Superhuman Fairness

Omid Memarrast [1] **Linh Vu** [1] **Brian Ziebart** [1]

Abstract

The fairness of machine learning-based decisions has become an increasingly important focus in the design of supervised machine learning methods. Most fairness approaches optimize a specified trade-off between performance measure(s) (e.g., accuracy, log loss, or AUC) and fairness measure(s) (e.g., demographic parity, equalized odds). This begs the question: are the right performance-fairness trade-offs being specified? We instead recast fair machine learning as an imitation learning task by introducing *superhuman fairness*, which seeks to simultaneously outperform human decisions on multiple predictive performance and fairness measures. We demonstrate the benefits of this approach given suboptimal decisions.

1. Introduction

The social impacts of algorithmic decisions based on machine learning have motivated various group and individual fairness properties that decisions should ideally satisfy (Calders et al., 2009; Hardt et al., 2016). Unfortunately, impossibility results prevent multiple common group fairness properties from being simultaneously satisfied (Kleinberg et al., 2016). Thus, no set of decisions can be universally fair to all groups and individuals for all notions of fairness. Instead, specified weightings, or trade-offs, of different criteria are often optimized (Liu & Vicente, 2022). Identifying an appropriate trade-off to prescribe to these fairness methods is a daunting task open to application-specific philosophical and ideological debate that could delay or completely derail the adoption of algorithmic methods.

We consider the motivating scenario of multiple (error-prone) stakeholders with different notions of fairness and desired performance-fairness trade-offs collaboratively producing decisions. Preference elicitations (Hiranandani et al., 2020) is of limited use since knowing the stakeholder trade-offs still leaves the question of how different stakeholders preferences should be prioritized. Rather than seeking optimal decisions for specific performance-fairness (meta-)trade-offs, we propose a more modest, yet more practical objective: **produce decisions preferred by all stakeholders over human-produced decisions with maximal frequency.** This provides an opportunity for **superhuman decisions** that Pareto dominate human decisions across predictive performance and fairness measures (Figure 1) *without identifying an explicit desired trade-off*. We argue that for many algorithmic fairness tasks, frequently outperforming human decisions across all relevant predictive performance and fairness measures may be sufficient for replacing human decision-makers with algorithmic decision-makers.

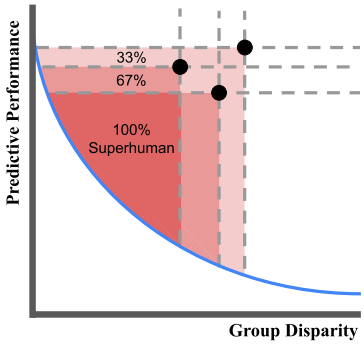

Figure 1. Three sets of decisions (black dots) with different predictive performance and group disparity values defining the sets of 100%-, 67%-, and 33%-superhuman fairness-performance values (red shades) based on Pareto dominance.

To the best of our knowledge, this paper is the first to define fairness objectives for supervised machine learning with respect to noisy human decisions rather than using prescriptive trade-offs or hard constraints. We leverage and extend a recently-developed imitation learning method for **subdominance minimization** (Ziebart et al., 2022). Instead of using the subdominance to identify a target trade-off, as previous work does in the inverse optimal control setting of sequential decision-making to estimate a cost function, we use it to directly optimize our fairness-aware classifier. We develop a method based on policy gradient optimization (Sutton & Barto, 2018) that allows flexible classes of probabilistic decision policies (e.g., aware or unaware of protected group membership status) to be optimized for given sets of performance/fairness measures and demonstrations.

We conduct extensive experiments on standard fairness datasets (`Adult` and `COMPAS`) using accuracy as a per-

[1]Department of Computer Science, University of Illinois Chicago, Chicago, USA. Correspondence to: Omid Memarrast <omemar2@uic.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

formance measure and three conflicting fairness definitions: Demographic Parity (Calders et al., 2009), Equalized Odds (Hardt et al., 2016), and Predictive Rate Parity (Chouldechova, 2017). Though our motivation is to outperform human decisions, we employ a synthetic decision-maker with differing amounts of label and group membership noise to identify sufficient conditions for superhuman fairness of varying degrees. We find that our approach achieves high levels of superhuman performance that increase rapidly with reference decision noise and significantly outperform the superhumanness of other methods that are based on more narrow fairness-performance objectives.

2. Fairness, Elicitation, and Imitation

2.1. Group Fairness Measures

Group fairness measures are primarily defined by confusion matrix statistics (based on labels $y_i \in \{0,1\}$ and decisions/predictions $\hat{y}_i \in \{0,1\}$ produced from inputs $\mathbf{x}_i \in \mathbb{R}^M$) for examples belonging to different protected groups (e.g., $a_i \in \{0,1\}$).

We focus on three prevalent fairness properties in this paper:

- **Demographic Parity** (DP) (Calders et al., 2009) requires equal positive rates across protected groups:

$$\mathrm{P}(\hat{Y}=1|A=1) = \mathrm{P}(\hat{Y}=1|A=0);$$

- **Equalized Odds** (EqOdds) (Hardt et al., 2016) requires equal true positive rates and false positive rates across groups, i.e.,

$$\mathrm{P}(\hat{Y}=1|Y=y,A=1) = \mathrm{P}(\hat{Y}=1|Y=y,A=0), \quad y \in \{0,1\};$$

- **Predictive Rate Parity** (PRP) (Chouldechova, 2017) requires equal positive predictive value ($\hat{y}=1$) and negative predictive value ($\hat{y}=0$) across groups:

$$\mathrm{P}(Y=1|A=1,\hat{Y}=\hat{y}) = \mathrm{P}(Y=1|A=0,\hat{Y}=\hat{y}), \quad \hat{y} \in \{0,1\}.$$

Violations of these fairness properties can be measured as differences:

$$\texttt{D.DP}(\hat{\mathbf{y}},\mathbf{a}) = \left| \frac{\sum_{i=1}^N \mathbb{I}\left[\hat{y}_i=1, a_i=1\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=1\right]} - \frac{\sum_{i=1}^N \mathbb{I}\left[\hat{y}_i=1, a_i=0\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=0\right]} \right|; \tag{1}$$

$$\texttt{D.EqOdds}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a}) = \max_{y\in\{0,1\}} \left| \frac{\sum_{i=1}^N \mathbb{I}\left[\hat{y}_i=1, y_i=y, a_i=1\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=1, y_i=y\right]} - \frac{\sum_{i=1}^N \mathbb{I}\left[\hat{y}_i=1, y_i=y, a_i=0\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=0, y_i=y\right]} \right|; \tag{2}$$

$$\texttt{D.PRP}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a}) = \max_{y\in\{0,1\}} \left| \frac{\sum_{i=1}^N \mathbb{I}\left[y_i=1, \hat{y}_i=y, a_i=1\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=1, \hat{y}_i=y\right]} - \frac{\sum_{i=1}^N \mathbb{I}\left[y_i=1, \hat{y}_i=y, a_i=0\right]}{\sum_{i=1}^N \mathbb{I}\left[a_i=0, \hat{y}_i=y\right]} \right|. \tag{3}$$

2.2. Performance-Fairness Trade-offs

Numerous fair classification algorithms have been developed over the past few years, with most targeting one or two fairness measures (Zafar et al., 2015; Hardt et al., 2016; Goel et al., 2018; Aghaei et al., 2019). With some exceptions (Blum & Stangl, 2019), predictive performance and fairness are typically competing objectives in supervised machine learning approaches (Menon & Williamson, 2018). Thus, though satisfying many fairness properties simultaneously may be naïvely appealing, doing so often significantly degrades predictive performance or even creates infeasibility (Kleinberg et al., 2016).

