Envy-Free Classification

Maria-Florina Balcan [1] **Travis Dick** [2] **Ritesh Noothigattu** [1] **Ariel D. Procaccia** [2]

Abstract

In classic fair division problems such as cake cutting and rent division, *envy-freeness* requires that each individual (weakly) prefer his allocation to anyone else's. On a conceptual level, we argue that envy-freeness also provides a compelling notion of fairness for classification tasks. Our technical focus is the *generalizability* of envy-free classification, i.e., understanding whether a classifier that is envy free on a sample would be almost envy free with respect to the underlying distribution with high probability. Our main result establishes that a small sample is sufficient to achieve such guarantees, when the classifier in question is a mixture of deterministic classifiers that belong to a family of low Natarajan dimension.

1. Introduction

The study of fairness in machine learning is driven by an abundance of examples where learning algorithms were perceived as discriminating against protected groups (Sweeney, 2013; Datta et al., 2015). Addressing this problem requires a conceptual — perhaps even philosophical — understanding of what fairness means in this context. In other words, the million dollar question is (arguably[1]) this: What are the formal constraints that fairness imposes on learning algorithms? On a very high level, most of the answers proposed so far (Luong et al., 2011; Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Hardt et al., 2016; Joseph et al., 2016; Zafar et al., 2017a;b) fall into two (partially overlapping) categories: individual fairness notions, and group fairness notions.

In the former category, the best known example is the influential fair classification model of Dwork et al. (2012). The model involves a set of individuals and a set of outcomes. It is instructive to think of financially-motivated settings where the outcomes are, say, credit card offerings or displayed advertisements, and a loss function represents the benefit (e.g., in terms of revenue) of mapping a given individual to a given outcome. The centerpiece of the model is a *similarity metric* on the space of individuals; it is specific to the classification task at hand, and ideally captures the ethical ground truth about relevant attributes. For example, a man and a woman who are similar in every other way should be considered similar for the purpose of credit card offerings, but perhaps not for lingerie advertisements. Assuming such a metric is available, fairness can be naturally formalized as a Lipschitz constraint, which requires that individuals who are close according to the similarity metric be mapped to distributions over outcomes that are close according to some standard metric (such as total variation). The algorithmic problem is then to find a classifier that minimizes loss, subject to the Lipschitz constraint.

As attractive as this model is, it has one clear weakness from a practical viewpoint: the availability of a similarity metric. Dwork et al. (2012) are well aware of this issue; they write that justifying this assumption is "one of the most challenging aspects" of their approach. They add that "in reality the metric used will most likely only be society's current best approximation to the truth." But, despite recent progress on automating ethical decisions in certain domains (Noothigattu et al., 2018; Freedman et al., 2018), the task-specific nature of the similarity metric makes even a credible approximation thereof seem unrealistic. In particular, if one wanted to learn a similarity metric, it is unclear what type of examples a relevant dataset would consist of.

An alternative notion of individual fairness, therefore, is called for. And our proposal draws on an extensive body of work on rigorous approaches to fairness, which — modulo one possible exception (see Section 1.2) — has not been tapped by machine learning researchers: the literature on *fair division* (Brams & Taylor, 1996; Moulin, 2003). The most prominent notion is that of *envy-freeness* (Foley, 1967; Varian, 1974), which, in the context of the allocation of goods, requires that the utility of each individual for his allocation be at least as high as his utility for the allocation of any other individual; this is the gold standard of fairness for problems such as cake cutting (Robertson & Webb, 1998; Procaccia, 2013) and rent division (Su, 1999; Gal et al.,

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA [2]Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Ariel D. Procaccia <arielpro@cs.cmu.edu>.

[1]Recent work takes a somewhat different view (Kilbertus et al., 2017).

2017).

Similarly, in the classification setting, envy-freeness would simply mean that the utility of each individual for his distribution over outcomes is at least as high as his utility for the distribution over outcomes assigned to any other individual. For example, it may well be the case that Bob is offered a worse credit card than that offered to Alice, but this outcome is not unfair if Bob is genuinely more interested in the card offered to him because he does not qualify for Alice's card. Of course, as before, envy-freeness requires access to individuals' utility functions, but — in stark contrast to the similarity metric of Dwork et al. (2012) — there are a variety of techniques for learning utility functions (Chajewska et al., 2001; Nielsen & Jensen, 2004; Balcan et al., 2012). It is also worth noting that the classification setting is different from classic fair division problems in that the "goods" (outcomes) are non-excludable. In fact, one envy-free solution simply assigns each individual to his favorite outcome; but when the loss function disagrees with the utility functions, it may be possible to achieve smaller loss without violating the envy-freeness constraint.

In summary, we view envy-freeness as a compelling, well-established, and, importantly, practicable notion of individual fairness for classification tasks. Our goal is to understand its learning-theoretic properties.

1.1. Our Results

The technical challenge we face is that the space of individuals is potentially huge, yet we seek to provide universal envy-freeness guarantees. To this end, we are given a sample consisting of individuals drawn from an unknown distribution. We are interested in learning algorithms that minimize loss, subject to satisfying the envy-freeness constraint, *on the sample*. Our primary technical question is that of generalizability, that is,

> *given a classifier that is envy free on a sample, is it approximately envy free on the underlying distribution?*

Surprisingly, Dwork et al. (2012) do not study generalizability in their model, and subsequent work, which does take a learning-theoretic viewpoint, does not give theoretical guarantees (see Section 1.2). We therefore view this question as part of our conceptual contribution.

In Section 3, we do not constrain the classifier in question. Therefore, we need some strategy to extend a classifier that is defined on a sample; assigning an individual the same outcome as his *nearest neighbor* in the sample is a popular choice. However, we show that *any* strategy for extending a classifier from a sample, on which it is envy free, to the entire set of individuals is unlikely to be approximately envy free on the underlying distribution, unless the sample is exponentially large.

For this reason, in Section 4, we focus on structured families of classifiers. On a high level, our goal is to relate the combinatorial richness of the family to generalization guarantees. One obstacle is that standard notions of dimension do not extend to the analysis of randomized classifiers, whose range is *distributions* over outcomes (equivalently, real vectors). We circumvent this obstacle by considering mixtures of *deterministic* classifiers that belong to a family of bounded Natarajan dimension (an extension of the well-known VC dimension to multi-class classification). Our main technical result asserts that, under this assumption, envy-freeness on a sample does generalize to the underlying distribution, even if the sample is relatively small (its size grows almost linearly in the Natarajan dimension). Finally, we discuss the implications of this result in Section 5.

1.2. Related Work

Conceptually, our work is most closely related to very recent, independent work by Zafar et al. (2017b). They are interested in group notions of fairness, and advocate preference-based notions instead of parity-based notions. In particular, they assume that each group has a utility function for *classifiers*, and define the *preferred treatment* property, which requires that the utility of each group for its own classifier be at least its utility for the classifier assigned to any other group. Their model and results focus on the case of binary classification where there is a desirable outcome and an undesirable outcome, so the utility of a group for a classifier is simply the fraction of its members that are mapped to the desirable outcome. Although, at first glance, this notion seems similar to envy-freeness, it is actually fundamentally different. To intuitively understand why, observe that in the binary classification setting of Zafar et al. (where everyone agrees on the desirable outcome), envy-freeness would constrain the classifier to map each and every individual to the desirable outcome with the exact same probability.[2] Our paper is also completely different from that of Zafar et al. in terms of technical results; theirs are purely empirical in nature, and focus on the increase in accuracy obtained when parity-based notions of fairness are replaced with preference-based ones.

Another related paper is the one by Zemel et al. (2013); although the connection is superficial, it is worth clarifying. Like us, their starting point is the paper of Dwork

[2] On a philosophical level, the fair division literature deals exclusively with individual notions of fairness. In fact, even in group-based extensions of envy-freeness (Manurangsi & Suksompong, 2017) the allocation is shared by groups, but individuals must not be envious. We subscribe to the view that group-oriented notions (such as statistical parity) are objectionable, because the outcome can be patently unfair to individuals.

et al. (2012), and the concern that the similarity metric is unrealistic. However, fairness in their framework is derived indirectly from statistical parity. Specifically, they learn two mappings, one from individuals to representations (which satisfies statistical parity), and one from representations to outcomes (which is unconstrained). Zemel et al. write that their learning approach "permits generalization to new examples distinct from those in the training set," but they do not provide theoretical guarantees.

