Fair for All: Best-effort Fairness Guarantees for Classification

Anilesh K. Krishnaswamy[1], Zhihao Jiang[2], Kangning Wang[1], Yu Cheng[3], and Kamesh Munagala[1]

[1]Duke University
[2]Tsinghua University
[3]University of Illinois at Chicago

December 21, 2020

Abstract

Standard approaches to group-based notions of fairness, such as *parity* and *equalized odds*, try to equalize absolute measures of performance across known groups (based on race, gender, etc.). Consequently, a group that is inherently harder to classify may hold back the performance on other groups; and no guarantees can be provided for unforeseen groups. Instead, we propose a fairness notion whose guarantee, on each group g in a class $\mathcal{G}$, is relative to the performance of the best classifier on g. We apply this notion to broad classes of groups, in particular, where (a) $\mathcal{G}$ consists of all possible groups (subsets) in the data, and (b) $\mathcal{G}$ is more streamlined.

For the first setting, which is akin to groups being completely unknown, we devise the PF (Proportional Fairness) classifier, which guarantees, on any possible group g, an accuracy that is proportional to that of the optimal classifier for g, scaled by the relative size of g in the data set. Due to including all possible groups, some of which could be too complex to be relevant, the worst-case theoretical guarantees here have to be proportionally weaker for smaller subsets.

For the second setting, we devise the BeFair (Best-effort Fair) framework which seeks an accuracy, on every $g \in \mathcal{G}$, which approximates that of the optimal classifier on g, independent of the size of g. Aiming for such a guarantee results in a non-convex problem, and we design novel techniques to get around this difficulty when $\mathcal{G}$ is the set of linear hypotheses. We test our algorithms on real-world data sets, and present interesting comparative insights on their performance.

1 Introduction

Machine learning is playing an ever-increasing role in making decisions that have a significant impact on our lives. Of late, we have seen the deployment of machine learning methods to provide advice for decisions pertaining to criminal justice (Angwin et al., 2016; Berk et al., 2018), credit/lending (Koren, 2016), health/medicine (Rajkomar et al., 2018), etc. Given the concerns of disparate impact and bias in this regard (Angwin et al., 2016; Barocas and Selbst, 2016), it is imperative that machine learning models are fair.

The question of defining notions of fairness, and developing methods to achieve them, has received a great deal of attention (Barocas et al., 2017; Binns, 2018). A common theme among the many approaches proposed thus far (Kleinberg, 2018; Chouldechova, 2017) is to fix beforehand a list of protected groups, and then ask for the (approximate) equality of some statistical measure across them. For example, *parity* seeks to equalize the accuracy across the given groups (Calders et al., 2009), while *equalized odds* seeks to equalize false positive or false negative rates (Hardt et al., 2016).

Classical definitions of fairness from microeconomics have also found application in machine learning (Balcan et al., 2019; Chen et al., 2019b; Hossain et al., 2020). In particular, there has been recent work (Zafar et al., 2017; Ustun et al., 2019) on adapting the notion of *envy-freeness*, which is born out of fair division theory (Brams and Taylor, 1996), to a group-based variant tailored to (binary) classification – every given pre-defined group should prefer the way it is classified (on aggregate) in comparison to how it would have been if it assumed the identity of some other group.

A major drawback of the aforementioned approaches is that they aim for an absolute guarantee: when some of the groups are inherently harder to classify than others, trying to achieve a particular measure of fairness, say equalized odds (Hardt et al., 2016), could do more harm than good by bringing down the accuracy on a group that is easier to classify (see Figure 1 for an example). In this paper, we take a more relative *best-effort* approach: aiming for guarantees that are defined in terms of how well each group can be classified in itself.

Another drawback of the standard approaches to fairness is that they depend critically on the specification of groups (via sensitive features such as race, gender, etc.). In many cases, the sensitive features are either missing (Chen et al., 2019a), or unusable, considering the need to adhere to *treatment parity* and anti-discrimination laws (Barocas and Selbst, 2016). Even if they can be used, it is sometimes not clear what the right categorization within them should be. For instance, it could be that a particular demographic group, which is defined on the basis of a shared cultural or ethnic feature, is actually a collection of hidden subgroups that are otherwise quite heterogeneous in terms of other socio-economic indicators (Meier and Melton, 2012; Chang, 2011). Therefore, mis-specifying or mis-calibrating the protected groups could end up hurting some groups within the data, potentially leading to unintended consequences such as a feeling of resentment among them (Hoggett et al., 2013).

We use the following instructive albeit stylized example to illustrate the

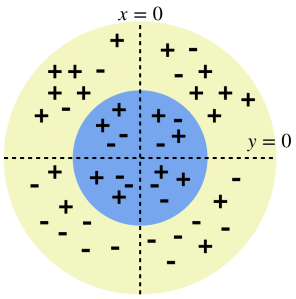

Figure 1: Given two groups Blue and Yellow (with true labels as shown), we have to choose just between the two classifiers $x = 0$ and $y = 0$. The Blue group is inherently harder to classify. Equalized odds makes us choose the classifier $x = 0$, thereby hurting the Yellow group. We could choose $y = 0$ with no aggregate effect on Blue, doing much better on Yellow.

a	1	1	0	0	1	1	0	0
b	1	0	1	0	1	0	1	0
c	1	1	1	1	0	0	0	0
y	1	1	0	0	1	0	1	0

Table 1: a, b are visible features, c is a hidden "demographic" feature, and y is the target label.

effect of missing group information.

