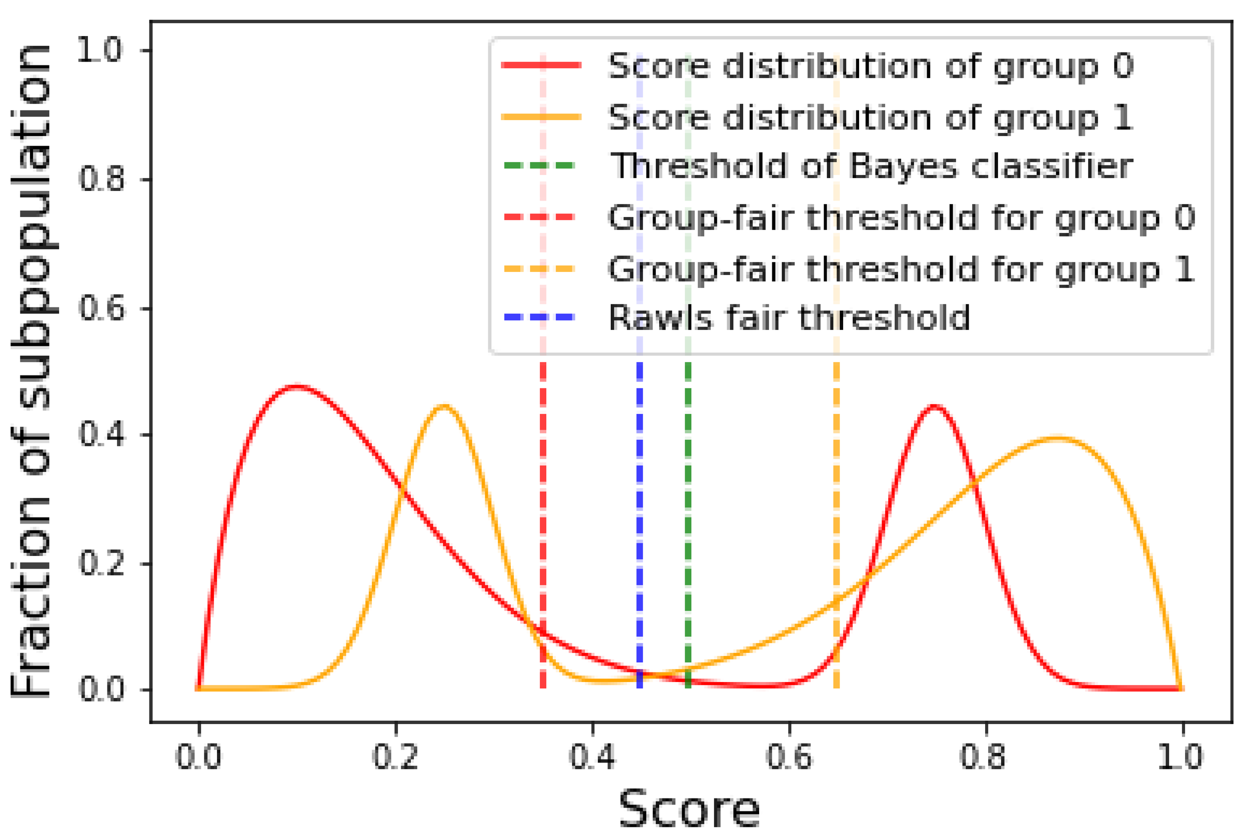
Rawlsian Fair Adaptation of Deep Learning Classifiers

Kulin Shah[1], Pooja Gupta[2], Amit Deshpande[1], Chiranjib Bhattacharyya[2]

[1] Microsoft Research, [2] Indian Institute of Science

Motivation



- **Bayes Classifier** (accuracy maximizer): Threshold at 0.5 on the score defined by $\eta(x) = \Pr(Y = 1 | X = x)$
- **Group-fair Classifier** (accuracy maximizer subject to demographic parity, equal opportunity etc.):
  - Group-aware: Group-dependent threshold on $\eta(x)$ [1]
  - Group-blind: Instance-dependent threshold $t(x)$ on $\eta(x)$ [2]
- **Rawlsian Fair Threshold**: Threshold that maximizes the minimum group-wise class-wise accuracy

Rawls Classifier

:

- Minimize the error rate on the most disadvantaged sensitive sub-population, i.e., label $Y = i$, group $Z = j$

$$\arg\min_{f} \max_{i,j} \Pr\left(f(X) \neq Y | \, Y = i, Z = j\right)$$

- *Theoretical property*: Rawlsian classifier is a threshold classifier on an ideal score function (not equal to $\eta(x)$)
- Our characterization holds even under different weights for different groups and classes

Our Contribution

- Characterization of a fair classifier under Pareto efficiency and Rawlsian least-difference principle
- Rawlsian fair adaptation to learn a threshold (on score) or a linear threshold classifier (on embedding)
- Experiments on real and synthetic datasets show that Rawlsian fair adaptataion is comparable or better than group-fair classifiers trained on the entire data

