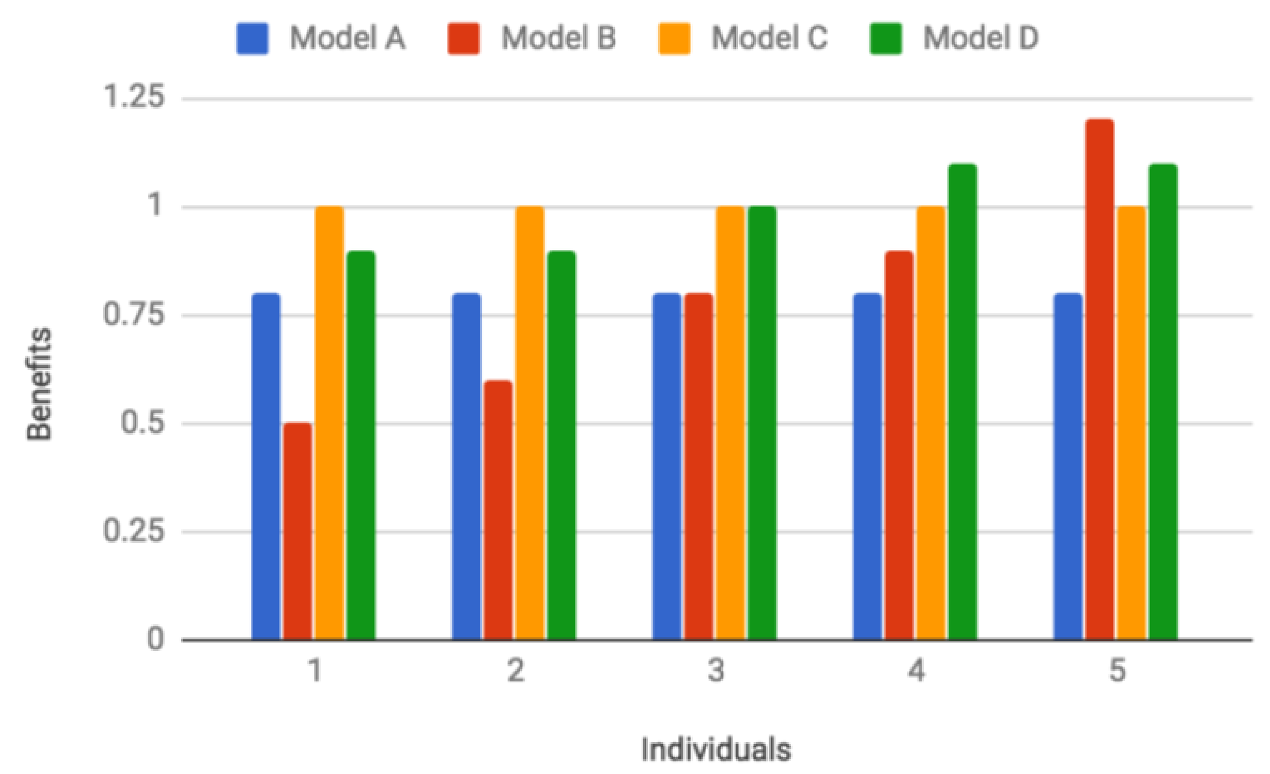
Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making

HODA HEIDARI CLAUDIO FERRARI KRISHNA GUMMADI ANDREAS KRAUSE
ETH Zürich, MPI-SWS

Fairness = Equality?

- The "leveling down" objection to equality
- Example: 5 individuals, 4 predictive models, different benefit distributions



- **According to inequality:** $C > D$ and $A > B$ and $A > D$ (!!!)
- **According to our measure:** $C > D > A > B$

Benefit Function

- $\mathbf{x}_i \in \mathcal{X}$ is the feature vector for individual i
- $y_i \in \mathcal{Y}$, the ground truth label for him/her
- $\hat{y}_i = h(\mathbf{x}_i)$ prediction for i
- $b(y, \hat{y})$ the benefit obtained by an individual with true label y and predicted label $\hat{y}$.
- We assume $b(y, \hat{y})$ linear in $\hat{y}$ (WLOG for binary classification!). E.g.

