Carnegie Mellon University J.P.Morgan

FACT: A Diagnostic for Group Fairness Trade-offs

Joon Kim, CMU (joonsikk@cs.cmu.edu)
Jiahao Chen, JPMorgan AI Research (jiahao.chen@jpmchase.com)
Ameet Talwalkar, CMU (talwalkar@cmu.edu)

ICML 2020

Fairness in ML is becoming more important

- More application areas with societal impact
  - Credit decision/Loan approval
  - Healthcare provision
  - Recidivism prediction
  - Facial recognition

- Quantitative notions of fairness:
  - Individual fairness
  - Group fairness
  - Representation fairness
  - Counterfactual fairness …

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

ICML 2020

Why Group Fairness?

- Widely studied both in social sciences as a concept of *disparate impact*
- Practical instantiations
  - *p-percent rule*: among the accepted subjects, the ratio between the subjects having a certain sensitive attribute to the subjects that do not have the attribute, should be no less than p:100. (U.S. Equal Employment Opportunity Commission)
- Intuitive to understand, even for non-ML experts
- Active area of research in ML
  - predictive equality, predictive parity, demographic parity, equalized odds, equal opportunity, class balance, calibration, conditions accuracy equality …

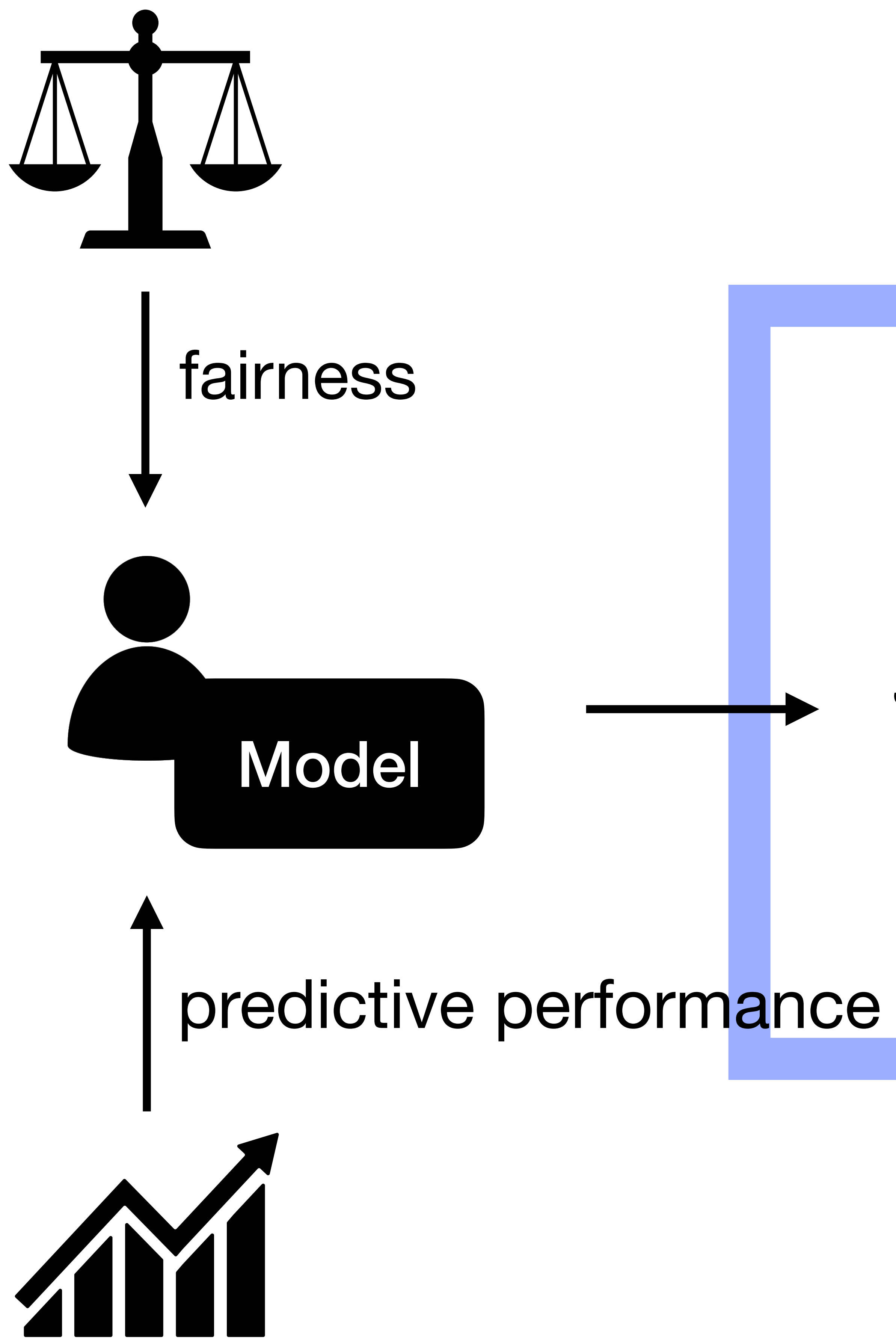
… but it comes with several trade-offs.