Example 1. As shown below in Table 1, there are two binary features $a, b \in \{0, 1\}$, and a hidden demographic feature $c \in \{0, 1\}$. The target label y follows the formula $y = (a \wedge c) \vee (b \wedge \neg c)$: if the hidden feature $c = 1$, then a is a perfect classifier, and if $c = 0$, then b is a perfect classifier. For brevity, we define three groups: $P = \{(1, 0, 0), (0, 1, 0)\}$, $Q = \{(1, 0, 1), (0, 1, 1)\}$, and $R = (P \cup Q)^{\mathsf{c}}$.

As a concrete example, suppose each data point corresponds to a job candidate. The hidden feature c corresponds to *gender*, (a, b) correspond to measures of two different traits, and y the assignment to one of two jobs. Suppose the family of available classifiers is $\mathcal{H} = \{a, b\}$. It can be seen that any of these classifiers does poorly in terms of fairness for groups based on the hidden feature c.

Consider the classifier a (in this case, a solution to the standard Empirical Risk Minimization (ERM) with 0-1 loss), which correctly classifies all data except those in P. Therefore, the ERM classifier a is unfair to those of gender 0. Our PF classifier (which we shall see later) gets around this issue by randomizing between a and b. This allows us to classify P and Q correctly with probability 0.5, and R correctly with probability 1. Note that the PF classifier is able to treat every gender equally in expectation, even without access to the gender labels.

In order to deal with the above issues, we take a best-effort approach to fairness, one that can be applied to broad classes of groups. If the group identities were well defined and limited in number, simple solutions work: for example, one could perhaps train decoupled classifiers (Ustun et al., 2019; Dwork et al., 2018). With the unavailability or mis-specification of group information, however, the problem is much more interesting. In this regard, we look at two settings – where the groups taken into account are given by (a) all possible subsets in the data, and (b) a more streamlined class of groups, such as all linearly separable ones.

In the former setting, we are effectively reasoning about fairness even though there are no pre-specified groups. Standard statistical notions are of no use in this regard, for, as noted by Kearns et al. (2018), "we cannot insist on any notion of statistical fairness for every subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it mis-classified."

In fact, such a limitation also applies to any deterministic classifier. Therefore, focusing on randomized classifiers, the questions we consider first (Section 2) are: *What is the best possible best-effort guarantee that can be achieved for all groups simultaneously?* We will see that, on account of taking all groups into account, some of which could be too complex to be meaningful in practice, we have to settle for guarantees that are proportionally weaker for smaller subsets. *Are there algorithms that achieve such a guarantee?* We answer this question in the affirmative by devising the Proportional Fairness (PF) classifier.

The next natural question is: *can we do better if we consider a more streamlined class of groups?* We will see (Section 3) that this is indeed the case. We note here that standard fairness notions (such as *parity*) can also be applied in such settings, by effectively solving a convex optimization problem (Kearns et al., 2018). In our best-effort fairness (BeFair) approach, even when we consider linearly separable groups, we need to solve a non-convex problem. A major contribution of our work is devising a way of dealing with this difficulty, and at that, one that works well in practice. In Section 4, we *evaluate all our algorithms on real-world datasets.* We see that our BeFair approach is able to achieve strong best-effort guarantees, significantly better than standard ERM classifiers. We also present several empirical insights on the performance on PF, mostly in line with our theoretical results.

A more detailed overview of our results are provided in Section 1.2. For want of space, all our proofs are deferred to the Supplement.

1.1 Related literature

The extant literature on fairness in machine learning Hardt et al. (2016); Kamiran and Calders (2012); Hajian and Domingo-Ferrer (2012); Chouldechova (2017); Corbett-Davies et al. (2017) primarily considers statistical notions of fairness which require the protected groups to be specified as input to the (binary) classification problem. Many of these notions are further known to be incompatible with one another Kleinberg (2018); Friedler et al. (2016). Individ-

ual notions of fairness, which loosely translate to asking for "similar individuals" to be "treated similarly", have also been studied Dwork et al. (2012). However, this requires additional assumptions to be made about the problem at hand, in the form of, e.g., a "similarity metric" defined on pairs of data points. Another related notion is envy-freeness (Hossain et al., 2020), which isn't very useful without group information (more in the Supplement).

Several issues have been raised with respect to defining the demographic groups that need to be considered for fairness. Chen et al. (2019a) assess the prevalence of disparity when missing demographic identities are imputed from the data. Hashimoto et al. (2018) look at a model where user retention among different groups is linked to the accuracy achieved on them respectively, and design algorithms that improve the user retention among minority groups based on distributionally robust optimization. Their methods, while oblivious to the identity of the groups, operate under the assumption that there are a fixed number K of groups, and work well in practice for small K. Kearns et al. (2018) study the problem of auditing classifiers for statistical parity (or other related fairness concepts) across a (possibly infinite) collection of groups of bounded VC dimension. However, they do not consider the fact that some groups could be inherently harder to classify than others, and instead work with standard statistical notions such as statistical parity. Doing so results in a non-convex problem – something that we deal with in our work.

