Superhuman Fairness

Omid Memarrast, Linh Vu, Brian Ziebart
Department of Computer Science
University of Illinois Chicago
Chicago, IL 60607, USA
{omemar2,lvu5,bziebart}@uic.edu

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 metric(s) (e.g., demographic parity, equalized odds). This begs the question: are the right performance-fairness trade-offs being specified? We instead re-cast 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.

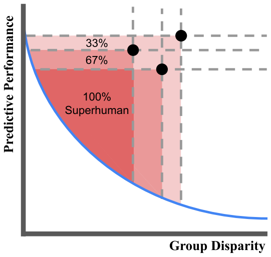

Figure 1: Three sets of decisions (black dots) with different predictive performance and group disparity values defining the sets of 100%-, 67%-, and 33%-superhuman fairness-performance values (red shades) based on Pareto dominance.

We consider the motivating scenario of a fairness-aware decision task currently being performed by well-intentioned, but inherently error-prone human decision makers. Rather than seeking optimal decisions for specific performance-fairness trade-offs, which may be difficult to accurately elicit, we propose a more modest, yet more practical objective: **outperform human decisions across all performance and fairness measures with maximal frequency**. We implicitly assume that available human decisions reflect desired performance-fairness trade-offs, but are often noisy and suboptimal. This provides an opportunity for **superhuman decisions** that Pareto dominate human decisions across predictive performance and fairness metrics (Figure 1) *without identifying an explicit desired trade-off*.

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 to estimate a cost function, we use it to directly optimize our fairness-aware classifier. We develop policy gradient optimization methods Sutton & Barto (2018) that allow flexible classes of probabilistic decision policies 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 performance 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 and Imitation Learning

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.,
$$\mathrm{P}(\hat{Y}\!=\!1|Y\!=\!y,A\!=\!1) = \mathrm{P}(\hat{Y}\!=\!1|Y\!=\!y,A\!=\!0), \ \ y \in \{0,1\};$$
- **Predictive Rate Parity** (PRP) Chouldechova (2017) requires equal positive predictive value ($\hat{y}=1$) and negative predictive value ($\hat{y}=0$) across groups:
$$\mathrm{P}(Y\!=\!1|A\!=\!1,\hat{Y}\!=\!\hat{y}) = \mathrm{P}(Y\!=\!1|A\!=\!0,\hat{Y}\!=\!\hat{y}), \ \ \hat{y} \in \{0,1\}.$$

2.2 Performance-Fairness Trade-offs

Numerous fair classification algorithms have been developed over the past few years, with most targeting one fairness metric Hardt et al. (2016). With some exceptions Blum & Stangl (2019), predictive performance and fairness are typically competing objectives in supervised machine learning approaches. 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 metrics (demographic parity, equalized odds, and predictive rate parity) to the best extent possible:

$$\min_\theta \mathbb{E}_{\hat{\mathbf{y}}\sim P_\theta}\Big[\mathrm{loss}(\hat{\mathbf{y}},\mathbf{y}) + \alpha_{\texttt{DP}}\texttt{D.DP}(\hat{\mathbf{y}},\mathbf{a}) + \alpha_{\texttt{Odds}}\texttt{D.EqOdds}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a}) + \alpha_{\texttt{PRP}}\texttt{D.PRP}(\hat{\mathbf{y}},\mathbf{y},\mathbf{a})\Big]. \quad (1)$$

2.3 Imitation Learning

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, $\tilde{\xi} = (\tilde{s}_1, \tilde{a}_1, \tilde{s_2}, \dots, \tilde{s}_T)$. 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:

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

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

$\mathbb{E}_{\xi\sim\pi}\left[f_k(\xi)\right] \leq \frac{1}{N}\sum_{i=1}^{N} f_k\left(\tilde{\xi}_i\right), \forall k,$ strict for at least one feature. Subdominance minimization Ziebart et al. (2022) seeks to produce trajectories that outperform each demonstration by a margin: $f_k(\xi) + m_k \leq f_k(\tilde{\xi}_i), \forall i, k,$ 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:

$$\text{subdom}_{\alpha}(\xi, \tilde{\xi}) \triangleq \sum_{k} \left[\alpha_k \left(f_k(\xi) - f_k(\tilde{\xi})\right) + 1\right]_{+}, \tag{2}$$

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:

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

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 metrics.

