Accounting for Model Uncertainty in Algorithmic Discrimination
Junaid Ali, Preethi Laho*, Krishna P. Gummadi
Max Planck Institute for Software Systems, Max Planck Institute for Informatics*

1. Motivation: Limitation of existing fairness approaches
• Current group fairness methods treat all errors equally
• Our proposal: Account for types of uncertainty
  - Aleatoric uncertainty (irreducible) due to inherent noise or stochasticity in the task, e.g., overlapping classes
  - Model uncertainty a.k.a. epistemic uncertainty (reducible) due lack of knowledge about the best model or lack of data

2. Our proposal
Equalize only epistemic errors (E)
Any datapoint could be affected

3. Characterizing model uncertainty
Idea: Use existing methods on predictive multiplicity to identify errors due to model uncertainty

4. Fairness under model uncertainty
Idea: Reuse the highly accurate classifiers used to identify the ambiguous region
Approach: Stochastically pick the classifiers to minimize disparity in group error rates in the ambiguous region.

5. Key Contributions
• Key idea: only equalize errors occurring due to model uncertainty.
  - Formalize this problem
  - Convex formulation to equalize epistemic errors
• Scalable convex proxies to capture predictive multiplicity
  - For linear/nonlinear classifiers unlike the state-of-the-art
  - Equally good as the state-of-the-art in identifying the ambiguous regions
  - 4 orders of magnitude faster than the state-of-the-art
• Empirical results using SQF dataset, COMPAS dataset and a synthetic dataset

6. Experimental Results
• Synthetic dataset: Group fair classifier makes several unjustifiable mistakes to equalize all errors.
• Please refer to the paper for detailed results.
• Our fairness method only equalizes errors in the regions more prone to model uncertainty.
• We only change decisions of the datapoints whose decisions are ambiguous or uncertain in the first place.
• Existing fairness methods could lead to trading-off unfairness in different regions.
• Our method equalize errors only in the ambiguous regions while being highly accurate.