2. The Model

We suppose that there is a space $\mathcal{X}$ of individuals, a finite space $\mathcal{Y}$ of outcomes, and a utility function $u : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ encoding the preferences of each individual for the outcomes in $\mathcal{Y}$. In the advertising example, individuals are users, outcomes are advertisements, and the utility function reflects the benefit an individual derives from being shown a particular advertisement. For any distribution $p \in \Delta(\mathcal{Y})$ (where $\Delta(\mathcal{Y})$ is the set of distributions over $\mathcal{Y}$) we let $u(x, p) = \mathbb{E}_{y\sim p}[u(x, y)]$ denote individual x's expected utility for an outcome sampled from p. We refer to a function $h : \mathcal{X} \to \Delta(\mathcal{Y})$ as a *classifier*, even though it can return a distribution over outcomes.

2.1. Envy-Freeness

Roughly speaking, a classifier $h : \mathcal{X} \to \Delta(\mathcal{Y})$ is envy free if no individual prefers the outcome distribution of someone else over his own.

Definition 1. A classifier $h : \mathcal{X} \to \Delta(\mathcal{Y})$ is *envy free (EF)* on a set S of individuals if $u(x, h(x)) \geq u(x, h(x'))$ for all $x, x' \in S$. Similarly, h is (α, β)-*EF* with respect to a distribution P on $\mathcal{X}$ if

$$\Pr_{x,x'\sim P}\big(u(x, h(x)) < u(x, h(x')) - \beta\big) \leq \alpha.$$

Finally, h is (α, β)-*pairwise EF* on a set of pairs of individuals $S = \{(x_i, x_i')\}_{i=1}^n$ if

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u(x_i, h(x_i)) < u(x_i, h(x_i')) - \beta\} \leq \alpha.$$

Any classifier that is EF on a sample S of individuals is also (α, β)-pairwise EF on any pairing of the individuals in S, for any $\alpha \geq 0$ and $\beta \geq 0$. The weaker pairwise EF condition is all that is required for our generalization guarantees to hold.

2.2. Optimization and Learning

Our formal learning problem can be stated as follows. Given sample access to an unknown distribution P over individuals $\mathcal{X}$ and their utility functions, and a known loss function $\ell : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, find a classifier $h : \mathcal{X} \to \Delta(\mathcal{Y})$ that is (α, β)-EF with respect to P minimizing expected loss $\mathbb{E}_{x\sim P}[\ell(x, h(x))]$, where for $x \in \mathcal{X}$ and $p \in \Delta(\mathcal{Y})$, $\ell(x, p) = \mathbb{E}_{y\sim p}[\ell(x, y)]$.

We follow the empirical risk minimization (ERM) learning approach, i.e., we collect a sample of individuals drawn i.i.d from P and find an EF classifier with low loss on the sample. Formally, given a sample of individuals $S = \{x_1, \ldots, x_n\}$ and their utility functions $u_{x_i}(\cdot) = u(x_i, \cdot)$, we are interested in a classifier $h : S \to \Delta(\mathcal{Y})$ that minimizes $\sum_{i=1}^n \ell(x_i, h(x_i))$ among all classifiers that are EF on S. The algorithmic problem itself is beyond the scope of the current paper; see Section 5 for further discussion.

2.3. Deterministic Versus Randomized Classifiers

We consider classifiers of the form $h : \mathcal{X} \to \Delta(\mathcal{Y})$, which can assign a distribution over outcomes to each of the individuals. However, one might wonder whether the EF classifier that minimizes loss on a sample happens to always be deterministic. Or, at least, the optimal deterministic classifier on the sample might incur a loss that is very close to that of the optimal randomized classifier. If this were true, we could restrict ourselves to classifiers of the form $h : \mathcal{X} \to \mathcal{Y}$, which would be much easier to analyze.

Unfortunately, it turns out that this is not the case. In fact, there could be an arbitrary (multiplicative) gap between the optimal randomized EF classifier and the optimal deterministic EF classifier. The intuition behind this is as follows. A deterministic classifier that has very low loss on the sample, but is not EF, would be completely discarded in the deterministic setting. On the other hand, a randomized classifier could take this loss-minimizing deterministic classifier and mix it with a classifier with high "negative envy", so that the mixture ends up being EF and at the same time has low loss. This is made concrete in Example 1 in the appendix.

More generally, it turns out that deterministic EF classifiers are very restrictive in our setting. Indeed, the following theorem (whose proof is relegated to the appendix) asserts that any deterministic EF classifier is equivalent to picking a subset of the outcomes, and assigning every individual in the sample to their favorite outcome in this subset.

Theorem 1. *A deterministic classifier $h : \mathcal{X} \to \mathcal{Y}$ is EF on sample S if and only if there exists a subset $\mathcal{Z} \subseteq \mathcal{Y}$ such that $h(x) \in \arg\max_{y\in\mathcal{Z}} u(x, y)$ for every $x \in S$.*

By contrast, there is no analogous characterization for randomized classifiers. In particular, even if an individual x does not get their favorite outcome and another individual x' is assigned this outcome with non-zero probability, this can be balanced by giving x' an outcome that x dislikes, thus eliminating any envy. Moreover, even if x is assigned his favorite outcome with some probability, he can also be given other outcomes with non-zero probability, making the

randomized classifier quite flexible.

3. Arbitrary Classifiers

An important (and typical) aspect of our learning problem is that the classifier h needs to provide an outcome distribution for every individual, not just those in the sample. For example, if h chooses advertisements for visitors of a website, the classifier should still apply when a new visitor arrives. Moreover, when we use the classifier for new individuals, it must continue to be EF. In this section, we consider two-stage approaches that first choose outcome distributions for the individuals in the sample, and then extend those decisions to the rest of $\mathcal{X}$.

In more detail, we are given a sample $S = \{x_1, \ldots, x_n\}$ of individuals and a classifier $h : S \to \Delta(\mathcal{Y})$ assigning outcome distributions to each individual. Our goal is to extend these assignments to a classifier $\overline{h} : \mathcal{X} \to \Delta(\mathcal{Y})$ that can be applied to new individuals as well. For example, h could be the loss-minimizing EF classifier on the sample S.

For this section, we assume that $\mathcal{X}$ is equipped with a distance metric d. Moreover, we assume in this section that the utility function u is L-Lipschitz on $\mathcal{X}$. That is, for every $y \in \mathcal{Y}$ and for all $x, x' \in \mathcal{X}$, we have $|u(x, y) - u(x', y)| \leq L \cdot d(x, x')$.

Under the foregoing assumptions, one natural way to extend the classifier on the sample to all of $\mathcal{X}$ is to assign new individuals the same outcome distribution as their nearest neighbor in the sample. Formally, for a set $S \subset \mathcal{X}$ and any individual $x \in \mathcal{X}$, let $\mathrm{NN}_S(x) \in \arg\min_{x' \in S} d(x, x')$ denote the nearest neighbor of x in S with respect to the metric d (breaking ties arbitrarily). The following simple result (whose proof is relegated to the appendix) establishes that this approach preserves envy-freeness in cases where the sample is exponentially large.