Given this, many approaches seek to choose parameters θ for (probabilistic) classifier P_θ that balance the competing predictive performance and fairness objectives (Kamishima et al., 2012; Hardt et al., 2016; Menon & Williamson, 2018; Celis et al., 2019; Martinez et al., 2020; Rezaei et al., 2020). Recently, Hsu et al. (2022) proposed a novel optimization framework to satisfy three conflicting fairness measures (demographic parity, equalized odds, and predictive rate parity) to the best extent possible:

$$\min_\theta \mathbb{E}_{\hat{\mathbf{y}}\sim P_\theta}\Big[\mathrm{loss}(\hat{\mathbf{y}},\mathbf{y}) + \alpha_{\texttt{DP}}\texttt{D.DP}(\hat{\mathbf{y}},\mathbf{a}) + \alpha_{\texttt{Odds}}\texttt{D.EqOdds}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a}) + \alpha_{\texttt{PRP}}\texttt{D.PRP}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a})\Big]. \tag{4}$$

2.3. Preference Elictation & Imitation Learning

Preference elicitation (Chen & Pu, 2004) is one natural approach to identifying desirable performance-fairness trade-offs. Preference elicitation methods typically query users for their pairwise preference on a sequence of pairs of options to identify their utilities for different characteristics of the options. This approach has been extended and applied to fairness measure elicitation (Hiranandani et al., 2020), allowing efficient learning of linear (e.g., Eq. (4)) and non-linear measures from finite and noisy preference feedback.

When decisions are made jointly by multiple stakeholders (Donaldson & Preston, 1995) rather than a single individual, preference elicitation may not be very informative. Each stakeholder's preferences could be elicited, for example, but how those sets of preferences should be prioritized to determine joint outcomes can remain unclear without strong additional assumptions about the decision-making process (e.g., outcomes determined by a majority vote) (Dowling et al., 2016).

Imitation learning (Osa et al., 2018) is a type of supervised

machine learning that seeks to produce a general-use policy $\hat{\pi}$ based on demonstrated trajectories of states and actions, $\tilde{\xi} = (\tilde{s}_1, \tilde{a}_1, \tilde{s}_2, \ldots, \tilde{s}_T)$. Inverse reinforcement learning methods (Abbeel & Ng, 2004; Ziebart et al., 2008) seek to rationalize the demonstrated trajectories as the result of (near-) optimal policies on an estimated cost or reward function. Feature matching (Abbeel & Ng, 2004) plays a key role in these methods, guaranteeing if the expected feature counts match, the estimated policy $\hat{\pi}$ will have an expected cost under the demonstrator's unknown fixed cost function weights $\tilde{w} \in \mathbb{R}^K$ equal to the average of the demonstrated trajectories:

$$\mathbb{E}_{\xi\sim\hat{\pi}}\left[f_k(\xi)\right] = \frac{1}{N}\sum_{i=1}^{N} f_k\left(\tilde{\xi}_i\right), \forall k \tag{5}$$

$$\implies \mathbb{E}_{\xi\sim\hat{\pi}}\left[\text{cost}_{\tilde{w}}(\xi)\right] = \frac{1}{N}\sum_{i=1}^{N} \text{cost}_{\tilde{w}}\left(\tilde{\xi}_i\right),$$

where $f_k(\xi) = \sum_{s_t\in\xi} f_k\left(s_t\right)$.

Syed & Schapire (2007) seeks to outperform the set of demonstrations when the signs of the unknown cost function are known, $\tilde{w}_k \geq 0$, by making the inequality,

$$\mathbb{E}_{\xi\sim\pi}\left[f_k(\xi)\right] \leq \frac{1}{N}\sum_{i=1}^{N} f_k\left(\tilde{\xi}_i\right), \forall k, \tag{6}$$

strict for at least one feature. Subdominance minimization (Ziebart et al., 2022) seeks to produce trajectories that outperform each demonstration by a margin:

$$f_k(\xi) + m_k \leq f_k(\tilde{\xi}_i), \forall i, k, \tag{7}$$

under the same assumption of known cost weight signs. However, since this is often infeasible, the approach instead minimizes the subdominance, which measures the α-weighted violation of this inequality:

$$\text{subdom}_\alpha(\xi, \tilde{\xi}) \triangleq \sum_k \left[\alpha_k\left(f_k(\xi) - f_k(\tilde{\xi})\right) + 1\right]_+, \tag{8}$$

where $[f(x)]_+ \triangleq \max(f(x), 0)$ is the hinge function and the per-feature margin has been reparameterized as α_k^{-1}. Previous work (Ziebart et al., 2022) has employed subdominance minimization in conjunction with inverse optimal control:

$$\min_{\mathbf{w}} \min_{\alpha} \sum_{i=1}^{N}\sum_{k=1}^{K} \text{subdom}_\alpha(\xi^*(\mathbf{w}), \tilde{\xi}_i), \text{where:}$$

$$\xi^*(\mathbf{w}) = \underset{\xi}{\text{argmin}} \sum_k w_k f_k(\xi),$$

learning the cost function parameters $\mathbf{w}$ for the optimal trajectory $\xi^*(\mathbf{w})$ that minimizes subdominance. One contribution of this paper is extending subdominance minimization to the more flexible prediction models needed for fairness-aware classification that are not directly conditioned on cost features or performance/fairness measures.

3. Subdominance Minimization for Improved Fairness-Aware Classification

We approach fair classification from an imitation learning perspective (Ziebart et al., 2022). We assume vectors of (human-provided) reference decisions are available that may have been produced collaboratively by multiple stakeholders with competing predictive performance-fairness tradeoffs. Our goal is to construct a fairness-aware classifier that outperforms reference decisions on all performance and fairness measures on withheld data as frequently as possible, which also provides guarantees to all stakeholders.

3.1. Superhumanness and Subdominance

We consider reference decisions $\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^{\text{M}}$ that are drawn from an (unknown) human decision-making process or baseline method $\tilde{\mathbb{P}}$, on a set of M items, $\mathbf{X}_{\text{M}\times\text{L}} = \{\mathbf{x}_j\}_{j=1}^{\text{M}}$, where L is the number of attributes in each of M items $\mathbf{x}_j$. Group membership attributes a_m from vector $\mathbf{a}$ indicate to which group item m belongs.

The predictive performance and fairness of decisions $\hat{\mathbf{y}}$ for each item are assessed based on ground truth $\mathbf{y}$ and group membership $\mathbf{a}$ using a set of predictive loss and unfairness measures[1] $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ (e.g., Equations 1, 2, 3). Without loss of generality, we assume that larger values for these measures are less desirable. Ideally, the set of these measures should cover the bases of all stakeholder preference functions (i.e., stakeholder cost functions for evaluating different vectors of decisions can be expressed as summed monotonic transformations of $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ measures).

Definition 3.1. A fairness-aware classifier is considered γ-superhuman for a given set of predictive loss and unfairness measures $\{f_k\}$ if its decisions $\hat{\mathbf{y}}$ satisfy:

$$P\left(\boldsymbol{f}\left(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right)\right) \geq \gamma.$$

If strict Pareto dominance is required to be γ-superhuman, which is often effectively true for continuous domains, then by definition, at most $(1-\gamma)\%$ of human decision makers could be γ-superhuman. However, far fewer than $(1-\gamma)$ may be γ-superhuman if pairs of human decisions do not Pareto dominate one another in either direction (i.e., neither $\boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}', \mathbf{y}, \mathbf{a}\right)$ nor $\boldsymbol{f}\left(\tilde{\mathbf{y}}', \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right)$ for pairs of human decisions $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$). From this perspective, any decisions with γ-superhuman performance more

[1]These measures take the place of features used to define cost/reward function in imitation learning methods. We instead use features to describe the inputs to our fairness-aware decision model, $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$.

than $(1-\gamma)\%$ of the time exceed the performance limit for the distribution of human demonstrators.

Unfortunately, directly maximizing γ is difficult in part due to the discontinuity of Pareto dominance ($\preceq$). The subdominance (Ziebart et al., 2022) serves as a convex upper bound for non-dominance in each measure $\{f_k\}$ and on $1-\gamma$ in aggregate:

$$\begin{aligned}&\text{subdom}^k_{\alpha_k}(\hat{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{y},\mathbf{a}) \triangleq \\ &\qquad [\alpha_k\left(f_k(\hat{\mathbf{y}},\mathbf{y},\mathbf{a}) - f_k(\tilde{\mathbf{y}},\mathbf{y},\mathbf{a})\right)+1]_+; \\ &\text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{y},\mathbf{a}) \triangleq \sum_k \text{subdom}^k_{\alpha_k}(\hat{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{y},\mathbf{a}).\end{aligned} \tag{9}$$

Given N vectors of reference decisions as demonstrations, $\tilde{\boldsymbol{\mathcal{Y}}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^{\text{N}}$, the subdominance for decision vector $\hat{\mathbf{y}}$ with respect to the set of demonstrations is:[2]

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

where $\hat{\mathbf{y}}_i$ is the predictions produced by our model for the set of items $\mathbf{X}_i$, and $\hat{\boldsymbol{\mathcal{Y}}}$ is the set of these prediction sets, $\hat{\boldsymbol{\mathcal{Y}}} = \{\hat{\mathbf{y}}_i\}_{i=1}^{\text{N}}$.