3 Subdominance Minimization for Fairness-Aware Classification

We approach fair classification from an imitation learning view. We assume vectors of (human-provided) reference decisions are available that roughly reflect desired fairness-performance tradeoffs, but are also noisy. 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.

3.1 Superhumanness and Subdominance

We consider reference decisions $\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^{\text{M}}$ that are drawn from a human decision-maker 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 $\mathbf{x}_j$. 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 $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$.

Definition 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 $\gamma-$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 $\gamma-$superhuman performance more 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 metric $\{f_k\}$ and on $1-\gamma$ in aggregate:

$$\begin{aligned} \text{subdom}_{\alpha_k}^{k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) &\triangleq \left[\alpha_k\left(f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) - f_k(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})\right) + 1\right]_{+}. \\ \text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) &\triangleq \sum_{k} \text{subdom}_{\alpha_k}^{k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}). \end{aligned} \tag{3}$$

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[1]

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

[1] 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.

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, $\hat{\boldsymbol{\mathcal{Y}}} = \{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. 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 feature dimension, k, are referred to as *support vectors* and denoted as: $\tilde{\boldsymbol{\mathcal{Y}}}_{\mathrm{SV}_k}(\hat{\mathbf{y}}, \alpha_k) = \Big\{\tilde{\mathbf{y}} | \alpha_k(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}})) + 1 \geq 0\Big\}$.

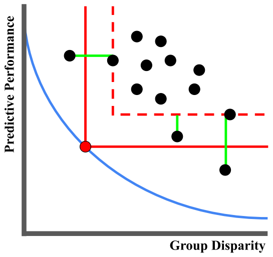

Figure 2: A Pareto frontier for possible $\hat{P}_\theta$ (blue) optimally trading off predictive performance (e.g., inaccuracy) and group unfairness. The model-produced decision (red point) defines dominance boundaries (solid red) and margin boundaries (dashed red), which incur subdominance (green lines)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

3.2 Performance-Fairness Subdominance Minimization

We consider probabilistic predictors $\mathbb{P}_\theta : \mathcal{X}^M \rightarrow \Delta_{\mathcal{Y}^M}$ 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.

Definition 2. *The minimally subdominant fairness-aware classifier* $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ *has model parameters* $\boldsymbol{\theta}$ *chosen by:*

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

Hinge loss slopes $\boldsymbol{\alpha} \triangleq \{\alpha_k\}_{k=1}^{\mathrm{K}}$ are also learned during training. α_k value defines by how far a produced decision does not sufficiently outperform the demonstrations on the k_{th} feature. When the α_k is large, the model chooses heavily weights support vector reference decisions for that particular k when minimizing subdominance.

We use the subgradient of 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 $\hat{\mathbb{P}}_{\boldsymbol{\theta}^*}$. A commonly used model like logistic regression can be used for $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$.

Theorem 1. *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:*

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

where the optimal α_k *for each* γ_k *is obtained from:*

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

using $\alpha_k^{(j)} = \frac{1}{f_k\left(\hat{\mathbf{y}}^{(j)}\right) - f_k\left(\tilde{\mathbf{y}}^{(j)}\right)}$ *to represent the* α_k *value that would make the demonstration with the* j_{th} *smallest* f_k *feature,* $\tilde{\mathbf{y}}^{(j)}$, *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 $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$:

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

Algorithm 1 shows the steps required for the training of our model. *Reference decisions* $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from a baseline method $\tilde{\mathbb{P}}$ are provided as input to the algorithm. In each iteration, 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 1.

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.

