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

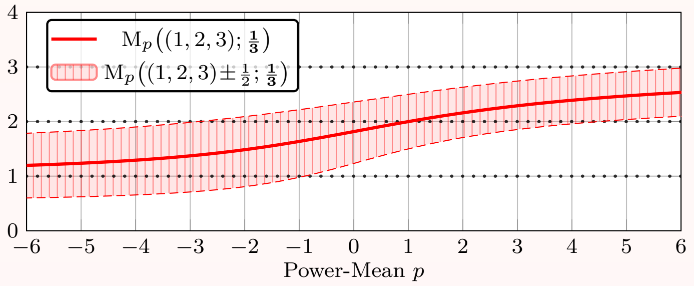
Cyrus Cousins of Brown University

Welfare, Malfare, and the Power Mean

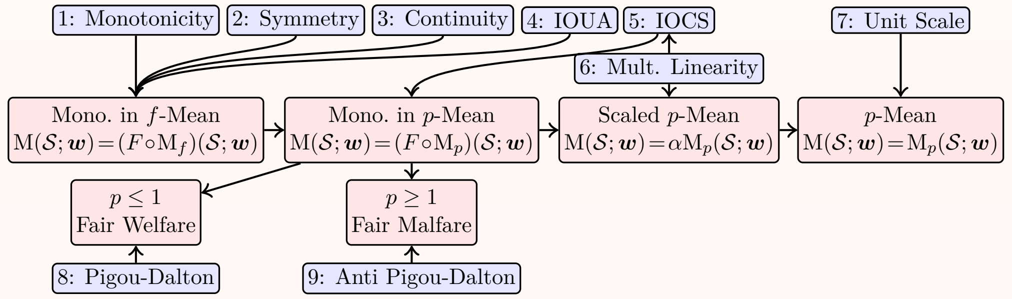
- The power-mean for $p \in \mathbb{R} \setminus \{0\}$ summarizes g *values* $\mathcal{S}$ with *weights* $\boldsymbol{w}$:

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

- Fair welfare: $p \leq 1$, $p = \infty$ is *maximin* over groups (egalitarianism)
  - Measure *overall wellbeing* given *utility values* (income, accuracy)
- Fair malfare: $p \geq 1$, $p = \infty$ is *minimax* 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 $\mathcal{S}$
  3. *Statistically Stable*
     (for $p \in [-\infty, 0) \cup [1, \infty]$)



Axioms of Cardinal Welfare



Relationships between population-mean axioms and properties.
Assumptions and axioms are shown in pastel blue, and properties in pastel red.

Estimating Malfare Values

1. Assuming only *monotonicity*:
   Suppose $\forall \omega \in \Omega : \hat{\mathcal{S}}(\omega) - \boldsymbol{\varepsilon}(\omega) \leq \mathcal{S}(\omega) \leq \hat{\mathcal{S}}(\omega) + \boldsymbol{\varepsilon}(\omega)$. Then

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

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

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

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

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

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

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

Empirical Malfare Minimization

Empirical risk and risk of hypothesis h given loss ℓ:

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

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

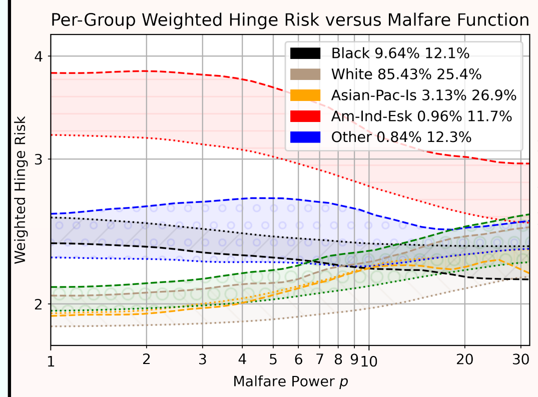
Theorem 1 (Generalization Guarantees for Malfare Estimation)**.** Suppose fair *power-mean malfare* $\mathrm{M}_p(\cdot; \cdot)$ (i.e., $p \geq 1$), *probability vector* $\boldsymbol{w} \in \mathbb{R}_+^g$, *loss function* $\ell : (\mathcal{Y} \times \mathcal{Y}) \to [0, r]$, samples $\boldsymbol{z}_i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

