepsilon$ can be written as the $\pi_1 = \pi_0 + 1 - 2\varepsilon$

Proof: Disparate Impact → Predictability

- Let us now fix the desired DI threshold τ, corresponding to the line $\pi_1 = \frac{\pi_0}{\tau}$
- Notice: Region $\{(\pi_0, \pi_1) | \pi_1 \geq \frac{\pi_0}{\tau}\}$ is the region where one would make a finding of disparate impact $(\tau = 0.8)$
- Consider the point $(\beta, \frac{\beta}{\tau})$, at which the line $\pi_0 = \beta$ intersects the DI curve $\pi_1 = \frac{\pi_0}{\tau}$
- This point lies on the BER contour:

$$\begin{aligned} (1 + \beta - \frac{\beta}{\tau})/2 &= \varepsilon \\ \varepsilon &= \frac{1}{2} - \frac{\beta(\frac{1}{\tau} - 1)}{2} \end{aligned} \tag{18}$$

- in particular for the DI threshold of $\tau = 0.8$, the desired BER threshold is: $\varepsilon = \frac{1}{2} - \frac{\beta}{8}$. So disparate impact implies predictability

Proof: Predictability → Disparate Impact

- Suppose there is a function $f : Y \to X$ such that $BER(f(y), x) \leq \varepsilon$
- Let $\psi^{-1} : X \to C$ be the inverse purely biased mapping i.e. $\psi^{-1}(1) = YES$ and $\psi^{-1}(0) = NO$
- Let $g : Y \to C = \psi^{-1} of$
- Recall $BER(\phi) = \frac{(1+\pi_0-\pi_1)}{2}$, this gives us $\pi_1 \geq 1 + \pi_0 - 2\varepsilon$. Therefore:

$$\frac{\pi_0}{\pi_1} \leq \frac{\pi_0}{1+\pi_0-2\varepsilon} = 1 - \frac{1-2\varepsilon}{1+\pi_0-2\varepsilon} \tag{19}$$

- Recall $DI(g) = \frac{\pi_0}{\pi_1}$ and $\pi_0 = \beta$ yields:

$$DI(g) \leq 1 - \frac{1-2\varepsilon}{1+\beta-2\varepsilon} = \tau \tag{20}$$

for $\tau = 0.8$, gives us BER threshold of: $\varepsilon = \frac{1}{2} - \frac{\beta}{8}$

Algorithm:

We run a classifier that optimizes BER on the given data set, attempting to predict the X from Y. Suppose the error in this prediction is ε. Using the estimate of β from the data, we can substitute this into the equation $\varepsilon = 1/2 - \beta/8$ and obtain a threshold ε'. If $\varepsilon' > \varepsilon$ then data set is free from disparate impact.

Certifying disparate Impact - Feldman et al

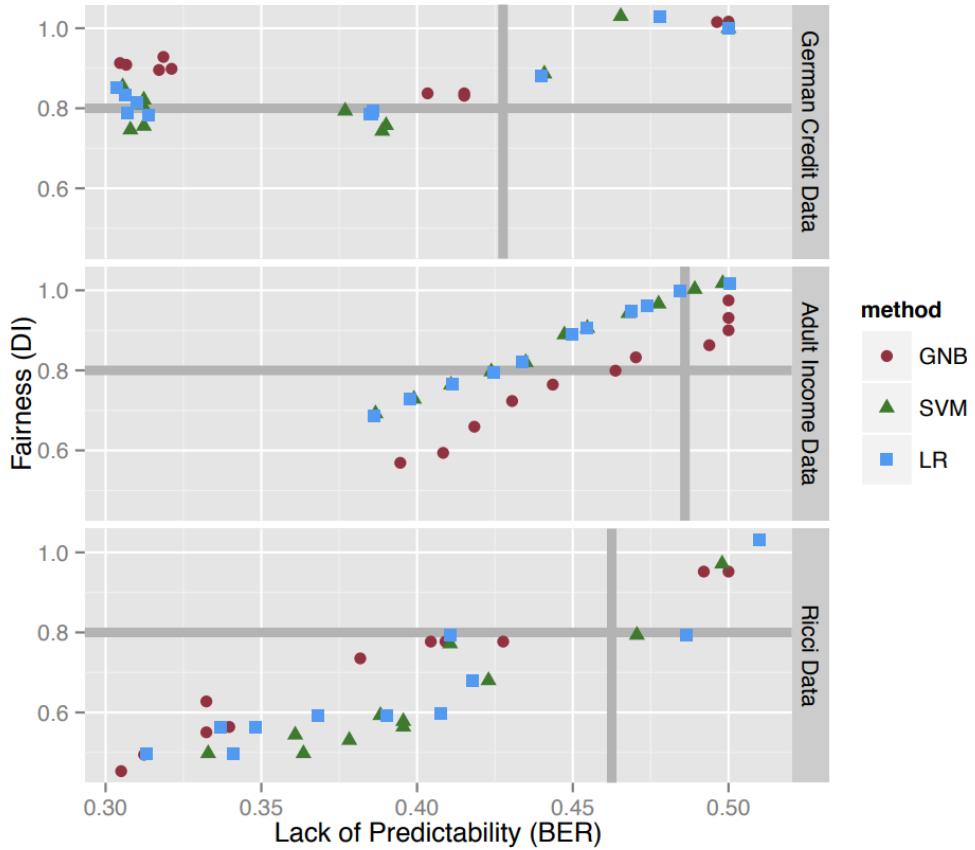


Figure: Lack of predictability (BER) of the protected attributes on the data set as compared to the disparate impact found in the test set when the class is predicted from the non-protected attributes

We want to change D so that it is no longer predictable. How can we do this? Given (X,Y), we want to construct a repaired data set (X,$\overline{Y}$) such that for all g: Y$\rightarrow$X, BER(g(Y),X) $>$ ε, where ε depends on the strength of guarantee we want.

Claim: Perfect repair is always possible.
Proof: Just set Y to 0 for every individual. For example, in a hiring decision, do not hire anyone. That is a fair choice.
Recall that

$$BER(f(Y),X) = \frac{Pr[f(Y) = 0|X = 1] + Pr[f(Y) = 1|X = 0]}{2}$$

Then on the repaired data, the balanced error rate of any classifier is 1/2, which is the maximum possible balanced error rate.

Cumulative Distribution Function (CDF) refers to the Probabilities of X being smaller than or equal to some value x: $F_X(x) = Pr(X \leq x) = p$. This function takes as input x and returns values from the [0,1] denoted as P. **Quantile Function** refers to the The inverse of the cumulative distribution function tells us what x would make $F_X(x)$ return some value p: $F^{-1}(p) = x$

Removing disparate Impact - Feldman et al

Given protected attribute X and a single numerical attribute Y, let Y_x = Pr($Y|X$ = x) denote the marginal distribution on Y conditioned on $X = x$. Let $F_x : Y_x \rightarrow [0,1]$ be the cumulative distribution function for values $y \in Y_x$ and let $F_x^{-1} : [0,1] \rightarrow Y_x$ be the associated quantile function. We will say that F_x ranks the values of Y_x.

Let $\overline{Y}$ be the repaired version of Y in $\overline{D}$. We will say that $\overline{D}$ strongly preserves rank if for any $y \in Y_x$ and $x \in X$, its repaired counterpart $\bar{y} \in \overline{Y}_x$ has $F_x(y) = F_x(\bar{y})$.
We define a "median" distribution A in terms of its quantile function $F_A^{-1} : F_A^{-1}(u) = median_{x \in X} F_x^{-1}(u)$

Removing disparate Impact - Feldman et al

Lemma: Let A be a distribution such that $F_A{}^{-1}(u) = median_{x \in X} F_x{}^{-1}(u)$. Then A is also the distribution minimizing $\sum_{x \in X} d(Y_x, C)$ over all distributions C, where $d(\cdot,\cdot)$ is the earthmover distance on R.

Algorithm: The proposed repair algorithm creates $\overline{Y}$, such that for all $y \in Y_x$, the corresponding $\bar{y} = F_A{}^{-1}(F_x(y))$. The resulting $\overline{D} = (X, \overline{Y}, C)$ changes only Y while the protected attribute and class remain the same as in the original data, thus preserving the ability to predict the class.

Rank preserving repair means that if we have a case where person 1 is selected and person 2 is not selected before repair then in no case will person 1 not be selected if person 2 is selected after repair.