- *Type1. Fairness vs. Fairness (impossibility and incompatibility)*
  - “It is not possible to satisfy certain multiple notions of fairness simultaneously unless some strong assumptions about the data and the model are satisfied.”
  - Kleinberg et al. 2017, Chouldechova 2017, etc.

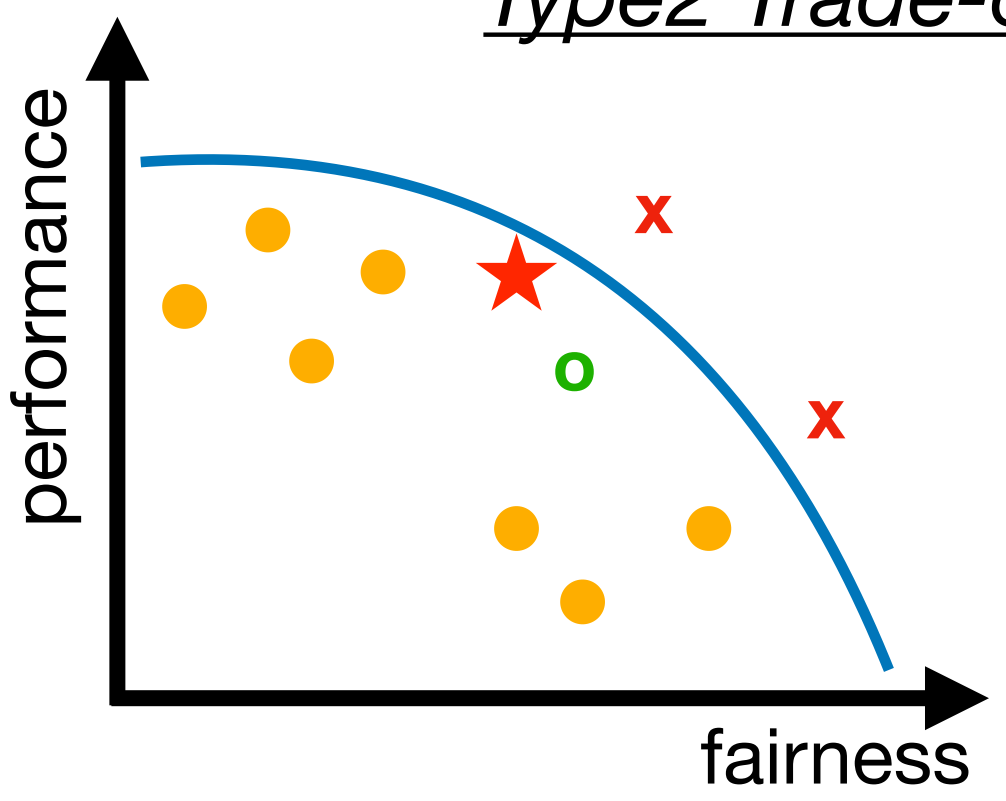
- *Type2. Fairness vs. Performance*
  - “Imposing fairness conditions tend to decrease the model’s predictive performance.”
  - Zafar et al. 2015, Menon and Williamson 2018, etc.

