# Revisiting Fair-PAC Learning and the Axioms of Cardinal Welfare

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

AIStats 2023
Valencia, Spain

## What is Welfare-Centric Fair Machine Learning?

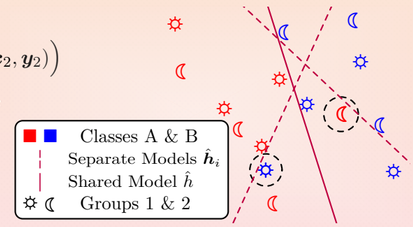
♣ Fair machine learning considers *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$
♣ We can handle each group individually
♢ *Empirical utility maximization*

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

♣ 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!



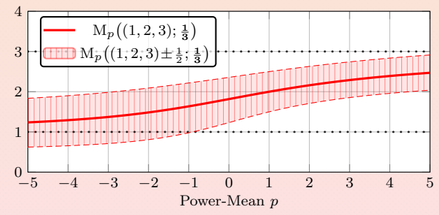

## Utilitarian, Egalitarian, and the Power-Mean Welfare

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

$$\mathrm{M}_{p\neq 0}(\mathcal{S}; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p}\ , \quad \mathrm{M}_0(\mathcal{S}; \boldsymbol{w}) \doteq \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \log(\mathcal{S}_i^p)\right) = \prod_{i=1}^{g} \mathcal{S}_i^{\boldsymbol{w}_i}$$

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



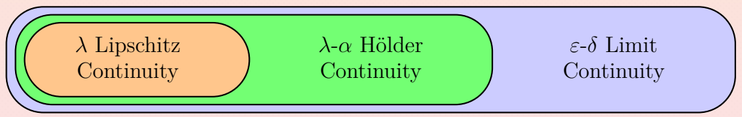

## A Hierarchy of Continuity Concepts

**Definition 1** (Hölder Continuity)
$\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ *is* Hölder continuous *in* $\mathcal{S}$ *with respect to norm* $\|\cdot\|_{\mathrm{M}}$ *if there exist some*
1. *scale* $\lambda \geq 0$,
2. *power* $\alpha \in (0, 1]$,

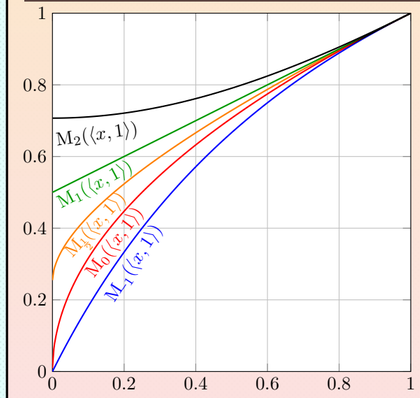
*such that for all* $\mathcal{S}, \mathcal{S}'$, *it holds that*

$$\left|\mathrm{M}(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}(\mathcal{S}'; \boldsymbol{w})\right| \leq \lambda \left\|\mathcal{S} - \mathcal{S}'\right\|_{\mathrm{M}}^{\alpha} \ .$$

♣ Such a function is $\lambda$-$\alpha$-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous.
♢ Bound the impact of *small changes*
♣ If $\alpha = 1$, it is $\lambda$-$\|\cdot\|_{\mathrm{M}}$ Lipschitz continuous.
♢ Bound the impact of *infinitessimal changes*




## Small-Scale Behavior of Power-Means

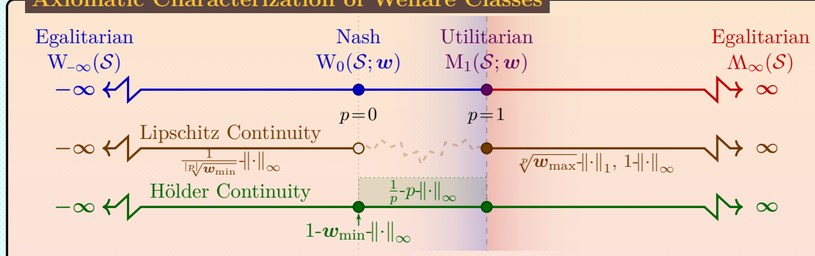



$$\frac{\partial}{\partial \mathcal{S}_i} \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) = \boldsymbol{w}_i \underbrace{\left(\frac{\mathcal{S}_i}{\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})}\right)^{p-1}}_{\text{Relative utility}}$$