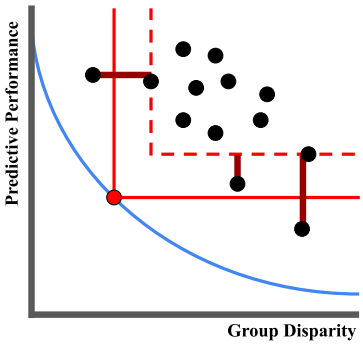

Figure 2. A Pareto frontier for possible $\hat{P}_\theta$ (blue) optimally trading off predictive performance (e.g., inaccuracy) and group unfairness. The model-produced decision (red point) defines dominance boundaries (solid red) and margin boundaries (dashed red), which incur subdominance (maroon lines) on three examples.

The subdominance is illustrated by Figure 2. Following concepts from support vector machines (Cortes & Vapnik, 1995), reference decisions $\tilde{\mathbf{y}}$ that actively constrain the predictions $\hat{\mathbf{y}}$ in a particular measure dimension, k, are referred to as *support vectors* and denoted as:

$$\tilde{\boldsymbol{\mathcal{Y}}}_{\text{SV}_k}(\hat{\mathbf{y}},\alpha_k) \triangleq \left\{\tilde{\mathbf{y}} | \alpha_k(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}})) + 1 \geq 0\right\}.$$

3.2. Performance-Fairness Subdominance Minimization

We consider probabilistic predictors $\mathbb{P}_\theta : \mathcal{X}^M \to \Delta_{\mathcal{Y}^M}$ that make structured predictions over the set of items in the most general case, but can also be simplified to make conditionally independent decisions for each item.

[2]For notational simplicity, we assume all demonstrated decisions $\tilde{\mathbf{y}} \in \tilde{\boldsymbol{\mathcal{Y}}}$ correspond to the same M items represented in $\mathbf{X}$. Generalization to unique $\mathbf{X}$ for each demonstration is straightforward.

Definition 3.2. The minimally subdominant fairness-aware classifier $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ 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 $\boldsymbol{\alpha} \triangleq \{\alpha_k\}_{k=1}^{\text{K}}$ are also learned from the data during training. For the subdominance of the k_{th} measure, α_k indicates the degree of sensitivity to how much the algorithm fails to sufficiently outperform demonstrations in that measure. When α_k value is higher, reducing underperformance on that measure minimizes the overall subdominance more than reducing underperformance on other measures.

The bi-level optimization of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ differs from single-level support vector machine optimization (of $\boldsymbol{\theta}$ alone), which is a convex optimization problem (Cortes & Vapnik, 1995). Instead, subdominance is a quasi-convex function, which similarly implies that there are no local optima as a function of the realized predictive performance/fairness measures.

Theorem 3.3. *The α_k-minimized subbdominance,*

$$\sum_k \overbrace{\min_{\alpha_k\geq 0}\left(\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right) + \lambda_k\alpha_k\right)}^{\Gamma_k(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a})}, \tag{10}$$

is a quasiconvex function in terms of the set of measures, $\{f_k(\hat{\mathbf{y}},\mathbf{y},\mathbf{a})\}$.

We use the gradient of the expected subdominance with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ to update these variables iteratively, and after convergence, the best learned weights $\boldsymbol{\theta}^*$ are used in the final model $\hat{\mathbb{P}}_{\boldsymbol{\theta}^*}$. Though subdominance minimization is not necessarily quasiconvex in terms of model parameters $\boldsymbol{\theta}$, particularly for complex, nonlinear models, stochastic gradient methods are often effective in avoiding local optima. A commonly used linear model like logistic regression can be used for $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ to simplify the overall optimization.

Theorem 3.4. *The gradient of expected subdominance under $\hat{\mathbb{P}}_\theta$ with respect to the set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$ is:*

$$\begin{aligned}&\nabla_\theta \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\sum_k \Gamma_k\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right)\right] \\ &= \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\left(\sum_k \Gamma_k(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a})\right)\nabla_\theta \log\hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X})\right],\end{aligned}$$

where the optimal α_k for each Γ_k (10) is obtained from:

$$\alpha_k = \underset{\alpha_k^{(m)}}{\text{argmin}}\, m \text{ such that } f_k\left(\hat{\mathbf{y}}\right) + \lambda \leq \frac{1}{m}\sum_{j=1}^m f_k\left(\tilde{\mathbf{y}}^{(j)}\right),$$

using $\alpha_k^{(j)} = \frac{1}{f_k(\hat{\mathbf{y}}^{(j)}) - f_k(\tilde{\mathbf{y}}^{(j)})}$ *to represent the* α_k *value that would make the demonstration with the* j_{th} *smallest* f_k *measure,* $\tilde{\mathbf{y}}^{(j)}$*, a support vector with zero subdominance.*

Using gradient descent, we update the model weights $\boldsymbol{\theta}$ using an approximation of the gradient based on a set of sampled predictions $\hat{\mathbf{y}} \in \hat{\boldsymbol{\mathcal{Y}}}$ from the model $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \left(\sum_{\hat{\mathbf{y}} \in \hat{\boldsymbol{\mathcal{Y}}}} \left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right),$$

Algorithm 1 Subdominance policy gradient optimization

Draw N set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ 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 $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$
  for $k \in \{1, ..., K\}$ **do**
    Sort reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ in ascending order by k^{th} measure value $f_k(\tilde{\mathbf{y}}_i)$: $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{\mathrm{N}}$
    Compute $\alpha_k^{(j)} = \frac{1}{f_k(\tilde{\mathbf{y}}^{(j)}) - f_k(\hat{\mathbf{y}}^{(j)})}$
    $\alpha_k = \underset{\alpha_k^{(m)}}{\operatorname{argmin}}\, m$ 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 $\Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\eta}{N} \sum_i \left(\sum_k \Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}_i|\mathbf{X}_i)$;

We show the steps for training our model in Algorithm 1. *Reference decisions* $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from a human decision-making process or baseline method $\tilde{\mathbb{P}}$ are provided as input to the algorithm. $\boldsymbol{\theta}$ is set to an initial value. In each iteration of the algorithm, we first sample a set of *model predictions* $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used for *reference decisions* $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. We then find the new $\boldsymbol{\theta}$ (and $\boldsymbol{\alpha}$) based on the algorithms discussed in Theorem 3.4.

3.3. Generalization Bounds

A fairness-aware classifier with a relatively small number of support vectors has important generalization guarantees under iid assumptions.

Theorem 3.5. *A classifier* $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ *from a family with a convex realizable space of measures* $\{f_k(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ *minimizing* $\sum_i \mathrm{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}_i, \mathbf{y}_i, \mathbf{a})$ *on a set of* N *iid reference decisions with support vector sets* $\left\{\tilde{\boldsymbol{\mathcal{Y}}}_{SV_k}(\hat{\mathbf{y}}, \alpha_k)\right\}$ *is on average* γ-**superhuman** *on the population distribution with:* $\gamma = 1 - \frac{1}{N} \left\| \bigcup_{k=1}^{K} \tilde{\boldsymbol{\mathcal{Y}}}_{S\,V_k}(\hat{\mathbf{y}}, \alpha_k) \right\|$.

The proof for this generalization bound (see Appendix A) is an extension to our setting of the generalization bound based on support vectors developed for inverse optimal control subdominance minimization (Ziebart et al., 2022). It requires that the realizable set of measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ is convex and that the (deterministic) P_θ with measures globally minimizing subdominance can be found. This may be unrealistic for complex P_θ models (e.g., multilayer neural networks).

Importantly, superhuman performance provides comparative satisfaction guarantees for stakeholders. Specifically, stakeholders will prefer the algorithmic decisions with at least γ frequency for a fairly wide range of cost functions defined in terms of the measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$.

