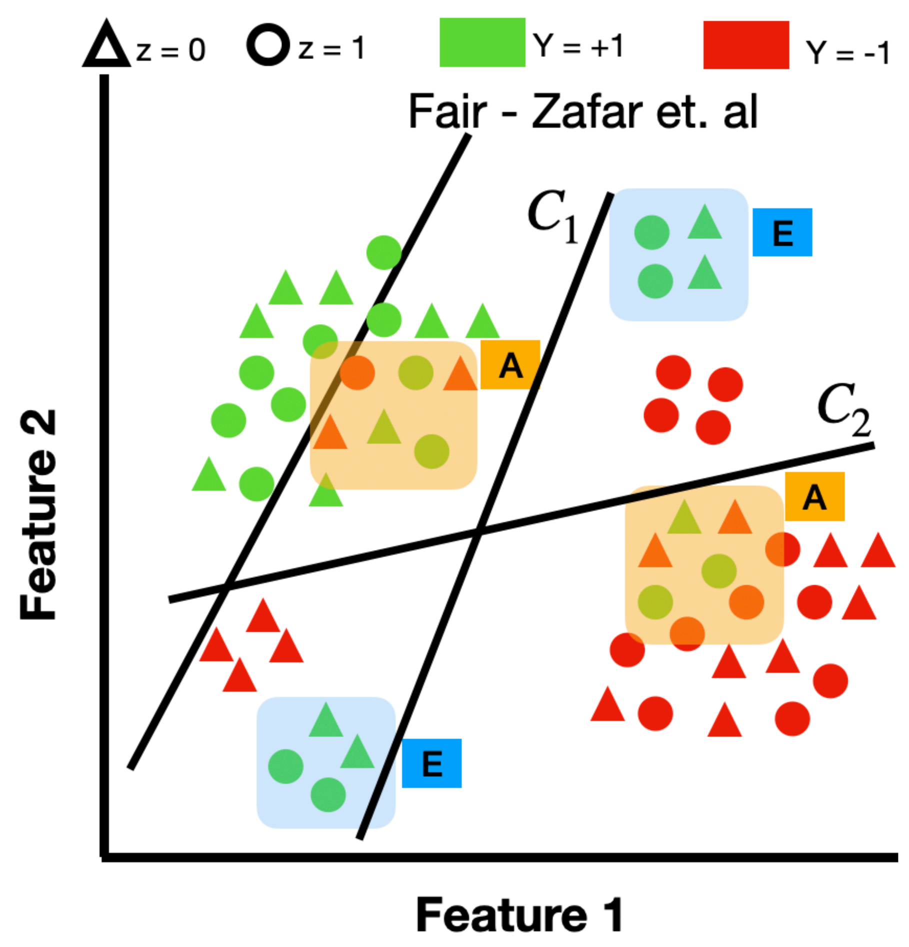
Accounting for Model Uncertainty in Algorithmic Discrimination

Junaid Ali, Preethi Lahoti*, 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
- 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



Aleatoric Errors (A): i.e., due to inherent noisy data (Region A)

Epistemic errors (E): i.e., due to lack of data, or lack of knowledge about the model. (Region E)

Existing methods: equalize all errors (A & E)

↓

Any datapoint could be affected

2. Our proposal

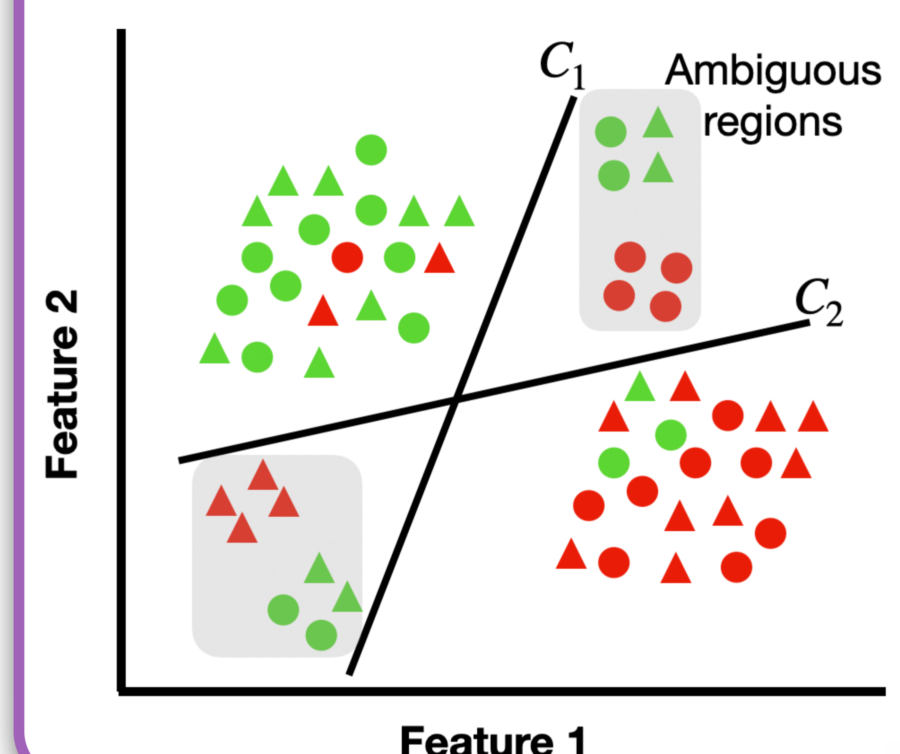
Equalize only epistemic errors (E) → Only the datapoints whose decisions are uncertain due to methodological limitations are affected

Key Idea

Ignore errors due to inherent noise. Focus only on the errors occurring due model uncertainty.