Algorithm:

1. Let p_y^x be the percentage of agents with protected status x whose numerical score is at most y.

2. We take a data point (x_i, y_i) and calculate $p_{y_i}^{x_i}$.

3. We find y_i^{-1} such that $p_{y_i^{-1}}^{1-x_i} = p_{y_i}^{x_i}$.

4. We repair $\bar{y}_i = median(y_i, y_i^{-1})$.

Removing disparate Impact - Feldman et al

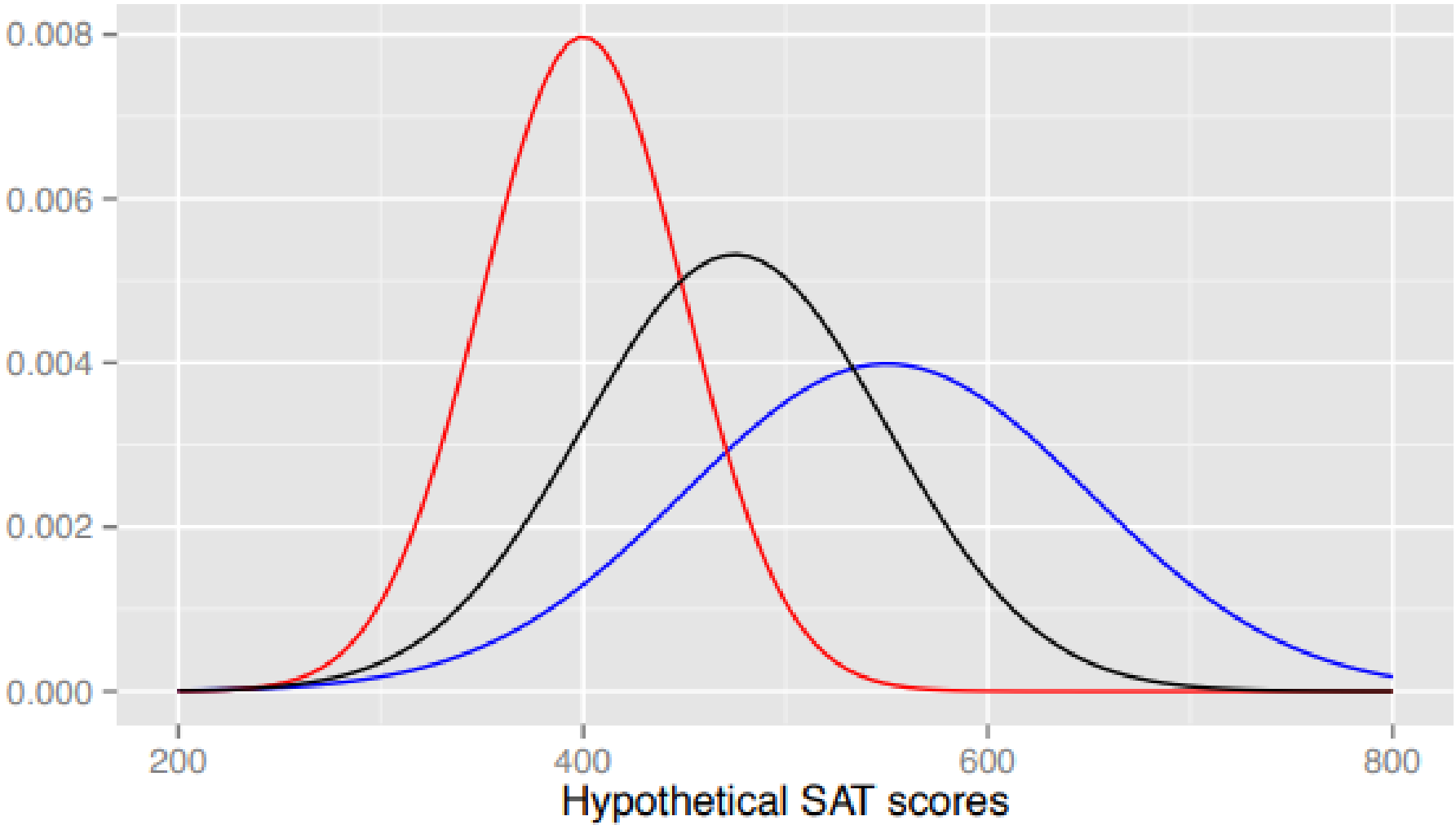


Figure: Hypothetical SAT scores

Full Repair Example

Blue curve: Distribution of SAT scores for X = female, $\mu = 550, \sigma = 100$
Red curve: Distribution of SAT scores for X = male, $\mu = 400, \sigma = 50$

p^F_{700} = 1/2 = $Pr(y \leq 700 | x = F)$ Percentage of females getting $\leq$ 700 marks is 50%.[y1 = 700]
p^F_{750} = 3/4 = $Pr(y \leq 750 | x = F)$ Percentage of females getting $\leq$ 750 marks is 75%. [y2 = 750]
p^M_{500} = 1/2 = $Pr(y \leq 500 | x = M)$ Percentage of males getting $\leq$ 500 marks is 50%. [y_1^{-1} = 500]
p^M_{550} = 3/4 = $Pr(y \leq 550 | x = M)$ Percentage of males getting $\leq$ 550 marks is 75%. [y_2^{-1} = 550]

Removing disparate Impact - Feldman et al

We obtain repair $\bar{y_1} = median(y_1, y_1^{-1})$ = (700+500)/2 = 600
We obtain repair $\bar{y_2} = median(y_2, y_2^{-1})$ = (750+550)/2 = 650

Also note that since $y_1 \leq y_2$ before repair, we will always have $\bar{y_1} \leq \bar{y_2}$ after repair.

Black curve: Fully repaired data is the distribution in black, with $\mu = 475, \sigma = 75$

Male score in 95th percentile: 500$\rightarrow$625
Female score in 95th percentile: 750$\rightarrow$625

$\bar{y_2} = \alpha_1 y_1 + \alpha_2 y_2 = \alpha y_1 + (1-\alpha) y_2$
Here α acts as the knob to control disparate impact in the dataset. If $\alpha > \frac{1}{2}$ then there will be reverse disparate impact. The order will still be preserved but in their own segments.

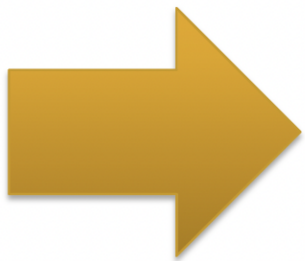
Fairness Constraints: A Flexible Approach for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi

How Machines Learn?

1. By training over historical data.
2. Example task: Predict who will return loan.

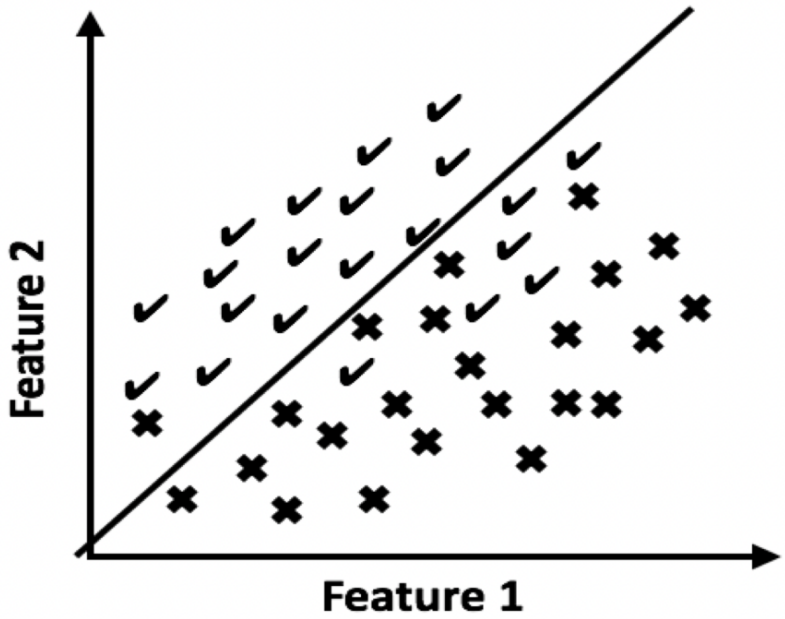
	F_1	F_2	...	F_m	Z
$User_1$	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$	Z_1
$User_2$	$x_{2,1}$			$x_{2,m}$	Z_2
$User_3$	$x_{3,1}$			$x_{3,m}$	Z_3
...	...			...	...
$User_n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$	Z_n



Past Outcomes
Returned
Defaulted
Defaulted
...
Returned

1. **Learning challenge:** Learn a decision boundary (W) in the feature space separating the two classes

Predict who will return loans?



1. Optimal (most accurate / least loss) linear boundary
2. But, how do machines find (compute) it?

Learning (computing) the optimal boundary

1. Define & optimize a loss (accuracy) function
   1. The loss function captures inaccuracy in prediction

$$L(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \qquad L(\mathbf{w}) = \sum_{i=1}^{N} -\log p(y_i | \mathbf{x}_i, \mathbf{w})$$

   2. Minimize (optimize) it over all examples in training data
2. Central challenge in machine learning
   1. Finding loss function that capture prediction loss, yet be efficiently optimized
   2. Many loss functions used in learning are convex

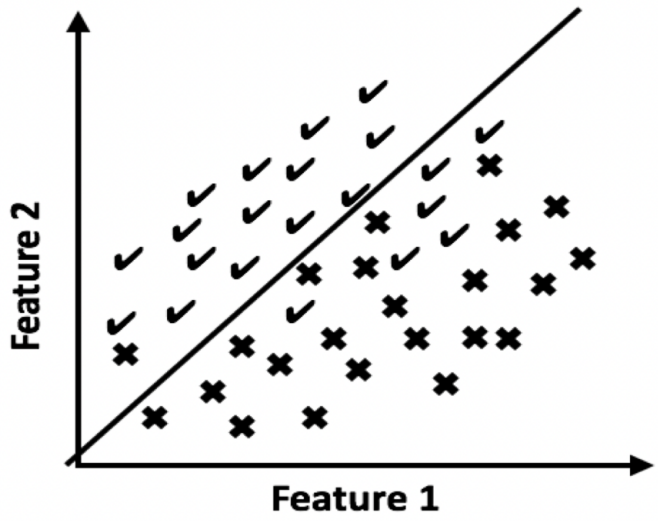
Convex-boundary based loss functions

Squared loss $\sum_{i=1}^{N}(y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$

Logistic loss $-\sum_{i=1}^{N}\log(1 + e^{-y_i d_{\mathbf{w}}(\mathbf{x}_i)})$

SVM loss $||\mathbf{w}||^2 + C\sum_{i=1}^{N}\max(0, 1 - y_i d_{\mathbf{w}}(\mathbf{x}_i))$

Predict who will return loans?



