Fairness, Part II

Giulia Fanti

Fall 2019

Based in part on slides by Anupam Datta

Administrative

- HW4 out today
  - Fairness + anonymous communication (next unit)
  - You will have ~3 weeks
- Presentations starting on Wednesday
  - Upload your slides to Canvas by **midnight** the night before so we can download them in the morning
  - Sign up for groups on Canvas so that we can assign group grades
  - Presentation rubric on Canvas!
  - Volunteer in SV to share their laptop on Wednesday?

In-class Quiz

- On Canvas

Last time

- Group fairness
  - Statistical parity
  - Demographic parity
  - Ensures that same ratio of people from each group get the "desirable" outcome
- Individual fairness
  - Ensures that similar individuals are treated similarly
  - Can learn a fair classifier by solving linear program

Today

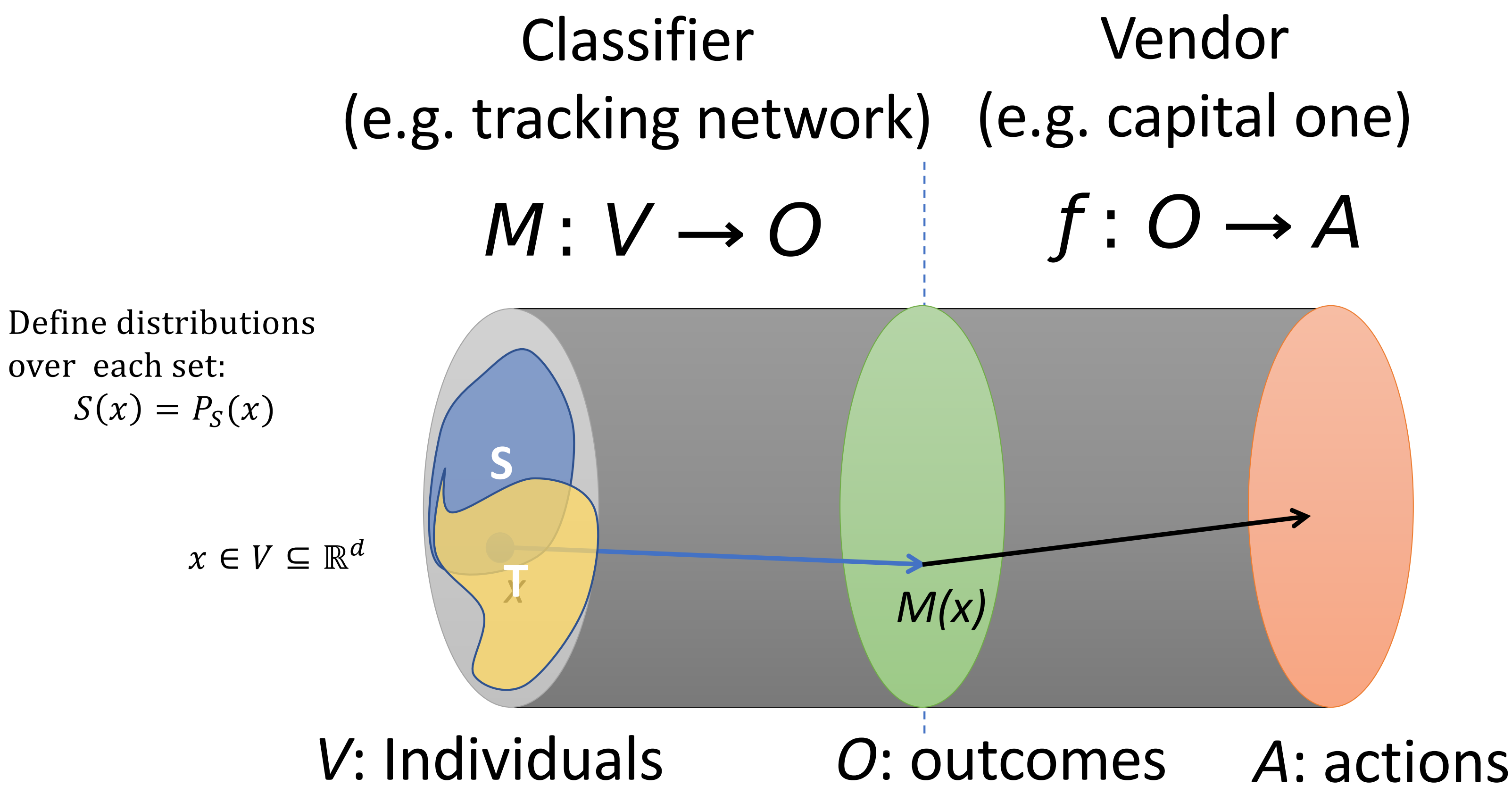
- When does individual fairness imply group fairness?
- Connections to differential privacy
- How do we take already-trained classifiers and make them fair?

