Economic Notions of Fairness for Machine Learning

Anonymous Author(s)
Submission Id: 1482

ABSTRACT

The past decade has witnessed a rapid growth of research on fairness in machine learning. In contrast, fairness has been formally studied for almost a century in microeconomics in the context of resource allocation, during which many general-purpose notions of fairness have been proposed. This paper explore the applicability of two such notions — envy-freeness and equitability — in machine learning. We propose novel relaxations of these fairness notions which apply to groups rather than to individuals, and are compelling in a broad range of settings. Our approach provides a unifying framework by incorporating several recently proposed fairness definitions as special cases. We provide generalization bounds for our approach, and theoretically and experimentally evaluate the tradeoff between loss minimization and our fairness guarantees.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*.

KEYWORDS

Group envy-freeness, group equitability, fairness, generalization

ACM Reference Format:
Anonymous Author(s). 2018. Economic Notions of Fairness for Machine Learning. In *Proceedings of WebConf '20: The Web Conference (WebConf '20)*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Machine learning algorithms are now ubiquitously used to automate decisions which affect human lives (e.g. deciding credit ratings, filtering resumes of job applicants, or making decisions regarding bails or loan applications). Their proliferation raises concerns that these algorithms might amplify human biases or introduce new sources of unfairness [6]. Such concerns have led to a recent explosion in research on fairness in machine learning [10, 14, 17, 22, 27, 28].

While yielding insights on how to make algorithms fairer, it has also led to plethora of fairness definitions [32], many of which are incompatible [14, 27]. There is a general lack of consensus on which is the *right* definition, and this choice is often application-dependent [25]. Further, most popular definitions such as statistical parity [10, 17] and equalized odds [22] only apply to restrictive *binary* settings (e.g. where a loan application can be either approved or rejected); there are few definitions or general frameworks for considering fairness across a broad range of settings [24].

While fairness in machine learning is a recent phenomenon, fairness has been formally studied in microeconomics (especially in fair resource allocation) since almost a century [37]. Initiated by studying the canonical *cake-cutting* setting, it has since focused on proposing general-purpose definitions such as proportionality [37], envy-freeness [19], equitability, the core [41], and Rawlsian egalitarian fairness [34], which apply to a broad range of settings. For example, the core is not only applicable in cake-cutting [41], but also in participatory budgeting [18], housing markets [36], matching markets [20], public goods allocation [18], and even clustering [13].

Recently, a number of papers emerged using these definitions to design fair machine learning algorithms [5, 40, 43]. One central fairness notion adopted by them all is *envy-freeness*, which mandates that no individual should envy another individual. Formally, this is written as $\forall i, j : u_i(o_i) \geq u_i(o_j)$, where u_i is the utility function of individual i and o_i is the outcome experienced by her.

Envy-freeness is compelling because it is simple, intuitive, and requires no information beyond individuals' utility functions, which can be easily learned easily from their actions [4, 12]; this is in contrast to definitions like *individual fairness* [17], which requires access to a task-specific similarity metric between people. However, it has a significant drawback. While it can be exactly satisfied in classic resource allocation settings like cake-cutting [38] or rent division [39], it is often too stringent for many machine learning applications. For example, in *binary* settings with only two outcomes where all individuals prefer the same outcome (e.g. prefer receiving a loan/bail than not receiving it), envy-freeness would require that all individuals receive the same outcome. In applications like targeted advertising where people have heterogeneous preferences, envy-freeness is less restrictive, but only when randomized assignments are allowed [5].

Almost all of this discussion applies to another key fairness definition, equitability, which is formally stated as $\forall i, j : u_i(o_i) = u_j(o_j)$. That is, all individuals must have the same utility for their own outcome. To illustrate its distinction from envy-freeness, consider a hypothetical setting where individual 1 has utility 1 for outcome A but 0 for outcome B, while individual 2 has utility 1 for B but 0 for A. Assigning outcome B to both individuals is envy-free (indeed, individual 1 does not envy individual 2 as they both receive the same outcome), but not equitable (individual 1 receives utility 0 whereas individual 2 receives utility 1).

The research question we address in this paper is: *Are there relaxations of envy-freeness and equitability which are more appropriate for machine learning settings?*

1.1 Our Contributions

We propose novel relaxations of envy-freeness and equitability in style of classical group fairness notions in machine learning. We

WebConf '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

are interested in ensuring fairness between a pair of groups $(G, \widehat{G})$, where groups are defined as arbitrary subsets of individuals. A classifier is group envy-free for this pair of groups if, on average, individuals in G have no less utility for their own classification outcome than for the outcomes of individuals in $\widehat{G}$. For the classifier to be group equitable for this pair of groups, the average utility of individuals in G for their own outcome should equal the average utility of individuals in $\widehat{G}$ for their own outcome.

Section 4 shows that both these notions capture several prior fairness notions such as statistical parity, equal opportunity, equalized odds, and equalized financial impact, as special cases, and extend them beyond binary classification. For equalized odds, Hardt et al. [22] show that a non-discriminatory classifier can be post-processed to satisfy equalized odds without access to feature vectors. We show such post-processing is not generally possible in our setting without access to utility functions (which depend on feature vectors).

Section 5 provides generalization error bounds for both our fairness notions using Rademacher complexity. We show that that classifiers that provide good fairness guarantees on polynomially large training set can also provide good fairness guarantees on the population, even when the pairs of groups for which fairness is sought is exponential.

In Section 6, we show that in the worst case, fairness and loss minimization are not very compatible: simply minimizing the empirical loss without any regards to group envy-freeness or group equitability can quite unfair, whereas imposing either fairness requirement can significantly increase the loss of the classifier.

Section 7 qualifies these observations by performing simulations in the targeted advertising setting. First, we derive efficient methods for training classifiers with low violations of group envy-freeness and group equitability constraints. Next, we observe that our method provides a good tradeoff between empirical risk minimization (which has very low loss but highly unfair) and trivial methods for achieving fairness (which is highly fair, but has very high loss). We also observe that empirically, group envy-freeness is much less imposing than group equitability or (individual) envy-freeness, indicating it may be better suited for practical applications.

1.2 Related Work

Envy-freeness in machine learning. Closely related to ours is the work of Balcan et al. [5], who consider envy-free classifiers in targeted advertising context. As described above, envy-freeness is a stringent constraint for machine learning; we also empirically observe this in our experiments (Section 7). Further, envy-freeness places a constraint for *every pair of individuals*, thus generating an extremely large number of constraints.

Envy-freeness with decoupled classifiers. Zafar et al. [43] and Ustun et al. [40] also adapt envy-freeness to the machine learning context. Though they define a notion of envy-freeness among groups that is similar to our group envy-freeness notion, there is a key difference. They work with group-conditional or decoupled classifiers, where the principal trains a potentially different classifier h_G for each group G. Then, it is said that group G does not envy group $\widehat{G}$ if $\mathbb{E}_{x \sim G}\, u(x, h_{\widehat{G}}(x) - u(x, h_G(x)) \leq 0$. That is, on average, an individual x from group G should prefer the outcome of h_G on x than the outcome of $h_{\widehat{G}}$ on x. We argue that this can allow the principal to satisfy the fairness requirement without actually being fair to groups. For example, consider a scenario in which all individuals prefer class 1 over class 0. The principal trains $(h_G, h_{\widehat{G}})$ such that h_G assigns class 0 to everyone in group G, whereas $h_{\widehat{G}}$ assigns class 1 to everyone in group $\widehat{G}$. When both groups are equally deserving of class 1, this is clearly unfair. However, $h_{\widehat{G}}$ may be a classifier which, when applied on any individual x from group G, detects membership in G using the feature vector and returns class 0 in that case. Then, these decoupled classifiers will satisfy envy-freeness according to the definitions of Ustun et al. [40], Zafar et al. [43]. In contrast, note that we require that individuals in group G not prefer the classification given to individuals in group $\widehat{G}$ (and not just the classification that would be given to them if the classifier for group $\widehat{G}$ were used for them). Hence, these unfair classifiers would significantly violate our group envy-freeness notion. We also note that auditing for fairness is much more difficult under their notion than ours. Checking group envy-freeness under our notion simply requires knowing the *classification outcomes*, which is often public information, whereas checking it under their notion requires access to the actual *classifiers*, which are often kept private [25].

Welfare-equalizing fairness. Concurrently to (and independently of) our work, Ben-Porat et al. [8] propose welfare-equalizing fairness, which coincides with our group equitability notion. They also argue that this subsumes classic fairness notions like statistical parity, equal opportunity, and equalized odds like we do for both group envy-freeness and group equitability in Section 4. However, their main focus is observing that equalized odds may harm even the disadvantaged group when utilities are not considered [24], whereas this cannot happen under group equitability, lending it further credibility. They also identify the structure of optimal group equitable classifiers in a certain context. Although one of our notions coincides with their proposal, our contribution is entirely different. We provide generalization error bounds for our fairness notions, and theoretically and empirically evaluatethe tradeoff between loss minimization and fairness, which they do not do.

Fair division. Envy-freeness originated in microeconomics literature on fair allocation of resources [19]. In that context, envy-freeness is easy to achieve either exactly [41] or approximately [9, 11, 29]. Hence, the literature has focused on group-level notions of fairness which *strengthen* (i.e. logically imply) envy-freeness, such as group envy-freeness [7] or group fairness [15]. These should not be confused with our notion of group envy-freeness, which is a *relaxation* of individual envy-freeness. We term it group envy-freeness because in machine learning, notions that relax individual-level fairness to groups are referred to as group fairness notions, and there are no notions that strengthen individual fairness because individual fairness is already severely restrictive.

In classical fair division, fairness notions which average utilities across individuals are usually not considered as individual utilities can be on different scales (or expressed in different units). Interpersonal comparisons of utilities is thus avoided [30]; an exception to this is the work of Aleksandrov and Walsh [2]. In machine learning contexts, however, utility can often be interpreted in terms of the probability of receiving the preferred outcome [43] or financial impact [33], and thus can be on the same scale.

We also remark that there is significant potential of importing further ideas from this literature. For example, the recent work of Gölz et al. [21] considers the implications of requiring monotonicity axioms from fair division in the classification context, and show that some axioms are easy to guarantee in conjunction with equalized odds, while others are effectively incompatible.