Our BeFair approach assumes black-box access to an agnostic learning oracle. Such reductions are commonplace in recent work on fairness in machine learning (Kearns et al., 2018). For example, Agarwal et al. (2018) reduce fair classification to a sequence of cost-sensitive classifications, the solutions of which can be achieved using out-of-the-box classification methods.

The study of fairness has had a much longer history in economics, in particular, the literature on fair division and cake-cutting (Brams and Taylor, 1996; Robertson and Webb, 1998). We include a fuller discussion of this literature in the Supplement.

1.2 Our model and results

We are given a set of n data points denoted by $\mathcal{N}$, with their features given by $\{x_i\}_{i\in\mathcal{N}}$, and their true binary labels by $\{y_i\}_{i\in\mathcal{N}}$. The hypothesis space at hand will be denoted by $\mathcal{H}$, a set of (deterministic) classifiers. The Boolean variable $u_i(h) \in \{0,1\}$ denotes whether the classifier $h \in \mathcal{H}$ correctly classifies data point i. In other words, $u_i(h) = \mathbb{1}[h(x_i) = y_i]$. A classification instance is defined by a pair $(\mathcal{N}, \mathcal{H})$. We assume that for any classifier $h \in \mathcal{H}$, its *complement* $\bar{h}$, defined by flipping the classification outcomes of h (i.e., $h(x_i) = 1 - \bar{h}(x_i)$), is also in $\mathcal{H}$. This assumption is valid for most natural families of binary classifiers. We denote by $\Delta(\mathcal{H})$ the space of all randomized classifiers over $\mathcal{H}$. If $h \in \Delta(\mathcal{H})$ is obtained via a distribution D_h over $\mathcal{H}$, then for a data point $i \in \mathcal{N}$, we defined the utility $u_i(h) \triangleq \mathbb{E}_{h'\sim D_h}[u_i(h')]$.

We are given $\mathcal{G}$, a class of groups, each element of which is of the form $g : \mathcal{N} \to \{0,1\}$. We also use g to denote the subset given by $\{i \in \mathcal{N} : g(i) = 1\}$

and $|g|$ as its size $|\{i \in \mathcal{N} : g(i) = 1\}|$. For any such g, its utility under h is $u_g(h) = \frac{1}{|g|} \sum_{i \in g} u_i(h)$.

For each $g \in \mathcal{G}$, define $h_g^* \triangleq \arg\max_{h \in \mathcal{H}} u_g(h)$ to be the best classifier for the group g. The best-effort fairness guarantee is captured via a constraint of the form $f(u_g(h), u_g(h'), |g|) \geq 0$. The function $f(\cdot)$ constrains the accuracy $u_g(h)$ of h, the classifier at hand, to that of the optimal classifier h_g^* for g, with a possible dependence on the size $|g|$ of the group g. Applying such a constraint for all $g \in \mathcal{G}$ gives us a uniform best-effort fairness guarantee. What sort of $f(\cdot)$ is workable depends on the class $\mathcal{G}$ considered.

In Section 2, we consider the case where $\mathcal{G}$ includes all the subsets of $\mathcal{N}$, i.e., there is no specific information about $\mathcal{G}$. Via a theoretical worst-case bound (Theorem 1), we show that the best we can do in this case is to choose $f(\cdot) = u_g(h) - \frac{|g|}{|\mathcal{N}|}[u_g(h')]^2$. For any group g that can be *perfectly classified* by some $h' \in \mathcal{H}$ ($u_g(h') = 1$), the same constraint boils down to $u_g(h) \geq |g|/|\mathcal{N}|$: in other words, a utility of at least $|g|/|\mathcal{N}|$ should be guaranteed on such a set. Such a guarantee can be interpreted as fairness: If g is a potentially hidden demographic that can be perfectly classified using some features, our classifier should not ignore those features entirely. We show that our PF classifier in fact achieves this guarantee (Theorem 2).

In Section 3, we consider a more streamlined class of groups: in particular, $\mathcal{G}$ contains all linearly separable groups. For ease of exposition, we define the error $\text{err}_g(h) = \sum_{i \in g}[1 - u_i(h)]$, and recast the discussion in terms of it.* In this case, we seek a much stronger guarantee: $\text{err}_g(h) \leq \text{err}_g(h_g^*) + \gamma$. The general form of the optimization problem we solve is as follows:

$$\begin{aligned} \min_{h \in \Delta(\mathcal{H})} \quad & \text{err}_{\mathcal{N}}(h) \\ \text{such that } \forall g \in \mathcal{G}, \quad & \text{err}_g(h_g^*) - \text{err}_g(h) + \gamma \geq 0. \end{aligned}$$

As discussed in more detail later, to solve the above problem we need to deal with the non-convex constraints. We outline a method (BeFair) to do so when $\mathcal{G}$ consists of linearly separable groups. We will also look for a slightly weaker guarantee as follows: $\text{err}_g(h) \leq \delta \cdot \text{err}_g(h_g^*) + \gamma$, for some $\delta \geq 1$ – weaker because now $\text{err}_g(h)$ has a slightly larger target $\delta u_g(h_g^*)$ to approximate. Our techniques extend seamlessly to such a formulation also.