Paper from last time:

Fairness Through Awareness

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

November 30, 2011


Classifier
(e.g. tracking network)
Vendor
(e.g. capital one)
$M : V \to O$
$f : O \to A$
Define distributions
over each set:
$S(x) = P_S(x)$
S
T
$x \in V \subseteq \mathbb{R}^d$
$M(x)$
V: Individuals
O: outcomes
A: actions

Individual fairness formulation:

Maximize utility

$$\max_{M_x} \mathbb{E}_{x \sim V} \mathbb{E}_{o \sim M_x} [U(x, o)]$$

Subject to fairness constraint

$$\text{s.t. } \|M_x - M_y\| \leq d(x, y) \ \forall\, x, y \in V$$

Q: What are the downsides to this formulation?

- Need a similarity metric between users
- Very high-dimensional LP may be difficult to solve
- Classifier must be trained *a priori* with fairness

When does Individual Fairness imply Group Fairness?

Suppose we enforce a metric *d.*

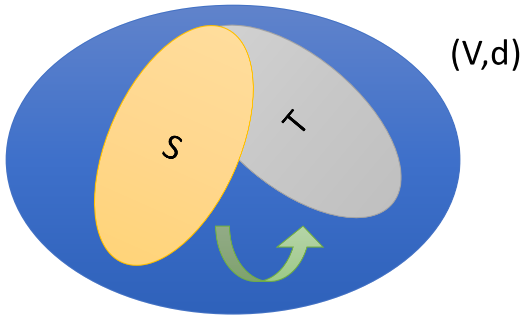
Question: Which *groups of individuals* receive (approximately) equal outcomes?

Answer is given by **Earthmover distance** (w.r.t. *d*) between the two groups.

How different are *S* and *T*?

Earthmover Distance:

"Cost" of transforming one distribution to another, by "moving" probability mass ("earth").



$$
\begin{aligned}
d_{EM}(S,T) = \quad & \min_h \quad \sum_{x,y\in V} h(x,y)d(x,y) \\
& \text{s.t.} \quad \sum_{y\in V} h(x,y) = S(x), \quad \forall x \in S \\
& \qquad\; \sum_{y\in V} h(y,x) = T(x), \quad \forall x \in T \\
& \qquad\; h(x,y) \geq 0
\end{aligned}
$$

h(x,y) – how much probability of x in S to move to y in T

Example: Compute Earth-Mover's Distance

- On document cam

$$
\begin{aligned}
d_{EM}(S,T) = \quad & \min_h \quad \sum_{x,y \in V} h(x,y) d(x,y) \\
& \text{s.t.} \quad \sum_{y \in V} h(x,y) = S(x), \quad \forall x \in S \\
& \qquad\quad \sum_{y \in V} h(y,x) = T(x), \quad \forall x \in T \\
& \qquad\quad h(x,y) \geq 0
\end{aligned}
$$

$$
\begin{aligned}
d_{EM}(S,T) = \quad \min_h \quad & \sum_{x,y \in V} h(x,y) d(x,y) \\
\text{s.t.} \quad & \sum_{y \in V} h(x,y) = S(x) \\
& \sum_{y \in V} h(y,x) = T(x) \\
& h(x,y) \geq 0
\end{aligned}
$$

$$
\text{bias}(d,S,T) = \max_{\text{M:d}-\text{Lipschitz model}} \mathrm{P}[M(x) = o | x \in S] - \mathrm{P}[M(x) = o | x \in T]
$$

Theorem:

Any Lipschitz mapping M satisfies group fairness up to bias(d,S,T).

