From Parity to Preference-based Notions of Fairness in Classification

Muhammad Bilal Zafar

joint work with

Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, Adrian Weller

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

UNIVERSITY OF
CAMBRIDGE

Existing notions of fairness in classification

- Inspired by anti-discrimination legislation
  - Rooted in social sciences ideas of parity (or equality)

Existing notions of fairness in classification

- Inspired by anti-discrimination legislation
  - Rooted in social sciences ideas of parity (or equality)

- Parity in treatment (avoid *disparate treatment*)
  - Outcome should not depend on the sensitive group membership (e.g., men, women)

Existing notions of fairness in classification

- Inspired by anti-discrimination legislation
  - Rooted in social sciences ideas of parity (or equality)

- Parity in treatment (avoid *disparate treatment*)
  - Outcome should not depend on the sensitive group membership (e.g., men, women)

- Parity in impact (avoid *disparate impact*)
  - Similar fraction of beneficial outcomes received by different groups (e.g., men, women)
  - Beneficial outcome: Granted loan, etc.

Parity can be a very stringent criterion

- Impossibility results
  - Certain parities cannot be achieved together
  - [Chouldechova, Big Data'17; Kleinberg et al., ITCS'17]

Parity can be a very stringent criterion

- Impossibility results
  - Certain parities cannot be achieved together
  - [Chouldechova, Big Data'17; Kleinberg et al., ITCS'17]
- High cost of parity fairness
  - Inherent tradeoff between accuracy and fairness
  - [Corbett-Davies et al., KDD'17]

Parity can be a very stringent criterion

- Impossibility results
  - Certain parities cannot be achieved together
  - [Chouldechova, Big Data'17; Kleinberg et al., ITCS'17]
- High cost of parity fairness
  - Inherent tradeoff between accuracy and fairness
  - [Corbett-Davies et al., KDD'17]
- All groups may end up losing
  - Benefits for all groups decrease
  - [Zafar et al., AISTATS'17]

From parity to preference

- Explore preference-based notions of fairness

- Inspired by fair division in economics
  - Parity treatment to preferred treatment
  - Parity impact to preferred impact

- Outcomes may not follow parity
  - Groups involved prefer their respective outcomes
  - Gains in accuracy over parity fairness

This talk

- Defining preference-based fairness notions

- Formalizing preference-based notions

- Training preferentially-fair classifiers
  - Mechanism design
  - Case study: NYPD SQF dataset

This talk

- **Defining preference-based fairness notions**
- Formalizing preference-based notions
- Training preferentially-fair classifiers
  - Mechanism design
  - Case study: NYPD SQF dataset

Some preliminaries

- Groups based on sensitive feature
  - Gender (men, women)
  - Race (African-American, Hispanic, White, …)
- Classification outcomes
  - Beneficial (+ve class) and not beneficial (-ve class)
- Group preference
  - Each group collectively prefers classifier with most beneficial outcomes

Some preliminaries

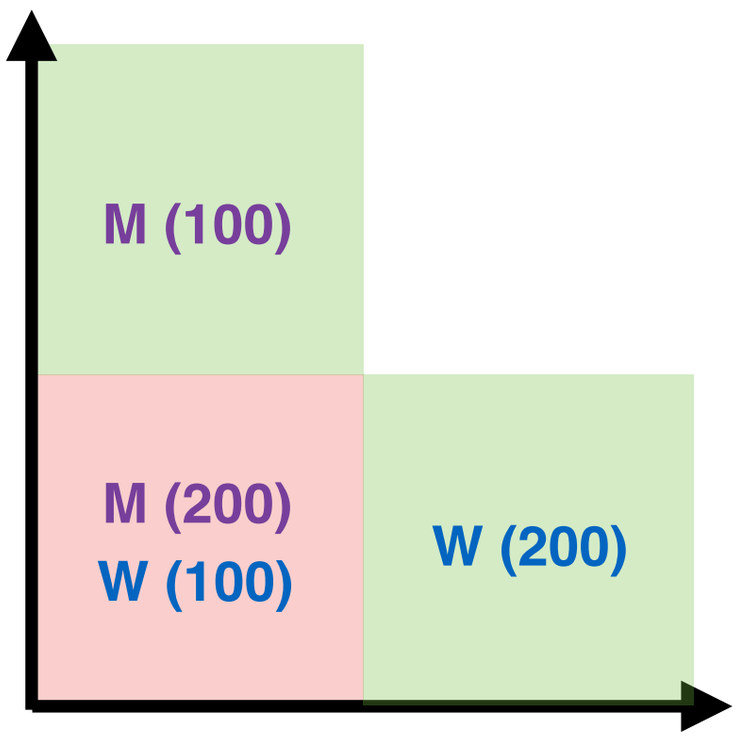
- Groups based on sensitive feature
  - Gender (men, women)
  - Race (African-American, Hispanic, White, …)
- Classification outcomes
  - Beneficial (+ve class) and not beneficial (-ve class)
- Group preference
  - Each group collectively prefers classifier with most beneficial outcomes
  - Classifier 1 (10% beneficial outcomes)
  - Classifier 2 (30% beneficial outcomes) ✔
  - Classifier 3 (15% beneficial outcomes)

