Economic Notions of Fairness for Machine Learning

Anonymous Author(s)
Submission Id: XXX-XXX-XXX

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. We explore the applicability of two such notions — envy-freeness and equitability — in machine learning contexts. We propose their relaxations that are compelling in a broad range of settings, and provide a unifying framework by incorporating several recently proposed fairness definitions as special cases. We provide generalization error bounds for our fairness definitions, and theoretically and experimentally evaluate the trade-offs between loss minimization and fairness guarantees.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

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, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Today, machine learning algorithms are 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). Growing with their use is the concern 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 [7, 10, 12, 16, 19, 20].

While this exercise has yielded many insights into the ways in which algorithms can be made fairer, it has also yielded a vast number of fairness definitions [21], many of which are incompatible [10, 19]. There is a general lack of consensus as to which is the *right* definition of fairness, and this choice is often application-dependent [18]. Further, most popular definitions such as statistical parity [7, 12] and equalized odds [16] 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 thinking about fairness across a broad range of settings [17].

While exploration of fairness in machine learning literature is a recent phenomenon, fairness has been the subject of formal study in microeconomics (especially in fair resource allocation) since almost a century [26]. While the literature was initiated with a study of the canonical setting of *cake-cutting*, it has focused on proposing general-purpose definitions such as proportionality [26], envy-freeness [14], equitability, the core [30], and Rawlsian egalitarian fairness [23] which apply to a broad range of settings.

For example, the core is not only applicable in cake-cutting [30], but also in participatory budgeting [13], housing markets [25], matching markets [15], public goods allocation [13], and even clustering [9].

Recently, there has been a stream of papers using these definitions to design fair machine learning algorithms [5, 29, 31]. One central notion of fairness adopted by all of these papers 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 a compelling definition because it is simple and intuitive, and does not require access to any information other than the utility functions of the individuals, which can be learned easily from their actions [4, 8]; this is in contrast to definitions like *individual fairness* [12], 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 [27] or rent division [28],[1] it is too stringent for many machine learning settings. For example, in *binary* settings, where there are only two outcomes and all individuals prefer the same outcome (e.g. prefer having loan applications approved than rejected, or prefer receiving bail than not), envy-freeness would require that all individuals have the same outcome. In applications like targeted advertising, where people have heterogeneous preferences, envy-freeness is less restrictive, but only when randomized ad 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. The reason this is different from envy-freeness is highlighted in the following example. Suppose there are two outcomes A and B, and two individuals 1 and 2 have the following utility functions: $u_1(A) = 1$, $u_1(B) = 0$, $u_2(A) = 0$, $u_2(B) = 1$. If both individuals are assigned outcome A, this is clearly envy-free. However, individual 1 obtained her preferred outcome for which she has utility 1, whereas individual 2 obtained her less preferred outcome for which she has

[1] Indeed, in resource allocation, it is usually therefore *strengthened* into group envy-freeness or group fairness [11].

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

utility 0. This intuitive unfairness is captured by the violation of equitability.

The research question we address is the following:

Are there relaxations of envy-freeness and equitability which are more appropriate for machine learning settings?

1.1 Our Contributions

[NS: We propose new relaxations along the line of group fairness definitions in machine learning which use average across a group. Give definitions.]

[NS: We show that these relaxations are broadly applicable in many settings from GT settings to non-GT ad setting. Generalize many previous defs as special cases. Thus, they provide a unifying framework.]

[NS: Also give ways to directly generalize those notions to multiclass.]

[NS: Allow auditing an algorithm for fairness by simply observing its output. Does not require knowledge of the algorithm itself, unlike in some other notions of fairness [29].]

[NS: Generalization results]

[NS: Experimental results show that fairness can be achieved with minimal increase in the loss function.]

1.2 Related Works

[NS: Fair division. Utility subjective. Usually not averaged for avoiding interpersonal comparisons (cite Moulin book). Less of a problem in ML. Typically utility measured in financial impact (cite equalized financial impact paper) or probability of the preferred outcome (e.g. in classification).]

[NS: Previous work importing EF (efClass, fairnessNoHarm, parityToPreference). But EF is too stringent, as described by efClass paper. Average definition conceptually simple relaxation but new to the best of our knowledge. (Didn't we find some recent papers where this was done?) Also, work on leveraging other ideas from fair division literature such as the axiomatic requirements of monotonicity to machine learning (cite Ariel's paper).]

[NS: Generalization: Uses Rademacher complexity approach instead of the pseudo-dimension approach of efClass. Able to provide fairness across exponentially many groups with only polynomial sample complexity requirement. Nicely works with Lipschitz utility functions. Providing fairness for exponentially many groups is similar to fairnessGerrymandering and multiCalibration papers.]

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 the training dataset for the purpose of evaluating classifiers. 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 a ground truth (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 allow 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.

[SH: Can we use subscript k.] [NS: Why? I used superscript because in h_d^k, k comes in superscript.] 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 to define $\ell(x^+, \mathcal{P}^{\mathcal{Y}}) = \mathbb{E}_{y \sim \mathcal{P}^{\mathcal{Y}}}[\ell(x^+, y)]$.

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 $x^+ \in \mathcal{X}^+$, $\sum_{y \in \mathcal{Y}} u(x^+, y) = 1$. [NS: Do we only really need this to be ≤ 1? If so, then this is WLOG.] [SH: yeah it's just easier if this always sums to 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). [NS: Add some discussion on when this could be useful.] That said, we assume that utility function of an individual only depends on features captured in x^+. In practice, two individuals with identical x^+ may have slightly different utilities, but our results hold approximately if one can find close approximations of individuals' utility functions which only depend 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.

