Aleatoric and Epistemic Discrimination:
Fundamental Limits of Fairness Interventions

Hao Wang, Luxi (Lucy) He, Rui Gao, Flavio P. Calmon

hao@ibm.com

Spotlight

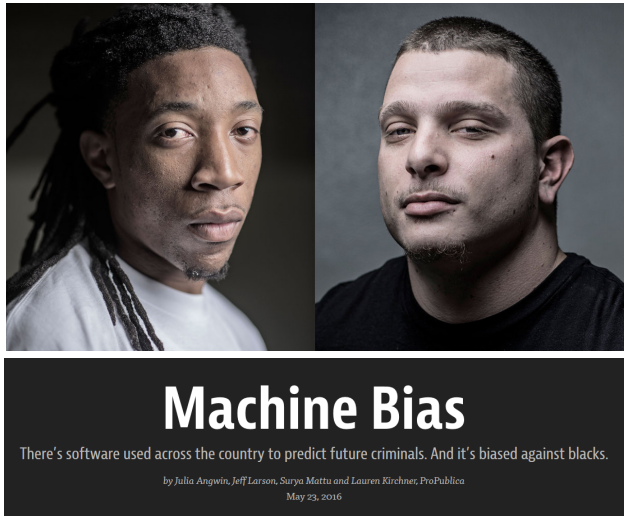
Harvard John A. Paulson
School of Engineering
and Applied Sciences

MIT-IBM
Watson
AI Lab

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Algorithmic discrimination and fairness interventions

UK's A-level grading algorithm



ProPublica'16

Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey

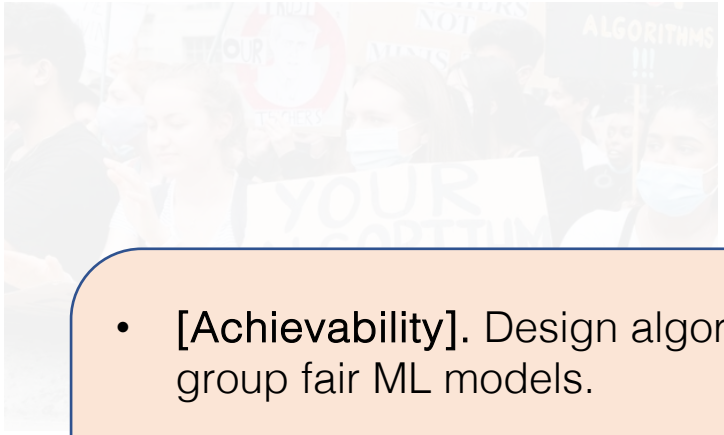
Max Hort, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Mark Harman

Abstract—This paper provides a comprehensive survey of bias mitigation methods for achieving fairness in Machine Learning (ML) models. We collect a total of 341 publications concerning bias mitigation for ML classifiers. These methods can be distinguished based on their intervention procedure (i.e., pre-processing, in-processing, post-processing) and the technology they apply. We investigate how existing bias mitigation methods are evaluated in the literature. In particular, we consider datasets, metrics and benchmarking. Based on the gathered insights (e.g., What is the most popular fairness metric? How many datasets are used for evaluating bias mitigation methods?). We hope to support practitioners in making informed choices when developing and evaluating new bias mitigation methods.

Index Terms—fairness, bias mitigation, machine learning

Recent survey found 341 fair learning algorithms!

An information-theoretic perspective



- **[Achievability].** Design algorithms to find both accurate and group fair ML models.
- **[Converse].** For a fixed data distribution, what is the information-theoretic limit of accuracy and group fairness, beyond which no model can achieve.

main focus of this paper

Fairness Pareto frontier for classification tasks

Definition. For a data distribution $P_{\mathsf{S},\mathsf{X},\mathsf{Y}}$ and $\alpha \geq 0$, we define

$$\mathsf{FairFront}(\alpha) \triangleq \max_{h} \text{ accuracy}$$
$$\text{s.t. fairness metrics} \leq \alpha$$

where the maximum is taken over all probabilistic classifiers h.