Parity treatment to preferred treatment

Parity treatment (Existing notion)

- Anti-discrimination notion of parity
- Parity: Individuals in G_1 get same decisions by posing as G_2
- Changing sensitive feature should not change benefits
- Cannot use group-conditional classifiers

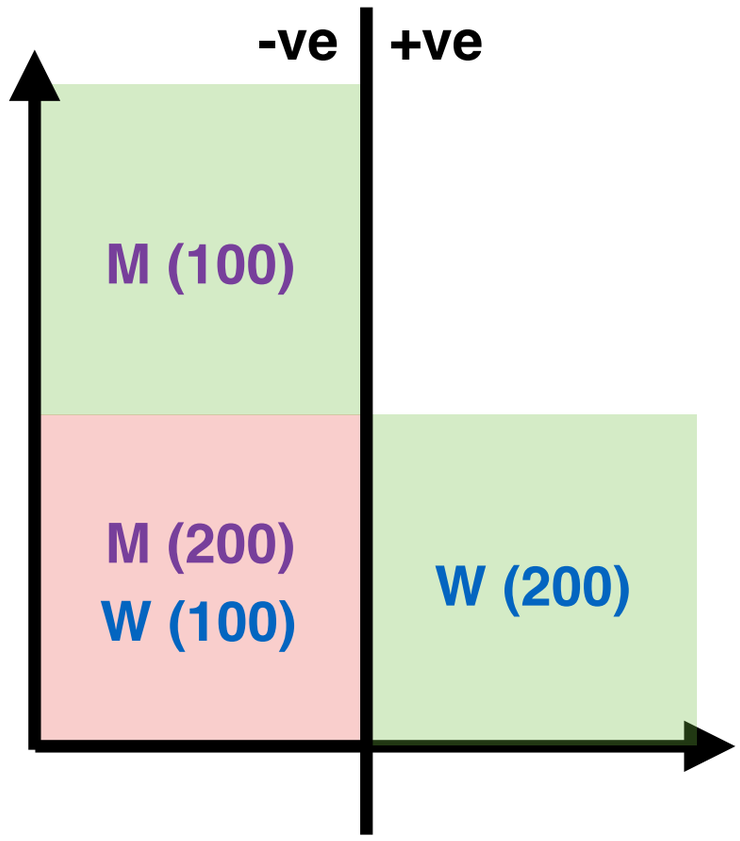
An example of parity treatment



An example of parity treatment

Parity treatment

Acc: 0.83 (500/600)



Parity treatment to preferred treatment

Parity treatment (Existing notion)

- Anti-discrimination notion of parity
- Parity: Individuals in G_1 get same decisions by posing as G_2
- Changing sensitive feature should not change benefits
- Cannot use group-conditional classifiers

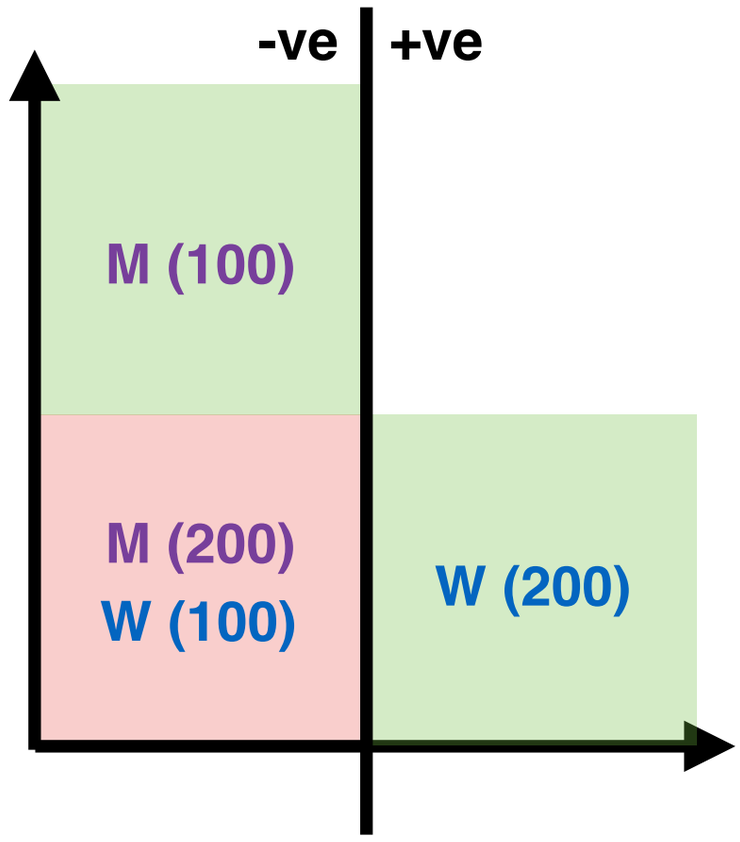
Preferred treatment (New notion)