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

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 the groups to be defined arbitrarily by the principal. 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 family 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 $\forall x^+, \widehat{x}^+ \in \mathcal{X}^+ : u(x^+, h(x)) \geq u(x^+, h(\widehat{x}))$. 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]

[SH: To Nisarg: I also wonder if it is better to define group envy instead or ϵ-group envy free. Can decide this after we go everything.] [NS: I thought about it. But taking sum across $(G, \widehat{G})$ and using max$(\cdot, 0)$ may seem weird. Best to keep it in experiments, and introduce our notion in a way that resembles other notions in the literature — a hard constraint.] [SH: I agree - see below and let me know if you like that]

Another way of viewing the envy-freeness is that the envy of any individual for any other individual is non-negative: $u(x^+, h(\widehat{x})) - u(x^+, h(x)) \leq 0, \forall x^+, \widehat{x}^+ \in \mathcal{X}^+$. We propose a group-level relaxation of this constraint, following a similar relaxation proposed by Aleksandrov and Walsh [2], in which the constraint is only required when averaged across individuals from a pair of 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-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. [SH: We denote this difference to be the the empirical group envy between $\mathcal{S}^G$ and $\mathcal{S}$ and] 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 see in Section 4, in certain cases it may be desirable to require group envy-freeness in only one direction.

[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].

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 family 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. 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 is also a stringent requirement in machine learning contexts similarly to envy-freeness. We relax this to a groupwise fairness notion as follows.

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

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

[SH: We denote this difference as the empirical group inequity between S^G and $S^{\widehat{G}}$ and] When $\epsilon = 0$, we simply refer to this as empirical group-equitability. [SH: Maybe mention that Group Equity is symmetric, unlike envy free]

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 family 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 $(G, \widehat{G}) \in \mathcal{G}$.

4 PARALLELS TO OTHER FAIRNESS NOTIONS

[NS: TAKING THE LOCK ON THIS SECTION]

[NS: First a general note about ground truth vs non ground truth setting, and our notion working across both.]

4.1 Statistical Parity, Equal Opportunity, and Equalized Odds

We now show that group envy-freeness and group equitability capture three popular group fairness notions proposed in the literature — statistical parity, equal opportunity, and equalized odds — as special cases, thus both providing a unifying framework to view them under a single umbrella and generalizing them beyond the binary classification setting.

Let us 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 family $\mathcal{G}$ of pairs of groups and utility functions of individuals such that group envy-freeness and group equitability with respect to $\mathcal{G}$ coincide with 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 class. In our notation, $u(x^+, 1) = 1$ 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 $\mathbb{E}[Y = 1|X \in \widehat{G}] = \mathbb{E}[Y = 1|X \in G]$, which is the condition for the classifier to satisfy statistical parity with respect to $(G, \widehat{G})$. Hence, the choice of $\mathcal{G} = \left\{(G, \widehat{G}), (\widehat{G}, G)\right\}$ suffices.

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

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

While we provide a proof for equivalence of the population versions of the group fairness notions, it is easy to see that the equivalence also holds for the empirical versions defined on a training set. Similarly, note that our definitions can also easily 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 our 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 reductions approach and generalization proof of Agarwal et al. [1]; however, their framework is still 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 [22] 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 in spirit 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 [22] 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 $[\Pr[Y = 0] \cdot \psi(b(X))|X \in G, Y^* = 1] = [\Pr[Y = 0] \cdot \psi(b(X))|X \in \widehat{G}, Y^* = 1]$.

Similarly to Theorem 1, it is easy to see that this is a special case of group envy-freeness and group equitability where

we are given a function ψ which takes as input a subset of features representing the wealth of the individual. Let $b(x^+)$ For an individual represented by x^+, let $b(x^+)$ denote the wealth.

4.3 Impossibility of Post-Processing

[NS: After the result, add an interpretation. (a) Cite the paper which shows that anyway post-processing is not sample-efficient. (b) What this shows is that unless you are in the edge case where all utilities are identical, we really cannot rely on post-processing. In that sense, it just shows that's possible, and is not against the definition.]

Secondly, note that **?**] showed that starting from a classifier that simply minimized loss, it is possible to attain a derived classifier that is optimal subject to equalized odds through only post processing. In other words, an algorithm with access to only group identity can convert an ERM classifer to the ERM classifer subject to equalized odds. Thought desirable, this is clearly difficult in our general setting as agent utilities are heterogeneous and fundamental to the fairness definition. Nonetheless, it is a pertinent question to ask if a group equitable or group envy free solution can always be constructed without access to utilities. The theorems below show that there

is and outline their sufficient and necessary conditions. Note for simplicity, that these results assume group identity is mutually exclusive ($G_i \cap G_j \; \forall \, i, j$) and $\mathcal{G}$ is the set of pair of all possible groups.

Theorem 2. *A randomized classifier $h \in \mathcal{H}_k$ is trained without access to utilities. Then h is always empirically group envy free if and only if $\forall x_1, x_2 \in \mathcal{X}$, $h(x_1) = h(x_2)$.*

