Envy-Free Classification

Maria-Florina Balcan Travis Dick Ritesh Noothigattu Ariel D. Procaccia

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, especially when individuals have heterogeneous preferences. 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?

In this paper, we propose a new measure of algorithmic fairness. It 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., 2017). 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.

It is important to say upfront that envy-freeness is *not* suitable for several widely-studied problems where there are only two possible outcomes, one of which is 'good' and the other 'bad'; examples include predicting whether an individual would default on a loan, and whether an offender would recidivate. In these degenerate cases, envy-freeness would require that the classifier assign each and every individual the exact same probability of obtaining the 'good' outcome, which, clearly, is not a reasonable constraint.

By contrast, we are interested in situations where there is a diverse set of possible outcomes, and individuals have diverse preferences for those outcomes. For example, consider a system responsible for displaying credit card advertisements to individuals. There are many credit cards with different eligibility requirements, annual rates, and reward programs. An individual's utility for seeing a card's advertisement will depend on his eligibility, his benefit from the rewards programs, and potentially other factors. It may well be the case that an envy-free advertisement assignment shows Bob advertisements for a card with worse annual rates than those shown to Alice; this outcome is not unfair if Bob is genuinely more interested in the card offered to him. Such rich utility functions are also evident in the context of job advertisements (Datta et al., 2015): people generally want higher paying jobs, but would presumably have higher utility for seeing advertisements for jobs that better fit their qualifications and interests.

A second appealing property of envy-freeness is that its fairness guarantee binds at the level of individuals. Fairness notions can be coarsely characterized as being either individual notions, or group notions, depending on whether they provide guarantees to specific individuals, or only on average to a protected subgroup. The majority of work on

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Travis Dick <tdick@cs.cmu.edu>.

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

fairness in machine learning focuses on group fairness (Luong et al., 2011; Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Joseph et al., 2016; Zafar et al., 2017).

There is, however, one well-known example of individual fairness: the influential fair classification model of Dwork et al. (2012). The model involves a set of individuals and a set of outcomes. 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).

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.

In place of a metric, envy-freeness requires access to individuals' utility functions, but — by contrast — we do not view this assumption as a barrier to implementation. Indeed, there are a variety of techniques for learning utility functions (Chajewska et al., 2001; Nielsen & Jensen, 2004; Balcan et al., 2012). Moreover, in our running example of advertising, one can use standard measures like expected click-through rate (CTR) as a good proxy for utility.

It is 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 this solution may be severely suboptimal according to another (standard) component of our setting, the *loss function*, which, in the examples above, might represent the expected revenue from showing an ad to an individual. Typically the loss function is not perfectly aligned with individual utilities, and, therefore, it may be possible to achieve smaller loss than the naïve solution 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 with a diverse set of outcomes when individuals have heterogeneous preferences. Our goal is to understand its learning-theoretic properties.

1.1. Our Results

The challenge 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 we are aware of only one subsequent paper that takes a learning-theoretic viewpoint on individual fairness and gives theoretical guarantees (see Section 1.2).

In Section 3, we do not constrain the classifier. 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 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 theoretical 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, in Section 5, we design and implement an algorithm that learns (almost) envy-free mixtures of linear one-vs-all classifiers. We present empirical results that validate our computational approach, and indicate good generalization properties even when the sample size is small.

1.2. Related Work

Conceptually, our work is most closely related to work by Zafar et al. (2017). 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.[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.

Concurrent work by Rothblum & Yona (2018) provides generalization guarantees for the metric notion of individual fairness introduced by Dwork et al. (2012), or, more precisely, for an approximate version thereof. There are two main differences compared to our work: first, we propose envy-freeness as an alternative notion of fairness that circumvents the need for a similarity metric. Second, they focus on randomized *binary* classification, which amounts to learning a real-valued function, and so are able to make use of standard Rademacher complexity results to show generalization. By contrast, standard tools do not directly apply in our setting. It is worth noting that several other papers provide generalization guarantees for notions of group fairness, but these are more distantly related to our work (Zemel et al., 2013; Woodworth et al., 2017; Donini et al., 2018; Kearns et al., 2018; Hébert-Johnson et al., 2018).

2. The Model

We assume 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]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.

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.

Recall that we consider randomized classifiers that 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.

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 Appendix B) establishes that this approach preserves envy-freeness in cases where the sample is exponentially large.