Theorem 2. *Let d be a metric on $\mathcal{X}$, P be a distribution on $\mathcal{X}$, and u be an L-Lipschitz utility function. Let S be a set of individuals such that there exists $\hat{\mathcal{X}} \subset \mathcal{X}$ with $P(\hat{\mathcal{X}}) \geq 1 - \alpha$ and $\sup_{x \in \hat{\mathcal{X}}}(d(x, \mathrm{NN}_S(x)) \leq \beta/(2L)$. Then for any classifier $h : S \to \Delta(\mathcal{Y})$ that is EF on S, the extension $\overline{h} : \mathcal{X} \to \Delta(\mathcal{Y})$ given by $\overline{h}(x) = h(\mathrm{NN}_S(x))$ is (α, β)-EF on P.*

The conditions of Theorem 2 require that the set of individuals S is a $\beta/(2L)$-net for at least a $(1 - \alpha)$-fraction of the mass of P on $\mathcal{X}$. In several natural situations, an exponentially large sample guarantees that this occurs with high probability. For example, if $\mathcal{X}$ is a subset of $\mathbb{R}^q$, $d(x, x') = \|x - x'\|_2$, and $\mathcal{X}$ has diameter at most D, then for any distribution P on $\mathcal{X}$, if S is an i.i.d. sample of size $O(\frac{1}{\alpha}(\frac{LD\sqrt{q}}{\beta})^q(q \log \frac{LD\sqrt{q}}{\beta} + \log \frac{1}{\delta}))$ will satisfy the conditions of Theorem 2 with probability at least $1 - \delta$. This sampling result is folklore, but, for the sake of completeness, we prove it in Lemma 5 of Appendix B.

However, the exponential upper bound given by the nearest neighbor strategy is as far as we can go in terms of generalizing envy-freeness from a sample (without further assumptions). Specifically, our next result establishes that *any* algorithm — even randomized — for extending classifiers from the sample to the entire space $\mathcal{X}$ requires an exponentially large sample of individuals to ensure envy-freeness on the distribution P.

Theorem 3. *There exists a space of individuals $\mathcal{X} \subset \mathbb{R}^q$, and a distribution P over $\mathcal{X}$ such that, for every randomized algorithm $\mathcal{A}$ that extends classifiers on a sample to $\mathcal{X}$, there exists an L-Lipschitz utility function u such that, when a sample of individuals S of size $n = 4^q/2$ is drawn from P without replacement, there exists an EF classifier on S for which, with probability at least $1 - 2\exp(-4^q/100) - \exp(-4^q/200)$ jointly over the randomness of $\mathcal{A}$ and S, its extension by $\mathcal{A}$ is not (α, β)-EF with respect to P for any $\alpha < 1/25$ and $\beta < L/8$.*

We remark that a similar result would hold even if we sampled S with replacement; we sample here without replacement purely for ease of exposition.

Proof of Theorem 3. Let the space of individuals be $\mathcal{X} = [0, 1]^q$ and the outcomes be $\mathcal{Y} = \{0, 1\}$. We partition the space $\mathcal{X}$ into cubes of side length $s = 1/4$. So, the total number of cubes is $m = (1/s)^q = 4^q$. Let these cubes be denoted by $c_1, c_2, \ldots c_m$, and let their centers be denoted by $\mu_1, \mu_2, \ldots \mu_m$. Next, let P be the uniform distribution over the centers $\mu_1, \mu_2, \ldots \mu_m$. For brevity, whenever we say "utility function" in the rest of the proof, we mean "L-Lipschitz utility function."

To prove the theorem, we use Yao's minimax principle (Yao, 1977). Specifically, consider the following two-player zero sum game. Player 1 chooses a deterministic algorithm $\mathcal{D}$ that extends classifiers on a sample to $\mathcal{X}$, and player 2 chooses a utility function u on $\mathcal{X}$. For any subset $S \subset \mathcal{X}$, define the classifier $h_{u,S} : S \to \mathcal{Y}$ by assigning each individual in S to his favorite outcome with respect to the utility function u, i.e. $h_{u,S}(x) = \arg\max_{y \in \mathcal{Y}} u(x, y)$ for each $x \in S$, breaking ties lexicographically. Define the cost of playing algorithm $\mathcal{D}$ against utility function u as the probability over the sample S (of size $m/2$ drawn from P without replacement) that the extension of $h_{u,S}$ by $\mathcal{D}$ is not (α, β)-EF with respect to P for any $\alpha < 1/25$ and $\beta < L/8$. Yao's minimax principle implies that for any randomized algorithm $\mathcal{A}$, its expected cost with respect to the worst-case utility function u is at least as high as the expected cost of any distribution over utility functions that is played against the best deterministic algorithm $\mathcal{D}$ (which is tailored for that distribution). Therefore, we establish the

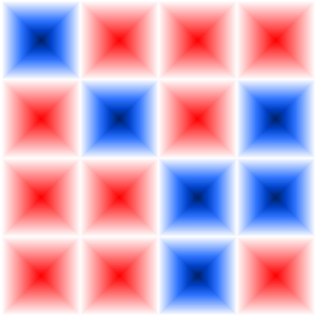

Figure 1. Illustration of $\mathcal{X}$ and an example utility function u for $d = 2$. Red shows preference for 1, blue shows preference for 0, and darker shades correspond to more intense preference. (The gradients are rectangular to match the L_∞ norm, so, strangely enough, the misleading X pattern is an optical illusion.)

desired lower bound by choosing a specific distribution over utility functions, and showing that the best deterministic algorithm against it has an expected cost of at least $1 - 2\exp(-m/100) - \exp(-m/200)$.

To define this distribution over utility functions, we first sample outcomes $y_1, y_2, \ldots, y_m$ i.i.d. from Bernoulli$(1/2)$. Then, we associate each cube center μ_i with the outcome y_i, and refer to this outcome as the *favorite* of μ_i. For brevity, let $\neg y$ denote the outcome other than y, i.e. $\neg y = (1 - y)$. For any $x \in \mathcal{X}$, we define the utility function as follows. Letting c_j be the cube that x belongs to,

$$u(x, y_j) = L\left[\frac{s}{2} - \|x - \mu_j\|_\infty\right]; \quad u(x, \neg y_j) = 0. \quad (1)$$

See Figure 1 for an illustration.

We claim that the utility function of Equation (1) is indeed L-Lipschitz with respect to any L_p norm. This is because for any cube c_i, and for any $x, x' \in c_i$, we have

$$\begin{aligned}|u(x, y_i) - u(x', y_i)| &= L\,|\|x - \mu_i\|_\infty - \|x' - \mu_i\|_\infty| \\ &\le L\|x - x'\|_\infty \le L\|x - x'\|_p.\end{aligned}$$

Moreover, for the other outcome, we have $u(x, \neg y_i) = u(x', \neg y_i) = 0$. It follows that u is L-Lipschitz within every cube. At the boundary of the cubes, the utility for any outcome is 0, and hence u is also continuous throughout $\mathcal{X}$. Because it is piecewise Lipschitz and continuous, u must be L-Lipschitz throughout $\mathcal{X}$, with respect to any L_p norm.

Next, let $\mathcal{D}$ be an arbitrary deterministic algorithm that extends classifiers on a sample to $\mathcal{X}$. We draw the sample S of size $m/2$ from P without replacement. Consider the distribution over favorites of individuals in S. Each individual in S has a favorite that is sampled independently from Bernoulli$(1/2)$. Hence, by Hoeffding's inequality, the fraction of individuals in S with a favorite of 0 is between $\frac{1}{2} - \epsilon$ and $\frac{1}{2} + \epsilon$ with probability at least $1 - 2\exp(-m\epsilon^2)$. The same holds simultaneously for the fraction of individuals with favorite 1.

Given the sample S and the utility function u on the sample (defined by the instantiation of their favorites), consider the classifier $h_{u,S}$, which maps each individual μ_i in the sample S to his favorite y_i. This classifier is clearly EF on the sample (by Theorem 1). Consider the extension $h^{\mathcal{D}}_{u,S}$ of $h_{u,S}$ to the whole of $\mathcal{X}$ as defined by algorithm $\mathcal{D}$. Define two sets Z_0 and Z_1 by letting $Z_y = \{\mu_j \notin S \mid h^{\mathcal{D}}_{u,S}(\mu_j) = y\}$, and let y_* denote an outcome that is assigned to at least half of the out-of-sample centers, i.e., an outcome for which $|Z_{y_*}| \ge |Z_{\neg y_*}|$. Furthermore, let θ denote the fraction of out-of-sample centers assigned to y_*. Note that, since $|S| = m/2$, the number of out-of-sample centers is also exactly $m/2$. This gives us $|Z_{y_*}| = \theta\frac{m}{2}$, where $\theta \ge \frac{1}{2}$.