Proof. If h assigns all individuals the same classification, it is envy-free and also group envy-free. We now show that without access to utilities, this is the only classifier that guarantees group envy-free. Consider two groups, G and $\widehat{G}$ and recall that $h_p(x)$ returns a vector of probabilities. Let P denote a vector of the average classification probabilities given to individuals in group S^G. More specifically: $P = \frac{1}{|S^G|} \sum_{x_i^+ \in S^G} h(x_i)$ and the c^{th} element $P_c = \frac{1}{|S^G|} \sum_{x_i^+ \in S^G} \mathcal{P}(y_{x_i} = c)$ where $c \in [d]$. Now for any $\widehat{x}_j^+ \in S^{\widehat{G}}$ such that $h(\widehat{x}_j^+) \neq P$, we can always find a class $\widehat{t}_j$ such that $\mathcal{P}(y_{\widehat{x}_j^+} = \widehat{t}_j) < P_{\widehat{t}_j}$. For such individuals, define their utility function as: $\forall \, \widehat{x}_j^+ \in S^{\widehat{G}} \mid h(\widehat{x}_j^+) \neq P$, $u(\widehat{x}_j^+, \widehat{t}_j) = 1$ and 0 for the other classes. For the remaining individuals in $S^{\widehat{G}}$ let them have utility 1 for class 1 and 0 for the remaining classes. Then, $\widehat{G}$ is group envy free of G if:

$$\frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ | h(\widehat{x}_j^+) \neq P} \mathcal{P}(y_{\widehat{x}_j^+} = \widehat{t}_j) + \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ | h(\widehat{x}_j^+) = P} \mathcal{P}(y_{\widehat{x}_j^+} = 1) \geq \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ | h(\widehat{x}_j^+) \neq P} P_{\widehat{t}_j} + \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ | h(\widehat{x}_j^+) = P} P_1$$

where we note that the second term in both sides of the inequality are equal since $h(\widehat{x}_j^+) = P$. This leads to a contradiction since we started by picking each $\widehat{t}_j$ such that $\mathcal{P}(y_{\widehat{x}_j^+} = \widehat{t}_j) < P_{\widehat{t}_j}$. This implies that $\forall \widehat{x}_j^+ \in S^{\widehat{G}}$, $h(\widehat{x}_j^+) = P$. Now we can use the same argument in reverse and illustrate that $\forall x_i^+ \in S^G$, $h_p(x_i^+) = \widehat{P}$. Now we have that: $\widehat{P} = \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ \in S^{\widehat{G}}} h(\widehat{x}_j^+) = \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}_j^+ \in S^{\widehat{G}}} P = P$ Therefore, $\forall x_i^+ \in S^G$, $\forall \widehat{x}_j^+ \in S^{\widehat{G}}$, $h(x_i^+) = h(\widehat{x}_j^+) = P = \widehat{P}$. Applying this to all possible group pairs concludes the proof. □

Theorem 3. *A randomized classifier $h \in \mathcal{H}_k$ is trained without access to utilities. Then h is always empirically group equitable if and only if $\forall \, x$, $h(x) = [\frac{1}{d}, \ldots, \frac{1}{d}]$.*

Proof. If h_p meets the conditions, then the average utility for S^G is: $\frac{1}{|S^G|} \sum_{x \in S^G} \frac{1}{d}(u(x,1) + \ldots, +u(x,d))$ where we note that each individual utility must sum to 1. Thus, each group has average utility $\frac{1}{d}$ and this classifier is empirically equitable. Now consider two group G and $\widehat{G}$. $\forall x_i^+ \in S^G$, let $t_i^{min} = \arg\min_c \mathcal{P}(y_{x_i} = c)$. Similarly, for each $\widehat{x}_j^+ \in \widehat{G}$, let $\widehat{t}_j^{max} = \arg\max_c \mathcal{P}(y_{\widehat{x}_j} = c)$. Note that for each x_i, $\mathcal{P}(y_{x_i} = t_i^{min}) \leq \frac{1}{d}$ and for each $\widehat{x}_j$, $\mathcal{P}(y_{\widehat{x}_j} = \widehat{t}_j^{max}) \geq \frac{1}{d}$. Now consider the following utilities: $\forall x_i^+ \in S^G : u(x_i^+, t_i^{min}) = 1$ and 0 for the other classes and $\forall \widehat{x}_j^+ \in \widehat{G}$, $u(\widehat{x}_j^+, \widehat{t}_j^{max}) = 1$ and 0 for the other classes. Under these utilities, group G and $\widehat{G}$ being equitable means:

$$\frac{1}{|S^G|} \sum_{x_i^+ \in S^G} \mathcal{P}(y_{x_i} = t_i^{min}) = \frac{1}{|S^{\widehat{G}}|} \sum_{\widehat{x}^+ \in \widehat{G}} \mathcal{P}(y_{\widehat{x}_j} = \widehat{t}_j^{max})$$

which implies that both LHS and RHS must be equal to $\frac{1}{d}$. Due to property of t_i^{min} and $\widehat{t}_j^{max}$, it must be that $\forall x_i^+ \in S^G$, $\mathcal{P}(y_{x_i} = t_i^{min}) = \frac{1}{d}$, and $\forall \widehat{x}_j^+ \in \widehat{G}$, $\mathcal{P}(y_{\widehat{x}_j} = \widehat{t}_j^{max}) = \frac{1}{d}$. This implies that $\forall x_i^+, \forall \, c \, \mathcal{P}(y_{x_i} = c) = \frac{1}{d}$. Same individuals in group $\widehat{G}$. Applying this result to all group pairs, we have the result to be proven. □

We note that the classifier that is always group equitable is unique, whereas for group envy-free it is much less restrictive. We conclude by noting that the fairness concepts shown to be subsumed by our proposed notions is not an exhaustive list and expect others.

5 GENERALIZATION

Our learning problem seeks a classifier that minimize loss subject to attaining group envy free or equitability. As this learned classifier needs to act on any given individual and not just those in the training sample, the question of whether our proposed notions generalize becomes crucial. In other words, how closely aligned are the group envy or group equity violations between the training set and the whole population. Bounding this deviation to be small implies that if we learn a classifier that achieves good fairness on the training sample, it will also be fair on unseen data.