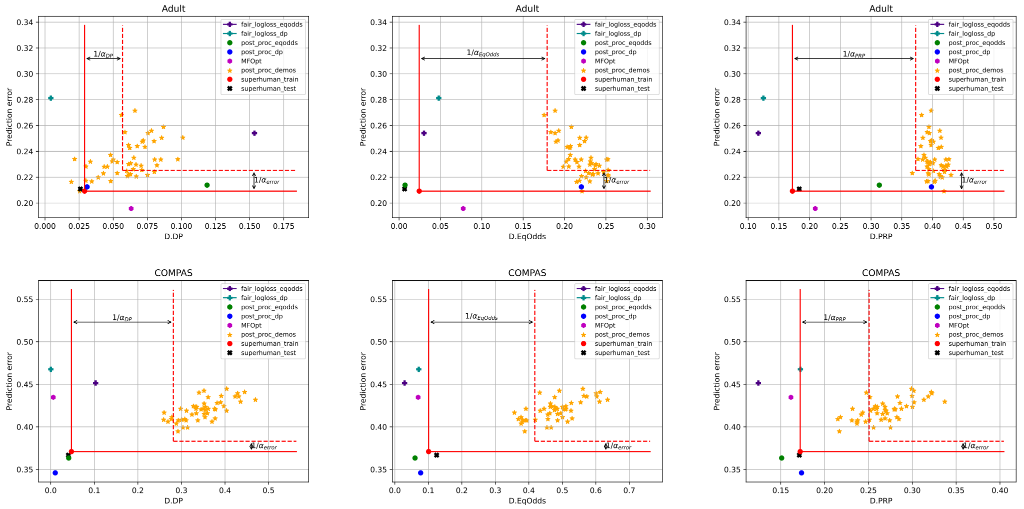

Figure 3: *Prediction error* versus *difference of*: *Demographic Parity* (D.DP), *Equalized Odds* (D.EqOdds) and *Predictive Rate Parity* (D.PRP) on test data using noiseless training data ($\epsilon = 0$) for Adult (top row) and COMPAS (bottom row) datasets.

4.1 Training and Testing Dataset

With repeated randomized splits of benchmark fairness datasets, we apply fair learning methods over noise-added ground truth data to emulate human decisions. We will describe this process in detail.

Datasets We perform experiments on two benchmark fairness datasets:

- UCI Adult Dheeru & Karra Taniskidou (2017).
- ProPublica's COMPAS Larson et al. (2016).

Partitioning the data We first split the entire dataset randomly into a disjoint train (tr-all) and test (ts-all) set of equal size. The test set (ts-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 (tr-all) set, randomly into a disjoint train (tr-demo) and test (ts-demo) set of equal size.