1. Optimal (most accurate / least loss) linear boundary
2. But, how do machines find (compute) it?
3. The boundary was computed using

$$\min \quad \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$

How to learn to avoid discrimination

1. Specify discrimination measures as constraints on learning.
2. Optimize for accuracy under those constraints.

$$\begin{aligned} &\textit{minimize } L(\mathbf{w}) \\ &\textit{subject to } \quad P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1) \end{aligned}$$

3. The constraints embed ethics & values when learning.
4. No free lunch: Additional constraints lower accuracy.
5. Tradeoff between performance & ethics (avoid discrimination)

A few observations

1. Any discrimination measure could be a constraint.

$$
\begin{aligned}
&\textit{minimize } L(\mathbf{w}) \\
&\textit{subject to} \quad P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x}) \\
&\qquad\qquad\quad P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1) \\
&\qquad\qquad\quad P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)
\end{aligned}
$$

2. Might not need all constraints at the same time.
   1. E.g., drop disp. impact constraint when no bias in data.
   2. When avoiding disparate impact/mistreatment, we could achieve higher accuracy without disparate treatment

Key technical challenge

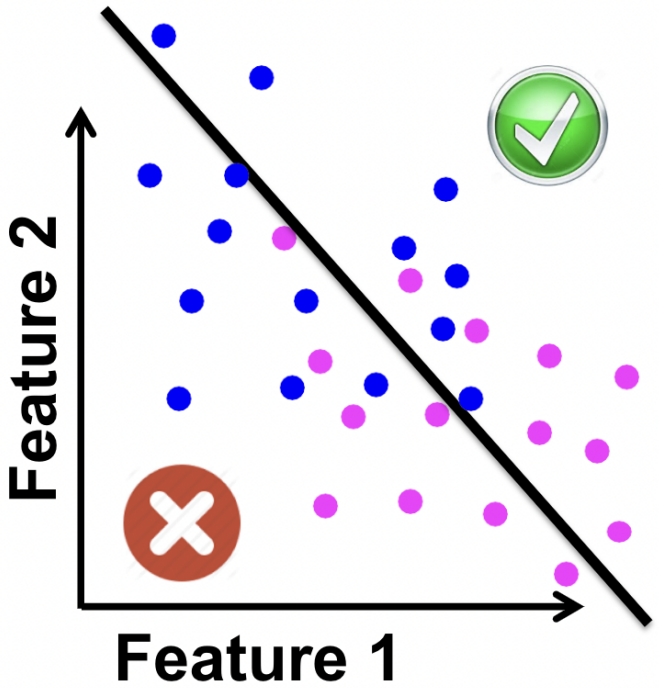
1. How to learn efficiently under these constraints?

$$\begin{aligned} &\textit{minimize } L(\mathbf{w}) \\ &\textit{subject to } \quad P(\hat{y}=1|z=0) = P(\hat{y}=1|z=1) \end{aligned}$$

$$\begin{aligned} &\textit{minimize } L(\mathbf{w}) \\ &\textit{subject to } \quad P(\hat{y}\neq y|z=0) = P(\hat{y}\neq y|z=1) \end{aligned}$$

2. Problem: The above formulations are not convex!
    1. Can't learn them efficiently.
3. Need to find a better way to specify the constraints.
4. So that loss function under constraints remains convex.

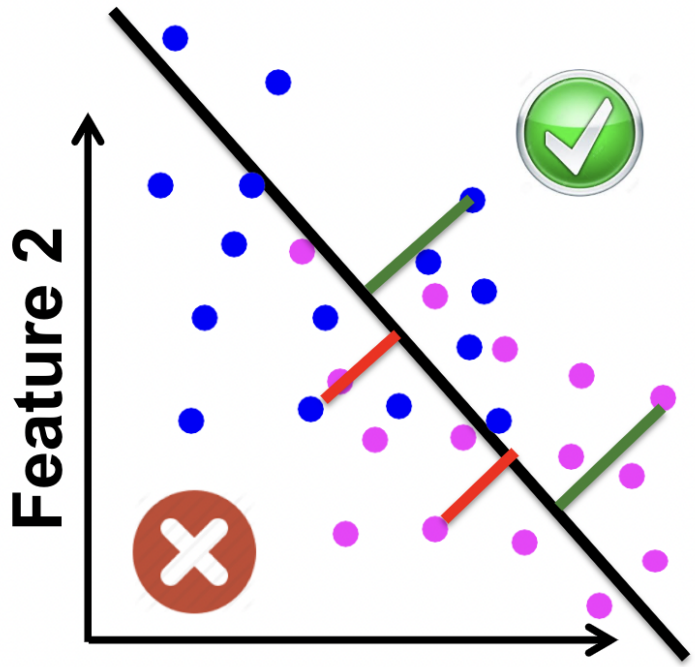
Disparate impact constraints: Intuition



$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

Limit the differences in the acceptance (or rejection) ratios across members of different sensitive groups

Disparate impact constraints: Intuition



$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

Limit the differences in the average strength of acceptance and rejection across members of different sensitive groups.

Specifying disparate impact constraints

1. Instead of requiring:

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

2. Bound covariance between items' sensitive feature values and their signed distance from classifier's decision boundary to less than a threshold

$$\boxed{\left| \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{z}_i - \bar{\mathbf{z}} \right) \mathbf{w}^{\mathbf{T}} \mathbf{x}_i \right| \leq \mathbf{c}}$$

Learning classifiers w/o disparate impact

1. Previous formulation: Non-convex, hard-to-learn

$$\begin{aligned} &\textit{minimize } L(\mathbf{w}) \\ &\textit{subject to } \quad P(\hat{y}=1|z=0) = P(\hat{y}=1|z=1) \end{aligned}$$

2. New formulation: Convex, easy-to-learn

$$\begin{aligned} &\textit{minimize } \quad L(\mathbf{w}) \\ &\textit{subject to } \quad \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\,\mathbf{w}^T\mathbf{x}_i \leq \mathbf{c} \\ &\qquad\qquad\quad \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\,\mathbf{w}^T\mathbf{x}_i \geq -\mathbf{c} \end{aligned}$$

A few observations

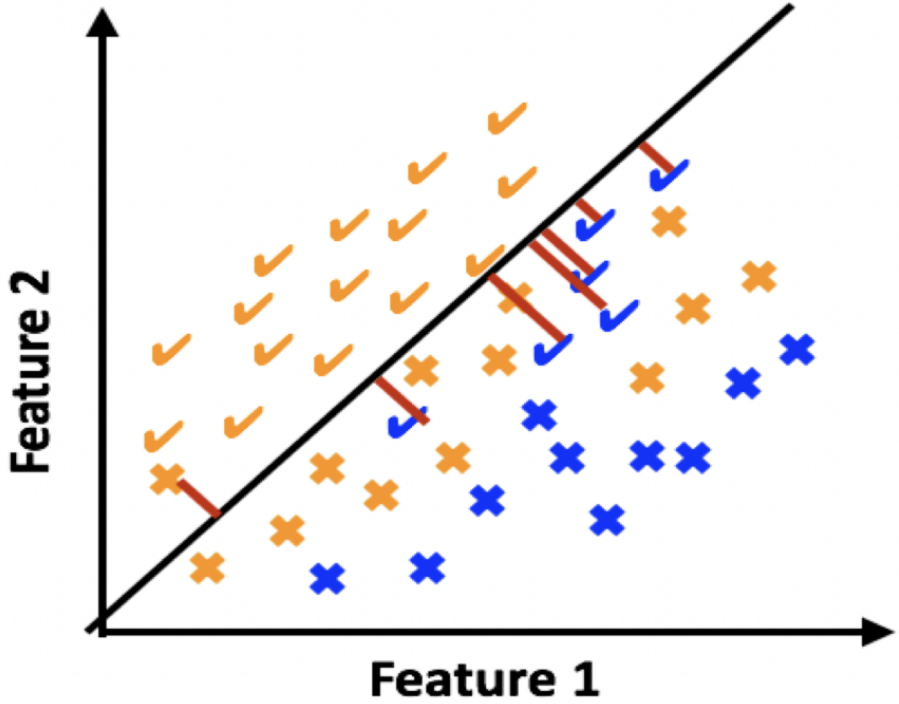
1. Our formulation can be applied to any convex- margin (loss functions) based classifiers
    1. hinge-loss, logistic loss, linear and non-linear SVM
2. Can easily change our formulation to optimize for fairness under accuracy constraints
    1. Useful in practice, when you want to be fair but have business necessity to meet a certain accuracy threshold

Specifying mistreatment constraints

$$\min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i))$$

Concave

($d_w(x)$ is affine)



Idea: Avg. misclassification distance from boundary for both groups should be the same.