We now define the sampling procedure and adjust our fairness definitions accordingly. Our training set $\mathcal{S}$ consist of independent samples z_i, where $z_i = (x_1^+, x_2^+) \sim \mathcal{P}^{\mathcal{X} \times \mathcal{X}}$. We define an *inclusion function* $b_{ij}(x_1^+, x_2^+)$ that returns 1 if $x_1^+ \in G_i \wedge x_2^+ \in G_j$ and 0 otherwise. The family of inclusion function is given by $\mathcal{B}$, and we note $|\mathcal{B}| = |\mathcal{G}|$. Using this, we now define the following:

$$U_{a,b}(h, b_{ij}) = \frac{1}{|\mathcal{S}|} \sum_{(x_1^+, x_2^+) \in \mathcal{S}} u(x_a^+, h(x_b)) b_{ij}(x_1^+, x_2^+)$$

$$U^*_{a,b}(h, b_{ij}) = \mathop{\mathbb{E}}_{(x_1^+, x_2^+) \sim \mathcal{S}} \left[u(x_a^+, h(x_b)) b_{ij}(x_1^+, x_2^+) \right]$$

where $a \in \{0, 1\}$ and $b \in \{0, 1\}$ and h is a randomized classifier, $h \in \mathcal{H}_k$. $U_{12}(\cdot, b_{ij})$ consider samples where the pair belongs to group i and j respectively and computes the total utility individuals in group i have for the classification given to those in group j. $U_{11}(\cdot, b_{ij})$ looks at the same set of samples as $U_{12}(\cdot, b_{ij})$ (since they both have the same inclusion function $b_{ij}(x_1^+, x_2^+)$) and computes the total utility individuals in group i have for their own classification. As such, G_i is empirically ϵ-group envy free of G_j if $U_{11}(h, b_{ij}) - U_{12}(h, b_{ij}) \leq \epsilon$ and population ϵ-group envy free if $U^*_{11}(h, b_{ij}) \geq U^*_{12}(h, b_{ij}) \; \forall i, j$. Similarly, a classifier is empirically ϵ-group equitable on G_i and G_j, if $|U_{11}(h, b_{ij}) = U_{22}(h, b_{ij})| \leq \epsilon$ and population group equitable if $|U^*_{11}(h, b_{ij}) = U^*_{22}(h, b_{ij})| \leq \epsilon$.

The generalization result for group equitability follows the same procedure and achieves the same bound as that of group envy-free: as such, we explicitly illustrate the latter. Let $V(h, b_{ij}) = U_{12}(h, b_{ij}) - U_{11}(h, b_{ij})$ denote the empirical envy between groups G_i and G_j and $V^*(h_p, b_{ij}) = U^*_{12}(h, b_{ij}) - U^*_{11}(h, b_{ij})$ denote the population envy between those groups (for equitability, we will define V and V^* to measure empirical and population inequity). Generalization of a classifier h is then equivalent to showing that

with high probability, the difference between empirical and population envy (or inequity) is small for any group pairing. We first give some pertinent definitions and lemmas before presenting the formal theorem.

Definition 1. *For a function class $\mathcal{F}$ mapping to $\mathbb{R}$, a set of samples $\mathcal{S} = z_1, \ldots, z_m$, and a vector $\sigma \in \{-1, +1\}^m$, where each σ_i is an independent Rademacher random variable ($\mathcal{P}(\sigma_i = 1) = \mathcal{P}(\sigma_i = -1) = \frac{1}{2}$), the **Rademacher complexity** of $\mathcal{F}$, denoted $\mathscr{R}(\mathcal{F})$, is given by [24]:*

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

Definition 2. *[SH: Not that important] A set $\mathcal{S} = \{z_1, \ldots, z_m\}$ is multi-class shattered by a function class $\mathcal{F}$, if there exists two functions f_1 and f_2 such that: (1) $\forall z \in \mathcal{S}$, $f_1(z) \neq f_2(z)$ and (2) for every $B \subset \mathcal{S}$, there exists a function $f \in \mathcal{F}$ such that: $\forall z \in B$, $f(z) = f_1(z)$ and $\forall z \in \mathcal{S}$ B, $f(z) = f_2(z)$. The **Natarajan dimension** of $\mathcal{F}$ is the cardinality of the largest set of points that can be multiclass shattered by $\mathcal{F}$.*

Lemma 1. *For some function class $\mathcal{F}$ mapping to reals and a binary function $b : z \to \{0, 1\}$, define $\mathcal{F}_b = \{z \to f(z) \cdot b(z) : f \in \mathcal{F}\}$. Then $\mathscr{R}(\mathcal{F}_b) \leq \mathscr{R}(\mathcal{F})$.*

Proof. Note that to compute $\mathscr{R}(\mathcal{F})$, for each σ an $f \in \mathcal{F}$ is chosen so as to attain the supremum - call it f^*_σ. So we can equivalently write: $\mathscr{R}(\mathcal{F}) = \mathbb{E}_\sigma \left[\sum \sigma_i f^*_\sigma(z_i)\right]$. For each σ, denote the vector τ where $\tau_i = \sigma_i$ if $b(z_i) = 1$ and $\tau_i = 0$ if $b(z_i) = 0$. Now consider for each σ, choosing f^*_τ instead. Then we have: $\mathbb{E}_\sigma \left[\sum \sigma_i f^*_\tau(z_i)\right] \leq \mathscr{R}(\mathcal{F})$. Note that if σ changes only in components where $p(z_i) = 0$, then f^*_τ does not change. Moreover, since σ_i is 1 or -1 with equal probability, the expectation over the components where $p(z_i) = 0$ goes to 0, and we have: $\mathbb{E}_\sigma \left[\sum_{i=1}^m \sigma_i f^*_\tau(z_i)\right] = \mathbb{E}_\sigma \left[\sum_{i|b(z_i)=1} \sigma_i f^*_\tau(z_i)\right]$. This is exactly the same as the complexity of $\mathcal{F}_b$, as $\mathscr{R}(\mathcal{F}_b) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i|b(z_i)=1} \sigma_i f(z_i)\right]$. □

Lemma 2. *Consider the following function class:*

$$\mathcal{F} = \left\{(x_1^+, x_2^+) \to u(x_1^+, h_p(x_2)) b_{ij}(x_1^+, x_2^+) : h \in \mathcal{H}_k\right\}$$

Then $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{H})$, where $\mathcal{H}$ is the family of deterministic classifiers over whom a randomized mixture is learned.

