CY 7790

Special Topics in Security and Privacy: Machine Learning Security and Privacy Fall 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science

November 29 2021

CY 7790

Special Topics in Security and Privacy:
Machine Learning Security and Privacy

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

By: Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
(IW3C2 2017)

Presented by Pablo Kvitca, for CY7790 - Fall 2021

November 29, 2021

Problem Statement

- Introduce a new notion of unfairness: **disparate mistreatment**

- Show how to apply it to decision-boundary based classifiers, with convex loss

Model

Objectives	**Knowledge**
Fairness on Classification Models	Predicted label class
(with respect to disparate mistreatment)	Ground-truth label class
	Sensitive Attribute labels

Contributions

- Disparate Mistreatment definition and explanation
- Application of Disparate Mistreatment on decision boundary-based classifiers
  - Such as logistic regression
  - Through Disciplined Convex-Concave Program (DCCP)
- Notion on avoiding disparate mistreatment
- Satisfying multiple fairness notions simultaneously

Disparate Treatment

- Intuitive notion of fairness (Two similar people should not be treated differently because the belong to different protected groups)
- Arises when a system provides ***different outputs*** for groups of people with the **same** (similar) values for **non-sensitive** attributes but **different** values for **sensitive** attributes

User Attributes		
Sensitive	**Non-sensitive**	
Gender	**Clothing Bulge**	**Prox. Crime**
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

Disparate Impact

- Arises when a system provides outputs that **benefit/hurt** a group of people sharing the **same sensitive** attributes, more frequently than other groups
- Independent of the ground truth for the label
  - If available, can be misleading
  - If used, required decision outcomes to be proportional: risks introducing *reverse-discrimination*

User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

C1: Stopped rate: 1.00 for male vs. 0.66 for female

Disparate Mistreatment (new)

- Needs the system to not have perfectly accurate predictions
- Misclassification rates may be different for groups of people with **different sensitive** attributes
  - Misclassification rate can vary: overall, false negative/positive, omission. discovery, ...

User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

C2: Misclassification (FPR) rate: 0.0 (M) vs. 1.0 (F)
C2: Misclassification (FNR) rate: 0.0 (M) vs. 0.5 (F)

C1: Misclassification (FNR) rate: 0.0 (M) vs. 0.5 (F)

When to apply Disparate Mistreatment

Measure	Ground Truth	Risk Reverse Discrimination
Disparate Treatment	INDEPENDENT	NO
Disparate Impact	INDEPENDENT	YES
Disparate Mistreatment	DEPENDENT	NO

Avoiding Disparate Treatment

- A classifier does **not** "suffer" from *disparate treatment* if:
  - Sensitive attribute: *z*

$$P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x})$$

Avoiding Disparate Impact

- A classifier does **not** "suffer" from *disparate impact* if:
  - Sensitive attribute: *z*

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

Avoiding Disparate Mistreatment

- A classifier does **not** “suffer” from *disparate mistreatment* if:

“**<u>Misclassification rates</u>** for *different groups* of people with *different* values of *sensitive attribute z* are the *same*”

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Misclassification Rates (shown)

- Overall Misclassification Rate (OMR):

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

- False Positive Rate (FPR):

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

- False Negative Rate (FNR):

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

Misclassification Rates (for future work)

- False Omission Rate (FOR):

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

- False Discovery Rate (FDR):

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

Training Classifiers without Disparate Mistreatment

- Can train a decision boundary-based classifier so that it does not suffer from *disparate mistreatment*
- Given a **convex** loss $L(\theta)$
  - ensure global optimum can be found efficiently
- Solve:

$$\begin{aligned}
\text{minimize} \quad & L(\boldsymbol{\theta}) \\
\text{subject to} \quad & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\
& P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon,
\end{aligned}$$

Training Classifiers without Disparate Mistreatment (cont)

- The disparate mistreatment can be measured using **covariance** between
  - Sensitive attributes
  - Signed distance between the features of misclassified samples and the decision boundary
- NOTE: covariance is actually computed by Monte-Carlo covariance

$$\begin{aligned} \mathrm{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \cancel{\mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))]} \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}), \end{aligned}$$

$$\mathbb{E}[(z - \bar{z})] = 0$$

$$\begin{aligned} g_{\boldsymbol{\theta}}(y, \mathbf{x}) &= min(0, y d_{\boldsymbol{\theta}}(\mathbf{x})), \\ g_{\boldsymbol{\theta}}(y, \mathbf{x}) &= min\left(0, \frac{1-y}{2} y d_{\boldsymbol{\theta}}(\mathbf{x})\right), \text{ or} \\ g_{\boldsymbol{\theta}}(y, \mathbf{x}) &= min\left(0, \frac{1+y}{2} y d_{\boldsymbol{\theta}}(\mathbf{x})\right), \end{aligned}$$

Training Classifiers without Disparate Mistreatment (cont)

- Resulting in:
  - Which... is not convex

$$
\begin{aligned}
\text{minimize} \quad & L(\boldsymbol{\theta}) \\
\text{subject to} \quad & \tfrac{1}{N} \textstyle\sum_{(\mathbf{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c, \\
& \tfrac{1}{N} \textstyle\sum_{(\mathbf{x},y,z)\in\mathcal{D}} (z-\bar{z})\, g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c,
\end{aligned}
$$

Aside: Disciplined Convex-Concave Program

- Disciplined Convex-Concave Program (DCCP) - 2016
  - https://arxiv.org/abs/1604.02639
- “Combines the ideas of disciplined convex programming (DCP) with convex-concave programming (CCP)”
  - “CCP is an organized heuristic for solving non-convex problems that involve objective and constraint functions that are a sum of a convex and a concave term”
  - “DCP is a structured way to define convex optimization problems, based on a family of basic convex and concave functions and a few rules for combining them.”
  - “Problems expressed using DCP can be automatically converted to standard form and solved by a generic solver”

[Shen et al]

Training Classifiers without Disparate Mistreatment (cont)

- Convert to a Disciplined Convex-Concave Program (DCCP)

$$
\begin{array}{ll}
\text{minimize} & L(\boldsymbol{\theta}) \\
\text{subject to} & \frac{-N_1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\
& + \frac{N_0}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \\
& \frac{-N_1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\
& + \frac{N_0}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c,
\end{array}
$$

$$
\frac{-N_1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) + \frac{N_0}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \sim c
$$