Further,

$$
\text{bias}(d,S,T) \leq d_{EM}(S,T)
$$

Some observations

$$\text{bias}(d, S, T) = \max_{\text{M:d}-\text{Lipschitz model}} \text{P}[M(x) = o | x \in S] - \text{P}[M(x) = o | x \in T]$$

Theorem:
Any Lipschitz mapping M satisfies group fairness up to $\text{bias}(d, S, T)$.

- By definition, the bias is the maximum deviation from group fairness that can be achieved!
- Indeed, for TV distance between distributions and binary classification,
$$\text{bias}(d, S, T) = d_{EM}(S, T)$$
- Takeaway message: If your groups are very far away (in EMD), the Lipschitz condition can only get you so far in terms of group fairness!

Connections to Differential Privacy

$$\max_{M_x} \mathbb{E}_{x\sim V}\,\mathbb{E}_{o\sim M_x}\,[U(x,o)]$$
$$\text{s.t. } \left\|M_x - M_y\right\| \le d(x,y) \;\; \forall\, x,y \in V$$

What if we don't use TV distance for $\left\|M_x - M_y\right\|$?

$$\|P-Q\|_\infty \triangleq \sup_{a\in A} \log\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right)$$

A mapping M satisfies ϵ-differential privacy iff it satisfies the Lipschitz property!

Summary: Individual Fairness

- Formalized fairness property based on treating similar individuals similarly
  - Incorporated vendor's utility
- Explored relationship between individual fairness and group fairness
  - Earthmover distance

Individual fairness formulation:

Maximize utility

$$\max_{M_x} \mathbb{E}_{x\sim V}\mathbb{E}_{o\sim M_x}\left[U(x,o)\right]$$

Subject to fairness constraint

$$\text{s.t. } \left\|M_x - M_y\right\| \leq d(x,y)\ \ \forall\, x,y \in V$$

Q: What are the downsides to this formulation?

- Need a similarity metric between users
- Very high-dimensional LP may be difficult to solve
- Classifier must be trained *a priori* with fairness

Lots of open problems/direction

- **Metric**
  - Social aspects, who will define them?
  - How to generate metric (semi-)automatically?
- **Earthmover characterization** when probability metric is not statistical distance

- Next: How can we compute a fair classifier from an already-computed unfair one?

