Blind Pareto Fairness and Subgroup Robustness

Natalia Martinez [*1] **Martin Bertran** [*1] **Afroditi Papadaki** [2] **Miguel Rodrigues** [2] **Guillermo Sapiro** [1]

Abstract

Much of the work in the field of group fairness addresses disparities between *predefined* groups based on protected features such as gender, age, and race, which need to be available at train, and often also at test, time. These approaches are static and retrospective, since algorithms designed to protect groups identified *a priori* cannot anticipate and protect the needs of different at-risk groups in the future. In this work we analyze the space of solutions for worst-case fairness beyond demographics, and propose *Blind Pareto Fairness* (BPF), a method that leverages no-regret dynamics to recover a fair minimax classifier that reduces worst-case risk of *any* potential subgroup of sufficient size, and guarantees that the remaining population receives the best possible level of service. BPF addresses *fairness beyond demographics*, that is, it does not rely on predefined notions of at-risk groups, neither at train nor at test time. Our experimental results show that the proposed framework improves worst-case risk in multiple standard datasets, while simultaneously providing better levels of service for the remaining population. The code is available at github.com/natalialmg/BlindParetoFairness.

1. Introduction

A large body of literature has shown that machine learning (ML) algorithms trained to maximize average performance on existing datasets may present discriminatory behaviour across pre-defined demographic groups (Barocas & Selbst, 2016; Hajian et al., 2016), meaning that segments of the overall population are measurably under-served by the ML model. This has sparked interest in the study on why these disparities arise, and on how they can be addressed (Mitchell et al., 2018; Chouldechova & Roth, 2018; Barocas et al., 2019). One popular notion is *group fairness*, where the algorithm has access to a set of predefined demographic groups during training, and the goal is to learn a model that satisfies a certain notion of fairness across these groups (e.g., statistical parity, equality of opportunity) (Dwork et al., 2012; Hardt et al., 2016); this is usually achieved by adding a constraint to the standard optimization objective. It has been shown that optimality may be in conflict with some notions of fairness (e.g., the optimal risk is different across groups) (Kaplow & Shavell, 1999; Chen et al., 2018), and perfect fairness can, in general, only be achieved by degrading the performance on the benefited groups without improving the disadvantaged ones. This conflicts with notions of no-harm fairness such as in (Ustun et al., 2019), which are appropriate where quality of service is paramount. Notions such as minimax fairness, commonly known as Rawlsian max-min fairness from an utility maximization perspective (Rawls, 2001; 2009), combined with Pareto efficiency, naturally address this no-harm concern (Martinez et al., 2020; Diana et al., 2020).

Recent works study fairness in ML when no information about the protected demographics is available, for example, due to privacy or legal regulations (Kallus et al., 2019). This is an important research direction and has been identified as a major industry concern (Veale & Binns, 2017; Holstein et al., 2019), since many applications and datasets in ML currently lack demographic records. We therefore study the problem of building minimax Pareto fair algorithms beyond demographics, meaning that not only we lack group membership records but also have no prior knowledge about the demographics to be considered (e.g., any subset of the population can be a valid protected group, see computationally identifiable groups (Hébert-Johnson et al., 2018)). This has the advantage of making the model robust to any potential demographic even if they are unknown at the time of design, or change through time; it is also efficient, since the model offers the best level of service to all the remaining (i.e., non-critical) population.

Main Contributions. We analyze *subgroup robustness*, where a model is minimax fair w.r.t. any group of sufficient size, regardless of any preconceived notion of protected groups; we also adhere to the notion of *no-harm* fairness by requiring our minimax model to be Pareto efficient (Mas-

[*]Equal contribution [1]Duke University, Durham, NC, USA [2]University College London, London, UK. Correspondence to: Natalia Martinez <natalia.martinez@duke.edu>, Martin Bertran <martin.bertran@duke.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Colell et al., 1995) by providing the best level of service to the remaining population. A model with these characteristics has a performance guarantee even for unidentified protected classes.

We show that being *subgroup robust* w.r.t. an unknown number of groups, where no individual group is smaller than a certain size, is mathematically equivalent in terms of worst group performance to solving a simplified two-group problem, where the population is divided into high and low risk groups, thereby providing a clear means to design "universal" minimax fair ML models. We further show the critical role of the minimum group size by proving that, for standard classification losses (cross-entropy and Brier score), there is a limit to the smallest group size we can consider before the solution degenerates to a trivial, uniform classifier, a similar result is also demonstrated for losses where there is a preferred outcome; a common scenario in fair ML. We additionally study the cost of blind subgroup robustness when compared to learning a model that is minimax fair w.r.t. predefined demographics.

We then propose *Blind Pareto Fairness* (BPF), a simple learning procedure that leverages recent methods in no-regret dynamics (Chen et al., 2017) to solve *subgroup robustness* subject to a user-defined minimum subgroup size. Our method is provably convergent and can be used on classification and regression tasks. We experimentally evaluate our method on a variety of standard ML datasets and show that it effectively reduces worst-case risk and compares favourably with previous works in the area. Although our work is motivated by fairness, subgroup robustness has applications beyond this important problem, see for example (Sohoni et al., 2020a; Duchi et al., 2020).

2. Related Work

A body of work has addressed fairness without explicit demographics by using proxy variables to impute the protected population labels (Elliott et al., 2008; Gupta et al., 2018; Zhang, 2018). These methods contrast with our assumptions by relying on a preconceived notion on what the protected demographics are (i.e., the protected demographics are known, but unobserved), since prior knowledge is needed to design useful proxy variables. Moreover, it has been reported that these approaches can exacerbate disparities by introducing undesired bias (Chen et al., 2019; Kallus et al., 2019); aiming to be fair by inferring protected attributes may be in conflict with privacy or anonymity concerns. These works might need re-training if new protected classes are identified, since a model trained under these conditions may be considerably harmful on an unknown population. This phenomena further supports the value of blind subgroup robustness.