Logistic Regression without Disparate Mistreatment

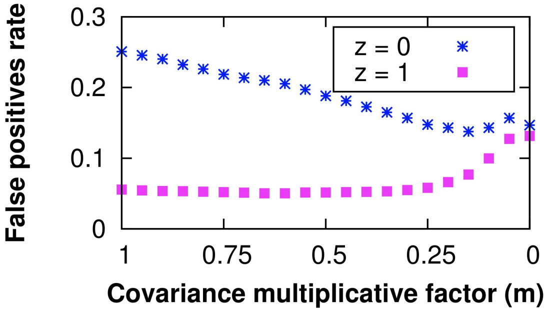
$$\begin{array}{ll} \text{minimize} & -\sum_{(\mathbf{x},y)\in\mathcal{D}} \log p(y_i|\mathbf{x}_i,\boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \le c \\ & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \ge -c \end{array}$$

Observations - Synthetic Data

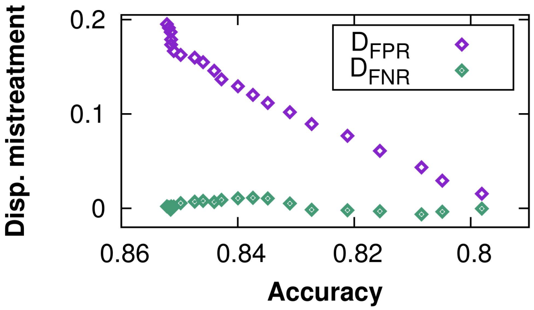
- Synthetic Data: only False Positive Rate
- Synthetic Data: only False Negative Rate
- Methods:
  - Train unconstrained classifier
  - Train constrained by FPR
  - Train constrained by FNR

Observations:

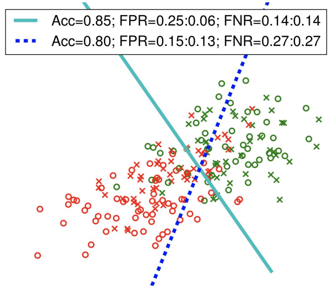
- As the fairness constraint value goes to zero, FPR for both groups converge
  - (ie. more fair)
- Causes a drop in accuracy



(a) Cov. vs FPR



(b) Fairness vs. Acc.



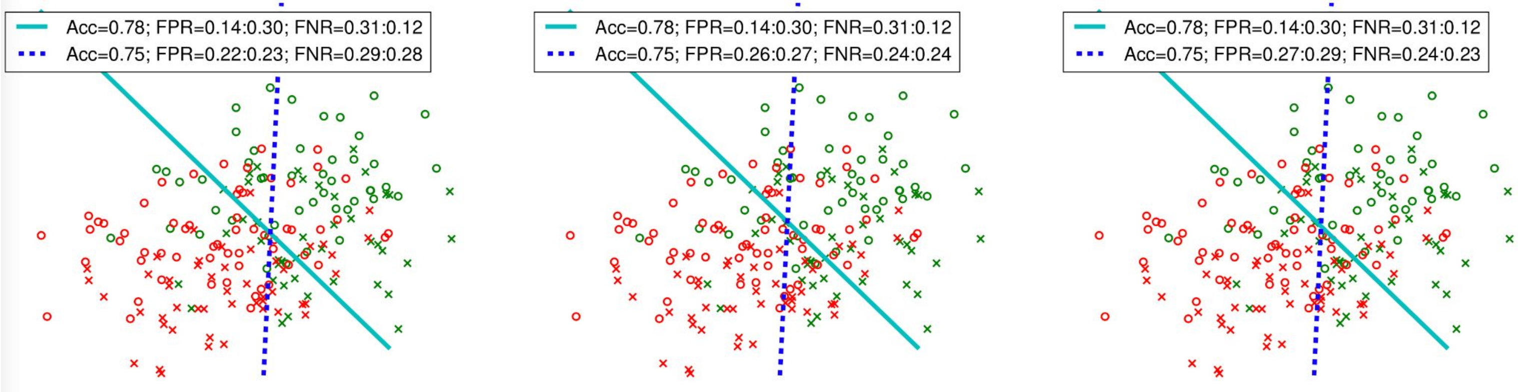
(c) Boundaries

Observations - Synthetic Data

- Synthetic Data: both FPR and FNR
  - **Case I**: have opposite signs
    - Train unconstrained
    - Train constrained by FPR, by FNR, and both
  - Case II: have same sign
    - Train unconstrained
    - Train constrained by FPR, by FNR, and both

Observations:

- Removing disparate mistreatment for just FPR causes rotation of the decision boundary
  - Decreases FPR, Increases FNR
- So controlling for FPR also removes FNR mistreatment
- Similar on FNR case, similar results to both
- (this is due to the data distribution, not always)



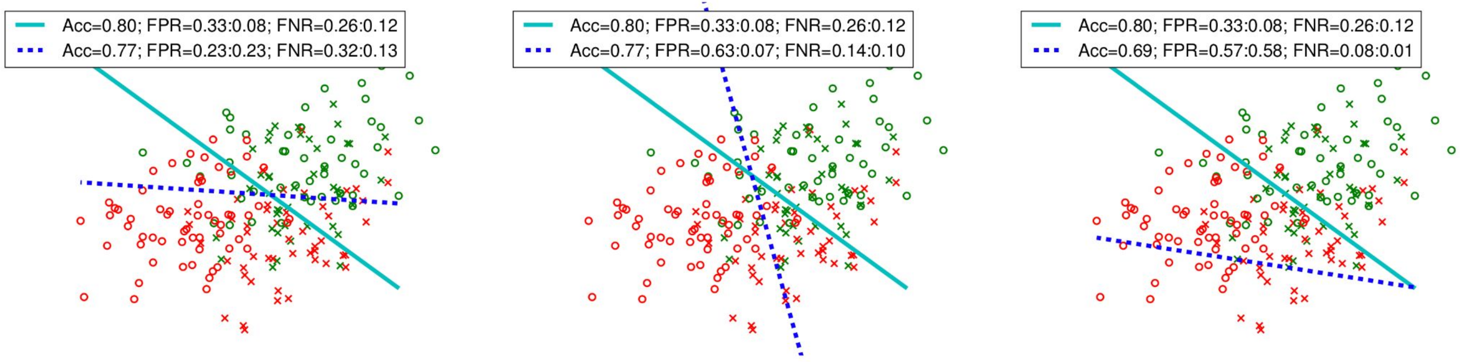
(a) FPR constraints (b) FNR constraints (c) Both constraints