Proof. Let $\mathcal{F}_1 = \left\{(x_1^+, x_2^+) \to u_p(x_1^+, h_p(x_2)) : h_p \in \mathcal{H}_p\right\}$ and by immediate application of Lemma 1, we have that: $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{F}_1)$. Expanding u, we have that: $u(x_1^+, h(x_2)) = \sum_{i=1}^k \eta_i u(x_1^+, h_d^i(x_2))$ which is a convex hull of the utilities attained under the k deterministic classifiers. Rademacher complexity of the convex hull of a set is equal to that of the set. Thus, we have that:

$$\mathscr{R}(\mathcal{F}_1) = \mathop{\mathbb{E}}_{\sigma} \left[\sup_{h_d \in \mathcal{H}_d} \sum \sigma_i u(x_i^+, h_d(x_2)) \right] \tag{1}$$

Note that $h_d : \mathcal{X} \to [d]$. Let $\alpha_i = u(x_i^+, i)$, and $\sum_i \alpha_i = 1$. We now define a surrogate $\tilde{u}(x_i^+, y)$ where $y \in [1, d]$ as:

$$\tilde{u}(x_i^+, y) = \alpha_{l(y)} + \left(\alpha_{c(y)} - \alpha_{c(y)}\right)(y - l(y)) \tag{2}$$

where $c(\cdot)$ and $l(\cdot)$ represent the ceiling and floor functions respectively. Note that $\tilde{u}$ is 1-Lipschitz in $[0, d]$, and $\tilde{u}(x_i^+, y) = u(x_i^+, y)$, when $y \in [d]$. As such, we can replace u with $\tilde{u}$ in equation 1, and since each $\tilde{u}(x_i^+, \cdot)$ is 1-Lipschitz, we can apply the contraction lemma [24]: $\mathscr{R}(\mathcal{F}_1) \leq \mathbb{E}_\sigma \left[\sup_{h_d \in \mathcal{H}_d} \sum \sigma_i h_d(x_2)\right] = \mathscr{R}(\mathcal{H}_d)$ □

Lemma 1 illustrates that for a given function class, multiplying it with fixed binary function does not increase it's Rademacher Complexity. Lemma 1 shows that the Rademacher complexity of the function class representing the summand of $\widehat{R}_{12}(h_p, b_{ij})$ can be upper bounded by the complexity of the class of deterministic classifiers. With the help of these two lemmas, we now present the following generalization theorem:

Theorem 4. *Let $\mathcal{H}$ represent the class of deterministic classifiers, over which a randomized mixture is learned. Then, for $\epsilon > \mathscr{R}(\mathcal{H})$ and $\delta > 0$, if $|\mathcal{S}| = \mathcal{O}\left(\frac{\log(|\mathcal{G}|)+\log(\delta)}{(\epsilon - \mathscr{R}(\mathcal{H}_d))^2}\right)$, then with probability at least $1 - \delta$, $\sup_{\mathcal{H}, \mathcal{B}} |V(h, b_{ij}) - V^*(h, b_{ij})| \leq \epsilon$*

Proof. By expanding V and V^*, we can re-write our bound as:

$$\mathcal{P}\left(\sup_{\mathcal{H}_k, \mathcal{B}}, |U_{12}(h, b_{ij}) - U_{11}(h, b_{ij}) + U^*_{11}(h, b_{ij}) - U^*_{12}(h, b_{ij})| \leq \epsilon\right)$$

$$\geq \mathcal{P}\left(\sup_{H_k, \mathcal{B}} |U_{12}(h, b_{ij}) - U^*_{12}(h, b_{ij})| + |U^*_{11}(h, b_{ij}) - U_{11}(h, b_{ij}))| \leq \epsilon\right) \tag{3}$$

where the last inequality follows from the triangle inequality. Let the event $A = \sup_{\mathcal{H}_k, \mathcal{B}} |U_{12}(h, b_{ij}) - U^*_{12}(h, b_{ij})| \geq \frac{\epsilon}{2}$ and the event $C = \sup_{\mathcal{H}_k, \mathcal{B}} |U_{11}(h, b_{ij}) - U^*_{11}(h, b_{ij})| \geq \frac{\epsilon}{2}$ Now using the identity, $\sup(X + Y) \leq \sup(X) + \sup(Y)$, we can write that equation $3 \geq \mathcal{P}(\neg A \wedge \neg C) = \mathcal{P}(\neg(A \vee C))$. For event A: $\mathcal{P}(A) \leq \sum_{b_{ij} \in \mathcal{B}} \mathcal{P}\left(\sup_{\mathcal{H}_k} |U^{12}_{ij}(h, b_{ij}) - U^{*\,12}_{ij}(h, b_{ij})| \geq \frac{\epsilon}{2}\right)$ by union bound. Let $\mathcal{F} = \left\{(x_1^+, x_2^+) \to u(x_1^+, h(x_2)) * b_{ij}(x_1^+, x_2^+) : h_p \in \mathcal{H}_p\right\}$. By appealing to generalization bound given in [24], we can write:

$$\mathcal{P}\left(\sup_{\mathcal{H}_k} |U_{12}(h, b_{ij}) - U^*_{12}(h, b_{ij})| \geq \frac{\epsilon}{2}\right) \leq 4\exp\left(\frac{-|\mathcal{S}|}{2}\left(\frac{\epsilon - 4\mathscr{R}(\mathcal{F})}{8}\right)^2\right)$$