Consider the distribution of favorites in Z_{y_*} (these are independent from the ones in the sample since Z_{y_*} is disjoint from S). Each individual in this set has a favorite sampled independently from Bernoulli$(1/2)$. Hence, by Hoeffding's inequality, the fraction of individuals in Z_{y_*} whose favorite is $\neg y_*$ is at least $\frac{1}{2} - \epsilon$ with probability at least $1 - \exp(-\frac{m}{2}\epsilon^2)$. We conclude that with a probability at least $1 - 2\exp(-m\epsilon^2) - \exp(-\frac{m}{2}\epsilon^2)$, the sample S and favorites (which define the utility function u) are such that: (i) the fraction of individuals in S whose favorite is $y \in \{0, 1\}$ is between $\frac{1}{2} - \epsilon$ and $\frac{1}{2} + \epsilon$, and (ii) the fraction of individuals in Z_{y_*} whose favorite is $\neg y_*$ is at least $\frac{1}{2} - \epsilon$.

We now show that for such a sample S and utility function u, $h^{\mathcal{D}}_{u,S}$ cannot be (α, β)-EF with respect to P for any $\alpha < 1/25$ and $\beta < L/8$. To this end, sample x and x' from P. One scenario where x envies x' occurs when (i) the favorite of x is $\neg y_*$, (ii) x is assigned to y_*, and (iii) x' is assigned to $\neg y_*$. Conditions (i) and (ii) are satisfied when x is in Z_{y_*} and his favorite is $\neg y_*$. We know that at least a $\frac{1}{2} - \epsilon$ fraction of the individuals in Z_{y_*} have the favorite $\neg y_*$. Hence, the probability that conditions (i) and (ii) are satisfied by x is at least $(\frac{1}{2} - \epsilon)|Z_{y_*}|\frac{1}{m} = (\frac{1}{2} - \epsilon)\frac{\theta}{2}$. Condition (iii) is satisfied when x' is in S and has favorite $\neg y_*$ (and hence assigned $\neg y_*$), or, if x' is in $Z_{\neg y_*}$. We know that at least a $\left(\frac{1}{2} - \epsilon\right)$ fraction of the individuals in S have the favorite $\neg y_*$. Moreover, the size of $Z_{\neg y_*}$ is $(1 - \theta)\frac{m}{2}$. So, the probability that condition (iii) is satisfied by x' is at least

$$\frac{\left(\frac{1}{2} - \epsilon\right)|S| + |Z_{\neg y_*}|}{m} = \frac{1}{2}\left(\frac{1}{2} - \epsilon\right) + \frac{1}{2}(1 - \theta).$$

Since x and x' are sampled independently, the probability that all three conditions are satisfied is at least

$$\left(\frac{1}{2} - \epsilon\right)\frac{\theta}{2} \cdot \left[\frac{1}{2}\left(\frac{1}{2} - \epsilon\right) + \frac{1}{2}(1 - \theta)\right].$$

This expression is a quadratic function in θ, that attains its minimum at $\theta = 1$ irrespective of the value of ϵ. Hence,

irrespective of $\mathcal{D}$, this probability is at least $\left[\frac{1}{2}\left(\frac{1}{2}-\epsilon\right)\right]^2$. For concreteness, let us choose ϵ to be $1/10$ (although it can be set to be much smaller). On doing so, we have that the three conditions are satisfied with probability at least $1/25$. And when these conditions are satisfied, we have $u(x, h^{\mathcal{D}}_{u,S}(x)) = 0$ and $u(x, h^{\mathcal{D}}_{u,S}(x')) = Ls/2$, i.e., x envies x' by $Ls/2 = L/8$. This shows that, when x and x' are sampled from P, with probability at least $1/25$, x envies x' by $L/8$. We conclude that with probability at least $1 - 2\exp(-m/100) - \exp(-m/200)$ jointly over the selection of the utility function u and the sample S, the extension of $h_{u,S}$ by $\mathcal{D}$ is not (α, β)-EF with respect to P for any $\alpha < 1/25$ and $\beta < L/8$.

To convert the joint probability into expected cost in the game, note that for two discrete, independent random variables X and Y, and for a Boolean function $\mathcal{E}(X, Y)$, it holds that

$$\Pr_{X,Y}(\mathcal{E}(X,Y)=1) = \mathbb{E}_X\left[\Pr_Y(\mathcal{E}(X,Y)=1)\right]. \quad (2)$$

Given sample S and utility function u, let $\mathcal{E}(u, S)$ be the Boolean function that equals 1 iff the extension of $h_{u,S}$ by $\mathcal{D}$ is not (α, β)-EF with respect to P for any $\alpha < 1/25$ and $\beta < L/8$. From Equation (2), $\Pr_{u,S}(\mathcal{E}(u, S) = 1)$ is equal to $\mathbb{E}_u\left[\Pr_S(\mathcal{E}(u, S) = 1)\right]$. The latter term is exactly the expected value of the cost, where the expectation is taken over the randomness of u. It follows that the expected cost of (any) $\mathcal{D}$ with respect to the chosen distribution over utilities is at least $1 - 2\exp(-m/100) - \exp(-m/200)$. □

4. Low-Complexity Families of Classifiers

In this section we show that (despite Theorem 3) generalization for envy-freeness is possible using much smaller samples of individuals, as long as we restrict ourselves to choosing a classifier from a family of relatively low complexity.

In more detail, two classic complexity measures are the VC-dimension (Vapnik & Chervonenkis, 1971) for binary classifiers, and the Natarajan dimension (Natarajan, 1989) for multi-class classifiers. However, to the best of our knowledge, there is no suitable dimension directly applicable to functions ranging over distributions, which in our case can be seen as $|\mathcal{Y}|$-dimensional real vectors. One possibility would be to restrict ourselves to deterministic classifiers of the type $h : \mathcal{X} \to \mathcal{Y}$. However, we have seen in Section 2 that envy-freeness is a very strong constraint on deterministic classifiers. Instead, we will consider a family $\mathcal{H}$ consisting of randomized mixtures of deterministic classifiers belonging to a family $\mathcal{G} \subset \{g : \mathcal{X} \to \mathcal{Y}\}$ of low Natarajan dimension. This allows us to adapt Natarajan-dimension-based generalization results to our setting while still working with randomized classifiers.

4.1. Natarajan Dimension Primer

Before presenting our main result, we briefly summarize the definition and relevant properties of the Natarajan dimension. For more details, we refer the reader to (Shalev-Shwartz & Ben-David, 2014).

We say that a family $\mathcal{G}$ *multi-class shatters* a set of points $x_1, \ldots, x_n$ if there exist labels $y_1, \ldots y_n$ and $y'_1, \ldots, y'_n$ such that for every $i \in [n]$ we have $y_i \neq y'_i$, and for any subset $C \subset [n]$ there exists $g \in \mathcal{G}$ such that $g(x_i) = y_i$ if $i \in C$ and $g(x_i) = y'_i$ otherwise. The Natarajan dimension of a family $\mathcal{G}$ is the cardinality of the largest set of points that can be multi-class shattered by $\mathcal{G}$.

For example, suppose we have a feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^q$ that maps each individual-outcome pair to a q-dimensional feature vector, and consider the family of functions that can be written as $g(x) = \arg\max_{y\in\mathcal{Y}} w^\top \Psi(x, y)$ for weight vectors $w \in \mathbb{R}^q$. This family has Natarajan dimension at most q.