Algorithm 1: Subdominance 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 based on their $k^{\text{th}}$
          feature value $f_k(\tilde{\mathbf{y}}_i)$: $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{\mathrm{N}}$;
        Compute $\alpha_k^{(j)} = \frac{1}{f_k(\tilde{\mathbf{y}}^{(j)}) - f_k(\hat{\mathbf{y}}^{(j)})}$;
        $\alpha_k = \operatorname{argmin}_{\alpha_k^{(m)}} m$
         such that
          $f_k\left(\hat{\mathbf{y}}^{(j)}\right) \leq \frac{1}{m}\sum_{j=1}^{m} f_k\left(\tilde{\mathbf{y}}^{(j)}\right)$;
        Compute $\Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})$;
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} +$
    $\frac{\eta}{N}\sum_i \left(\sum_k \Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})\right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}_i|\mathbf{X}_i)$;
```

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

Fair classifier $\tilde{\mathbb{P}}$ Using the noisy data, we provide existing fairness-aware methods with labeled tr-demo data and unlabeled ts-demo to produce decisions on the ts-demo data as demonstrations $\tilde{\mathbf{y}}$. Specifically, we employ the **Post-processing** method of Hardt et al. (2016) with *DP* as the fairness constraint for Adult dataset and **Robust fair logloss** method of Rezaei et al. (2020) with *EqOdds* as the fairness constraint for COMPAS dataset. We repeat the process of partitioning tr-all $N = 50$ times to create randomized partitions of tr-demo and ts-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 $K = 4$ measures: *inaccuracy* (Prediction error), *difference of demographic parity* (D.DP), *difference of equalized odds* (D.EqOdds), *difference of predictive rate parity* (D.PRP).

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

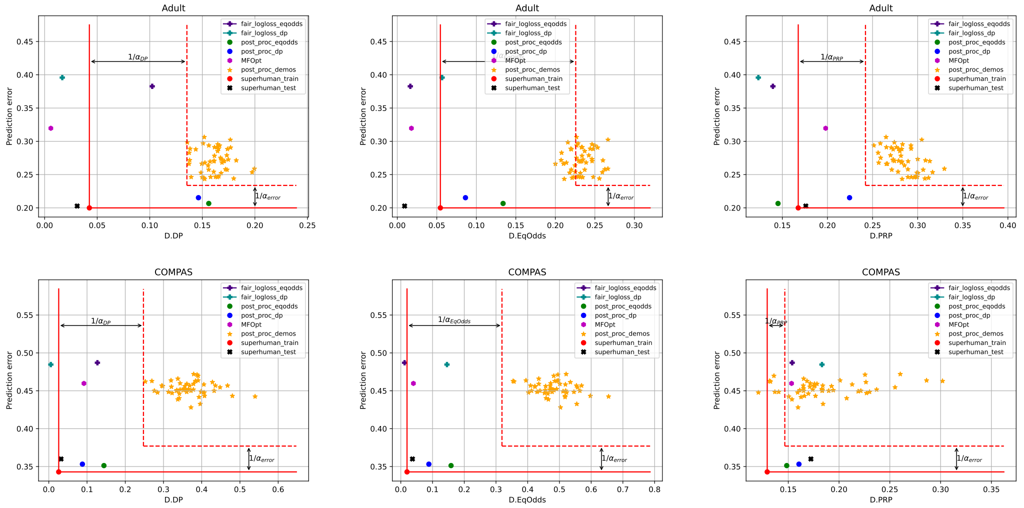

Figure 4: Experimental results on the `Adult` and `COMPAS` datasets with noisy demonstrations ($\epsilon = 0.2$).

- The **Post-processing** model of Hardt et al. (2016) with {DP and EqOdds} as the fairness constraint (`post_proc_dp` and `post_proc_eqodds`).
- The **Robust Fair-logloss** model of Rezaei et al. (2020) with {DP and EqOdds} 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 metrics {DP, EqOdds and PRP} to the best extent possible (`MFOpt`).

4.3 Superhuman Model Specification and Updates

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* $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in demonstrations $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. 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 $\mathbf{X}_i$. 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 $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$.

Update model parameters $\boldsymbol{\theta}$ We update $\boldsymbol{\theta}$ until convergence using Algorithm 1.