Individual fairness (Dwork et al., 2012) provides guarantees beyond protected attributes, but requires predefined similarity functions which may be hard or infeasible to design for real-world tasks. The works of (Hébert-Johnson et al., 2017; Kearns et al., 2018) address fairness w.r.t. subgroups based solely on input features, and while these works greatly extend the scope of the protected demographics, they still rely on labeled protected features for guidance. The work of (Sohoni et al., 2020b) partitions the input space to address robust accuracy. We note that partitions,[1] based only on the input space of the model do not modify the solution of risk-based Pareto optimal models, since the optimal classifier for any input value remains unchanged (i.e., there is no conflict between objectives for any value of the input space, see Theorem 4.1 in (Martinez et al., 2020)). In our work we consider subgroups based on both outcome and all input features, which broadens the scope to all conceivable subgroups based on the information available to the trainer. For many risk-based measures of utility, such as crossentropy, Brier score, or ℓ_2 regression loss, the optimal classifier can be expressed as a function of the conditional output probability $p(Y|X)$, X being the input (features) and Y the output. In particular, if we only consider groups that introduce covariate shift (i.e., $p(X|A)$ varies across different values of the group membership A) but do not change the conditional target distribution ($p(Y|X, A) = p(Y|X)$ for all A), then the set of Pareto classifiers only contain one element and the Pareto curve degenerates to the utopia point. By specifically taking outcomes into account in our partition function, we allow for robustness to perturbations on the conditional distribution $p(Y|X, A)$.

There are two recent approaches that are the closest to our objective (protecting unknown and unobserved demographics). One is distributionally robust optimization (DRO) (Hashimoto et al., 2018; Duchi et al., 2020), where the goal is to achieve minimax fairness for unknown populations of sufficient size. Similar to our work, they minimize the risk of the worst-case group for the worst-case group partition, they use results from distributional robustness that focus the attention of the model exclusively on the high-risk samples (i.e., their model reduces the tail of the risk distribution). However, they do not explicitly account for Pareto efficiency, meaning that their solution may be sub-optimal on the population segment that lies below their high-risk threshold, doing unnecessary harm. The other recent method that tackles the minimax objective is adversarially reweighted learning (ARL) (Lahoti et al., 2020), where the model is trained to reduce a positive linear combination of the sample errors, these weighting coefficients are proposed by an adversary (implemented as a neural network), with the goal of maximizing the weighted empirical error. This method focuses

[1] In this work we consider "partition" and "subgroup" interchangeable.

on computationally identifiable subgroups, meaning that they can be characterized by a function $f : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, (Hébert-Johnson et al., 2018). However, they do not provide an optimality guarantee on the adversary, nor do they pose a constraint on the computationally identifiable subgroups. Our theoretical results indicate that adding an easily interpretable group size constraint on this subgroup is necessary so that the worst-case partition does not yield a trival, uniform classifier for the optimal adversary; this observation is validated in experimental results.

3. Problem Formulation

3.1. Minimax Fairness

We first consider the supervised group fairness classification scenario (Barocas et al., 2019), where we have access to a dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^n \sim p(X, Y, A)^{\otimes n}$ containing n i.i.d. triplets. Here $X \in \mathcal{X}$ denotes the input features, $Y \in \mathcal{Y}$ the categorical target variable, and $A \in \mathcal{A}$ group membership. We consider a classifier $h \in \mathcal{H}$ belonging to an hypothesis class $\mathcal{H}$ whose goal is to predict Y from X, $h : \mathcal{X} \to \Delta^{|\mathcal{Y}|-1}$; note that $h(X)$ can take any value in the simplex and is readily interpretable as a distribution over labels Y. Given a loss function $\ell : \Delta^{|\mathcal{Y}|-1} \times \Delta^{|\mathcal{Y}|-1} \to \mathbb{R}^+$, fairness is considered in the context of a Multi-Objective Optimization Problem (MOOP), where the objective is to learn a classifier that minimizes the conditional group risks $\mathbf{r}(h) = \{r_a(h)\}_{a\in\mathcal{A}}$,

$$\begin{aligned} &\min_{h\in\mathcal{H}}(r_1(h), ..., r_{|\mathcal{A}|}(h)), \\ &r_a(h) = \mathbb{E}_{X,Y|A=a}[\ell(h(X), Y)]. \end{aligned} \tag{1}$$

The solution to this MOOP may not be unique (e.g., the optimal classifier of different groups differs), and therefore there is a set of optimal (Pareto) solutions that can be achieved. It is possible that none of these Pareto solutions satisfy some group fairness criteria (e.g., equality of risk), meaning that achieving perfect fairness comes at the cost of optimality (Kaplow & Shavell, 1999; Bertsimas et al., 2011). In this work we do not to compromise optimality, meaning that we do not degrade the performance of a low-risk group if it does not directly benefit another, and consider a minimax fairness approach (Rawls, 2001; 2009), where the goal is to find a Pareto optimal classifier that minimizes the worst-case group risk,

$$\min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{a\in\mathcal{A}} r_a(h). \tag{2}$$

$\mathcal{H}_{P_\mathcal{A}}$ represents the set of properly Pareto optimal classifiers in $\mathcal{H}$ given a group set $\mathcal{A}$ as defined next.

Definition 3.1. An hypothesis $h^* \in \mathcal{H}$ is Pareto optimal if h^* is Pareto efficient, meanining that $\nexists h' \in \mathcal{H} : \mathbf{r}(h') \prec \mathbf{r}(h)$, [2]. Given a partition set $\mathcal{A}$ we denote the set of Pareto hypothesis in $\mathcal{H}$ as $\mathcal{H}_{P_\mathcal{A}}$.

[2] $\mathbf{r}(h') \prec \mathbf{r}(h)$ if $r_a(h') \leq r_a(h) \forall a \wedge \exists a' : r_{a'}(h') < r_{a'}(h)$

Note that Definition 3.1 establishes that there is no other model in the hypothesis class whose associated group risks are uniformly better for all groups.