Examples of group fairness metrics

FAIRNESS METRIC	DEFINITION
Statistical Parity	$\lvert \Pr(\hat{\mathsf{Y}} = \hat{y} \mid \mathsf{S} = s) - \Pr(\hat{\mathsf{Y}} = \hat{y} \mid \mathsf{S} = s') \rvert \leq \alpha_{\mathsf{SP}}$
Equalized Odds	$\lvert \Pr(\hat{\mathsf{Y}} = \hat{y} \mid \mathsf{S} = s, \mathsf{Y} = y) - \Pr(\hat{\mathsf{Y}} = \hat{y} \mid \mathsf{S} = s', \mathsf{Y} = y) \rvert \leq \alpha_{\mathsf{EO}}$
Overall Accuracy Equality	$\lvert \Pr(\hat{\mathsf{Y}} = \mathsf{Y} \mid \mathsf{S} = s) - \Pr(\hat{\mathsf{Y}} = \mathsf{Y} \mid \mathsf{S} = s') \rvert \leq \alpha_{\mathsf{OAE}}$

S: (sensitive) group attributes, X: input features

Y: true label, $\hat{\mathsf{Y}}$: predicted outcome

Rewrite fairness Pareto frontier

$$\text{FairFront}(\alpha) \triangleq \max_{\text{classifier } h} \text{accuracy} \quad \text{s.t. fairness metrics} \le \alpha$$

functional optimization

$$\stackrel{?}{=} \max_{P_{\hat{Y}|Y,S}} \text{accuracy} \quad \text{s.t. fairness metrics} \le \alpha, \quad P_{\hat{Y}|Y,S} \in [0,1]^8 \cap \{\text{TNR} + \text{FPR} = 1\} \cap \{\text{FNR} + \text{TPR} = 1\}$$

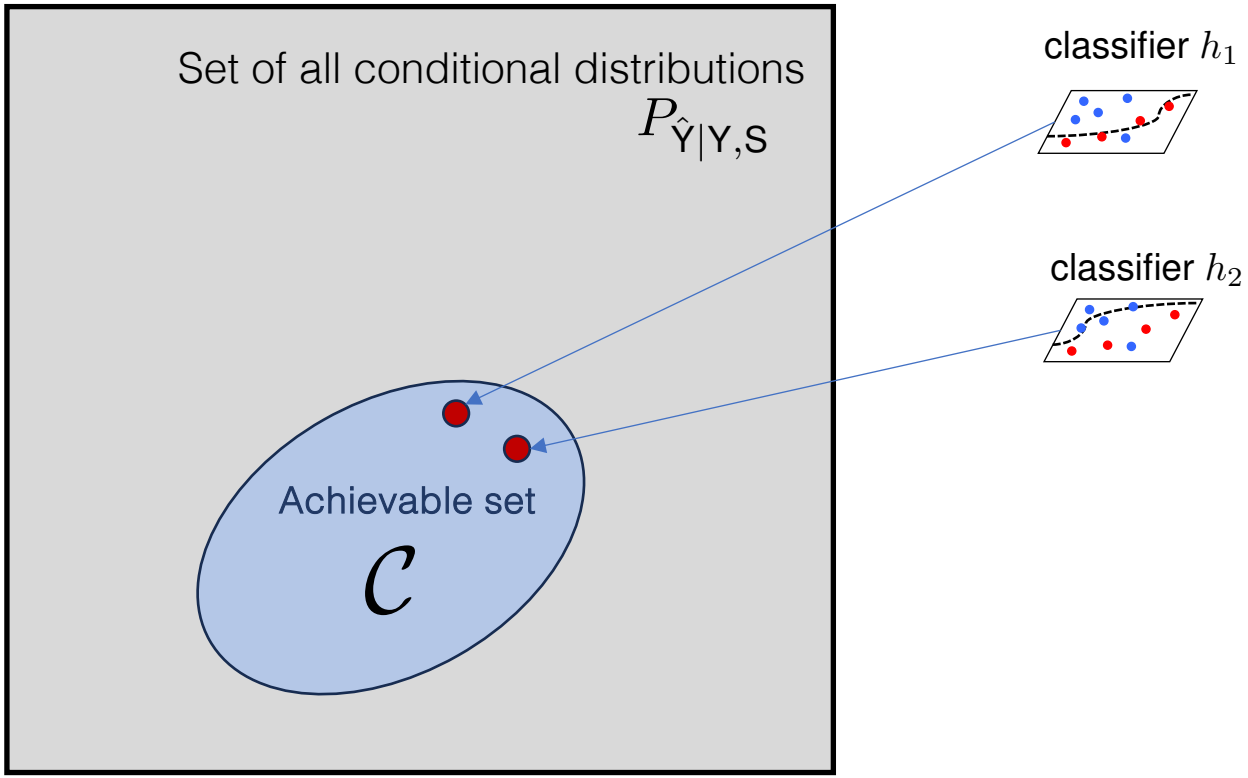
linear program with 8 variables!