4.4 Experimental Results

After obtaining the best model weight $\boldsymbol{\theta}^*$ from the training data (`tr-all`), we evaluate our model on unseen test data (`ts-all`). We employ hard predictions (i.e., the most probable label) using our approach at test time rather than random sampling.

Noise-free reference decisions Our first set of experiments considers learning from reference decisions with no added noise. The results are shown in Figure 3. We observe that our approach outperforms demonstrations in all fairness metrics and shows comparable performance in *accuracy*. Note that the margin boundaries (dotted red lines) in Figure 3 are equal to $\frac{1}{\alpha_k}$ for feature k, hence there is reverse relation between α_k and margin boundary for feature k. We observe larger values of α_k for *prediction error* and *demographic parity difference*. The reason is that these features are already optimized in demonstrations and our model has to increase α_k values for those features to sufficiently outperform them.

Noisy reference decisions In this set of experiments, we introduce significant amounts of noise ($\epsilon = 0.2$) into our reference decisions and also training data of baseline methods. The results are shown in Figure 4. In the case of learning from noisy demonstrations, our approach still outperforms the reference decisions. Due to the noisy setting, demonstrations have worse *prediction error* but regardless of this issue, our approach still can achieve a competitive *prediction error*.

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 to 1-superhuman performance. In Table 1 we show the percentage of demonstrations that each method can outperform across all prediction/fairness measures (i.e., the $\gamma-$superhuman value).

Table 1: Percentage of demonstrations that each method outperforms in all prediction/fairness measures.

Method	`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%

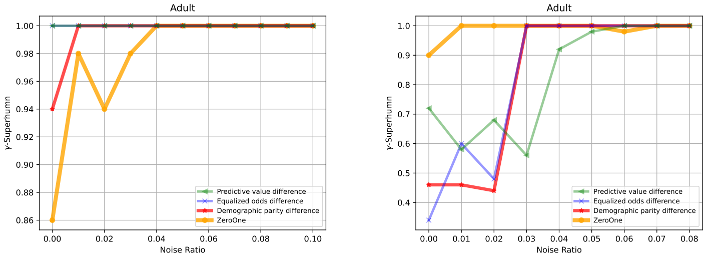

Figure 5: The relationship between the ratio of augmented noise in the label and the protected attribute of reference decisions produced by post-processing (left) and fair-logloss (right) and achieving γ-superhuman performance in our approach.

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. 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. However, our approach 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 trade-offs 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 metrics 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.

6 Acknowledgements

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

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1–8, 2004.

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