Fair Adaptation Method

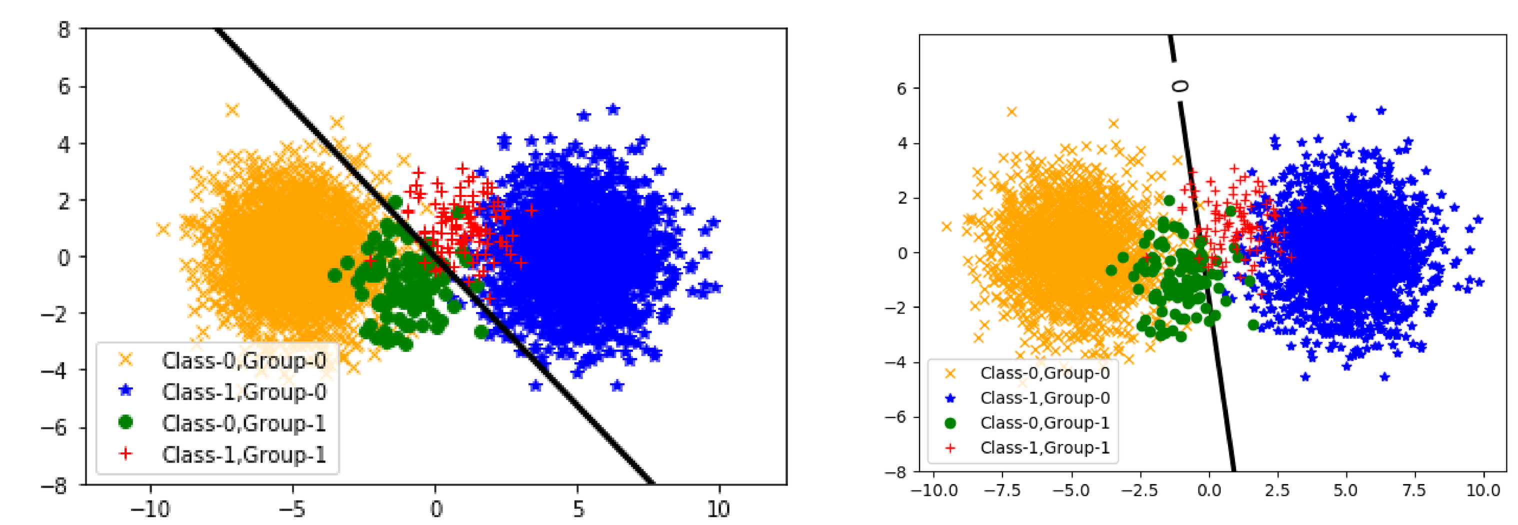
- Difficult to fix or adapt deep learning classifiers for fairness with limited access to training data
- Most deep learning models give scores or feature embeddings that are difficult to retrain
- **Problem:** Restricted Rawls classifier to learn a threshold (on score) or a linear threshold classifier (on embedding) using only 2nd order statistics of score or feature distribution over label $Y = i$, group $Z = j$.
- **Solution:** Formulate using ambiguous chance constraints, define and solve a convex optimization

References

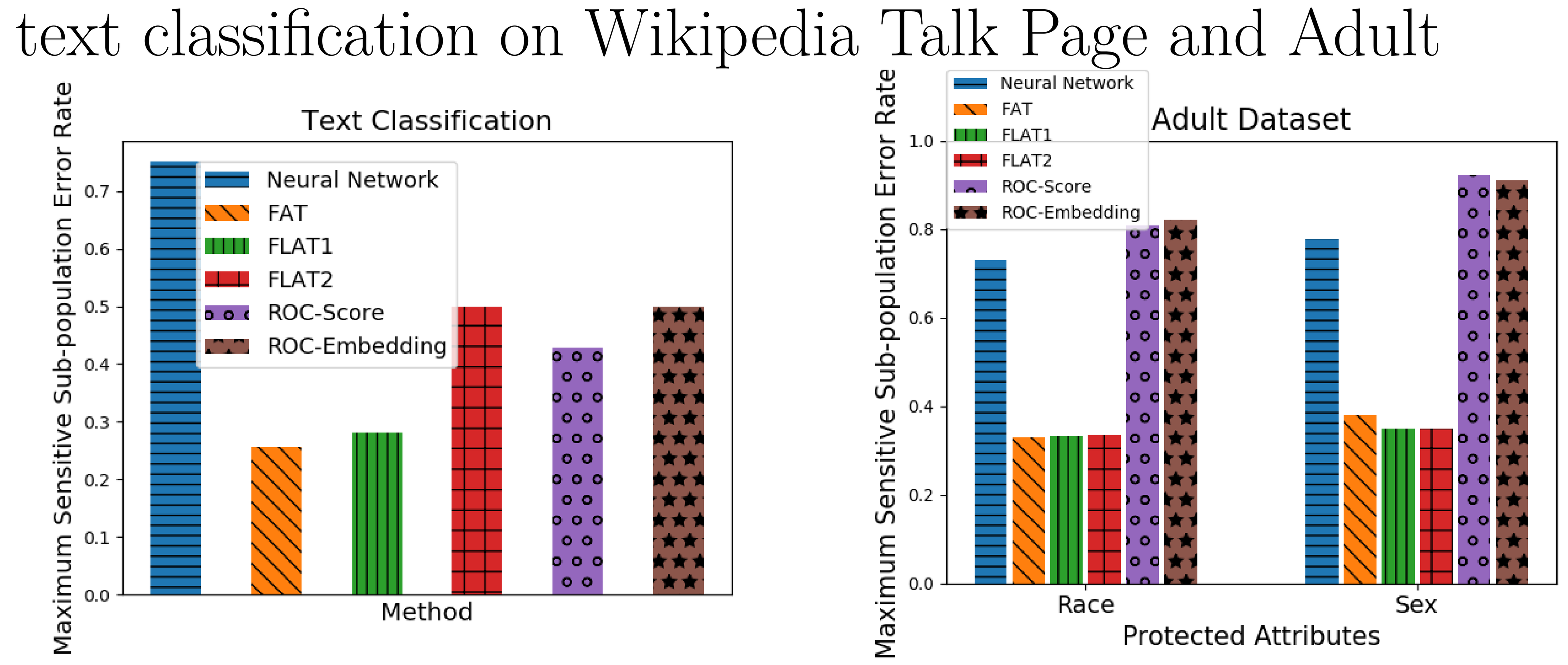
1. Algorithmic Decision Making and the Cost of Fairness, Corbett-Davies et al.
2. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees, Celis et al.