2 Best-effort guarantee for all groups

The first question to ask is whether there is a fundamental limit on how well one can hope to do with respect to fairness in the setting where $\mathcal{G} = 2^{\mathcal{N}}$. Since we are dealing with a notion of fairness that is measured relative to the family of classifiers at hand, we first want to understand what the best guarantee that can be given is (in the form of a worst-case bound), with no conditions on the type of classifiers used.

*While a fundamentally similar discussion can be done in terms of the utilities, using errors instead leads to an easier handling of the constants involved.

Theorem 1. *On any data set $\mathcal{N}$, there is no randomized classifier h (for some $\mathcal{H}$) such that for all $g \subseteq \mathcal{N}$ admitting a perfect classifier $h_g^* \in \mathcal{H}$ (i.e., $u_g(h_g^*) = 1$), we have $u_g(h) > \frac{|g|}{|\mathcal{N}|}$.*

This theorem shows that, in terms of how much utility is accrued by each of the perfectly classified sets, the best bound we can hope to target is one proportional to the fractional size of the given set of data points. Note that for every instance, there exists some $\mathcal{H}$, such that the claim of the theorem holds – this is not true more generally in the sense that there could exist some $\mathcal{H}$ for which the claim does not hold as shown by Example 2 (in the Supplement).

2.1 Proportional Fairness (PF) Classifier

We now demonstrate a classifier that matches the above bound as long as the utilities $u_i(h_j)$'s are binary (which holds in our model, but could also be encountered in other scenarios involving resource allocation discussed in Section 1.1); as mentioned before, this captures multi-class classification as well. The Proportional Fairness classifier is defined as follows:

Definition 1 (Proportional Fairness (PF))**.** Given an instance $(\mathcal{N}, \mathcal{H})$, the PF classifier h_{PF} is the one that maximizes $f(h) \triangleq \sum_{i \in \mathcal{N}} \ln u_i(h)$ over all $h \in \Delta(\mathcal{H})$.

As mentioned before, the proportional fairness objective has had a long history in network resource allocation literature (Kelly et al., 1998). However, to the best of our knowledge, its applicability to the classification problem, and the implications thereof, have never been established before.

We now show that the PF classifier achieves a guarantee matching the worst-case bound in Theorem 1.

Theorem 2. *For any subset $g \subseteq \mathcal{N}$ that admits a perfect classifier $h_g^* \in \mathcal{H}$ (i.e., $u_g(h_g^*) = 1$) we have $u_g(h_{\text{PF}}) \geq \frac{|g|}{|\mathcal{N}|}$.*

Thus, the PF classifier achieves, on any subset, an accuracy that is *proportional* to the accuracy of the best classifier on that subset scaled by the fractional size of the subset. As mentioned earlier, the use of perfectly classifiable subsets in our analysis is just for the ease of exposition. The results can be suitably translated to using all possible subsets. For example, the following is a simple corollary of Theorem 2:

Corollary 2.1. *For any subset $g \subseteq \mathcal{N}$, with its best classifier $h_g^* = \arg\max_{h \in \mathcal{H}} u_g(h)$, we have $u_g(h_{\text{PF}}) \geq \alpha \left[u_g(h_g^*)\right]^2$, where $\alpha = \frac{|g|}{n}$.*

Assuming black-box access to an agnostic learning oracle, the PF classifier can be computed using a primal dual style algorithm (details in the Supplement, or see Bhalgat et al. (2013) for similar results). In our experiments, we just use a heuristic instead (see Section 4, and also the Supplement). We also do not explicitly discuss the generalization properties – but we would expect that PF is not prone to overfitting, since all possible groups have to be given a guarantee on performance (details in the Supplement).

Interpreting the results: Theorem 2 and Corollary 2.1 neatly characterize how PF achieves the best possible theoretical bound . One drawback of applying PF in practice is that the theoretical guarantee is proportionally lower for smaller groups, notwithstanding the fact that, in practice, the accuracy of PF on small groups is much better than what is given by these bounds (see Section 4). As far as the bounds as concerned, the reason that we have to settle for an accuracy proportionally lower for smaller subsets is that the guarantee has to hold for all possible subsets. Some of these subsets could be extremely complex, and possibly unreasonable in most practical settings. As will see next, we can do much better with more restricted classes of groups.

3 Best-effort guarantees for linearly separable groups

In this section, we limit $\mathcal{G}$ to be a more streamlined class of groups, and aim for a much stronger guarantee. We then devise an algorithm that achieves such a guarantee. As mentioned in Section 1.2, we want to find a a randomized classifier $h \in \Delta(\mathcal{H})$ that, for every group $g \in \mathcal{G}$, achieves an absolute error which is within an additive factor γ from that of the optimal classifier h^*_g for g. In particular, the optimization problem we would like to solve is the following:

Problem 1 (BeFair(γ))**.** For a given hypothesis space $\mathcal{H}$, a class of groups $\mathcal{G}$, and $\gamma \geq 0$,

$$\begin{aligned} \min_{h \in \Delta(\mathcal{H})} \quad & \mathrm{err}_{\mathcal{N}}(h) \\ \text{such that } \forall g \in \mathcal{G}, \quad & \mathrm{err}_g(h^*_g) - \mathrm{err}_g(h) + \gamma \geq 0. \end{aligned}$$

In particular, we consider $\mathcal{H}$ to be the space of linear hypotheses, and $\mathcal{G}$ the class of all linearly separable groups.[†] As mentioned earlier, despite using linear hypotheses and groups, we are faced with a non-convex problem. It can be seen that the non-convexity stems from the best-effort constraint – while the terms $\mathrm{err}_g(h)$ and $\mathrm{err}_g(h^*_g)$ can be individually made convex by using standard surrogate loss functions, their combination obtained by subtracting one from the other cannot. Note that such a difficulty does not arise for the more absolute notions of fairness such as *parity*, as is the case with the techniques in Kearns et al. (2018). Also, even in our setting, if $\mathcal{G}$ were a small finite set, then all the optimal classifiers h^*_g could be calculated offline, and the corresponding constraints listed to form a simpler convex optimization problem.