More definitions of fair classifiers

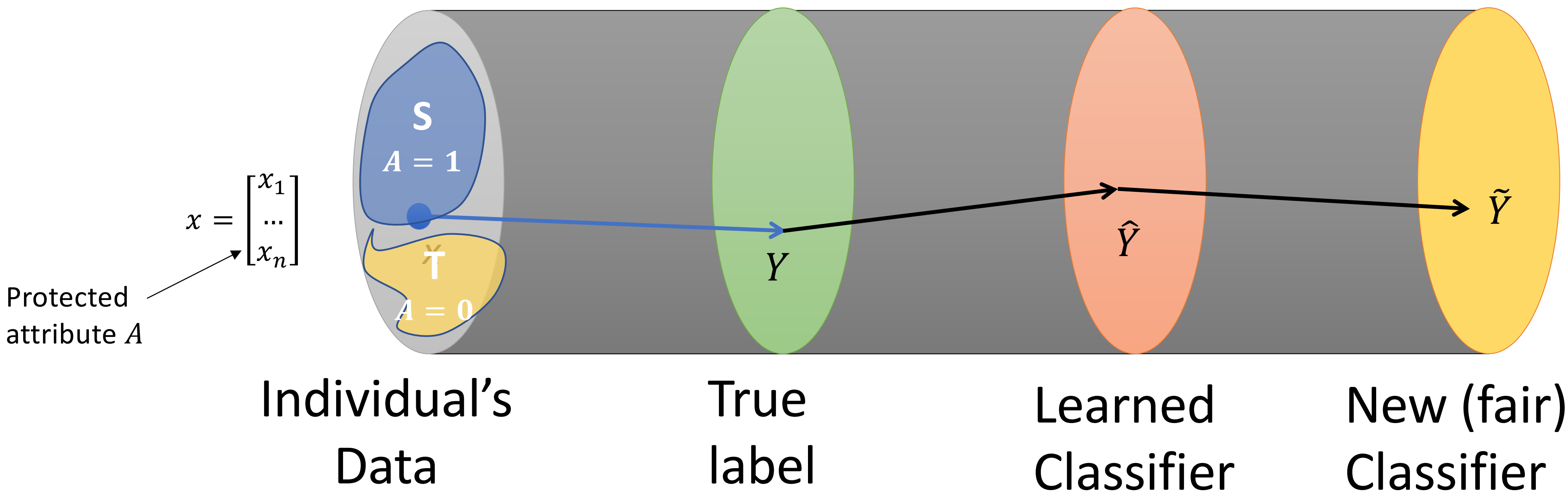
- NeurIPS 2016

Equality of Opportunity in Supervised Learning

Moritz Hardt
Google
m@mrtz.org

Eric Price*
UT Austin
ecprice@cs.utexas.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu



S
$A = 1$
T
$A = 0$
$x = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$
Protected attribute A
Y
$\hat{Y}$
$\tilde{Y}$
Individual's Data
True label
Learned Classifier
New (fair) Classifier

Equalized odds

- Consider binary classifiers
- We say a classifier $\hat{Y}$ has equalized odds if for all true labels y,

$$P\left[\hat{Y} = 1 | A = 0, Y = y\right] = P\left[\hat{Y} = 1 | A = 1, Y = y\right]$$

Q: How would this definition look if we only wanted to enforce group fairness?

A: $P\left[\hat{Y} = 1 | A = 0\right] = P\left[\hat{Y} = 1 | A = 1\right]$

Equal opportunity

- Suppose $Y = 1$ is the desirable outcome
  - E.g., getting a loan
- We say a classifier $\hat{Y}$ has equal opportunity if

$$P\left[\hat{Y} = 1 | A = 0, Y = 1\right] = P\left[\hat{Y} = 1 | A = 1, Y = 1\right]$$

Interpretation: True positive rate is the same for both classes

Weaker notion of fairness → can enable better utility

How can we create a predictor that meets these definitions?

- Key property: Should be oblivious

- A property of predictor $\hat{Y}$ is oblivious if it only depends on the joint distribution of $(Y, A, \hat{Y})$

- What does this mean?
- It does not depend on training data X

Need 4 parameters to define $\tilde{Y}$ from $(\hat{Y}, A)$

Protected attribute A

Predicted Label $\hat{Y}$

	0	1
0	$p_{00} = P(\tilde{Y} = 1 \mid A = 0, \hat{Y} = 0)$	$p_{01} = P(\tilde{Y} = 1 \mid A = 1, \hat{Y} = 0)$
1	$p_{10} = P(\tilde{Y} = 1 \mid A = 0, \hat{Y} = 1)$	$p_{11} = P(\tilde{Y} = 1 \mid A = 1, \hat{Y} = 1)$

Once our p_{ii}'s are defined…

- How do we check that equalized odds are satisfied?