3. Characterizing model uncertainty

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



Predictive multiplicity
Classifiers C1 and C2 are equally accurate classifiers that disagree on a subset of the data (Ambiguous region).

Assumption:
Hypothesis class for finding the classifiers is sufficiently complex.

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.

$$minimize_w \quad \left| \sum_{\theta \in C} w_\theta \cdot (Err_{z=1}(\theta) - Err_{z=0}(\theta)) \right|$$

$$st \quad 0 \leq w_\theta \leq 1 \quad \text{and} \quad \sum_{\theta} w_\theta = 1$$

C: is the set of classifiers exhibiting predictive multiplicity

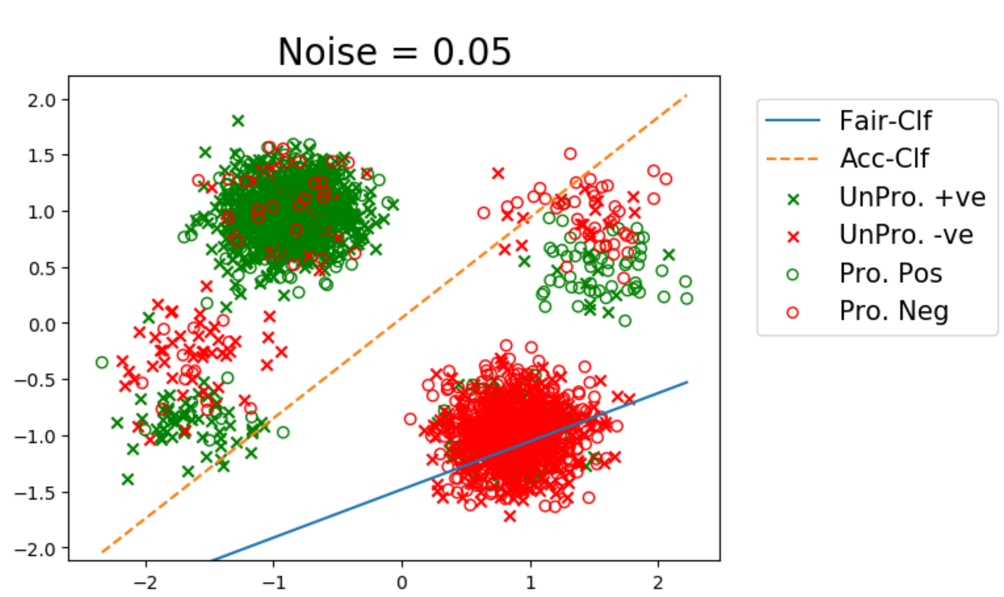
Err: False positive rate or false negative rates in ambiguous regions

z: represents the sensitive attribute

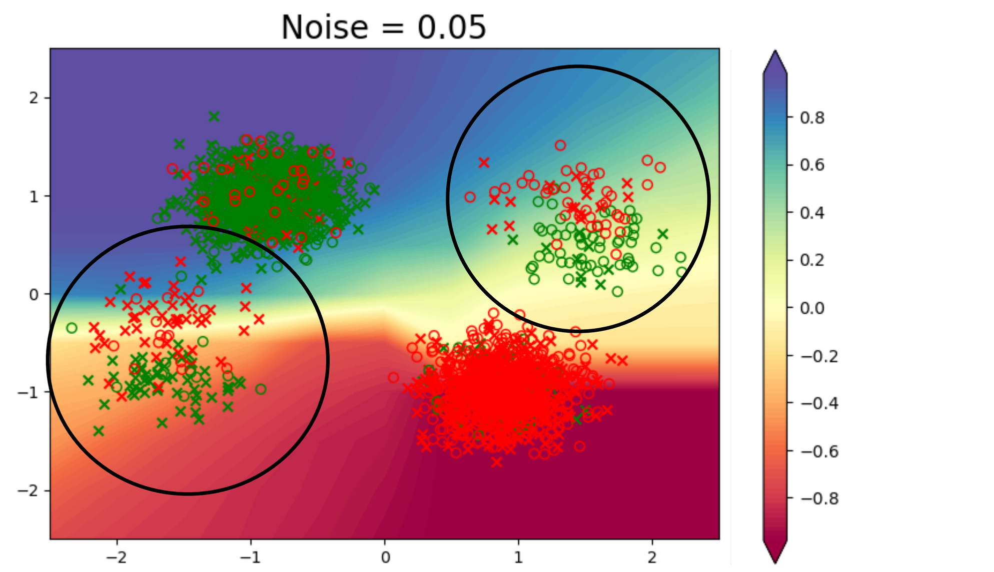
5. Key Contributions

- **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



Color represents the expected predicted class



Synthetic dataset: Equalizing FPR/FNR

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.13/-0.14	0.05/-0.06	0.46/-0.45	0.89
Fair	0.03/-0.02	0.05/-0.06	-0.14/0.29	0.77/0.89
Uniform	0.04/-0.04	0.05/-0.06	-0.22/0.20	0.89 / 0.89
Ours	0.07/-0.07	0.05/-0.06	**0.0/-0.01**	0.89/0.89

Compas dataset: Equalizing FPR/FNR

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.19/0.33	-0.24/0.54	-0.11/0.15	0.66
Fair	0.02/0.03	-0.24/0.54	0.34/0.-0.42	0.66/0.65
Uniform	-0.19/0.34	-0.24/0.54	-0.11/0.15	0.66/ 0.66
Ours	-0.14/0.26	-0.24/0.54	**-0.01/0.03**	0.66/ 0.66

- 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.