For a set $S \subset \mathcal{X}$ of points, we let $\mathcal{G}\big|_S$ denote the restriction of $\mathcal{G}$ to S, which is any subset of $\mathcal{G}$ of minimal size such that for every $g \in \mathcal{G}$ there exists $g' \in \mathcal{G}\big|_S$ such that $g(x) = g'(x)$ for all $x \in S$. The size of $\mathcal{G}\big|_S$ is the number of different labelings of the sample S achievable by functions in $\mathcal{G}$. The following Lemma is the analogue of Sauer's lemma for binary classification.

Lemma 1 (Natarajan)**.** *For a family $\mathcal{G}$ of Natarajan dimension d and any subset $S \subset \mathcal{X}$, we have $\left|\mathcal{G}\big|_S\right| \leq |S|^d |\mathcal{Y}|^{2d}$.*

Classes of low Natarajan dimension also enjoy the following uniform convergence guarantee.

Lemma 2. *Let $\mathcal{G}$ have Natarajan dimension d and fix a loss function $\ell : \mathcal{G} \times \mathcal{X} \to [0, 1]$. For any distribution P over $\mathcal{X}$, if S is an i.i.d. sample drawn from P of size $O(\frac{1}{\epsilon^2}(d \log |\mathcal{Y}| + \log \frac{1}{\delta}))$, then with probability at least $1-\delta$ we have $\sup_{g\in\mathcal{G}} \left|\mathbb{E}_{x\sim P}[\ell(g, x)] - \frac{1}{n}\sum_{x\in S} \ell(g, x)\right| \leq \epsilon$.*

4.2. Main Result

We consider the family of classifiers that can be expressed as a randomized mixture of m deterministic classifiers selected from a family $\mathcal{G} \subset \{g : \mathcal{X} \to \mathcal{Y}\}$. Our generalization guarantees will depend on the complexity of the family $\mathcal{G}$, measured in terms of its Natarajan dimension, and the number m of functions we are mixing. More formally, let $\vec{g} = (g_1, \ldots, g_m) \in \mathcal{G}^m$ be a vector of m functions in $\mathcal{G}$ and $\alpha \in \Delta_m$ be a distribution over $[m]$, where $\Delta_m = \{p \in \mathbb{R}^m : p_i \geq 0, \sum_i p_i = 1\}$ is the m-dimensional probability simplex. Then consider the function $h_{\vec{g},\alpha} : \mathcal{X} \to \Delta(\mathcal{Y})$ with assignment probabilities given by

$$\Pr(h_{\vec{g},\alpha}(x) = y) = \sum_{i=1}^{m} \mathbb{I}\{g_i(x) = y\}\alpha_i.$$

Intuitively, for a given individual x, $h_{\vec{g},\alpha}$ chooses one of the g_i randomly with probability α_i, and outputs $g_i(x)$. Let

$$\mathcal{H}(\mathcal{G}, m) = \{h_{\vec{g},\alpha} : \mathcal{X} \to \Delta(\mathcal{Y}) \,:\, \vec{g} \in \mathcal{G}^m, \alpha \in \Delta_m\}$$

be the family of classifiers that can be written this way. Our main technical result shows that envy-freeness generalizes for this class.

Theorem 4. *Suppose $\mathcal{G}$ is a family of deterministic classifiers of Natarajan dimension d, and let $\mathcal{H} = \mathcal{H}(\mathcal{G}, m)$ for $m \in \mathbb{N}$. For any distribution P over $\mathcal{X}$, $\gamma > 0$, and $\delta > 0$, if $S = \{(x_i, x_i')\}_{i=1}^n$ is an i.i.d. sample of pairs drawn from P of size*

$$n \geq O\left(\frac{1}{\gamma^2}\left(dm^2 \log \frac{dm|\mathcal{Y}|\log(m|\mathcal{Y}|/\gamma)}{\gamma} + \log\frac{1}{\gamma}\right)\right),$$

then with probability at least $1-\delta$, every classifier $h \in \mathcal{H}$ that is (α, β)-pairwise-EF on S is also $(\alpha+7\gamma, \beta+4\gamma)$-EF on P.

Our proof of Theorem 4 consists of two steps. First, we show that envy-freeness generalizes for finite classes. Second, we show that $\mathcal{H}(\mathcal{G}, m)$ can be approximated by a finite subset.

Lemma 3. *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \Delta(\mathcal{Y})\}$ be a finite family of classifiers. For any $\gamma > 0$, $\delta > 0$, and $\beta \geq 0$ if $S = \{(x_i, x_i')\}_{i=1}^n$ is an i.i.d. sample of pairs from P of size $n \geq \frac{1}{2\gamma^2} \ln \frac{|\mathcal{H}|}{\delta}$, then with probability at least $1-\delta$, every $h \in \mathcal{H}$ that is (α, β)-pairwise-EF on S (for any α) is also $(\alpha+\gamma, \beta)$-EF on P.*

Proof. Let $f(x, x', h) = \mathbb{I}\{u(x, h(x)) < u(x, h(x')) - \beta\}$ be the indicator that x is envious of x' by at least β under classifier h. Then $f(x_i, x_i', h)$ is a Bernoulli random variable with success probability $\mathbb{E}_{x,x' \sim P}[f(x, x', h)]$. Applying Hoeffding's inequality to any fixed hypothesis $h \in \mathcal{H}$ guarantees that $\Pr_S(\mathbb{E}_{x,x'\sim P}[f(x,x',h)] \geq \frac{1}{n}\sum_{i=1}^n f(x_i, x_i', h) + \gamma) \leq \exp(-2n\gamma^2)$. Therefore, if h is (α, β)-EF on S, then it is also $(\alpha+\gamma, \beta)$-EF on P with probability at least $1 - \exp(-2n\gamma^2)$. Applying the union bound over all $h \in \mathcal{H}$ and using the lower bound on n completes the proof. □

Next, we show that $\mathcal{H}(\mathcal{G}, m)$ can be covered by a finite subset. Since each classifier in $\mathcal{H}$ is determined by the choice of m functions from $\mathcal{G}$ and mixing weights $\alpha \in \Delta_m$, we will construct finite covers of $\mathcal{G}$ and Δ_m. Our covers $\hat{\mathcal{G}}$ and $\hat{\Delta}_m$ will guarantee that for every $g \in \mathcal{G}$, there exists $\hat{g} \in \hat{\mathcal{G}}$ such that $\Pr_{x\sim P}(g(x) \neq \hat{g}(x)) \leq \gamma/m$. Similarly, for any mixing weights $\alpha \in \Delta_m$, there exists $\hat{\alpha} \in \Delta_m$ such that $\|\alpha - \hat{\alpha}\|_1 \leq \gamma$. If $h \in \mathcal{H}(\mathcal{G}, m)$ is the mixture of $g_1, \ldots, g_m$ with weights α, we let $\hat{h}$ be the mixture of $\hat{g}_1, \ldots, \hat{g}_m$ with weights $\hat{\alpha}$. This approximation has two sources of error: first, for a random individual $x \sim P$, there is probability up to γ that at least one $g_i(x)$ will disagree with $\hat{g}_i(x)$, in which case h and $\hat{h}$ may assign completely different outcome distributions. Second, even in the high-probability event that $g_i(x) = \hat{g}_i(x)$ for all $i \in [m]$, the mixing weights are not identical, resulting in a small perturbation of the outcome distribution assigned to x.