3.2. Blind Pareto Fairness

In this work we consider a more challenging problem, namely *Blind Pareto Fairness* (BPF), where the group variable A and the conditional distribution $p(A|X, Y)$ are completely unknown (not just unobserved), even at training time. Here the goal is to learn a model that has the best performance on the worst-group risk of the worst partition density $p(A|X, Y)$ ("sensitive" group assignment), subject to a group size constraint ($p(A=a) \geq \rho, \forall a$). We formulate the following new problem,

$$R^* = \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{p(A|X,Y) \\ s.t. p(A) \succeq \rho}} \max_{a\in\mathcal{A}} r_a(h). \tag{3}$$

Here R^* is the minimum worst group error achieved for the worst partition density with known number of partitions $|\mathcal{A}|$. Since $\mathcal{H}_{P_\mathcal{A}}$ is explicitly dependent on $p(A|X, Y)$, the objective presented in Problem 3 seems ill-defined because the learner has to pick h from the Pareto hypotheses $\mathcal{H}_{P_\mathcal{A}}$ before the adversary can pick $p(A|X, Y)$. However, this will not be a problem because we consider scenarios where both the loss function and the hypothesis set are convex,[4], making the minimax and maximin formulations equivalent. Therefore, $p(A|X, Y)$ can be picked before h, making $\mathcal{H}_{P_\mathcal{A}}$ a well-defined set.

One issue with the formulation in Problem 3 is that it is undetermined in the sense that it admits several worst partition densities and classifiers for $|\mathcal{A}| > 2$. Fortunately, Lemma 3.1 shows that the minimum worst group error R^* in Problem 3 is the same as the one achieved if we were to consider an alternative formulation where a variable $A \in \{0, 1\}$ represents the worst-group risk membership. This makes the study of the binary problem attractive when we wish to minimize the number of assumptions we make on the protected groups. Here the objective becomes

$$R^* = \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{a\in\{0,1\} \\ p(A|X,Y) \\ s.t.\ p(A) \succeq \rho}} r_a(h), \tag{4}$$

with $h^*, p^*(A|X, Y)$ achieving this solution. We overload the notation $\mathcal{H}_{P_\mathcal{A}}$ in the context of Problem 4 to refer to the Pareto set for a binary group distribution. Figure 1 shows an example of the risks for the worst and best partitions at different sizes ρ achieved with a method that optimizes for

[3] $p(A) \succeq \rho$ if $p(A = a) \geq \rho\ \forall a \in \mathcal{A}$

[4] Meaning that for any $h, h' \in \mathcal{H}$ and $\lambda \in [0, 1]$, exists $h_\lambda \in \mathcal{H} : h_\lambda(x) = \lambda h(x) + (1 - \lambda)h'(x) \forall x \in \mathcal{X}$.

the worst case partition, like DRO or our proposed BPF (the latter shows better performance on the remaining partition owing to the Pareto constraint), versus deploying a baseline model that minimizes the empirical risk.

Lemma 3.1 shows that the minimum worst risk R^* is the same for problems 3 and 4, hence, we focus our analysis on the latter throughout the text. There are two main advantages of working with the binary problem, the first is that finding the worst partition $p(A \mid X, Y)$ for a given h is straightforward when $|\mathcal{A}| = 2$. The second is that, in general, we may not know the number of groups we wish to be fair to, and this equivalence shows that it is sufficient to specify the minimum size a group must have before it is considered for the purposes of minimax fairness. Moreover, restricting the minimum size of the partitions to be considered is an interpretable way of constraining the adversary.

Lemma 3.1. *Given an hypothesis class $\mathcal{H}$ and a finite alphabet, $\mathcal{A} : |\mathcal{A}| \geq 2$, problems 3 and 4 have the same minimum worst-group risk solution R^* for all ρ where problem 3 is defined ($\rho \leq \frac{1}{|\mathcal{A}|}$).*

A question that arises from Problem 4 is how the optimal classifier and partition function depend on the partition size. In Lemma 3.2, we show the existence of a critical size ρ^* for standard classification losses (cross-entropy and Brier score) whereby solving Problem 4 for partitions smaller than ρ^* leads to a uniformly random classifier. This result shows that attempting to be minimax fair w.r.t. arbitrarily small group sizes yields a trivial classifier with limited practical utility. Therefore, if one where to consider an adversary without any capacity restriction this would be it's optimal solution, something that we corroborate empirically.

Lemma 3.2. *Given Problem 4 with $p(Y|X) > 0 \; \forall X, Y$,[5] and let the classification loss be cross-entropy or Brier score. Let*

$$\bar{h}(X) : \bar{h}_i(X) = \frac{1}{|\mathcal{Y}|} \forall X, \forall i \in \{0, ..., |\mathcal{Y}| - 1\},$$

be the uniform classifier, and let $\bar{h} \in \mathcal{H}$. There exists a critical partition size

$$\rho^* = |\mathcal{Y}| \mathbb{E}_X[\min_y p(y \mid X)] \leq 1$$

such that solutions to Problem 4, $\forall \rho \leq \rho^*$, are $h^* = \bar{h}$ and

$$R^* = \bar{R} = \begin{cases} \log |\mathcal{Y}| & \text{if } \ell = \ell_{CE} \\ \frac{|\mathcal{Y}|-1}{|\mathcal{Y}|} & \text{if } \ell = \ell_{BS} \end{cases}.$$

That is, any partitions smaller than ρ^* yield the uniform classifier with constant risk $\bar{R}$.

In some circumstances, we may wish to address fairness in situations where there may be a preferred outcome. It is possible to capture this preference with other convex surrogate losses such as weighted crossentropy (wCE), defined as $-\sum_{i=1}^{|\mathcal{Y}|} w_i \mathbf{1}(y=i) \log h_i(x)$. In this scenario, it is also worthwhile to analyze the existence of critical classifiers and partition sizes. The proof presented in Supplementary Material Section A.1 shows a more general statement for a broader class of loss functions, including wCE, which may yield different critical classifiers and partition sizes, the restriction $p(Y|X) > 0$ is also lifted.