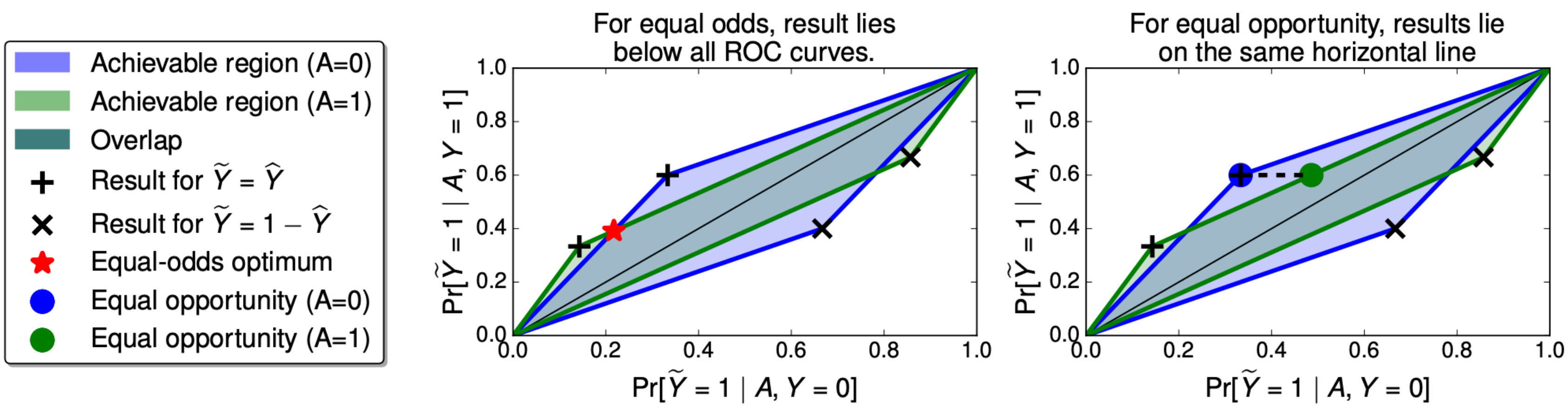
$$\gamma_a\big(\tilde{Y}\big) \triangleq \big(P\{\tilde{Y} = 1|A = a, Y = 0\}, P\{\tilde{Y} = 1|A = a, Y = 1\}\big)$$

Compute $\gamma_1(\tilde{Y})$ and $\gamma_0(\tilde{Y})$. (Depends on joint distribution of $\big(Y, A, \hat{Y}\big)$)
They should be equal (to satisfy equalized odds)

Q: What condition do we need for an equal opportunity classifier?

A: The 2nd entries of $\gamma_1(\tilde{Y})$ and $\gamma_0\big(\tilde{Y}\big)$ should match

Geometric Interpretation via ROC curves



Write equalized odds as an optimization

- Let's define a loss function $\ell(\tilde{Y}_p, Y)$ describing loss of utility from using $\tilde{Y}_p$ instead of Y

- Now we can optimize:

$$\begin{aligned} \min_{p} \quad & \mathbb{E}\ell(\widetilde{Y}_p, Y) \\ \text{s.t.} \quad & \gamma_0(\widetilde{Y}_p) = \gamma_1(\widetilde{Y}_p) \\ & \forall_{y,a} 0 \leqslant p_{ya} \leqslant 1 \end{aligned}$$

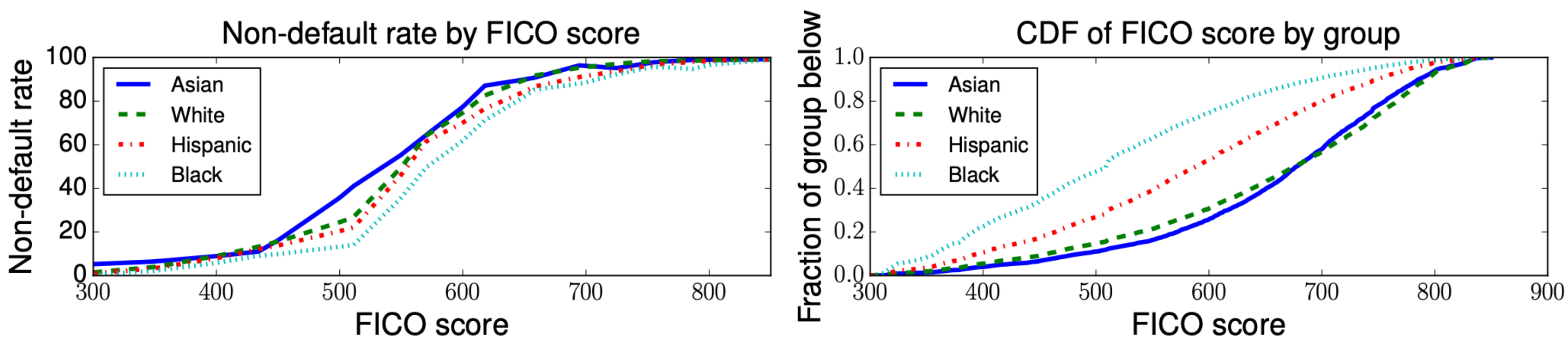
- Objective and constraints are both linear in vector of p values!

