An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins ♡ ♢ Brown University ♣ ♠ Department of Computer Science

cyrus_cousins@brown.edu ♡ ♢ ♣ ♠ http://cs.brown.edu/people/ccousins/projects/fairness/home.html

NeurIPS 2021

35th Conference on Neural Information Processing Systems

Online • December 6–14, 2021

Welfare, Malfare, and the Power Mean

♣ The power-mean for $p \in \mathbb{R} \setminus \{0\}$ summarizes g *values* $\mathcal{S}_{1:g}$ with *weights* $\boldsymbol{w}$:

$$\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p} \ .$$

♣ Fair welfare: $p \le 1$, $p = \infty$ is *minimum* over groups (egalitarian optimization)

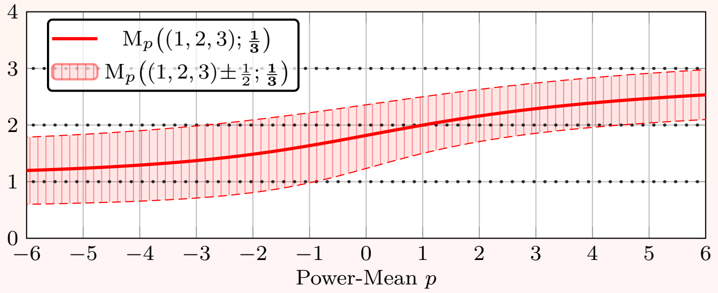
♠ Measure *overall wellbeing* given *utility values* (accuracy, income)

♣ Fair malfare: $p \ge 1$, $p = \infty$ is *minimum* over groups (robust minimization)

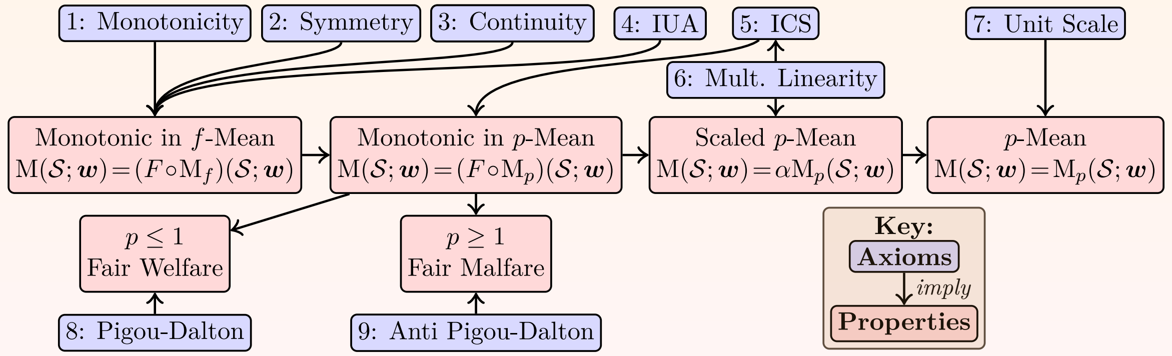
♠ Measure *overall illbeing* given *disutility values* (loss, harm)

♣ Power-means are:

1. *Axiomatically Justified*
2. *Interpretable*
   $\mathrm{M}_p(\mathcal{S};\boldsymbol{w})$ units match $\mathcal{S}_{1:g}$
3. *Stochastically Stable*
   (for $p \in [-\infty, 0) \cup [1, \infty]$)



Axioms of Cardinal Welfare



Relationships between aggregator function axioms and properties (cardinal welfare theory).

Estimating Malfare Values

1. Assuming only *monotonicity*:

   Suppose $\forall \omega \in \Omega : \hat{\mathcal{S}}(\omega) - \boldsymbol{\varepsilon}(\omega) \le \mathcal{S}(\omega) \le \hat{\mathcal{S}}(\omega) + \boldsymbol{\varepsilon}(\omega)$. Then

   $$\mathrm{M}_p(\mathbf{0} \vee (\hat{\mathcal{S}} - \boldsymbol{\varepsilon}); \boldsymbol{w}) \le \mathrm{M}_p(\mathcal{S};\boldsymbol{w}) \le \mathrm{M}_p(\hat{\mathcal{S}} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \ ,$$

   where $\boldsymbol{a} \vee \boldsymbol{b}$ denotes the (elementwise) maximum.