$\Rightarrow$ How to view them under a simple unified perspective?

Towards a systematic characterization of trade-offs



Type2 Trade-off



Type1 Trade-off

Incompatible fairness notions
{CG, PP, DP, and any of EOp, PE, PCB, NCB, EFOR}
{CG, DP, and any of EOp, PE, PCB, NCB, EFOR}
{CG,EOp}, {CG,PCB}, {CG,EOp,PCB},{CG,EFOR,EOp}, {CG,EFOR,PCB},{CG,EFOR,EOp,PCB}
{CG,PE}, {CG,NCB}, {CG,EOp,NCB}, {CG,EFOR,PE}, {CG,EFOR,NCB}, {CG,EFOR,EOp,NCB}
{CG,EOd}, {CG, PCB, NCB},{CG,EOd,PCB,NCB}, {CG,EFOR,EOd}, {CG,EFOR,PCB,NCB},{CG,EFOR,EOd,PCB,NCB}

FACT Diagnostic

ICML 2020

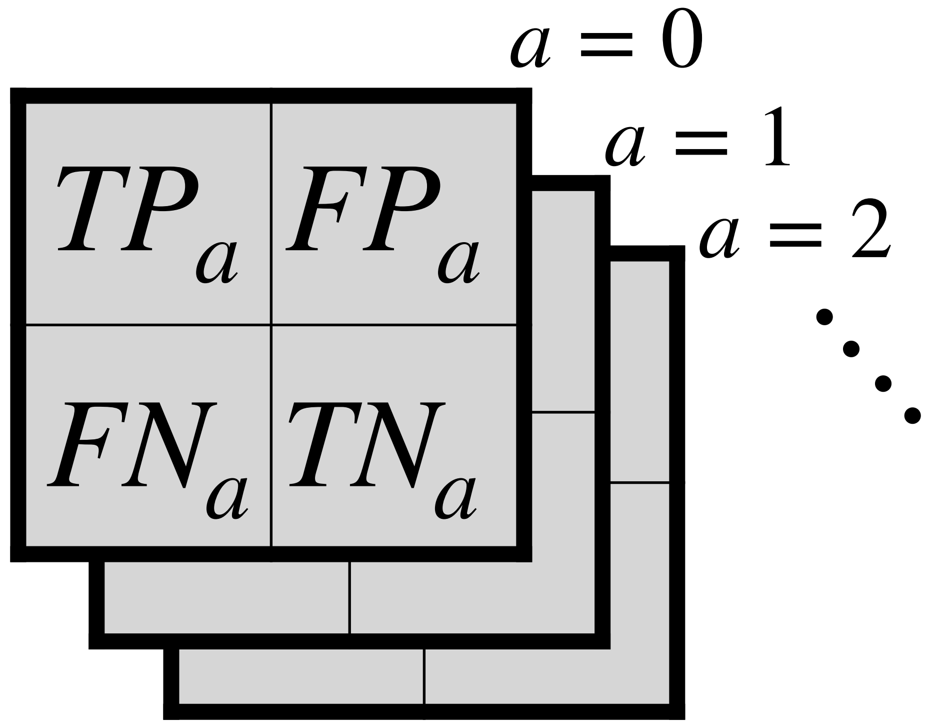
We will cover…

- Fairness-confusion tensor (FACT)
  - Provides a linear/quadratic characterization of group fairness notions
- Optimization problems over the fairness-confusion tensor
  - Solutions reflect the boundaries of the trade-off
  - One instance shows a general method for deriving fairness incompatibilities
  - One instance shows a connection to post-processing methods
- Demonstration on use cases