What about continuous values?

- E.g., suppose we use a numeric credit score R to predict binary value Y
- You can threshold the score to get a comparable definition of equalized odds
- If R satisfies equalized odds, then so does any predictor $\hat{Y} = I\{R > t\}$, where t is some threshold

Case study: FICO Scores

- Credit scores R range from 300 to 850
- Binary variable Y = whether someone will default on loan



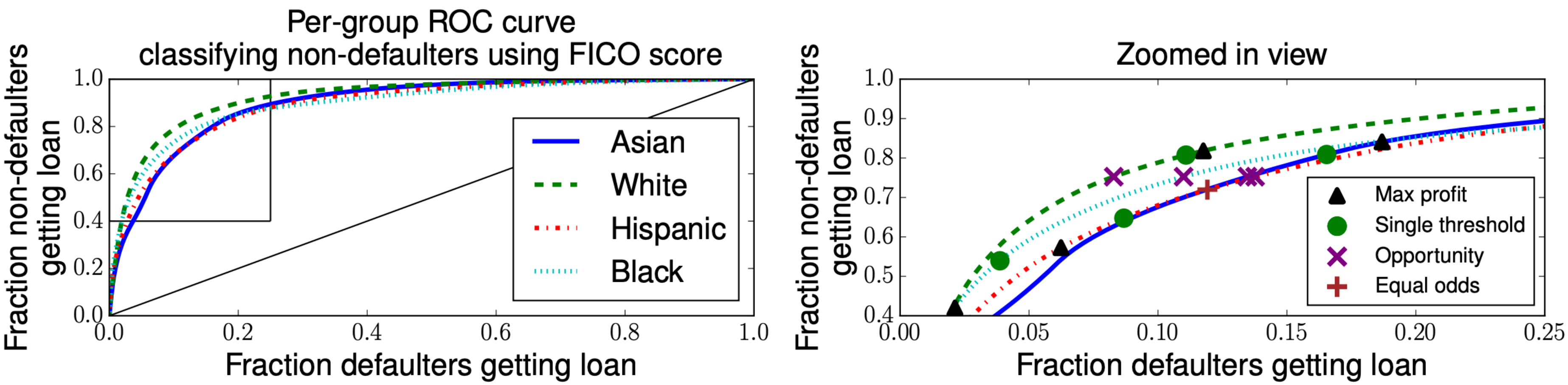
Experiment

- False positive – giving a loan to someone who will default
- False negative – not giving a loan to someone who will not default
- Loss function = false positives are 4.5x as expensive as false negatives

Baselines

- **Max profit** – no fairness constraint
- **Race blind** – uses same FICO threshold for all groups
- **Group fairness** – picks for each group a threshold such that the fraction of group members that qualify for loans is the same
- **Equal opportunity** – picks a threshold for each group s.t. fraction of non-defaulting group members is the same
- **Equalized odds** – requires both the fraction of non-defaulters that qualify for loans and the fraction of defaulters that qualify for loans to be constant across groups

ROC Curve Results



Profit Results

Fair by some definition

Method	Profit (% relative to max profit)
Max profit	100
Race blind	99.3
Equal opportunity	92.8
Equalized odds	80.2
Group fairness (demographic parity)	69.8