- Fair division notion of group envy-freeness
- Envy-freeness: G_1 collectively will not prefer to be posing as G_2
- Changing sensitive feature should not increase benefits
- Can use group-conditional classifiers

Parity treatment vs. preferred treatment

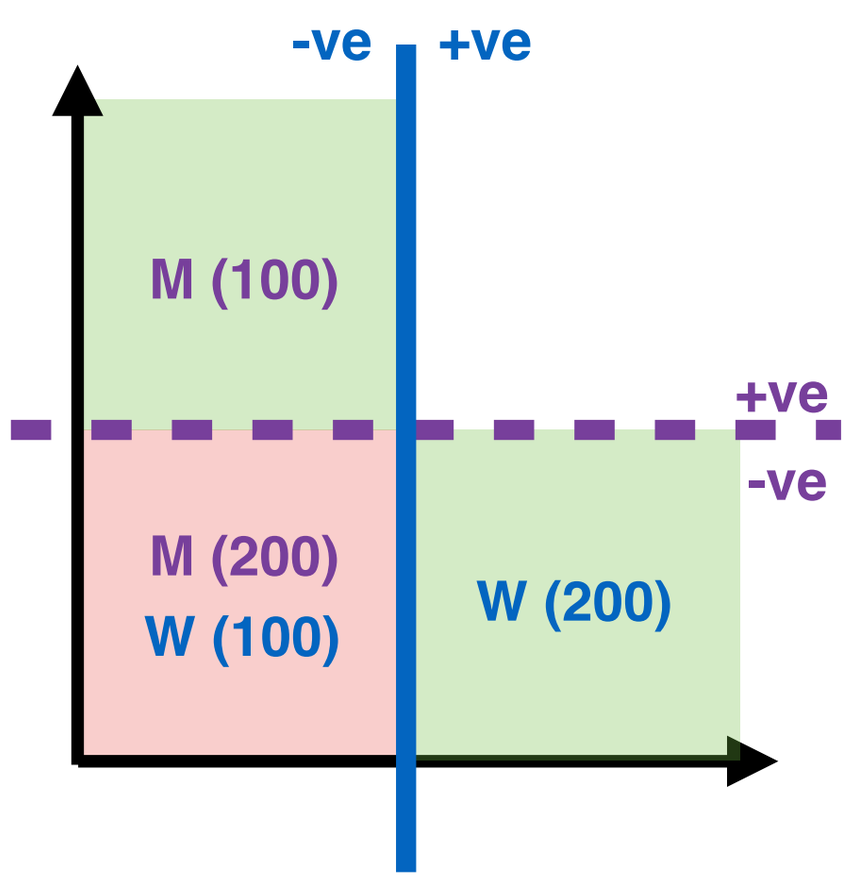
Parity treatment

Acc: 0.83 (500/600)



Preferred treatment

Acc: 1.00 (600/600)



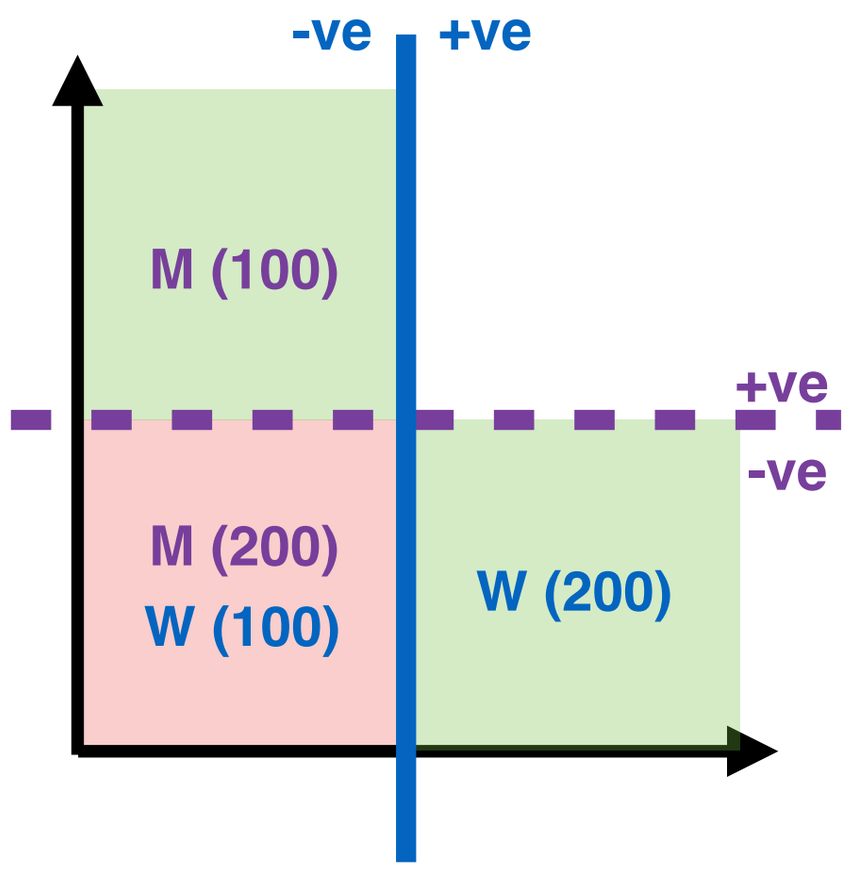
Parity treatment vs. preferred treatment