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

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

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *ACM FAT**, 2019.

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

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

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

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: Who's the fairest of them all? In *Advances in Neural Information Processing Systems*, 2022.

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

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, 9, 2016.

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

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto fairness: A multi objective perspective. In *Proceedings of the International Conference on Machine Learning*, pp. 6755–6764. PMLR, 13–18 Jul 2020.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *ACM FAT**, 2018.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2): 1–179, 2018.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5511–5518, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

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

Brian Ziebart, Sanjiban Choudhury, Xinyan Yan, and Paul Vernaza. Towards uniformly superhuman autonomy via subdominance minimization. In *International Conference on Machine Learning*, pp. 27654–27670. PMLR, 2022.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pp. 1433–1438, 2008.

A PROOFS OF THEOREMS

Proof of Theorem 1. The gradient of the training objective with respect to model parameters θ is:

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

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

$$\nabla_\theta \log\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) = \frac{1}{\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})}\nabla_\theta\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) \implies \nabla_\theta\hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})\nabla_\theta\log\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}). \tag{4}$$

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:

$$\alpha_k(f_k(\hat{\mathbf{y}}_i) - f_k(\tilde{\mathbf{y}}_i)) + 1 = 0 \implies \alpha_k = \frac{1}{f_k(\tilde{\mathbf{y}}_i) - f_k(\hat{\mathbf{y}}_i)}. \tag{5}$$

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:

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

$$= \lambda + \frac{1}{N}\sum_{i=1}^{j-1}\left(f_k(\hat{\mathbf{y}})-f_k(\tilde{\mathbf{y}}^{(i)})\right) + \left[0, f_k(\hat{\mathbf{y}})-f_k(\tilde{\mathbf{y}}^{(j)})\right],$$