Observations - Synthetic Data

- Synthetic Data: both FPR and FNR
  - Case I: have opposite signs
    - Train unconstrained
    - Train constrained by FPR, by FNR, and both
  - **Case II**: have same sign
    - Train unconstrained
    - Train constrained by FPR, by FNR, and both

Observations:

- Controlling for only FPR leads to drop in accuracy, can exacerbate FNR
- Controlling for only FNR leads to drop in accuracy, can exacerbate FPR
- Controlling for both: FPR/FNR go to zero, but accuracy drops



(a) FPR constraints (b) FNR constraints (c) Both constraints

Evaluation - Performance to other methods*

- This method (through DCCP for avoiding disparate mistreatment)
  - Sensitive features are used a learnable features
- Hardt et al.
  - Adds post-processing on an unfair classifier to decide on the right threshold for each group so that it is fair.
  - Needs sensitive attribute information.
  - Cannot avoid disparate treatment
- Baseline
  - Tries to remove disparate mistreatment by introduction penalties for misclassied data points with different sensitive attribute values, during training
    - First, trains an unfair classifier
    - Select set of misclassificated points for a sensitive attribute with a higher error rate
    - Iteratively re-trains with increasingly higher penalties on this set until a given unfairness level is met

Evaluation - Performance to other methods*

		FPR constraints			FNR constraints			Both constraints		
		Acc.	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$	**Acc.**	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$	**Acc.**	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$
Synthetic setting 1 (Figure 2)	**Our method**	0.80	0.02	0.00	–	–	–	–	–	–
	Our method$_{\text{sen}}$	0.85	0.00	0.25	–	–	–	0.83	0.07	0.01
	Baseline	0.65	0.00	0.00	–	–	–	–	–	–
	Hardt et al.	0.85	0.00	0.21	–	–	–	0.80	0.00	0.02
Synthetic setting 2 (Figure 3)	**Our method**	0.75	−0.01	0.01	0.75	−0.01	0.01	0.75	−0.01	0.01
	Our method$_{\text{sen}}$	0.80	0.00	0.03	0.80	0.02	0.01	0.80	0.01	0.02
	Baseline	0.59	−0.01	0.15	0.59	−0.15	0.01	0.76	−0.04	0.03
	Hardt et al.	0.80	0.00	0.03	0.80	0.03	0.00	0.79	0.00	−0.01
Synthetic setting 3 (Figure 4)	**Our method**	0.77	0.00	0.19	0.77	0.55	0.04	0.69	−0.01	0.06
	Our method$_{\text{sen}}$	0.78	0.00	0.42	0.79	0.38	0.03	0.77	0.14	0.06
	Baseline	0.57	0.01	0.09	0.67	0.44	0.01	0.38	−0.43	0.01
	Hardt et al.	0.78	0.01	0.44	0.79	0.41	0.02	0.67	0.02	0.00
ProPuclica COMPAS (Section 5.2)	**Our method$_{\text{sen}}$**	0.660	0.06	−0.14	0.662	0.03	−0.10	0.661	0.03	−0.11
	Baseline	0.643	0.03	−0.11	0.660	0.00	−0.07	0.660	0.01	−0.09
	Hardt et al.	0.659	0.02	−0.08	0.653	−0.06	−0.01	0.645	−0.01	−0.01

Observations - Real Data

- From ProPublica COMPAS dataset
  - (simplified to subset with black/white races)
- Method:
  - Train unconstrained
  - Train constrained by FPR, by FNR, by both

Observations:

- Similar to synthetic data:
  - Constraint on FPR reduces FNR
  - Constraint on FNR reduces FPR
- All 3 methods achieve similar accuracy for similar levels of fairness
- Does not completely remove disparate mistreatment (probably due to small dataset)
  - Hardt does but low accuracy

Strengths

- Intuitive and easily checkable
- Works with respect to misclassification rates
    - Possibly more than one!
- Low degradation of accuracy
- Can be used simultaneously with *disparate treatment* constraints

Limitations

- Requires ground truth knowledge about labels.
- Requires demographic information (mention this might not be necessary)
- Needs to be shown for FDR and FOR
- Requires the Disciplined Convex-Concave Program (DCCP)
  - Can be complicated to convert a problem to DCCP
  - Work well in practice, do not have a **guarantee** of finding a global optimum
- Analytical covariance was done through Monte Carlo covariance
  - This is an approximation, inaccurate on smaller dataset
- Applies only to decision boundary-based classifiers
- Applies only if the loss is convex, to be solved as DCCP

Discussion

- The *disparate mistreatment* measurement can be applied to any classifier
    - Though it is only defined for the binary case, it is not clear how it extends to multi-class
- The “baseline” method described on the paper might be applicable to any classifier, but has higher complexity

Related Work - DCCP

Disciplined Convex-Concave Programming

Xinyue Shen Steven Diamond Yuantao Gu Stephen Boyd

April 12, 2016

Abstract

In this paper we introduce *disciplined convex-concave programming* (DCCP), which combines the ideas of disciplined convex programming (DCP) with convex-concave programming (CCP). Convex-concave programming is an organized heuristic for solving nonconvex problems that involve objective and constraint functions that are a sum of a convex and a concave term. DCP is a structured way to define convex optimization problems, based on a family of basic convex and concave functions and a few rules for combining them. Problems expressed using DCP can be automatically converted to standard form and solved by a generic solver; widely used implementations include `YALMIP`, `CVX`, `CVXPY`, and `Convex.jl`. In this paper we propose a framework that combines the two ideas, and includes two improvements over previously published work on convex-concave programming, specifically the handling of domains of the functions, and the issue of nondifferentiability on the boundary of the domains. We describe a Python implementation called `DCCP`, which extends `CVXPY`, and give examples.