Men using **men's** classifier
33% (100/300) get benefits

Men using **women's** classifier
0% get benefits

Preferred treatment

Acc: 1.00 (600/600)



Parity treatment vs. preferred treatment

Men using **men's** classifier
33% (100/300) get benefits

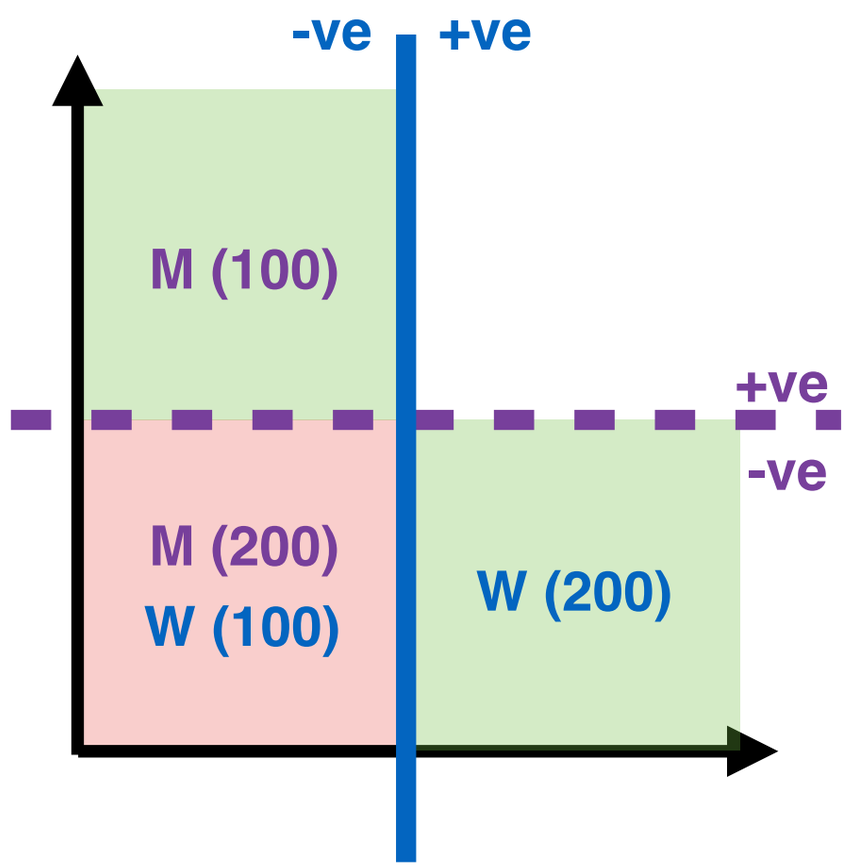
Men using **women's** classifier
0% get benefits

Women using **women's** classifier
66% (200/300) get benefits

Women using **men's** classifier
0% get benefits

Preferred treatment

Acc: 1.00 (600/600)