[5]This restriction can be lifted and a similar result holds, see Supplementary Material for details.

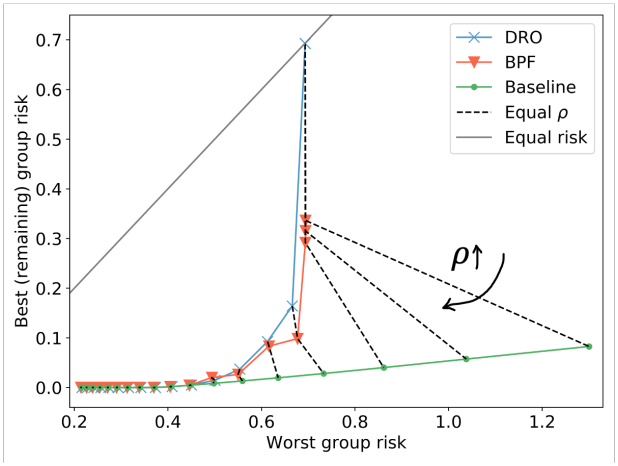

Figure 1. Worst and best crossentropy risks achieved by DRO, BPF and a baseline classifier for varying minimum size partitions (ρ) on a synthetic example (see Section A.2). The tradeoffs shown for DRO and BPF correspond to model pairs that were optimized for a specific ρ value (i.e., different points in the tradeoff curve correspond to different BPF and DRO classifiers). Lower worst group risks correspond to larger group sizes (ρ); results for the same ρ value are connected with a dashed line. We observe that BPF is able to achieve the same worst-case group performance that DRO achieves, but with better results on the non-critical partition owing to its Pareto optimality constraint, this is especially apparent on small group sizes. The baseline classifier suffers significantly larger worst group errors for small partition sizes.

It is straightforward to prove that R^* is non-increasing with ρ (see Supplementary Material A.1). A natural question that arises is what is the additional cost in optimality we pay if we apply subgroup robustness instead of optimizing for a known partition. Lemma 3.3 provides an upper bound for the cost of blind fairness, showing that it is at most the difference between R^* and the risk of the baseline model. Moreover, the upper bound decreases with larger group size, and is in no scenario larger than the difference between the risk of the uniform classifier and the baseline classifier for BS and CE losses.

Lemma 3.3. *Given a distribution $p(X, Y)$ and any predefined partition group $p(A'|X, Y)$ with $A' \in \mathcal{A}'$,$|\mathcal{A}'|$ finite. Let $\hat{h}, \hat{R} = \{\arg\} \min_{h \in \mathcal{H}} \max_{a' \in \mathcal{A}'} r_{a'}(h)$ be the minimax fair solution for this partition and its corresponding minimax risk.*

Let h^* and R^* be the classifier and risks that solve Problem 4 with $\rho = \min_{a' \in \mathcal{A}'} p(a')$. Then the price of minimax fairness can be upper bounded by

$$\max_{a' \in \mathcal{A}'} r_{a'}(h^*) - \hat{R} \leq R^* - \min_{h \in \mathcal{H}} r(h).^{6} \tag{5}$$

In the following section, we provide a practical algorithm that asymptotically,[7] solves Problem 4 and yields a classifier that both minimizes the worst-group risk and is also Pareto-efficient w.r.t. the remaining population. (All proofs are presented in the Supplementary Material A.1.)

4. Optimization

In order to develop our optimization approach, we begin by showing that for group sizes $\rho \leq \frac{1}{2}$, we can drop the innermost max operator in Problem 4, and we only need to consider the risk for the $a = 1$ partition on a partition of size exactly equal to ρ (i.e., $p(A = 1) = \rho$). This is presented in the following lemma.

Lemma 4.1. *Given Problem 4 with minimum group size $\rho \leq \frac{1}{2}$, the following problems are value equivalent:*

$$\begin{aligned} R^{\text{I}} &= \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \max_{p(A|X,Y)} \max_{a \in \{0,1\}} r_a(h), \\ &\qquad s.t. p(A) \succeq \rho \\ R^{\text{II}} &= \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \max_{p(A|X,Y)} r_1(h), \\ &\qquad s.t.\ p(A{=}1){=}\rho \\ R^{\text{I}} &= R^{\text{II}}. \end{aligned} \tag{6}$$

We note that the second problem in Eq. 6 is in itself an interesting optimization problem for $\rho > 1/2$, since it equates to efficiently minimizing the risk of the at-risk majority. Next, Lemma 4.2 shows that the Pareto optimality constraint is easy to enforce if the base loss function $\ell(h(x), y)$ is both bounded (i.e., $\ell(h(x), y) \leq C\ \forall x, y, h \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$), and strictly convex w.r.t. model h, and if the hypothesis class $\mathcal{H}$ is a convex set as well.

Lemma 4.2. *Given the problem on the right hand side of Eq. 6, a convex hypothesis class $\mathcal{H}$, and a bounded loss function $0 \leq \ell(h(x), y) \leq C\ \forall x, y, h \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$ that is strictly convex w.r.t its first input $h(x)$, the following problems are equivalent:*

$$\begin{aligned} \mathcal{H}^{\text{I}}, R^{\text{I}} &= \{\arg\} \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \max_{p(A|X,Y)} r_1(h) \\ &\qquad s.t.\ p(A{=}1) = \rho \\ \mathcal{H}^{\text{II}}, R^{\text{II}} &= \{\arg\} \min_{h \in \mathcal{H}} \sup_{p(A|X,Y)} r_1(h), \\ &\qquad s.t.\ p(A{=}1) = \rho \\ &\qquad p(A{=}1|X,Y) > 0,\ \forall X, Y \\ R^{\text{I}} &= R^{\text{II}},\ \mathcal{H}^{\text{I}} \supseteq \mathcal{H}^{\text{II}}, \end{aligned} \tag{7}$$

where we explicitly add $\{\arg\}$ to discuss the hypotheses achieving these minimax solutions. The set of hypotheses $\mathcal{H}^{\text{II}}$ are valid solutions to our problem, and correspond to the set of properly Pareto hypothesis (Geoffrion, 1968) which have the property of not admitting unbounded tradeoff between risks, see Definition 2.8.5 in (Miettinen, 2012). The BS loss satisfies both conditions in Lemma 4.2, CE loss also satisfies these conditions if the classifier assigns a minimum label probability for all values. The ℓ_2 regression loss over a bounded set also satisfies these conditions.