Related Work - Fairness

On the Applicability of Machine Learning Fairness Notions

Karima Makhlouf
Université du Québec à Montréal
Montréal, Canada
makhlouf.karima@courrier.uqam.ca

Sami Zhioua
Higher Colleges of Technology
Dubai, UAE
szhioua@hct.ac.ae

Catuscia Palamidessi
INRIA, École Polytechnique, IPP
Paris, France
catuscia@lix.polytechnique.fr

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

Pieter Delobelle
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
pieter.delobelle@kuleuven.be

Paul Temple,
Gilles Perrouin,
Benoit Frénay,
Patrick Heymans
PReCISE, NaDi, University of Namur
Namur, Belgium
firstname.lastname@unamur.be

Bettina Berendt
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
Faculty of Electrical Engineering and Computer Science, TU Berlin
Berlin, Germany
bettina.berendt@kuleuven.be

Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms

Gareth P. Jones
Experian DataLabs UK&I and EMEA
London, United Kingdom

James M. Hickey*
James.Hickey@experian.com
Experian DataLabs UK&I and EMEA
London, United Kingdom

Pietro G. Di Stefano
Experian DataLabs UK&I and EMEA
London, United Kingdom

Charanpal Dhanjal
Experian DataLabs UK&I and EMEA
London, United Kingdom

Laura C. Stoddart
Experian DataLabs UK&I and EMEA
London, United Kingdom

Vlasios Vasileiou
Experian DataLabs UK&I and EMEA
London, United Kingdom

https://github.com/Trusted-AI/AIF360

Related Work - Fairness - AI Fairness 360

- Same group as: Adversarial Robustness Toolbox (ART)
  - https://github.com/Trusted-AI/adversarial-robustness-toolbox
- AI Fairness 360:
  - https://github.com/Trusted-AI/AIF360
- They also have AI Explainability 360 (AIX360)
- (From IBM Research)

References

All images on this presentation are extracted from original paper:

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi Max Planck Institute for Software Systems (MPI-SWS) (2017). International World Wide Web Conference Committee (IW3C2), c published under Creative Commons CC BY 4.0 License. WWW 2017, April 3–7, 2017, Perth, Australia. ACM 978-1-4503-4913-0/17/04. http://dx.doi.org/10.1145/3038912.3052660

EqualityOf Opportunity In Supervised Learning

Moritz Hardt, Eric Price, Nathan Srebro

Presented by Zohair Shafi - CY 7790 | 29th November 2021

“Types” Of Fairness?

Parities

$\hat{Y}$: Classifier
Y : Labels
A : Protected Attribute

Demographic Parity :

Same accuracy

$$Pr(\hat{Y} = 1 \,|\, A = 1) = Pr(\hat{Y} = 1 \,|\, A = 0)$$

True Positive Parity :

$$Pr(\hat{Y} = 1 \,|\, Y = 1,\, A = 1) = Pr(\hat{Y} = 1 \,|\, Y = 1,\, A = 0)$$

Equal opportunity

False Positive Parity :

Equalized Odds

$$Pr(\hat{Y} = 1 \,|\, Y = 0,\, A = 1) = Pr(\hat{Y} = 1 \,|\, Y = 0,\, A = 0)$$

“Types” Of Fairness?

Parities

$\hat{Y}$: Classifier
Y : Labels
A : Protected Attribute

Demographic Parity :

$$Pr(\hat{Y} = 1 \,|\, A = 1) = Pr(\hat{Y} = 1 \,|\, A = 0)$$

- Not very nice - does not ensure fairness.
- Can accept random individuals in a demographic so long as percentages of acceptance match - can arise when low training data is available for a demographic

“Types” Of Fairness?

Parities

$\hat{Y}$: Classifier
Y : Labels
A : Protected Attribute

True Positive Parity :

$$Pr(\hat{Y} = 1 \mid Y = 1, A = 1) = Pr(\hat{Y} = 1 \mid Y = 1, A = 0)$$

False Positive Parity :

$$Pr(\hat{Y} = 1 \mid Y = 0, A = 1) = Pr(\hat{Y} = 1 \mid Y = 0, A = 0)$$

- Nice
- Equal bias and equal accuracy in all demographics

Achieving Fairness

Binary Classifier

- Post learning
- Derived Predictor : Predictor $\tilde{Y}$ is a derived from a random variable $\hat{Y}$ and protected attribute A if it is a possibly randomized function of the random variables $(\hat{Y}, A)$ alone - i.e., $\tilde{Y}$ is independent of X conditional on $(\hat{Y}, A)$
- Geometric solution - intersecting ROC curves
  $\gamma_a(\hat{Y}) = (FPR_a, TPR_a)$
- Consider a 2D convex polytope :
  $P_a(\hat{Y}) = convhull\ \{(0,0), \gamma_a(\hat{Y}), \gamma_a(1 - \hat{Y}), (1,1)\}$
- Predictor $\tilde{Y}$ is "derived" if and only if
  $\gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \quad \forall a \in [0,1]$
- Equalized Odds : $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$
  Equal Opportunity : $\gamma_0(\hat{Y})_2 = \gamma_1(\hat{Y})_2$, i.e., $TPR_0 = TPR_1$

Achieving Fairness

Binary Classifier

$$\min_{\widetilde{Y}} \quad \mathbb{E}\ell(\widetilde{Y}, Y)$$

$$\text{s.t.} \quad \forall a \in \{0,1\} : \gamma_a(\widetilde{Y}) \in P_a(\widetilde{Y})$$