We redefine the BeFair(γ) as follows to explicitly factor the hidden optimization problem of finding $h^*(g)$ into the corresponding constraint for g; by using the fact that if $\mathrm{err}_g(h^*_g) - \mathrm{err}_g(h) + \gamma \geq 0$, then $\mathrm{err}_g(h') - \mathrm{err}_g(h) + \gamma \geq 0$ for any $h' \in \mathcal{H}$.

[†]If $g \in \mathcal{G}$, then g and $\mathcal{N} \setminus g$ are linearly separable.

Problem 2 (BeFair(γ))**.**

$$\min_{h \in \Delta(\mathcal{H})} \quad \mathrm{err}_{\mathcal{N}}(h)$$
$$\text{such that } \forall g \in \mathcal{G}, h' \in \mathcal{H}, \quad \mathrm{err}_g(h') - \mathrm{err}_g(h) + \gamma \geq 0.$$

We first define the partial Lagrangian corresponding to Problem 2. Let $\phi(g, h, h') \triangleq -\mathrm{err}_g(h') + \mathrm{err}_g(h) - \gamma$. With dual variables $\lambda_{g,h'}$ for every $g \in \mathcal{G}$ and $h' \in \mathcal{H}$:

$$L(h, \lambda) \triangleq \mathrm{err}(h) + \sum_{g \in \mathcal{G}, h' \in \mathcal{H}} \lambda_{g,h'} \phi(g, h, h').$$

In order to have a convergent algorithm for our optimization, we will restrict the dual space to the bounded set $\Lambda = \{\lambda \in \mathcal{R}_+^{|\mathcal{G} \times \mathcal{H}|} : \|\lambda\|_1 \leq C\}$, where C will be a parameter in our algorithm. Then, by the Minimax Theorem, solving Problem 2 is equivalent to solving the following:

$$\min_{h \in \Delta(\mathcal{H})} \max_{\lambda \in \Lambda} L(h, \lambda) = \max_{\lambda \in \Lambda} \min_{h \in \Delta(\mathcal{H})} L(h, \lambda). \tag{1}$$

The minmax problem can be viewed as a two player zero-sum game: The set of pure strategies for the *learner* (corresponding to the primal) corresponds to $\mathcal{H}$ – each deterministic classifier $h \in \mathcal{H}$ is a valid pure strategy. For the *adversary* (corresponding to the dual), the pure strategies in Λ can be either the all zeros vectors, or a particular choice of $(g, h') \in \mathcal{G} \times \mathcal{H}$. Then, solving Problem 2, via the minmax formulation in Equation 1, is the same as finding an equilibrium of the corresponding two-player zero-sum game with $L(h, \lambda)$ as the payoff for the dual player.

3.1 Solving the BeFair(γ) problem via a convex relaxation:

The equilibrium of a two-player zero-sum game can be found using Fictitious Play, an iterative algorithm which is guaranteed to converge[‡] given that we can solve for the best responses of both players (Robinson, 1951). Fictitious Play (Brown, 1949) proceeds in rounds alternating between the primal and dual player: in each round, each player chooses a best response to the the mixed strategy that randomizes uniformly over the empirical history of the other's strategies. A formal description is given in Algorithm 1.

Learner's best response: For a given mixed strategy λ of the adversary, the learner needs to solve:

$$\min_{h \in \Delta(\mathcal{H})} \mathrm{err}(h) + \sum_{g \in \mathcal{G}, h' \in \mathcal{H}} \lambda_{g,h'} \mathrm{err}_g(h).$$

[‡]The asymptotic convergence is usually fast in practice, and especially so in our experiments.

Since the optimum is obtained at the corner points of the feasible region of the strategy space, we need only consider pure strategies for the optimization problem above. The learner's problem then becomes:

$$\min_{h \in \mathcal{H}} \sum_i w_i \mathbb{1}[h(x_i) \neq y_i],$$

where $w_i \triangleq 1 + \sum_{g \in \mathcal{G}, h' \in \mathcal{H}} \lambda_{g,h'} \mathbb{1}[g(x_i) = 1]$, and this can be solved since we assume black-box access to a weighted ERM oracle. In practice, many heuristics (like Logistic Regression, Boosting, etc.) are used effectively for this problem, even though it is known to be hard in the worst case (Feldman et al., 2012).