Both accuracy and group fairness metrics can be written in terms of $P_{\hat{Y}|Y,S}$.

$$P_{\hat{Y}|Y,S} = \left(\begin{array}{cc} \text{TNR} & \text{FPR} \\ \text{FNR} & \text{TPR} \end{array} \right)_{\text{Group 0}}, \left(\begin{array}{cc} \text{TNR} & \text{FPR} \\ \text{FNR} & \text{TPR} \end{array} \right)_{\text{Group 1}} \in [0,1]^8 \cap \{\text{TNR} + \text{FPR} = 1\} \cap \{\text{FNR} + \text{TPR} = 1\}$$

There is an issue…

Not every $P_{\hat{Y}|Y,S}$ in $[0,1]^8 \cap \{\text{TNR} + \text{FPR} = 1\} \cap \{\text{FNR} + \text{TPR} = 1\}$ corresponds to a feasible classifier h.



Main theoretical result

COMPARISON OF EXPERIMENTS

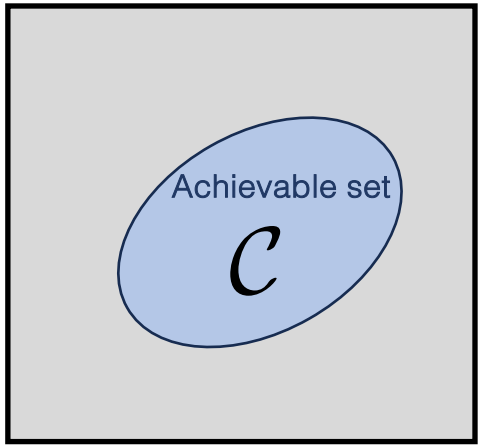
DAVID BLACKWELL
HOWARD UNIVERSITY

1. Summary

Bohnenblust, Shapley, and Sherman [2] have introduced a method of comparing two sampling procedures or experiments; essentially their concept is that one experiment α is more informative than a second experiment β, $\alpha \supset \beta$, if, for every possible risk function, any risk attainable with β is also attainable with α. If α is a sufficient statistic for a procedure equivalent to β, $\alpha \succ \beta$, it is shown that $\alpha \supset \beta$. In the case of dichotomies, the converse is proved. Whether $\succ$ and $\supset$ are equivalent in general is not known. Various properties of $\succ$ and $\supset$ are obtained, such as the following: if $\alpha \succ \beta$ and γ is independent of both, then the combination $(\alpha, \gamma) \succ (\beta, \gamma)$. An application to a problem in 2×2 tables is discussed.

Apply Blackwell's results

to characterize the achievable set.