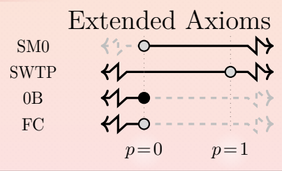
Assuming *unit range*, power-means are:
1. $p \geq 1$: 1-$\|\cdot\|_\infty$ Lipschitz
2. $p \in (0, 1)$: $\frac{1}{p}$-$p$-$\|\cdot\|_\infty$ Hölder
3. $p = 0$: 1-$\boldsymbol{w}_{\min}$-$\|\cdot\|_\infty$ Hölder
4. $p < 0$: $\frac{1}{\sqrt[p]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_\infty$ Lipschitz

The "difficult cases" occur as:
♣ $p \to 0$
♣ $\boldsymbol{w}_{\min} \to 0$ for $p < 1$

## Axiomatic Characterization of Welfare Classes




♣ Axiomatic characterization of welfare functions
♢ Uniquely satisfied by $\mathrm{W}_p(\cdot; \boldsymbol{w})$ for $p \leq 1$
♣ Additional axioms further restrict $p$
♢ Continuity properties vary by region
♢ Impact on sample efficiency of learning




## Fair-PAC Learning

**Definition 2** (Fair-PAC Learning)
*Suppose*
1. *hypothesis class* $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$
2. *utility function* $\mathrm{U}: \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$
3. *welfare class* $\mathcal{W} \subseteq \mathbb{R}_{0+}^g \to \mathbb{R}_{0+}$

$\mathcal{H}$ *is* fair-PAC-learnable *if there exists an algorithm* $\mathcal{A}$ *such that for any*
1. *distributions* $\mathcal{D}_{1:g}$ *over* $(\mathcal{X} \times \mathcal{Y})$
2. *welfare function* $\mathrm{W}(\cdot; \boldsymbol{w}) \in \mathcal{W}$
3. *additive error* $\varepsilon > 0$
4. *failure probability* $\delta \in (0, 1)$

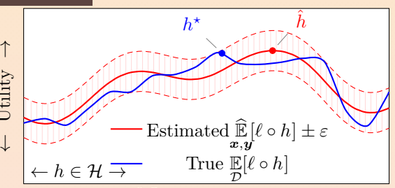
$\mathcal{A}$ *can identify a hypothesis* $\hat{h} \in \mathcal{H}$ *such that*
1. $\mathcal{A}$ *has* $\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon, \delta, \mathrm{W}, g)$ *sample complexity (per-group)*
2. *with probability at least* $1 - \delta$, $\hat{h}$ *obeys*

$$\underbrace{\mathrm{W}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}\big[\mathrm{U}(\hat{h}(x), y)\big], \ldots; \boldsymbol{w}\right)}_{\text{Learned model welfare}} \geq \underbrace{\operatorname*{argmax}_{h^\star\in\mathcal{H}} \mathrm{W}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}\big[\mathrm{U}(h^\star(x), y)\big], \ldots; \boldsymbol{w}\right)}_{\text{Optimal model welfare}} - \varepsilon$$

## Uniform Convergence and PAC-Learnability

♣ Suppose that for hypothesis class $\mathcal{H}$, with a sample $(\boldsymbol{x}, \boldsymbol{y})$ of size $\mathrm{m}_{\mathcal{H}}(\varepsilon, \delta)$, it holds

$$\mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}\left(\sup_{h\in\mathcal{H}} \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell \circ h] - \mathop{\widehat{\mathbb{E}}}_{\boldsymbol{x},\boldsymbol{y}}[\ell \circ h]\right| > \varepsilon\right) < \delta$$

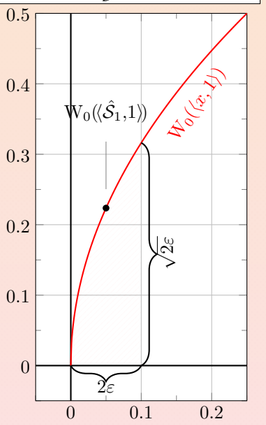



♣ Many ways to show this for various $\mathcal{H}$
♢ Vapnik-Chervonenkis dimension
♢ Rademacher averages