Theorem 1. *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 1 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 $O(\frac{1}{\alpha}(\frac{LD\sqrt{q}}{\beta})^q(q \log \frac{LD\sqrt{q}}{\beta} + \log \frac{1}{\delta}))$, it will satisfy the conditions of Theorem 1 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. The proof of Theorem 2 can be found in Appendix B.

Theorem 2. *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.

4. Low-Complexity Families of Classifiers

In this section we show that (despite Theorem 2) generalization for envy-freeness is possible using much smaller samples of individuals, as long as we restrict ourselves to classifiers 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 $\left|\mathcal{G}\big|_S\right| \le |S|^d |\mathcal{Y}|^{2d}$.*

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 $\sup_{g\in\mathcal{G}} \left|\mathbb{E}_{x\sim P}[\ell(g, x)] - \frac{1}{n}\sum_{x\in S} \ell(g, x)\right| \le \epsilon$.*

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 $\eta \in \Delta_m$ be a distribution over $[m]$, where $\Delta_m = \{p \in \mathbb{R}^m : p_i \ge 0, \sum_i p_i = 1\}$ is the m-dimensional probability simplex. Then consider the function $h_{\vec{g},\eta} : \mathcal{X} \to \Delta(\mathcal{Y})$ with assignment probabilities given by $\Pr(h_{\vec{g},\eta}(x) = y) = \sum_{i=1}^m \mathbb{I}\{g_i(x) = y\}\eta_i$. Intuitively, for a given individual x, $h_{\vec{g},\eta}$ chooses one of the g_i randomly with probability η_i, and outputs $g_i(x)$. Let

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

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

Theorem 3. *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*

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

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.

Theorem 3 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.

We now turn to the theorem's proof, which 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. The proofs of both lemmas are relegated to Appendix C.

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 $\beta \ge 0$ if $S = \{(x_i, x'_i)\}_{i=1}^n$ is an i.i.d. sample of pairs from P of size $n \ge \frac{1}{2\gamma^2}\ln\frac{|\mathcal{H}|}{\delta}$, 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.*

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 $\eta \in \Delta_m$, 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 $\Pr_{x\sim P}(g(x) \neq \hat{g}(x)) \le \gamma/m$. Similarly, for any mixing weights $\eta \in \Delta_m$, there exists $\hat{\eta} \in \Delta_m$ such that $\|\eta - \hat{\eta}\|_1 \le \gamma$. If $h \in \mathcal{H}(\mathcal{G}, m)$ is the mixture of $g_1, \ldots, g_m$ with weights η, we let $\hat{h}$ be the mixture of $\hat{g}_1, \ldots, \hat{g}_m$ with weights $\hat{\eta}$. 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 $\hat{h} \in \mathcal{H}$ satisfying:*

1. *$\Pr_{x\sim P}(\|h(x) - \hat{h}(x)\|_1 > \gamma) \le \gamma$.*
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. $\ge 1 - \delta$, we have $\|h(x) - \hat{h}(x)\|_1 \le \gamma$ for all but a 2γ-fraction of $x \in S$.*

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 3. 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 $\beta' \ge 0$, 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 $\|h(x) - \hat{h}(x)\|_1 \le \gamma$ 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 $\Pr_{x_1,x_2\sim P}(\max_{i=1,2} \|h(x_i) - \hat{h}(x_i)\|_1 > \gamma) \le 2\gamma$.

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 $\max_{x=x_i,x_i'} \|h(x) - \hat{h}(x)\|_1 \le \gamma$ 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

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

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.

5. Implementation and Empirical Validation

So far we have not directly addressed the problem of *computing* the loss-minimizing envy-free classifier from a given family on a given sample of individuals. We now turn to this problem. Our goal is not to provide an end-all solution, but rather to provide evidence that computation will not be a long-term obstacle to implementing our approach.

In more detail, our computational problem is to find the loss-minimizing classifier h from a given family of randomized classifiers $\mathcal{H}$ that is envy free on a given a sample of individuals $S = \{x_1, \dots, x_n\}$. For this classifier h to generalize to the distribution P, Theorem 3 suggests that the class $\mathcal{H}$ to use is of the form $\mathcal{H}(\mathcal{G}, m)$, where $\mathcal{G}$ is a class of deterministic classifiers of low Natarajan dimension.

