# Superhuman Fairness

Omid Memarrast, Linh Vu, Brian D. Ziebart

UNIVERSITY OF ILLINOIS CHICAGO

ICML International Conference On Machine Learning

## Motivation

Defining desired fairness-predictive performance trade-offs precisely is difficult:

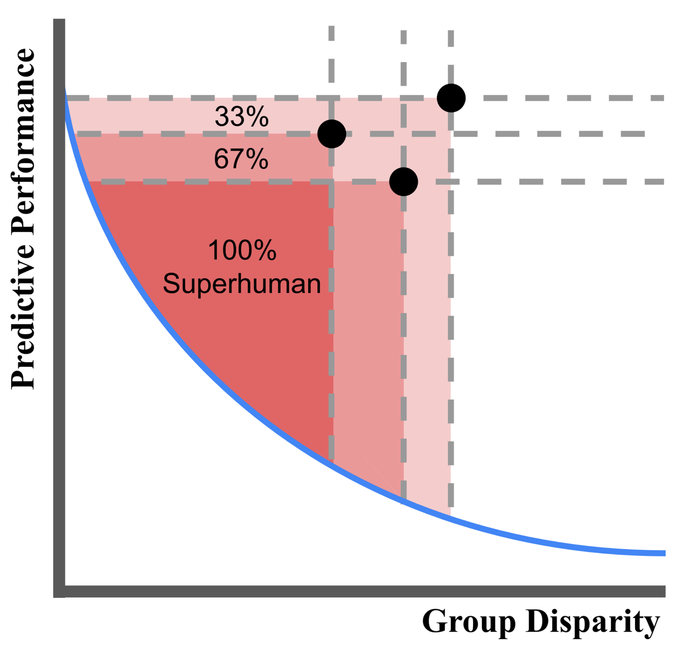
- **Multiple fairness metrics [dp, eqodds, eqopp, prp, ...]**
- **One (or more) predictive performance metrics [acc, f-meas, ...]**

To produce desirable decisions on actual data, fine-tuning any hand-specified trade-off is often required.

Human decisions (i.e., reference decisions) are often available, but the fairness trade-offs they are based on are typically unknown.

**A new fairness question:** Can algorithmic decisions be produced that **all stakeholders with different notions of fairness and desired performance-fairness trade-offs** prefer over human decisions?

**Our approach:** seek decisions that outperform reference human decisions across all fairness/performance metrics of interest.

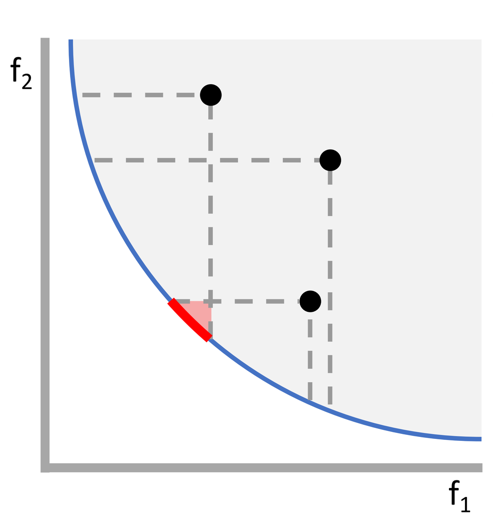



Three sets of decisions (**black dots**) with different predictive performance and group disparity values defining the sets of 100%-, 67%-, and 33%-superhuman fairness-performance values (red shades) based on Pareto dominance.

**Why not elicit preferences [1]?** Multiple stakeholders often influence decisions, and eliciting their preferences does not resolve how their competing preferences should be prioritized.

**Why not use inverse reinforcement learning methods [2, 3] (i.e., feature-matching)?** Noise in the reference decisions can make estimating demonstrated fairness-performance trade-offs error prone, leading to decisions that some stakeholders prefer less than reference decisions even when decisions that all stakeholders prefer exist.