Generalization. In proving generalization of our notions, we use the Rademacher complexity framework, but tie it to the Natarajan dimension of the family of multiclass classifiers. Balcan et al. [5] also used the Natarajan dimension when analyzing generalization of envy-freeness. However, they only establish that a large fraction of constraints will be approximately satisfied with high probability, whereas in our setting, all constraints are approximately satisfied with high probability.[1] We also note that our approach can provide fairness across exponentially many pairs of *arbitrarily defined* groups with polynomially many training data points; similar approaches in the literature for providing fairness across exponentially many groups require the family of groups to have "low complexity" in a certain sense [23, 26].

2 PRELIMINARIES

For a natural number $k \in \mathbb{N}$, define $[k] = \{1, \ldots, k\}$. For a set T, let $\Delta(T)$ denote the set of all distributions over T.

We are interested in a classification setting where the task is to learn to classify individuals into appropriate classes. Typically, an individual is represented by a feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^m$, which is accessible to the classifier. In some classification settings, there is also side information (e.g. a ground truth label y^* for each individual), which, while not available to the classifier, could be used as part of training. To capture this general setting, we represent individuals by the *extended feature vector* $x^+ \triangleq (x, y^*) \in \mathcal{X}^+$, where y^* is any side information. We use this abstract notation to convey the fact that our definitions and framework apply to machine learning settings with a ground truth (e.g. the loan or bail setting) as well as those without (e.g. the targeted ad setting). Let $\mathcal{P}^{\mathcal{X}^+}$ denote a distribution over individuals.

Let there be a finite set of classes (a.k.a. labels) $\mathcal{Y} = [d]$. Note that throughout the paper, we consider multiclass classifiers. As before, let $\mathcal{P}^{\mathcal{Y}} \in \Delta(\mathcal{Y})$ denote a distribution over classes.

Classifiers: A *deterministic classifier* $h_d : \mathcal{X} \to \mathcal{Y}$ assigns a class to each individual. A *randomized classifier* $h_r : \mathcal{X} \to \Delta(\mathcal{Y})$ assigns a distribution over classes to each individual. We use $\mathcal{H}$ to denote a family of classifiers. For a family $\mathcal{H}$ of deterministic classifiers, we use $\Delta^k(\mathcal{H})$ to denote the family of randomized classifiers which can be expressed as a mixture over k deterministic classifiers from $\mathcal{H}$, i.e., $\Delta^k(\mathcal{H})$ is the set of all h_r such that $h_r = \sum_{t=1}^k \eta^t h_d^t$ for some $h_d^1, \ldots, h_d^k \in \mathcal{H}$ and $\eta^1, \ldots, \eta^k \in [0, 1]$ with $\sum_{t=1}^k \eta^t = 1$.

Loss functions: A *loss function* is a function $\ell : \mathcal{X}^+ \times \mathcal{Y} \to [0, 1]$, where $\ell(x^+, y)$ return the loss in predicting label y for individual x^+.[2] For a randomized prediction $\mathcal{P}^{\mathcal{Y}}$, we extend the loss function and define $\ell(x^+, \mathcal{P}^{\mathcal{Y}}) = \mathbb{E}_{y \sim \mathcal{P}^{\mathcal{Y}}}[\ell(x^+, y)]$.

[1] This is because in an infinite population, envy-freeness imposes infinitely many constraints, whereas our approach imposes finitely many constraints.

[2] The loss must be bounded, but the restriction to $[0, 1]$ is without loss of generality.

Given a loss function ℓ and a finite dataset $S \subseteq \mathcal{X}^+$, the empirical risk of a classifier h is given by $R_S(h) = \frac{1}{|S|} \cdot \sum_{x_i^+ \in S} \ell(x_i^+, h(x_i))$. The classifier which minimizes this empirical risk is termed the *empirical risk minimizer* (ERM). The expected loss of the classifier on the population, defined by a distribution $\mathcal{P}^{\mathcal{X}^+}$ over individuals, is $R(h) = \mathbb{E}_{x^+ \sim \mathcal{P}^{\mathcal{X}^+}}[\ell(x^+, h(x))]$.

Utility: A utility function is given by $u : \mathcal{X}^+ \times \mathcal{Y} \to [0, 1]$, where $u(x^+, y)$ encodes the utility of individual x^+ being assigned class y.[3] We assume that individual utilities are normalized: for each $x^+ \in \mathcal{X}^+$, $\sum_{y \in \mathcal{Y}} u(x^+, y) = 1$. Again, with a slight abuse of notation, we define $u(x^+, \mathcal{P}^{\mathcal{Y}}) = \mathbb{E}_{y \sim \mathcal{P}^{\mathcal{Y}}}[u(x^+, y)]$.

Note that we allow the utility function of an individual represented by x^+ to depend on the side information (such as a ground truth label for the individual). That said, we assume that utility function of an individual only depends on features captured by x^+. In practice, two individuals with identical x^+ may have slightly different utilities, but our results hold approximately if a close approximation of their individuals' utility functions can be found which depend only on x^+.

3 GROUP ENVY-FREENESS AND GROUP EQUITABILITY

Our main conceptual contribution in this work is to propose two group fairness notions for machine learning, inspired by the literature on fair division. For this, we first define the notion of groups.

Groups: Unlike much prior literature on fairness in machine learning where groups are defined based on certain *sensitive attribute* (e.g. race, gender, ethnicity, etc.), our framework allows groups to be defined arbitrarily. A *group* of individuals G is identified by a subset of extended feature vectors, i.e., $G \subseteq \mathcal{X}^+$. Our fairness guarantees apply to pairs of groups. Let $\mathcal{G}$ denote a set of pairs of groups; we want to ensure fairness across all pairs of groups $(G, \widehat{G}) \in \mathcal{G}$. We are now ready to define our group fairness notions.

Group envy-freeness: In the fair division literature, envy-freeness is a notion of individual fairness, which requires that no individual should envy any other individual. This was adapted to the classification context by Balcan et al. [5], and formally translates to the following: a classifier h is *envy-free* if $\forall x^+, \widehat{x}^+ \in \mathcal{X}^+ : u(x^+, h(x)) \geq u(x^+, h(\widehat{x}))$. Another way of viewing the envy-freeness is that the envy of any individual for any other individual is non-negative: $u(x^+, h(\widehat{x})) - u(x^+, h(x)) \leq 0, \forall x^+, \widehat{x}^+ \in \mathcal{X}^+$. As argued in the introduction, this is a very stringent requirement in most applications. For example, in the loan/bail domain, this requires either granting all loan/bail applications or denying them all.[4] For the targeted advertisement domain, this translates to showing each individual her most preferred ad out of all ads shown to anyone.[5]

We propose a group-level relaxation of this constraint, following a similar relaxation proposed by Aleksandrov and Walsh [2]. Instead of mandating each individual prefer their outcome to anyone else's,

[3] Once again, the utility must be bounded, but the restriction to $[0, 1]$ is without loss of generality.

[4] This assumes every individual prefers receiving loan/bail to not receiving it. For randomized classifier, this would translate to granting loan/bail to each individual with exactly equal probability.

[5] The requirement becomes a bit less stringent for randomized classifiers, as observed by Balcan et al. [5].

we require that the average preference (i.e. utility) of individuals in a group for their outcome be higher than their average preference for outcomes given to another group. Formally, given a pair of groups $G, \widehat{G} \subseteq \mathcal{X}^+$, a dataset $S \subseteq \mathcal{X}^+$, and $\epsilon \geq 0$, we say that classifier h is empirically ϵ-group-envy-free on $(G, \widehat{G})$ with respect to S if

$$\frac{1}{|S^G| \cdot |S^{\widehat{G}}|} \sum_{x^+ \in S^G, \widehat{x}^+ \in S^{\widehat{G}}} u(x^+, h(\widehat{x})) - u(x^+, h(x)) \leq \epsilon,$$

where $S^G = S \cap G$ and $S^{\widehat{G}} = S \cap \widehat{G}$ represent restrictions of S to groups G and $\widehat{G}$, respectively. We refer to this difference as the empirical group envy of G for $\widehat{G}$ on S. When $\epsilon = 0$, we simply refer to this as empirical group envy-freeness. Note that while we want the group envy to be non-negative (or minimally positive), having large negative group envy is not necessarily desirable. Also, like envy-freeness, group envy-freeness is not symmetric: group envy-freeness on $(G, \widehat{G})$ does not imply group envy-freneess on $(\widehat{G}, G)$. In fact, as we argue in Section 8, in certain applications it may be desirable to impose asymmetric group envy-freeness constraints.

The population version is simply given by the expectation over a distribution of individuals $\mathcal{P}^{\mathcal{X}^+}$: we say that classifier h is population ϵ-group-envy-free on $(G, \widehat{G})$ if

$$\mathbb{E}\left[u(x^+, h(\widehat{x})) - u(x^+, h(x)) \,\middle|\, x^+ \in G, \widehat{x}^+ \in \widehat{G}\right] \leq \epsilon.$$

Again, when $\epsilon = 0$, we simply refer to this as population group-envy-freeness.

Our goal is to ensure that this style of group fairness is maintained across a set of pairs of groups $\mathcal{G}$. We say that h is empirically (resp. population) ϵ-group-envy-free on $\mathcal{G}$ with respect to S (resp. $\mathcal{P}^{\mathcal{X}^+}$) if it is ϵ-group-envy-free on $(G, \widehat{G})$ with respect to S (resp. $\mathcal{P}^{\mathcal{X}^+}$) for all $(G, \widehat{G}) \in \mathcal{G}$.

Group equitability: In the classical fair division literature, equitability requires that all agents have identical utility for their outcomes. This translates to the following definition in our classification context: a classifier h is *equitable* if $\forall x^+, \widehat{x}^+ : u(x^+, h(x)) = u(\widehat{x}^+, h(\widehat{x}))$. Unfortunately, this too is a stringent requirement in machine learning contexts similarly to envy-freeness.