The supremum constraint on Lemma 4.2 is handled by slightly limiting adversary capacity and ensuring $p(A = 1|X, Y) \geq \epsilon > 0\, \forall X, Y \in \mathcal{X} \times \mathcal{Y}$, which leads to a minimax formulation. Furthermore, in the conditions of Lemma 4.2, we can exchange the minimum and the maximum without changing the problem, this leads to a feasible problem formulation based on importance weighting

$$\max_{\substack{p(A|X,Y) \\ s.t. p(A=1)=\rho \\ p(A=1|X,Y) \geq \epsilon}} \min_{h \in \mathcal{H}} \mathbb{E}_{X,Y}\left[\tfrac{p(A=1|X,Y)}{p(A=1)} \ell(h(X), Y)\right]. \tag{8}$$

Note that ϵ ensures that a minimal priority is assigned to each sample, even if belongs to the low-risk group. We solve Problem 8 using no-regret dynamics (Freund & Schapire, 1999), the solution is the Nash equilibrium of a two-player zero-sum game, where one player, the adversary, iteratively proposes partition distributions $p(A|X, Y)$, the modeler then responds near optimally with a model h, and incurs loss $r_1(h)$. Based on the history of losses, the adversary iteratively refines its proposed partition function into the worst-case partition.

To solve the above problem with parameter $\epsilon > 0$, we leverage the theoretical results presented in (Chen et al., 2017) for improper robust optimization of infinite loss sets with oracles. We present the results in terms of a finite dataset with n samples; assume that both players have access to $\{x_i, y_i\}_{i=1}^n \sim P(X, Y)^{\otimes n}$, and let $t \in \{0, \ldots, T\}$ indicate the current round of the zero-sum game. In each round t, the modeler produces a classifier h^t and the adversary proposes an empirical distribution of $p(A|X, Y)$, denoted as $\boldsymbol{\alpha}^t = \{\alpha_i^t\}_{i=1}^n \in \mathcal{U}_{\epsilon,\rho}$ such that $\mathcal{U}_{\epsilon,\rho} = \{\boldsymbol{\alpha} : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}$, where ρ is the minimum partition size. The empirical risk (cost) of round t is $L^t = L(h^t, \boldsymbol{\alpha}^t)$, with

$$L(h, \boldsymbol{\alpha}) \quad = \frac{\sum_{i=1}^n \alpha_i \ell(h(x_i), y_i)}{n\rho}. \tag{9}$$

Note that $L(h, \boldsymbol{\alpha})$ is the importance-weighted estimate of $r_1(h)$ over the samples in the dataset, that is $\mathbb{E}_{X,Y}[\frac{p(X,Y|A=1)}{P(X,Y)} \ell(h(X), Y)] = \mathbb{E}_{X,Y}[\frac{p(A=1|X,Y)}{p(A=1)} \ell(h(X), Y)]$.

[6] $r(h) = \mathbb{E}_{X,Y}[\ell(h(X), Y)]$.

[7] The algorithm is iterative, we prove convergence to the optimal solution with the number of iterations.

In order to find the Nash equilibrium of this game, we use projected gradient ascent on the adversary, while the modeler uses approximate best response with a γ-approximate Bayesian oracle $h^t = M(\alpha^t)$,[8]. In particular, we use a variant proposed in (Chen et al., 2017) for robust non-convex optimization. Algorithm 1 shows the proposed approach.

Algorithm 1 Blind Pareto Fairness

Require: Inputs: Dataset $\{(x_i, y_i)\}_{i=1}^n$, partition size ρ
Require: Hyper-parameters: rounds T, parameter η, adversary boundary coefficient $\epsilon > 0$, γ-approximate Bayesian solver $\mathrm{M}(\cdot) \simeq \arg\min_{h\in\mathcal{H}} L(h, \cdot)$
Initialize $\boldsymbol{\alpha}^0 = \hat{\boldsymbol{\alpha}} = \{\rho\}_{i=1}^n$
Initialize classifier and loss
$h^0 = \mathrm{M}(\hat{\boldsymbol{\alpha}}), L^0 = L(h^0, \hat{\boldsymbol{\alpha}})$
for round $t = 1, \ldots, T$ **do**
Adversary updates partition function by projected gradient ascent:
$\boldsymbol{\alpha}^t \leftarrow \boldsymbol{\alpha}^{t-1} + \eta \nabla_{\boldsymbol{\alpha}} L(h^t, \hat{\boldsymbol{\alpha}}) = \boldsymbol{\alpha}^{t-1} + \eta \frac{\ell(h^t, y)}{n\rho}$
$\hat{\boldsymbol{\alpha}} \leftarrow \prod_{\mathcal{U}_{\epsilon,\rho}} (\boldsymbol{\alpha}^t), \mathcal{U}_{\epsilon,\rho} = \{\boldsymbol{\alpha} : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}$
Solver approximately solves for the current partition:
$h^t \leftarrow \mathrm{M}(\hat{\boldsymbol{\alpha}})$
end for
Return: Classifier h^T

The proposed Algorithm 1 is an instantiation of Algorithm 3 in (Chen et al., 2017) for oracle efficient improper robust optimization with infinite loss sets. To implement the projection operator $\prod_{\mathcal{U}_{\epsilon,\rho}}(\cdot)$, we use a variant of the algorithm proposed in (Duchi et al., 2008) for projections onto the simplex that also contemplates the hypercube constraint $[\epsilon, 1]$. This projection update only requires access to the last risk evaluation on each sample during the training stage, making it a scalable and lightweight addition to the standard supervised training scenario. We can then immediately leverage the results in Theorem 7 in (Chen et al., 2017) to show that the algorithm converges.