where $\epsilon >= 4\mathscr{R}(\mathcal{F})$. Denote this upper bound to be δ_1. As such: $\mathcal{P}(A) \leq |\mathcal{G}|\delta'$. Secondly, note that due to Lemma 2, $\mathscr{R}(\mathcal{F}) \leq \mathscr{R}(\mathcal{H})$. We perform the exact same procedure for event C and write: $\mathcal{P}(C) \leq |\mathcal{G}|\delta'$. Combining these results, we have that $\mathcal{P}(\neg(A \vee C)) \geq 1 - |\mathcal{G}|^2(\delta') \triangleq 1 - \delta$. Solving for the sample complexity, we have that: $|\mathcal{S}| = \mathcal{O}\left(\frac{\log(|\mathcal{G}|)+\log(\delta)}{(\epsilon - \mathscr{R}(\mathcal{H}_d))^2}\right)$ □

Theorem 4 implies that largest possible group envy (or equivalently group inequity) for any two groups is bounded by ϵ with probability $1 - \delta$. The sample complexity for this bound scales logarithmically with $|\mathcal{G}|$, allowing for large, possibly exponential, number of groups with only a linear increase in samples. Secondly, we note that the generalization error and sample complexity both scale poorly with increasing complexity of $\mathcal{H}$. As such, deterministic classifiers of low complexity are desired. As we are in a multiclass classification setting, we rely on analysis based on Natarajan dimension to show that using a linear multiclass classifer leads to good generalization. We first state two lemmas, first of which is a multiclass analogue of Sauer's lemma. It bounds the number of different labellings of the sample $\mathcal{S}$ achievable by a function class $\mathcal{F}$, $|\mathcal{F}|_{\mathcal{S}}$

Lemma 3. *(Natarajan) For a family $\mathcal{F}$ of Natarajan dimension j, and any subset $\mathcal{S} \subseteq \mathcal{X}$, we have that $|\mathcal{F}|_{\mathcal{S}} = |\mathcal{S}|^j d^{2j}$*

Lemma 4. *(Massart) Let $A \subseteq \mathbb{R}^m$ be a finite set of points with $r = max_{z \in A}||z||_2$. Then $\frac{1}{m}\mathbb{E}_\sigma[max_{x \in A}\sum_{i=1}^m z_i\sigma_i] \leq \frac{r}{m}\sqrt{2\log(|A|)}$*

Theorem 5. *Let $\mathcal{H}$ be the family of linear one-vs-all classifiers.*

$$\mathcal{H}_d = \left\{ x \in \mathcal{X} \rightarrow \arg\max_{y \in [d]} w_y^T x \ : \ w_y \in \mathbb{R}^m \right\}$$

be the family of linear one-vs-all classifiers. Then the Rademacher complexity of $\mathcal{H}$ is given by: $\mathscr{R}(\mathcal{H}) = \frac{2d}{|\mathcal{S}|}\sqrt{|\mathcal{S}|md\log(|\mathcal{S}|d)}$

Proof. From [24], we have that the Natarajan dimension of $\mathcal{H}_d \leq md$. Then by lemma 3, for a sample $\mathcal{S}$, $|\mathcal{S}|^{md}d^{2md}$ possible labellings that can be achieved by $\mathcal{H}$. For some any of samples $\mathcal{S}$, let $h(\bar{x}) = \left[h(x_1)\dots h(x_{|\mathcal{S}|})\right]$, and note that $||h_d(\bar{x})||_2 \leq d\sqrt{|\mathcal{S}|}$. Since $\mathcal{H}$ achieves a finite set of labellings and $h(\bar{x})$ is bounded, we can apply Massert's lemma and get: $\mathscr{R}(\mathcal{H}_d) = \frac{2d}{|\mathcal{S}|}\sqrt{|\mathcal{S}|md\log(|\mathcal{S}|d)}$. □

[SH: To Nisarg: Please add the comment on how changing the value of class labellings will change the lipschitz constant and some discussion on that]
[SH: To Nisarg: Please use that trick you mentioned and use the final complexity to express the sample complexity as max()]
[SH: To Nisarg: Add a little discussion on the end about the generality of the result. No need to assume groups are mutually exclusive, exponential number of groups, can be across any pairs ...]

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:

Lemma 5. *Consider an array $[a_1, a_2, \dots, a_n]$ sorted in increasing order, with $a_i \in [0..1]$. Then: $\sum_{i=1}^n \sum_{j=i}^n |a_i - a_j| \leq \left\lceil \frac{n}{2} \right\rceil \left\lfloor \frac{n}{2} \right\rfloor$*

Proof. Notice that $\sum_{i=1\dots n} \sum_{j=i\dots n} |a_i - a_j| = \sum_{i=1\dots n}(2i - n - 1)a_i \leq \left\lceil \frac{n}{2} \right\rceil \left\lfloor \frac{n}{2} \right\rfloor$ and that equality is achieved when $\forall\, i = 1, \dots, \left\lceil \frac{n}{2} \right\rceil$, $a_i = 0$ and $\forall\, i = \left\lceil \frac{n}{2} \right\rceil + 1, \dots, n$, $a_i = 1$ □