♣ What does *uniform convergence* give us?
♢ *Asymptotic consistency* of empirical utility maximizer $\hat{h}$
♢ Finite-sample convergence rate bounds
♡ UC $\implies$ PAC with sample complexity $\mathrm{m}_{\mathcal{H}}(\frac{\varepsilon}{2}, \delta)$

♣ By $\varepsilon$-$\delta$ limit continuity alone:
♢ Consistency of empirical welfare maximizer $\hat{h}$
♢ Finite-sample convergence rate bounds?
♡ Convergence rate depends on welfare function
♡ Asymptotic bounds in terms of $\nabla_{\mathcal{S}^\star} \mathrm{W}(\mathcal{S}^\star; \boldsymbol{w})$ for welfare-maximal utility vector $\mathcal{S}^\star$

♣ By how much does $\mathrm{W}(\cdot; \boldsymbol{w})$ magnify error?
♢ Hölder continuity analysis




## Characterizing Fair-PAC Learnability

**Theorem 3** (Welfare Optimization Sample Complexity)
*Suppose that for sample size* $\mathrm{m}_{\mathcal{H}}(\varepsilon, \delta)$, *it holds that*

$$\mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}\left(\sup_{h\in\mathcal{H}} \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell \circ h] - \mathop{\widehat{\mathbb{E}}}_{\boldsymbol{x},\boldsymbol{y}}[\ell \circ h]\right| > \varepsilon\right) < \delta \ .$$

*Then* $\mathcal{H}$ *is FPAC-learnable with sample complexity* $\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon, \delta, \mathrm{W}, g) \leq \mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{2\lambda}}, \frac{\delta}{g}\right)$ .

♣ Sample complexity of $\varepsilon$-$\delta$ learning bounded objectives is usually $\mathrm{m}_{\mathcal{H}}(\varepsilon, \delta) \in \mathbf{O}\left(\frac{\ln \frac{1}{\delta}}{\varepsilon^2}\right)$
♣ Fair-learning the class of *all weighted power-means* is thus

$$\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon, \delta, \mathrm{W}, g) \leq \mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{2\lambda}}, \frac{\delta}{g}\right) \in \mathbf{O}\left(\frac{\lambda^{\frac{2}{\alpha}} \ln \frac{g}{\delta}}{\varepsilon^{\frac{2}{\alpha}}}\right) \subseteq \underbrace{\mathbf{O}\left(\frac{\ln \frac{g}{\delta}}{\varepsilon^2 \boldsymbol{w}_{\min}^2}\right)}_{\text{Worst-Case Bound!}}$$

♣ For any constant $c \in (0, 1)$, if $\boldsymbol{w}_{\min} \geq \frac{c}{g}$ and $|p| \geq c$, then

$$\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon, \delta, \mathrm{W}, g) \in \mathbf{O}\left(\frac{g^{\frac{2}{c}} \ln \frac{g}{\delta}}{\varepsilon^{\frac{2}{c}}}\right) \subseteq \mathrm{Poly}^{\frac{1}{c}}\left(\frac{1}{c}, \frac{1}{\varepsilon}, \frac{1}{\delta}, \log \frac{1}{\delta}\right) \ .$$

## A New Paradigm of Statistically Sound Fair Machine Learning

♣ Axiomatically characterize class of fair welfare functions
♢ Fairness varies interpersonally, but "reasonable axioms" describe "reasonable people"
♢ The power-mean family (with $p \leq 1$): $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p}$
♣ Analyze continuity properties of fair welfare functions
♢ Lipschitz and Hölder continuity: $\left|\mathrm{M}(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}(\mathcal{S}'; \boldsymbol{w})\right| \leq \lambda \left\|\mathcal{S} - \mathcal{S}'\right\|_{\mathrm{M}}^{\alpha}$
♣ Fair-PAC learnability for all wefare functions W in class $\mathcal{W}$
♢ Uniform convergence $\implies$ FPAC-Learnability