$$\gamma_0(\widetilde{Y}) = \gamma_1(\widetilde{Y})$$

"Derived" Constraint

Equalized Odds

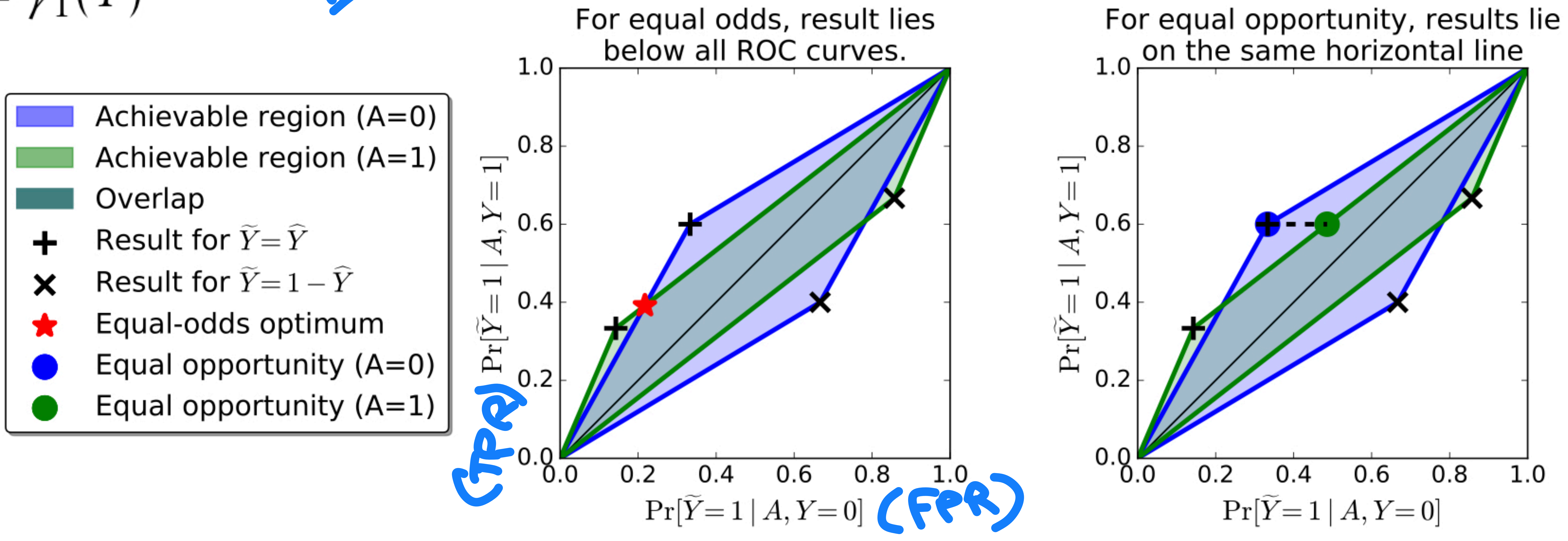


Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

Achieving Fairness

Real Valued Score Function

- For a real valued score R, choose a threshold t such that :
  $\hat{Y} = I\{R > t\}$
- Optimal threshold should be chosen to balance FPR and TPR to minimize expected loss.
- Can have multiple thresholds : $t_a \ \forall a \in A$
- $C_a(t) = (Pr(\hat{R} > t \mid Y = 0,\ A = a),\ Pr(\hat{R} > t \mid Y = 1,\ A = a))$
  $C_a(t) = (FPR_{a,t}, TPR_{a,t})$
- Equalized Odds : $C_a(t) = C_{a'}(t) \quad \forall a, a' \in A,\ t \in [0,1]$

Achieving Fairness

Real Valued Score Function

- ROC curves might not intersect except trivially at (0, 0) and (1, 1)
- Use randomization to fill the span of possible derived predictors and allow for more intersection (not too sure what this means)
- $D_a = convhull\ \{C_a(t) : t \in [0,1]\}$
- Only consider points above the diagonal - better than random guessing

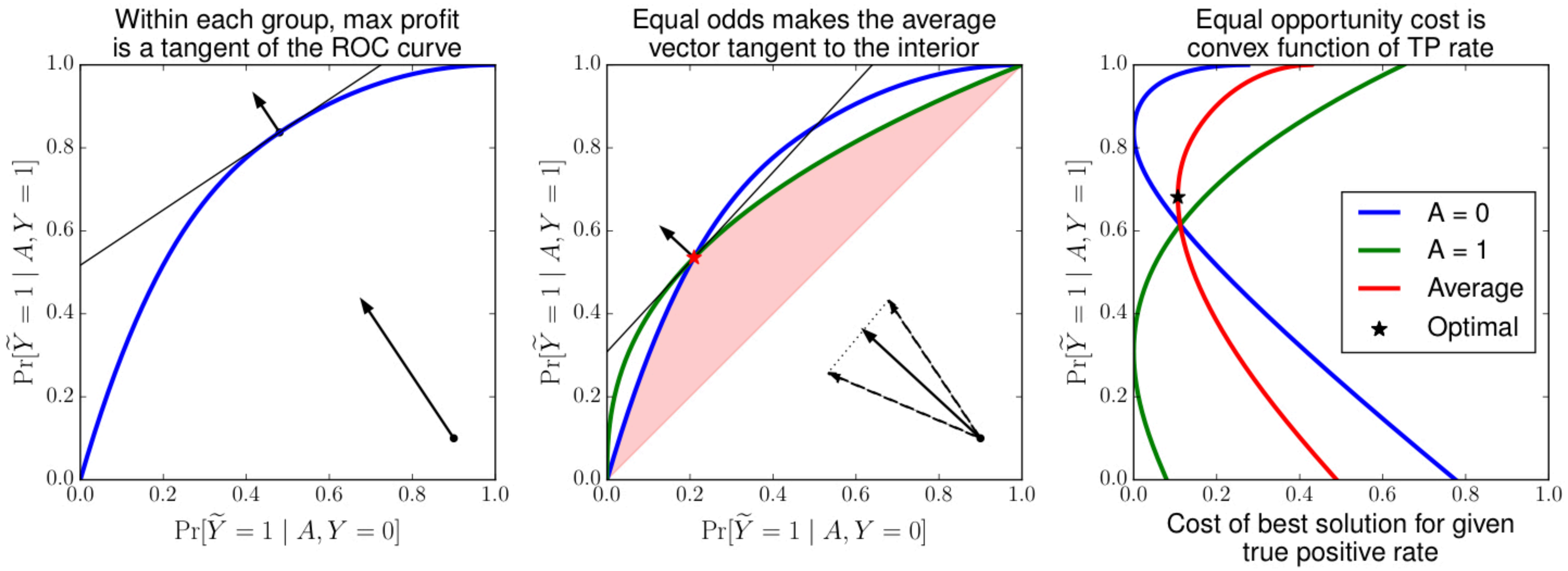


Figure 2: Finding the optimal equalized odds threshold predictor (middle), and equal opportunity threshold predictor (right). For the equal opportunity predictor, within each group the cost for a given true positive rate is proportional to the horizontal gap between the ROC curve and the profit-maximizing tangent line (i.e., the two curves on the left plot), so it is a convex function of the true positive rate (right). This lets us optimize it efficiently with ternary search.

