Algorithms and Analysis for Optimizing Robust Objectives in Fair Machine Learning

Cyrus Cousins
cbcousins@umass.edu
University of Amherst Massachusetts
Department of Computer Science

December 2023
New York, New York
Fairness in Operations and AI Workshop
Columbia University

What is Welfare-Centric Fair Machine Learning?

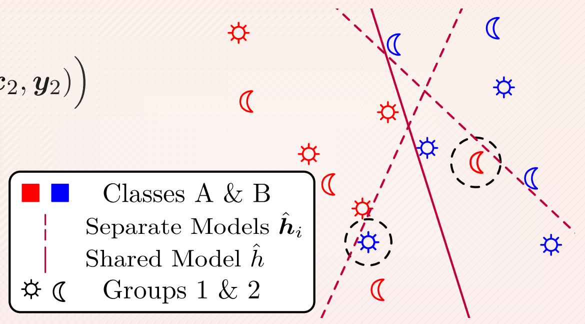
- ♣ Fair machine learning considers *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$
- ♣ We can handle each group individually
  - ♢ *Empirical utility maximization*

$$\hat{\mathrm{U}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i) \doteq \frac{1}{m}\sum_{j=1}^{m} \mathrm{U}\big(h(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \quad \forall i: \ \hat{\boldsymbol{h}}_i \doteq \operatorname*{argmax}_{h\in\mathcal{H}} \mathrm{U}(h; \boldsymbol{x}_i, \boldsymbol{y}_i)$$

- ♣ What is the best classifier *overall?*
  - ♢ *Empirical welfare maximization*

$$\hat{h} \doteq \operatorname*{argmax}_{h\in\mathcal{H}} \mathrm{W}\left(\hat{\mathrm{U}}(h; \boldsymbol{x}_1, \boldsymbol{y}_1), \hat{\mathrm{U}}(h; \boldsymbol{x}_2, \boldsymbol{y}_2)\right)$$

- ♣ Welfare functions encode *social values*
  - ♢ Optimize a *given welfare function*
  - ♢ Objectives specify tradeoffs!



Power Means and the Social Planner's Problem

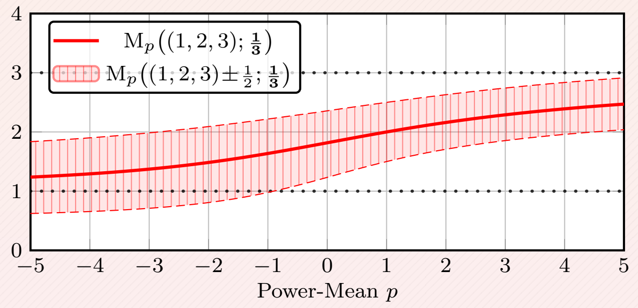
- ♣ A *social planner* arranges society to the benefit of all
- ♣ How should we aggregate utility or disutility across groups?
- ♣ The power-mean for $p \in \mathbb{R}$ summarizes g *values* $\boldsymbol{s}_{1:g}$ with *weights* $\boldsymbol{w}_{1:g}$ as

$$\mathrm{M}_{p\neq 0}(\boldsymbol{s}; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^p}$$

for $p \neq 0$, or

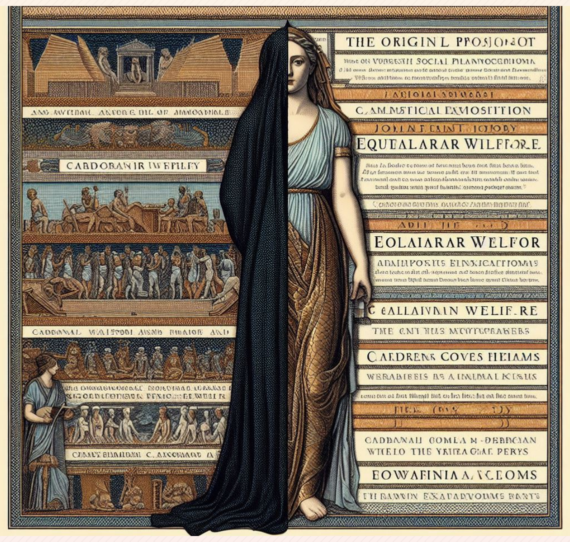
$$\mathrm{M}_0(\boldsymbol{s}; \boldsymbol{w}) \doteq \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \log(\boldsymbol{s}_i^p)\right)$$