We relax this to a group setting by requiring the average utility of individuals in a group for their outcome be equal across all groups. Formally, given a pair of groups $G, \widehat{G} \subseteq \mathcal{X}^+$, a dataset $S \subseteq \mathcal{X}^+$, and $\epsilon \geq 0$, we say that classifier h is empirically ϵ-group-equitable on $(G, \widehat{G})$ with respect to S if

$$\left| \frac{1}{|S^G|} \sum_{x^+ \in S^G} u(x^+, h(x)) - \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}^+ \in S^{\widehat{G}}} u(\widehat{x}^+, h(\widehat{x})) \right| \leq \epsilon.$$

We refer to this difference as the empirical group equitability violation between G and $\widehat{G}$ on S. When $\epsilon = 0$, we simply refer to this as empirical group equitability. Given a distribution of individuals $\mathcal{P}^{\mathcal{X}^+}$, we say that classifier h is population ϵ-group-equitable on $(G, \widehat{G})$ if

$$\left| \mathbb{E}\left[u(x^+, h(x)) \,\middle|\, x^+ \in G\right] - \mathbb{E}\left[u(\widehat{x}^+, h(\widehat{x})) \,\middle|\, \widehat{x}^+ \in \widehat{G}\right] \right| \leq \epsilon.$$

Again, when $\epsilon = 0$, we simply refer to this as population group-equitability.

Given a set of pairs of groups $\mathcal{G}$, we say that h is empirically (resp. population) ϵ-group-equitable on $\mathcal{G}$ with respect to S (resp. $\mathcal{P}^{\mathcal{X}^+}$) if it is ϵ-group-equitable on $(G, \widehat{G})$ with respect to S (resp. $\mathcal{P}^{\mathcal{X}^+}$) for all $(G, \widehat{G}) \in \mathcal{G}$. Lastly, we note that unlike group envy free, group equitability is symmetric: group equitable on $(G, \widehat{G})$ implies group equitable on $(\widehat{G}, G)$.

4 CLASSICAL FAIRNESS NOTIONS AS SPECIAL CASES

In this section, we discuss the connection of group envy-freeness and group equitability to several classical notions of fairness proposed in the fair machine learning literature.

4.1 Statistical Parity, Equal Opportunity, and Equalized Odds

First, we show that three popular group fairness notions for binary classification — statistical parity, equal opportunity, and equalized odds — are special cases of our definitions. Thus, our definitions provide a unifying framework for viewing classical definitions under one umbrella and generalizing them to multiclass classification.

Recall the binary classification framework. Each individual x also has a ground truth label y^*. Recall that y_x denotes the class assigned to the individual. When the individual is sampled from population, we use X, Y^*, and Y to denote the corresponding random variables. In this setting, there is a positive class (say $y = 1$), which is preferred by all individuals. For example, this may correspond to receiving a loan or bail. Given a pair of groups $(G, \widehat{G})$, the three aforementioned notions of fairness — which treat both groups symmetrically — are defined as follows.

- *Statistical parity* demands an equal probability of getting a positive classification, regardless of group identity: $\Pr[Y = 1|X \in G] = \Pr[Y = 1|X \in \widehat{G}]$.
- *Equal opportunity* is similar to statistical parity, except that we now condition on both group identity and positive ground truth class: $\Pr[Y = 1|X \in G, Y^* = 1] = \Pr[Y = 1|X \in \widehat{G}, Y^* = 1]$.
- *Equalized odds* is similar to equal opportunity, except that we seek fairness across both positive and negative ground truth classes: $\Pr[Y = 1|X \in G, Y^* = a] = \Pr[Y = 1|X \in \widehat{G}, Y^* = a]$ for all $a \in \{0, 1\}$.

Theorem 1. *Given a pair of groups $(G, \widehat{G})$, there is a set $\mathcal{G}$ of pairs of groups and individual utility functions such that group envy-freeness and group equitability with respect to $\mathcal{G}$ coincide with statistical parity with respect to $(G, \widehat{G})$. The same holds for equal opportunity and equalized odds.*

Proof. Let all individuals have utility 1 for the preferred class, and 0 for the less preferred one. That is, $u(x^+, 1) = 1$ and $u(x^+, 0) = 0$. Then, for a random class Y, we have $u(x^+, Y) = \Pr[Y = 1]$. For a pair of groups $(G, \widehat{G})$, a classifier h is group envy-free with respect to $(G, \widehat{G})$ if

$$\mathbb{E}\left[\Pr[h(\widehat{x}) = 1] - \Pr[h(x) = 1] \,\middle|\, x^+ \in G, \widehat{x}^+ \in \widehat{G}\right] \leq 0$$
$$\Leftrightarrow \Pr[Y = 1|X \in \widehat{G}] \leq \Pr[Y = 1|X \in G].$$

Hence, the classifier is group envy-free with respect to both $(G, \widehat{G})$ and $(\widehat{G}, G)$ if and only if $\Pr[Y = 1|X \in \widehat{G}] = \Pr[Y = 1|X \in G]$, which is the condition for the classifier to satisfy statistical parity with respect to $(G, \widehat{G})$. Hence, $\mathcal{G} = \left\{(G, \widehat{G}), (\widehat{G}, G)\right\}$ suffices.

It is easy to see that for equitability, simply $\mathcal{G} = (G, \widehat{G})$ suffices as equitability is already a symmetric condition.

Finally, equal opportunity with respect to $(G, \widehat{G})$ is simply statistical parity with respect to $(G_1, \widehat{G}_1)$, where $G_a = G \cap \left\{x^+ \in \mathcal{X}^+ : y^* = a\right\}$, $a \in \{0, 1\}$. Hence, given the above proof, this can be obtained as group envy-freeness with respect to $\mathcal{G} = \{(G_1, \widehat{G}_1), (\widehat{G}_1, G_1)\}$ and group equitability with respect to $\mathcal{G} = \{(G_1, \widehat{G}_1)\}$. Similarly, equalized odds with respect to $(G, \widehat{G})$ is simply statistical parity with respect to both $(G_1, \widehat{G}_1)$ and $(G_0, \widehat{G}_0)$, which is group envy-freeness with respect to $\mathcal{G} = \{(G_1, \widehat{G}_1), (\widehat{G}_1, G_1), (G_0, \widehat{G}_0), (\widehat{G}_0, G_0)\}$ and group equitability with respect to $\mathcal{G} = \{(G_1, \widehat{G}_1), (G_0, \widehat{G}_0)\}$. ◻

While we provide a proof for equivalence of the population versions of these notions, it is easy to see that the equivalence also holds for the empirical versions defined on a training set. Similarly, our definitions can also capture statistical parity, equal opportunity, or equalized odds with respect to multiple pairs of groups.

The fact that these three classical fairness definitions are subsumed by group envy-freeness and group equitability has two key implications. First, our framework provides a methodical approach to extend these fairness definitions from binary classification to multi-class classification. Second, our generalization guarantee from Section 5 using the Rademacher complexity provides a single generalization proof for all three classic fairness definitions. This is similar to the approach and generalization proof of Agarwal et al. [1]; however, their framework is limited to binary classification.

4.2 Equalized Financial Impact

Sometimes, even in binary classification, simply equalizing error rates across subpopulations defined by sensitive attributes (and possibly the ground truth classes) may not be sufficient. Ramnarayan [33] considers the loan setting, and argues that the financial impact of an error made by the classifier (i.e. where the outcome differs from the ground truth label) may vary across individuals within a group depending on their wealth, and proposes equalizing the financial impact rather than error rates across the protected groups. This is similar to the harm-reduction approach of Altman et al. [3].

Formally, let $b(x^+)$ denote some measure of wealth of an individual represented by x^+, and let $\psi(b(x^+))$ denote the financial impact when individual x^+ receives a loan. Ramnarayan [33] assumes that $b(x^+)$ is simply one of the features; however, our approach allows b to be any function of the features. Then, a classifier satisfies equalized financial impact with respect to a pair of groups $(G, \widehat{G})$ if $\mathbb{E}[\Pr[Y = 0] \cdot \psi(b(X))|X \in G, Y^* = 1] = \mathbb{E}[\Pr[Y = 0] \cdot \psi(b(X))|X \in \widehat{G}, Y^* = 1]$. Similarly to Theorem 1, it is easy to see that this is also a special case of group envy-freeness and group equitability, where ψ defines the utility functions.

Theorem 2. *Given a pair of groups $(G, \widehat{G})$, there is a set $\mathcal{G}$ of pairs of groups and utility functions of individuals such that group envy-freeness and group equitability with respect to $\mathcal{G}$ coincide with equalized financial impact with respect to $(G, \widehat{G})$.*

4.3 Impossibility of Post-Processing

Hardt et al. [22] show that given any (possibly discriminatory) binary classifier, one can derive from it a binary classifier satisfying equalized odds (or equal opportunity) using a simple post-processing step. This post-processing step does not require access to any feature vector information from the training data except for group membership and ground truth labels. It achieves fairness by simply taking an appropriate convex combination of the given classifier, its inverse (which flips the prediction on each individual), and trivial constant classifiers.

While such post-processing is clearly desirable, we show that when we move beyond the binary classification setting, we cannot hope to post-process an arbitrary given classifier and achieve fairness. For example, if we start from the empirical risk minimizer (ERM) which is obtained without accessing utilities, and perform a post-processing step which also does not access utilities, then any classifier derived is ultimately obtained without accessing utilities. We show that such classifiers can only guarantee group envy-freeness or group equitability in trivial cases. For these results, we assume that group membership is exclusive, i.e., we want to ensure fairness with respect to a pair of groups $(G, \widehat{G})$ where $G \cap \widehat{G} = \emptyset$; note that we do not generally require this in our framework, although this is a common use case.

Theorem 3. *Suppose h is a (possibly randomized) classifier obtained without access to utilities, $(G, \widehat{G})$ is a pair of groups with $G \cap \widehat{G} = \emptyset$, and S is a finite set of individuals. Then:*

1. *h is guaranteed to be empirically group envy-free on S with respect to $(G, \widehat{G})$ if and only if $h(x)$ is identical for all $x^+ \in S^G$, given by the following equation:*

$$\Pr[h(x) = c] = \frac{1}{|S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} \Pr[h(\widehat{x}) = c], \ \forall x^+ \in S^G, c \in \mathcal{Y}.$$

2. *h is guaranteed to be empirically group equitable on S with respect to $(G, \widehat{G})$ if and only if for all $x^+ \in S^G$ and $\widehat{x}^+ \in S^{\widehat{G}}$, we have that $h(x) = h(\widehat{x}) = \mathcal{U}(\mathcal{Y})$, where $\mathcal{U}(\mathcal{Y})$ represents the uniform distribution over the set of classes $\mathcal{Y}$.*

Proof. Let us first consider group envy-freeness. For any h satisfying the equation given in part (1), we can see that the average empirical envy of any $x^+ \in S^G$ towards $\widehat{G}$ is

$$\frac{1}{|S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} u(x^+, h(\widehat{x}) - u(x^+, h(x))$$

$$= \frac{1}{|S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} \sum_{c \in \mathcal{Y}} u(x^+, c) \cdot (\Pr[h(\widehat{x}) = c] - \Pr[h(x) = c])$$

$$= \sum_{c \in \mathcal{Y}} u(x^+, c) \cdot \left(\left(\frac{1}{|S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} \Pr[h(\widehat{x}) = c] \right) - \Pr[h(x) = c] \right) = 0,$$

where the last equality uses the condition in part (1). Since this holds for all $x^+ \in S^G$, clearly h is empirically group envy-free with respect to $(G, \widehat{G})$.