Achieving Fairness

Optimal Equalized Odds Threshold Predictor

- Consider optimal predictor $\tilde{Y}$ as a mixture of two threshold predictors :
  $\tilde{Y} = I\{R > T_a\}$ where

  $T_a = \underline{t}_a$ with probability $\underline{p}_a$
  $T_a = \bar{t}_a$ with probability $\bar{p}_a$

- For each group "*a*", use :

  - Fixed threshold $T_a = t_a$

  - Mixture of two thresholds $\underline{t}_a < \bar{t}_a$:

    $R < \underline{t}_a \implies \tilde{Y} = 0$
    $R > \bar{t}_a \implies \tilde{Y} = 1$
    $\underline{t}_a < R < \bar{t}_a \implies \tilde{Y} = 1$ with probability $\underline{p}_a$

$$\min_{\forall a:\, \gamma \in D_a} \gamma_0 \overbrace{\ell(1,0)}^{\text{For any loss function}} + (1-\gamma_1)\ell(0,1)$$

Note: $\gamma \in D_a$

Achieving Fairness

Bayes Optimal Regressor

- Given random variables (X, A) and target variable Y, the Bayes Optimal Regressor is given by

  $R = argmin_{r(x,a)} \, \mathbb{E}[(Y - r(X, A))^2]$
  where
  $r(x, a) = \mathbb{E} \, [Y | X = x, \, A = a]$

- Bayes optimal classifier is hence a threshold predictor of R where the threshold depends on the loss function.

Achieving Fairness

Bayes Optimal Regressor - Non Discriminating Classifier

- Given a bounded loss function ℓ, a Bayes optimal regressor $R*$, there is an optimal equalized odds predictor $Y*$ and an equalized odds predictor $\hat{Y}$ derived from $(\hat{R}, A)$ such that :

$$\mathbb{E}\ell(\hat{Y}, Y) \leq \mathbb{E}\ell(Y*, Y) + 2\sqrt{2}\ d_K(\hat{R}, R*)$$

where $d_K(R, R')$ is the conditional Kolmogorov distance between two random variables defined as

$$d_{\mathrm{K}}(R, R') \stackrel{\mathrm{def}}{=} \max_{a,y \in \{0,1\}} \sup_{t \in [0,1]} \left|\Pr\{R > t \mid A = a, Y = y\} - \Pr\{R' > t \mid A = a, Y = y\}\right|$$

Results - FICO Scores

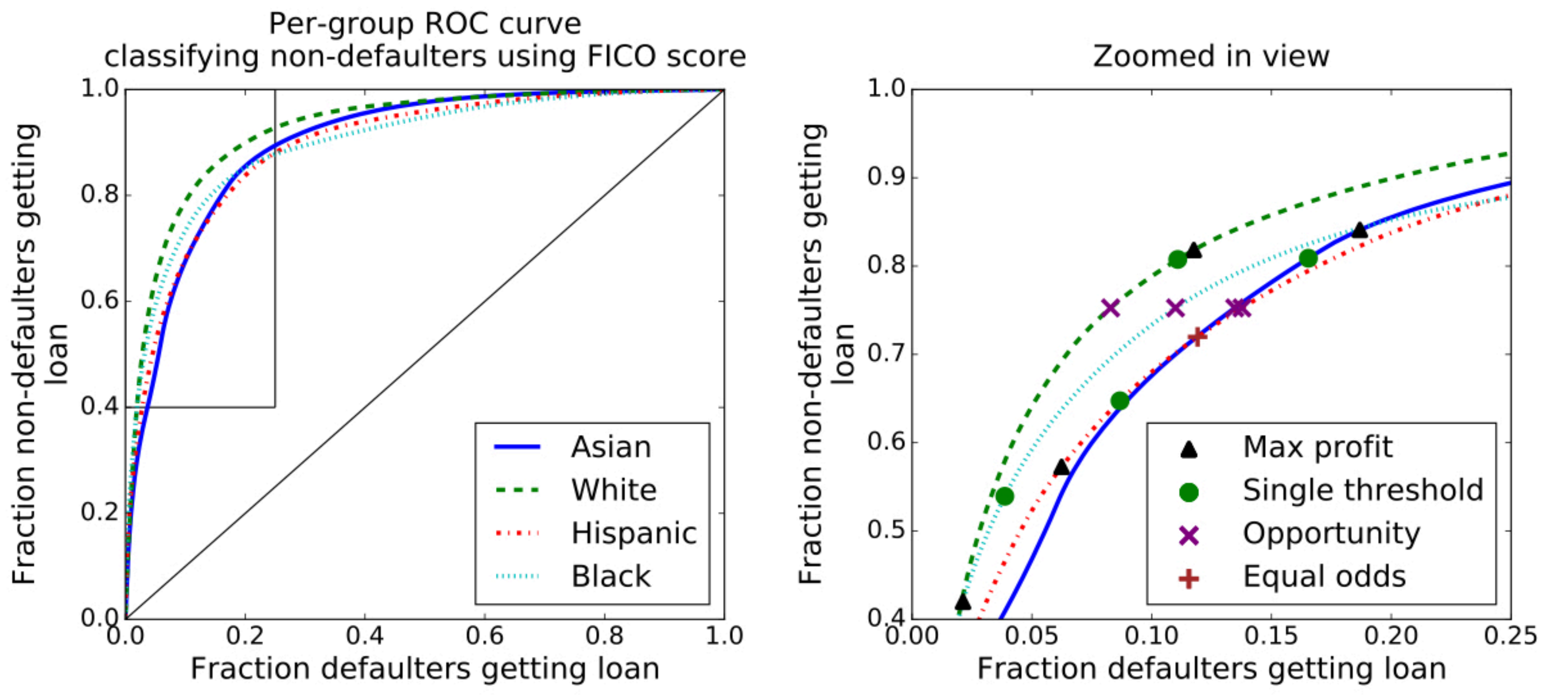


Figure 10: The ROC curve for using FICO score to identify non-defaulters. Within a group, we can achieve any convex combination of these outcomes. Equality of opportunity picks points along the same horizontal line. Equal odds picks a point below all lines.