Fairness-confusion Tensor & Group Fairness

- *Fairness-confusion tensor* = stacked confusion matrix per protected attributes (a)

$\mathbf{z} =$



$$= (TP_1, FN_1, FP_1, TN_1, TP_0, FN_0, FP_0, TN_0)^T / N \qquad \in \mathscr{K}$$

- Group fairness takes the form : $(\text{value } r_1 \text{ from group } 1) - (\text{value } r_0 \text{ from group } 0) = 0$
  - The values are derived from the elements of the fairness-confusion tensor

Linear/Quadratic Group Fairness

- Fairness conditions can be rewritten as a condition $\phi(\mathbf{z}) = 0$ where

  - Linear fairness: $\phi(\mathbf{z}) = \mathbf{A}\mathbf{z}$

Demographic parity (DP)	$\Pr(\hat{y} = 1 \mid \mathbf{a} = 1) = \Pr(\hat{y} = 1 \mid \mathbf{a} = 0)$ $\mathbf{A}_{\text{DP}} = \frac{1}{N}\begin{pmatrix} N_0 & 0 & N_0 & 0 & -N_1 & 0 & -N_1 & 0 \end{pmatrix}$
Equality of opportunity (EOp)	$\Pr(\hat{y} = 1 \mid y = 1, \mathbf{a} = 1) = \Pr(\hat{y} = 1 \mid y = 1, \mathbf{a} = 0)$ $\mathbf{A}_{\text{EOP}} = \frac{1}{N}\begin{pmatrix} M_0 & 0 & 0 & 0 & -M_1 & 0 & 0 & 0 \end{pmatrix}$
Predictive equality (PE)	$\Pr(\hat{y} = 1 \mid y = 0, \mathbf{a} = 1) = \Pr(\hat{y} = 1 \mid y = 0, \mathbf{a} = 0)$ $\mathbf{A}_{\text{PE}} = \frac{1}{N}\begin{pmatrix} 0 & 0 & N_0 - M_0 & 0 & 0 & 0 & -N_1 + M_1 & 0 \end{pmatrix}$
Equalized odds (EOd)	EOp $\wedge$ PE
Equal false negative rate (EFNR)	$\Pr(\hat{y} = 0 \mid y = 1, \mathbf{a} = 1) = \Pr(\hat{y} = 0 \mid y = 1, \mathbf{a} = 0)$ $\mathbf{A}_{\text{EFNR}} = \frac{1}{N}\begin{pmatrix} 0 & M_0 & 0 & 0 & 0 & -M_1 & 0 & 0 \end{pmatrix}$
Calibration within groups (CG)	$\Pr(y = 1 \mid P_\theta(\mathbf{x}) = s, \mathbf{a} = 1) = \Pr(y = 1 \mid P_\theta(\mathbf{x}) = s, \mathbf{a} = 0) = s$ $\mathbf{A}_{\text{CG}} = \begin{pmatrix} 1 - v_1 & 0 & -v_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 - v_0 & 0 & -v_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 - v_1 & 0 & -v_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 - v_0 & 0 & -v_0 \end{pmatrix}$
Positive class balance (PCB)	$\mathbb{E}(P_\theta \mid y = 1, \mathbf{a} = 1) = \mathbb{E}(P_\theta \mid y = 1, \mathbf{a} = 0)$ $\mathbf{A}_{\text{PCB}} = \min_a(M_a)\begin{pmatrix} \frac{v_1}{M_1} & \frac{v_0}{M_1} & 0 & 0 & -\frac{v_1}{M_0} & -\frac{v_0}{M_0} & 0 & 0 \end{pmatrix}$
Negative class balance (NCB)	$\mathbb{E}(P_\theta \mid y = 0, \mathbf{a} = 1) = \mathbb{E}(P_\theta \mid y = 0, \mathbf{a} = 0)$ $\mathbf{A}_{\text{NCB}} = \min_a(N_a - M_a)\begin{pmatrix} 0 & 0 & \frac{v_1}{N_1 - M_1} & \frac{v_0}{N_1 - M_1} & 0 & 0 & -\frac{v_1}{N_0 - M_0} & -\frac{v_0}{N_0 - M_0} \end{pmatrix}$