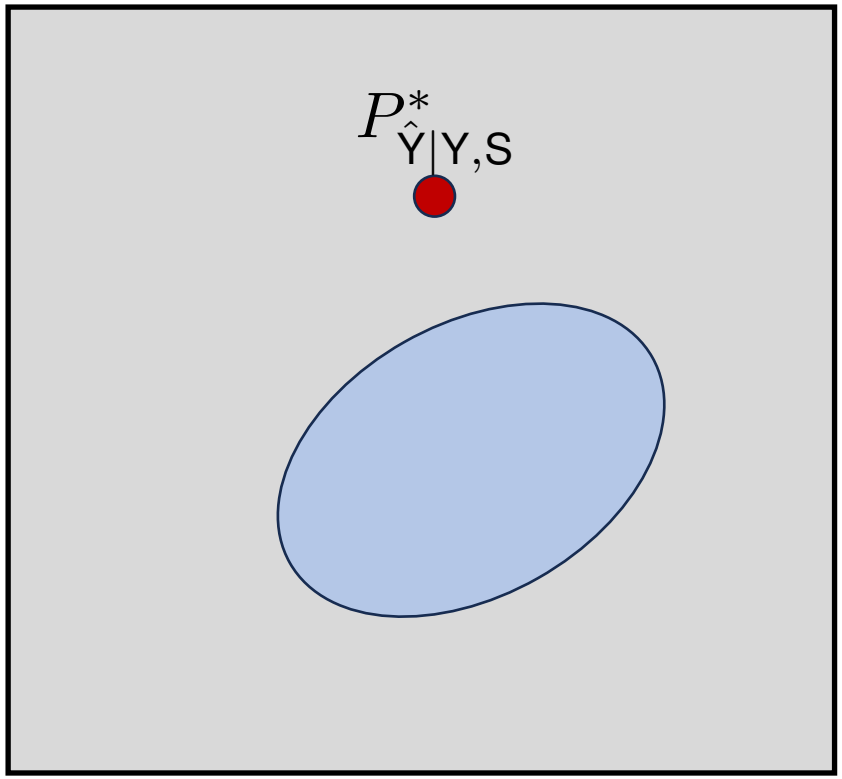
Theorem. The set $\mathcal{C}$ is the collection of all conditional distributions $P_{\hat{\mathsf{Y}}|\mathsf{S},\mathsf{Y}}$ s.t. For any $k \in \mathcal{N}$ and any $\{\boldsymbol{a}_i \mid \boldsymbol{a}_i \in [-1,1]^{AC}, i \in [k]\}$,

$$\sum_{\hat{y}=1}^{C} \max_{i\in[k]} \left\{\boldsymbol{a}_i^T \boldsymbol{\Lambda}_\mu \boldsymbol{p}_{\hat{y}}\right\} \leq \mathbb{E}\left[\max_{i\in[k]} \{\boldsymbol{a}_i^T \boldsymbol{g}(\mathsf{X})\}\right],$$

where $\boldsymbol{p}_{\hat{y}} \triangleq (P_{\hat{\mathsf{Y}}|\mathsf{S},\mathsf{Y}}(\hat{y}|1,1), \cdots, P_{\hat{\mathsf{Y}}|\mathsf{S},\mathsf{Y}}(\hat{y}|A,C))^T$, $\boldsymbol{\Lambda}_\mu = \mathsf{diag}(\mu_{1,1}, \cdots, \mu_{A,C})$ with $\mu_{s,y} \triangleq \Pr(\mathsf{S}=s, \mathsf{Y}=y)$, and $\boldsymbol{g}(x) = \big(P_{\mathsf{S},\mathsf{Y}|\mathsf{X}}(1,1|x), \cdots, P_{\mathsf{S},\mathsf{Y}|\mathsf{X}}(A,C|x)\big)$.

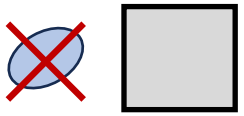
An iterative algorithm to approximate FairFront

Step 1. Find an optimal conditional distribution over a relaxed achievable set.



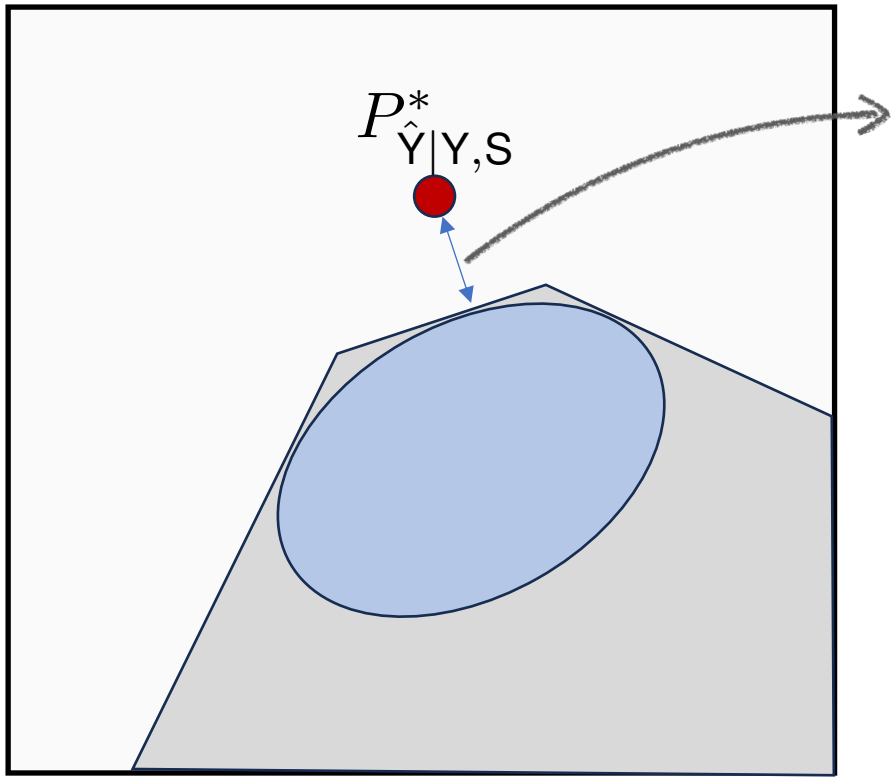
$$P^*_{\hat{Y}|Y,S} = \underset{P_{\hat{Y}|Y,S}}{\operatorname{argmax}} \text{ accuracy}$$