## Superhuman behavior: an ideal objective?



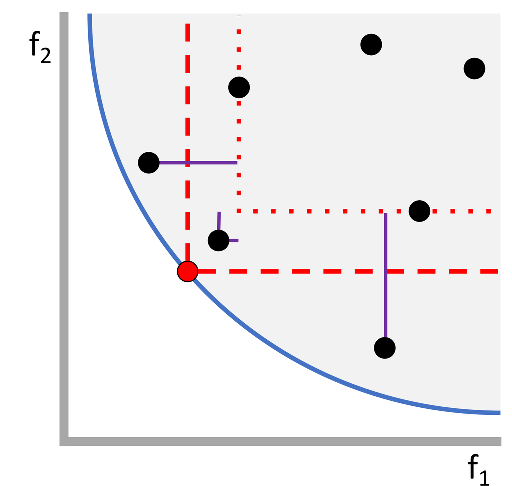

A **policy** is **superhuman** if it has smaller **cost features** $f_1, f_2, \ldots$ for all **human demonstrations** [4]

Guarantees lower cost than demonstration costs for family of additive cost functions

Set of **superhuman policies** on the **Pareto frontier** shrinks as demonstrations grow

Unfortunately, this set can often become empty!

## Subdominance Minimization




A **policy** is **γ-superhuman** if it has smaller **metrics** $f_1$, $f_2$, … than γ% of **human demonstrations**

**Subdominance** measures how far a policy is from superhuman by some **margins,** bounding the **superhuman percentile**.

**Minimum Subdominance Inverse Optimal Control [4]** seeks policies on the Pareto frontier minimizing it

**Subdominance** in each measure $\{f_k\}$ for a set of:
reference decisions (human demonstration) $\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^{M}$ and
model predictions $\hat{\mathbf{y}} = \{\hat{y}_j\}_{j=1}^{M}$ is measured as:

$$\text{subdom}_{\alpha_k}^{k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \triangleq \left[\alpha_k \left(f_k(\hat{\mathbf{y}}, \mathbf{a}) - f_k(\tilde{\mathbf{y}}, \mathbf{a})\right) + 1\right]_+$$

The subdominance for decision vector $\hat{\mathbf{y}}$ with respect to the set of demonstrations (N vectors of reference decisions) aggregated over k measure can be measure as:

$$\text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) = \frac{1}{N} \sum_{\tilde{\mathbf{y}} \in \tilde{\boldsymbol{\mathcal{Y}}}} \sum_{k} \text{subdom}_{\alpha_k}^{k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})$$

The minimally subdominant fairness-aware classifier $P_{\boldsymbol{\theta}}$ has model parameters $\boldsymbol{\theta}$ chosen by:

$$\underset{\boldsymbol{\theta}}{\text{argmin}} \min_{\boldsymbol{\alpha} \succeq 0} \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim P_{\boldsymbol{\theta}}} \left[\text{subdom}_{\boldsymbol{\alpha}}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right)\right] + \lambda \|\boldsymbol{\alpha}\|_1$$

Hinge loss slopes $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^{K}$ are also learned during training. $\alpha_k$ value defines by how far a produced decision does not sufficiently outperform the demonstrations in measure $\{f_k\}$.

We use policy gradient to obtain $\boldsymbol{\theta}$ :

$$\nabla_\theta \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[ \sum_k \overbrace{\min_{\alpha_k} \left(\text{subdom}_{\alpha_k}^{k}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right) + \lambda_k \alpha_k\right)}^{\Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})} \right] = \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[ \left( \sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right]$$

And solve for $\boldsymbol{\alpha}$ analytically given y $\alpha_k = \underset{\alpha_k^{(m)}}{\text{argmin}}\, m$ *such that* $f_k(\hat{\mathbf{y}}) + \lambda \leq \frac{1}{m} \sum_{j=1}^{m} f_k\left(\tilde{\mathbf{y}}^{(j)}\right)$

where $\alpha_k^{(j)} = \frac{1}{f_k\left(\hat{\mathbf{y}}^{(j)}\right) - f_k\left(\tilde{\mathbf{y}}^{(j)}\right)}$

Note: When the $\alpha_k$ is large, the model heavily weights support vector reference decisions for that particular k when minimizing subdominance.

**Generalization:** On average, the **minimally subdominant policy** is $\gamma$-superhuman on the population distribution (under IID assumptions) with:

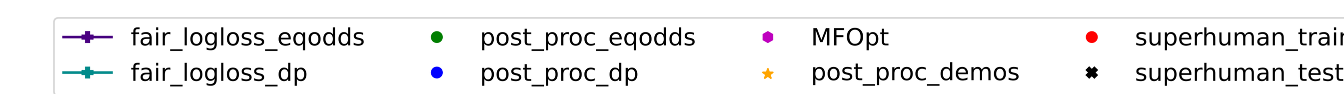
$$\gamma = 1 - \frac{1}{N} \left\| \bigcup_{k=1}^{K} \tilde{\boldsymbol{\mathcal{Y}}}_{SV_k}\left(\hat{\mathbf{y}}, \alpha_k\right) \right\|$$