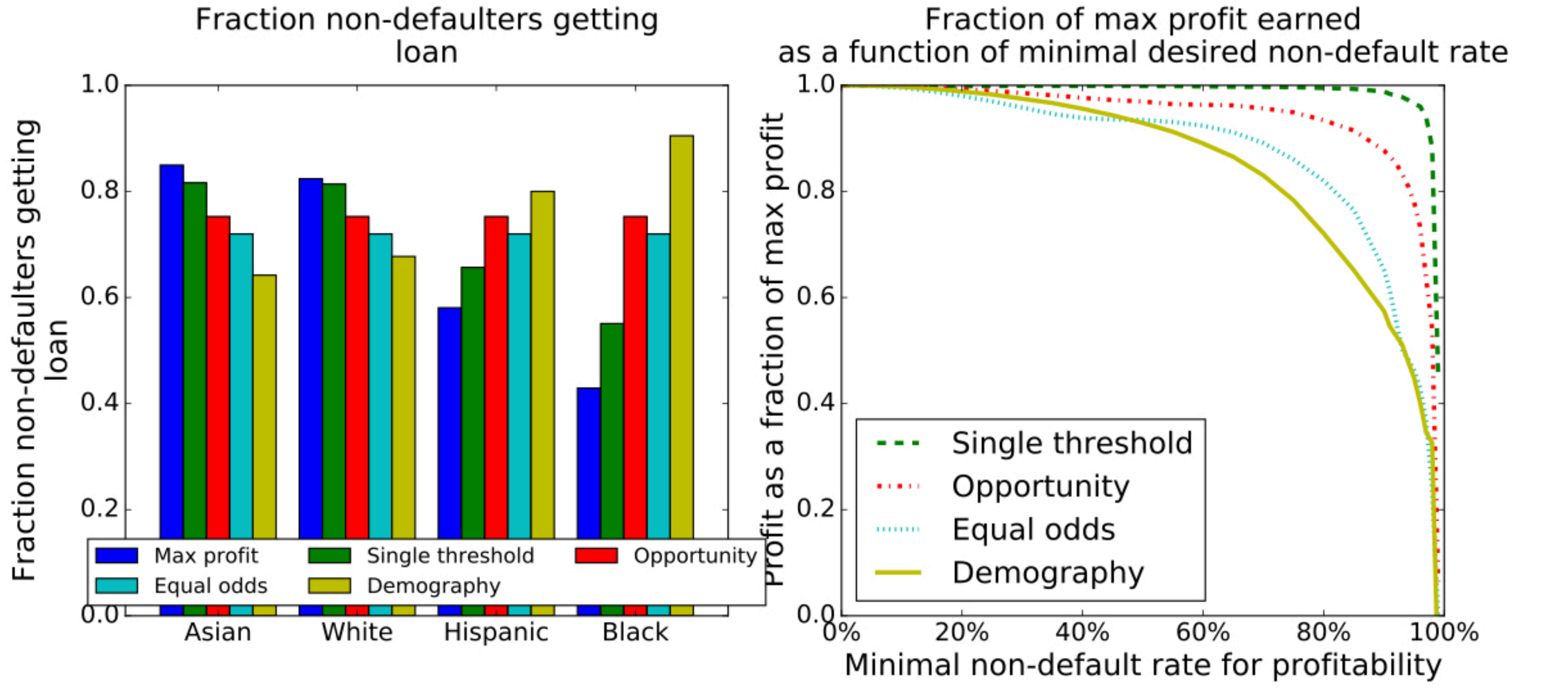


Figure 11: On the left, we see the fraction of non-defaulters that would get loans. On the right, we see the profit achievable for each notion of fairness, as a function of the false positive/negative trade-off.

- Protected Attribute : Race
  Asian
  Black
  White
  Hispanic

- Loss function where False Positives are 82/18 as expensive as False Negatives

- Five Constraints :

  Max Profit - No fairness constraint

  Race Blind - Requires same threshold for each group **(99.3% of max profit)**

  Demographic Parity - Picks a threshold for each group such that fraction of group members qualifying for loans is the same **(69.8% of max profit)**

  Equal Opportunity - Picks a threshold for each group such that fraction of non-defaulting members is the same across groups **(92.8% of max profit)**

  Equalized Odds - Requires fraction of non-defaulters and defaulters that qualify for loans to be the same across groups **(80.2% of max profit)**