s.t. fairness metrics $\leq \alpha$

$P_{\hat{Y}|Y,S} \in$ 

An iterative algorithm to approximate FairFront

Step 2. Use Blackwell's results to refine the relaxed achievable set.



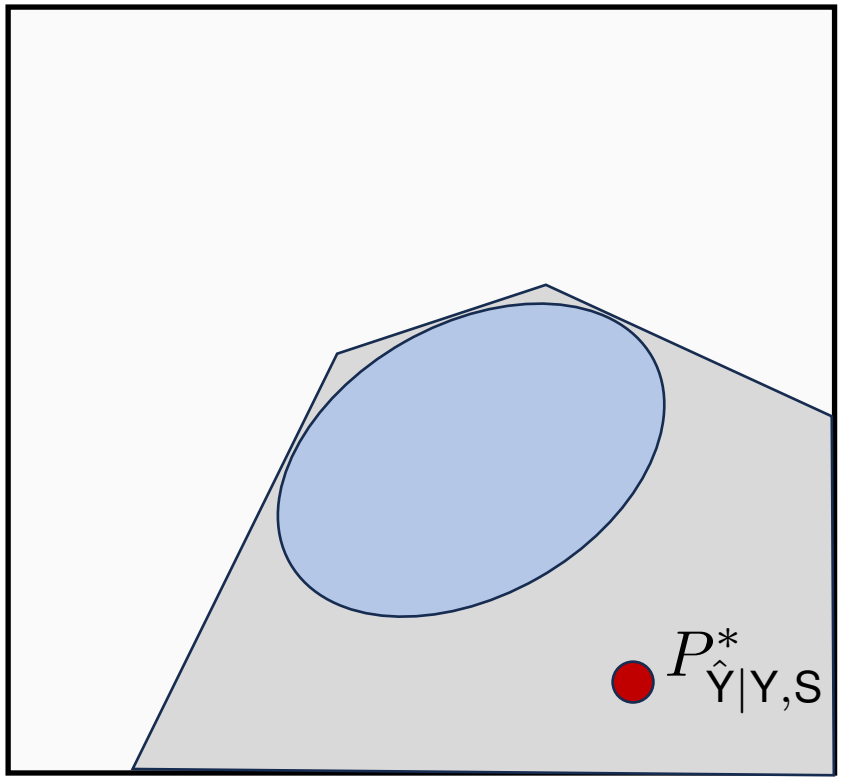
Solve a DC program to max this distance

$$\min_{\substack{\boldsymbol{a}_i \in [-1,1]^{AC} \\ i \in [k]}} \mathbb{E}\left[\max_{i\in[k]}\{\boldsymbol{a}_i^T \boldsymbol{g}(X)\}\right] - \sum_{\hat{y}=1}^{C} \max_{i\in[k]} \left\{\boldsymbol{a}_i^T \boldsymbol{\Lambda}_\mu \boldsymbol{p}_{\hat{y}}^t\right\}.$$

Intuition: find a piecewise linear function that separates $P^*_{\hat{Y}|Y,S}$ from the achievable set.

An iterative algorithm to approximate FairFront

Step 1. Find an optimal conditional distribution over a relaxed achievable set.