To see the converse, suppose for contradiction that there exists $x_i^+ \in S^G$ violating the condition in part (1). Then, there exists $c_i \in \mathcal{Y}$ such that $\Pr[h(x_i) = c_i] < \frac{1}{|S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} \Pr[h(\widehat{x}) = c_i]$. Since h

was constructed without access to utility functions, the underlying utilities could have been such that $u(x_i^+, c_i) = 1$, $u(x_i^+, c) = 0$ for all $c \in \mathcal{Y} \setminus \{c_i\}$, and $u(x^+, c) = 1/|\mathcal{Y}|$ for all $x^+ \in S^G \setminus \{x_i^+\}, c \in \mathcal{Y}$. Then, the group envy of G towards $\widehat{G}$ is

$$\begin{aligned}
&\frac{1}{|S^G| \cdot |S^{\widehat{G}}|} \cdot \sum_{x^+ \in S^G, \widehat{x}^+ \in S^{\widehat{G}}} u(x^+, h(\widehat{x}) - u(x^+, h(x)) \\
&= \frac{1}{|S^G| \cdot |S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} u(x_i^+, h(\widehat{x}) - u(x_i^+, h(x_i)) \\
&= \frac{1}{|S^G| \cdot |S^{\widehat{G}}|} \cdot \sum_{\widehat{x}^+ \in S^{\widehat{G}}} (\Pr[h(\widehat{x}) = c_i] - \Pr[h(x_i) = c_i]) \\
&= \frac{1}{|S^G|} \cdot \left(\frac{\sum_{\widehat{x}^+ \in S^{\widehat{G}}} \Pr[h(\widehat{x}) = c_i]}{|S^{\widehat{G}}|} - \Pr[h(x_i) = c_i] \right) > 0,
\end{aligned}$$

where the first transition follows from the fact that $u(x^+, \mathcal{P}^{\mathcal{Y}}) = 1/d$ for every $x^+ \in S^G \setminus \{x_i^+\}$ and every classification $\mathcal{P}^{\mathcal{Y}}$ [SH: maybe say and satisfying part 1?], and the last transition follows from the way x_i^+ was constructed.

For group equitability, note that the condition in part (2) is very strong. If h satisfies this condition, then $u(x^+, h(x)) = u(\widehat{x}^+, h(\widehat{x})) = 1/d$, where $d = |\mathcal{Y}|$. Hence, group equitability is trivially satisfied. We now show that this is also necessary. For each $x_i^+ \in S^G$, define $t_i^{\min} = \arg\min_c \Pr[h(x_i) = c]$. For each $\widehat{x}_j^+ \in S^{\widehat{G}}$, define $\widehat{t}_j^{\max} = \arg\max_c \Pr[h(\widehat{x}_j) = c]$. Note that for each $x_i^+ \in S^G$, $\Pr[h(x_i) = t_i^{\min}] \le 1/d$, and for each $\widehat{x}_j^+ \in S^{\widehat{G}}$, $\Pr[h(\widehat{x}_j) = \widehat{t}_j^{\max}] \ge 1/d$.

Now, consider the following utilities: $\forall x_i^+ \in S^G : u(x_i^+, t_i^{\min}) = 1$ and $u(x_i^+, c) = 0$ for all $c \neq t_i^{\min}$, and $\forall \widehat{x}_j^+ \in S^{\widehat{G}}$, $u(\widehat{x}_j^+, \widehat{t}_j^{\max}) = 1$ and $u(\widehat{x}_j^+, c) = 0$ for all $c \neq \widehat{t}_j^{\max}$. Under these utilities, it is easy to check that h is empirically group equitable on S with respect to $(G, \widehat{G})$ if and only if $(1/|S^G|) \cdot \sum_{x_i^+ \in S^G} \Pr[h(x_i) = t_i^{\min}] = (1/|S^{\widehat{G}}|) \cdot \sum_{\widehat{x}_j^+ \in S^{\widehat{G}}} \Pr[h(\widehat{x}_j) = \widehat{t}_j^{\max}]$. However, this requires $\Pr[h(x_i) = t_i^{\min}] = \Pr[h(\widehat{x}_j) = \widehat{t}_j^{\max}] = 1/d$ for each $x_i^+ \in S^G, \widehat{x}_j^+ \in S^{\widehat{G}}$. By the definitions of $t_i^{\min}$ and $\widehat{t}_j^{\max}$, we get that $h(x_i) = h(\widehat{x}_j) = \mathcal{U}(\mathcal{Y})$ for all $x_i^+ \in S^G, \widehat{x}_j^+ \in S^{\widehat{G}}$, as desired. □

Note that if we require the classifier h to be empirically group envy-free on S with respect to both $(G, \widehat{G})$ and $(\widehat{G}, G)$ (to make the requirement symmetric between the groups), then we obtain that $h(x) = h(\widehat{x})$ must hold for all $x^+ \in S^G, \widehat{x}^+ \in S^{\widehat{G}}$. Though less strict than the requirement for group equitability, it is still too restrictive in practice. Hence, post-processing cannot produce reasonable classifiers in our more general setting, without accessing individual utilities. We remark that even in the binary classification setting with homogeneous preferences, it has been observed that any post-processing which does not access the features can be very suboptimal in performance [42].

5 GENERALIZATION

Our learning problem seeks a classifier that has low empirical risk and satisfies (or minimally violates) group envy-freeness or group equitability constraints on the training data. However, this classifier is then used to classify all individuals in the population. Hence, it is crucial that our fairness definitions generalize well. That is, we seek to establish that classifiers which are approximately fair on training data are also approximately fair on the population according to our fairness definitions. For this purpose, we use the Rademacher complexity approach.

Let $\mathcal{G}$ denote a finite set of pairs of groups. Let us denote the k^{th} pair be denoted by $(G_{k,1}, G_{k,2})$.[6] Let b_k denote the corresponding membership function: for a pair of individuals $z = (x_1^+, x_2^+)$, we have the indicator $b_k(z) \triangleq b_k(x_1^+, x_2^+) = \mathbb{1}[x_1^+ \in G_{k,1} \wedge x_2^+ \in G_{k,2}]$. Let $\mathcal{B} = \{b_1, \ldots, b_{|\mathcal{G}|}\}$ denote the family of membership functions, with $|\mathcal{B}| = |\mathcal{G}|$. Let $\mathcal{P}^{\mathcal{X}^+ \times \mathcal{X}^+}$ denote any joint distribution over pairs of individuals, and let S denote a finite training set of iid pairs $z_i = (x_{i,1}^+, x_{i,2}^+)$ sampled from this joint distribution and $|S| = n$.[7]

Let us now define empirical and population violations of group envy-freeness and group equitability constraints in this framework. For this, we need the following quantities. For $a, b \in \{1, 2\}$, $i \in \{1, \ldots, |\mathcal{G}|\}$, and classifier h, let

$$U_{ab}^S(h, b_k) = \frac{1}{|S|} \sum_{(x_1^+, x_2^+) \in S} u(x_a^+, h(x_b)) \cdot b_k(x_1^+, x_2^+),$$
$$U_{ab}(h, b_k) = \mathbb{E}_{(x_1^+, x_2^+) \sim \mathcal{P}^{\mathcal{X}^+ \times \mathcal{X}^+}} \left[u(x_a^+, h(x_b)) \cdot b_k(x_1^+, x_2^+) \right].$$

Note that $U_{12}^S(h, b_k)$ (resp. $U_{12}(h, b_k)$) effectively measures the average (resp. expected) utility of an individual in group $G_{k,1}$ for the classification given to an individual in group $G_{k,2}$. Similarly, $U_{11}^S(h, b_k)$ and $U_{22}^S(h, b_k)$ (resp. $U_{11}(h, b_k)$ and $U_{22}(h, b_k)$) effectively measure the average (resp. expected) utility of individuals in groups $G_{k,1}$ and $G_{k,2}$.

We can now define the empirical and population group envy-freeness and group equitability violations in terms of these quantities. Hereinafter, we focus on group envy-freeness. The definitions and proofs for group equitability are almost identical. Let the empirical and population group envy be defined as $V^S(h, b_k) = U_{12}^S(h, b_k) - U_{11}^S(h, b_k)$ and $V(h, b_k) = U_{12}(h, b_k) - U_{11}(h, b_k)$.[8] To establish generalization, our goal is to show that given a sufficiently large training dataset S, with high probability, the difference $|V^S(h, b_k) - V(h, b_k)|$ is small for all h and all b_k.

We first introduce two necessary definitions.

Definition 1. *For a function class $\mathcal{F}$ containing functions mapping $\mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R}$ and a set of samples $S = \{z_1, \ldots, z_m\}$, the **Rademacher complexity** of $\mathcal{F}$ is*

$$\mathcal{R}(\mathcal{F} \circ S) = \mathbb{E}_{\sigma \in \{-1,+1\}^n} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \cdot f(z_i) \right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ and each σ_i is an independent random variable with $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = 1/2$. For simplicity of notation, we omit S and simply denote this as $\mathcal{R}(\mathcal{F})$.

[6] In this section, we switch from $(G, \widehat{G})$ to $(G_{k,1}, G_{k,2})$ for notational convenience.

[7] When the training data consists of n individuals sampled iid from a distribution $\mathcal{P}^{\mathcal{X}^+}$, we can simply pair up individuals to create a dataset consisting of $n/2$ pairs sampled iid from the product distribution $\mathcal{P}^{\mathcal{X}^+} \times \mathcal{P}^{\mathcal{X}^+}$.