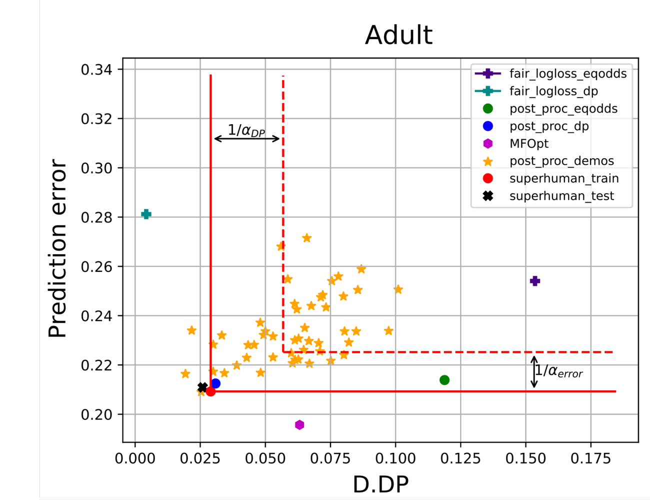
**Algorithm 1** Subdominance policy gradient optimization

Draw N set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{N}$ from a human decision-maker or baseline method $\tilde{\mathbb{P}}$. Initialize: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$
**while** $\boldsymbol{\theta}$ *not converged* **do**
 Sample model predictions $\{\hat{\mathbf{y}}_i\}_{i=1}^{N}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{N}$
 **for** $k \in \{1, ..., K\}$ **do**
  Sort reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{N}$ in ascending order by $k^{\text{th}}$ measure value $f_k(\tilde{\mathbf{y}}_i)$: $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{N}$
  Compute $\alpha_k^{(j)} = \frac{1}{f_k\left(\tilde{\mathbf{y}}^{(j)}\right) - f_k\left(\hat{\mathbf{y}}^{(j)}\right)}$
  $\alpha_k = \underset{\alpha_k^{(m)}}{\text{argmin}}\, m$ such that
  $f_k(\hat{\mathbf{y}}) + \lambda \leq \frac{1}{m} \sum_{j=1}^{m} f_k\left(\tilde{\mathbf{y}}^{(j)}\right)$
  Compute $\Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})$
 $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\eta}{N} \sum_i \left(\sum_k \Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})\right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}_i|\mathbf{X}_i)$;

## Experiments

We create 50 **synthetic** demonstrations using post-processing fairness method (Hardt et al. 2016) for demographic parity. Then we train our model to find $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ that minimize the Subdominance value. We use a logistic regression model with weights $\boldsymbol{\theta}$ as our decision model. We perform experiments on Adult and COMPAS datasets.



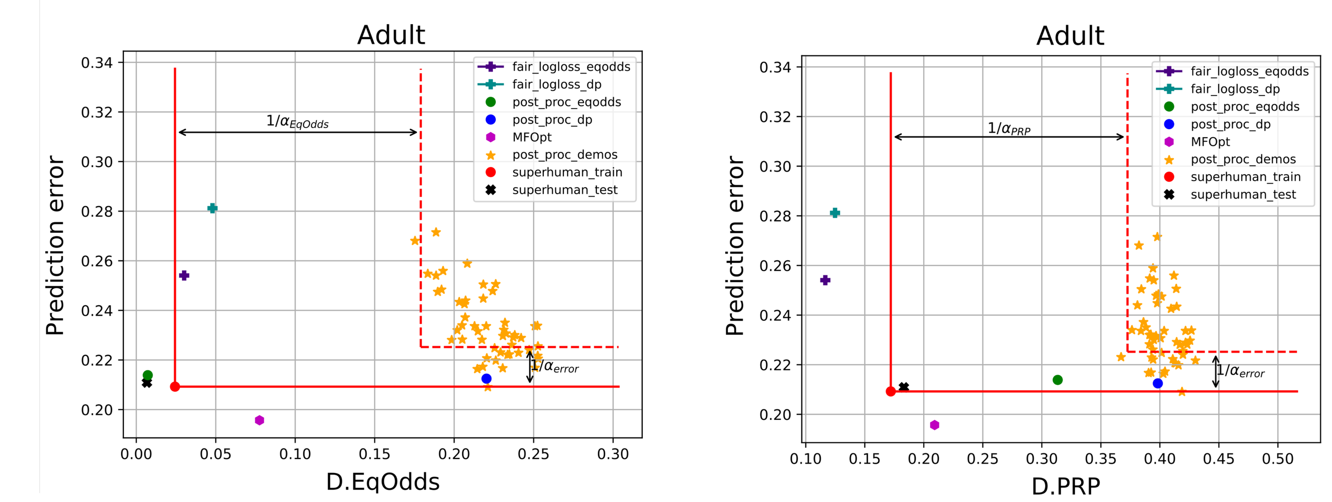

**What Metrics to use:** Three Conflicting Fairness Measures (Impossibility Theorem)

(In)Accuracy + [Demographic Parity, Equalized Odds, Predictive Rate Parity]