$$b_i = \hat{y}_i - y_i + 1$$

Fairness Behind a Veil of Ignorance

- Core idea: social welfare as fairness behind a veil of ignorance
- Axiomatic characterization:
  - **Monotonicity:** $\mathbf{b}' \succ \mathbf{b} \Rightarrow \mathcal{W}(\mathbf{b}') > \mathcal{W}(\mathbf{b})$.
  - **Independence of unconcerned agents:** $\forall \mathbf{b}, \mathbf{b}', a, c$,
  $$(\mathbf{b}|^i a) \succeq (\mathbf{b}'|^i a) \Leftrightarrow (\mathbf{b}|^i c) \succeq (\mathbf{b}'|^i c).$$
  - **Independence of common scale:** $\forall c > 0$,
  $$\mathcal{W}(\mathbf{b}) \geqslant \mathcal{W}(\mathbf{b}') \Leftrightarrow \mathcal{W}(c\mathbf{b}) \geqslant \mathcal{W}(c\mathbf{b}').$$
  - **Anonymity**
  - **Progressive transfers principle**
- According to Debreu-Groman Theorem, $\mathcal{W}_\alpha(b_1, \ldots, b_n) = \sum_{i=1}^n w_\alpha(b_i)$, where
  - for $0 < \alpha \leqslant 1$, $w_\alpha(b) = b^\alpha$;
  - for $\alpha = 0$, $w_\alpha(b) = \ln(b)$;
  - for $\alpha < 0$, $w_\alpha(b) = -b^\alpha$

A Convex Formulation

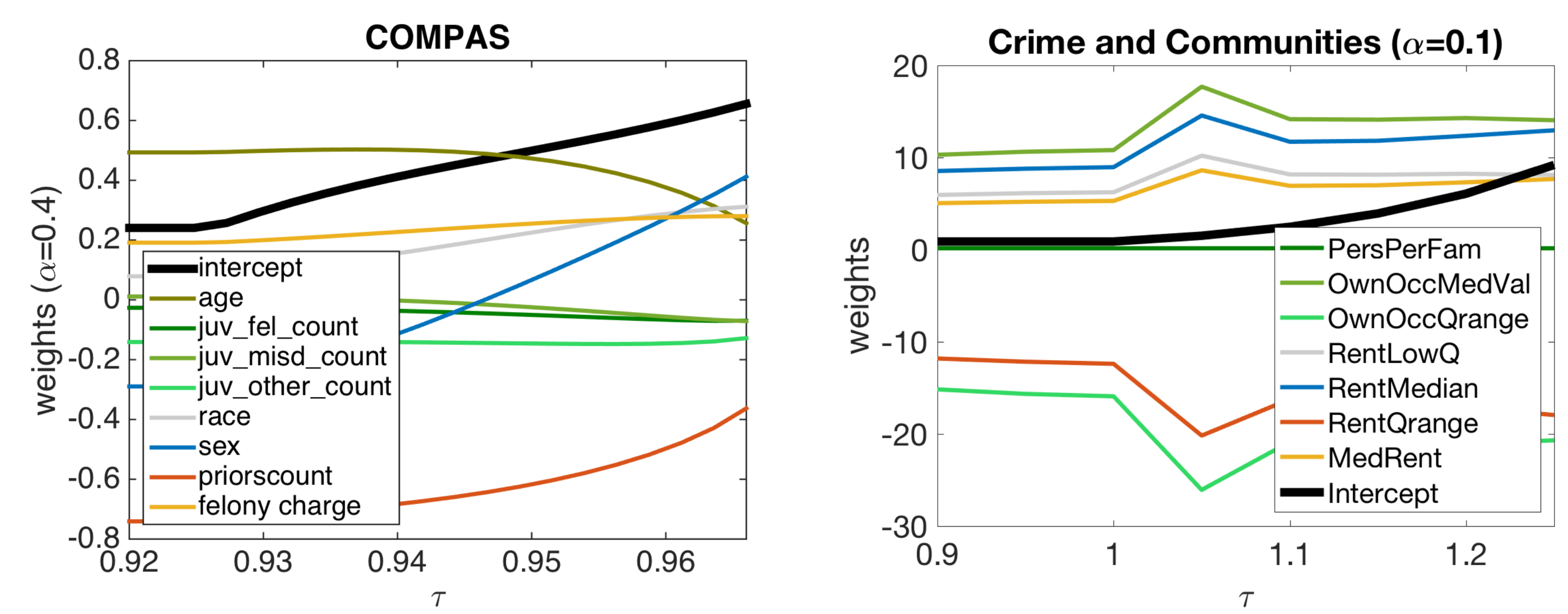
- Our formulation:

$$\min_{h \in \mathcal{H}} \mathcal{L}(h, D) \text{ s.t. } \mathcal{W}_\alpha(\mathbf{b}) \geqslant \tau$$