[8] Note that in the difference $U_{12}^S - U_{11}^S$, the denominators of both terms are the same ($|S|$), and in the numerator, we measure the envy for every z with $b_k(z) = 1$ and ignore the term corresponding to every z with $b_k(z) = 0$. Hence, this difference is proportional to the group envy that $G_{k,1}$ has towards $G_{k,2}$. For group equitability, we would need to consider both $U_{11}^S(h, b_k) - U_{22}^S(h, b_k)$ and $U_{22}^S(h, b_k) - U_{11}^S(h, b_k)$ for bounding the empirical violation, and the two similar quantities for bounding the population violation.

Definition 2. *A set $S = \{z_1, \dots, z_m\}$ is multi-class shattered by a function class $\mathcal{F}$, if there exist two functions f_1 and f_2 such that: (1) $\forall z \in S, f_1(z) \neq f_2(z)$ and (2) for every $B \subseteq S$, there exists a function $f \in \mathcal{F}$ such that $f(z) = f_1(z)$ for all $z \in B$ and $f(z) = f_2(z)$ for all $z \in S \setminus B$. The* ***Natarajan dimension*** *of $\mathcal{F}$ is the cardinality of the largest set of points that can be multi-class shattered by $\mathcal{F}$.*

Let $\mathcal{H}$ be a family of deterministic classifiers and recall that $\Delta^k(\mathcal{H})$ contains all randomized classifiers that are mixtures of k classifiers from $\mathcal{H}$. To obtain generalization, we need to bound the Rademacher complexity of

$$\{g : g(x_1^+, x_2^+) = u(x_a^+, h(x_b)) \cdot b_k(x_1^+, x_2^+), h \in \Delta^k(\mathcal{H}), b_k \in \mathcal{B}\}.$$

Our approach is as follows. First, in Lemma 1, we eliminate the dependence on b_k in the product $u(x_a^+, h(x_b)) \cdot b_k(x_1^+, x_2^+)$. Next, in Lemma 2, we eliminate the dependence on both u and the randomized nature of $h \in \Delta^k(\mathcal{H})$, and express our bound directly in terms of $\mathscr{R}(\mathcal{H})$. Combining these results, in Theorem 4, we prove our generalization bound in terms of $|\mathcal{G}|$ and $\mathscr{R}(\mathcal{H})$. Finally, we observe in Theorem 5 that function classes with low Natarajan dimension have low Rademacher complexity. In particular, we show that linear one-vs-all classifiers generalize well. We begin with the first result.

Lemma 1. *Given a function class $\mathcal{F}$ containing functions mapping $\mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R}$ and a binary function $b : \mathcal{X}^+ \times \mathcal{X}^+ \to \{0, 1\}$, define $\mathcal{F}_b = \{f_b : f_b(z) = f(z) \cdot b(z), f \in \mathcal{F}\}$. Then $\mathscr{R}(\mathcal{F}_b) \leq \mathscr{R}(\mathcal{F})$.*

Proof. For each σ, let $f^*_\sigma \in \mathcal{F}$ be where $\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) b(z_i)$ is attained (if the supremum is not attained, we can take a sequence of f^*_σ that converge to the supremum) and define $\overline{\sigma}$ where $\overline{\sigma}_i = \sigma_i$ if $b(z_i) = 1$ and $\overline{\sigma}_i = 0$ otherwise. Then, we can write

$$\begin{aligned}\mathscr{R}(\mathcal{F}_b) &= \mathbb{E}_\sigma \left[\sum_i \sigma_i f^*_\sigma(z_i) b(z_i)\right] = \mathbb{E}_\sigma \left[\sum_{i:b(z_i)=1} \sigma_i f^*_\sigma(z_i)\right] \\ &= \mathbb{E}_\sigma \left[\sum_{i:b(z_i)=1} \sigma_i f^*_{\overline{\sigma}}(z_i)\right] = \mathbb{E}_\sigma \left[\sum_i \sigma_i f^*_{\overline{\sigma}}(z_i)\right] \\ &\leq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i)\right] = \mathscr{R}(\mathcal{F}),\end{aligned}$$

where the third transition holds because $\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) b(z_i) = \sup_{f \in \mathcal{F}} \sum_{i:b(z_i)=1} \sigma_i f(z_i)$ only depends on the values of σ_i where $b(z_i) = 1$, and the fourth transition holds because the expected value of the term corresponding to each i with $b(z_i) = 0$ is 0 (this is because σ_i takes values $+1$ and -1 with probability $1/2$ each, while $f_{\overline{\sigma}}$ does not change). □

Using Lemma 1, we are now express our desired Rademacher complexity directly in terms of the complexity of $\mathcal{H}$, eliminating the dependence on utilities, group membership functions, and randomization over deterministic classifiers. However, classifiers from $\mathcal{H}$ act on individuals, whereas training set S consists of pairs of individuals. Hence, we define $\mathcal{H}_b = \{h_b : h_b(x_1, x_2) = h(x_b); h \in \mathcal{H}\}$, where $b \in \{1, 2\}$. Now, $\mathscr{R}(\mathcal{H}_b) \triangleq \mathscr{R}(\mathcal{H}_b \circ S)$ is well-defined.

Lemma 2. *Given a family of deterministic classifiers $\mathcal{H}$, function $b_k : \mathcal{X}^+ \times \mathcal{X}^+ \to \{0, 1\}$, and $a, b \in \{1, 2\}$, define*

$$\mathcal{F} = \{f : f(x_1^+, x_2^+) = u(x_a^+, h_p(x_b)) \cdot b_k(x_1^+, x_2^+); h \in \Delta^k(\mathcal{H})\}.$$

Then, $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{H}_b)$.

Proof. Let $\mathcal{F}_1 = \left\{f : f(x_1^+, x_2^+) = u(x_a^+, h(x_b)) : h \in \Delta^k(\mathcal{H})\right\}$. By Lemma 1, we have that $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{F}_1)$. For a mixture $h = \sum_{t=1}^k \eta^t h^t_d$, note that $u(x_a^+, h(x_b)) = \sum_{t=1}^k \eta^t u(x_a^+, h^t_d(x_b))$, which is in the convex hull of $\left\{u(x_a^+, h_d(x_b)) : h_d \in \mathcal{H}\right\}$. Because the Rademacher complexity of the convex hull of a set is equal to that of the set [35],

$$\mathscr{R}(\mathcal{F}_1) = \mathbb{E}_\sigma \left[\sup_{h_d \in \mathcal{H}} \sum_i \sigma_i \cdot u(x^+_{i,a}, h_d(x_{i,b}))\right]. \tag{1}$$

Recall that $h_d : \mathcal{X} \to [d]$. To apply the contraction lemma [35], we need u to be Lipschitz continuous in its second argument. For this, we extend u to a surrogate $\overline{u}$ as follows. For all $y \in [1, d]$,

$$\overline{u}(x^+_{i,a}, y) = u(x^+_{i,a}, \lfloor y \rfloor) + \left(u(x^+_{i,a}, \lceil y \rceil) - u(x^+_{i,a}, \lfloor y \rfloor)\right)(y - \lfloor y \rfloor). \tag{2}$$

Note that $\overline{u}$ is 1-Lipschitz in its second argument (because utilities sum to 1),[9] and matches u when $y \in [d]$. As such, we can replace u with $\overline{u}$ in Equation (1), and applying the contraction lemma [35], obtain: $\mathscr{R}(\mathcal{F}_1) \leq \mathbb{E}_\sigma \left[\sup_{h_d \in \mathcal{H}_d} \sum \sigma_i h_d(x_{i,b})\right] = \mathscr{R}(\mathcal{H}_b)$. □

With the help of these two lemmas, we now present the following generalization theorem:

Theorem 4. *Let $\mathcal{H}$ be a family of deterministic classifiers, S be a finite training set such that $\mathscr{R}(\mathcal{H}_b \circ S) \leq \epsilon/8$ for each $b \in \{1, 2\}$, and $\delta > 0$. If $|S| \geq 512 \ln(8|\mathcal{G}|/\delta)/\epsilon^2$, then with probability at least $1 - \delta$, we have $\sup_{h \in \Delta^k(\mathcal{H}), b_k \in \mathcal{B}} |V^S(h, b_k) - V(h, b_k)| \leq \epsilon$.*

Proof. By expanding V^S and V, we have

$$\begin{aligned}&\Pr\left[\sup_{h \in \Delta^k(\mathcal{H}), b_k \in \mathcal{B}} |V^S(h, b_k) - V(h, b_k)| \geq \epsilon\right] \\ &= \Pr\left[\sup_{\substack{h \in \Delta^k(\mathcal{H}), \\ b_k \in \mathcal{B}}} |U^S_{12}(h, b_k) - U^S_{11}(h, b_k) - U_{12}(h, b_k) + U_{11}(h, b_k)| \geq \epsilon\right] \\ &\leq \Pr\left[\sup_{h \in \Delta^k(\mathcal{H}), b_k \in \mathcal{B}} |U_{S\,12(h, b_k) - U_{12}(h, b_k)| \geq \epsilon/2}\right] \\ &\quad + \Pr\left[\sup_{h \in \Delta^k(\mathcal{H}), b_k \in \mathcal{B}} + |U_{11}(h, b_k) - U^S_{11}(h, b_k))| \geq \epsilon/2\right] \\ &\leq \sum_{b_k \in \mathcal{B}} \left(\Pr\left[\sup_{h \in \Delta^k(\mathcal{H})} |U_{12}(h, b_k) - U^*_{12}(h, b_k)| \geq \epsilon/2\right]\right. \\ &\quad \left. + \Pr\left[\sup_{h \in \Delta^k(\mathcal{H})} + |U^*_{11}(h, b_k) - U_{11}(h, b_k))| \geq \epsilon/2\right]\right),\end{aligned} \tag{3}$$

where the second transition follows from the triangle inequality, and the last transition follows from the union bound.