- ♣ Fair welfare requires $p \leq 1$, extremes are interesting special cases
  - ♢ $p = 1$ is *weighted sum*, a.k.a. utilitarian welfare, over groups (well-studied case)
  - ♢ $p = 0$ is the *Nash social welfare* over groups
  - ♢ $p = -\infty$ limit is the *minimum* over groups (egalitarian or robust optimization)
- ♣ Power-means are:
  1. *Axiomatically Justified*
  2. *Interpretable*
     $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})$ units match $\boldsymbol{s}_{1:g}$
  3. *Stochastically Stable*
     (for $p \in [-\infty, 0) \cup [1, \infty]$)



John Rawls, the Original Position, and the Veil of Ignorance

- ♣ Justice, fairness, and societal wellbeing should be *objective concepts*
  - ♢ Should not depend on *our own identities*
- ♣ The "original position argument"
  - ♢ We should assess a situation from behind a "veil of ignorance"
- ♣ Rawls argues for *worst-case* robust or pessimistic analysis
  - ♢ Egalitarian welfare (malfare) is born!
  - ♢ Given utility $\boldsymbol{s} \in \mathbb{R}^g$, assess welfare as $\mathrm{M}_{-\infty}(\boldsymbol{s}) = \min_{i=1}^{g} \boldsymbol{s}_i$
  - ♢ Given disutility $\boldsymbol{s} \in \mathbb{R}^g$, assess malfare as $\mathrm{M}_{\infty}(\boldsymbol{s}) = \max_{i=1}^{g} \boldsymbol{s}_i$



This work: We pose the original problem setting as an *adversarial game*

- ♣ Robust fair objectives arise as solution concepts against specific adversaries

Egalitarianism and Adversarial Games

The Rawlsian Game

g: *Number of inhabitants* in the game world
Θ: *Parameter space* of worlds Dæmon can create
$\boldsymbol{s}(\theta): \Theta \to \mathbb{R}^g_{0+}$: (Dis)utility vector of the inhabitants of world θ
$\mathcal{S} \doteq \{\boldsymbol{s}(\theta) \,|\, \theta \in \Theta\} \subseteq \mathbb{R}^g_{0+}$: Set of *feasible (dis)utility vectors* the Dæmon can create
$\mathcal{A}_{\mathrm{D\ae}} \doteq \mathcal{S}$: Dæmon's action space
$\mathcal{A}_{\mathrm{Ang}} \doteq \mathcal{W} \subseteq \triangle_g$: Angel's action space
$\mathrm{P}(\boldsymbol{s}, i) \doteq \pm\langle \boldsymbol{s}_i, -\boldsymbol{s}_i\rangle$: 0-sum payoff function

Strategic gameplay (Dæmon goes first):

$$\arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{A}_{\mathrm{D\ae}}} \operatorname*{\substack{\min\\\max}}_{i\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\boldsymbol{s}, i) = \arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{S}} \operatorname*{\substack{\min\\\max}}_{i\in 1,\ldots,g} \boldsymbol{s}_i$$

Weaker Adversaries, Robust Utilitarianism, and the Gini Social Welfare

The Constrained Rawlsian Game

$g, \Theta, \boldsymbol{s}(\theta), \mathcal{S}$: As above
$\mathcal{A}_{\mathrm{D\ae}} \doteq \mathcal{S}$: Dæmon's action space
$\mathcal{A}_{\mathrm{Ang}} \doteq \mathcal{W} \subseteq \triangle_g$: Angel's action space (*inhabitant* i becomes *distribution* $\boldsymbol{w}$)
New 0-sum payoff function uses *expected (dis)utility:*

$$\mathrm{P}(\boldsymbol{s}, \boldsymbol{w}) \doteq \langle \boldsymbol{w}\cdot\boldsymbol{s}, -\boldsymbol{w}\cdot\boldsymbol{s}\rangle$$