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 2. We extend the leave-one-out generalization bound of Ziebart et al. (2022) by considering the set of reference decisions that are support vectors for any learner decisions with non-zero probability. For the remaining reference decisions that are not part of this set, removing them from the training set would not change the optimal model choice and thus contribute zero error to the leave-one-out cross validation error, which is an almost unbiased estimate of the generalization error Vapnik & Chapelle (2000).

□

B GENERALIZATION BOUNDS

With a small effort, we extend the generalization bounds based on support vectors developed for inverse optimal control subdominance minimization Ziebart et al. (2022).

Theorem 2. *A classifier $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ trained to minimize* $\operatorname{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}_i)$ *on a set of N iid reference decisions has the support vector set* $\left\{\bigcup_{\hat{\mathbf{y}}:P_\theta(\hat{\mathbf{y}}|\mathbf{X})>0} \tilde{\boldsymbol{\mathcal{Y}}}_{SV_k}(\hat{\mathbf{y}}, \alpha_k)\right\}$ *defined by the union of support vectors for any decision with support under $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$. Such a classifier is on average γ-***superhuman** *on the population distribution with:* $\gamma = 1 - \frac{1}{N}\left\|\bigcup_{k=1}^{K}\bigcup_{\hat{\mathbf{y}}:P_\theta(\hat{\mathbf{y}}|\mathbf{X})>0} \tilde{\boldsymbol{\mathcal{Y}}}_{S\,V_k}(\hat{\mathbf{y}}, \alpha_k)\right\|$.

This generalization bound requires overfitting to the training data so that the $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ has restricted support (i.e., $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(\hat{\mathbf{y}}|\mathbf{X}) = 0$ for many $\hat{\mathbf{y}}$) or becomes deterministic.

C Additional Results

We show the numerical results for noiseless and noisy experiements shown in Figures 3 and 4 in Tables 2 and 3, respectively.

Table 2: Experimental results on noise-free datasets, along with the α_k values learned for each feature in subdominance minimization.

Dataset / Method	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
$\gamma-$**superhuman**	98%	94%	100%	100%	100%	100%	100%	100%
MinSub-Fair (ours)	0.210884	0.025934	**0.006690**	0.183138	0.366806	0.040560	0.124683	0.171177
MFOpt	**0.195696**	0.063152	0.077549	0.209199	0.434743	0.005830	0.069519	0.161629
post_proc_dp	0.212481	0.030853	0.220357	0.398278	0.345964	0.010383	0.077020	0.173689
post_proc_eqodds	0.213873	0.118802	0.007238	0.313458	**0.363395**	0.041243	0.060244	0.151040
fair_logloss_dp	0.281194	**0.004269**	0.047962	0.124797	0.467610	**0.000225**	0.071392	0.172418
fair_logloss_eqodds	0.254060	0.153543	0.030141	**0.116579**	0.451496	0.103093	**0.029085**	**0.124447**

Table 3: Experimental results on datasets with noisy demonstrations, along with the α_k values learned for each feature.