Let $\mathcal{F} = \left\{g : g(x_1^+, x_2^+) = u(x_1^+, h(x_2)) \cdot b_k(x_1^+, x_2^+), h \in \Delta^k(\mathcal{H})\right\}$. From Lemma 2, we have $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{H}_2)$. Using the standard generalization bound for Rademacher complexity [35],

$$\begin{aligned}&\Pr\left[\sup_{h \in \Delta^k(\mathcal{H})} |U^S_{12}(h, b_k) - U_{12}(h, b_k)| \geq \epsilon/2\right] \\ &\leq 4e^{-\frac{|S|}{2}\left(\frac{\epsilon - 4\mathscr{R}(\mathcal{F})}{8}\right)^2} \leq 4e^{-\frac{|S|}{2}\left(\frac{\epsilon - 4\mathscr{R}(\mathcal{H})}{8}\right)^2} \leq 4e^{-\frac{|S|}{2}\left(\frac{\epsilon}{16}\right)^2},\end{aligned}$$

[9] If we examined the proof of Rademacher calculus with Lipschitz continuous functions, we notice that it would only access the value of $\overline{u}$ at integral values of y. Hence, such an extension is technically not required. However, we construct it to use the Lipschitz continuity result as a black-box.

where the last inequality holds because we are given $\mathscr{R}(\mathcal{H}_2) \leq \epsilon/8$. The same argument applies to the second term in Equation (3). Hence, the probability in Equation (3) is at most $8|\mathcal{G}| \cdot e^{-|S|\epsilon^2/512}$. Setting this to δ and solving for $|S|$ completes the proof. □

Theorem 4 implies that regardless of the classifier trained, with high probability, the difference between group envy-freeness (or group equitability) violation between training and test will be small.

Finally, note that Theorem 4 requires $|S|$ to be large enough such that $\mathscr{R}(\mathcal{H}_b \circ S) \leq \epsilon/8$ for each $b \in \{1, 2\}$. Hence, for small $|S|$ to suffice, we also seek a family of deterministic classifiers $\mathcal{H}$ for which the Rademacher complexity quickly vanishes as $|S|$ grows. We show that the family of linear one-vs-all classifiers that we use in our experiments has this property.

Theorem 5. *Let $\mathcal{H}$ be the family of linear one-vs-all classifiers given by $\mathcal{H} = \{h_{\vec{w}} : h_{\vec{w}}(x) = \arg\max_{y\in[d]} w_y^T x\ ;\ w_y \in \mathbb{R}^m\}$. If $|S| \geq \frac{4096 d^3 m \ln(6dm/\epsilon)}{\epsilon^2}$, then $\mathscr{R}(\mathcal{H}_b \circ S) \leq \epsilon/8$ for each $b \in \{1, 2\}$.*

Proof. It is well-known that the Natarajan dimension of $\mathcal{H}$ is at most md [35]. Using the version of Sauer's lemma for Natarajan dimension [35], we get that the number of possible labelings of S is at most $|S|^{md} d^{2md}$. Fix $b \in \{1, 2\}$. Then, $||(h_b(z_1), \ldots, h_b(z_{|S|}))||_2 \leq d\sqrt{|S|}$. Hence, using Massart's lemma [35], we get $\mathscr{R}(\mathcal{H}_b) \leq \frac{2d}{|S|} \cdot \sqrt{|S| md \log(|S|d)}$. Using simple algebra, it can be checked that our lower bound on $|S|$ is sufficient to get $\mathscr{R}(\mathcal{H}_b) \leq \epsilon/8$. □

Combining Theorems 4 and 5, we get that when learning a mixture of linear one-vs-all classifiers, training set of size $|S| = O\left(\frac{\ln(|\mathcal{G}|/\delta)+d^3 m \ln(dm/\epsilon)}{\epsilon^2}\right)$ suffices to get generalization error bound of ϵ with respect to $\mathcal{G}$ with probability at least $1 - \delta$.

Note that our sample complexity scales logarithmically with $|\mathcal{G}|$, which allows achieving fairness with respect to exponentially many pairs of groups with only polynomial training sample size. This is reminiscent of similar results due to Hébert-Johnson et al. [23], Kearns et al. [26]. However, crucially, their approach requires the groups over which fairness is desired to be of small complexity or computable via small circuits. In contrast, our approach works with arbitrarily defined groups, and is therefore significantly stronger. Note that we do not place any assumptions on the groups (e.g. that they be defined based on certain sensitive attributes) or pairs of groups (e.g. that they be disjoint) involved in our constraints.

6 TRADEOFF BETWEEN LOSS AND FAIRNESS

How does ensuring fairness affect loss minimization? This is a key question as machine learning algorithms are usually designed to minimize risk and subjecting them to fairness may create a tension in that objective. Similarly, it is apt to ponder exactly how fair is a solution that simply minimizes risk. We now looks to answer these question analytically in the worst cases for the proposed notions of group envy free and group equitability. Without loss of generality, for this section we assume that loss must be bounded between 0 and 1. We also assume that the fairness constraints are enforced across all possible group pairs. As before, utilities are linear and must sum to 1. For simplicity, we assume in this section that there are g groups which are mutually exclusive ($G_i \cap G_j \,\forall\, i, j$) and $\mathcal{G}$ is the set of pairs of all possible groups ($|\mathcal{G}| = g(g-1)$). We show our bounds on the training set; generalization implies these bounds hold approximately with high probability on the population.

6.1 Unfairness of Risk Minimization

First we consider the standard expected risk minimization algorithm and consider the maximum possible group envy and group inequity possible. Let h_{ERM} denote the ERM classifier, h_{GEF} to the classifier that minimizes risk subject to group envy constraint and h_{GEQ} the classifier that minimizes risk subject to group equitability. Then, we consider the following: $\sum \widehat{G}_{envy}(i, j, h_{ERM}) - \sum \widehat{G}_{envy}(i, j, h_{GEF})$ and $\sum \widehat{G}_{ineq}(i, j, h_{ERM}) - \sum \widehat{G}_{ineq}(i, j, h_{GEQ})$. We briefly note that since our classifiers are probabilistic, both constraints can always be satisfied, meaning that $\sum \widehat{G}_{envy}(i, j, h_{GEF}) = 0$ and $\sum \widehat{G}_{ineq}(i, j, h_{GEQ}) = 0$. Thus, we look to analyze the unfairness of ERM in the worst case and give the following results.

Theorem 6. *The maximum group envy of an ERM classifier, $\sum \widehat{G}_{envy}(i, j, h_{ERM})$ is given by: $\Theta(g^2)$*

Theorem 7. *The maximum empirical group inequity of an ERM classifier, $\sum \widehat{G}_{ineq}(i, j, h_{ERM})$ is given by: $\Theta\left(\left\lceil \frac{g}{2} \right\rceil * \left\lfloor \frac{g}{2} \right\rfloor\right)$*

Proof. Note that group inequity is always between 0 and 1 and since it is symmetric, total group inequity is upper bounded by lemma **??**. We now show this is tight by constructing an instance that achieves this. First, number the groups $1, \ldots, k$ Now we define utilities. For every individual x in group i where $i \leq \left\lceil \frac{|\mathcal{G}|}{2} \right\rceil$, $u(x, 1) = 0$ and utility $\frac{1}{d-1}$ for the remaining classes. For every individual x in group j where $j > \left\lceil \frac{|\mathcal{G}|}{2} \right\rceil$, $u(x, 1) = 1$ and utility is 0 for the remaining classes. For all $x \in \mathcal{X}$, we define the loss function as: $\ell(x, 1) = 0$ and $\ell(x, y) = 1, \forall y \neq 1$. As such, the ERM classifier will assign class 1 to all individuals. Under this classifier, the average utility for groups $i < \left\lfloor \frac{g}{2} \right\rfloor$ is 0 and for groups $j > \left\lfloor \frac{g}{2} \right\rfloor$ is 1. Then by application of Lemma **??**, we have the bound desired. □

[SH: Write some discussion here]

6.2 Inefficiency of Fair Classifiers

We now turn to the second related question: what is the maximum possible difference in efficiency between an ERM classifier, and a classifier that minimizes risk subject to fairness? Using the same notation from the preceding section, we are interested in the following quantities: $\sum L(x, \widehat{y}, h_{GEF}) - \sum L(x, \widehat{y}, h_{ERM})$ and $\sum L(x, \widehat{y}, h_{GEQ}) - \sum L(x, \widehat{y}, h_{ERM})$. We give the following results:

Theorem 8. *The maximum efficiency difference between an ERM classifier subject to group envy free, h_{GEF} and the ERM classifier, h_{ERM}, is $\Omega\left(\frac{n(g-1)}{g}\right)$*

Proof. Consider an instance when the number of groups is equal to the number of classes $g = d$ and the sample contains an equal number of individuals in each group n/g. Numbering the groups $1, \ldots, g$, let $S_1, \ldots, S_g$ denote the samples belonging to each group. The utility for an individual $x^+ \in S_i$ is: $u(x^+, i) = 0$ and $u(x^+, c) = \frac{1}{d-1}$ where $\forall\, c \neq i$. Similarly, the loss for an individual $x^+ \in S_i$ is: $\ell(x^+, i) = 0$ and $\ell(x^+, c) = 1, \forall\, c \neq i$. Thus h_{ERM} assigns every $x \in S_i$, class i and attains 0 loss. However, it is not group envy free. We now show that h_{GEF} assigns: $\forall x \in S_i, \mathcal{P}(y_x = i) = \frac{1}{d}$. This is clearly envy free and achieves loss $\frac{n(d-1)}{d}$. To show this achieves the lowest possible subject to group envy free, we introduce $P(S_i, h, c) = \frac{1}{|S_i|}\sum_{x^+ \in S_i} \mathcal{P}(y_x = c)$ denoting the average probability individuals in group i have of achieving class c under classifier h. Now by contradiction, assume there is a classifier h' that is group envy free and achieves strictly lower loss. This implies that $\sum_{i=1}^{g} P(S_i, h', i) > 1$. Also note that for each group i to be group envy free under our utilities: $P(G_j, h', i) \geq P(G_i, h', i), \forall j \neq i$. Combining these two, we have that for group 1: $\sum_{c=1}^{g} P(S_1, h', c) \geq 1$ which is a contradiction. □

Theorem 9. *The maximum efficiency difference between an ERM classifier subject to group equitability, h_{GEQ} and the ERM classifier, h_{ERM}, is $\Omega\left(\lfloor \frac{g}{2} \rfloor * \frac{n}{g}\right)$.*