Rewriting mistreatment constraints

$$\min \quad \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$

$$\text{s.t.} \quad P(y_{true} \neq y_{pred} \mid ♀) = P(y_{true} \neq y_{pred} \mid ♂)$$

Rewriting mistreatment constraints

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & -\epsilon \leq \frac{1}{|\male|} \sum_{\male} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) - \frac{1}{|\female|} \sum_{\female} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) \leq \epsilon \end{aligned}$$

Concave Concave

$P(y_{true} \neq y_{pred} \mid \male)$ $P(y_{true} \neq y_{pred} \mid \female)$

1. Can be solved efficiently.
2. Using Disciplined Convex-Concave Programming.

1. New formulation: Convex-concave, can learn efficiently using convex-concave programming.

$$
\begin{aligned}
&\textit{minimize} \quad L(\mathbf{w}) \\
&\textit{subject to} \quad \frac{-N_1}{N}\sum_{i=1}^{N_0} g_{\mathbf{w}}(y_i,\mathbf{x}_i) + \frac{N_0}{N}\sum_{i=1}^{N_1} g_{\mathbf{w}}(y_i,\mathbf{x}_i) \leq \mathbf{c} \\
&\qquad\qquad \frac{-N_1}{N}\sum_{i=1}^{N_0} g_{\mathbf{w}}(y_i,\mathbf{x}_i) + \frac{N_0}{N}\sum_{i=1}^{N_1} g_{\mathbf{w}}(y_i,\mathbf{x}_i) \geq -\mathbf{c}
\end{aligned}
$$

All misclassifications

$$g_{\mathbf{w}}(y,\mathbf{x}) = min(0, yd_{\mathbf{w}}(\mathbf{x})),$$

False negatives

$$g_{\mathbf{w}}(y,\mathbf{x}) = min\left(0, \frac{1+y}{2} yd_{\mathbf{w}}(\mathbf{x})\right), \text{ or}$$

False positives

$$g_{\mathbf{w}}(y,\mathbf{x}) = min\left(0, \frac{1-y}{2} yd_{\mathbf{w}}(\mathbf{x})\right),$$

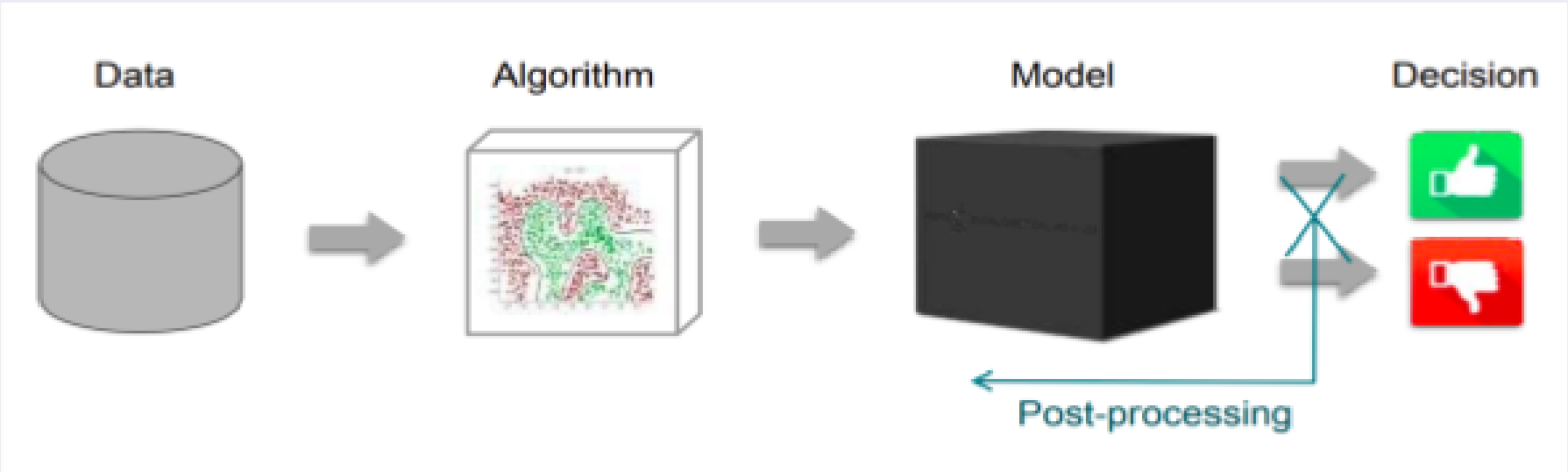
Fair Classification - Postprocessing

Paper being followed

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).

Fair Classification - Postprocessing

- We discuss Hardt et al., a post-processing algorithm that uses the paradigm of score transformation.
- The algorithm adjusts (post-process) the scores produced by an arbitrary classifier to remove discrimination from the data.



Contributions of the Paper

The paper proposes a simple, interpretable, and actionable framework for measuring and removing discrimination based on protected attributes.

Contributions of the Paper

The paper proposes a simple, interpretable, and actionable framework for measuring and removing discrimination based on protected attributes.

- **Contribution 1 -** The paper proposes an easily checkable and interpretable notion of avoiding discrimination based on protected attributes. The paper's notion enjoys a natural interpretation in terms of graphical dependency models.

Contributions of the Paper

The paper proposes a simple, interpretable, and actionable framework for measuring and removing discrimination based on protected attributes.

- **Contribution 1 -** The paper proposes an easily checkable and interpretable notion of avoiding discrimination based on protected attributes. The paper's notion enjoys a natural interpretation in terms of graphical dependency models.
- **Contribution 2 -** The paper gives a simple and effective framework for constructing classifiers satisfying our criterion from an arbitrary learned predictor. Rather than changing a possibly complex training pipeline, the result follows via a simple post-processing step that minimizes the loss in utility.

Objective

We assume a source distribution over (Y, X, A), where Y is the target or true outcome (e.g. “default on loan”), X is the available features, and A is the protected attribute.
The objective of supervised learning is to construct a (possibly randomized) predictor $\hat{Y} = f(X, A)$ that predicts Y as is typically measured through a loss function. Furthermore, we would like to require that $\hat{Y}$ does not discriminate with respect to A, and the goal of this paper is to formalize this notion.
To do so, a post-processing needs to be done on $\hat{Y}$ to get $\tilde{Y}$.

Equalized Odds

- We say that a predictor $\hat{Y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.
  Unlike demographic parity, equalized odds allows $\hat{Y}$ to depend on A but only through the target variable Y. As such, the definition encourages the use of features that allow to directly predict Y but prohibits abusing A as a proxy for Y.

$$Pr\{\hat{Y} = 1|A = 0, Y = y\} = Pr\{\hat{Y} = 1|A = 1, Y = y\};\ y \in \{0, 1\} \quad (21)$$

Equal Opportunity

- We say that a binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if

$$Pr\{\hat{Y} = 1 | A = 0, Y = 1\} = Pr\{\hat{Y} = 1 | A = 1, Y = 1\} \tag{22}$$

Equal opportunity is a weaker, though still interesting, notion of non-discrimination, and thus typically allows for stronger utility as we shall see in our case study.

Oblivious Measures

- A property of a predictor $\hat{Y}$ or score R is said to be oblivious if it only depends on the joint distribution of (Y, A,$\hat{Y}$) or (Y, A, R), respectively.

Derived Predictor

- A predictor $\tilde{Y}$ is derived from a random variable R and the protected attribute A if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\tilde{Y}$ is independent of X conditional on (R, A).

Derived Predictor

The derived predictor $\tilde{Y}$ can be derived from in the following way (we are using $\hat{Y}$ instead of R:

$$Pr\{\tilde{Y}=1|A=a,Y=y\}=<LINEAR\ TRANSFORM>(Pr\{\hat{Y}=1|A=a,Y=y\}) \tag{23}$$

such that, $Pr\{\tilde{Y}=1|A=a,Y=y\}$ CAN POTENTIALLY TAKE ALL VALUES IN $[0,1]$

Note it is assumed that there is access of ground truth (Y) during the training session.

Derived Predictor - Intuition

Intuition

To ensure that the resultant $Pr\{\tilde{Y} = 1|A = a, Y = y\}$ can **take values from** 0 **to** 1, an appropriate **affine combination** (explained in the next few slides) could be a solution for a linear transformation.
NOTE THAT -

- $Pr\{\tilde{Y} = 1|A = a, Y = y\} = 0$ (no one gets the job) is FAIR
- $Pr\{\tilde{Y} = 1|A = a, Y = y\} = 1$ (all get the job) is FAIR
- But, both the above cases are **not efficient**.

Convex Combination Representation

Vertices of the Parallelogram

A general point (x, y) (don't confuse with X, X represents input features) inside the parallelogram can be expressed as a **convex combination** of its four vertices:

$$(0,0), \quad (u,v), \quad (1,1), \quad (1-u, 1-v)$$

Barycentric Coordinates

Any point (x, y) inside the parallelogram can be represented in terms of **barycentric coordinates** or as an **AFFINE COMBINATION** of the vertices:

$$(x,y) = \lambda_1(0,0) + \lambda_2(u,v) + \lambda_3(1,1) + \lambda_4(1-u, 1-v)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weights satisfying:

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$$

Same result if $\lambda_1 = 0$

Expanded Form

Equations for x and y

Expanding the equation:

$$x = \lambda_2 u + \lambda_3 + \lambda_4(1-u)$$

$$x = (\lambda_2 - \lambda_4) \cdot u + (\lambda_3 + \lambda_4)$$

Let $\alpha = \lambda_2, \beta = \lambda_3, \gamma = \lambda_4$ $(\alpha + \beta + \gamma = 1)$

$$x = (\alpha - \gamma) \cdot u + (\beta + \gamma)$$

Similarly,

$$y = (\alpha - \gamma) \cdot v + (\beta + \gamma)$$

Deriving from a Binary Predictor

We will first develop an intuitive geometric solution in the case where we adjust a binary predictor $\hat{Y}$ and A is a binary protected attribute. For convenience, we introduce the notation :

$$y_a(\hat{Y}) \overset{\text{def}}{=} \Big(Pr\{\hat{Y} = y | A = a, Y = 0\}, Pr\{\hat{Y} = y | A = a, Y = 1\} \Big) \tag{24}$$

The first component of $y_a(\hat{Y})$ is the false positive rate of $\hat{Y}$ within the demographic satisfying A = a. Similarly, the second component is the true positive rate of $\hat{Y}$ within A = a.