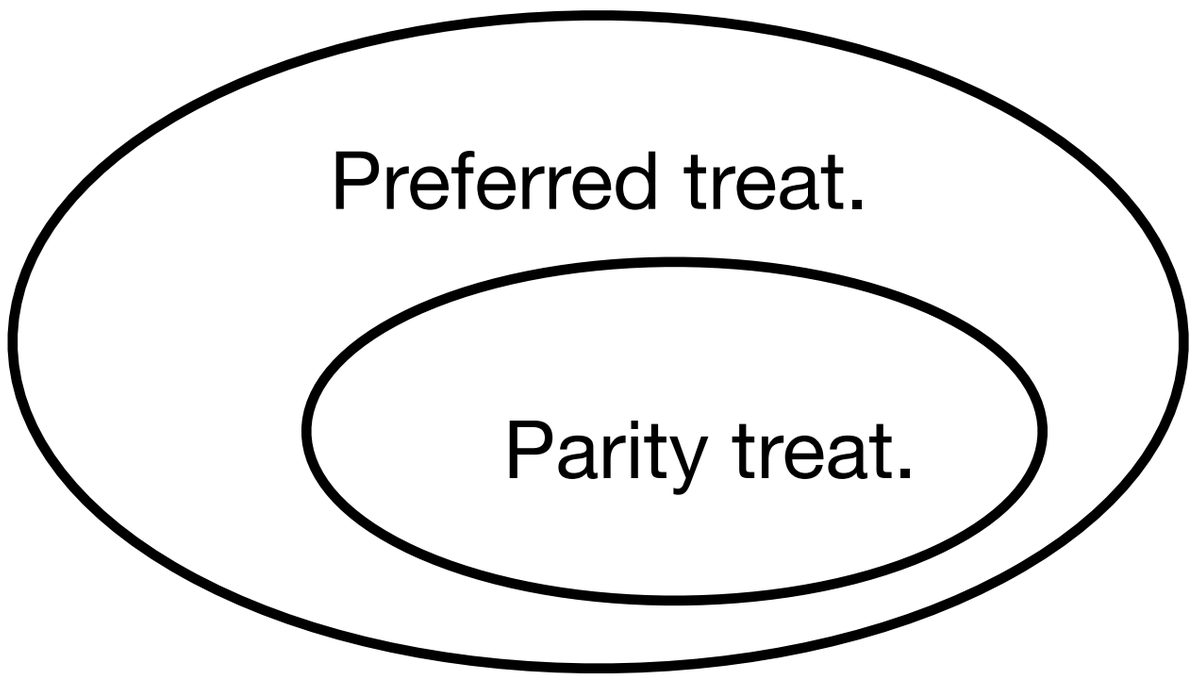


No incentive to change group

Avoids reverse discrimination (or lowering the bar) claims

Preferred treatment: A relaxation of parity treatment

- **Parity treatment:** Changing sensitive feature does not change group benefits
- **Preferred treatment:** Changing sensitive feature does not increase group benefits



Each solution satisfying parity treatment also satisfies preferred treatment
(Room for more accurate solutions)

Parity impact to preferred impact

Parity impact (Existing notion)

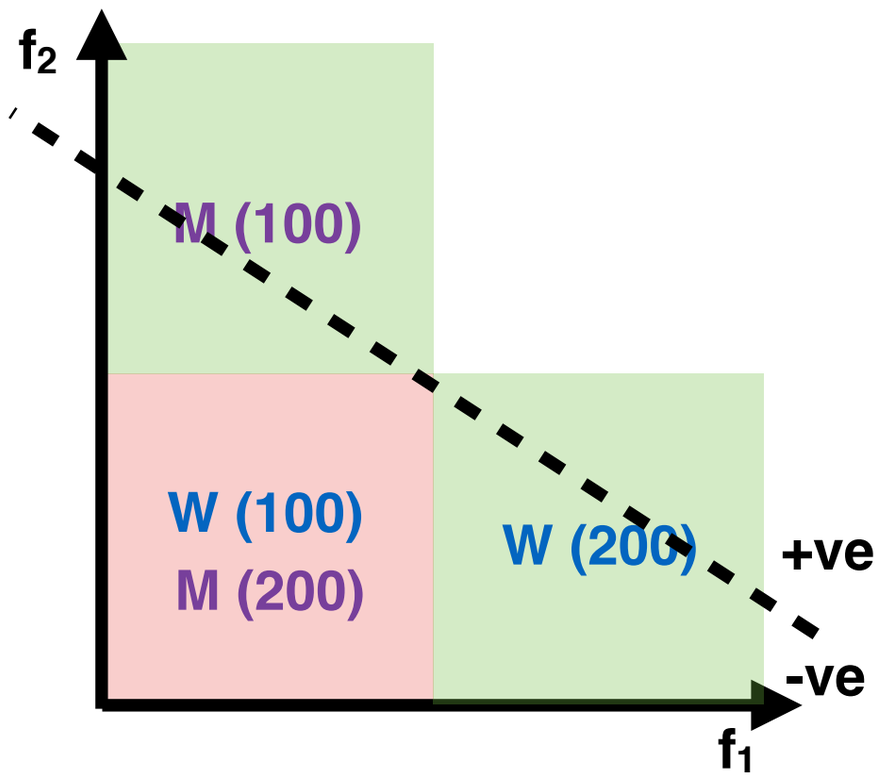
- Anti-discrimination notion of parity
- Parity: 20% from G1 accepted, 20% from G2 accepted
- All groups gets equal fraction of beneficial outcomes

An example of parity impact

Parity impact

Acc: 0.72

Benefit: 22% (M), 22%(W)



Parity impact to preferred impact

Parity impact (Existing notion)

- Anti-discrimination notion of parity
- Parity: 20% from G1 accepted, 20% from G2 accepted
- All groups gets equal fraction of beneficial outcomes