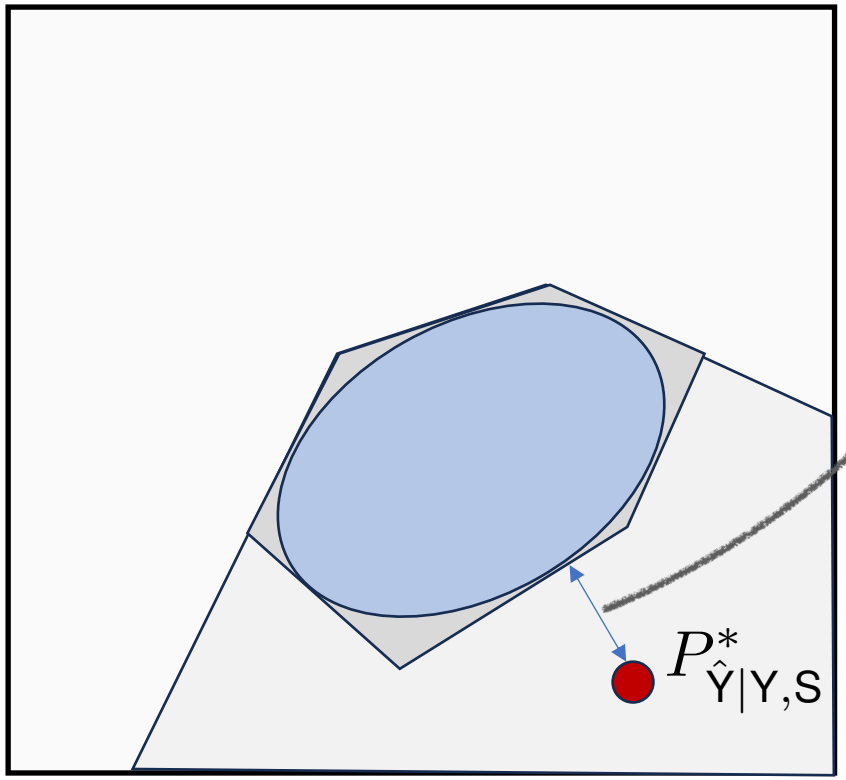
$$P^*_{\hat{Y}|Y,S} = \underset{P_{\hat{Y}|Y,S}}{\operatorname{argmax}} \text{ accuracy}$$

s.t. fairness metrics $\leq \alpha$

$P_{\hat{Y}|Y,S} \in$ 

An iterative algorithm to approximate FairFront

Step 2. Use Blackwell's results to refine the relaxed achievable set.



Solve a DC program to max this distance

$$\min_{\substack{\boldsymbol{a}_i \in [-1,1]^{AC} \\ i \in [k]}} \mathbb{E}\left[\max_{i\in[k]}\{\boldsymbol{a}_i^T \boldsymbol{g}(X)\}\right] - \sum_{\hat{y}=1}^{C} \max_{i\in[k]} \left\{\boldsymbol{a}_i^T \boldsymbol{\Lambda}_\mu \boldsymbol{p}_{\hat{y}}^t\right\}.$$

Tightness of our upper bound

- Our algorithm provides an upper bound estimate of FairFront.
- Existing (group) fairness interventions provide lower bounds.