Corollary 3.6. *For any stakeholder with a cost function,* $cost(\mathbf{f}, \mathbf{X})$ *such that:*

$$\mathbf{f}_1 \preceq \mathbf{f}_2 \implies cost(\mathbf{f}_1, \mathbf{X}) \leq cost(\mathbf{f}_2, \mathbf{X}),$$

a γ*-superhuman classifier will be preferable in expectation with probability at least:*

$$P\left(cost(\mathbf{f}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}), X) \leq cost(\mathbf{f}(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})), \mathbf{X})\right) \geq \gamma.$$

4. Experiments

The goal of our approach is to produce a fairness-aware prediction method that outperforms reference (human) decisions on multiple fairness/performance measures. In this section, we discuss our experimental design to synthesize reference decisions with varying levels of noise, evaluate our method, and provide comparison baselines.[3]

4.1. Training and Testing Dataset Construction

To emulate human decision-making with various levels of noise, we add noise to benchmark fairness datasets and apply fair learning methods over repeated randomized dataset splits. We describe this process in detail in the following section.

Datasets We perform experiments on two benchmark fairness datasets:

- UCI `Adult` dataset (Dheeru & Karra Taniskidou, 2017) considers predicting whether a household's income exceeds \$50K/yr based on census data. Group membership is based on gender. The dataset consists of 45,222 items.
- `COMPAS` dataset (Larson et al., 2016) considers predicting recidivism with group membership based on race. It consists of 6,172 examples.

[3] Our code is publicly available at `https://github.com/omidMemari/superhumn-fairness`.

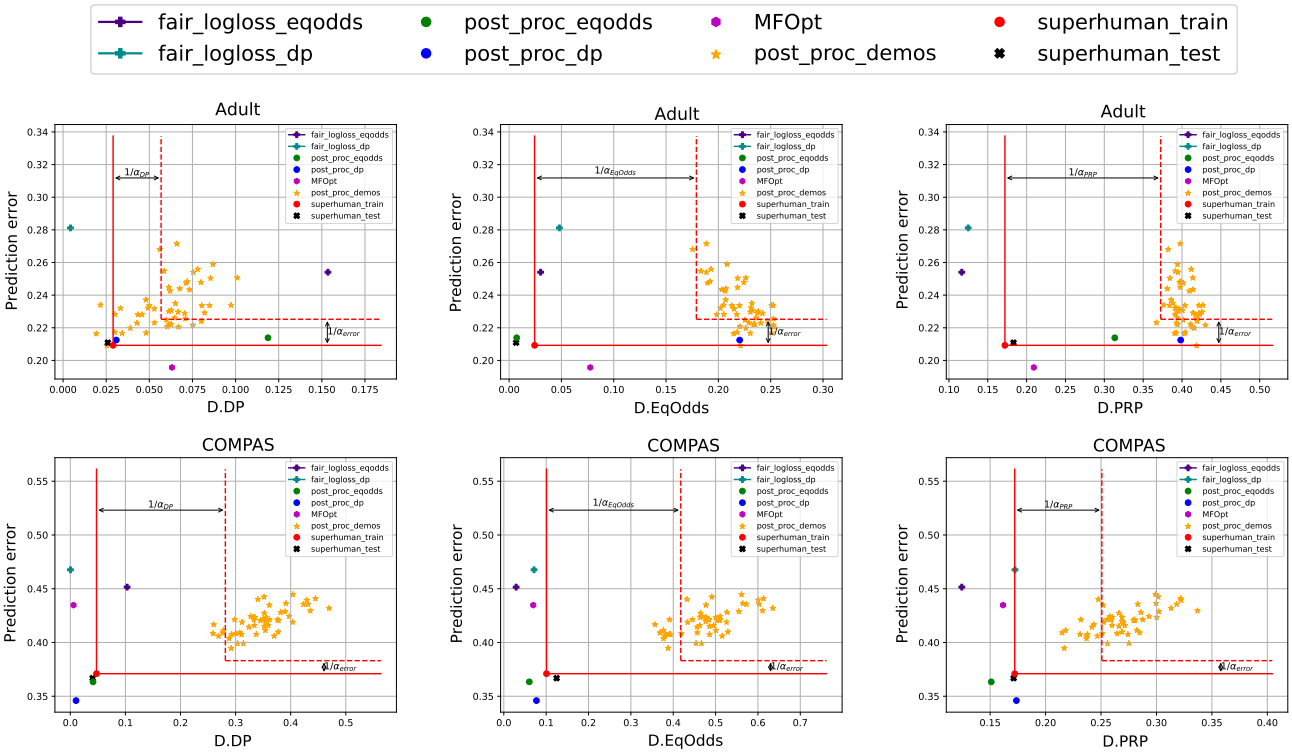

Figure 3. Prediction error versus *difference of*: *Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`) and *Predictive Rate Parity* (`D.PRP`) on test data using noiseless training data ($\epsilon = 0$) for `Adult` (top row) and `COMPAS` (bottom row) datasets.

Partitioning the data We first split the entire dataset randomly into a disjoint train (`train-all`) and test (`test-all`) set of equal size. The test set (`test-all`) is entirely withheld from the training procedure and ultimately used solely for evaluation. To produce each demonstration (a vector of reference decisions), we split the (`train-all`) set randomly into a disjoint train (`train-demo`) and test (`test-demo`) set of equal size.

Noise insertion We randomly flip $\epsilon\%$ of the ground truth labels $\mathbf{y}$ and group membership attributes $\mathbf{a}$ to add noise to our demonstration-producing process.

Fair classifier $\tilde{\mathbb{P}}$: Using the noisy data, we provide existing fairness-aware methods with labeled `train-demo` data and unlabeled `test-demo` to produce decisions on the `test-demo` data as demonstrations $\tilde{\mathbf{y}}$. Specifically, we employ:

- The **Post-processing** method of Hardt et al. (2016), which aims to reduce both *prediction error* and {*demographic parity* or *equalized odds*} at the same time. We use *demographic parity* as the fairness constraint. We produce demonstrations using this method for `Adult` dataset.
- **Robust fairness for logloss-based classification** (Rezaei et al., 2020) employs distributional robustness to match target fairness constraint(s) while robustly minimizing the log loss. We use *equalized odds* as the fairness constraint. We employ this method to produce demonstrations for `COMPAS` dataset.

We repeat the process of partitioning `train-all` $N = 50$ times to create randomized partitions of `train-demo` and `test-demo` and to then produce a set of demonstrations $\{\tilde{\mathbf{y}}\}_{i=1}^{50}$.

4.2. Evaluation Metrics and Baselines

Predictive Performance and Fairness Measures Our focus for evaluation is on outperforming demonstrations in multiple fairness and performance measures. We use $K = 4$ measures: *inaccuracy* (`Prediction error`), *difference from demographic parity* (`D.DP`), *difference from equalized odds* (`D.EqOdds`), *difference from predictive rate parity* (`D.PRP`).

Baseline methods As baseline comparisons, we train five different models on the entire train set (`train-all`) and then evaluate them on the withheld test data (`test-all`):

- The **Post-processing** model of (Hardt et al., 2016) with {*demographic parity* or *equalized odds*} as the fairness constraint (`post_proc_dp` and `post_proc_eqodds`).