In this section, we let $\mathcal{G}$ to be the class of *linear one-vs-all classifiers*. In particular, denoting $\mathcal{X} \subset \mathbb{R}^q$, each $g \in \mathcal{G}$ is parameterized by $\vec{w} = (w_1, w_2, \dots, w_{|\mathcal{Y}|}) \in \mathbb{R}^{|\mathcal{Y}|\times q}$, where

$$g(x) = \operatorname*{argmax}_{y\in\mathcal{Y}} \left(w_y^\top x\right).$$

This class $\mathcal{G}$ has a Natarajan dimension of at most $q|\mathcal{Y}|$. The optimization problem to solve in this case is

$$\begin{aligned} &\min_{h\in\mathcal{H}(\mathcal{G},m)} \quad \sum_{i=1}^{n} L(x_i, h(x_i)) \\ &\text{s.t.} \quad u(x_i, h(x_i)) \ge u(x_i, h(x_j)) \quad \forall (i, j) \in [n]^2 \end{aligned}$$

Rewriting this in terms of the mixture components and weights, the exact optimization problem to solve is

$$\begin{aligned} &\min_{\vec{g}\in\mathcal{G}^m,\eta\in\Delta_m} \quad \sum_{i=1}^{n}\sum_{k=1}^{m} \eta_k L(x_i, g_k(x_i)) \\ &\text{s.t.} \quad \sum_{k=1}^{m} \eta_k u(x_i, g_k(x_i)) \ge \sum_{k=1}^{m} \eta_k u(x_i, g_k(x_j)) \quad \forall (i, j) \in [n]^2. \end{aligned} \tag{1}$$

In (1), we impose one EF constraint for every pair of individuals, rather than just the $n/2$ pairs in some matching among the individuals. We expect this to produce classifiers with lower envy in practice.

5.1. Algorithm

Observe that optimization problem (1) is highly non-convex and non-differentiable as formulated, because of the argmax computed in each of the $g_k(x_i)$. Another challenge is the combinatorial nature of the problem, as we need to find m functions from $\mathcal{G}$ along with their mixing weights. In designing an algorithm, therefore, we employ several tricks of the trade to achieve tractability.

Learning the mixture components. We first assume predefined mixing weights $\tilde{\eta}$, and *iteratively* learn mixture components based on them. Specifically, let $g_1, g_2, \dots g_{k-1}$ denote the classifiers learned so far. To compute the next component g_k, we solve the optimization problem (1) with these components already in place (and assuming no future ones). This induces the following optimization problem.

$$\begin{aligned} \min_{g_k \in \mathcal{G}} \quad & \sum_{i=1}^{n} L(x_i, g_k(x_i)) \\ \text{s.t.} \quad & USF_{ii}^{(k-1)} + \tilde{\eta}_k u(x_i, g_k(x_i)) \\ & \geq USF_{ij}^{(k-1)} + \tilde{\eta}_k u(x_i, g_k(x_j)) \quad \forall (i,j) \in [n]^2, \end{aligned} \tag{2}$$

where $USF_{ij}^{(k-1)}$ denotes the expected utility i has for j's assignments so far, i.e.,

$$USF_{ij}^{(k-1)} = \sum_{c=1}^{k-1} \tilde{\eta}_c u(x_i, g_c(x_j)).$$

Solving the optimization problem (2) is still non-trivial because it remains non-convex and non-differentiable. To resolve this, we relax the constraints,[3] and obtain the following convex problem.

$$\begin{aligned} &\min_{\vec{w} \in \mathbb{R}^{|\mathcal{Y}| \times q} \quad \sum_{i=1}^{n} \max_{y \in \mathcal{Y}} \left\{ L(x_i, y) + w_y^\top x_i - w_{y_i}^\top x_i \right\} \\ &+ \lambda \sum_{i \neq j} \max \Bigg(USF_{ij}^{(k-1)} + \tilde{\eta}_k \max_{y \in \mathcal{Y}} \left\{ u(x_i, y) + w_y^\top x_j - w_{s_i}^\top x_j \right\} \\ &- USF_{ii}^{(k-1)} + \tilde{\eta}_k \max_{y \in \mathcal{Y}} \left\{ -u(x_i, y) + w_y^\top x_i - w_{b_i}^\top x_i \right\}, 0 \Bigg), \end{aligned} \tag{3}$$

where $y_i = \text{argmin}_{y \in \mathcal{Y}} L(x_i, y)$, $s_i = \text{argmin}_{y \in \mathcal{Y}} u(x_i, y)$, $b_i = \text{argmax}_{y \in \mathcal{Y}}\, u(x_i, y)$, and λ is a parameter which defines the trade-off between loss and envy-freeness. The rationale behind this relaxation (3) is explained in more detail in Appendix D. On a very high-level, it is inspired by multi-class SVMs.