2. Suppose range r. Then with probability at least $1 - \delta$ over choice of $\boldsymbol{x}$:

   $$\left| \mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w}) \right| \le r \sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \ .$$

3. Suppose range r and variances $\mathbb{V}_{\mathcal{D}_i}[\ell]$. With probability at least $1 - \delta$:

   $$\left| \mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}};\boldsymbol{w}) \right| \le \frac{r \ln \frac{2g}{\delta}}{3m} + \max_{i \in 1,\dots,g} \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2g}{\delta}}{m}} \ .$$

N.b.: 2 & 3 hold for all fair malfare functions ($p \ge 1$), but *not all* fair welfare functions.

Empirical Malfare Minimization

Empirical risk and risk of hypothesis h given loss ℓ:

$$\hat{\mathrm{R}}(h;\ell,\boldsymbol{z}) \doteq \hat{\mathop{\mathbb{E}}_{(x,y)\in \boldsymbol{z}}}\left[\ell(y, h(x))\right] \quad \& \quad \mathrm{R}(h;\ell,\mathcal{D}) \doteq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y,h(x))\right] \ .$$

We define *empirical malfare minimization* (EMM), given $\mathrm{M}(\cdot;\boldsymbol{w})$, $\mathcal{D}_{1:g}$, and $\boldsymbol{z}_{1:g}$, with proxy and optimal models

$$\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}\Big(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i); \boldsymbol{w}\Big) \quad \& \quad h^* \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}\Big(i \mapsto \mathrm{R}(h;\ell,\mathcal{D}_i); \boldsymbol{w}\Big) \ .$$

Under what conditions is $\hat{h}$ a *good proxy* for h^*?

Theorem 1 (Uniform Convergence of Malfare)
Suppose fair malfare $\mathrm{M}_p(\cdot;\cdot)$ (i.e., $p \ge 1$), probability vector $\boldsymbol{w} \in \mathbb{R}_+^g$, loss function $\ell : (\mathcal{Y} \times \mathcal{Y}) \to [0, r]$, samples $\boldsymbol{z}_i \sim \mathcal{D}_i^m$, and hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$.
Then with probability at least $1 - \delta$ over choice of $\boldsymbol{z}$:

$$\sup_{h\in\mathcal{H}} \left| \mathrm{M}_p\big(i \mapsto \mathrm{R}(h;\ell,\mathcal{D}_i);\boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i);\boldsymbol{w}\big) \right|$$
$$\le \mathrm{M}_p\left(i \mapsto 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3r\sqrt{\frac{\ln \frac{g}{\delta}}{2m}}; \boldsymbol{w}\right) \ .$$

Experiments

♣ Training *linear models* on *adult* (census data) dataset

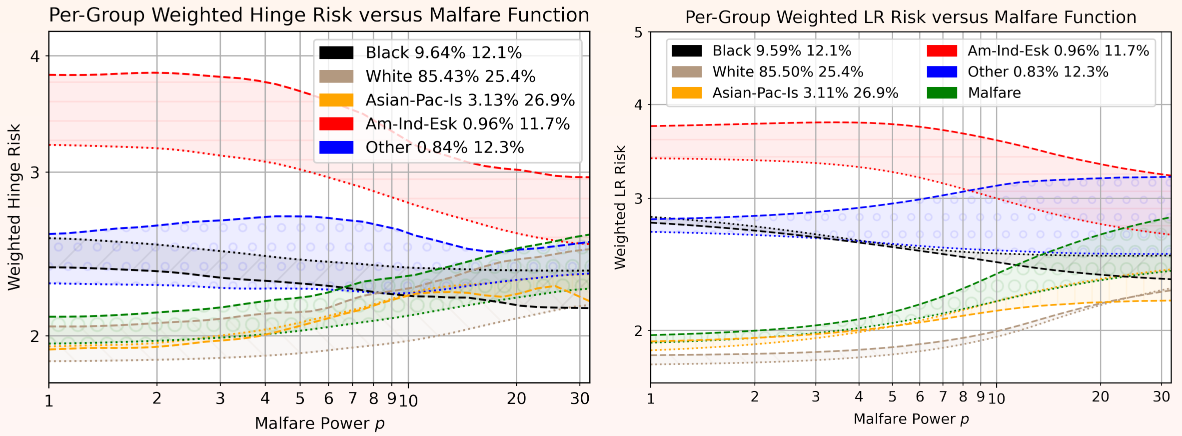
♠ Support vector machine (hinge loss)

♠ Logistic regression (cross entropy loss)

♠ Losses weighted by group-conditional label frequencies

♣ Predict whether income exceeds 50,000\$ per annum

♣ Minimize power-mean malfare over $g = 5$ ethnic groups



♣ Higher $p \implies$ fairer model, closer to *egalitarianism*

♣ $p = 1$ favors *large groups* (at the expense of minorities)

♠ *This is the default* (if minority groups are even considered during training)!