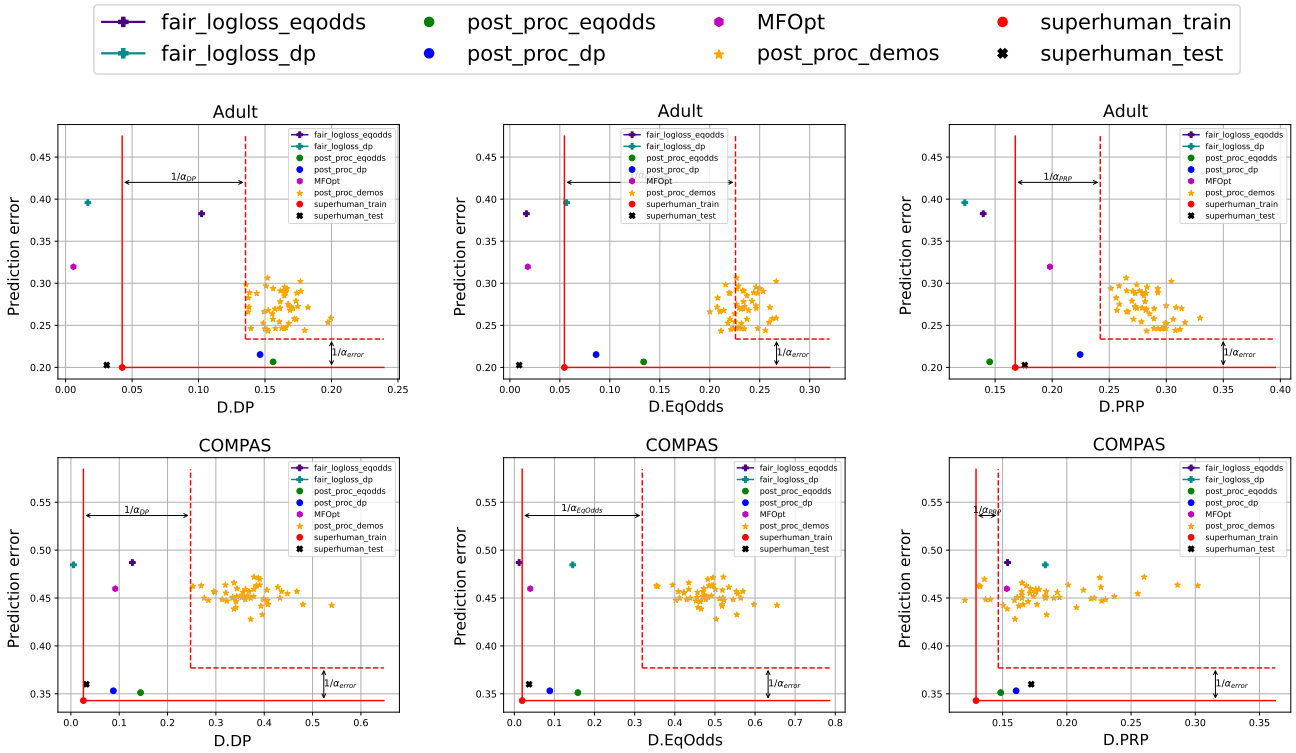

Figure 4. Experimental results on the `Adult` and `COMPAS` datasets with noisy demonstrations ($\epsilon = 0.2$). Margin boundaries are shown with dotted red lines. Each plot shows the relationships between two measures.

- The **Robust Fair-logloss** model of (Rezaei et al., 2020) with {*demographic parity* or *equalized odds*} as the fairness constraint (`fair_logloss_dp` and `fair_logloss_eqodds`).

- The **Multiple Fairness Optimization** framework of Hsu et al. (2022) which is designed to satisfy three conflicting fairness measures {*demographic parity, equalized odds*, and *predictive rate parity*} to the best extent possible (`MFOpt`).

Hinge Loss Slopes As discussed previously, each α_k value corresponds to the hinge loss slope, which defines the sensitivity of produced decision not sufficiently outperforming the demonstrations on the k_{th} measure. When the α_k is large, the model heavily weights support vector reference decisions for that particular k when minimizing subdominance. We report these values in our experiments.

4.3. Superhuman Model Specification and Updates

We use a *logistic regression* model $\mathbb{P}_{\boldsymbol{\theta}_0}$ with first-order moment feature functions, $\phi(y, \mathbf{x}) = [x_1 y, x_2 y, \dots x_m y]^\top$, and weights $\boldsymbol{\theta}$ applied independently on each item as our decision model. During the training process, we update the model parameter $\boldsymbol{\theta}$ to reduce subdominance.

Sample from Model $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ In each iteration of the algorithm, we first sample *prediction vectors* $\{\hat{\mathbf{y}}_i\}_{i=1}^{\text{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in demonstrations $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\text{N}}$. In the implementation, to produce the i_{th} sample, we look up the indices of the items used in $\tilde{\mathbf{y}}_i$, which constructs item set $\mathbf{X}_i$. Now we make predictions using our model on this item set $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$. The model produces a probability distribution for each item which can be sampled and used as a prediction $\{\hat{\mathbf{y}}_i\}_{i=1}^{\text{N}}$.

Update model parameters $\boldsymbol{\theta}$ We update $\boldsymbol{\theta}$ until convergence using Algorithm 1. For our logistic regression model, the gradient is:

$$\nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \phi(\hat{\mathbf{y}}, \mathbf{X}) - \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\phi(\hat{\mathbf{y}}, \mathbf{X})\right],$$

where ϕ denotes the feature function and $\phi(\hat{\mathbf{y}}, \mathbf{X}) = \sum_{m=1}^{\text{M}} \phi(\hat{\mathbf{y}}_m, \mathbf{x}_m)$ is the corresponding feature function for the i^{th} set of reference decisions. We employ a learning rate of $\eta = 0.01$.

4.4. Experimental Results

After training each model, e.g., obtaining the best model weight $\boldsymbol{\theta}^*$ from the training data (`train-all`) for `superhuman`, we evaluate each on unseen test data (`test-all`). We employ hard predictions (i.e., the most

Table 1. Experimental results on noise-free datasets, along with the α_k values learned for each measure in subdominance minimization.

Method \ Dataset	Adult				COMPAS			
	Prediction error	DP diff	EqOdds diff	PRP diff	Prediction error	DP diff	EqOdds diff	PRP diff
α_k	62.62	35.93	6.46	4.98	82.5	4.27	3.15	12.72
γ-**superhuman**	98%	94%	100%	100%	100%	100%	100%	100%
MinSub-Fair (ours)	0.2109	0.0259	**0.0067**	0.1831	0.3668	0.0406	0.1247	0.1712
MFOpt	**0.1957**	0.0632	0.0775	0.2092	0.4347	0.0058	0.0695	0.1616
post_proc_dp	0.2125	0.0309	0.2204	0.3983	**0.3460**	0.0104	0.0770	0.1737
post_proc_eqodds	0.2139	0.1188	0.0072	0.3135	0.3634	0.0412	0.0602	0.1510
fair_logloss_dp	0.2812	**0.0043**	0.0480	0.1248	0.4676	**0.0002**	0.0714	0.1724
fair_logloss_eqodds	0.2541	0.1535	0.0301	**0.1166**	0.4515	0.1031	**0.0291**	**0.1244**

Table 2. Experimental results on datasets with noisy demonstrations, along with the α_k values learned for each measure.

Method \ Dataset	Adult				COMPAS			
	Prediction error	DP diff	EqOdds diff	PRP diff	Prediction error	DP diff	EqOdds diff	PRP diff
α_k	29.63	10.77	5.83	13.42	29.33	4.51	3.34	57.74
γ-**superhuman**	100%	100%	100%	100%	100%	100%	100%	98%
MinSub-Fair (ours)	**0.1937**	0.0310	**0.0093**	0.1760	0.3600	0.0320	0.0367	0.1723
MFOpt	0.3157	**0.0132**	0.0225	0.2092	0.4597	0.0919	0.0397	0.1533
post_proc_dp	0.2265	0.1442	0.0879	0.2304	0.3532	0.0879	0.0884	0.1605
post_proc_eqodds	0.2176	0.1572	0.1396	0.1451	**0.3513**	0.1442	0.1584	**0.1485**
fair_logloss_dp	0.3835	0.0246	0.0577	**0.1158**	0.4846	**0.0053**	0.1455	0.1832
fair_logloss_eqodds	0.3776	0.1179	0.0238	0.1380	0.4870	0.1272	**0.0119**	0.1539

probable label) using our approach at test time rather than randomly sampling.

Noise-free reference decisions Our first set of experiments considers learning from reference decisions with no added noise.[4] The results are shown in Figure 3. We observe that our approach outperforms demonstrations in all fairness measures and shows comparable performance in *accuracy*. The (post_proc_dp) performs comparably to the average of demonstrations in all dimensions, hence our approach can outperform it in all fairness measures. In comparison to (post_proc_dp), our approach can outperform in all fairness measures but is slightly worse in *prediction error*.

We show the experiment results along with α_k values in Table 1. Note that the margin boundaries (dotted red lines) in Figure 3 are equal to $\frac{1}{\alpha_k}$ for measure k, hence there is reverse relation between α_k and margin boundary for measure k. We observe larger values of α_k for *prediction error* and *demographic parity difference*. The reason is that these measures are already optimized in demonstrations and our model has to increase α_k values for those measures to sufficiently outperform them.