Strategic gameplay (Dæmon goes first):

$$\arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{A}_{\mathrm{D\ae}}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{S}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s}$$

A Few Special Cases

- ♣ Constant weights $\mathcal{W} = \{\boldsymbol{w}^*\}$: Optimize *utilitarian welfare* $\mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}^*)$
- ♣ Norm ball $\mathcal{W} = \boldsymbol{w}^* + \{\boldsymbol{w} \,|\, \|\boldsymbol{w}\| \leq \gamma\}$: Optimize robust utilitarian welfare
- ♣ Weights *bounded from below* $\mathcal{W} \doteq \{\boldsymbol{w} \in \triangle_g \,|\, \boldsymbol{w} \succeq \gamma\boldsymbol{w}^*\}$: Optimize *utilitarian-maximin social welfare function* (UMSWF), i.e.,

$$\operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}) = \gamma\mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}^*) + (1-\gamma)\mathrm{M}_{\mp\infty}(\boldsymbol{s})$$

- ♣ Given some *sorted weights vector* $\boldsymbol{w}^{\downarrow}$ or $\boldsymbol{w}^{\uparrow}$, let $\boldsymbol{s}^{\uparrow}$ denote *sorted* $\boldsymbol{s}$
  - ♢ $\mathrm{M}_{\boldsymbol{w}^{\downarrow}}(\boldsymbol{s}) \doteq \boldsymbol{w}^{\downarrow}\cdot\boldsymbol{s}^{\uparrow}$ is *Gini social welfare*, and $\mathrm{M}_{\boldsymbol{w}^{\uparrow}}(\boldsymbol{s}) \doteq \boldsymbol{w}^{\uparrow}\cdot\boldsymbol{s}^{\uparrow}$ is *Gini social malfare*
  - ♢ If Angel has actions $\mathcal{W} \doteq \{\pi(\boldsymbol{w}^{\downarrow}) | \pi \in \Pi_g\}$, where Π_g is all permutations $\{1, \ldots, g\}$:

$$\max_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\boldsymbol{w}^{\downarrow}}(\boldsymbol{s}) \quad \text{or} \quad \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\boldsymbol{w}^{\uparrow}}(\boldsymbol{s})$$

A Divine Symmetry: Welfare and Malfare as Game-Theoretic Equilibria

A Game of Dæmonic Justice

- ♣ Suppose Dæmon wants to optimize $\substack{\textit{welfare}\\\textit{malfare}}$ fairness concept $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*)$
- ♣ Angel remains adversarial over some $\mathcal{W}$
- ♣ Payoff function

$$\mathrm{P}(\boldsymbol{s}, \boldsymbol{w}) \doteq \big\langle \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}), -\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})\big\rangle$$

- ♣ Strategic gameplay (Dæmon goes first):

$$\arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{A}_{\mathrm{D\ae}}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{S}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})$$

  - ♢ Optimizes a *robust power-mean!*

A Game of Angelic Justice

- ♣ Suppose Angel wants to optimize $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*)$ for some $p > 0$ with action space $\mathcal{W} = \triangle_g$
- ♣ We have the payoff function $\mathrm{P}(\boldsymbol{s}, \boldsymbol{w}) \doteq \big\langle \boldsymbol{w}\cdot\boldsymbol{s}, \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*)\big\rangle$
- ♣ Angel strategy: Play $\boldsymbol{w}_i \propto \boldsymbol{w}_i^* \boldsymbol{s}_i^{p-1}$
- ♣ Dæmon strategy: Select $\boldsymbol{s}$ to optimize $\boldsymbol{s}\cdot\boldsymbol{w} = \sum_{i=1}^{g} \boldsymbol{w}_i^* \boldsymbol{s}_i^p = \mathrm{M}_p^p(\boldsymbol{s}; \boldsymbol{w})$
- ♣ Play at this Nash equilibrium also optimizes a power-mean
- ♣ Angel can modify strategy to incorporate *robustness* $\boldsymbol{w}^* \in \mathcal{W}^*$