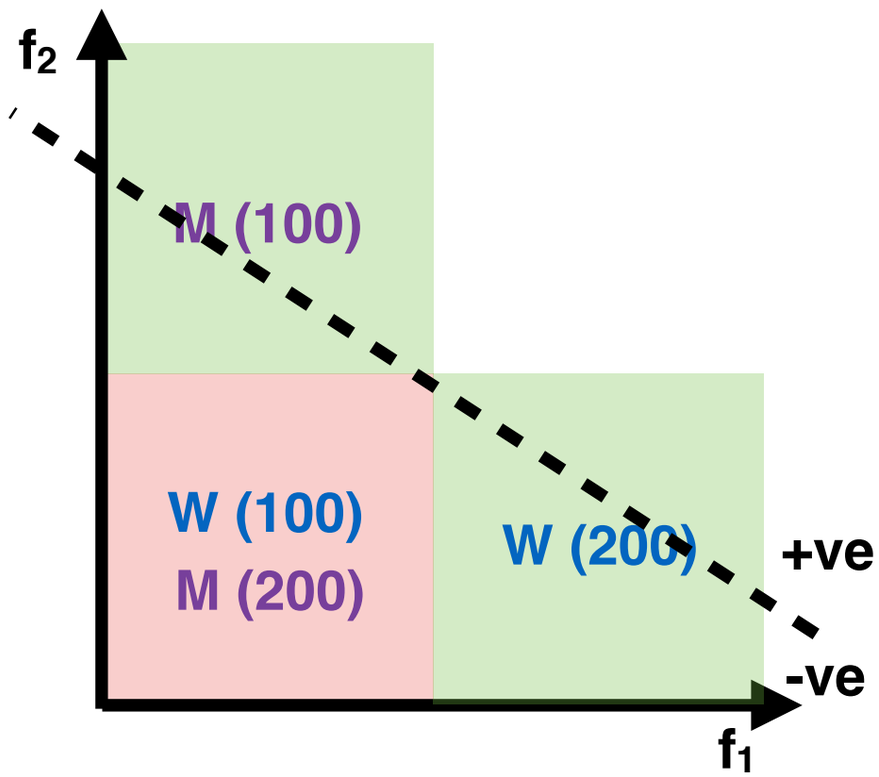
Preferred impact (New notion)

- Fair division notion of bargaining solution
- Bargaining solution: 25% from G1 accepted, 40% from G2 accepted (or, revert to parity)
- All group gets at least as much benefits as parity impact

Parity impact vs. preferred impact

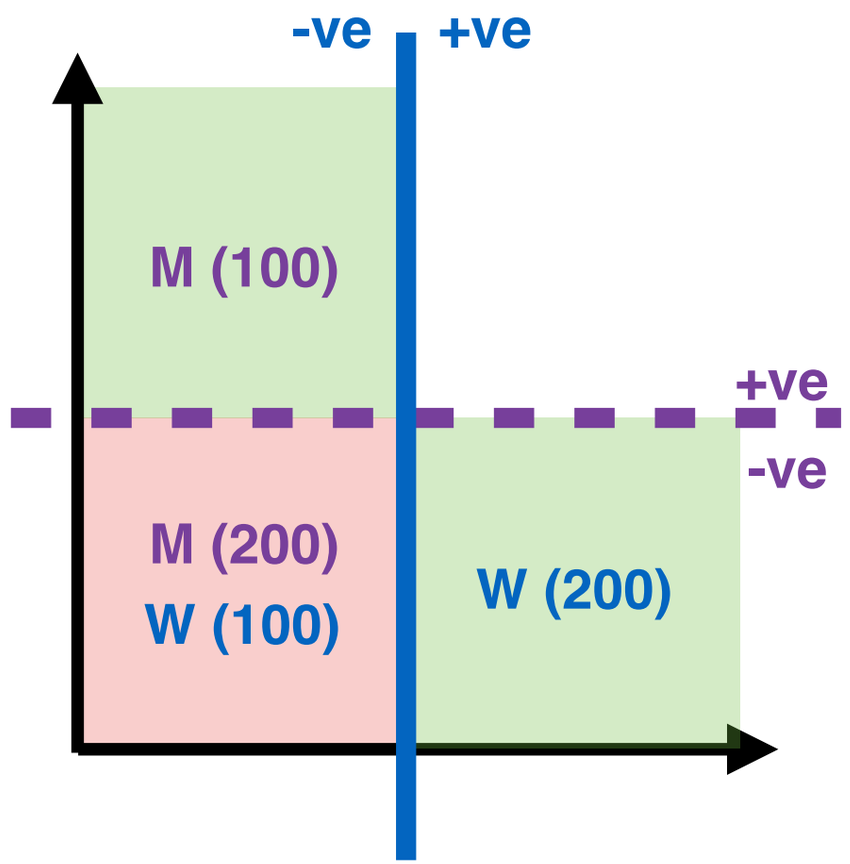
Parity impact

Acc: 0.72
Benefit: 22% (M), 22%(W)