Lemma 4. *Let $\mathcal{G}$ be a family of deterministic classifiers with Natarajan dimension d, and let $\mathcal{H} = \mathcal{H}(\mathcal{G}, m)$ for some $m \in \mathbb{N}$. For any $\gamma > 0$, there exists a subset $\hat{\mathcal{H}} \subset \mathcal{H}$ of size $O\big(\frac{(dm|\mathcal{Y}|^2 \log(m|\mathcal{Y}|/\gamma))^{dm}}{\gamma^{(d+1)m}}\big)$ such that for every $h \in \mathcal{H}$ there exists $\hat{h} \in \hat{\mathcal{H}}$ satisfying:*

1. $\Pr_{x\sim P}(\|h(x) - \hat{h}(x)\|_1 > \gamma) \leq \gamma$.
2. *If S is an i.i.d. sample of individuals of size $O(\frac{m^2}{\gamma^2}(d\log|\mathcal{Y}| + \log\frac{1}{\delta}))$ then w.p. $\geq 1-\delta$, we have $\|h(x) - \hat{h}(x)\|_1 \leq \gamma$ for all but a 2γ-fraction of $x \in S$.*

Proof. As described above, we begin by constructing finite covers of Δ_m and $\mathcal{G}$. First, let $\hat{\Delta}_m \subset \Delta_m$ be the set of distributions over $[m]$ where each coordinate is a multiple of γ/m. Then we have $|\hat{\Delta}_m| = O((\frac{m}{\gamma})^m)$ and for every $p \in \Delta_m$, there exists $q \in \hat{\Delta}_m$ such that $\|p - q\|_1 \leq \gamma$.

In order to find a small cover of $\mathcal{G}$, we use the fact that it has low Natarajan dimension. This implies that the number of effective functions in $\mathcal{G}$ when restricted to a sample S' grows only polynomially in the size of S'. At the same time, if two functions in $\mathcal{G}$ agree on a large sample, they will also agree with high probability on the distribution.

Formally, let S' be an i.i.d. sample drawn from P of size $O(\frac{m^2}{\gamma^2} d \log|\mathcal{Y}|)$, and let $\hat{\mathcal{G}} = \mathcal{G}\big|_{S'}$ be any minimal subset of $\mathcal{G}$ that realizes all possible labelings of S' by functions in $\mathcal{G}$. We now argue that with probability 0.99, for every $g \in \mathcal{G}$ there exists $\hat{g} \in \hat{\mathcal{G}}$ such that $\Pr_{x\sim P}(g(x) \neq \hat{g}(x)) \leq \gamma/m$. For any pair of functions $g, g' \in \mathcal{G}$, let $(g, g') : \mathcal{X} \to \mathcal{Y}^2$ be the function given by $(g, g')(x) = (g(x), g'(x))$, and let $\mathcal{G}^2 = \{(g, g') \,:\, g, g' \in \mathcal{G}\}$. The Natarajan dimension of $\mathcal{G}^2$ is at most $2d$ (see Lemma 6 in Appendix C). Moreover, consider the loss $c : \mathcal{G}^2 \times \mathcal{X} \to \{0, 1\}$ given by $c(g, g', x) = \mathbb{I}\{g(x) \neq g'(x)\}$. Applying Lemma 2 with the chosen size of $|S'|$ ensures that with probability at least 0.99 every pair $(g, g') \in \mathcal{G}^2$ satisfies

$$\left|\mathop{\mathbb{E}}_{x\sim P}[c(g, g', x)] - \frac{1}{|S'|}\sum_{x\in S'} c(g, g', x)\right| \leq \frac{\gamma}{m}.$$

By the definition of $\hat{\mathcal{G}}$, for every $g \in \mathcal{G}$, there exists $\hat{g} \in \hat{\mathcal{G}}$ for which $c(g, \hat{g}, x) = 0$ for all $x \in S'$, which implies that $\Pr_{x\sim P}(g(x) \neq \hat{g}(x)) \leq \gamma/m$.

Using Lemma 1 to bound the size of $\hat{\mathcal{G}}$, we have that

$$|\hat{\mathcal{G}}| \leq |S'|^d |\mathcal{Y}|^{2d} = O\left(\left(\frac{m^2}{\gamma^2} d|\mathcal{Y}|^2 \log|\mathcal{Y}|\right)^d\right).$$

Since this construction succeeds with non-zero probability, we are guaranteed that such a set $\hat{\mathcal{G}}$ exists. Finally, by an identical uniform convergence argument, it follows that if S is a fresh i.i.d. sample of the size given in Item 2 of the lemma's statement, then, with probability at least $1-\delta$, every g and $\hat{g}$ will disagree on at most a $2\gamma/m$-fraction of S, since they disagree with probability at most γ/m on P.

Next, let $\hat{\mathcal{H}} = \{h_{\vec{g},\alpha} : \vec{g} \in \hat{G}^m, \alpha \in \hat{\Delta}_m\}$ be the same family as $\mathcal{H}$, except restricted to choosing functions from $\hat{\mathcal{G}}$ and mixing weights from $\hat{\Delta}_m$. Using the size bounds above and the fact that $\binom{N}{m} = O((\frac{N}{m})^m)$, we have that

$$|\hat{\mathcal{H}}| = \binom{|\hat{\mathcal{G}}|}{m} \cdot |\hat{\Delta}_m| = O\left(\frac{(dm^2|\mathcal{Y}|^2 \log(m|\mathcal{Y}|/\gamma))^{dm}}{\gamma^{(2d+1)m}}\right).$$

Suppose that h is the mixture of $g_1, \ldots, g_m \in \mathcal{G}$ with weights $\alpha \in \Delta_m$. Let $\hat{g}_i$ be the approximation to g_i for each i, let $\hat{\alpha} \in \hat{\Delta}_m$ be such that $\|\alpha - \hat{\alpha}\|_1 \leq \gamma$, and let $\hat{h}$ be the random mixture of $\hat{g}_1, \ldots, \hat{g}_m$ with weights $\hat{\alpha}$. For an individual x drawn from P, we have $g_i(x) \neq \hat{g}_i(x)$ with probability at most γ/m, and therefore they all agree with probability at least $1-\gamma$. When this event occurs, we have $\|h(x) - \hat{h}(x)\|_1 \leq \|\alpha - \hat{\alpha}\|_1 \leq \gamma$.

The second part of the claim follows by similar reasoning, using the fact that for the given sample size $|S|$, with probability at least $1-\delta$, every $g \in \mathcal{G}$ disagrees with its approximation $\hat{g} \in \hat{\mathcal{G}}$ on at most a $2\gamma/m$-fraction of S. This means that $\hat{g}_i(x) = g_i(x)$ for all $i \in [m]$ on at least a $(1-2\gamma)$-fraction of the individuals x in S. For these individuals, $\|h(x) - \hat{h}(x)\|_1 \leq \|\alpha - \hat{\alpha}\|_1 \leq \gamma$. □

Combining the generalization guarantee for finite families given in Lemma 3 with the finite approximation given in Lemma 4, we are able to show that envy-freeness also generalizes for $\mathcal{H}(\mathcal{G}, m)$.

Proof of Theorem 4. Let $\hat{\mathcal{H}}$ be the finite approximation to $\mathcal{H}$ constructed in Lemma 4. If the sample is of size $|S| = O(\frac{1}{\gamma^2}(dm \log(dm|\mathcal{Y}| \log|\mathcal{Y}|/\gamma) + \log\frac{1}{\delta}))$, we can apply Lemma 3 to this finite family, which implies that for any $\beta' \geq 0$, with probability at least $1-\delta/2$ every $\hat{h} \in \hat{\mathcal{H}}$ that is (α', β')-pairwise-EF on S (for any α') is also $(\alpha'+\gamma, \beta')$-EF on P. We apply this lemma with $\beta' = \beta + 2\gamma$. Moreover, from Lemma 4, we know that if $|S| = O(\frac{m^2}{\gamma^2}(d \log|\mathcal{Y}| + \log\frac{1}{\delta}))$, then with probability at least $1-\delta/2$, for every $h \in \mathcal{H}$, there exists $\hat{h} \in \hat{\mathcal{H}}$ satisfying $\|h(x) - \hat{h}(x)\|_1 \leq \gamma$ for all but a 2γ-fraction of the individuals in S. This implies that on all but at most a 4γ-fraction of the pairs in S, h and $\hat{h}$ satisfy this inequality for both individuals in the pair. Assume these high probability events occur. Finally, from Item 1 of the lemma we have that $\Pr_{x_1,x_2\sim P}(\max_{i=1,2} \|h(x_i) - \hat{h}(x_i)\|_1 > \gamma) \leq 2\gamma$.