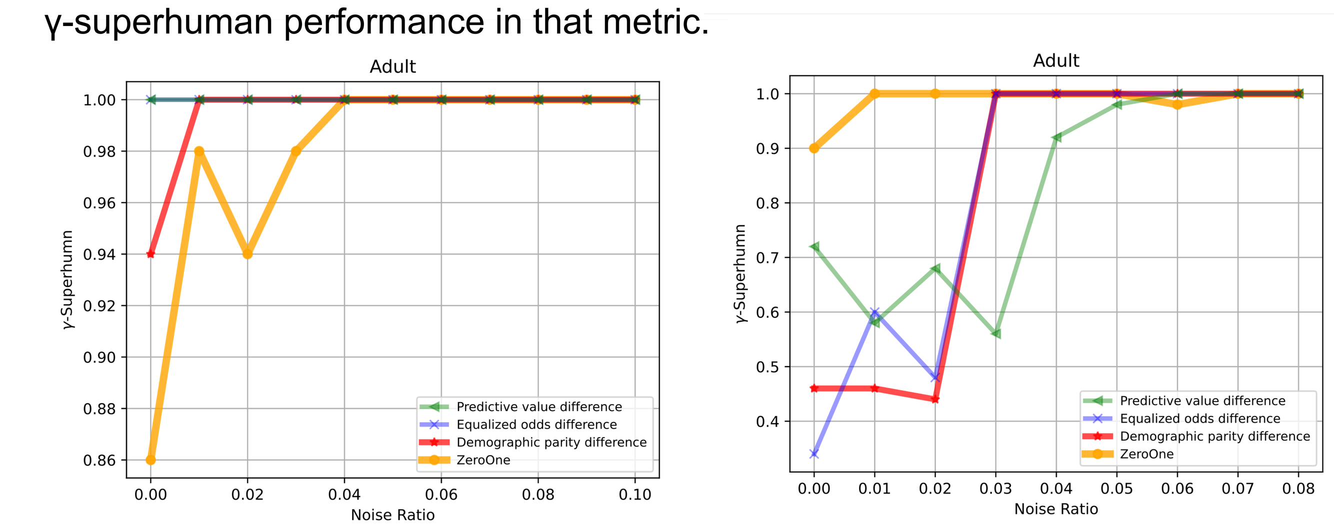
Our approach outperforms demonstrations and shows competitive performance with baselines.

★ Demonstrations (Ref Decisions)

● Our Model (Superhuman_train)




As we increase noise in the label and the protected attribute of reference decisions produced by post-processing (left) and fair-logloss (right) our approach achieves higher γ-superhuman performance in that metric.




In both noiseless and noisy settings our approach outperforms higher percentage of demonstrations in all prediction/fairness measure compared to other baselines.

| Method \ Dataset | Adult $\epsilon = 0.0$ | Adult $\epsilon = 0.2$ | COMPAS $\epsilon = 0.0$ | COMPAS $\epsilon = 0.2$ |
|---|---|---|---|---|
| MinSub-Fair (ours) | **96%** | **100%** | **100%** | **98%** |
| MFOpt | 42% | 0% | 18% | 18% |
| post_proc_dp | 16% | 86% | **100%** | 80% |
| post_proc_eqodds | 0% | 66% | **100%** | 88% |
| fair_logloss_dp | 0% | 0% | 0% | 0% |
| fair_logloss_eqodds | 0% | 0% | 0% | 0% |

Percentage of reference demonstrations that each method outperforms in all predictive performance and fairness measures.

### References

[1] Hiranandani, G., Narasimhan, H., and Koyejo, S. Fair performance metric elicitation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11083–11095, 2020..

[2] Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 1–8, 2004.

[3] Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, pp. 1449–1456, 2007.

[4] Ziebart, B., Choudhury, S., Yan, X., and Vernaza, P. Towards uniformly superhuman autonomy via subdominance minimization. In *International Conference on Machine Learning*, pp. 27654–27670. 2022.

This material is based upon work supported by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939743.