  - Quadratic fairness: $\phi(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T B \mathbf{z}$

Predictive parity (PP)	$\Pr(y = 1 \mid \hat{y} = 1, \mathbf{a} = 1) = \Pr(y = 1 \mid \hat{y} = 1, \mathbf{a} = 0)$ $\frac{1}{2}\mathbf{z}^T \mathbf{B}_{\text{PP}} \mathbf{z} = (TP_1 FP_0 - TP_0 FP_1)/N^2$
Equal false omission rate (EFOR)	$\Pr(y = 1 \mid \hat{y} = 0, \mathbf{a} = 1) = \Pr(y = 1 \mid \hat{y} = 0, \mathbf{a} = 0)$ $\frac{1}{2}\mathbf{z}^T \mathbf{B}_{\text{EFOR}} \mathbf{z} = (TN_1 FN_0 - TN_0 FN_1)/N^2$
Conditional accuracy equality (CA)	PP $\wedge$ EFOR

ICML 2020

Optimizing over the Fairness-confusion Tensor

- Least-squares Accuracy-Fairness Optimality Problem (LAFOP)

$$\arg\min_{\mathbf{z}\in\mathcal{K}} \boxed{(\mathbf{c}\cdot\mathbf{z})^2} \quad + \quad \lambda \boxed{\|\mathbf{A}\mathbf{z}\|_2^2} \qquad \mathbf{c} = (0,1,1,0,0,1,1,0)^T$$

performance criteria
= classification error (accuracy)

fairness criteria
= linear fairness

- (ϵ, δ)-solutions: $\left\{\mathbf{z} : \mathbf{c}\cdot\mathbf{z} \leq \delta, \quad \|\mathbf{A}\mathbf{z}\| \leq \epsilon\right\}$
  - Demonstrate how the achievable performance δ can change across different fairness conditions measured by ϵ

ICML 2020

Special Case I: Incompatibility among Fairness

- When λ approaches infinity, solving LAFOP is equivalent to solving the following:

$$\begin{pmatrix} \mathbf{A}^{(0)} \\ \vdots \\ \mathbf{A}^{(K-1)} \\ \mathbf{A}_{\text{const}} \end{pmatrix} \mathbf{z} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{b}_{\text{const}} \end{pmatrix}, \mathbf{z} \geq 0$$

- Incompatibility can be verified by the number of solutions to this linear system

Sets of fairness definitions	Necessary conditions
{CG, PP, DP, and any of EOp, PE, PCB, NCB, EFOR}	$M_0 = M_1$ and $N_0 = N_1$
{CG, DP, and any of EOp, PE, PCB, NCB, EFOR}	EBR only
{CG,EOp}, {CG,PCB}, {CG,EOp,PCB},{CG,EFOR,EOp}, {CG,EFOR,PCB},{CG,EFOR,EOp,PCB}	$v_0 = 0$ or EBR
{CG,PE}, {CG,NCB}, {CG,EOp,NCB}, {CG,EFOR,PE}, {CG,EFOR,NCB}, {CG,EFOR,EOp,NCB}	$v_1 = 1$ or EBR
{CG,EOd}, {CG, PCB, NCB}, {CG,EOd,PCB,NCB}, {CG,EFOR,EOd}, {CG,EFOR,PCB,NCB},{CG,EFOR,EOd,PCB,NCB}	($v_0 = 0$ and $v_1 = 1$) or EBR