Experimental Results

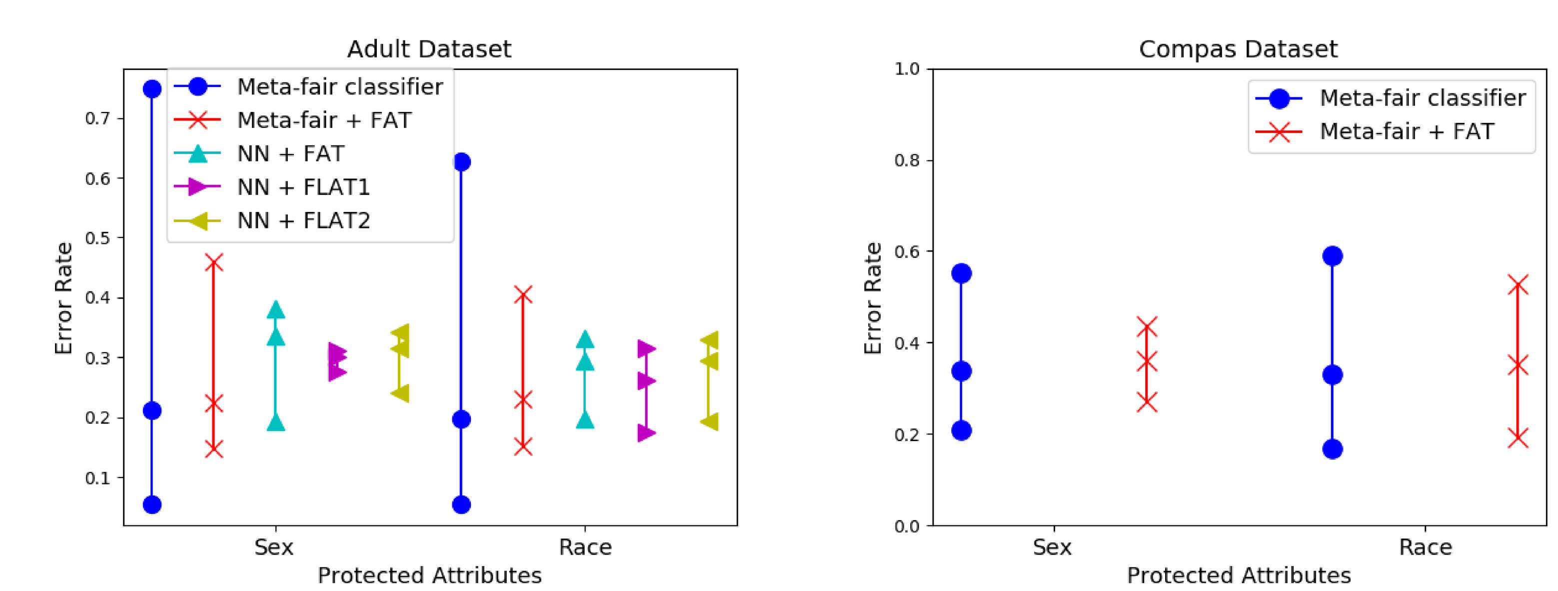
- Comparison of decision boundary on synthetic data
  - Left fig.: Decision boundary of our method
  - Right fig.: Decision boundary of Meta-fair classifier [2]



- Comparison of maximum group-wise class-wise error for text classification on Wikipedia Talk Page and Adult



- Comparison of FPR and FNR on Adult and COMPAS
  - Maximum (top point), average (middle point), minimum (bottom point) for FPRs and FNRs across all groups



- Our method (trained on 2nd order stats) significantly outperforms baselines (trained on complete data)