Lemma 4.3. *Consider the setting of Algorithm 1, with parameter $\epsilon > 0$, and $\eta = \max_{\boldsymbol{\alpha}\in\mathcal{U}_{\epsilon,\rho}} \frac{||\boldsymbol{\alpha}||_2}{\sqrt{2T}} \leq \sqrt{\frac{n\rho}{2T}}$ with $\mathcal{U}_{\epsilon,\rho} = \{\boldsymbol{\alpha} : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}$, and L a 1-Lipschitz function w.r.t. $\boldsymbol{\alpha}$, let P be a uniform distribution over the set of models $\{h^1, \ldots, h^T\}$, and let R^* be the minimax solution to the loss presented in Eq. 9. Then we have*

$$\max_{\boldsymbol{\alpha}\in\mathcal{U}_{\epsilon,\rho}} \mathbb{E}_{h\sim P} L(h, \boldsymbol{\alpha}) \leq \gamma R^* + \sqrt{\frac{2n\rho}{T}}.$$

As in (Chen et al., 2017), we use h^T instead of the ensemble $\{h^1, \ldots, h^T\}$. We use stochastic gradient descent (SGD) as our γ-approximate Bayesian oracle, in practice, we alternate a single epoch of SGD with the adversary update for simplicity. We note that the 1-Lipschitz constraint can be relaxed to any G-Lipschitz function by working through the no regrets guarantees for projected gradient descent of G-Lipschitz functions in the proof provided in (Chen et al., 2017).

Figure 2 shows how the performance of the recovered classifiers (trained for a given partition size ρ) is optimal for its own partition size, but sub-optimal for other ρ values. These curves give a better picture on how risks are being traded off across samples.

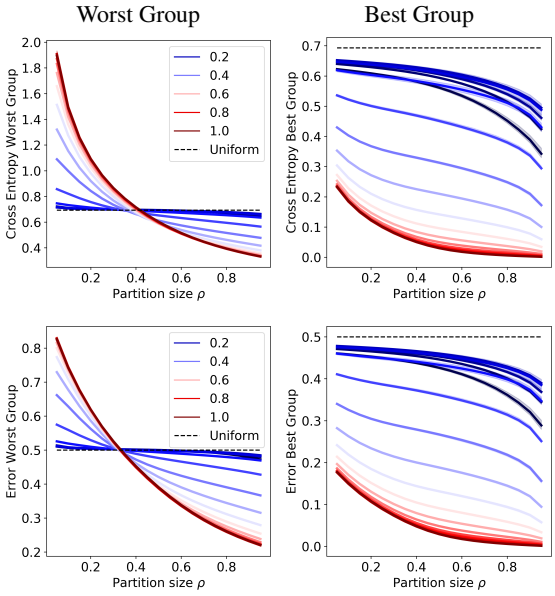

Figure 2. Crossentropy and error rates of BPF classifiers for varying (evaluation) partition sizes ρ evaluated on the UCI Adult dataset. Each individual curve denotes the performance of a unique BPF classifier, trained for a particular ρ, across a range of evaluation partition sizes on the test set. The uniform classifier is shown for reference. We observe how performance is traded off between low and high risk groups for varying partition sizes, in particular, smaller training partition sizes yield uniformly worse performance on the best group, no matter the evaluation partition size. Conversely, optimality on the worst group is dependent on how matched the train and test partition sizes are.

Generalization

For a fixed, known distribution $p(A=1|X, Y)$ it is straightforward to prove the following PAC bound for $r_1(h)$.

Lemma 4.4. *Given $p(A = 1|X, Y) \geq \epsilon, \forall X, Y$, $p(A = 1) = \rho$, and a bounded loss function $0 \leq \ell(h(X), Y) \leq C, \forall X, Y, h$ we denote the expected and empirical importance weighted risk as $r_1(h) = \mathbb{E}_{X,Y}[\frac{p(A=1|X,Y)}{p(A=1)}\ell(h(X), Y)]$ and $\hat{r}_1(h) =$*

[8] this produces an hypothesis with up to γ times more risk than the optimal solution for parameter α^t

$\sum_{i=1}^{n} \frac{p(A=1|x_i,y_i)}{np(A=1)} \ell(h(x_i), y_i)$ *respectively. Where* $h \in \mathcal{H}$ *and* $|\mathcal{H}|$ *is the dimension of the hypothesis set. Under these conditions, the following PAC bound holds*

$$P\left(\max_{h \in \mathcal{H}} |r_1(h) - \hat{r}_1(h)| \geq \frac{C}{\rho} \sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}} \right) \leq \delta$$

We observe that the upper bound on the generalization error degrades for small ρ values, since the ratio $\frac{p(A=1|X,Y)}{p(A=1)}$ is upper bound by $\frac{1}{\rho}$. Note that, for the special distribution presented in Lemma 3.2, there is no generalization error since the risk of the resulting classifier is, by construction, distribution invariant. Although this is the case for small ρ values in ideal scenarios, in practice we may still be able to over-fit to a small fraction of our samples in a way that still leads to larger-than-random errors in test samples.

5. Experimental Results

We experimentally validate our methods and theoretical results on a variety of standard datasets, we compare performance against DRO (Hashimoto et al., 2018), ARL (Lahoti et al., 2020), and a baseline classifier (empirical risk minimization). We show the trade-offs of each method on their worst group and the remaining population. As presented below (see Figure 3), the baseline method performs best on the low-risk population, but it suffers from large, fat tails in terms of loss distribution. We also show how both DRO and BPF empirically achieve the theoretical results laid in Lemma 3.2, with BPF having better results on the low-risk population than DRO, owing to the Pareto optimality constraint. Moreover, if the adversary's network on the ARL framework is given enough capacity, it degrades to the uniform (trivial) solution presented in Lemma 3.2 since it does not control for other restrictions on the partitions learned (e.g., group size). We also show that the performance of ARL can vary with adversarial network capacity (e.g., depth and width of the network). However, translating this to the effective size of the worst case partition being optimized is not easy to interpret or evaluate beforehand.