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

Proof. Since there are $g(g-1)$ pair of constraints and maximum possible group envy is 1, the total group envy is upper bounded by $g(g-1) = O(g^2)$. To show this bound is tight, we now construct an instance that achieves this. Numbering the groups $1, \dots, g$, let any individual in group $i \leq \left\lfloor \frac{g}{2} \right\rfloor$ have utility 1 for class 1 and 0 for any other classes. Let them also have loss 1 for all classes but class 2, for which they have 0 loss. Similarly, for individuals in group $i > \left\lfloor \frac{g}{2} \right\rfloor$, let them have utility 1 for class 2 and 0 for all other classes. Such individuals also have loss 1 for all classes, except class 1 for which they have 0 loss. In this instance, the ERM classifier assigns all individuals in group $i \leq \left\lfloor \frac{g}{2} \right\rfloor$ class 2 and all individuals in group $i > \left\lfloor \frac{g}{2} \right\rfloor$ class 1. It is then evident that the total group envy is $\left\lfloor \frac{g}{2} \right\rfloor \left\lceil \frac{g}{2} \right\rceil = \Omega(g^2)$. □

[SH: Can make this $\Theta(g^2)$ too actually]

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 5. We now show this is tight by constructing an instance that achieves this. First, number the groups $1, \dots, k$ 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 5, 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 $\Omega\left(\frac{n(g-1)}{g}\right)$*

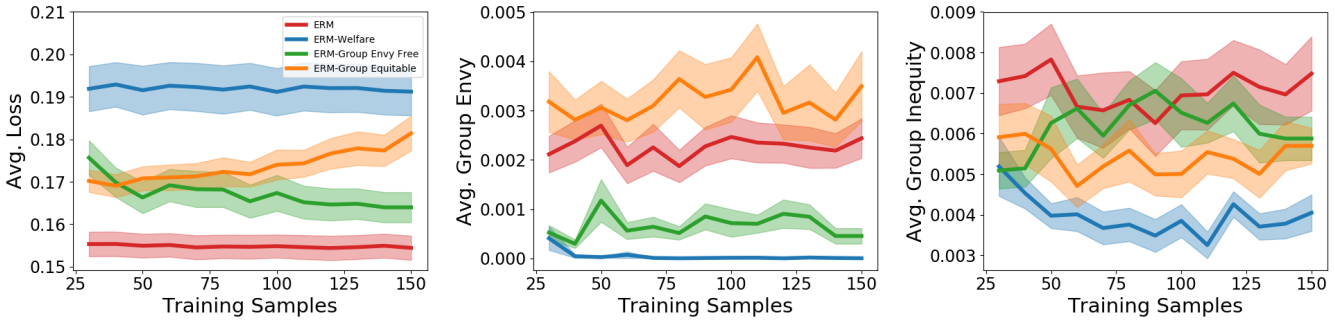

Figure 1: Plotting test values with 80% confidence intervals as a function of training set size

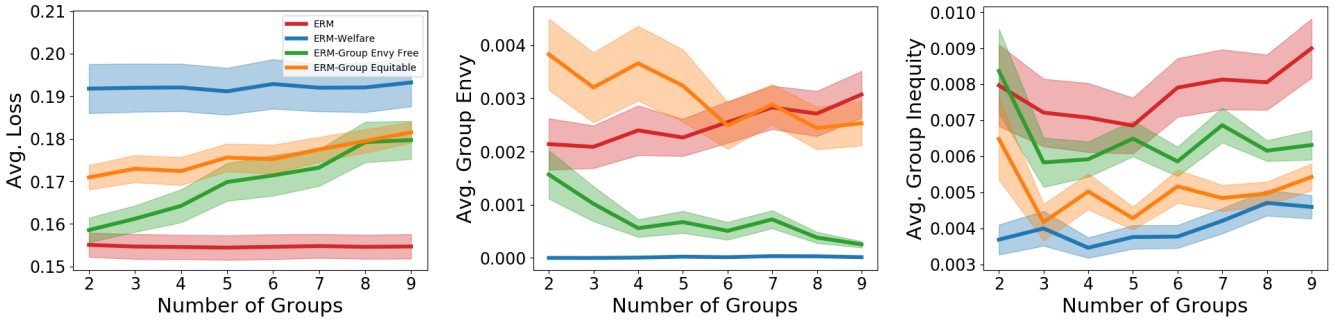

Figure 2: Plotting test values with 80% confidence intervals as a function of training set size

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* $\Omega\left(\left\lfloor \frac{g}{2} \right\rfloor * \frac{n}{g}\right)$.

Proof. Let the sample contain an equal number of individuals in each class n/g. Numbering the groups $1, \ldots, g$, let $S_1, \ldots, S_g$ 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 $\Omega\left(\left\lfloor \frac{g}{2} \right\rfloor * \frac{n}{g}\right)$ □

[SH: Write some discussion here on results]

7 IMPLEMENTATION AND EXPERIMENTS

We now consider the problem of implementing a classifier that achieves our proposed fairness notions. The approach outlined is similar to the one proposed by [5] for building envy-free classifiers. We extend their procedure to a group setting for both envy-free and equitability. We note that this is not meant to be exhaustive but rather illustrate that computation of such fair classifiers is feasible. We leave the design of more robust classifier with better guarantees for future work.

Formally, our goal is to learn a randomized classifier h_r from a family $\mathcal{H}$ that minimizes empirical risk while ensuring group envy free/group equitability. Note that we consider randomized classifiers that are a mixture over k deterministic classifiers. Let $\mathbf{h} = \left\{h_d^1, \ldots, h_d^k\right\}$ denote the vector of deterministic classifiers and $\boldsymbol{\eta}$ the vector of k mixing co-efficients. Section 5 suggests that for good generalization results, we learn deterministic multiclass

classifiers of low complexity, and theorem 5 given a result for a linear one-vs-all classifiers. Utilizing this in our experiments, our deterministic classifier family $\mathcal{H}$ is:

$$\mathcal{H}_d = \left\{ x \in \mathcal{X} \rightarrow \arg\max_{y \in [d]} w_y^T x \ : \ w_y \in \mathbb{R}^m \right\} \tag{4}$$