ICML 2020

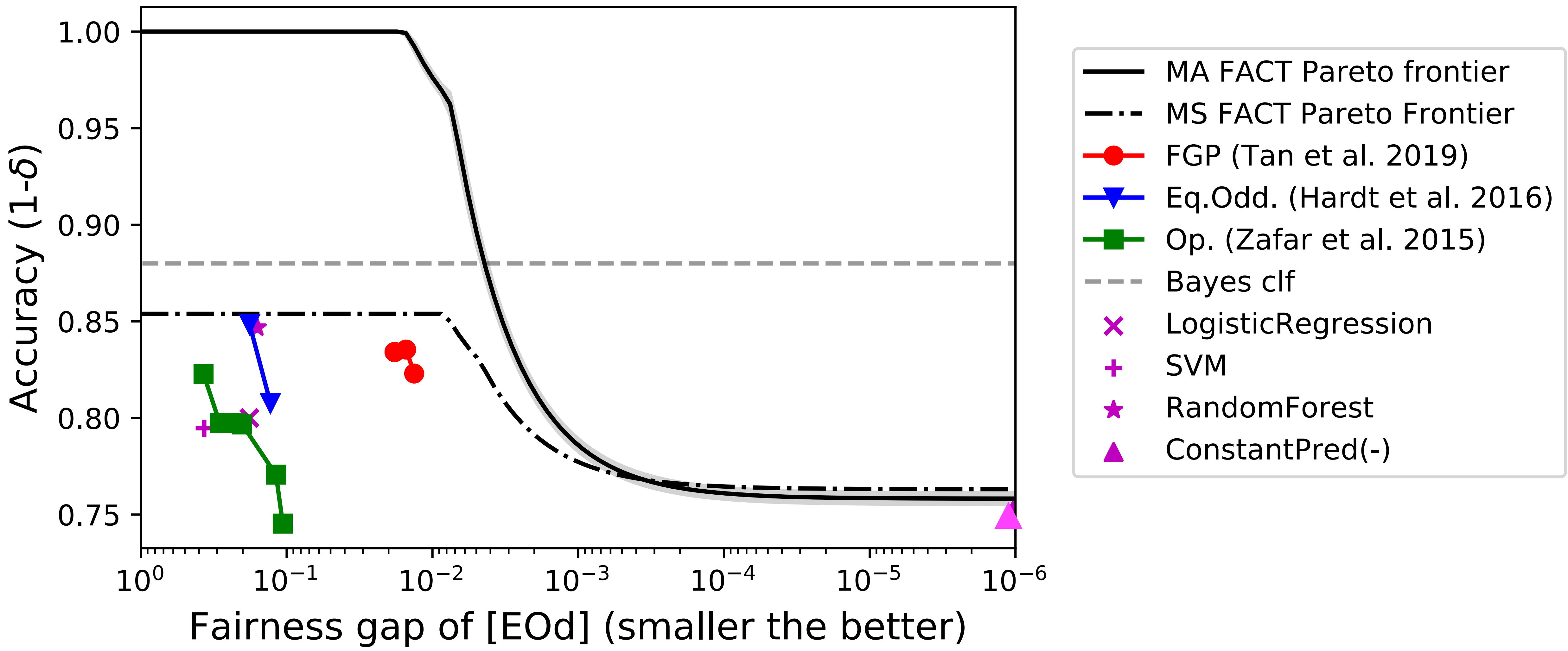
Special Case II: Post-processing

- Model-specific LAFOP (MS-LAFOP)

$$\arg\min_{\mathbf{z}\in\mathscr{K}} \underbrace{\boxed{(\mathbf{c}\cdot\mathbf{z})^2}}_{\substack{\text{performance criteria} \\ \text{= accuracy}}} + \lambda \underbrace{\boxed{\|\mathbf{A}\mathbf{z}\|_2^2}}_{\substack{\text{fairness criteria} \\ \text{= linear fairness}}} \quad \text{such that} \quad \underbrace{\boxed{\phi(\mathbf{z}) \in \Gamma(\hat{\mathbf{z}})}}_{\text{model-specific constraints on fairness}}$$