Noisy reference decisions In our second set of experiments, we introduce significant amounts of noise ($\epsilon = 0.2$) into our reference decisions. We similarly add this noise to the training datasets (train-all) of the baseline methods.

[4]Added noise does not imply the original dataset is noise-free.

Table 3. Percentage of reference demonstrations that each method outperforms in all prediction/fairness measures.

Method \ Dataset	Adult		COMPAS	
	$\epsilon = 0.0$	$\epsilon = 0.2$	$\epsilon = 0.0$	$\epsilon = 0.2$
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%

The results for these experiments are shown in Figure 4. We observe that in the case of learning from noisy demonstrations, our approach still outperforms the reference decisions.

The main difference here is that due to the noisy setting, demonstrations have worse *prediction error* but regardless of this issue, our approach still can achieve a competitive *prediction error*. We show the experimental results along with α_k values in Table 2.

Relationship of noise to superhuman performance We also evaluate the relationship between the amount of augmented noise in the label and protected attribute of demonstrations, with achieving γ-superhuman performance in our approach. As shown in Figure 5, with slightly increasing the amount of noise in demonstrations, our approach can outperform 100% of demonstrations and reach 1-superhuman performance. In Table 3 we show the percentage of demonstrations that each method can outperform across all predic-

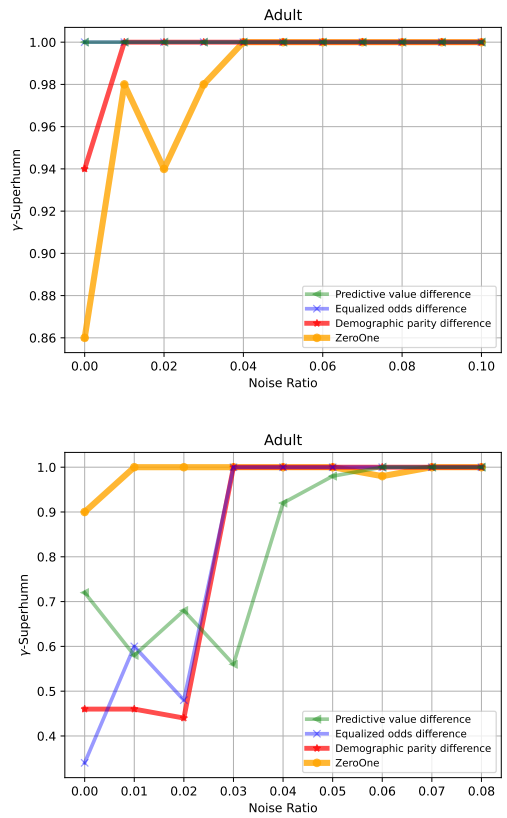

Figure 5. The relationship between the ratio of augmented noise in the label and the protected attribute of reference decisions produced by post-processing (upper) and fair-logloss (lower) and achieving γ-superhuman performance in our approach.

tion/fairness measures (i.e., the $\gamma-$superhuman value).

5. Conclusions

In this paper, we introduce superhuman fairness, an approach to fairness-aware classifier construction based on imitation learning. Our approach avoids explicit performance-fairness trade-off specification or elicitation. Instead, it seeks to unambiguously outperform human decisions across multiple performance and fairness measures with maximal frequency. When successful, this provides important guarantees for stakeholders with a broad set of possible preferences for performance and fairness measures. We develop a general framework for pursuing this based on subdominance minimization (Ziebart et al., 2022) and policy gradient optimization methods (Sutton & Barto, 2018) that enable a broad class of probabilistic fairness-aware classifiers to be learned. Our experimental results show the effectiveness of our approach in outperforming synthetic decisions corrupted by small amounts of label and group-membership noise when evaluated using multiple fairness criteria combined with predictive accuracy.

Societal impacts By design, our approach has the potential to identify fairness-aware decision-making tasks in which human decisions can frequently be outperformed by a learned classifier on a set of provided performance and fairness measures. This has the potential to facilitate a transition from manual to automated decisions that are preferred by all interested stakeholders, so long as their interests are reflected in some of those measures. Since the formulation only provides preference guarantees for stakeholders with nonnegatively-weighted combinations of performance and fairness measures, it may reduce the negative impact of stakeholders in human-produced decision-making from successfully seeking negative outcomes for specific groups.

Despite these benefits, our approach also has limitations. First, when performance-fairness tradeoffs can either be fully specified (e.g., based on first principles) or effectively elicited, fairness-aware classifiers optimized for those tradeoffs should produce better results than our approach, which operates under greater uncertainty cast by the noisiness of human decisions. Second, if target fairness concepts lie outside the set of measures we consider, our resulting fairness-aware classifier will be oblivious to them. Third, our approach assumes human-demonstrated decision are well-intentioned, noisy reflections of desired performance-fairness trade-offs. If this is not the case, then our methods could succeed in outperforming them across all fairness measures, but still not provide an adequate degree of fairness.

Future directions We have conducted experiments with a relatively small number of performance/fairness measures using a simplistic logistic regression model. Scaling our approach to much larger numbers of measures and classifiers with more expressive representations are both of great interest. Additionally, we plan to pursue experimental validation using human-provided fairness-aware decisions in addition to the synthetically-produced decisions we consider in this paper. More broadly, other techniques that can minimize subdominance or provide generalization guarantees for stakeholders adoption preferences of algorithmic decision-making are of significant interest.

Acknowledgements

This work was supported by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939743. We thank our reviewers for providing constructive feedback that helped to substantially improve the framing and presentation of the paper.

References

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

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 1418–1426, 2019.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*, 2019.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *ACM FAT**, 2019.

Chen, L. and Pu, P. Survey of preference elicitation methods. Technical report, EPFL, 2004.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Donaldson, T. and Preston, L. E. The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of management Review*, 20(1):65–91, 1995.

Dowling, A. W., Ruiz-Mercado, G., and Zavala, V. M. A framework for multi-stakeholder decision-making and conflict resolution. *Computers & Chemical Engineering*, 90:136–150, 2016.

Goel, N., Yaghini, M., and Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In *AAAI Conference on Artificial Intelligence*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323, 2016.

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

Hsu, B., Mazumder, R., Nandy, P., and Basu, K. Pushing the limits of fairness impossibility: Who's the fairest of them all? In *Advances in Neural Information Processing Systems*, 2022.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica*, 9, 2016.

Liu, S. and Vicente, L. N. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, pp. 1–25, 2022.

Martinez, N., Bertran, M., and Sapiro, G. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR, 13–18 Jul 2020.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *ACM FAT**, 2018.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2): 1–179, 2018.

Rezaei, A., Fathony, R., Memarrast, O., and Ziebart, B. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5511–5518, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 2018.

Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Vapnik, V. and Chapelle, O. Bounds on error expectation for support vector machines. *Neural computation*, 12(9): 2013–2036, 2000.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

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. PMLR, 2022.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008.

A. Proofs of Theorems

Proof of Theorem 3.3. We first establish that the average α_k-minimized subdominance of a single measure k,

$$\frac{1}{N}\sum_{\tilde{\mathbf{y}}}\min_{\alpha_k}\text{subdom}^k_{\alpha_k}(\hat{\mathbf{y}},\tilde{\mathbf{y}},\mathbf{y},\mathbf{a}) = \frac{1}{N}\sum_{\tilde{\mathbf{y}}}\left[\alpha_k^*\left(\hat{f}_k - f_k(\tilde{\mathbf{y}},\mathbf{y},\mathbf{a})\right)+1\right]_+., \tag{11}$$

is a monotonic (increasing) function of $\hat{f}_k \triangleq f_k(\hat{\mathbf{y}},\mathbf{y},\mathbf{a})$.