Lemma 4.2. A predictor $\hat{Y}$ satisfies:

1. Equalized odds if and only if $y_0(\hat{Y}) = y_1(\hat{Y})$, and
2. Equal opportunity if and only if $y_0(\hat{Y})$ and $y_1(\hat{Y})$ agree in the second component, i.e., $y_0(\hat{Y})_2 = y_1(\hat{Y})_2$.

For $a \in \{0,1\}$, consider the two-dimensional convex polytope defined as the convex hull of four vertices:

$$P_a(\hat{Y}) \stackrel{\text{def}}{=} \text{convhull}\Big\{(0,0), y_a(\hat{Y}), y_a(1-\hat{Y}), (1,1)\Big\} \tag{25}$$

Deriving from a Binary Predictor

Lemma 4.3. A predictor $\tilde{Y}$ is derived if and only if for all $a \in \{0,1\}$, we have $y_a(\tilde{Y}) \in P_a(\hat{Y})$.

Proof

Since a derived predictor $\tilde{Y}$ can only depend on $(\hat{Y}, A)$ and these variables are binary, the predictor $\tilde{Y}$ is completely described by four parameters in [0,1] corresponding to the probabilities $Pr\{\tilde{Y} = 1 | \hat{Y} = \hat{y}, A = a\}$ for $\hat{y}, a \in \{0,1\}$. Each of these parameter choices leads to one of the points in $P_a(\hat{Y})$ and every point in the convex hull can be achieved by some parameter setting.

Deriving from a Binary Predictor

Combining Lemma 4.2 with Lemma 4.3, we see that the following optimization problem gives the optimal derived predictor with equalized odds:

$$min_{\tilde{Y}}\ El(\tilde{Y}, Y)(El - ExpectedLoss) - ENSURES\ ACCURACY$$
$$s.t\ \forall a \in \{0,1\} : y_a(\tilde{Y}) \in P_a(\hat{Y}) \quad (derived) - POSTPROC.\ CONSTRAINT$$
$$y_0(\tilde{Y}) = y_1(\tilde{Y}) \quad (equalizedodds) - ENSURES\ FAIRNESS$$

Derived Predictor

The derived predictor $\tilde{Y}$ can NOW be derived from in the following way (we are using $\hat{Y}$ instead of R:

$$Pr\{\tilde{Y}=1|A=a,Y=y\} = (\alpha-\gamma)\cdot Pr\{\hat{Y}=1|A=a,Y=y\}+\beta+\gamma$$

$\alpha,\beta,\gamma\in[0,1]$ & $\alpha+\beta+\gamma=1$(26)
Note it is assumed that there is access of ground truth (Y) during the training session.

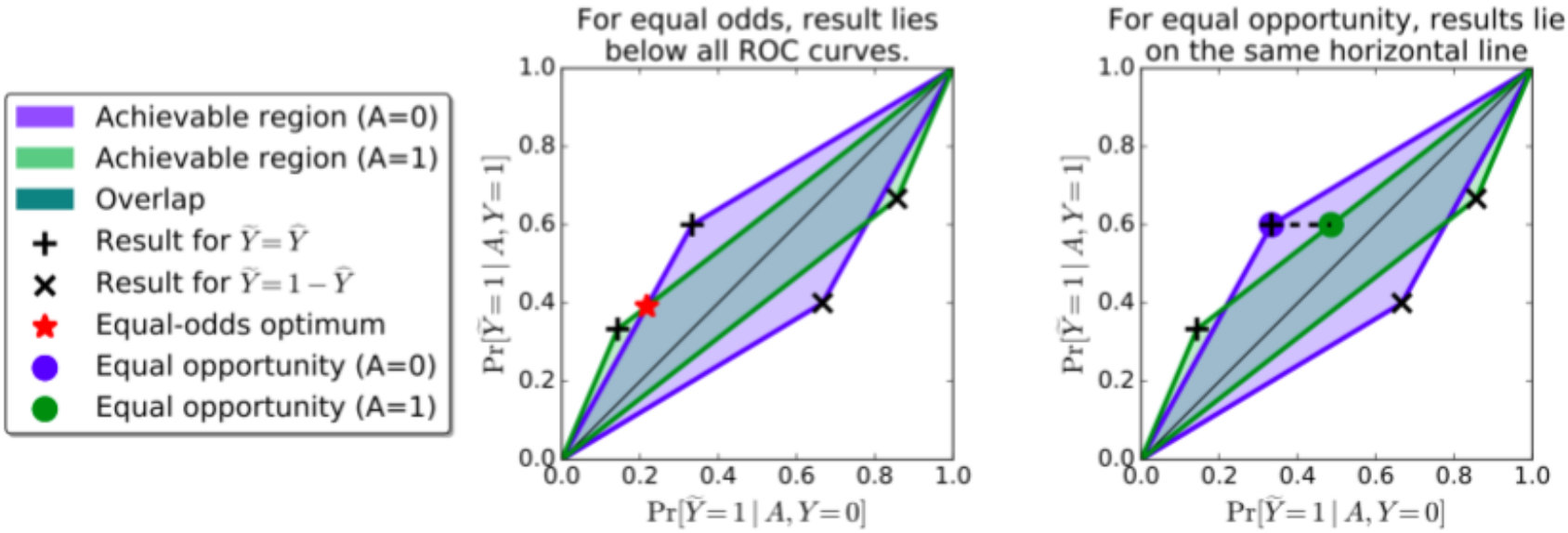


Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

During Inference ....

$$Pr\{\tilde{Y}=1|A=a\} = Pr\{\tilde{Y}=1|A=a,Y=1\} \cdot Pr\{Y=1\} + Pr\{\tilde{Y}=1|A=a,Y=0\} \cdot Pr\{Y=0\} \quad (27)$$

Once we calculate $Pr\{\tilde{Y}=1|A=a\}$, we can sample from this probability distribution to get $\tilde{Y}$.

Ensuring generalized fairness in batch classification

Manjish Pal, Subham Pokhriyal, Sandipan Sikdar and Niloy Ganguly
Scientific Reports volume 13, Article number: 18892 (2023)