Dataset / Method	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
$\gamma-$**superhuman**	100%	100%	100%	100%	100%	100%	100%	98%
MinSub-Fair (ours)	**0.202735**	0.030961	**0.009263**	0.176004	0.359985	0.031962	0.036680	0.172286
MFOpt	0.319696	**0.005651**	0.017385	0.198472	0.459731	0.091892	0.039745	0.153257
post_proc_dp	0.225462	0.064232	0.237852	0.400427	0.353164	0.087889	0.088414	0.160538
post_proc_eqodds	0.224561	0.103158	0.010552	0.310070	**0.351269**	0.144190	0.158372	**0.148493**
fair_logloss_dp	0.285549	0.007576	0.057659	**0.115751**	0.484620	**0.005309**	0.145502	0.183193
fair_logloss_eqodds	0.254577	0.147932	0.012778	0.118041	0.487025	0.127163	**0.011918**	0.153869

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.

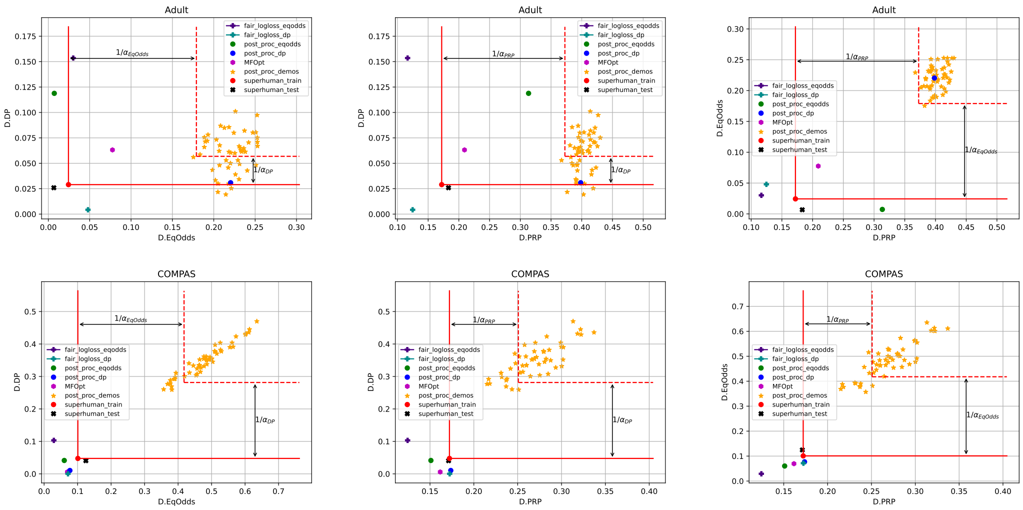

Figure 6: The trade-off between each pair of: *difference of Demographic Parity* (D.DP), *Equalized Odds* (D.EqOdds) and *Predictive Rate Parity* (D.PR) on test data using noiseless training data ($\epsilon = 0$) for Adult (top row) and COMPAS (bottom row) datasets.

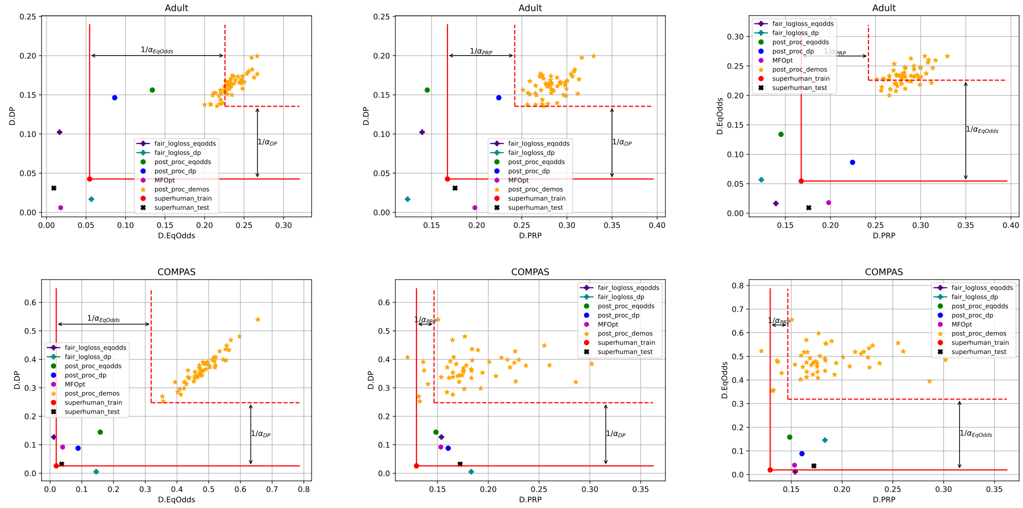

Figure 7: The trade-off between each pair of: *difference of Demographic Parity* (D.DP), *Equalized Odds* (D.EqOdds) and *Predictive Rate Parity* (D.PR) on test data using noiseless training data ($\epsilon = 0.2$) for Adult (top row) and COMPAS (bottom row) datasets.

C.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 $K = 5$ features. 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 features to outperform reference decisions on. The results are shown in Figure 8.

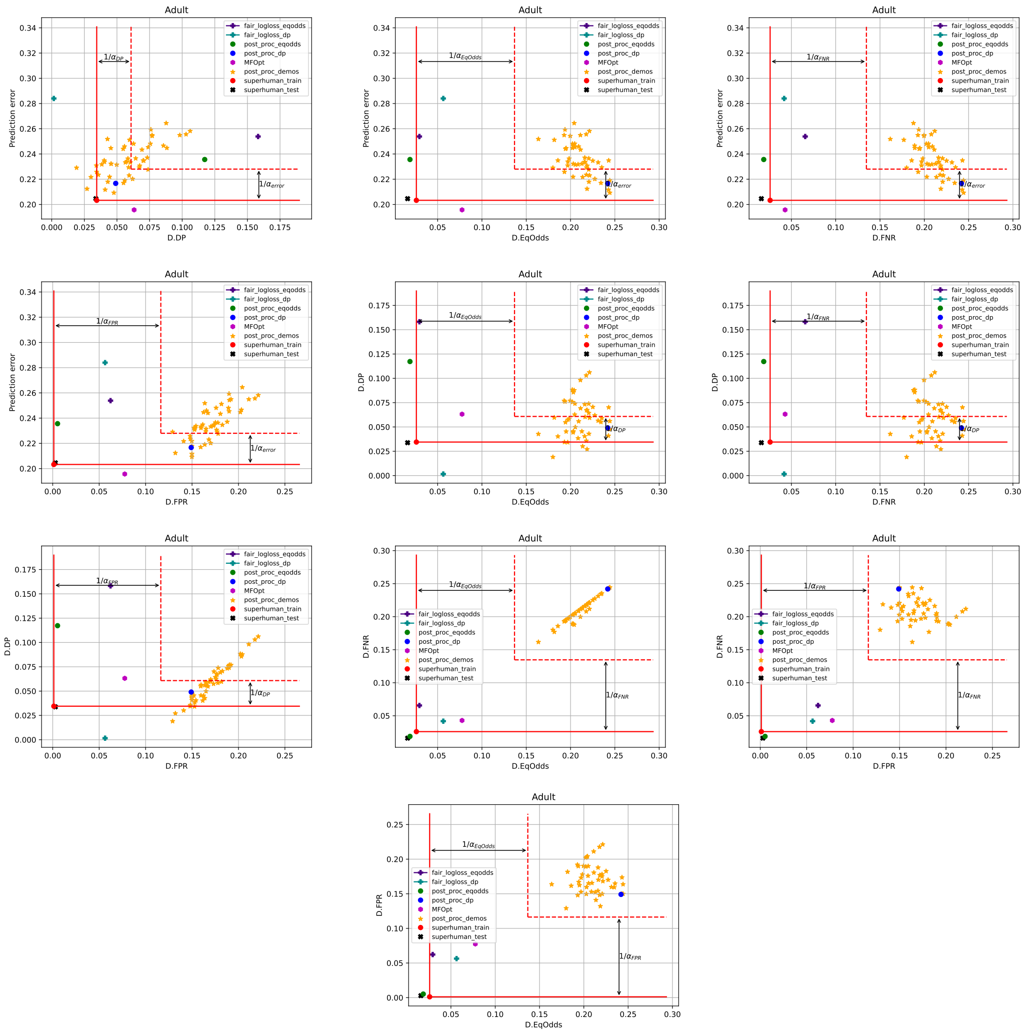

Figure 8: The trade-off between each pair of: *difference of Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`), *False Negative Rate* (`D.FNR`), *False Positive Rate* (`D.FPR`) and *Prediction Error* on test data using noiseless training data ($\epsilon = 0$) for `Adult` dataset.

D Group Fairness Violations

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:

$$\texttt{D.DP}(\hat{\mathbf{y}}, \mathbf{a}) = \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0\right]} \right|; \tag{6}$$

$$\texttt{D.EqOdds}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, y_i = y, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1, y_i = y\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, y_i = y, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0, y_i = y\right]} \right|; \tag{7}$$

$$\texttt{D.PRP}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[y_i = 1, \hat{y}_i = y, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1, \hat{y}_i = y\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[y_i = 1, \hat{y}_i = y, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0, \hat{y}_i = y\right]} \right|. \tag{8}$$