When $\alpha_k^* \geq 0$ is nonzero, it is minimized by defining a margin boundary at the largest support vector, $\tilde{\mathbf{y}}_{(j)}$:

$$\alpha_k^* = \frac{1}{f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \hat{f}_k}.$$

When summed over all examples, Equation (11) can be expressed as:

$$\frac{j}{N}\left(\frac{\left(\hat{f}_k - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}\right)}{f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \hat{f}_k}+1\right) = \frac{j}{N}\left(\frac{\left(f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}\right)}{f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \hat{f}_k}\right). \tag{12}$$

From the left-hand side of Eq. (12), we can see that when $\hat{f}_k$ is equal to the average features of the j (smallest) support vectors, $\overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}$, the subdominance is equal to the support vector frequency (j/N). This is also precisely the value of $\hat{f}_k$ at which a new support vector with measure value $f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a})$, is added. Starting from the left-hand side of Eq. (12), we show that this has the same value of j/N for the subdominance when $\hat{f}_k = \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}$:

$$\begin{aligned}
&\frac{j+1}{N}\left(\frac{\overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})} - \overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)},\mathbf{y},\mathbf{a})}}{f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}+1\right)\\
&= \frac{j+1}{N}\left(\frac{\overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})} - \overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)},\mathbf{y},\mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}{f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}\right)\\
&= \frac{j+1}{N}\left(\frac{-\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)},\mathbf{y},\mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}\right)\\
&= \frac{1}{N}\left(\frac{-(j+1)\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)},\mathbf{y},\mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) + jf_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}\right)\\
&\overset{(a)}{=} \frac{1}{N}\left(\frac{-j\overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})} + jf_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}}\right) = \frac{j}{N},
\end{aligned} \tag{13}$$

where step (a) follows from $(j+1)\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)},\mathbf{y},\mathbf{a})} - f_k(\tilde{\mathbf{y}}_{(j+1)},\mathbf{y},\mathbf{a}) = j\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}}$. This shows that at its non-smooth points, the subdominance is not decreasing.

Differentiating the right-hand side of Eq. (12) yields:

$$j\left(\frac{\left(f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}\right)}{(f_k(\tilde{\mathbf{y}}_{(j)},\mathbf{y},\mathbf{a}) - \hat{f}_k)^2}\right), \tag{14}$$

which is nonnegative as long as $f_k(\tilde{\mathbf{y}}_{(j)}) \geq \overline{f_k(\tilde{\mathbf{y}}_{(1:j)},\mathbf{y},\mathbf{a})}$, a condition that is always true by definition of the ordered support vectors. Thus, since subdominance is non-decreasing at both its smooth and nonsmooth portions, it is a monotonic (increasing) function of $\hat{f}_k$ in each dimension k.

Since the per-measure subdominances are independent and combined via summation over all the dimensions k to form the entire subdominance, the sublevel sets must be convex, and the subdominance overall is therefore a quasiconvex function of $\hat{\mathbf{f}}$. □

Proof of Theorem 3.4. The gradient of the training objective with respect to model parameters θ is:

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

which follows directly from a property of gradients of logs of function:

$$\nabla_\theta \log\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) = \frac{1}{\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})}\nabla_\theta\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) \implies \nabla_\theta\hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})\nabla_\theta\log\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}). \tag{15}$$

We note that this is a well-known approach employed by policy-gradient methods in reinforcement learning (Sutton & Barto, 2018).

Next, we consider how to obtain the $\alpha-$minimized subdominance for a particular tuple ($\hat{\mathbf{y}}$,$\tilde{\boldsymbol{\mathcal{Y}}}$,$\mathbf{y}$,$\mathbf{a}$), $\Gamma_k\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right) = \min_{\alpha_k}\left(\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right)+\lambda_k\alpha_k\right)$, analytically.

First, we note that $\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right)+\lambda_k\alpha_k$ is comprised of hinged linear functions of α_k, making it a convex and piece-wise linear function of α_k. This has two important implications: (1) any point of the function for which the subgradient includes 0 is a global minimum of the function (Boyd & Vandenberghe, 2004); (2) an optimum must exist at a corner of the function: $\alpha_k = 0$ or where one of the hinge functions becomes active:

$$\alpha_k(f_k(\hat{\mathbf{y}}_i) - f_k(\tilde{\mathbf{y}}_i)) + 1 = 0 \implies \alpha_k = \frac{1}{f_k(\tilde{\mathbf{y}}_i) - f_k(\hat{\mathbf{y}}_i)}. \tag{16}$$

The subgradient for the j^{th} of these points (ordered by f_k value from smallest to largest and denoted $f_k(\tilde{\mathbf{y}}^{(j)})$ for the demonstration) is:

$$\begin{aligned}
&\partial_{\alpha_k}\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a}\right)\Big|_{\alpha_k=(f_k(\hat{\mathbf{y}})-f_k(\tilde{\mathbf{y}}^{(j)}))^{-1}} = \partial_{\alpha_k}\left(\frac{1}{N}\sum_{i=1}^{j}\left[\alpha_k\left(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(i)})\right)+1\right]_+ + \lambda\alpha_k\right)\\
&= \lambda + \frac{1}{N}\sum_{i=1}^{j-1}\left(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(i)})\right) + \left[0, f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(j)})\right],
\end{aligned}$$

where the final bracketed expression indicates the range of values added to the constant value preceding it.

The smallest j for which the largest value in this range is positive must contain the 0 in its corresponding range, and is thus the provides the j value for the optimal α_k value. □

Proof of Theorem 3.5. We first recall generalization guarantees for support vector machines (SVMs) (Cortes & Vapnik, 1995) based on leave-one-out cross validation (LOOCV) that our approach leverages. For support vector machines, examples that are not support vectors incur zero loss and do not actively constrain the SVM parameters. Thus, when these examples are removed, the decision boundary does not change and therefore no cross validation loss is incurred on any left-out example during LOOCV. Due to this, the support vector frequency is an upper bound on the leave-one-out cross validation error, which is an (almost) unbiased estimate of the generalization inaccuracy (Vapnik & Chapelle, 2000).

Since subdominance is quasiconvex instead of convex, this analysis is slightly more complicated. Specifically, it requires the set of realizable $\mathbf{f}$ measures to be convex. The intersection of the sublevel sets of the quasiconvex subdominance (Theorem 3.3 with a convex set of feasible measures is also convex, so the constrained subdominance minimization problem (minimizing subdominance over the set of realizable features for the family of possible P_θ) is also quasiconvex. As a result, no local optima exist that are not global optima. Since the non-support vectors do not actively constrain the global optima, removing them does not change the global optima and therefore they do not contribute any loss to the leave-one-out cross validation error. The remaining argument then follows directly from the SVM LOOCV analysis.

□

B. Additional Results

In the main paper, we only included plots that show the relationship of a fairness metric with *prediction error*. To show the relation between each pair of fairness metrics, in Figures 6 and 7 we show the remaining plots removed from Figures 3 and 4 respectively.

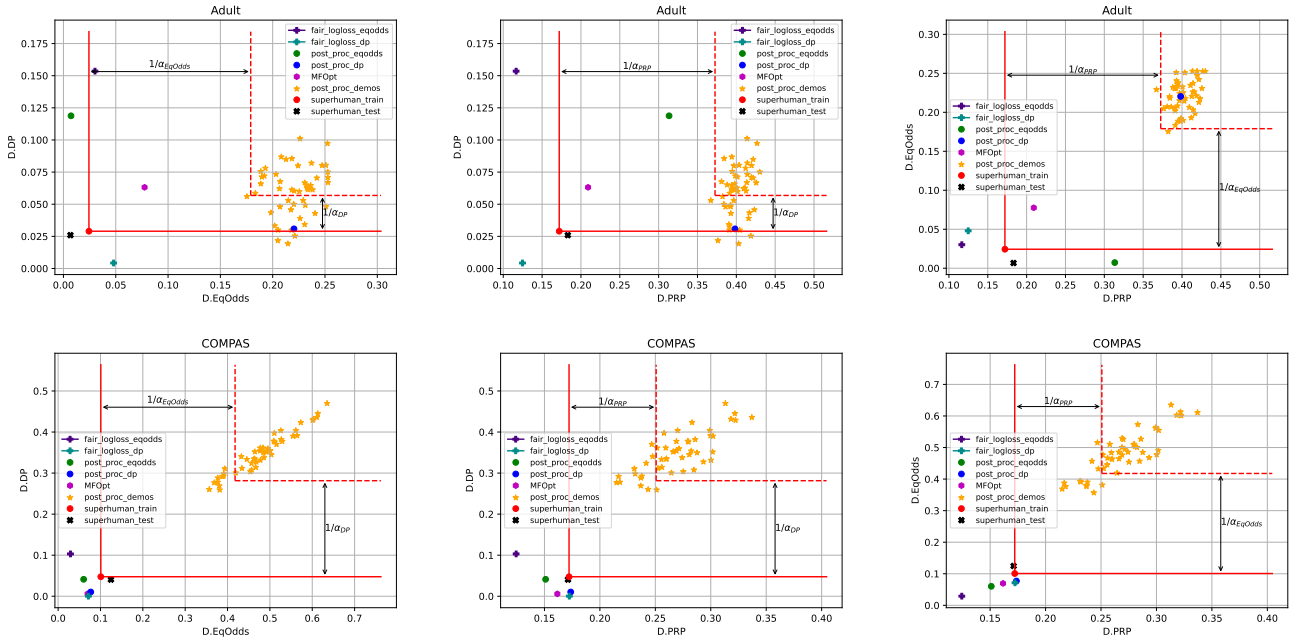

Figure 6. The trade-off between each pair of: *difference of Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`) and *Predictive Rate Parity* (`D.PR`) on test data using noiseless training data ($\epsilon = 0$) for `Adult` (top row) and `COMPAS` (bottom row) datasets.