Datasets. We used four standard fairness datasets for comparison. The UCI Adult dataset (Dua & Graff, 2017) which contains $48,000$ records of individual's annual income as well as 13 other attributes, including race, gender, relationship status, and education level. The target task is income prediction (binary, indicating above or below $50K$). The Law School dataset (Wightman, 1998) contains law school admission data used to predict successful bar exam candidates; in our examples we limit ourselves to UGPA, LSAT scores and family income as input covariates. The COMPAS dataset (Barenstein, 2019) which contains the criminal history, serving time, and demographic information such as sex, age, and race of convicted criminals. The goal is prediction of recidivism per individual.[9] Lastly we used the MIMIC-III dataset, which consists of clinical records collected from adult ICU patients at the Beth Israel Deaconess Medical Center (Johnson et al., 2016). The objective is predicting patient mortality from clinical notes. We analyze clinical notes acquired during the first 48 hours of ICU admission following the pre-processing methodology in (Chen et al., 2018), ICU stays under 48 hours and discharge notes are excluded from the analysis. Tf-idf statistics on the $1,000$ most frequent words in clinical notes are taken as input features.

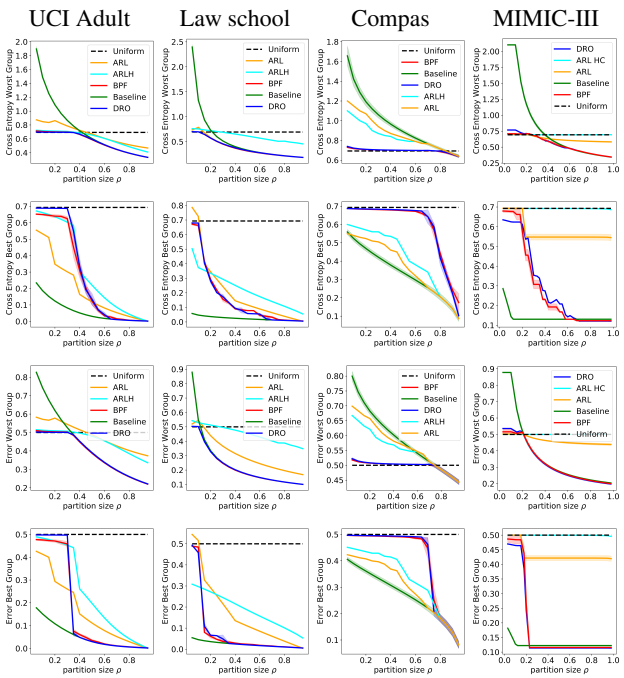

Figure 3. Cross-entropy (CE) and error rate (Error) metrics on best and worst groups as a function of group size for BPF, DRO, ARL, and baseline classifiers; results for very high capacity adversarial networks for ARL (ARL HC) are also shown, random classifier shown for reference. Results are provided for UCI adult, law school, COMPAS, and MIMIC-III datasets. Cross-entropy of both ARL and baseline classifiers for the worst group are very large for small group size, DRO and BPF both approximate the theoretical result shown in Lemma 3.2. The main experimental difference between DRO and the proposed BPF is that BPF exhibits better results on the best group partition than DRO for the same level of worst group performance, owing to the Pareto restriction on the BPF classifier resulting in no-unnecessary-harm for any group (see also Figure 1). Error results largely mimic the observations on the cross-entropy metric. Results are reported on held out (test) data.

[9] This dataset is the source of extensive and very legitimate controversy in the fairness community, and is here used for benchmarking only.

Setup and Results. We train BPF for 18 minimum group sizes $\rho = \{0.05, \dots, 0.9\}$ and $\epsilon = 0.01$, we report cross-entropy loss and error rate,[10] on the worst partition of the dataset (i.e., average over the worst $100 \times \rho\%$ samples based on cross-entropy loss), the values for the remaining low risk group is also reported to evaluate optimality. DRO models were trained on 18 equispaced values of their threshold parameter ($\eta \in [0, 1]$). For ARL, we tried four configurations for their adversarial network (adversary with 1 or 2 hidden layers with 256 or 512 units each), we additionally evaluated the same setup when the ARL adversary has access to the learned features of the classifier network, this setup still falls within the computationally identifiable scenario in their work, but offers a more challenging adversary, we denote these latter experiments as (ARL HC). The classifier architecture for BPF, ARL, and DRO was standardized to a single-layer MLP with 512 hidden units. In all cases we use cross-entropy loss and same input data. Results correspond to the best hyper-parameter for each group size; mean and standard deviations are computed using 5-fold cross-validation, all figures are reported on held out (test) data. Further implementation details are provided in Supplementary Material, Section A.4.

Figure 3 shows the performance of the best and worst groups across partition sizes. Both DRO and BPF recover results close to the random classifier for the smaller group sizes, which aligns with the results shown in Lemma 3.2, that is, below a certain partition size (e.g., $\rho \simeq 0.3$ for adult dataset) the average cross-entropy of the worst group is the risk of the uniform classifier ($\log 2$). We observe that the performance of ARL seems to be dataset dependent, but is generally able to reduce worst-case risks w.r.t. the baseline classifier for small partition sizes. The high capacity ARL (ARL HC) behaves as expected in most cases, producing results much more closely aligned with the uniform classifier, this supports our theoretical results stating that an unconstrained adversary should converge to the uniform classifier. Altogether, both BPF and DRO have the best performance; they are consistently able to recover the best worst group risk, and do well on preserving performance on the remaining samples. BPF also provides an explicit description of the worst group partition, and the option to control trade-offs between groups with the use of the ϵ parameter. In some situations, the difference between these two methods can become more pronounced, but this is dataset dependent. Additional details on this comparison are provided in Supplementary Material, Section A.2.