ICML 2020

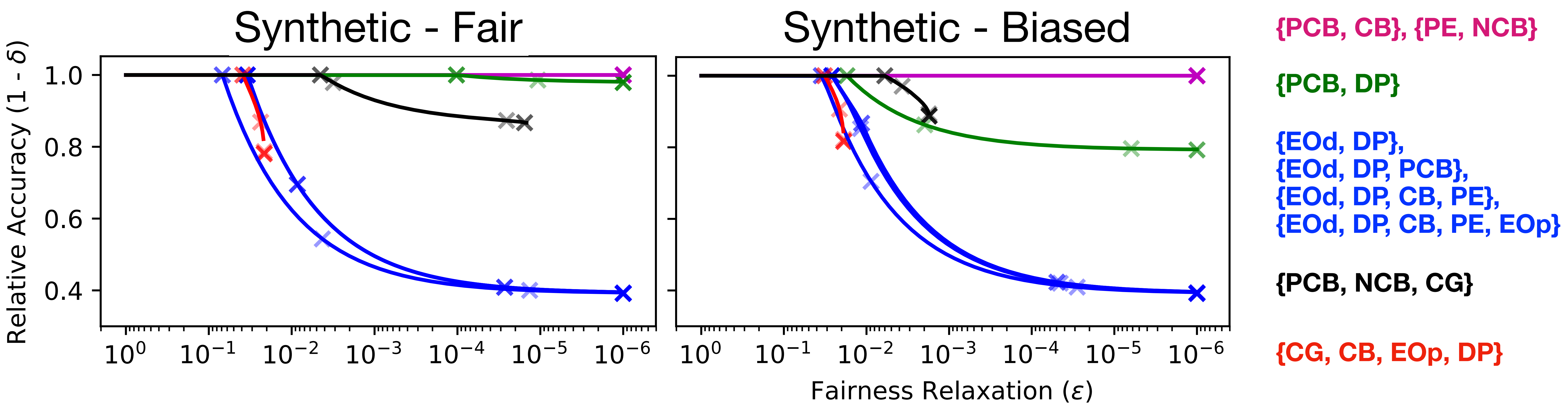
FACT Pareto Frontiers



- Set of (ϵ, δ)-solutions of LAFOP plotted over varying ϵ
- Model-agnostic case (MA): bounds should be interpreted w.r.t the Bayes error
- Model-specific case (MS): bounds are more realistic