Learning the mixing weights. Once the mixture components $\vec{g}$ are learned (with respect to the predefined mixing weights $\tilde{\eta}$), we perform an additional round of optimization to learn the optimal weights η for them. This can be done via the following linear program

$$\begin{aligned} &\min_{\eta \in \Delta_m, \xi \in \mathbb{R}_{\geq 0}^{n \times n}} \quad \sum_{i=1}^{n} \sum_{k=1}^{m} \eta_k L(x_i, g_k(x_i)) + \lambda \sum_{i \neq j} \xi_{ij} \\ &\text{s.t.} \quad \sum_{k=1}^{m} \eta_k u(x_i, g_k(x_i)) \geq \sum_{k=1}^{m} \eta_k u(x_i, g_k(x_j)) - \xi_{ij} \quad \forall (i,j). \end{aligned} \tag{4}$$

5.2. Methodology

To validate our approach, we have implemented our algorithm. However, we cannot rely on standard datasets, as we need access to both the features and the utility functions of individuals. Hence, we rely on synthetic data. All our code is included as supplementary material. Our experiments are carried out on a desktop machine with 16GB memory and an Intel Xeon(R) CPU E5-1603 v3 @ 2.80GHz×4 processor. To solve convex optimization problems, we use CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018).

In our experiments, we cannot compute the optimal solution to the original optimization problem (1), and there are no existing methods we can use as benchmarks. Hence, we generate the dataset such that we know the optimal solution upfront.

Specifically, to generate the whole dataset (both training and test), we first generate random classifiers $\vec{g}^\star \in \mathcal{G}^m$ by sampling their parameters $\vec{w}_1, \dots \vec{w}_m \sim \mathcal{N}(0,1)^{|\mathcal{Y}| \times q}$, and generate $\eta^\star \in \Delta_m$ by drawing uniformly random weights in $[0,1]$ and normalizing. We use $h_{\vec{g}^\star, \eta^\star}$ as the optimal solution of the dataset we generate. For each individual, we sample each feature value independently and u.a.r. in $[0,1]$. For each individual x and outcome y, we set $L(x,y) = 0$ if $y \in \{g_k^\star(x) : k \in [m]\}$ and otherwise we sample $L(x,y)$ u.a.r. in $[0,1]$. For the utility function u, we need to generate it such that the randomized classifier $h_{\vec{g}^\star, \eta^\star}$ is envy free on the dataset. For this, we set up a linear program and compute each of the values $u(x,y)$. Overall, $h_{\vec{g}^\star, \eta^\star}$ is envy free and has zero loss, so it is obviously the optimal solution. The dataset is split into 75% training data (to measure the accuracy of our solution to the optimization problem) and 25% test data (to check for generalizability).

For our experiments, we use the following parameters: $|\mathcal{Y}| = 10$, $q = 10$, $m = 5$, and $\lambda = 10.0$. We set the predefined weights to be $\tilde{\eta} = \left[\frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^{m-1}}, \frac{1}{2^{m-1}}\right]$.[4] In our experiments we vary the number of individuals, and each result is averaged over 25 runs. On each run, we generate a new ground-truth classifier $h_{\vec{g}^*, \eta^*}$, as well as new individuals, losses, and utilities.

5.3. Results

Figure 1 shows the time taken to compute the mixture components $\vec{g}$ and the optimal weights η, as the number of individuals in the training data increases. As we will see shortly, even though the η computation takes a very small fraction of the time, it can lead to non-negligible gains in terms of loss and envy.

Figure 2 shows the average loss attained on the training and test data by the algorithm immediately after computing the mixture components, and after the round of η optimization.

[3] This may lead to solutions that are not exactly EF on the sample. Nonetheless, Theorem 3 still guarantees that there should not be much additional envy on the testing data.

[4] The reason for using an exponential decay is so that the subsequent classifiers learned are different from the previous ones. Using smaller weights might cause consecutive classifiers to be identical, thereby 'wasting' some of the components.