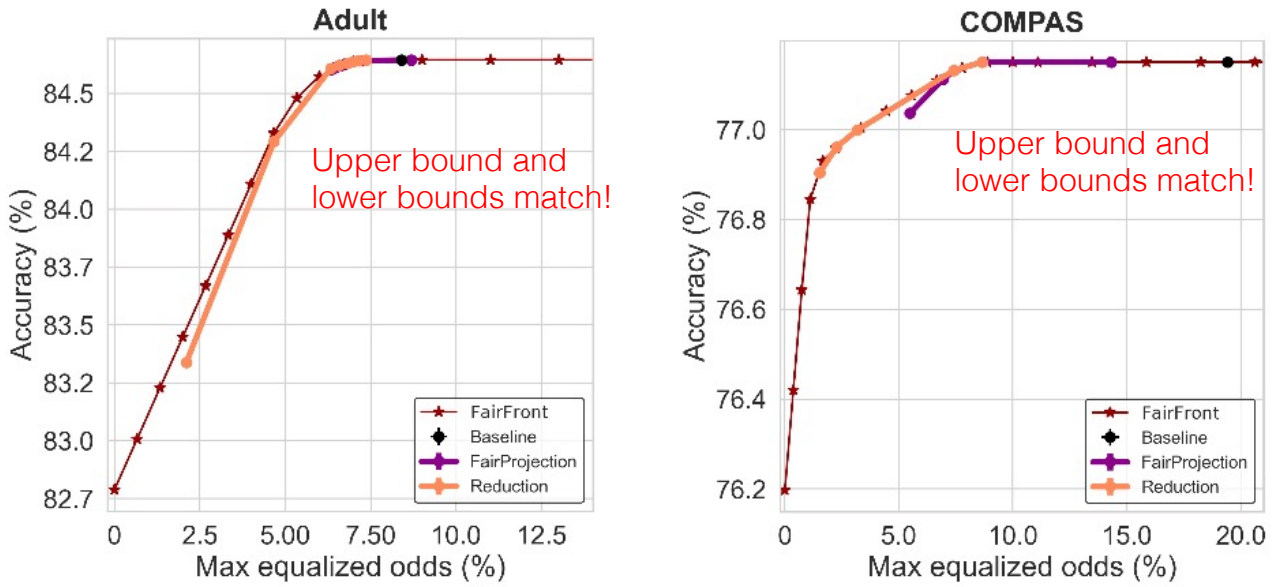


Figure 1: We compare `Reduction` and `FairProjection` with (our upper bound estimate of) FairFront on the Adult (Left) and COMPAS (Right) datasets. We train a classifier that approximates the Bayes optimal and use it as a basis for `Reduction` and `FairProjection`. This result not only demonstrates the tightness of our approximation but also shows that SOTA fairness interventions have already achieved near-optimal fairness-accuracy curves.

Our numerical experiments are semi-synthetic since we apply fairness interventions to train classifiers using the entire dataset and resample from it as the test set. This setup enables us to eliminate the estimation error associated with our algorithm.

Aleatoric and epistemic discrimination

- Aleatoric discrimination captures inherent biases in the data distribution that can lead to unfair decisions in downstream tasks.
- Epistemic discrimination is due to algorithmic choices made during model development and lack of knowledge about the optimal "fair" predictive model.

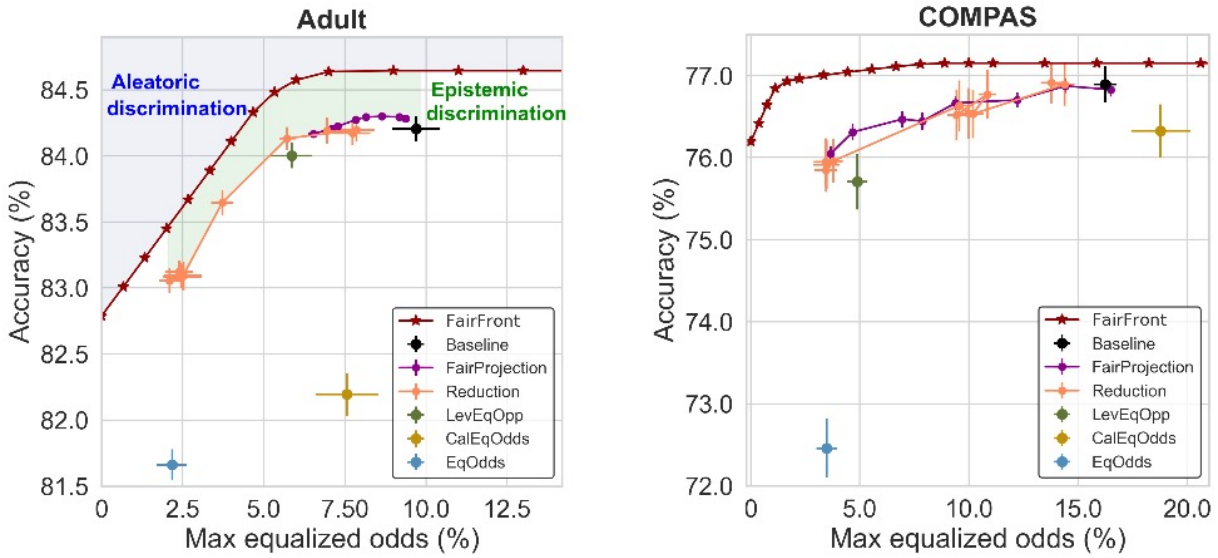


Figure 2: We benchmark existing fairness interventions using (our upper bound estimate of) FairFront. We use FairFront to quantify aleatoric discrimination and measure epistemic discrimination by comparing a classifier's accuracy and fairness violation with FairFront. The results show that SOTA fairness interventions are effective at reducing epistemic discrimination.

