AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

Can We Obtain Fairness For Free?

Rashidul Islam, Shimei Pan, and James R. Foulds
Department of Information Systems
University of Maryland, Baltimore County

Motivation

- There is growing awareness that AI and ML systems can in some cases learn to behave in unfair ways

J. Angwin et al., ProPublica, 2016

- AI community has invested a large amount of effort
- However, techniques for ensuring fairness have currently attained relatively little adoption in deployed AI systems
- Main barrier: Fairness brings a cost in performance!

"Big Tech refuses to prioritize solving these issues over their bottom line."

-- Kate Crawford. NYT, 2016

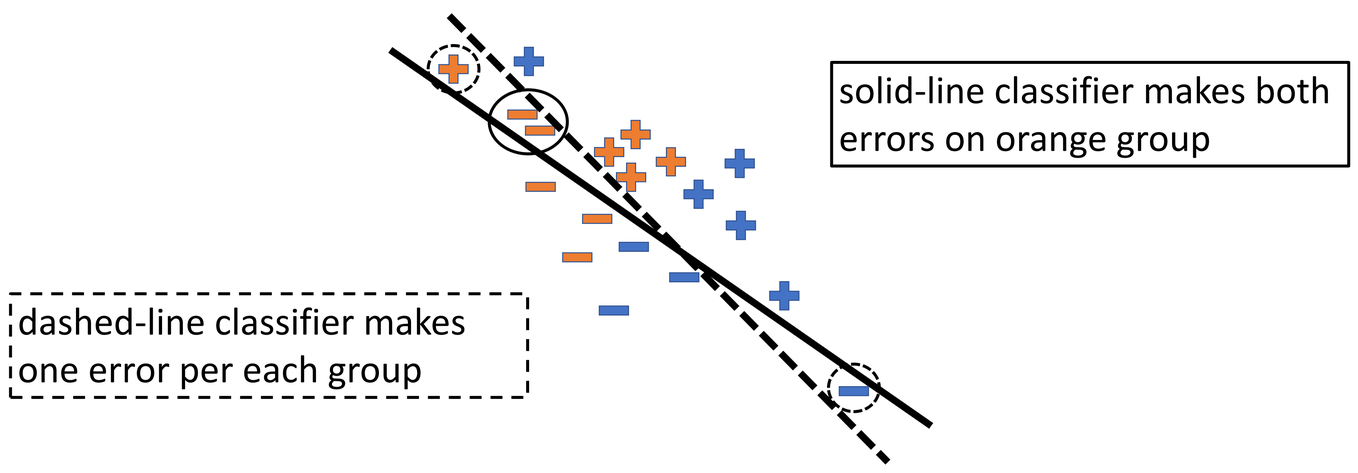
Our Research Questions?

- Clearly trade-offs exist, but are they inevitable?
- Is it possible to obtain some degree of improvement in fairness metrics for free?

*Our study shows the answer is **frequently yes!***

Fairness for Free

- We identify two mechanisms that can potentially lead to fairness for free:
  1. **The regularization benefits of fairness penalties**
     - It has potential to reduce overfitting
  2. **"Gerrymandering" the errors between protected groups**
     - Multiple classifiers can potentially obtain same or similar number of errors



Hyper-parameter Selection Strategy

- **Full Hyper-parameter Search (FHS)**
  our gold standard
  - Over all DNN hyper-parameters + λ
- **Stage-wise Hyper-parameter Search (SHS)**
  faster alternative
  - Over only the fairness trade-off λ

For both strategies, select the fair model with the best fairness metric, such that accuracy is at least as good as for best typical model (TM)

Fair Learning Algorithms

- Differential Fair Model (DFM) — J Foulds et al., ICDE, 2020

$$\min_\theta f(\mathbf{X};\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x}_i;\theta) + \lambda[\max(0, \epsilon(\mathbf{X};\theta) - \epsilon_t)]$$

Trade-off

- Adversarial Debiasing Model (ADM)

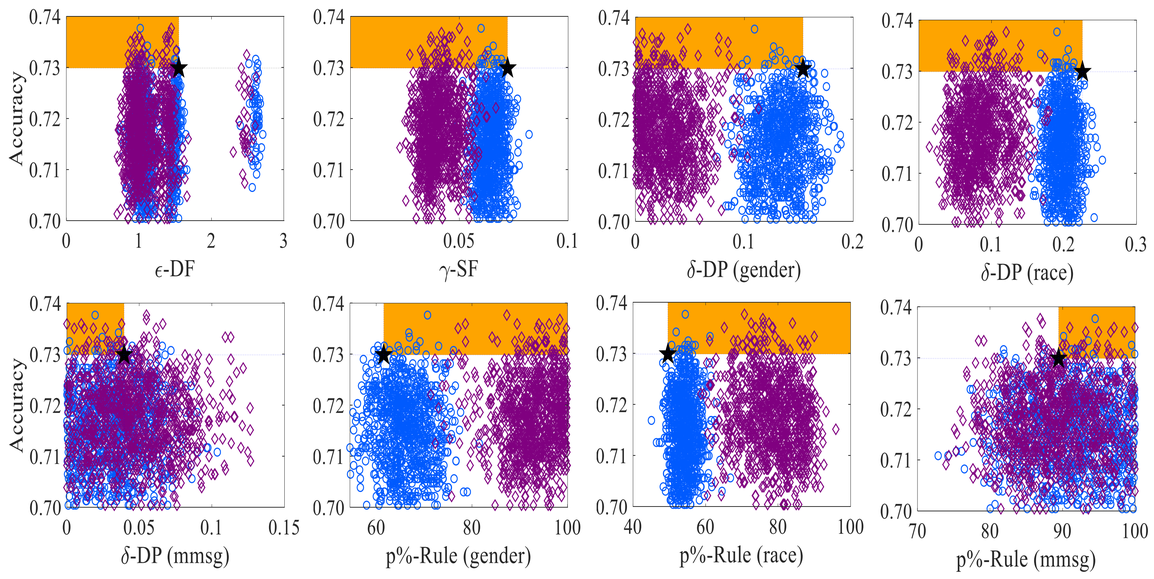
$$\min_\theta \max_\phi f(\mathbf{X};\theta,\phi) \triangleq \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x}_i;\theta) - \lambda L(\mathbf{X};\theta,\phi)$$