Using such classifiers, we look solve an optimization problem that minimizes empirical risk while satisfying group envy free or group equitability constraints for the pairs of groups in $\mathcal{G}$. Equation 5 expresses this formally for group envy free and equation 6 for group equitability.

$$\min_{\substack{\mathbf{h} \in \mathcal{H}^m \\ \boldsymbol{\eta} \in \Delta_k}} \sum_{t=1}^{k} \eta_t \sum_{x^+ \in \mathcal{S}} \ell(x^+, h_d^t(x)) \text{ such that } \forall\ (G, \widehat{G}) \in \mathcal{G} \tag{5}$$

$$\frac{1}{|\mathcal{S}^G|} \sum_{t=1}^{k} \eta_t \sum_{x^+ \in S^G} u(x^+, h_d^t(x)) \geq \frac{1}{|\mathcal{S}^G||\mathcal{S}^{\widehat{G}}|} \sum_{t=1}^{k} \eta_t \sum_{\substack{x^+ \in S^G \\ \widehat{x}^+ \in S^{\widehat{G}}}} u(x^+, h_d^t(\widehat{x}))$$

$$\min_{\substack{\mathbf{h} \in \mathcal{H}^m \\ \boldsymbol{\eta} \in \Delta_k}} \sum_{t=1}^{k} \eta_t \sum_{x^+ \in \mathcal{S}} \ell(x^+, h_d^t(x)) \text{ such that } \forall\ (G, \widehat{G}) \in \mathcal{G} \tag{6}$$

$$\frac{1}{|\mathcal{S}^G|} \sum_{t=1}^{k} \eta_t \sum_{x^+ \in S^G} u(x^+, h_d^t(x)) \geq \frac{1}{|\mathcal{S}^{\widehat{G}}|} \sum_{t=1}^{k} \eta_t \sum_{\widehat{x}^+ \in S^{\widehat{G}}} u(\widehat{x}^+, h_d^t(\widehat{x}))$$

Convex Relaxation: In both problems 5 and 6 $\ell(x^+, h_d^t(x))$ 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 relaxed version of these quantities that are convex. This relaxation approach was also utilized by Balcan et al. [5] and is based on multiclass-SVM. Note for any $c \in [d]$, $\ell(x^+, h_d^t(x)) \leq \ell(x^+, h_d^t(x)) + w_{h_d^T(x)}^T x - w_c^T x$. Therefore:

$$\therefore \ell(x^+, h_d^t(x)) \leq \max_{y \in \mathcal{Y}} \ell(x^+, y) + w_y^T x - w_c^T x \tag{7}$$

giving us a convex upper bound to $\ell(x^+, h_d^t(x))$. So perform a similar relaxation for $u(x^+, h_d^T(x))$.

Training: We still face an issue of tractability due to the combinatorial nature of the optimization problem: equation 5 and equation 6 requires we learn k classifiers and their mixing co-efficients simultaneously. To circumvent this, we use a two step process where we start by setting $\boldsymbol{\eta}$ to default values and learn the k classifiers one by one. In the second step, we solve a linear problem to learn the optimal $\boldsymbol{\eta}$ for the k learned classifiers. Secondly for tractability reasons, instead of enforcing group envy or group equity violations as hard constraints, we instead add a fairness penalty to the loss, controlling its effect with a hyperparameter λ. Letting $G_{envy}(S^G, S^{\widehat{G}})$ denote the group envy that S^G has for $S^{\widehat{G}}$ and noting that negative envy is not necessarily desirable, we minimize the following penalty term: $\max(G_{envy}(S^G, S^{\widehat{G}}), 0)$. Similarly using $G_{ineq}(S^G, S^{\widehat{G}})$ to denote the group inequity, we minimize the following penalty term: $\max(G_{ineq}(S^G, S^{\widehat{G}}), G_{ineq}(S^G, S^{\widehat{G}}))$. Using the convex relaxations for $u(x^+, h(x))$ renders these penalty terms to also be convex and thus our minimization objective as a whole.

7.1 Experiment

REFERENCES

[1] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69.

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

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

[4] Balcan, M. F., Constantin, F., Iwata, S., and Wang, L. (2012). Learning valuation functions. In *Conference on Learning Theory*, pages 4–1.

[5] Balcan, M.-F., Dick, T., Noothigattu, R., and Procaccia, A. D. (2019). Envy-free classification. In *Advances in neural information processing systems*. Forthcoming.

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

[7] 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.

[8] Chajewska, U., Koller, D., and Ormoneit, D. (2001). Learning an agent's utility function by observing behavior. In *ICML*, pages 35–42.

[9] Chen, X., Fain, B., Lyu, L., and Munagala, K. (2019). Proportionally fair clustering. In *International Conference on Machine Learning*, pages 1032–1041.

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

[11] Conitzer, V., Freeman, R., Shah, N., and Vaughan, J. W. (2019). Group fairness for the allocation of indivisible goods. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1853–1860.

[12] 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. ACM.

[13] Fain, B., Goel, A., and Munagala, K. (2016). The core of the participatory budgeting problem. In *International Conference on Web and Internet Economics*, pages 384–399. Springer.

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

[15] Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *Americal Mathematical Monthly*, 69(1):9–15.

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

[17] Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (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*, pages 181–190. ACM.

[18] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (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*, page 600. ACM.

[19] Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (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*, pages 43:1–43:23.

[20] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

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

[22] Ramnarayan, G. (2018). Equalizing financial impact in supervised learning. *arXiv preprint arXiv:1806.09211*.

[23] Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

[24] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

[25] Shapley, L. and Scarf, H. (1974). On cores and indivisibility. *Journal of mathematical economics*, 1(1):23–37.

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

[27] Stromquist, W. (1980). How to cut a cake fairly. *American Mathematical Monthly*, 87(8):640–644.

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

[29] 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.

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

[31] 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.