B.1. Experiment with more measures

Since our approach is flexible enough to accept wide range of fairness/performance measures, we extend the experiment on `Adult` to $K = 5$ measures. In this experiment we use *Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`), *False Negative Rate* (`D.FNR`), *False Positive Rate* (`D.FPR`) and *Prediction Error* as the measures to outperform reference decisions on. The results are shown in Figure 8.

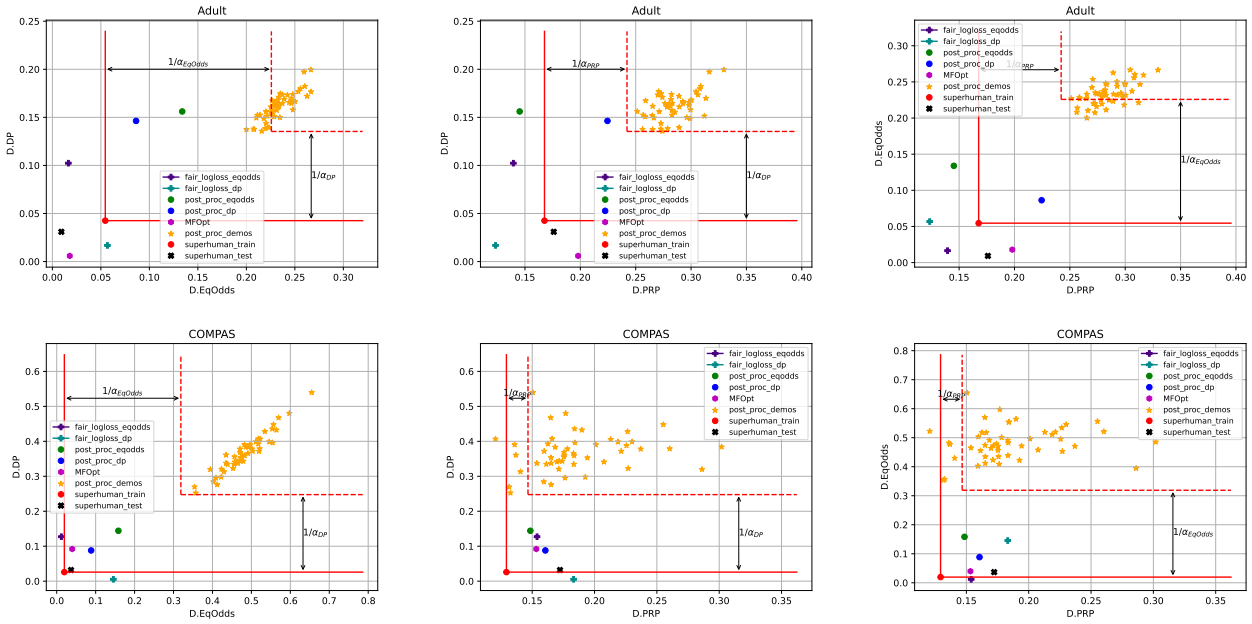

Figure 7. The trade-off between each pair of: *difference of Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`) and *Predictive Rate Parity* (`D.PR`) on test data using noiseless training data ($\epsilon = 0.2$) for `Adult` (top row) and `COMPAS` (bottom row) datasets.

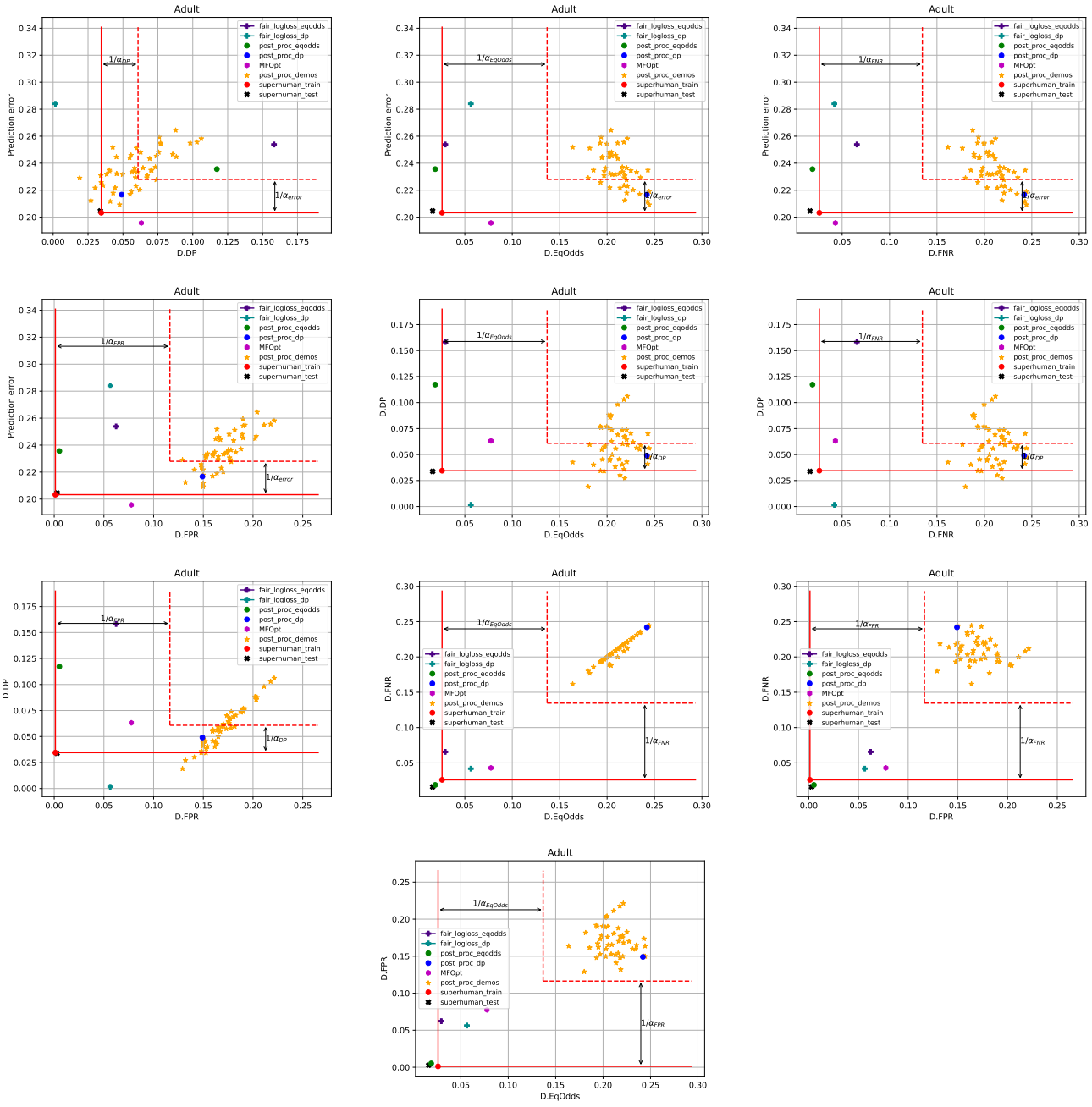

Figure 8. The trade-off between each pair of: *difference of Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`), *False Negative Rate* (`D.FNR`), *False Positive Rate* (`D.FPR`) and *Prediction Error* on test data using noiseless training data ($\epsilon = 0$) for `Adult` dataset.