♠ Dire need for fairness-sensitive learning objectives

Fair PAC Learnability

Definition 2 (Fair-PAC Learnability)
Hypothesis class sequence $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{X} \to \mathcal{Y}$ *is* fair-PAC-learnable *w.r.t.* loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$ *if there exists a (randomized) algorithm $\mathcal{A}$, such that for all:*

1. *sequence indices d;*
2. *g instance distributions $\mathcal{D}_{1:g}$;*
3. *probability vectors $\boldsymbol{w} \in \mathbb{R}_+^g$;*
4. *malfares M satisfying axioms 1-7+9;*
5. *additive appx. errors $\varepsilon > 0$; &*
6. *failure probabilities $\delta \in (0, 1)$;*

$\mathcal{A}$ can identify a hypothesis $\hat{h} \in \mathcal{H}$, i.e., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$, where

1. ***finite sample complexity**: $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ consumes no more than $\mathrm{m}(\varepsilon, \delta, d, g) : (\mathbb{R}_+ \times (0,1) \times \mathbb{N} \times \mathbb{N}) \to \mathbb{N}$ samples; &*
2. ***correctness**: with probability at least $1 - \delta$, $\hat{h}$ obeys*

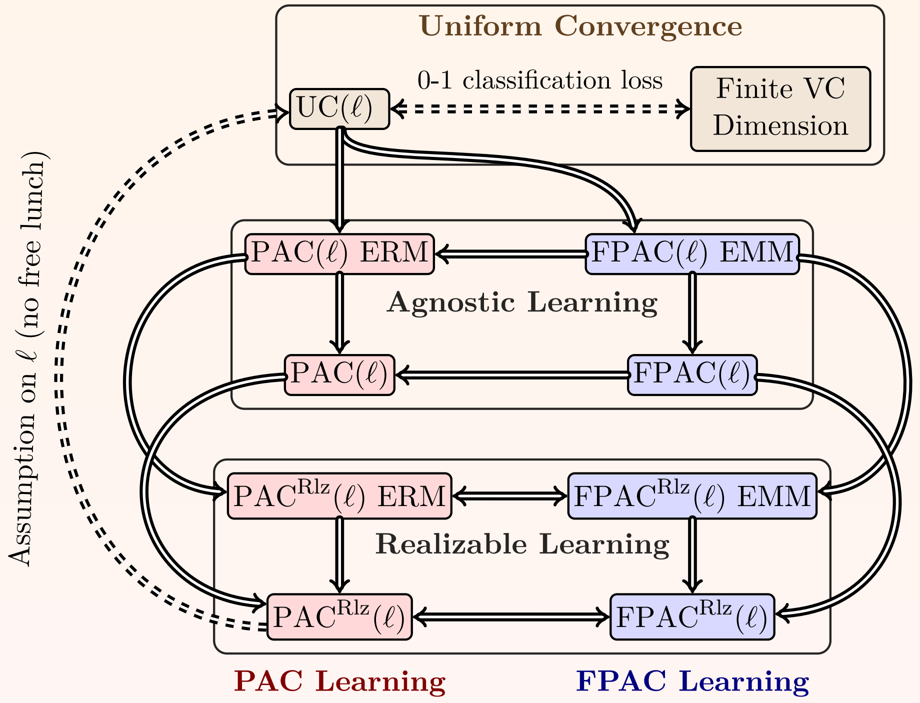
$$\mathrm{M}\left(i \mapsto \mathrm{R}(\hat{h};\ell,\mathcal{D}_i);\boldsymbol{w}\right) \le \inf_{h^*\in\mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{R}(h^*;\ell,\mathcal{D}_i);\boldsymbol{w}\right) + \varepsilon \ .$$

♣ The class of such fair-learning problems is denoted FPAC

♣ If $\forall d \in \mathbb{N}$, the space of $\mathcal{D}_{1:g}$ is restricted s.t. $\inf_{h\in\mathcal{H}_d} \max_{i\in 1,\dots,g} \mathrm{R}(h;\ell,\mathcal{D}_i) = 0$,

♠ then $(\mathcal{H}, \ell)$ is *realizable-FPAC-learnable*, denoted $(\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}$

Fundamental Theorem of Fair Statistical Learning



♣ Implications between membership in PAC and FPAC classes

♠ For fixed ℓ, $\Rightarrow$ denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment)

♣ Dashed implication arrows hold conditionally on ℓ

♠ When the no-free-lunch assumption on ℓ holds, the hierarchy collapses

♣ Under realizability, some classes are known to coincide