A Convex Framework for Fair Regression

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, A. Roth

Penn
University *of* Pennsylvania

Motivation

- Machine learning (ML) increasingly used to make critical decisions, e.g. hiring and sentencing
- Problem: there are many examples of ML that is discriminatory or *unfair*
- There is a large body of work on fair *classification*; we instead focus on fair *regression*

Fairness Definitions

- Adapts idea that similar individuals (similar ground-truth label) should be treated similarly (similar predicted label) [Dwork et. al.] by introducing sample fairness penalties
- Individual Fairness penalty:

$$\mathbf{f_1(w, S)} = \frac{\mathbf{1}}{\mathbf{n_1 n_2}} \sum_{\substack{(x_i,y_i)\in S_1 \\ (x_j,y_j)\in S_2}} \mathbf{d(y_i, y_j)\left(w \cdot x_i - w \cdot x_j\right)^2}$$

  - ▷ Each pair of similar examples classified dissimilarly adds loss – no "cancellation", most stringent fairness requirement
- Group Fairness penalty:

$$\mathbf{f_2(w, S)} = \left[\frac{\mathbf{1}}{\mathbf{n_1 n_2}} \sum_{\substack{(x_i,y_i)\in S_1 \\ (x_j,y_j)\in S_2}} \mathbf{d(y_i, y_j)\,(w \cdot x_i - w \cdot x_j)}\right]^2$$

  - ▷ Pairs of similar examples classified dissimilarly can be cancelled out by pairs classified dissimilarly in the opposite direction, least stringent fairness requirement
- Hybrid Fairness: cancellation only among cross-pairs within "buckets" – interpolates between individual and group fairness
- Fairness loss minimized by constant predictors, but this incurs bad accuracy loss
  - ▷ How to trade off accuracy and fairness losses?

The Optimization Problem

- Overall loss function to minimize is

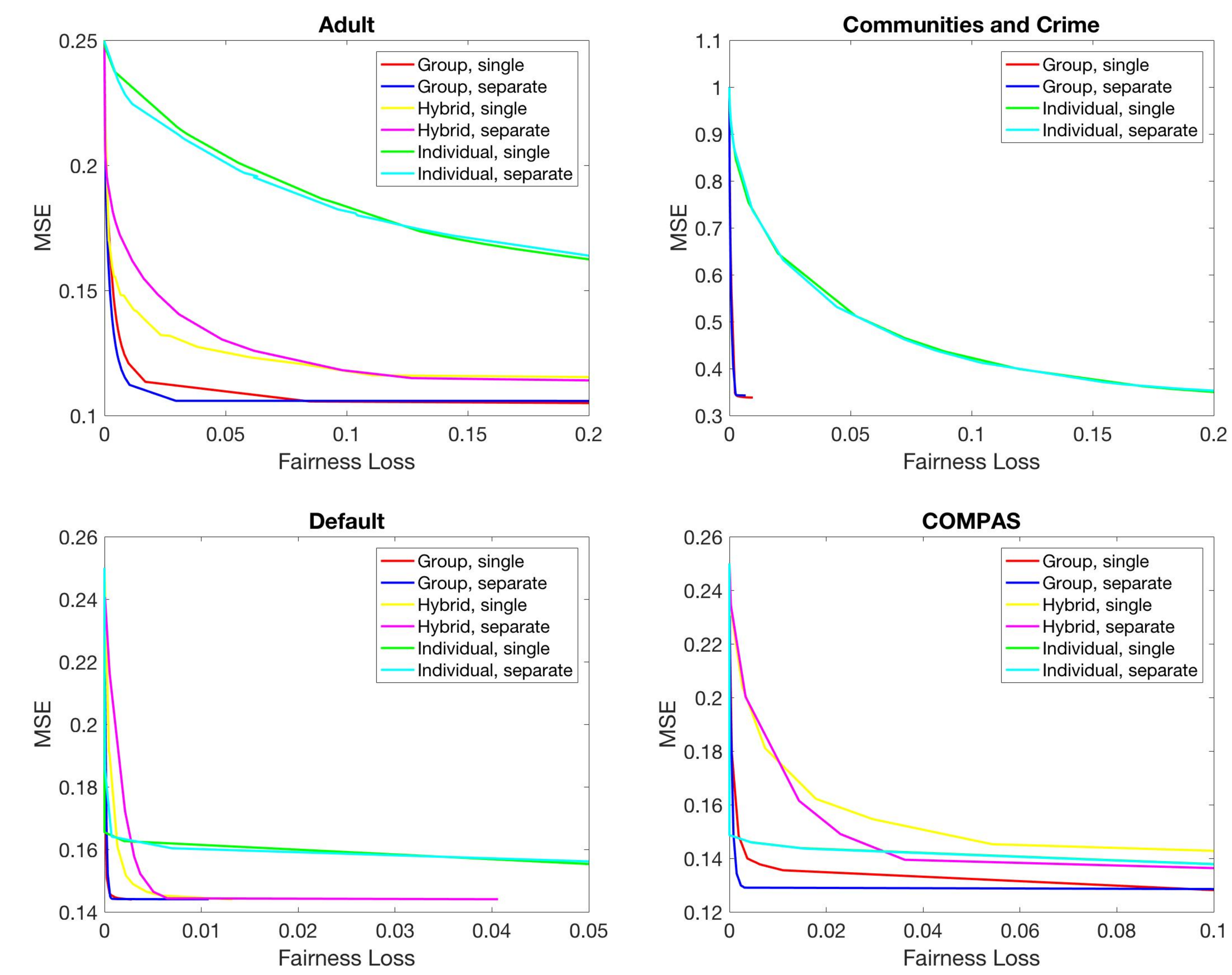
$$\min_{\mathbf{w}} \mathbb{E}_{(x,y)\sim\mathcal{P}}[(\mathbf{w \cdot x - y})^2] + \lambda \mathbf{f(w)} + \alpha(\lambda)\|\mathbf{w}\|_2$$