Now let $h \in \mathcal{H}$ be any classifier that is (α, β)-pairwise-EF on S. Since the utilities are in $[0,1]$ and $\max_{x=x_i,x_i'} \|h(x) - \hat{h}(x)\|_1 \leq \gamma$ for all but a 4γ-fraction of the pairs in S, we know that $\hat{h}$ is $(\alpha+4\gamma, \beta+2\gamma)$-pairwise-EF on S. Applying the envy-freeness generalization guarantee (Lemma 3) for $\hat{\mathcal{H}}$, it follows that $\hat{h}$ is also $(\alpha+5\gamma, \beta+2\gamma)$-EF on P. Finally, using the fact that

$$\Pr_{x_1,x_2\sim P}\left(\max_{i=1,2} \|h(x_i) - \hat{h}(x_i)\|_1 > \gamma\right) \leq 2\gamma,$$

it follows that h is $(\alpha+7\gamma, \beta+4\gamma)$-EF on P. □

It is worth noting that the (exponentially large) approximation $\hat{\mathcal{H}}$ is only used in the generalization analysis; importantly, an ERM algorithm need not construct it. Also note that the number of outcomes $|\mathcal{Y}|$ only appears in the logarithmic terms of the sample complexity bounds, allowing these results to handle very large outcome spaces, provided that we choose a suitable family $\mathcal{G}$ of deterministic classifiers.

5. Discussion

Theorem 4 is only effective insofar as families of classifiers of low Natarajan dimension are useful. And, indeed, several prominent families have low Natarajan dimension (Daniely et al., 2012), including one vs. all (which is a special case of the example given in Section 4.1), multiclass SVM, tree-based classifiers, and error correcting output codes. However, to make our approach truly practical, two optimization-related challenges must be addressed.

First, even if a family has good learning-theoretic properties with respect to envy, it may be bad in terms of optimizing the loss function. As an extreme example, envy-freeness with respect to a family that contains only constant classifiers generalizes wonderfully (they are always EF), but classifiers in this family are likely to lead to large loss compared to the optimal envy-free classifier. So the question is this: given a sample, how close is the loss of the optimal envy-free mixture of classifiers from a given family of low Natarajan dimension (say, multiclass SVMs) to that of the optimal envy-free classifier?

Second, there is the challenge of *computing* the loss-minimizing envy-free classifier from a given family. This problem is highly nontrivial even when the classifier is deterministic and belongs to one of the simple families listed above, and the loss and utility functions are linear. In fact, in some cases, the unconstrained loss minimization problem is fairly straightforward, but the envy-freeness constraint makes the feasible set non-convex. Going forward, this is, in our view, the main obstacle to implementing our approach, and the one that should be tackled next.

References

Balcan, M.-F., Constantin, F., Iwata, S., and Wang, L. Learning valuation functions. In *Proceedings of the 25th Conference on Computational Learning Theory (COLT)*, pp. 4.1–4.24, 2012.

Brams, S. J. and Taylor, A. D. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, 1996.

Chajewska, U., Koller, D., and Ormoneit, D. Learning an agent's utility function by observing behavior. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 35–42, 2001.

Daniely, A., Sabato, S., and Shalev-Shwartz, S. Multiclass learning approaches: A theoretical comparison with implications. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 485–493, 2012.

Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In *Proceedings of the 15th Privacy Enhancing Technologies Symposium (PETS)*, pp. 92–112, 2015.

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

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21st International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 259–268, 2015.

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

Freedman, R., Schaich Borg, J., Sinnott-Armstrong, W., Dickerson, J. P., and Conitzer, V. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018. Forthcoming.

Gal, Y., Mash, M., Procaccia, A. D., and Zick, Y. Which is the fairest (rent division) of them all? *Journal of the ACM*, 64(6): article 39, 2017.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 3315–3323, 2016.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 325–333, 2016.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 656–666, 2017.

Luong, B. T., Ruggieri, S., and Turini, F. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 502–510, 2011.

Manurangsi, P. and Suksompong, W. Asymptotic existence of fair divisions for groups. *Mathematical Social Sciences*, 89:100–108, 2017.

Moulin, H. *Fair Division and Collective Welfare*. MIT Press, 2003.

Natarajan, B. K. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Nielsen, T. D. and Jensen, F. V. Learning a decision maker's utility function from (possibly) inconsistent behavior. *Artificial Intelligence*, 160(1–2):53–78, 2004.

Noothigattu, R., Gaikwad, S. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A. D. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018. Forthcoming.

Procaccia, A. D. Cake cutting: Not just child's play. *Communications of the ACM*, 56(7):78–87, 2013.

Robertson, J. M. and Webb, W. A. *Cake Cutting Algorithms: Be Fair If You Can*. A. K. Peters, 1998.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

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

Sweeney, L. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

Vapnik, V. and Chervonenkis, A. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2): 264–280, 1971.

Varian, H. Equity, envy and efficiency. *Journal of Economic Theory*, 9:63–91, 1974.

Yao, A. C. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 17th Symposium on Foundations of Computer Science (FOCS)*, pp. 222–227, 1977.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 962–970, 2017a.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., Gummadi, K. P., and Weller, A. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 228–238, 2017b.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 325–333, 2013.

A. Appendix for Section 2

Example 1. Let $S = \{x_1, x_2\}$ and $\mathcal{Y} = \{y_1, y_2, y_3\}$. Let the loss function be such that

$$\ell(x_1, y_1) = 0 \qquad \ell(x_1, y_2) = 1 \qquad \ell(x_1, y_3) = 1$$
$$\ell(x_2, y_1) = 1 \qquad \ell(x_2, y_2) = 1 \qquad \ell(x_2, y_3) = 0$$

And let the utility function be such that

$$u(x_1, y_1) = 0 \qquad u(x_1, y_2) = 1 \qquad u(x_1, y_3) = \frac{1}{\gamma}$$
$$u(x_2, y_1) = 0 \qquad u(x_2, y_2) = 0 \qquad u(x_2, y_3) = 1$$

where $\gamma > 1$. Now, the only deterministic classifier with a loss of 0 is h_0 such that $h_0(x_1) = y_1$ and $h_0(x_2) = y_3$. But, this is not EF, since $u(x_1, y_1) < u(x_1, y_3)$. Furthermore, every other deterministic classifier has a total loss of at least 1, causing the optimal deterministic EF classifier to have loss of at least 1.

To show that randomized classifiers can do much better, consider the randomized classifier h_* such that $h_*(x_1) = (1 - 1/\gamma, 1/\gamma, 0)$ and $h_*(x_2) = (0, 0, 1)$. This classifier can be seen as a mixture of the classifier h_0 of 0 loss, and the deterministic classifier h_e, where $h_e(x_1) = y_2$ and $h_e(x_2) = y_3$, which has high "negative envy". One can observe that this classifier h_* is EF, and has a loss of just $1/\gamma$. Hence, the loss of the optimal randomized EF classifier is γ times smaller than the loss of the optimal deterministic one, for any $\gamma > 1$.

Theorem 1. *A deterministic classifier $h : \mathcal{X} \to \mathcal{Y}$ is EF on sample S if and only if there exists a subset $\mathcal{Z} \subseteq \mathcal{Y}$ such that $h(x) \in \arg\max_{y \in \mathcal{Z}} u(x, y)$ for every $x \in S$.*

Proof. Suppose first that there is $\mathcal{Z} \subseteq \mathcal{Y}$ such that $h(x) \in \arg\max_{y \in \mathcal{Z}} u(x, y)$ for every $x \in S$. Now, consider two arbitrary individuals $x, x' \in S$. By definition, $u(x, h(x)) \geq u(x, y)$ for every $y \in \mathcal{Z}$. Also, $h(x') \in \mathcal{Z}$. This gives us $u(x, h(x)) \geq u(x, h(x'))$ and implies that h EF on S.