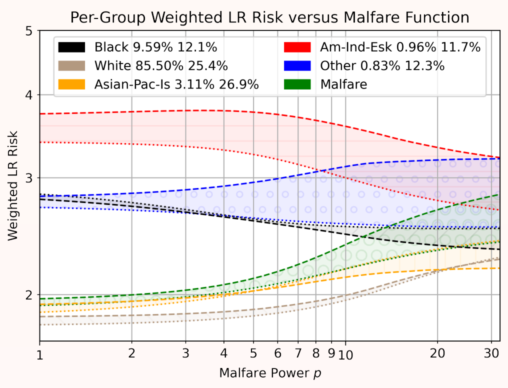
$$\sup_{h \in \mathcal{H}} \Big| \mathrm{M}_p\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) \Big|$$

$$\leq \mathrm{M}_p\bigg(i \mapsto 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3r\sqrt{\frac{\ln \frac{g}{\delta}}{2m}}; \boldsymbol{w}\bigg) \ .$$

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 is $\leq$ or $> 50,000\$$ per annum
- Minimize malfare over 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*, assuming minority groups are even considered during training!
  - Dire need for fairness-sensitive learning objectives

Fair PAC Learning

Definition 2 (Fair-PAC (FPAC) Learnability)**.** Suppose *hypothesis class sequence* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{X} \to \mathcal{Y}$, and *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$.

We say $\mathcal{H}$ is *fair PAC-learnable* w.r.t. *loss function* ℓ if $\exists$ 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 $\varepsilon > 0$; and
6. failure probabilities $\delta \in (0, 1)$;

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

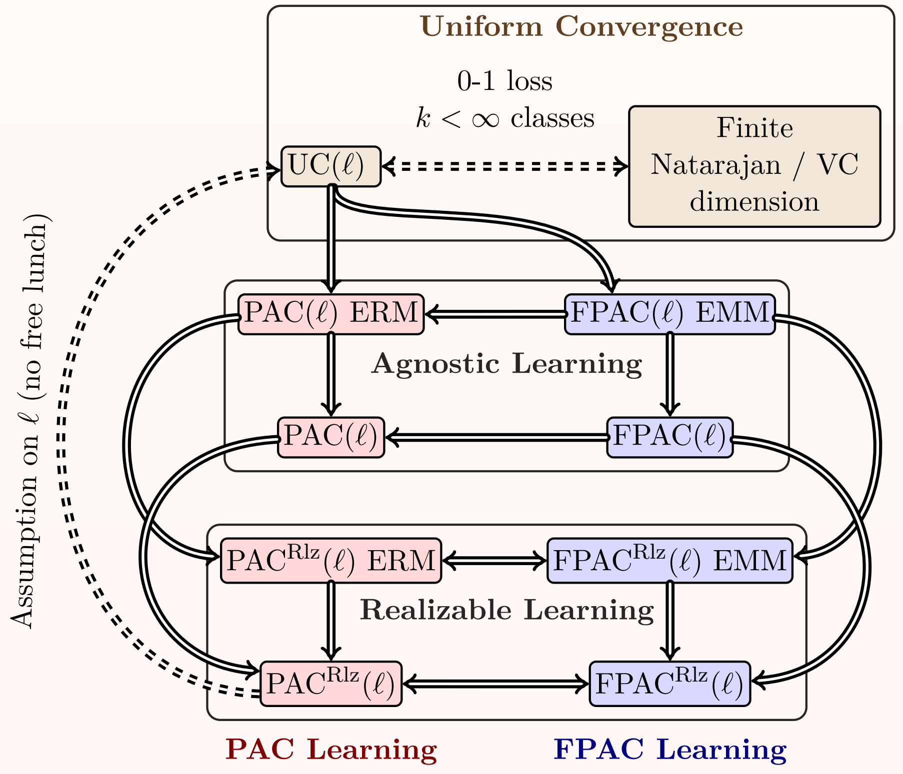
1. there exists some *sample complexity* function $\mathrm{m}(\varepsilon, \delta, d, g) : (\mathbb{R}_+ \times (0,1) \times \mathbb{N} \times \mathbb{N}) \to \mathbb{N}$ s.t. $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ consumes no more than $\mathrm{m}(\varepsilon,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

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

The class of such fair-learning problems is FPAC.

Finally, if for all d, the space of $\mathcal{D}$ is restricted such that $\exists h \in \mathcal{H}_d$ s.t. $\max_{i \in 1,\dots,g} \mathrm{R}(h; \ell, \mathcal{D}_i) = 0$, then $(\mathcal{H}, \ell)$ is *realizable-FPAC-learnable.*

Fair PAC Learnability



Implications between membership in PAC and FPAC classes. In particular, for arbitrary fixed ℓ, implication 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, and in general, under realizability, some classes are known to coincide.