- Accuracy loss + fairness loss + ℓ_2 regularizer
- Benefit: convex optimization problem $\Rightarrow$ tractable

Summary of Datasets

Data Set	Type	n	d	Minority	Protected
Adult	logit	32561	14	10771	gender
Comm. & Crime	linear	1994	128	227	race
COMPAS	logit	3373	19	1455	race
Default	logit	30000	24	11888	gender

Pareto Curves



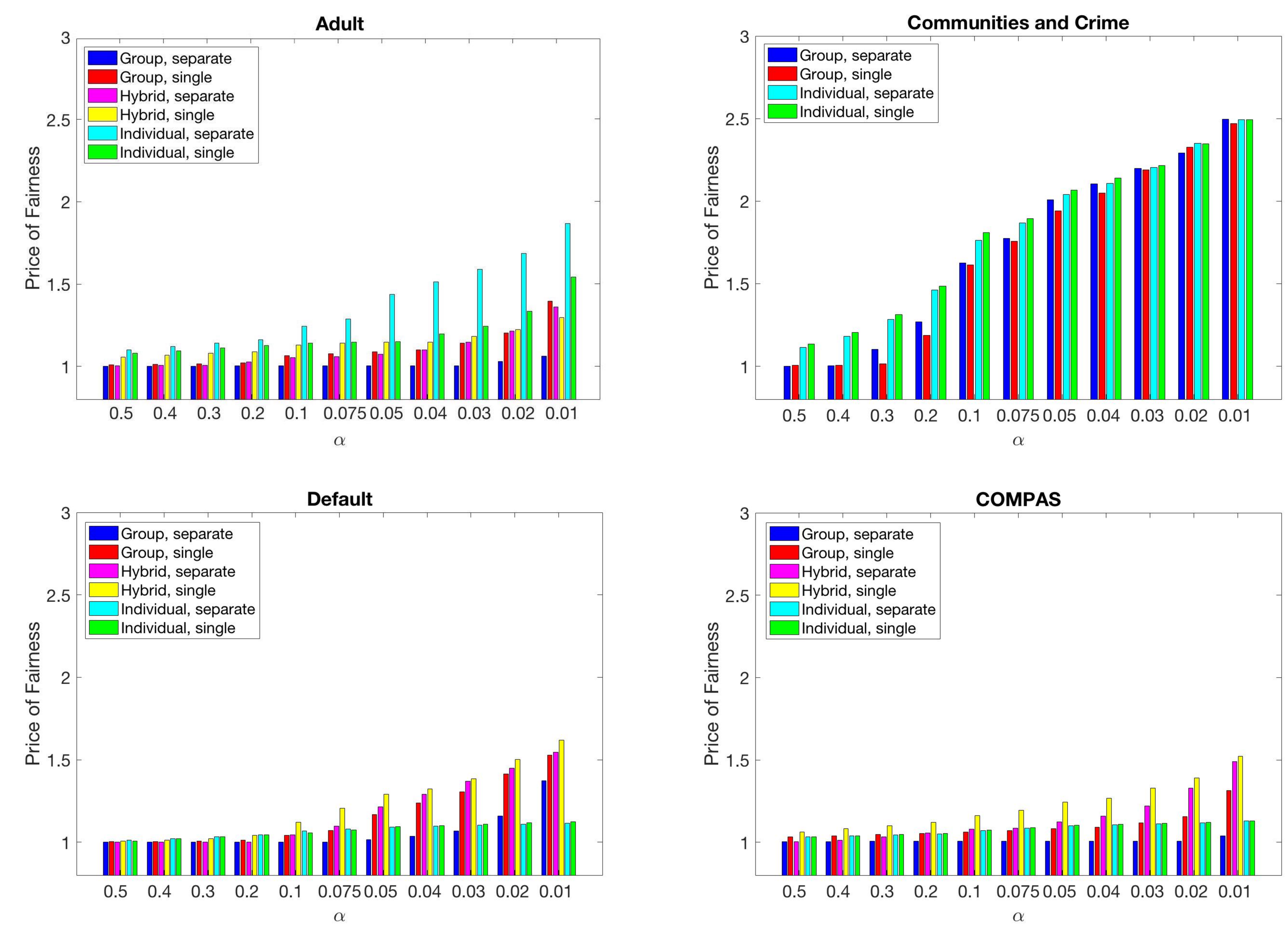
Quantitative Measure of Trade-off

- Price of Fairness

$$\mathrm{PoF}(\alpha) = \frac{\min_{\mathbf{w}} \mathbf{err(w)} \text{ subject to } \mathbf{f(w)} \le \alpha \mathbf{f(w^*)}}{\mathbf{err(w^*)}}$$

- The increment in error for any given fairness level of α compared to the best unfair predictor

Price of Fairness Curves



Takeaways

- Notion of fairness that's tractable to optimize
- The detailed trade-offs between fairness and accuracy and different notions of fairness appear to be quite data-dependent and lack *universals*
- Possibly consistent with emerging theoretical literature demonstrating the lack of a unified, comprehensive fairness definition