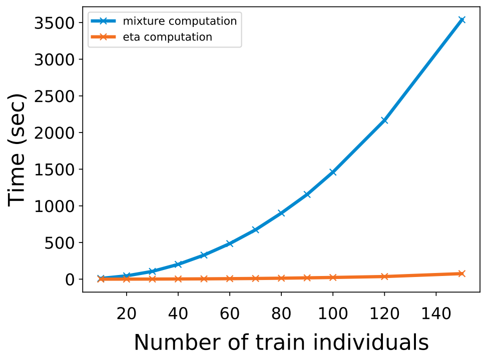

Figure 1. The algorithm's running time

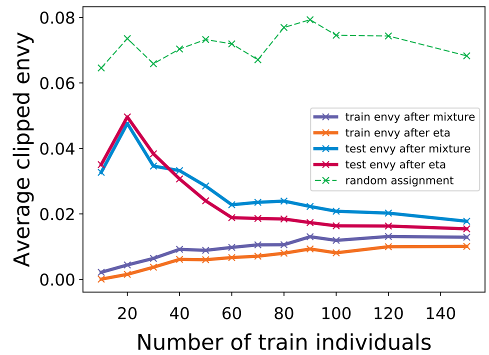

Figure 3. Training and test envy

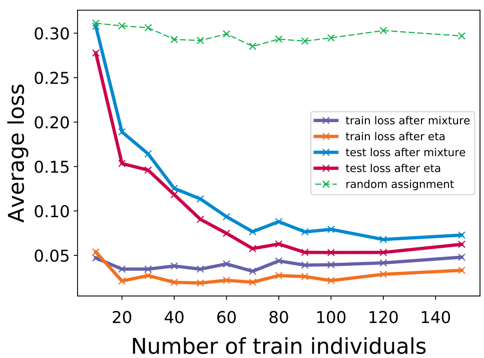

Figure 2. Training and test loss

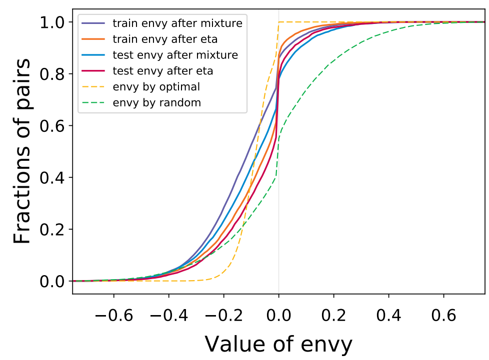

Figure 4. CDF of training and test envy for 100 training individuals

It also shows the average loss attained (on both the training and test data) by a random allocation, which serves as a naïve benchmark for calibration purposes. Recall that the optimal assignment $h_{\vec{g}^\star,\eta^\star}$ has loss 0. For both the training and testing individuals, optimizing η improves the loss of the learned classifer. Moreover, our algorithms achieve low training errors for all dataset sizes, and as the dataset grows the testing error converges to the training error.

Figure 3 shows the average envy among pairs in the training data and test data, where, for each pair, negative envy is replaced with 0, to avoid obfuscating positive envy. The graph also depicts the average envy attained (on both the training and test data) by a random allocation. As for the losses, optimizing η results in lower average envy, and as the training set grows we see the generalization gap decrease.

In Figure 4 we zoom in on the case of 100 training individuals, and observe the empirical CDF of envy values. Interestingly, the optimal randomized classifier $h_{\vec{g}^\star,\eta^\star}$ shows lower negative envy values compared to other algorithms, but as expected has no positive envy pairs. Looking at the positive envy values, we can again see very encouraging results. In particular, for at least a 0.946 fraction of the pairs in the train data, we obtain envy of at most 0.05, and this generalizes to the test data, where for at least a 0.939 fraction of the pairs, we obtain envy of at most 0.1.

In summary, these results indicate that the algorithm described in Section 5.1 solves the optimization problem (1) for linear one-vs-all classifiers almost optimally, and that its output generalizes well even when the training set is small.

6. Conclusion

In this work we propose EF as a suitable fairness notion for learning tasks with many outcomes over which individuals have heterogeneous preferences. We provide generalization guarantees for a rich family of classifiers, showing that if we find a classifier that is envy-free on a sample of individuals, it will remain envy-free when we apply it to new individuals from the same distribution. This result circumvents an exponential lower bound on the sample complexity suffered by any two-stage learning algorithm that first finds an EF assignment for the sample and then extends it to the entire space. Finally, we empirically demonstrate that finding low-envy and low-loss classifiers is computationally tractable. These results show that envy-freeness is a practical notion of fairness for machine learning systems.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Balcan, M.-F., Constantin, F., Iwata, S., and Wang, L. Learning valuation functions. In *Proc. of 25th 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 *Proc. of 18th ICML*, pp. 35–42, 2001.