Preferred impact

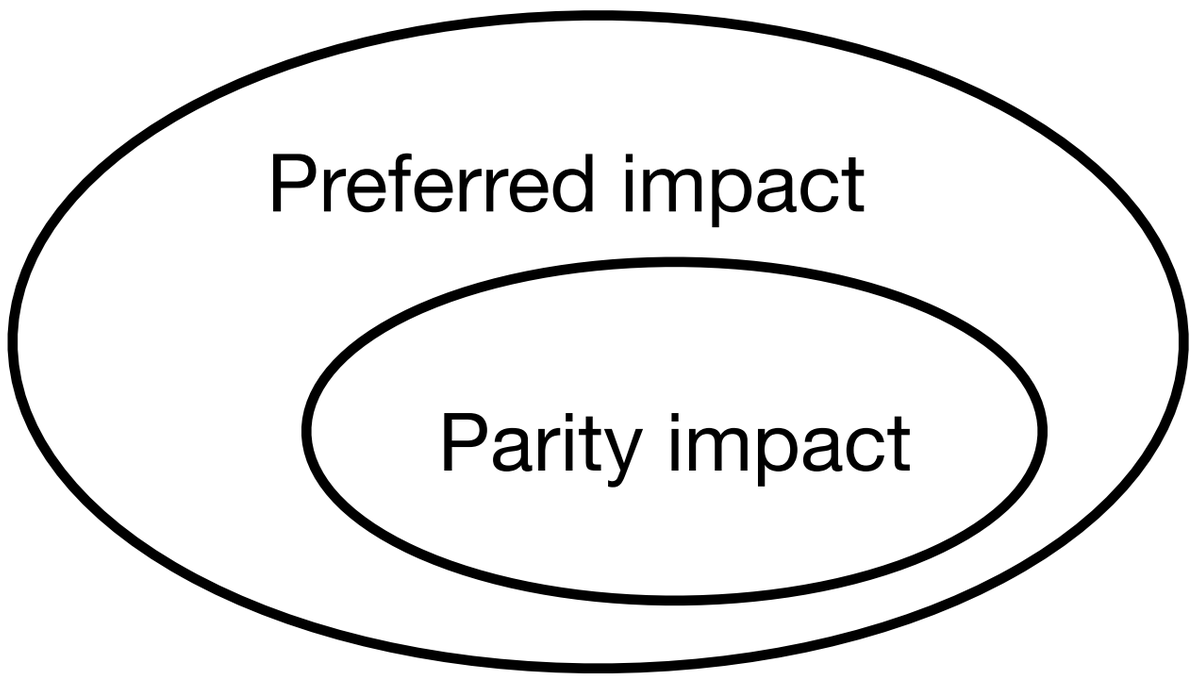
Acc: 1.00
Benefit: 33% (M), 67%(W)



Both groups have incentive to move to preferred impact

Preferred impact: A relaxation of parity impact

- **Parity impact:** All groups should get similar fraction of benefits
- **Preferred impact:** All groups should get at least as much benefits as parity impact



Each solution satisfying parity impact also satisfies preferred impact
(Room for more accurate solutions)

This talk

- Defining preference-based fairness notions

- **Formalizing preference-based notions**

- Training preferentially-fair classifiers
  - Mechanism design
  - Case study: NYPD SQF dataset

Formalizing preferential fairness

- Given
  - Sensitive feature group z
  - Classifier $\boldsymbol{\theta}$ with outcomes $\hat{y} \in [-1, 1]$
  - $B_z(\boldsymbol{\theta}) = P(\hat{y} = 1 \mid z, \boldsymbol{\theta})$

Formalizing preferential fairness

- Given
  - Sensitive feature group z
  - Classifier $\boldsymbol{\theta}$ with outcomes $\hat{y} \in [-1, 1]$
  - $B_z(\boldsymbol{\theta}) = P(\hat{y} = 1 \mid z, \boldsymbol{\theta})$

Benefits for men by $\boldsymbol{\theta}$

$$B_{\male}(\boldsymbol{\theta}) = P(\hat{y} = 1 \mid \male, \boldsymbol{\theta})$$

Benefits for women by $\boldsymbol{\theta}$

$$B_{\female}(\boldsymbol{\theta}) = P(\hat{y} = 1 \mid \female, \boldsymbol{\theta})$$

Formalizing preferential fairness

Preferred treatment

$$B_{♂}(\boldsymbol{\theta}_{♂}) \geq B_{♂}(\boldsymbol{\theta}_{♀})$$
$$B_{♀}(\boldsymbol{\theta}_{♀}) \geq B_{♀}(\boldsymbol{\theta}_{♂})$$

- Train separate classifiers for men ($\boldsymbol{\theta}_{♂}$) and women ($\boldsymbol{\theta}_{♀}$)
- Benefits for men with their classifier more than benefits for men with women's classifier
- Benefits for women with their classifier more than benefits for women with men's classifier

Formalizing preferential fairness

Preferred treatment

$$B_{♂}(\boldsymbol{\theta}_{♂}) \geq B_{♂}(\boldsymbol{\theta}_{♀})$$
$$B_{♀}(\boldsymbol{\theta}_{♀}) \geq B_{♀}(\boldsymbol{\theta}_{♂})$$

Preferred impact

Given parity impact classifiers $\boldsymbol{\theta}'_{♂}$ and $\boldsymbol{\theta}'_{♀}$

$$B_{♂}(\boldsymbol{\theta}_{♂}) \geq B_{♂}(\boldsymbol{\theta}'_{♂})$$
$$B_{♀}(\boldsymbol{\theta}_{♀}) \geq B_{♀}(\boldsymbol{\theta}'_{♀})$$