Results - FICO Scores

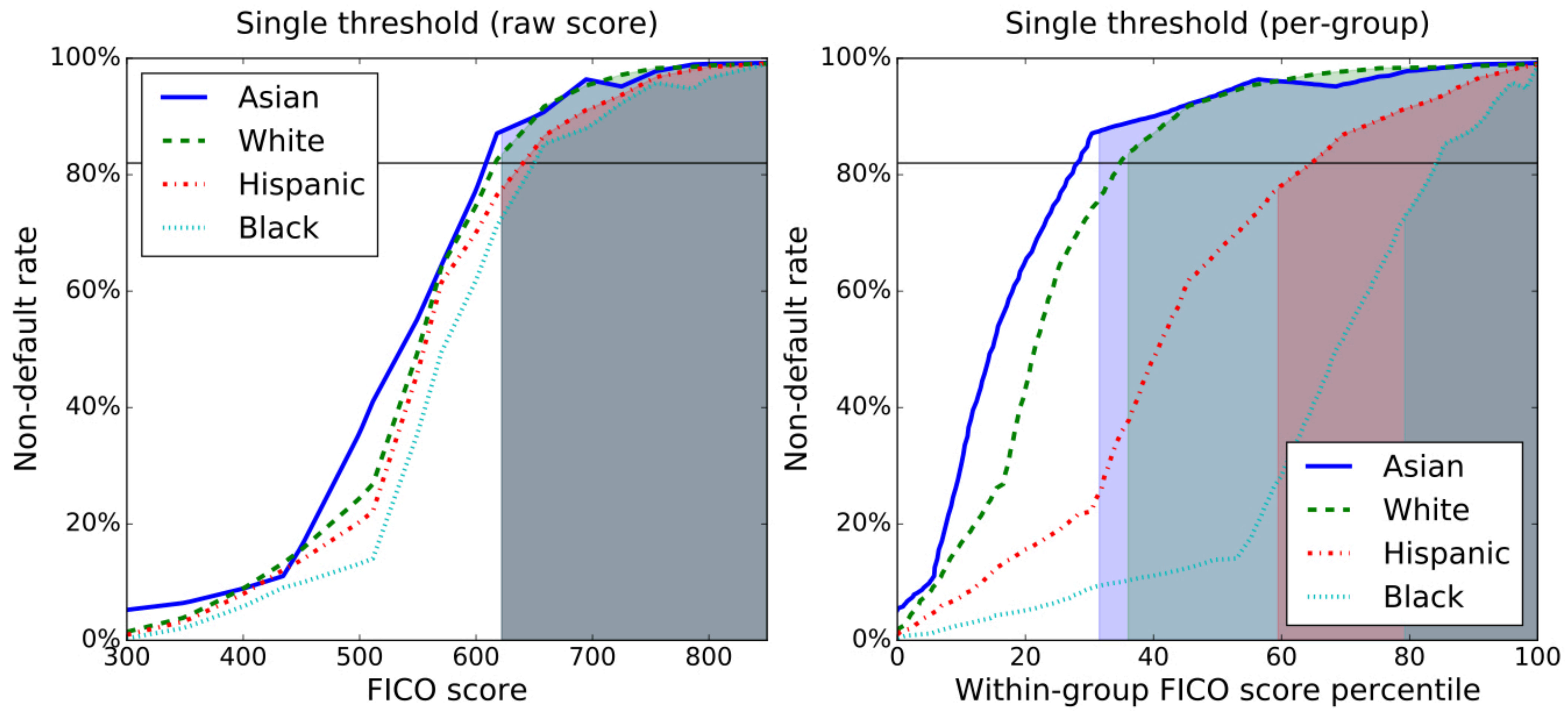


Figure 8: The common FICO threshold of 620 corresponds to a non-default rate of 82%. Rescaling the x axis to represent the within-group thresholds (right), $\Pr[\widehat{Y}=1 \mid Y=1, A]$ is the fraction of the area under the curve that is shaded. This means black non-defaulters are much less likely to qualify for loans than white or Asian ones, so a race blind score threshold violates our fairness definitions.

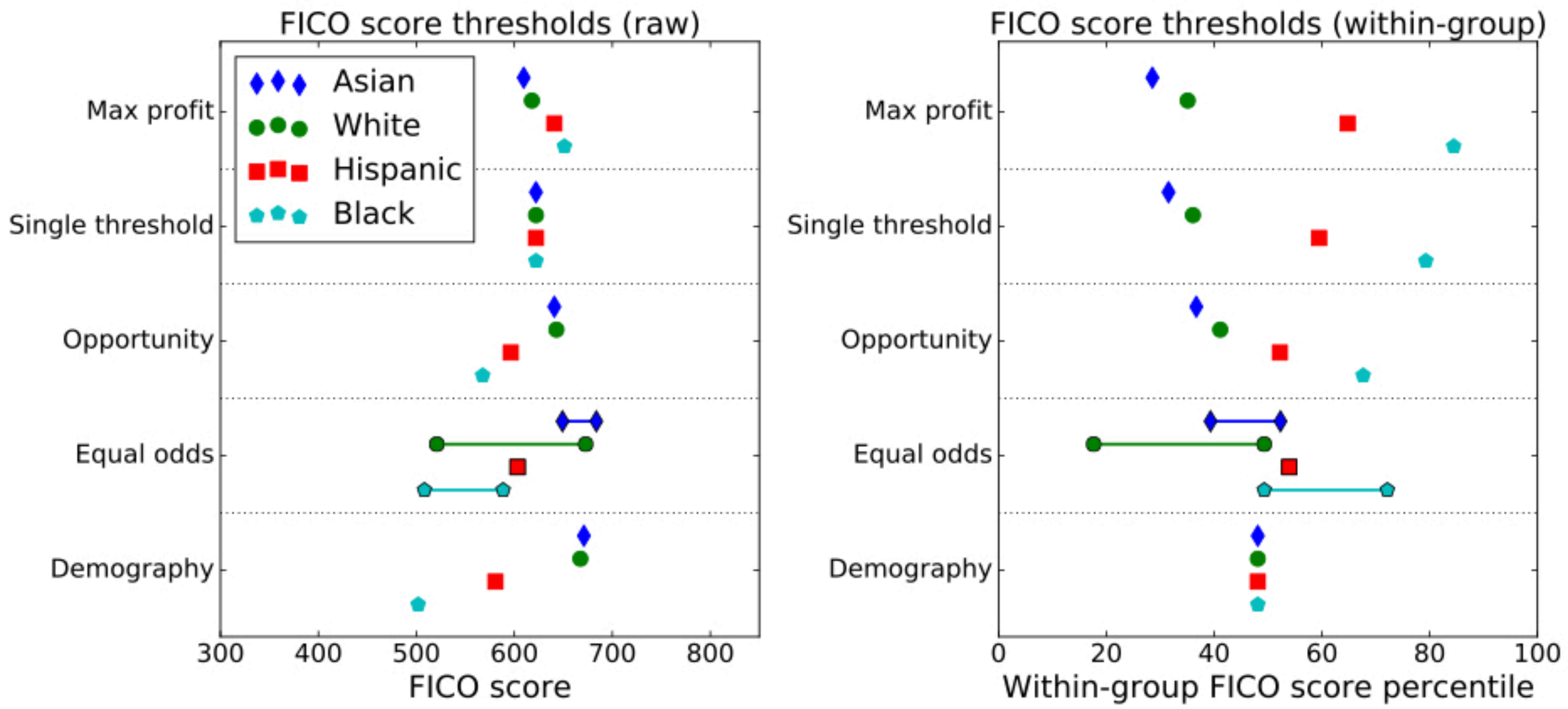


Figure 9: FICO thresholds for various definitions of fairness. The equal odds method does not give a single threshold, but instead $\Pr[\widehat{Y}=1 \mid R, A]$ increases over some not uniquely defined range; we pick the one containing the fewest people. Observe that, within each race, the equal opportunity threshold and average equal odds threshold lie between the max profit threshold and equal demography thresholds.