Daniely, A., Sabato, S., and Shalev-Shwartz, S. Multiclass learning approaches: A theoretical comparison with implications. In *Proc. of 25th 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 *Proc. of 15th PETS*, pp. 92–112, 2015.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical Risk Minimization under Fairness Constraints. arXiv:1802.08626, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Proc. of 3rd ITCS*, pp. 214–226, 2012.

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 *Proc. of 32nd AAAI*, pp. 1636–1645, 2018.

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 *Proc. of 30th NIPS*, pp. 3315–3323, 2016.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. In *Proc. of 35th ICML*, 2018. Forthcoming.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Proc. of 30th NIPS*, pp. 325–333, 2016.

Kearns, M., Neel, S., Roth, A., and Wu, S. Computing parametric ranking models via rank-breaking. In *Proc. of 35th ICML*, 2018.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Proc. of 31st 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 *Proc. of 17th 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 *Proc. of 32nd AAAI*, pp. 1587–1594, 2018.

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.

Rothblum, G. N. and Yona, G. Probably approximately metric-fair learning. arXiv:1803.03242, 2018.

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.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Proc. of 30th COLT*, pp. 1920–1953, 2017.

Yao, A. C. Probabilistic computations: Towards a unified measure of complexity. In *Proc. of 17th FOCS*, pp. 222–227, 1977.

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 *Proc. of 31st NIPS*, pp. 228–238, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proc. of 30th 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

$$\begin{array}{lll} \ell(x_1, y_1) = 0 & \ell(x_1, y_2) = 1 & \ell(x_1, y_3) = 1 \\ \ell(x_2, y_1) = 1 & \ell(x_2, y_2) = 1 & \ell(x_2, y_3) = 0 \end{array}$$

And let the utility function be such that

$$\begin{array}{lll} u(x_1, y_1) = 0 & u(x_1, y_2) = 1 & u(x_1, y_3) = \dfrac{1}{\gamma} \\ u(x_2, y_1) = 0 & u(x_2, y_2) = 0 & u(x_2, y_3) = 1 \end{array}$$

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 $\gamma > 1$.

B. Appendix for Section 3

Theorem 1. *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 $\overline{h}$. $\square$

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 $C_1, \dots, C_m$ 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. $\square$

Theorem 2. *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 $\beta < L/8$.*

Proof. 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 $c_1, c_2, \dots c_m$, and let their centers be denoted by $\mu_1, \mu_2, \dots \mu_m$. Next, let P be the uniform distribution over the centers $\mu_1, \mu_2, \dots \mu_m$. 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 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 $y_1, y_2, \dots, y_m$ 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,

$$u(x, y_j) = L\left[\frac{s}{2} - \|x - \mu_j\|_\infty\right]; \quad u(x, \neg y_j) = 0. \quad (5)$$

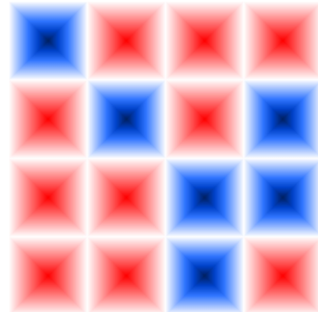

Figure 5. Illustration of $\mathcal{X}$ and an example utility function u for $d = 2$. Red shows preference for 1, blue shows preference for 0, and darker shades correspond to more intense preference. (The gradients are rectangular to match the L_∞ norm, so, strangely enough, the misleading X pattern is an optical illusion.)

See Figure 5 for an illustration.

We claim that the utility function of Equation (5) 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

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

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. 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 $\theta \geq \frac{1}{2}$.