Adversary's best response: The adversary's best response problem is more involved and will require some novel techniques to solve. Again, we need to optimize only over pure strategies. With a bit of analysis, the dual best response problem can be seen to be equivalent to solving, for a given $h \in \Delta(\mathcal{H})$:

$$\min_{g \in \mathcal{G}, h' \in \mathcal{H}} \mathrm{err}_g(h') - \mathrm{err}_g(h). \tag{2}$$

For all $i \in \mathcal{N}$, define $t_i \triangleq \mathbb{E}\mathbb{1}[h(x_i) \neq y_i]$. Then the above objective can be written as

$$\begin{aligned} &\sum_i \mathbb{1}[g(x_i) = 1] \left(\mathbb{1}[h'(x_i) \neq y_i] - \mathbb{E}\mathbb{1}[h(x_i) \neq y_i] \right) \\ =&\sum_i \mathbb{1}[g(x_i) \neq -1] \left(\mathbb{1}[h'(x_i) \neq y_i] - t_i \right), \end{aligned} \tag{3}$$

which is non-convex. For each i, we would like to convexify it differently when $t_i = 0$ or $t_i > 0$.

Since we only consider the case where both $\mathcal{G}$ and $\mathcal{H}$ consist of linear hypotheses[§], define $z_g \triangleq x^\intercal \theta_g$ and $z_{h'} \triangleq -y x^\intercal \theta_{h'}$, where θ_g and $\theta_{h'}$ are the coefficients of the linear hypotheses g and h' respectively. We consider the cases $t_i = 0$ and $t_i > 0$ separately.

If $t = 0$: Each such $i \in \mathcal{N}$ adds the term $\mathbb{1}[z_g > 0] \cdot \mathbb{1}[z_{h'} > 0]$ to the objective. Replace each indicator function with an exponential to get $e^{z_g} \cdot e^{z_{h'}} = e^{z_g + z_{h'}}$.

If $t > 0$: Each such $i \in \mathcal{N}$ add the term $\mathbb{1}[z_g > 0] \cdot \mathbb{1}[z_{h'} > 0] - \mathbb{1}[z_g > 0]$ to the objective. The first term is treated as before, and the second term, i.e., $\mathbb{1}[z_g > 0]$ is replaced by $(1 - e^{-z_g})$, whence the whole objective becomes $e^{z_g + z_{h'}} + e^{-z_g} - 1$, which is convex.

Therefore, we need to solve

$$\min_{\theta_g, \theta_{h'}} \sum_{i \in \mathcal{N}} e^{z_g + z_{h'}} + t_i (e^{-z_g} - 1), \tag{4}$$

which can be done via convex optimization methods.

[§]More general classes of $\mathcal{G}$ and $\mathcal{H}$ are an interesting open problem.

Algorithm 1 Solving BeFair(γ)

Input: data set $\mathcal{N}$, $\gamma \geq 0$, number of rounds T.
Initialize by setting h_0 to be some classifier in $\mathcal{H}$, and λ_0 to be the zero vector.

for $t = 1, \ldots T$: **do**
  $\bar{h} \leftarrow$ uniform distribution over $\{h_0, \ldots, h_{t-1}\}$
  $\bar{\lambda} \leftarrow \frac{1}{t} \sum_{t' < t} \lambda_{t'}$
  $h_t \leftarrow$ Learner's best response to $\bar{\lambda}$
  $\lambda_t \leftarrow$ Adversary's best response to $\bar{h}$
end for
Return: $\bar{\lambda}_t$

The solution returned by Algorithm 1 has to be checked for feasibility with respect to Problem 2 – this tells us if the problem is feasible to begin with. Also, the solution to the Adversary's problem (Equation 2) can potentially be improved by alternately optimizing for g and h', à la Expectation-Maximization: For a fixed h', we can optimize g by using Equation 3 and a convexification analogous to Equation 4. The converse problem of optimizing h', for a fixed g, is just a weighted ERM problem.

3.2 A more general version of BeFair(γ):

For $\delta \geq 1$, we can generalize Problem 2 as follows:

Problem 3 (δ-BeFair(γ))**.**

$$\min_{h \in \Delta(\mathcal{H})} \quad \mathrm{err}_{\mathcal{N}}(h)$$
$$\text{such that } \forall g \in \mathcal{G}, \quad \delta \cdot \mathrm{err}_g(h^*_g) - \mathrm{err}_g(h) + \gamma \geq 0.$$

The only difference from Problem 2 is that we have slightly weaker constraints: the error of h on g is compared with δ times the least possible error on g. With a straightforward modification, the overall technique in Section 3.1 works for this problem too.

For a fixed δ, computing the quantity $\max_{g \in \mathcal{G}} \mathrm{err}_g(h) - \delta \cdot \mathrm{err}_g(h^*_g)$ gives us a way of measuring the fidelity of any given classifier h. To do so, the Adversary's problem can be solved (as shown) to find the smallest γ for which h becomes feasible for the constraints in Problem 3 (i.e., satisfies the best-effort guarantees). We discuss this in more detail in Section 4.1.

4 Experiments

The primary goal of our experiments is to show that the BeFair algorithm works extremely well in practice. As we discuss below, BeFair achieves its intended purpose (as discussed in the previous section), by achieving strong

	LR	Ada	hPF	1.0-BeFair	1.1-BeFair
adult	0.83	0.84	0.78	0.79	0.80
compas	0.75	0.75	0.64	0.70	0.71

Table 2: Overall test accuracy of various methods.

best-effort fairness guarantees uniformly over all linearly separable groups. In particular, it is able to achieve a performance that is a close approximation of the best possible on these groups (as given in Problem 3) for small values of δ and γ. In addition, we will also evaluate the PF algorithm and show how it behaves differently from BeFair, on account on having to provide guarantees for all possible groups, even those corresponding to a high VC dimension. We also show how, in practice, the performance of PF seems to be better than what is suggested by the worst-case lower bound via Theorem 2 (which is proportionally weaker for smaller groups).