Fair Learning and Adversarial Optimization

We have reduced our fair and robust fair objectives to the common form

$$\arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{A}_{\mathrm{D\ae}}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{S}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})$$

> **Lemma 1** (Power-Mean Curvature)
> *For any $p \geq 1$, if $\boldsymbol{s}: \Theta \to \mathbb{R}^g$ is convex, then $\mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w}): (\Theta \times \triangle_g) \to \mathbb{R}_{0+}$ is convex in θ and concave in $\boldsymbol{w}$.*
> *For any $p \leq 1$, if $\boldsymbol{s}: \Theta \to \mathbb{R}^g$ is concave, then $\mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w}): (\Theta \times \triangle_g) \to \mathbb{R}_{0+}$ is concave in θ and convex in $\boldsymbol{w}$.*

Thus we can efficiently optimize robust fair objectives using *gradient* $\substack{\textit{ascent-descent}\\\textit{descent-ascent}}$ over $\boldsymbol{s} \in \mathcal{S}$

For indirect optimization tasks, we express the task in terms of parameter space Θ as

$$\arg \operatorname*{\substack{\max\\\min}}_{\boldsymbol{s}\in\mathcal{A}_{\mathrm{D\ae}}} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \arg \operatorname*{\substack{\max\\\min}}_{\theta\in\Theta} \operatorname*{\substack{\min\\\max}}_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w})$$

We can also efficiently optimize robust fair objectives using *gradient* $\substack{\textit{ascent-descent}\\\textit{descent-ascent}}$ over $\theta \in \Theta$

Robust Fair Objectives in Fair Machine Learning

- ♣ For any training point x, training label y, we define the *logistic loss* of model θ as

$$\ell\big(h_\theta(x), y\big) = \frac{1}{2}\big(1 + \tanh(y\, x\cdot\theta)\big)$$

- ♣ We then compute *disutility* for each group by averaging over training points

$$\boldsymbol{s}_i(\theta) = \frac{1}{\boldsymbol{m}_i}\sum_{j=1}^{\boldsymbol{m}_i} \ell\big(h_\theta(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big)$$

- ♣ Observe: $\boldsymbol{s}_i(\theta)$ is *strictly convex* over $\theta \in \Theta$: we can apply maximin methods
- ♣ Fair logistic regression objective is then

$$\operatorname*{argmin}_{\theta\in\Theta} \mathrm{M}\big(i \mapsto \boldsymbol{s}_i(\theta)\big) = \operatorname*{argmin}_{\theta\in\Theta} \sup_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_p\left(i \mapsto \frac{1}{\boldsymbol{m}_i}\sum_{j=1}^{\boldsymbol{m}_i} \ell\big(h_\theta(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \boldsymbol{w}\right)$$

Robust Fair Objectives in Fair Allocation

A Most Curious Game

Combining welfare or malfare objectives and robustness yields the *Gini-power-mean family*

$$\mathrm{M}_{\boldsymbol{w}^{\downarrow},p}(\boldsymbol{s}) \doteq \mathrm{M}_p(\boldsymbol{s}^{\uparrow}; \boldsymbol{w}^{\downarrow}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i^{\downarrow}(\boldsymbol{s}_i^{\uparrow})^p} \text{ for utility }, \quad \text{or}$$

$$\mathrm{M}_{\boldsymbol{w}^{\uparrow},p}(\boldsymbol{s}) \doteq \mathrm{M}_p(\boldsymbol{s}^{\uparrow}; \boldsymbol{w}^{\uparrow}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i^{\uparrow}(\boldsymbol{s}_i^{\uparrow})^p} \text{ for disutility}$$

- ♣ Generalizes power-mean and Gini families
  - ♢ Gini arises for $p = 1$
  - ♢ Power-mean (unweighted) for $\boldsymbol{w}^{\uparrow} = \frac{1}{g}\mathbf{1}$ or $\boldsymbol{w}^{\downarrow} = \frac{1}{g}\mathbf{1}$
- ♣ Arises from power-mean axioms and a robust original position game!