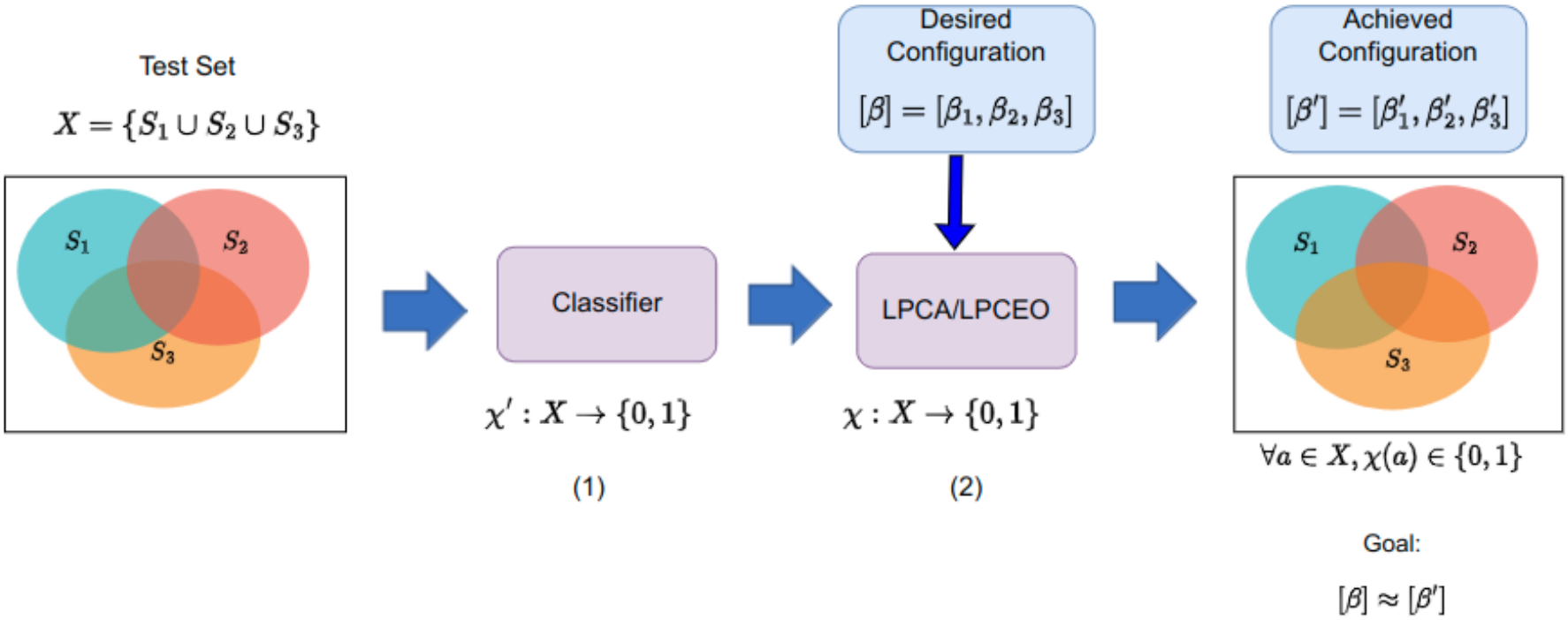
Problem Set-Up

- For a particular dataset X a set of sensitive attributes (like 'race', 'gender' etc.) are given.
- All the subpopulations are modeled as subsets S_j (e.g. 'black') where $S = \{S_1, S_2, \ldots, S_m\}$ is the set of all distinct subpopulations.
- Let the set of sensitive attributes corresponding to these populations be $A_1, A_2 \ldots, A_k$.
- Example $X = \{a,b,c,d,e\}, S_1(\text{male}) = \{a,b\}, S_2(\text{female}) = \{c,d,e\}, S_3(\text{black}) = \{b,d\}, S_4(\text{white}) = \{a,c,e\}$ and $A_1 = \text{gender}, A_2 = \text{race}$.
- **Fair-Classification:** Learn from the training data and ensure low unfairness in test data.

Fair Batch Classification

- Classification is point-wise i.e. elements are to be labelled one at a time once training is done.
- Instead we consider the problem of classification of entire test data (or of batches) as a postprocessing step.
- In this work we propose LP-based fair batch classification and compare with state of the art fair classification, ranking and subset selection algorithms.
- Fair batch classification is more suitable in recruitment and admission-like scenarios.

Proposed Postprocessing Step



Notions of Fairness

- Two most well studied and popular notions of fairness considered namely Demographic Parity and Equalized Odds.
- Our algorithms can handle the important case of multiple overlapping subpopulations for both DP and EO unlike many others in the literature.
- **Demographic Parity:** Ensure same acceptance rate / selection rate across sensitive subpopulations.
- **Equalized Odds:** Ensure same TPR and FPR across sensitive subpopulations.

Notions of Fairness

- **Single Sensitive Attributes** All the sensitive populations (subsets) belong to a single attribute e.g. $S_1 = \text{white}, S_2 = \text{black}, S_3 = \text{hispanic}$. Here $S_i \bigcap S_j = \phi \; \forall i,j \in [m]$.
- **Independent Groups** The sets S_j's can be intersecting e.g. $S_1 = \text{male}, S_2 = \text{black}, S_3 = \text{female}$.
- **Intersectional Groups** Elements of S_j are k-tuples of the form $(a_1, a_2, \ldots a_k) \in A_1 \times A_2 \cdots \times A_k$. In this case $S_i \bigcap S_j = \phi \; \forall i,j \in [m]$.

Overlapping Groups

- **Gerrymandering Groups** Elements of S_j are r-tuples of the form $(a_1, a_2, \ldots a_r) \in A_{i_1} \times A_{i_2} \cdots \times A_{i_r}$ for any $r < k$. In this case, generally, $S_i \bigcap S_j \neq \phi$.
- Simple exercise; show that
  - **Independent Groups** $\subseteq$ **Gerrymandering Groups**.
  - **Intersectional Groups** $\subseteq$ **Gerrymandering Groups**.

Defining Demographic Disparity (DDP_M)

- For multiple overlapping subpopulations, we can define Demographic Disparity (DDP_M) as the difference between the maximum and acceptance rates of a classifier $\hat{Y}$ across all the sensitve populations. Mathematically,
- For the case of a single sensitive attribute A with two subpopulations, $DDP_S = \max_j \mathbb{P}[\hat{Y} = 1|A = 0] - \min_j \mathbb{P}[\hat{Y} = 1|A = 1]$.
- For arbitrary sensitive subpopulations S_i's $DDP_M = \max_j \mathbb{P}[\hat{Y} = 1|S_j] - \min_j \mathbb{P}[\hat{Y} = 1|S_j]$

The Configuration Model

- We define a configuration as $[\beta] = \{\beta_1, \beta_2 \cdots \beta_m\}$ where $\beta_j = \mathbb{P}[\hat{Y} = 1 | S = S_j]$ and $j \in [m]$ is simply the acceptance rate of the subpopulation S_j.
- We also define a configuration based Demographic Disparity, DDP_C, which can be used to compare any two given configurations $[\beta]$ and $[\beta']$.
- $DDP_C([\beta], [\beta']) = l_\infty([\beta], [\beta'])$
- We can think of $[\beta]$ as the natural or **input configuration** (acceptance rates) and $[\beta']$ as the **obtained configuration** (acceptance rates) with low DDP_M.
- Under some conditions we can prove a relationhship between DDP_C and DDP_M.

lemma

Given a configuration [β] in which the acceptance rates across all the subpopulations are the same and equal to β, one can show $DDP_M([\beta'])/2 \leq DDP_C([\beta], [\beta'])$ where $[\beta'] = \{\beta'_1, \beta'_2, \ldots, \beta'_m\}$ is an arbitrary configuration and $\beta \in [\min_j \beta'_j, \max_j \beta'_j]$.

proof

Since, $\beta \in [\min_j \beta'_j, \max_j \beta'_j]$, we can write $DDP_M(\beta') = \max_j \beta'_j - \min_j \beta'_j = (\max_j \beta'_j - \beta) + (\beta - \min_j \beta'_j) \leq 2 \cdot \max_j |\beta'_j - \beta| = 2DDP_C([\beta], [\beta'])$.

Decidability Problem for Demographic Disparity

Problem (Demographic Disparity)

Let X be a universal set with $|X| = n$. Given a collection $S = \{S_1, S_2, \cdots S_m\}$ of subsets of X as the set of all subpopulations across the sensitive attributes and a configuration $[\beta] = (\beta_1, \beta_2, \ldots \beta_m)$, decide whether there exists a $0/1$ labeling of elements of the universal set X such that the configuration $[\beta]$ can be realized for all the m (possibly overlapping) sets.

Data without True Labels

We first consider the case when data is unlabelled and fairness is ensured by attaining low DDP_M using a linear program which we refer to as **LPC**.

LPC- LP-Configuration

$$\min \epsilon$$

$$\left(\sum_{a \in S_j} \chi(a)\right) \geq (\beta_j - \epsilon)|S_j| \; \forall S_j \in S$$

$$\left(\sum_{a \in S_j} \chi(a)\right) \leq (\beta_j + \epsilon)|S_j| \; \forall S_j \in S$$

$$0 \leq \chi(a) \leq 1 \; \forall a \in X$$

Data with True Labels

- Then we consider data has true labels and reduce DDP_C while maintaining accuracy, precision and recall.
- We crucially use the *confidence values* $\mathbb{P}[\hat{Y}_a = 1]$ returned by a classifier (say Random Forest or SVM) for every record a.
- Our approach provided flexibility to realize an arbitrary configuration [β] chosen by the user.
- Our approach can also provide flexibility to control DDP_M for multiple overlapping subpopulations unlike other algorithms.

LPCA

$$\min \sum_{a \in X} \chi(a) w(a)$$

$$\left(\sum_{a \in S_i} \chi(a) \right) \geq \beta_i |S_i| \; \forall S_i \in S$$

$$\left(\sum_{a \in S_i} \chi(a) \right) \leq (\beta_i + \varepsilon) |S_i| \; \forall S_i \in S_j$$

$$\beta_i = (\alpha) \beta_i^{\text{initial}} + (1 - \alpha) \hat{\beta}$$

$$0 \leq \chi(a) \leq 1 \; \forall a \in X$$

LPCA-Meaning of Variables and user defined Parameters

- $\chi(a)$: The variable that decides the post-processed label for data record a.
- $w(a)$ (user-input) : The weight associated with record a. Any weighing scheme can be used that gives more weight to records with high confidence scores of the classifier.
- ε (user-input): desired upper bound on DDP_C.
- β_i (user-input) : desired acceptance rate for the subgroup S_i. This can be given as direct input or using the equation.
- β_i^{initial} (user-input): The acceptance rates of subgroup S_i in the training data.
- $\hat{\beta}$ (user-input): Desired acceptance rate of all subgroups.
- α (user-input): Controls how much importance is given to β_i^{initial} and $\hat{\beta}$.