ICML 2020

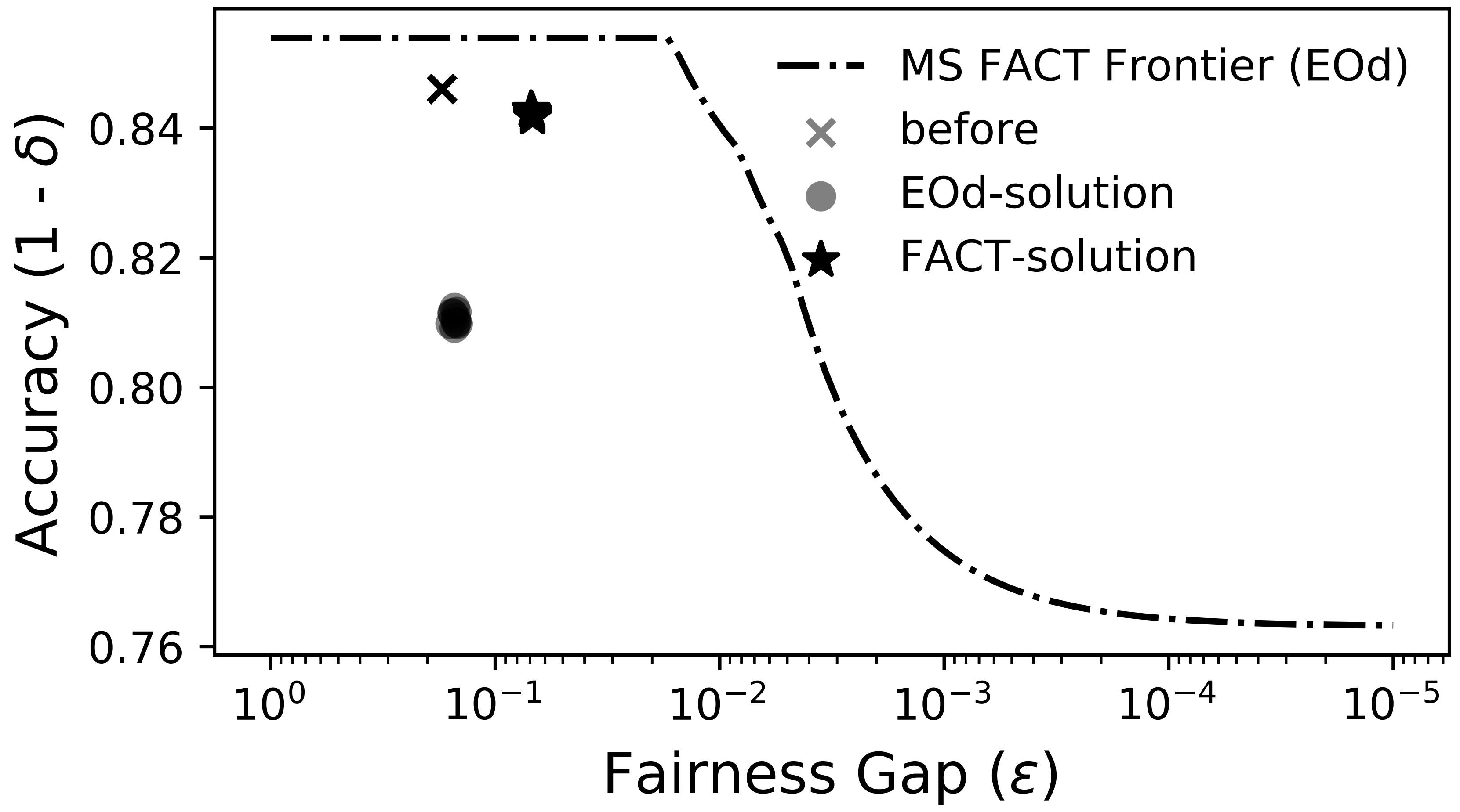
A model-agnostic scenario



- Equalized Odds (EOd) and Demographic Parity (DP) dominates the behaviors of the curves in blue.
- Halted trajectories for Black and Red lines indicate incompatibility.
- Fair dataset yields a better trade-off scheme than the biased dataset.

ICML 2020

A model-specific scenario: reduction to post-processing



- We can compute a *mixing ratio* for post-processing methods using the solutions from MS-LAFOP.

- FACT-solution finds a better classifier with a smaller trade-off.

ICML 2020

Discussions

- FACT diagnostic for systematic reasoning about type1 and type2 trade-offs involving group fairness.
- Fairness-confusion tensor provides a unified perspective on group fairness.
- Many results presented only involved linear fairness and accuracy (LAFOP, MS-LAFOP), but we can expect a more diverse results from the more general class of optimization problem presented in the paper.
- Post-processing via FACT can be generalized to other notions of fairness.

FACT: A Diagnostic for Group Fairness Trade-offs

Website: www.cs.cmu.edu/joonsikk

Paper: https://arxiv.org/abs/2004.03424

Joon Kim, CMU (joonsikk@cs.cmu.edu)
Jiahao Chen, JPMorgan AI Research (jiahao.chen@jpmchase.com)
Ameet Talwalkar, CMU (talwalkar@cmu.edu)

ICML 2020