Proof. Let the sample contain an equal number of individuals in each class n/g. Numbering the groups $1, \dots, g$, let $S_1, \dots, S_g$ denote the samples belonging to each group. $\forall x^+ \in \mathcal{X}^+$, let $u(x^+, 1) = 1$ and $u(x^+, c) = 0, \forall c \neq 1$. Let individuals x^+ in an odd group ($i\%2 = 1$) have loss: $\ell(x^+, 1) = 0$ and $\ell(x^+, c) = 1, \forall c \neq 1$. Similarly, let individuals x^+ in an even group ($i\%2 = 0$) have loss: $\ell(x^+, 2) = 0$ and $\ell(x^+, c) = 1, \forall c \neq 2$. By assigning members in odd groups class 1 and even goups class 2, h_{ERM} achieves 0 loss. For h_{GEQ}, we require that for any two groups i and j, $\frac{1}{|S_i|}\sum_{x^+ \in S_i} u(x^+, h(x)) = \frac{1}{|S_j|}\sum_{x^+ \in S_j} u(x^+, h(x)) = A$, where $A \in [0, 1]$. Thus, for an odd group, the loss is $\frac{n}{g}(1 - A)$ and for an even group, the loss is $\frac{n}{g}A$. Thus, the total loss is $\Omega\left(\lfloor \frac{g}{2} \rfloor * \frac{n}{g}\right)$ □

[SH: Write some discussion here on results]

7 IMPLEMENTATION AND EXPERIMENTS

In this section, we propose a method for training (almost) group envy-free and (almost) group equitable classifiers, and use it to empirically evaluate the tradeoff between our fairness desiderata and loss minimization. Our approach follows the convex relaxation approach proposed by Balcan et al. [5] for building (almost) envy-free classifiers. We emphasize that we do not view this approach as the end-all solution; rather, it simply illustrates feasibility of training good classifiers subject to our fairness guarantees.

Formally, our goal is to learn a mixture $\sum_{t=1}^{k} \eta^t h_d^t \in \Delta^k(\mathcal{H})$ which minimizes the empirical risk subject to group envy-freeness or group equitability constraints. Let $h = (h_d^1, \dots, h_d^k) \in \mathcal{H}^k$ and $\eta = (\eta^1, \dots, \eta^k) \in \Delta^k$, where Δ^k denotes the k-simplex containing probability distributions over k elements. Section 5 suggests when using the following family $\mathcal{H}$ of linear one-vs-all multiclass classifiers, we obtain good generalization due to its low Natarajan dimension: $\mathcal{H} = \left\{g_{\vec{w}} : g_{\vec{w}}(x) = \arg\max_{y \in [d]} w_y^T x \;;\; w_y \in \mathbb{R}^m\right\}$.

Given this family, a finite training set S of individuals, and a finite set of pairs of groups $\mathcal{G}$, our learning problem is the following. We only add one of the two sets of constraints, depending on whether we require group envy-freeness or group equitability.

$$\min_{\substack{h \in \mathcal{H}^k \\ \eta \in \Delta^k}} \sum_{t=1}^{k} \eta^t \sum_{x^+ \in S} \ell(x^+, h_d^t(x)) \text{ such that } \forall (G, \widehat{G}) \in \mathcal{G}$$

$$\frac{1}{|S^G||S^{\widehat{G}}|}\sum_{t=1}^{k} \eta_t \sum_{x^+ \in S^G, \widehat{x}^+ \in S^{\widehat{G}}} u(x^+, h_d^t(\widehat{x})) - u(x^+, h_d^t(x)) \leq 0$$
$$\text{// for group envy-freeness, or} \tag{4}$$
$$\frac{1}{|S^G|}\sum_{t=1}^{k} \eta_t \sum_{x^+ \in S^G} u(x^+, h_d^t(x)) = \frac{1}{|S^{\widehat{G}}|}\sum_{t=1}^{k} \eta_t \sum_{\widehat{x}^+ \in S^{\widehat{G}}} u(\widehat{x}^+, h_d^t(\widehat{x}))$$
$$\text{// for group equitability.}$$

Convex relaxation of loss and utilities: Note that in Equation (4), $\ell(x^+, h_d^t(x))$ and $u(x^+, h_d^t(x))$ are neither convex nor differentiable due to the use of arg max in h_d^t. As such, we consider the following multiclass-SVM-inspired convex relaxation, similarly to Balcan et al. [5]. Note for any $c \in [d]$, $\ell(x^+, h_d^t(x)) \leq \ell(x^+, h_d^t(x)) + w_{h_d^t(x)}^T x - w_c^T x$. Thus, we get $\ell(x^+, h_d^t(x)) \leq \max_{y \in \mathcal{Y}} \ell(x^+, y) + w_y^T x - w_c^T x$, which is a convex upper bound on $\ell(x^+, h_d^t(x))$. We use a similar convex upper bound on $u(x^+, h_d^t(x))$.

Training: Our problem is still not entirely convex due to the product of mixture probabilities with loss and utilities. To circumvent this, we use an iterative approach [5] whereby we first fix a default η and successively train for each h_d^t given previously learned $h_d^1, \dots, h_d^{t-1}$, and after learning full h, fix it and train for η. Second, for tractability, instead of enforcing group envy-freeness or group equitability as hard constraints, we add a penalty term encoding these violation and control its effect with a Lagrangian parameter λ. Let $T_{envy}(S^G, S^{\widehat{G}})$ denote the maximum[10] of 0 and LHS of the group envy-freeness constraint in Equation (4), and $T_{eq}(S^G, S^{\widehat{G}})$ denote the absolute difference between LHS and RHS in the group equitability constraint in Equation (4). Then, we add $\lambda \cdot T_{envy}(S^G, S^{\widehat{G}})$ penalty term for group envy-freeness, and $\lambda \cdot T_{eq}(S^G, S^{\widehat{G}})$ penalty term for group equitability. This converts our problem from constrained to unconstrained optimization, which is easier to solve.

7.1 Experiment Design

We consider the targeted advertising domain, where the classes represent different ads, $\ell(x^+, y)$ represents the loss to the principal for ad y being shown to individual x^+, and $u(x^+, y)$ represents the utility to the individual for setting ad y.

Our simulation setup is similar to that of Balcan et al. [5]. However, they are interested in measuring how accurate their convex relaxation approach is, and therefore generate instances where ERM is guaranteed to be fair. We are instead interested in measuring how our method compares to ERM, and therefore generate random instances where ERM is no longer magically guaranteed to be fair.

We set the number of classes to be $d = 5$. We create a finite training set S by sampling n iid feature vectors uniformly from $[0, 1]^m$, where $m = 14$. Let $X_{train} \in [0, 1]^{n \times m}$ denote the matrix of all feature vectors. We partition the individuals into g equal-sized groups (default is $g = 4$) based on the first feature coordinate, and add group envy-freeness or group equitability constraints for all pairs of groups. We implement loss and utilities as functions of the features as follows.[11] First, we sample matrices L and U uniformly

[10] We do this because strictly negative envy is not specifically desired.

[11] For generalization, it is necessary for loss and utilities to be correlated with features.

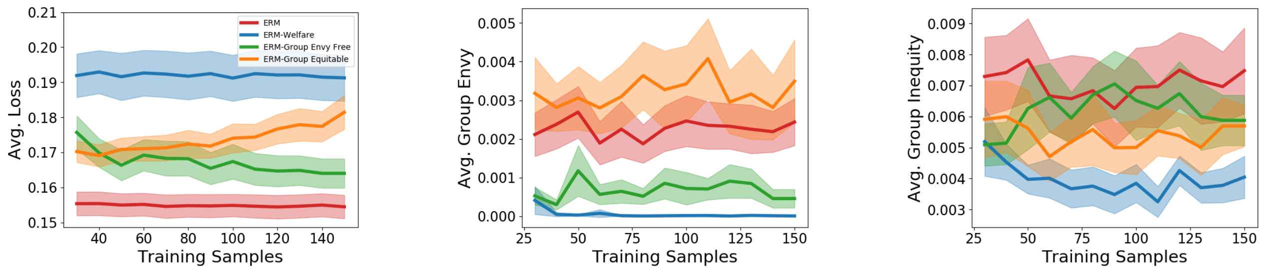

Figure 1: Comparing ERM (red) and ERM with welfare penalty (blue), group envy-freeness penalty (green), and group equitability penalty (orange) with varying training sample size and 4 groups. 90% confidence intervals are shown.

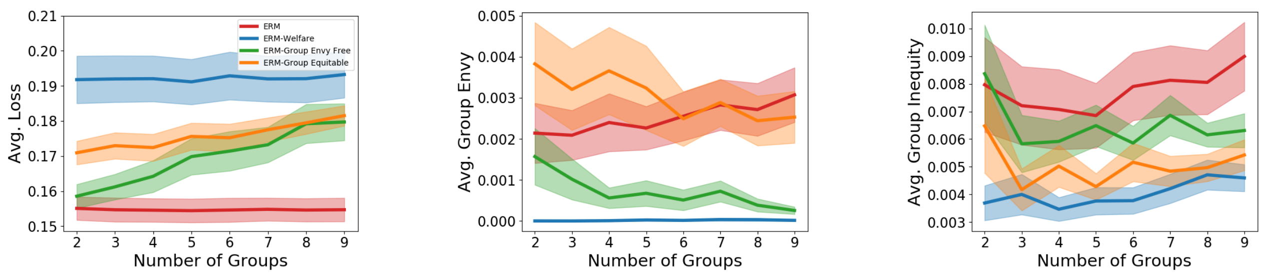

Figure 2: Comparing the same four approaches with varying number of groups and training sample size of 100. 90% confidence intervals are shown.