Benefits from parity impact classifier (constants)

This talk

- Defining preference-based fairness notions
- Formalizing preference-based notions
- Training preferentially-fair classifiers
  - **Mechanism design**
  - Case study: NYPD SQF dataset

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad B_{\male}(\boldsymbol{\theta}_{\male}) \geq \quad B_{\male}(\boldsymbol{\theta}_{\female})$$

$$B_{\female}(\boldsymbol{\theta}_{\female}) \geq \quad B_{\female}(\boldsymbol{\theta}_{\male})$$

Add preferred treatment constraints

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\male}) \geq \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\female})$$

$$P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\female}) \geq \quad P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\male})$$

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\male}) \geq \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\female})$$

$$P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\female}) \geq \quad P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\male})$$

- Non-convex for many well-known classifiers (logistic regression, SVM)
- Hard to compute efficiently

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\male}) \geq \quad P(\hat{y} = 1 | \male, \boldsymbol{\theta}_{\female})$$

$$P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\female}) \geq \quad P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\male})$$

Approximate positive class probability using ramp function

$$\max(0, \boldsymbol{\theta}^T \mathbf{x})$$

Non-zero when ŷ=1, zero otherwise

Training preferred treatment classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad \sum \max(0, \boldsymbol{\theta}_{\male}^T \mathbf{x}) \geq \sum \max(0, \boldsymbol{\theta}_{\female}^T \mathbf{x})$$

$$\sum_{\male} \max(0, \boldsymbol{\theta}_{\female}^T \mathbf{x}) \geq \sum_{\female} \max(0, \boldsymbol{\theta}_{\male}^T \mathbf{x})$$

Convex Convex

$$P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\female}) \quad P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\male})$$

Disciplined Convex-Concave Program (DCCP)

(can be approximated efficiently)

[Shen, Diamond, Gu, Boyd, 2016]

Training preferred impact classifier

Training preferred impact classifier

$\theta'_{\male}$ and $\theta'_{\female}$ parity impact classifiers

Training preferred impact classifier

Idea: Learn under constraints

$$\min \quad \sum_{\male} L(\mathbf{x}, y, \boldsymbol{\theta}_{\male}) + \sum_{\female} L(\mathbf{x}, y, \boldsymbol{\theta}_{\female})$$

$$\text{s.t} \quad \sum_{\male} \max(0, \boldsymbol{\theta}_{\male}^T \mathbf{x}) \geq \sum_{\male} \max(0, \boldsymbol{\theta}_{\male}^{'T} \mathbf{x})$$

$$\sum_{\female} \max(0, \boldsymbol{\theta}_{\female}^T \mathbf{x}) \geq \sum_{\female} \max(0, \boldsymbol{\theta}_{\female}^{'T} \mathbf{x})$$

$\boldsymbol{\theta}'_{\male}$ and $\boldsymbol{\theta}'_{\female}$ parity impact classifiers

Convex

Constant

$P(\hat{y} = 1 | \female, \boldsymbol{\theta}_{\female})$ $P(\hat{y} = 1 | \female, \boldsymbol{\theta}'_{\female})$

Disciplined Convex-Concave Program (DCCP)

This talk

- Defining preference-based fairness notions
- Formalizing preference-based notions
- Training preferentially-fair classifiers
  - Mechanism design
  - **Case study: NYPD SQF dataset**

Case study: NYPD SQF dataset

- Are the preferential fairness constraints effective?

- Does preferential fairness lead to accuracy gains over parity fairness?

Case study: NYPD SQF dataset

- Two sensitive feature groups
  - African-Americans and Whites
- Two classes
  - Pedestrian in possession of a weapon (non-beneficial)
  - No weapon (beneficial)

Toy classification task

Case study: NYPD SQF dataset

	Uncons.
Af. Americans	67% (34%)
Whites	12% (22%)
Accuracy	0.74

Dissimilar group benefits

Incentive to change group

Case study: NYPD SQF dataset

	Uncons.	Parity fairness
Af. Americans	67% (34%)	50% (50%)
Whites	12% (22%)	52% (52%)
Accuracy	0.74	0.61

Similar group benefits

No incentive to change group

Large drop in accuracy

Case study: NYPD SQF dataset

	Uncons.	Parity fairness	Preferred fairness
Af. Americans	67% (34%)	50% (50%)	78% (76%)
Whites	12% (22%)	52% (52%)	52% (33%)
Accuracy	0.74	0.61	0.68

Group benefits more than parity

No incentive to change group

Modest drop in accuracy

Case study: NYPD SQF dataset

- Are the preferential fairness constraints effective?
  - Yes. Each group prefers their outcome.

- Does preferential fairness lead to accuracy gains over parity fairness?
  - Yes. Smaller loss in accuracy.

Conclusion

- Preference-based notions of fairness
  - Each group prefers to pose as itself
  - Each group prefers the outcomes over parity impact
- Preference-based notions can lead to more accurate solutions

Paper at

https://tinyurl.com/preference-based-fairness