We borrow this notion from ML uncertainty literature (see Hüllermeier and Waegeman, 2021, for a survey).

Fairness in missing values

- Real-world data often have missing values, and the missing patterns can be different across different protected groups
- When population groups have disparate missing patterns, aleatoric discrimination escalates, leading to a sharp decline in the effectiveness of fairness intervention algorithms.

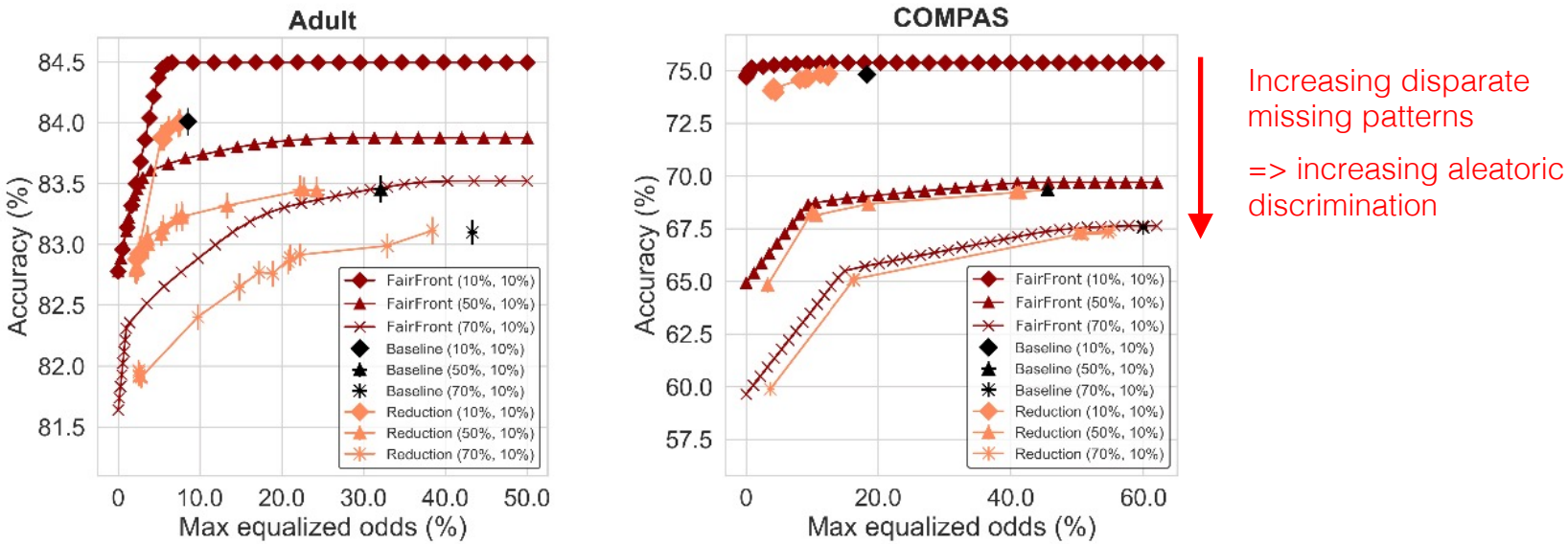


Figure 3: Fairness risks of disparate missing patterns. The missing probabilities of group 0 (female in Adult/African-American in COMPAS) and group 1 (male in Adult/Caucasian in COMPAS) are varying among $\{(10\%, 10\%), (50\%, 10\%), (70\%, 10\%)\}$. We apply `Reduction` and `Baseline` to the imputed data and plot their fairness-accuracy curves against FairFront. As shown, the effectiveness of fairness interventions substantially decrease with increasing disparate missing patterns in data.

Thanks!!!!

Poster

Aleatoric and Epistemic Discrimination: Fundamental Limits of Fairness Interventions

Hao Wang · Luxi He · Rui Gao · Flavio Calmon

Great Hall & Hall B1+B2 #1601

[Abstract]

[OpenReview]

Wed 13 Dec 5 p.m. CST — 7 p.m. CST **(Bookmark)**

Please visit our poster,
if you are interested! ☺