We work with two data sets: `adult`, the Adult[¶] dataset from the UCI Machine Learning Repository, and `compas`, the COMPAS[‖] Risk of Recidivism data set (Angwin et al., 2016). Both have binary labels and a mixture of numerical and categorical features. Using these data sets, we compare and contrast the following methods (recalling their definitions from earlier):

1. PF: As the exact solution of PF is computationally inefficient, we use hPF, a heuristic (details in the Supplement) inspired by Reweighted Approval Voting. (Aziz et al., 2017). In what follows, we refer to hPF as PF.
2. δ-BeFair as described in Section 3.
3. ERM methods: We use LR (Logisitic Regression), since it performs best here. We also compare overall accuracy with Ada (AdaBoost), an ensemble method.
4. The lower bound given by Corollary 2.1.

In Table 2, we present the overall accuracy of various methods, i.e., that measured on the entire test set. As there is a trade-off between ensuring fairness for groups and maximizing overall accuracy, PF has a lower overall accuracy compared to other methods. δ-BeFair is much closer to the ERM baselines (especially as seen on the `compas` dataset, even for a small value of $\delta = 1.1$. Larger values of δ can only increase accuracy as the fairness constraints become laxer.

4.1 Evaluating the performance of BeFair

We first define Maximum Additive Error (MAE_δ), parametrized by δ, of any given classifier h.

[¶] 48842 instances, 14 features, https://archive.ics.uci.edu/ml/datasets/Adult

[‖] 6172 instances, 8 features, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Definition 2 ($\text{MAE}_\delta(h)$)**.** For a given h, and δ, $\text{MAE}_\delta(h) = \max_{g \in \mathcal{G}} \text{err}_g(h) - \delta \cdot \text{err}_g(h^*_g)$.

For a given h, $\text{MAE}_\delta(h)$ specifies, for the worst-off group g, how much difference there is between the error of h and that of the best classifier for g scaled by δ.

For instance, $\text{MAE}_\delta(\text{BeFair})$ can be computed by searching over different values of γ to pick the smallest that gives a feasible solution for the δ-BeFair(γ) problem. On the other hand, $\text{MAE}_\delta(\text{ERM})$ can be computed by solving the Adversary's problem (Equation 3 modified as per δ) for $h = \text{ERM}$.

In Figure 2, we compare MAE_δ of ERM and BeFair for $\delta = 1.0, 1.05, \ldots, 1.30$. Errors are reported as a percentage of the entire data set.

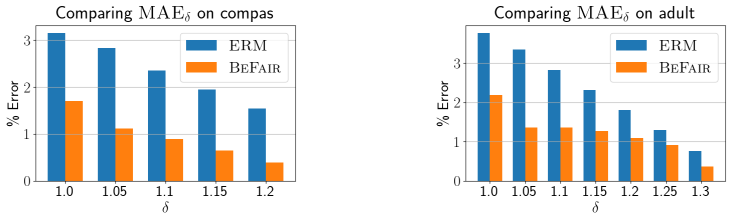

Figure 2: Comparing MAE_δ between ERM and BeFair for varying values of δ.

Many key observations can be made from this plot:

(a) As we increase δ, the MAE_δ of both ERM and BeFair decrease. This is because the best-effort constraints get laxer with increasing δ.

(b) Even for $\delta = 1.0$, BeFair achieves an improvement over ERM of close to 50% (since $\text{MAE}_\delta(\text{BeFair})$ is about half of $\text{MAE}_\delta(\text{ERM})$) on `compas`, and 33% on `adult`, in the MAE_δ value.

(c) For a slightly larger value of $\delta = 1.10$, we get an extremely low value for $\text{MAE}_\delta(\text{BeFair})$ of around 1%, which means BeFair gets a strong approximation. Therefore, BeFair is able to achieve an multiplicative error of 0.1, with an additive error of around 1%.

(d) $\text{MAE}_\delta(\text{ERM})$ decreases linearly with δ, while most of the improvement in $\text{MAE}_\delta(\text{BeFair})$ comes from increasing δ from 1 to 1.05. In other words, BeFair is able to extract a bigger improvement with a small increase of δ.

Overall, BeFair achieves low MAE_δ for small values of $\delta = 1.05, 1.1$.

4.2 Comparison of PF with BeFair:

In Figure 3, we order (on the x axis) the data points in the test set in ascending order of their scores (i.e., confidence of predicting the true label) given

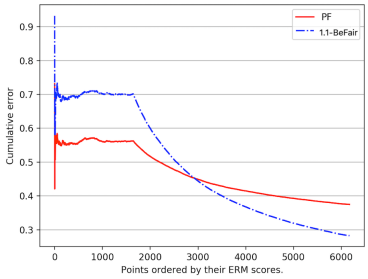

Figure 3: Error on subsets of varying sizes (`compas`): x axis denotes points in ascending order of LR scores. y axis denotes error accrued on the subset containing points up to x.

by LR. For each point x, the y axis shows the accuracy of various methods on the subset consisting of all points from 0 through x. If we imagine the LR scores as a measure of how easy the points are to classify correctly: then we see that PF gets a more uniform error on all these sets, whereas 1.1-BeFair has lower error on sets with higher LR scores. This is because the groups where BeFair has high error are probably too complex to be meaningful practically. On the other hand, PF must provide uniform guarantees over all groups, even those corresponding to large VC dimensions, and therefore has a uniformly higher error on them.