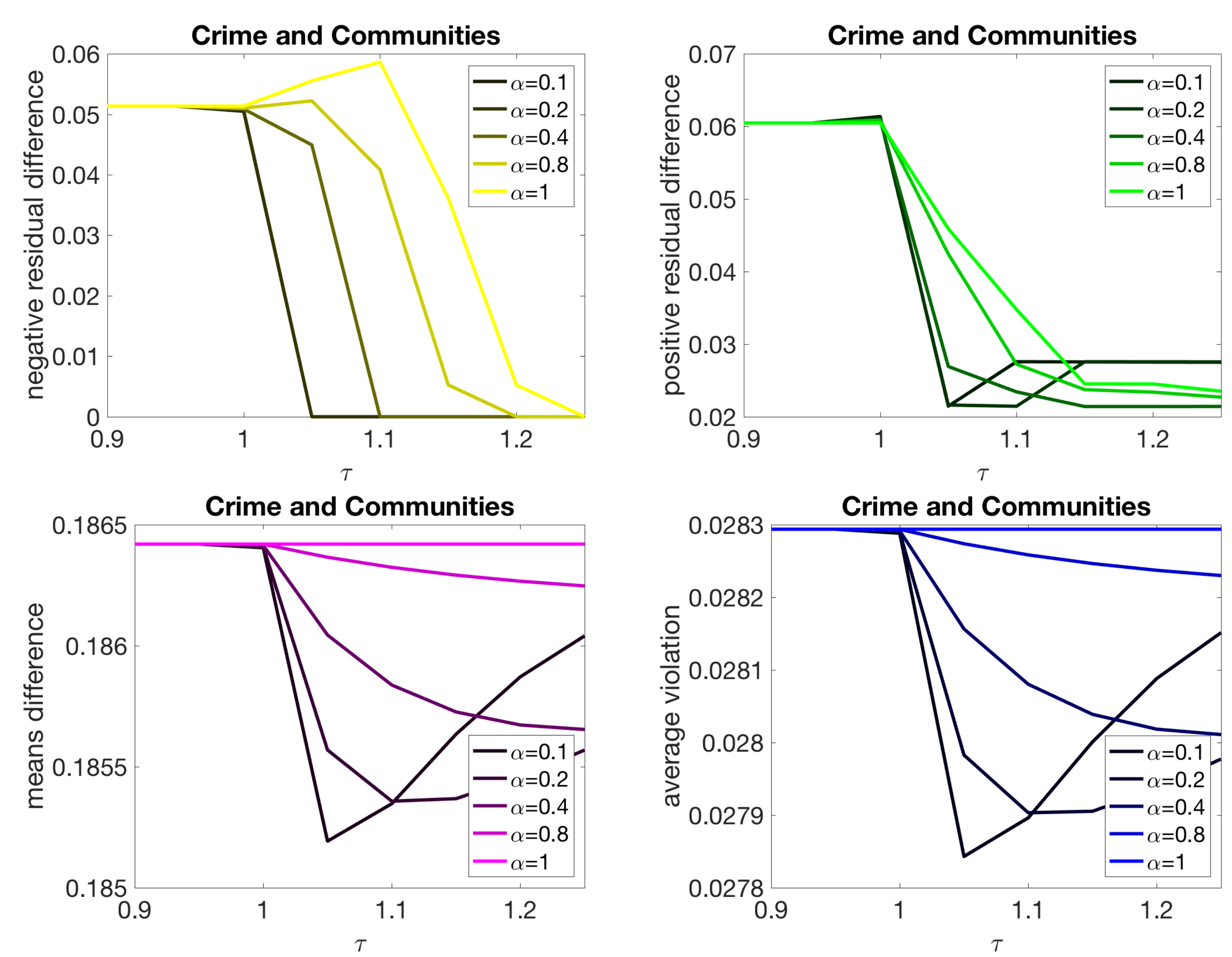
- Linear regression

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (\theta.\mathbf{x}_i - y_i)^2$$
$$\text{s.t. } \frac{1}{n} \sum_{i=1}^n (\theta.\mathbf{x}_i - y_i + 1)^\alpha \geqslant \tau$$

- Impact of our in-processing on model parameters:



Impact on Previous Notions of Fairness

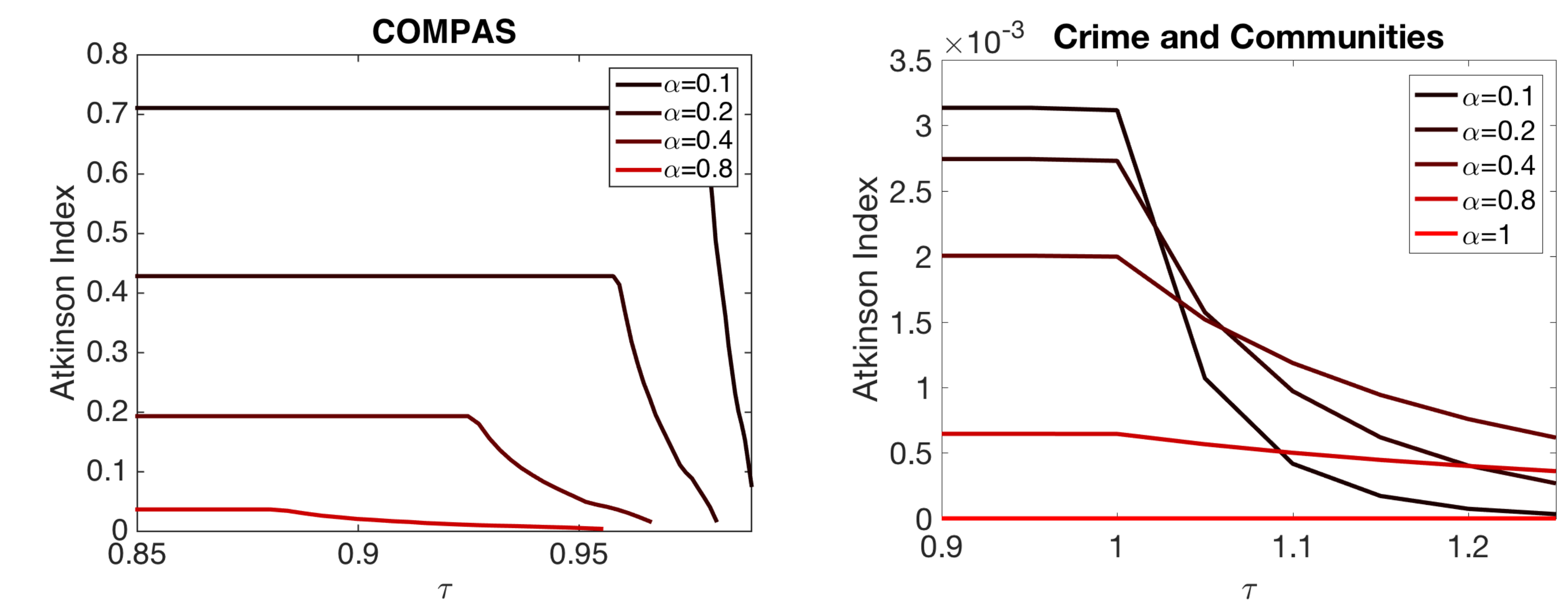


Connection to Inequality

- Atkinson Index is a *welfare*-based measure of inequality

$$A_\beta(b_1, \ldots, b_n) = 1 - \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^n b_i^{1-\beta}\right)^{1/(1-\beta)} \quad \text{for } 0 \leqslant \beta \neq 1$$

- μ, the mean benefit
- compared with the Equally Distributed Equivalent (EDE)



Proposition:
Consider two benefit vectors $\mathbf{b}, \mathbf{b}' \succ \mathbf{0}$ *with equal means* $(\mu = \mu')$. *For* $0 < \alpha < 1$, $A_{1-\alpha}(\mathbf{b}) \geqslant A_{1-\alpha}(\mathbf{b}')$ *if and only if* $\mathcal{W}_\alpha(\mathbf{b}) \leqslant \mathcal{W}_\alpha(\mathbf{b}')$.

- For a fixed mean benefit μ, our measure and Atkinson index $\Rightarrow$ the same indifference curves and total ordering.

Summary

Cardinal social welfare as a measure of fairness behind a veil of ignorance

- Addresses the leveling down objection to inequality
- Enjoys a convex formulation
- Often limits individual level inequality
- Previous notions only characterize *conditions* of fairness
- Our work: a principled way of generalizing to more complicated settings
  - Beyond binary classification
  - More than one group
- Useful for measuring both individual and group level fairness

Future Directions

- Extension to other learning tasks
- Extension to descriptive (as opposed to normative) behavioral theories
- Human perception of fairness in the context of automated decision making
- What is the right benefit function?
- ...

hheidari@inf.ethz.ch