Extending configuration models for Equalized Odds

Configuration Model $[\beta] = \{\beta_1, \beta_2 \cdots \beta_m\}$ where $\beta_j = \mathbb{P}[\hat{Y} = 1 | S = S_j]$ and $j \in [m]$ is simply the acceptance rate of the subpopulation S_j.
Ground Truth $[\eta] = \{\eta_1, \eta_2 \cdots \eta_m\}$

Difference of Equalized Odds (DEO_M)

- Analogous to DDP we can extend the definition of difference of equalized odds multiple subpopulation case.
- $DEO_M = (\max_j TPR_j - \min_j TPR_j) + (\max_j FPR_j - \min_j FPR_j)$.
- $DEO_M = (\max_j \beta_j \cdot \Delta_j - \min_j \beta_j \cdot \Delta_j) + (\max_j \beta_j \cdot \Delta'_j - \min_j \beta_j \cdot \Delta'_j)$.

  where $\Delta_j = \frac{\mathbb{P}[Y=1|\hat{Y}=1,S=S_j]}{\mathbb{P}[Y=1|S=S_j]}$ and $\Delta'_j = \frac{\mathbb{P}[Y=0|\hat{Y}=1,S=S_j]}{\mathbb{P}[Y=0|S=S_j]}$

Decidability problem for of Equalized Odds

Problem (Equalized Odds)

Let X be a universal set with $|X| = n$, such that each element $x \in X$ has a tag $r(x) \in \{0,1\}$. Given a collection $S = \{S_1, S_2, \cdots S_m\}$ of subsets of X as the set of all subpopulations across the sensitive attributes, a configuration $[\beta] = (\beta_1, \beta_2, \ldots \beta_m)$ and (tpr, fpr), decide whether there exists a $0/1$ labelling of the elements of X such that the configuration can be realized with the given tpr and fpr for all the m sets w.r.t the given tags $r(.)$.

*Combined Linear Program (***LPCA** *and* **LPCEO***)*

LPCEO

LPCA

$$\min \sum_{a \in X} \chi(a) w(a)$$

$$\left(\sum_{a \in S_i} \chi(a) \right) \geq \beta_i |S_i| \; \forall S_i \in S$$

$$\left(\sum_{a \in S_i} \chi(a) \right) \leq (\beta_i + \epsilon) |S_i| \; \forall S_i \in S_j$$

$$0 \leq \chi(a) \leq 1 \; \forall a \in X$$

$$\beta_{ij} = (\alpha) \beta_{ij}^{\text{initial}} + (1 - \alpha) \hat{\beta}$$

$$\hat{S}_i \sum_{a \in S_i; r(a)=1} \chi(a) r(a) \geq tpr_i \; \forall i \in [m]$$

$$\hat{S}_i \sum_{a \in S_i; r(a)=1} \chi(a) r(a) \leq tpr_i + \delta_1 \; \forall i \in [m]$$

$$\hat{S}'_i \sum_{a \in S_i; r(a)=0} \chi(a)(1 - r(a)) \geq fpr_i \; \forall i \in [m]$$

$$\hat{S}'_i \sum_{a \in S_i; r(a)=0} \chi(a)(1 - r(a)) \leq fpr_i + \delta_2 \; \forall i \in [m]$$

$$0 \leq tpr_i, fpr_i \leq 1 \; \forall i \in [m]$$

Results for **LPC**

We show that **LPC** can realize very low DDP_C (and DDP_M) for various real world datasets and also for synthetic dataset with a large number of sensitive attributes (k).

Datasets	$\beta = 0.1$	$\beta = 0.25$	$\beta = 0.4$	**Synthetic**	$\beta = 0.1$	$\beta = 0.25$	$\beta = 0.35$
Adult	0.004	0.0025	0.0045	$k = 2$	0.0005	0.0007	0.0004
COMPAS	0.002	0.002	0.0005	$k = 6$	0.0006	0.0006	0.0007
Bank	0.0004	0.0005	0.0006	$k = 10$	0.001	0.001	0.001
German	0.009	0.0025	0.008	$k = 20$	0.1809	0.0236	0.0177

Table 1. ϵ (DDP_C) that leads to feasibility of **LPC** for various datasets and acceptance rates (β). For Synthetic datasets results shown for different number of sensitive groups (k) and acceptance rates (β). In synthetic dataset each of the synthetic attributes are considered binary. So $k = 20$ implies that there are 20 sensitive attributes and in all 40 possible subpopulations.

Performance Metrics

- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$.
- **Precision** = $\frac{TP}{TP+FP}$
- **Recall** = $\frac{FP}{TP+FN}$
- Usually we consider Fairness vs Accuracy tradeoffs, but is accuracy always the best performance metric ?
- In the context of fairness when we are selecting items, it might be better to focus on the selected set and how many TPs we have in that set.
- How to characterize this ? Depends on the natural acceptance of each subpopulation in the training data.

Performance Metrics

- If the chosen acceptance rate $\beta_i < \beta_i^{\text{natural}}$, then precision is a better metric, otherwise recall is better. Why?
- Let us say our target configuration is $\{\beta, \beta, \ldots, \beta\}$ a particular set (S_j) has N elements, then precision = $\frac{TP}{\beta \cdot N}$. Since $\beta_i < \beta_i^{\text{natural}}$, it is better to look at how many TP are there in the selected set.
- Also recall = $\frac{FP}{\beta_j^{\text{natural}} \cdot N}$. Thus when $\beta_i > \beta_i^{\text{natural}}$, it is better to compare the total number of FPs.
- We consider the weighted versions of precision and recall which is averaged over the cardinalities of each of the subpopulations.

Comparative Results for DDP_M

We compare configuration-wise performance of **LPCA** with diverse fair classification, ranking and subset-selection algorithms.

	Baseline	D_{Acc}	D_{Prec}	D_{Recall}	**Baseline**	D_{Acc}	D_{Prec}	D_{Recall}
Adult	Agarwal	0.0346	0.0306	-0.025	Madras	0.0002	0.0288	0.1416
	Zafar	0.0013	0.0677	0.0497	Yang	0.0347	0.221	NA
	Padala	0.0173	0.0669	NA	Mehrotra	0.0196	0.0349	0.1214
	DELTR	0.1849	0.5333	0.6093	Greedy-Fair	0.1716	0.3582	0.5372
Bank	Agarwal	0.0042	0.060	NA	Madras	0.0019	-0.028	0.2695
	Zafar	0.0042	0.0664	NA	Yang	0.0043	0.0733	NA
	Padala	0.0119	0.0193	NA	Mehrotra	0.0024	0.0095	0.1222
	DELTR	0.083	0.5724	NA	Greedy-Fair	0.0904	0.578	0.4461
COMPAS	Agarwal	0.0039	NA	0.0102	Madras	0.0015	0.0078	0.0575
	Zafar	0.0013	0.0046	-0.008	Yang	0.0697	0.1103	0.1060
	Padala	0.0028	NA	0.0018	Mehrotra	0.0101	0.0168	-0.0035
	DELTR	0.1717	0.2804	0.1806	Greedy-Fair	0.14	0.0714	0.144
German	Agarwal	0.0232	0.0098	0.0377	Madras	0.0015	0.0078	0.0575
	Zafar	0.0199	0.0162	0.0022	Yang	0.0244	0.0158	0.0173
	Padala	0.055	0.0378	0.0417	Mehrotra	0.0300	0.0278	-0.0026
	DELTR	0.1434	NA	0.1044	Greedy-Fair	0.3278	0.0522	NA

Table 3. Difference in average weighted precision (of higher acceptance class), weighted recall (of lower acceptance class) and overall accuracy over the 4 configurations, with **LPCA** for different datasets. Positive values mean **LPCA** is better. DELTR, Greedy-Fair and Madras perform on single binary sensitive attributes.

Comparative Results for DDP_M

Comparison of least DDP_M achieved by different baselines.

	Adult		Bank		COMPAS		German	
Baseline	DDP_M	Accuracy	DDP_M	Accuracy	DDP_M	Accuracy	DDP_M	Accuracy
Zafar	0.1074	0.8357	0.0656	0.9027	0.0668	0.6559	0.0841	0.74
Agarwal	0.0711	0.8035	0.0335	0.9049	0.0251	0.6344	0.0895	0.75
Yang	0.018	0.7812	0.0069	0.8935	0.0356	0.5580	0.0655	0.7333
Padala	0.0658	0.8031	0.02	0.8735	0.0154	0.6256	0.0396	0.69
Mehrotra	0.0049	0.8061	0.0198	0.9044	0.0028	0.63	0.0025	0.7066
Madras	0.0332	0.8358	0.0403	0.9075	0.0403	0.6616	0.0534	0.7200
DELTR	0.0034	0.6529	0.0048	0.8291	0.0049	0.4981	0.0312	0.6066
Greedy-Fair	0.0004	0.56	0.0688	0.68	0.0014	0.5033	0.0073	0.4733
LPCA	**0.0049**	**0.8444**	0.008	**0.908**	**0.0009**	**0.6723**	0.0075	**0.7533**

Table 2. The *minimum* DDP_M achieved by various baselines along with the corresponding **test accuracies** for various datasets. In almost all the datasets, the least DDP_M and the highest accuracy in that configuration is achieved by **LPCA**. The results of DELTR (ranking) and Greedy-fair (subset-selection) are obtained for single (binary) sensitive attribute their values and hence are not compared with other algorithms performing on multi-attribute case.

Comparative Results for DEO_M

We compare configuration wise performance of **LPCEO** with various fair classification algorithms.