Although none of the compared models address disparities along predefined populations, we can nonetheless observe how each classifier performs on these groups. Table 1 shows accuracy conditioned on different demographics for each competing method on the Adult dataset. We observe that the different methods achieve results close to the uniform classifier for small partition sizes as expected. In many cases, the results for BPF are better than ARL and DRO for each protected attribute (at same ρ value). We also observe that on several minorities, the BPF model provides the best utility values out of all the competing methods, BPF is also the best model at preserving worst group performance.

Table 2 shows how target labels and predefined sensitive groups are represented in the high risk group identified by BPF, values reported on held out data. We observe that, for low partition sizes, outcomes are balanced across groups (in concordance with Lemma 3.2). As the partition size increases, the composition of the high risk group becomes more similar to the base distribution. Similar results to tables 1 and 2 are provided in Supplementary Material A.3 for the remaining datasets.

Method/ρ	White	Black	Asian-PacI	Other
Prop(%)	85.6%	9.6%	2.9%	1.8%
ARL .15	49.1%	51.9%	50.2%	52.8%
DRO .15	50.5%	50.0%	49.5%	50.1%
BPF .15	52.3%	52.5%	51.0%	53.5%
ARL .35	60.3%	65.9%	59.5%	63.8%
DRO .35	57.4%	59.4%	57.3%	59.5%
BPF .35	65.0%	69.7%	64.1%	68.5%
ARL .45	68.6%	77.5%	68.6%	74.3%
DRO .45	80.2%	88.0%	79.1%	86.9%
BPF .45	80.2%	88.0%	79.1%	86.9%

Table 1. Accuracy across demographic partitions (groups given no special consideration by the algorithms) in the Adult dataset for ARL, DRO and BPF models for varying partition sizes.

Group	Prop(%)	BPF .15	BPF .35	BPF .45
Proportion on Worst Partition, Ethnicity/Income				
White/0	64.2	43.8	44.8	53.5
White/1	21.4	45.3	44.9	35.6
Black/0	8.5	2.7	3.0	4.6
Black/1	1.1	2.9	2.5	1.8
Asian-PacI/0	2.1	1.8	1.7	1.9
Asian-PacI/1	0.8	2.0	1.8	1.4
Other/0	1.5	0.4	0.4	0.8
Other/1	0.3	1.0	0.8	0.5

Table 2. Demographic composition of worst groups as a function of minimum partition size on the Adult dataset. BPF homogenizes outcomes across partitions and protected attributes. For larger group sizes, the demographics of the partition approach that of the baseline population.

[10] Error rate is computed on the randomized classifier $\hat{Y} \sim h(X)$.

6. Discussion

In this work we formulate and analyze subgroup robustness, particularly in the context of fairness without demographics or labels. Our goal is to recover a model that minimizes the risk of the worst-case partition of the input data subject to a minimum size constraint, while we additionally constrain this model to be Pareto efficient w.r.t. the low-risk population as well. This means that we are optimizing for the worst unknown subgroup without causing unnecessary harm on the rest of the data. We show that it is possible to protect high risk groups without explicit knowledge of their number or structure, only the size of the smallest one, and that there is a minimum partition size under which the random classifier is the only minimax option for cross-entropy and Brier score losses.

We propose BPF, an algorithm that provably converges to a properly Pareto minimax solution, it requires minimal modifications to the standard learning pipeline of a standard model, and can scale easily to large datasets. Our results on a variety of standard fairness datasets show that this approach reduces worst-case risk as expected, and produces better models than competing methods for the low-risk population, thereby avoiding unnecessary harm. It identifies high risk samples and is easy to interpret since the user can control the optimal adversary through the use of a target worst partition size.

If a policymaker has a desired risk tradeoff instead of a target group size, we can search for the smallest partition size achieving this tradeoff using the proposed BPF; this now guarantees that the recovered model can satisfy this risk tradeoff for the worst possible partition up to size ρ, and for any smaller partition size there exists a partition such that this tradeoff is violated. Moreover, the tradeoffs between the worst and best group for a fixed group size ρ can be controlled with the weight's lower bound ϵ.

Future work includes incorporating additional domain-specific constraints on the worst partition and developing an algorithm that combines BPF with knowledge about some subgroups that must be protected as well.

References

Barenstein, M. Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*, 2019.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bertsimas, D., Farias, V. F., and Trichakis, N. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pp. 3539–3550, 2018.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, 2019.

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pp. 4705–4714, 2017.

Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Services Research*, 43 (5p1):1722–1736, 2008.

Freund, Y. and Schapire, R. E. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Geoffrion, A. M. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22(3):618–630, 1968.

Gupta, M., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Hajian, S., Bonchi, F., and Castillo, C. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2125–2126, 2016.

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

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2019.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

Kaplow, L. and Shavell, S. The conflict between notions of fairness and the Pareto principle. *American Law and Economics Review*, 1(1):63–77, 1999.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Martinez, N., Bertran, M., and Sapiro, G. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, 2020.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. *Microeconomic Theory*, volume 1. Oxford University Press New York, 1995.

Miettinen, K. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 2012.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Rawls, J. *Justice as Fairness: A Restatement*. Harvard University Press, 2001.

Rawls, J. *A Theory of Justice*. Harvard University Press, 2009.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., , and Re, C. *Addressing hidden stratification: Fine-grained robustness in coarse-grained classification problems*, 2020a. http://hazyresearch.stanford.edu/hidden-stratification,July2020.

Sohoni, N. S., Dunnmon, J. A., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *To Appear in Neural Information Processing Systems*, 2020b. http://stanford.edu/~nims/no_subclass_left_behind.pdf.

Ustun, B., Liu, Y., and Parkes, D. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382, 2019.

Veale, M. and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Wightman, L. F. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998.

Zhang, Y. Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197, 2018.