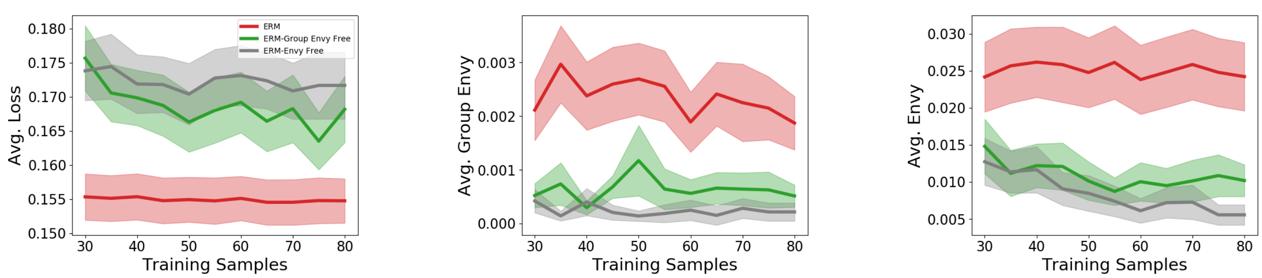

Figure 3: Comparing ERM (red), ERM with group envy-freeness penalty (green), and ERM with envy-freeness penalty (grey) with varying training sample size and 4 groups. 90% confidence intervals are shown.

from $[0,1]^{d\times m}$. Then, for each $x^+ \in S$, we set $\ell(x^+,[d]) = Lx^+$ and $u(x^+,[d]) = Ux^+$. Finally, we normalize both $u(x^+,\cdot)$ and $\ell(x^+,\cdot)$ to sum to 1 for each $x^+ \in S$. Our training set size varies, but in all our figures, we show the performance on a test set of size 100, which we generate using the same process.[12] A simulation consists of random sampling of a training set X_{train}, a test set X_{test}, and matrices L and U. Each data point plotted is the average over 40 simulations, and 90% confidence intervals are shown. We solved our unconstrained convex optimization problems using CVXPY [16] on workstations with Intel(R) Xeon(R) W-2133 CPU @ 3.60GHz and 16GB RAM, and parallelized computation across 16 such cores.

We train five classifiers: ERM (which simply minimizes the empirical loss), ERM-Welfare (which minimizes empirical loss minus λ times welfare), ERM-GroupEF and ERM-GroupEQ (unconstrained convex relaxations of optimization problem (4) with group envy-freeness and group equitability constraints), and ERM-EF (a similar method by Balcan et al. [5] for envy-freeness). In all methods, we set the Lagrangian parameter to $\lambda = 10$. We omit ERM-EF solution from the bulk of our experiments because it does not scale well with the training sample size. We instead show a separate comparison with this solution on smaller sample sizes.

7.2 Results of Experiments

We consider three key metrics: loss per individual, total group envy-freeness violation across all pairs of groups, and total group equitability violation across all pairs of groups.

Figure 1 shows the performance of ERM, ERM-Welfare, ERM-Group Envy Free, and ERM-Group Equitable with varying number of training samples. The number of groups is fixed to be 4. As expected, ERM attains the lowest loss, but at the cost of significant violation of group envy-freeness and group equitability. ERM-Welfare

[12] We confirm that in our experiments, all classifiers used have similar loss, group envy-freeness violation, and group equitability violation on training and test sets. We only show the test results for readability.

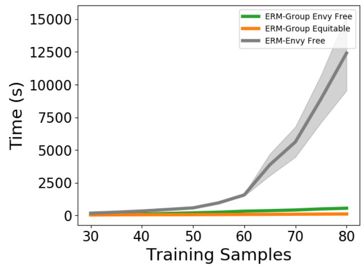

Figure 4: Running time of ERM with three different penalties (group envy-freeness, group equitability, and envy-freeness) with varying training sample size and 4 **groups.** 90% **confidence intervals are shown.**

expectedly performs very well in terms of the fairness metrics,[13], but at a significant cost in terms of the loss. ERM-GroupEF and ERM-GroupEQ provide reasonable tradeoffs, with the former achieving at least ~ 50% reduction in group envy-freeness violation compared to ERM and the latter achieving at least ~ 30% reduction in group equitability violation compared to ERM. Interestingly, ERM-GroupEF clearly outperforms ERM-GroupEQ for group envy-freeness, but the converse effect is less strong in group equitability.

We see a similar story in Figure 2, when we fix the training sample size to 100 but vary the number of groups. This time, ERM-GroupEF and ERM-GroupEQ, respectively, achieve at least ~ 50% and ~ 40% reduction in group envy-freeness and group equitability violations compared to ERM.

Finally, in Figure 3, we compare ERM-EF to ERM-GroupEF. Since individual envy-freeness is a stricter constraint than group envy-freeness, as expected, ERM-EF performs better in terms of the fairness metrics (both group envy-freeness and individual envy-freeness violations), whereas ERM-GroupEF performs better in terms of the loss. However, the difference is small, and both solutions perform very similarly. The biggest drawback of ERM-EF is the running time. As we see in Figure 4 computing ERM-EF takes significantly longer than computing ERM-GroupEF or ERM-GroupEQ. Indeed, this is why we did not include ERM-EF in Figure 1, where the training sample size was larger than what ERM-EF could handle.

8 DISCUSSION

In this paper, we explored the applicability of two prominent fairness notions from the economic literature on fair division — envy-freeness and equitability — in machine learning. We proposed novel relaxations of these definitions in a group setting, unifying several previously proposed ones under a single framework and extending them beyond the binary classification setting.

Group envy-freeness, in particular, allows placing asymmetric constraints.[14] This is a feature of very few fairness definitions in the literature, but one that could be useful in certain applications. For example, equalized odds demands $\Pr[Y = 1 | X \in G, Y^* = a] \geq \Pr[Y = 1 | X \in \widehat{G}, Y^* = a]$ for all $a \in \{0, 1\}$, capturing the intuition that individuals in groups G and $\widehat{G}$ with the same ground truth label deserve equal treatment. However, individuals with ground truth label 1 (e.g. individuals likely not to re-offend or likely to repay a loan) may also deserve treatment that is no worse than that given to individuals with ground truth label 0. Though it may emerge naturally from loss minimization, it can be imposed explicitly through group envy-freeness for appropriately defined pairs of groups.

Our approach leaves the choice of $\mathcal{G}$, the set of pairs of groups across which fairness is desired, to the designer. This allows application-dependent definitions of protected groups, but also raises an interesting challenge. Consider a multi-class version of the loan setting, in which there are d different types of loans (thus, $d + 1$ possible outcomes including "no loan"). In this case, it makes sense for the ground truth label to also be a vector $Y^* \in \{0, 1\}^d$, where Y_r^* denotes the individual's ability to repay a loan of type r, if it were given to the individual. How should one subdivide protected groups based on vector-valued ground truth labels?

In our view, this work only scratches the surface of exploring how the abundant economic literature on fairness can be applied to machine learning, and despite significant recent progress in this direction [24, 31], there is much left to explore and future work in this direction can discover novel challenges for the machine learning community to address.

[13]This is not surprising as it is a relaxation of minimizing loss subject to maximum welfare. In the targeted advertising context, maximum welfare is attained when each individual is shown their most preferred ad, which clearly performs well in terms of fairness.

[14]One can similarly define an asymmetric variant of group equitability, where the equality constraint is replaced by an inequality.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.

[2] Martin Aleksandrov and Toby Walsh. 2018. Group envy freeness and group pareto efficiency in fair division with indivisible items. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 57–72.

[3] Micah Altman, Alexandra Wood, and Effy Vayena. 2018. A harm-reduction framework for algorithmic fairness. *IEEE Security & Privacy* 16, 3 (2018), 34–45.

[4] Maria Florina Balcan, Florin Constantin, Satoru Iwata, and Lei Wang. 2012. Learning valuation functions. In *Conference on Learning Theory*. 4–1.

[5] Maria-Florina Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia. 2019. Envy-Free Classification. In *Advances in neural information processing systems*. Forthcoming.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] J. Bartholdi, C. A. Tovey, and M. A. Trick. 1992. How hard Is It to Control an Election. *Mathematical and Computer Modelling* 16 (1992), 27–40.

[8] Omer Ben-Porat, Fedor Sandomirskiy, and Moshe Tennenholtz. 2019. Protecting the Protected Group: Circumventing Harmful Fairness. *arXiv preprint arXiv:1905.10546* (2019).

[9] E. Budish. 2011. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy* 119, 6 (2011), 1061–1103.

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

[11] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D Procaccia, Nisarg Shah, and Junxing Wang. 2019. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)* 7, 3 (2019), 12.

[12] Urszula Chajewska, Daphne Koller, and Dirk Ormoneit. 2001. Learning an agent's utility function by observing behavior. In *ICML*. 35–42.

[13] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally Fair Clustering. In *International Conference on Machine Learning*. 1032–1041.

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

[15] Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer Wortman Vaughan. 2019. Group fairness for the allocation of indivisible goods. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*. 1853–1860.

[16] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[18] Brandon Fain, Ashish Goel, and Kamesh Munagala. 2016. The core of the participatory budgeting problem. In *International Conference on Web and Internet Economics*. Springer, 384–399.

[19] D. Foley. 1967. Resource allocation and the public sector. *Yale Economics Essays* 7 (1967), 45–98.

[20] D. Gale and L. S. Shapley. 1962. College Admissions and the Stability of Marriage. *Americal Mathematical Monthly* 69, 1 (1962), 9–15.

[21] Paul Gölz, Anson Kahng, and Ariel D Procaccia. 2019. Paradoxes in Fair Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Forthcoming.

[22] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[23] Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. 1944–1953.

[24] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 181–190.

[25] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 600.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*. 2569–2577.

[27] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*. 43:1–43:23.

[28] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[29] R. J. Lipton, E. Markakis, E. Mossel, and A. Saberi. 2004. On approximately fair allocations of indivisible goods. In *Proceedings of the 6th ACM Conference on Economics and Computation (EC)*. 125–131.

[30] H. Moulin. 2004. *Fair Division and Collective Welfare*. MIT Press.

[31] Sendhil Mullainathan. 2018. Algorithmic Fairness and the Social Welfare Function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 1–1.

[32] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[33] Govind Ramnarayan. 2018. Equalizing Financial Impact in Supervised Learning. arXiv:1806.09211.

[34] J. Rawls. 1971. *A Theory of Justice*. Harvard University Press.

[35] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

[36] Lloyd Shapley and Herbert Scarf. 1974. On cores and indivisibility. *Journal of mathematical economics* 1, 1 (1974), 23–37.

[37] H. Steinhaus. 1948. The problem of fair division. *Econometrica* 16 (1948), 101–104.

[38] W. Stromquist. 1980. How to cut a cake fairly. *Amer. Math. Monthly* 87, 8 (1980), 640–644.

[39] F. E. Su. 1999. Rental harmony: Sperner's lemma in fair division. *Amer. Math. Monthly* 106, 10 (1999), 930–942.

[40] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. 6373–6382.

[41] H. Varian. 1974. Equity, envy and efficiency. *Journal of Economic Theory* 9 (1974), 63–91.

[42] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.

[43] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*. 229–239.