	Adult		Bank		COMPAS		German	
Baseline	D_{DEO_M}	$\boldsymbol{D_{Acc}}$	D_{DEO_M}	$\boldsymbol{D_{Acc}}$	D_{DEO_M}	$\boldsymbol{D_{Acc}}$	D_{DEO_M}	$\boldsymbol{D_{Acc}}$
Zafar	0.1239	0.0354	0.3199	0.2866	0.083	0.0197	0.0695	-0.0022
Agarwal	0.0678	0.0332	-0.0992	0.0068	-0.0284	-0.0181	0.0360	0.0183
Padala	0.06281	0.0282	NA	NA	0.0089	0.0085	0.0165	0.0078
Yang	-0.081	0.0448	0.2817	0.0031	0.0893	-0.0032	0.0649	0.0167
Romano	0.1175	0.0360	0.1430	0.015	-0.0109	0.0052	0.0726	0.0092
Mary	0.0387	-0.0048	0.0602	0.0652	0.0224	0.0227	0.041	0.0025
Cho	0.0883	0.005	NA	NA	0.0311	0.0167	NA	NA
Hardt	0.0389	0.0071	NA	NA	-0.04	0.014	0.0003	0.0092

Table 4. Average Difference (over 4 configurations) of DEO_M and Accuracy between various baselines and **LPCEO**. Positive values imply that **LPCEO** is performing better. Zafar, Padala and Romano can handle only single binary sensitive attributes. NA entries refer to a scenario in which the baseline is giving trivial classification as output (all 1's or all 0's) that results in $DEO_M = 0$ and hence **LPCEO** also attains the same accuracy and DEO_M at that configuration.

Fair Gerrymandering

- Ensure fairness in several overlapping subpopulations proposed by Kearns et al. (ICML 2018). Studied seriously, in a paper by Yang et al. (NeurIPS 2020).
- **LPCA** and **LPCEO** show improved accuracy and DEO_M on configurations of Yang.

		Yang		**LPCEO**	
Baseline	DDP_M	Accuracy	DEO_M	Accuracy	DEO_M
Adult	0.0683	0.7899	0.2716	0.828	0.3261
Bank	0.0632	0.9016	0.1323	0.9073	0.1272
COMPAS	0.1135	0.6564	0.113	0.6581	0.1366
German	0.1343	0.7498	0.2071	0.7683	0.2207

Table 5. Comparison of test accuracies, and DEO_M between Yang et al. [49] and **LPCEO** averaged over four configurations generated by Yang on various datasets. The number of gerrymandering groups for Adult, Bank, COMPAS and German datasets are 13, 10, 8 and 8 respectively.

Perturbation of Training data

Unlike other classification and ranking algorithms, our results remain unaffected by change in sensitive data in the training data.

	Original Train-Test		Train Modified	
Baseline	Accuracy	DDP_M	Accuracy	DDP_M
Zafar	0.831	0.21	0.8218	0.18
Padala	0.8031	0.0658	0.8346	0.1766
Agarwal	0.8035	0.0711	0.9056	0.052
Yang	0.7812	0.018	0.7985	0.1195
Madras	0.8323	0.0185	0.8505	0.1972
DELTR	0.6529	0.0034	0.6126	0.0044
LPCA	0.8335	0.019	0.8326	0.019

	Original Train-Test		Train Modified	
Baseline	Accuracy	DEO_M	Accuracy	DEO_M
Zafar	0.811	0.23	0.805	0.29
Padala	0.8126	0.1633	0.8395	0.189
Agarwal	0.8543	0.1324	0.8617	0.3987
Yang	0.795	0.2564	0.795	0. 2581
Mary	0.8430	0.2787	0.8402	0.4333
Romano	0.8088	0.1612	0.8046	0.2239
Cho	0.8423	0.2176	0.8452	0.4073
Hardt	0.8396	0.2948	0.8645	0.4598
LPCEO	0.8637	0.1778	0.8637	0.1778

Table 6. Test accuracy and DDP_M and DEO_M of various baselines on the Adult dataset with change in the 'sex' and 'race' attribute in the training data keeping the test data unchanged. We have chosen our DDP_M similar to Yang by tuning $\alpha, \hat{\beta}$ to show comparative results. The DELTR algorithm can only cater to single binary sensitive attribute.

Parameter Study

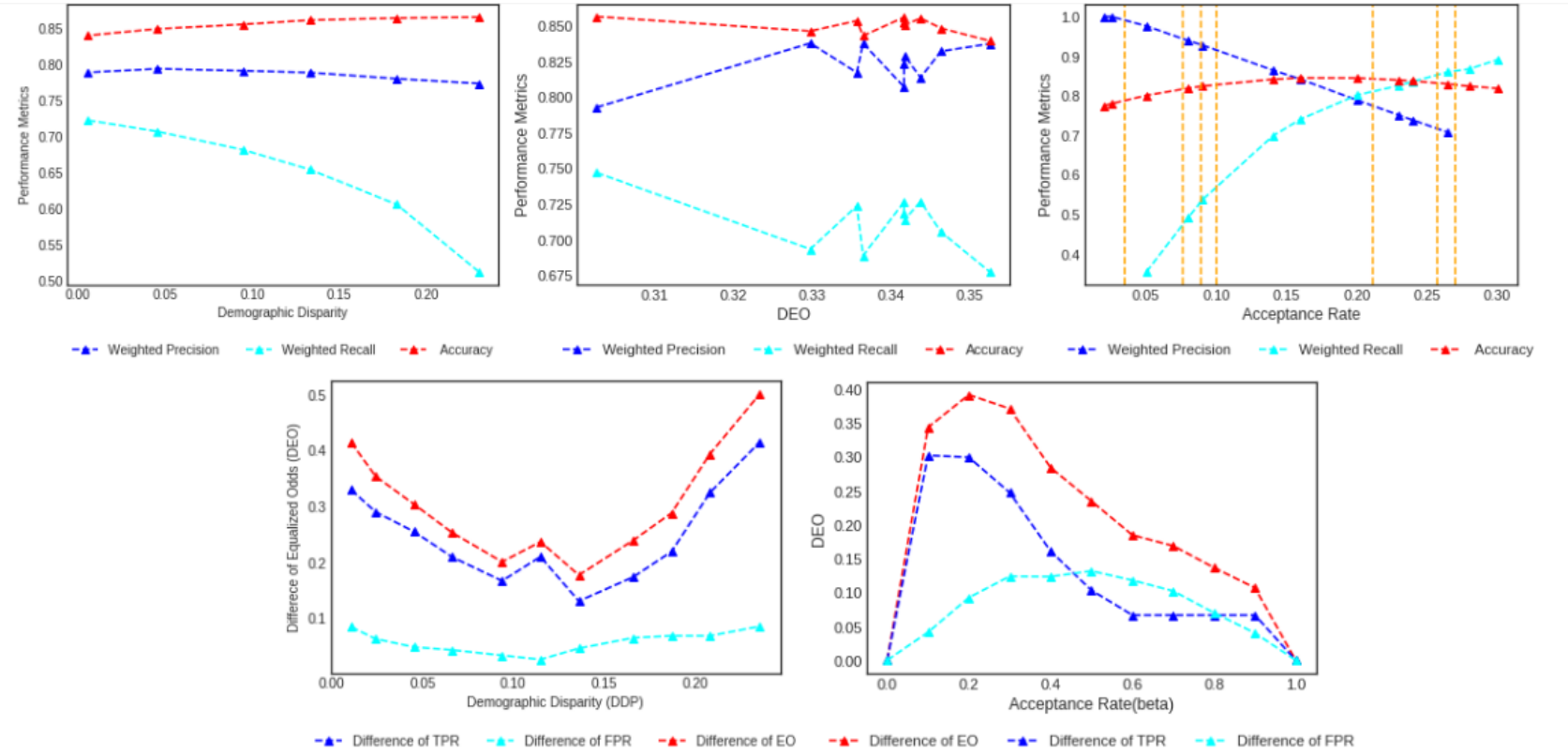


Fig. 1. (First Row) Variation of different performance metrics on configs generated by **LPCA** and **LPCEO** on Adult dataset with (a) increasing DDP_M and fixed $\hat{\beta} = 0.2$, (b) increasing DEO_M, (c) increasing $\hat{\beta}$. The seven vertical lines correspond to the initial configuration of acceptance rates of each of the seven subpopulations and fixed $\hat{\beta} = 0.2$. (Second Row) Variation of DEO_M on configurations generated by **LPCEO** with (d) increasing DDP_M by varying α and $\beta = 0.2$, (e) increasing $\hat{\beta}$ and $\alpha = 0$.

Future Work

- Investigate whether the configuration model can be applied to other notions of fairness like *counter-factual fairness* and *calibration*.
- A classifier is said to be counterfactually fair if it satisfies,

$$\mathbb{P}[\hat{Y}_{z\leftarrow z_0}(U) = y|\mathbf{x} = x, z = z_0] = \mathbb{P}[\hat{Y}_{z\leftarrow z_0'}(U) = y|\mathbf{x} = x, z = z_0]$$

- A classifier is said to satisfy calibration if

$$\mathbb{P}[\hat{Y} = 1|S = s, z = z_0] = \mathbb{P}[\hat{Y} = 1|S = s, z = z_0']$$

Future Work

- Can the notion of Equalized Odds be extended to unsupervised graph theoretic scenarios like *fair clustering* for multiple overlapping attributes?
- In the context of *interpretability* and *explainability* can we analyze and understand the relationship between *descriptive accuracy* and fairness ?
- Will it be helpful to use non-linear optimization paradigms like SDP or other cone programming techniques as has been used by Zafar et al. (JMLR 2019) in case of Equalized Odds?

Conclusion

- In this work, we have developed LP-based batch classification algorithms to ensure demographic parity and equalized odds.
- The configuration model efficiently handles the case of multiple overlapping subpopulations and gerrymandering and provides flexibility to the user.
- Our algorithms are LP-based and conceptually simpler and faster than state-of-the-art fair classification, ranking and subset selection algorithms and provide modest improvement in performance.