Louppe et al., NeurIPS, 2017

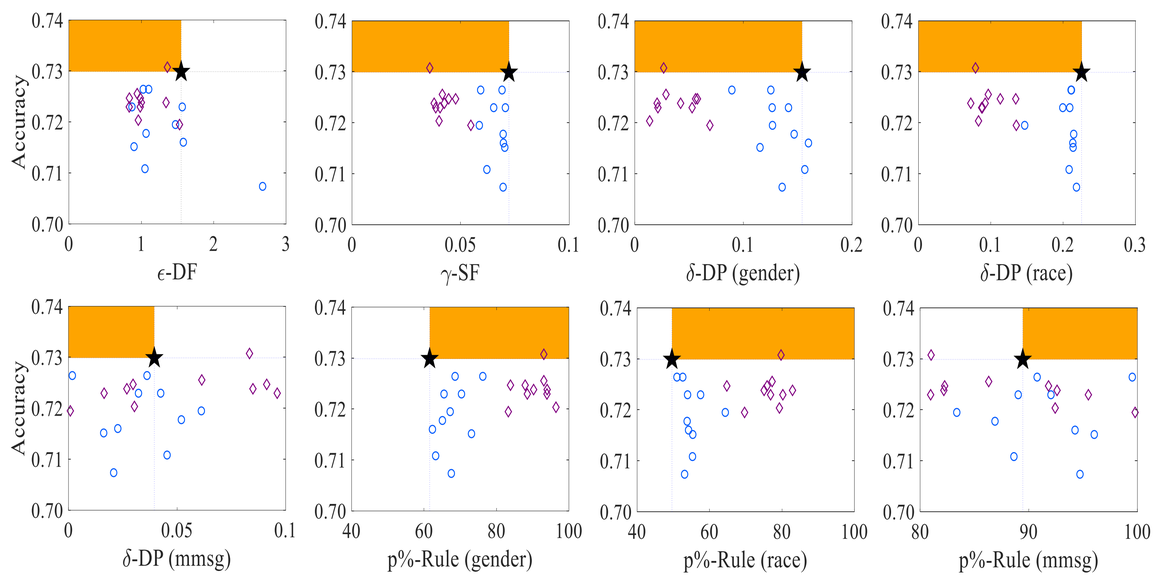
Analysis on Grid Search

★ TM ○ DFM ◇ ADM ▬ Fairness for Free Region

FHS on COMPAS



SHS on COMPAS

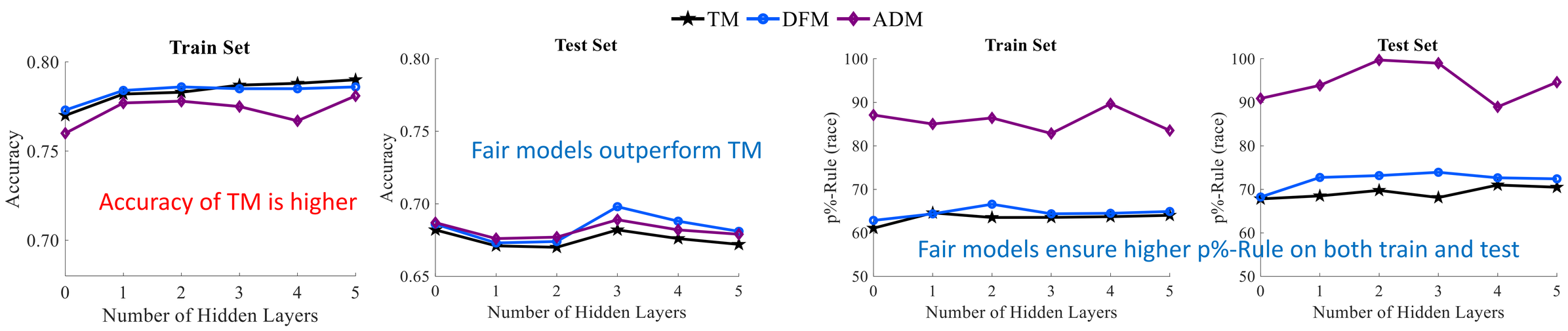


A large number of fair models satisfied the criteria of ``fairness for free'' in terms of all the fairness metrics

Only a single ADM satisfied our criteria of ``fairness for free" for most of the fairness metrics

Similar results for FHS and SHS approach on other benchmark datasets: **Adult**, **Bank**, and **HHP** data

Case Study on Overfitting



Fair models reduce overfitting which helps to improve both accuracy and fairness

We demonstrate that it is possible to improve fairness to some degree with no loss or even an improvement in accuracy via a sensible hyper-parameter selection strategy

Our results reveal a pathway toward increasing the deployment of fairness techniques in real systems