In the other direction, suppose $h : \mathcal{X} \to \mathcal{Y}$ is EF on sample S. Define $\mathcal{Z}$ as the range of h on S, i.e. $\mathcal{Z} = h(S)$. Suppose for the sake of contradiction that there exists $x_o \in S$ such that $h(x_o) \notin \arg\max_{y \in \mathcal{Z}} u(x_o, y)$. Let y_o be an arbitrary outcome in $\arg\max_{y \in \mathcal{Z}} u(x_o, y)$. This implies that $u(x_o, y_o) > u(x_o, h(x_o))$, and $y_o \in \mathcal{Z}$. But, since $\mathcal{Z}$ is the range of h, it follows that that there exists $x'_o \in S$ for which $h(x'_o) = y_o$. We conclude that $u(x_o, h(x_o)) < u(x_o, h(x'_o))$, contradicting envy-freeness. □

B. Appendix for Section 3

Theorem 2. *Let d be a metric on $\mathcal{X}$, P be a distribution on $\mathcal{X}$, and u be an L-Lipschitz utility function. Let S be a set of individuals such that there exists $\hat{\mathcal{X}} \subset \mathcal{X}$ with $P(\hat{\mathcal{X}}) \geq 1 - \alpha$ and $\sup_{x \in \hat{\mathcal{X}}}(d(x, \mathrm{NN}_S(x)) \leq \beta/(2L)$. Then for any classifier $h : S \to \Delta(\mathcal{Y})$ that is EF on S, the extension $\overline{h} : \mathcal{X} \to \Delta(\mathcal{Y})$ given by $\overline{h}(x) = h(\mathrm{NN}_S(x))$ is (α, β)-EF on P.*

Proof. Let $h : S \to \Delta(\mathcal{Y})$ be any EF classifier on S and $\overline{h} : \mathcal{X} \to \Delta(\mathcal{Y})$ be the nearest neighbor extension. Sample x and x' from P. Then, x belongs to the subset $\hat{\mathcal{X}}$ with probability at least $1 - \alpha$. When this occurs, x has a neighbor within distance $\beta/(2L)$ in the sample. Using the Lipschitz continuity of u, we have $|u(x, \overline{h}(x)) - u(\mathrm{NN}_S(x), h(\mathrm{NN}_S(x)))| \leq \beta/2$. Similarly, $|u(x, \overline{h}(x')) - u(\mathrm{NN}_S(x), h(\mathrm{NN}_S(x')))| \leq \beta/2$. Finally, since $\mathrm{NN}_S(x)$ does not envy $\mathrm{NN}_S(x')$ under h, it follows that x does not envy x' by more than β under $\overline{h}$. □

Lemma 5. *Suppose $\mathcal{X} \subset \mathbb{R}^q$, $d(x, x') = \|x - x'\|_2$, and let $D = \sup_{x, x' \in \mathcal{X}} d(x, x')$ be the diameter of $\mathcal{X}$. For any distribution P over $\mathcal{X}$, $\beta > 0$, $\alpha > 0$, and $\delta > 0$ there exists $\hat{\mathcal{X}} \subset \mathcal{X}$ such that $P(\hat{\mathcal{X}}) \geq 1 - \alpha$ and, if S is an i.i.d. sample drawn from P of size $|S| = O(\frac{1}{\alpha}(\frac{LD\sqrt{q}}{\beta})^q(d \log \frac{LD\sqrt{q}}{\beta} + \log \frac{1}{\delta}))$, then with probability at least $1 - \delta$, $\sup_{x \in \hat{\mathcal{X}}} d(x, \mathrm{NN}_S(x)) \leq \beta/(2L)$.*

Proof. Let C be the smallest cube containing $\mathcal{X}$. Since the diameter of $\mathcal{X}$ is D, the side-length of C is at most D. Let $s = \beta/(2L\sqrt{q})$ be the side-length such that a cube with side-length s has diameter $\beta/(2L)$. It takes at most $m = \lceil D/s \rceil^q$

cubes of side-length s to cover C. Let $C_1, \ldots, C_m$ be such a covering, where each C_i has side-length s.

Let C_i be any cube in the cover for which $P(C_i) > \alpha/m$. The probability that a sample of size n drawn from P does not contain a sample in C_i is at most $(1-\alpha/m)^n \leq e^{-n\alpha/m}$. Let $I = \{i \in [m] : P(C_i) \geq \alpha/m\}$. By the union bound, the probability that there exists $i \in I$ such that C_i does not contain a sample is at most $me^{-n\alpha/m}$. Setting

$$\begin{aligned} n &= \frac{m}{\alpha} \ln \frac{m}{\delta} \\ &= O\left(\frac{1}{\alpha}\left(\frac{LD\sqrt{q}}{\beta}\right)^q \left(q \log \frac{LD\sqrt{q}}{\beta} + \log \frac{1}{\delta}\right)\right) \end{aligned}$$

results in this upper bound being δ. For the remainder of the proof, assume this high probability event occurs.

Now let $\hat{\mathcal{X}} = \bigcup_{i \in I} C_i$. For each $j \notin I$, we know that $P(C_j) < \alpha/m$. Since there at most m such cubes, their total probability mass is at most α. It follows that $P(\hat{\mathcal{X}}) \geq 1 - \alpha$. Moreover, every point $x \in \hat{\mathcal{X}}$ belongs to one of the cubes C_i with $i \in I$, which also contains a sample point. Since the diameter of the cubes in our cover is $\beta/(2L)$, it follows that $\mathrm{dist}(x, \mathrm{NN}_S(x)) \leq \beta/(2L)$ for every $x \in \hat{\mathcal{X}}$, as required. □

C. Appendix for Section 4

Lemma 6. *Let $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{Y}\}$ have Natarajan dimension d. For $g_1, g_2 \in \mathcal{G}$, let $(g_1, g_2) : \mathcal{X} \to \mathcal{Y}^2$ denote the function given by $(g_1, g_2)(x) = (g_1(x), g_2(x))$ and let $\mathcal{G}^2 = \{(g_1, g_2) : g_1, g_2 \in \mathcal{G}\}$. Then the Natarajan dimension of $\mathcal{G}^2$ is at most $2d$.*

Proof. Let D be the Natarajan dimension of $\mathcal{G}^2$. Then we know that there exists a collection of points $x_1, \ldots, x_D \in \mathcal{X}$ that is shattered by $\mathcal{G}^2$, which means there are two sequences $q_1, \ldots, q_n \in \mathcal{Y}^2$ and $q'_1, \ldots, q'_n \in \mathcal{Y}^2$ such that for all i we have $q_i \neq q'_i$ and for any subset $C \subset [D]$ of indices, there exists $(g_1, g_2) \in \mathcal{G}^2$ such that $(g_1, g_2)(x_i) = q_i$ if $i \in C$ and $(g_1, g_2)(x_i) = q'_i$ otherwise.

Let $n_1 = \sum_{i=1}^D \mathbb{I}\{q_{i1} \neq q'_{i1}\}$ and $n_2 = \sum_{i=1}^D \mathbb{I}\{q_{i2} \neq q'_{i2}\}$ be the number of pairs on which the first and second labels of q_i and q'_i disagree, respectively. Since none of the n pairs are equal, we know that $n_1 + n_2 \geq D$, which implies that at at least one of n_1 or n_2 must be $\geq D/2$. Assume without loss of generality that $n_1 \geq D/2$ and that $q_{i1} \neq q'_{i1}$ for $i = 1, \ldots, n_1$. Now consider any subset of indices $C \subset [n_1]$. We know there exists a pair of functions $(g_1, g_2) \in \mathcal{G}^2$ with $(g_1, g_2)(x_i)$ evaluating to q_i if $i \in C$ and q'_i if $i \notin C$. But then we have $g_1(x_i) = q_{i1}$ if $i \in C$ and $g_1(x_i) = q'_{i1}$ if $i \notin C$, and $q_{i1} \neq q'_{i1}$ for all $i \in [n_1]$. It follows that $\mathcal{G}$ shatters $x_1, \ldots, x_{n_1}$, which consists of at least $D/2$ points. Therefore, the Natarajan dimension of $\mathcal{G}^2$ is at most $2d$, as required. □