4.3 Comparison of PF with the theoretical lower bound

In Figure 4 (left), we order (on the x axis) the data points in the test set in ascending order of their scores (i.e., confidence of predicting the true label) given by Ada. For each point x, the y axis shows the accuracy of various methods on the subset consisting of all points from 0 through x. Figure 4 (right) does the same with PF scores.

We see that the accuracy of PF is much higher than the worst-case lower bound. PF comes close to the lower bound for larger subsets, especially for those that are easy to classify (see the Supplement for more details). Note that the lower bound *is not monotonic* because it depends on both the size of the subset and the best possible classification accuracy on it (Corollary 2.1). Also, in Figure 4 (right), ERM methods do worse because the points with low hPF scores are inherently much harder to classify.

5 Conclusions

In this paper, we study group fairness in the (multi-class) classification setting. We propose a notion based on best-effort guarantees, which requires each group in a class $\mathcal{G}$ to have a classification accuracy that is as close as possible

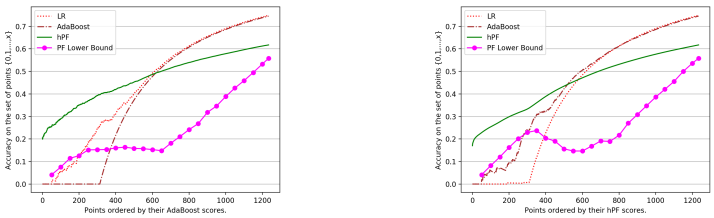

Figure 4: Accuracy on subsets of varying size (`compas`): x axis denotes points in ascending order of ADA (left) and HPF (right) scores. y axis denotes accuracy on the subset containing points up to x.

to the optimal for that group. When $\mathcal{G}$ consists of all possible groups, we show that PF achieves the theoretical optimum in our setting. When $\mathcal{G}$ consists of linearly separable groups, we can do much better via the BEFAIR algorithm, which crucially depends on convexification techniques to solve an essentially non-convex problem. We also test our methods on real-world datasets and show that they perform well in practice, especially the BEFAIR method.

One interesting question for future work is to extend our techniques for more involved classes of groups, say, for example, when $\mathcal{G}$ consists of all groups that can be identified by a fixed neural network. Similar extensions of the hypothesis space $\mathcal{H}$ are also worth looking at. Moreover, in some applications (bail/loan decisions, college admissions, etc.), false negatives and false positives play drastically different roles. Can our framework be extended to deal with such considerations? Can it also be extended to multi-class classification? Note that the guarantees of PF carry over to this setting directly. We would also like to point out that randomized classifiers are not always desirable and have some limitations in practice (Cotter et al., 2019). How to think about best-effort fairness of deterministic classifiers with unknown groups is another interesting open question.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23:2016.

Aziz, H., Brill, M., Conitzer, V., Elkind, E., Freeman, R., and Walsh, T. (2017). Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485.

Balcan, M.-F. F., Dick, T., Noothigattu, R., and Procaccia, A. D. (2019). Envy-free classification. In *Advances in Neural Information Processing Systems*, pages 1238–1248.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *NIPS Tutorial*.

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533.

Bhalgat, A., Gollapudi, S., and Munagala, K. (2013). Optimal auctions via the multiplicative weight method. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 73–90.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159.

Brams, S. J. and Taylor, A. D. (1996). *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press.

Brown, G. W. (1949). *Some Notes on Computation of Games Solutions*. RAND Corporation, Santa Monica, CA.

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

Chang, T. (2011). Debunking the myth of 'homogeneous' asian students. `https://www.educationworld.com/a_admin/debunking_myth_of_homogeneous_asian_students.shtml`.

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

Chen, X., Fain, B., Lyu, L., and Munagala, K. (2019b). Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1032–1041.

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

Cotter, A., Gupta, M., and Narasimhan, H. (2019). On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems*, pages 10910–10920.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. (2012). Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Hajian, S. and Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1934–1943.

Hoggett, P., Wilkinson, H., and Beedell, P. (2013). Fairness and the politics of resentment. *Journal of Social Policy*, 42(3):567–585.

Hossain, S., Mladenovic, A., and Shah, N. (2020). Designing fairly fair classifiers via economic fairness notions. In *Proceedings of the 29th International World Wide Web Conference*.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2569–2577.

Kelly, F. P., Maulloo, A. K., and Tan, D. K. (1998). Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252.

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 40–40.

Koren, J. R. (2016). What does that web search say about your credit. *Los Angeles Times*.

Meier, K. J. and Melton, E. K. (2012). Latino heterogeneity and the politics of education: The role of context. *Social science quarterly*, 93(3):732–749.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Robertson, J. and Webb, W. (1998). *Cake-cutting algorithms: Be fair if you can*. AK Peters/CRC Press.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301.

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

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239.