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

$$\Pr_{X,Y}(\mathcal{E}(X, Y) = 1) = \mathbb{E}_X\left[\Pr_Y(\mathcal{E}(X, Y) = 1)\right]. \quad (6)$$

Given sample S and utility function u, let $\mathcal{E}(u, S)$ be the Boolean function that equals 1 if and only if 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 (6), $\Pr_{u,S}(\mathcal{E}(u, S) = 1)$ 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 $1 - 2\exp(-m/100) - \exp(-m/200)$. ☐

C. Appendix for Section 4

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. ☐

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 $\hat{h} \in \mathcal{H}$ 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 below). 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 $\hat{\mathcal{H}} = \{h_{\vec{g}, \eta} \,:\, \vec{g} \in \hat{G}^m, \eta \in \hat{\Delta}_m\}$ 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 $\eta \in \Delta_m$. Let $\hat{g}_i$ be the approximation to g_i for each i, let $\hat{\eta} \in \hat{\Delta}_m$ be such that $\|\eta - \hat{\eta}\|_1 \leq \gamma$, and let $\hat{h}$ be the random mixture of $\hat{g}_1, \ldots, \hat{g}_m$ with weights $\hat{\eta}$. 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 $\|h(x) - \hat{h}(x)\|_1 \leq \|\eta - \hat{\eta}\|_1 \leq \gamma$.

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, $\|h(x) - \hat{h}(x)\|_1 \leq \|\eta - \hat{\eta}\|_1 \leq \gamma$. □

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. □

D. Appendix for Section 5

Here we describe details of the transformation of the optimization problem from (2) to (3). Firstly, softening constraints of (2) with slack variables, we obtain

$$\begin{aligned}
&\min_{g_k \in \mathcal{G}, \xi \in \mathbb{R}_{\geq 0}^{n \times n}} \quad \sum_{i=1}^{n} L(x_i, g_k(x_i)) + \lambda \sum_{i \neq j} \xi_{ij} \\
&\text{s.t.} \quad USF_{ii}^{(k-1)} + \tilde{\eta}_k u(x_i, g_k(x_i)) \\
&\qquad\qquad \geq USF_{ij}^{(k-1)} + \tilde{\eta}_k u(x_i, g_k(x_j)) - \xi_{ij} \quad \forall (i, j).
\end{aligned}$$

Here, ξ_{ij} basically captures how much i envies j under the selected assignments (note that, ξ_{ij} is 0 if the pair is non-envious, so that the algorithm does not go increasing negative envy at the cost of positive envy for someone else). Plugging in optimal values of the slack variables, we obtain

$$\begin{aligned}\min_{g_k \in \mathcal{G}} \quad & \sum_{i=1}^{n} L(x_i, g_k(x_i)) \\ & + \lambda \sum_{i \neq j} \max \Big(USF_{ij}^{(k-1)} + \tilde{\eta}_k u(x_i, g_k(x_j)) \\ & \qquad\qquad -USF_{ii}^{(k-1)} - \tilde{\eta}_k u(x_i, g_k(x_i)), 0 \Big) . \end{aligned} \tag{7}$$

Next, we perform convex relaxation of different components of this objective function. For this, let's observe the term $L(x_i, g_k(x_i))$. And, let $\vec{w}$ denote the parameters of g_k. By definition, we have

$$w_{g_k(x_i)}^\top x_i \geq w_{y'}^\top x_i$$

for any $y' \in \mathcal{Y}$. This implies that

$$\begin{aligned} L(x_i, g_k(x_i)) &\leq L(x_i, g_k(x_i)) + w_{g_k(x_i)}^\top x_i - w_{y'}^\top x_i \\ &\leq \max_{y \in \mathcal{Y}} \left\{ L(x_i, y) + w_y^\top x_i - w_{y'}^\top x_i \right\} , \end{aligned}$$

giving us a convex upper bound on the loss $L(x_i, g_k(x_i))$. As this holds for any $y' \in \mathcal{Y}$, we choose $y' = y_i$ as defined in the main body, since it leads to the lowest achievable loss value. Therefore, we have

$$L(x_i, g_k(x_i)) \leq \max_{y \in \mathcal{Y}} \left\{ L(x_i, y) + w_y^\top x_i - w_{y_i}^\top x_i \right\} .$$

This right hand side is basically an upper bound which apart from encouraging $\vec{w}$ to have the highest dot product with x_i at y_i, also penalizes if the margin by which this is higher is not enough (where the margin depends on other losses $L(x_i, y)$). This surrogate loss is very similar to multi-class support vector machines. We perform similar relaxations for the other two components of the objective function. In particular, for the $u(x_i, g_k(x_i))$ term, we have

$$-u(x_i, g_k(x_i)) \leq \max_{y \in \mathcal{Y}} \left\{ -u(x_i, y) + w_y^\top x_i - w_{b_i}^\top x_i \right\} ,$$

where b_i is as defined in the main body. Finally, for the remaining term, we have

$$u(x_i, g_k(x_j)) \leq \max_{y \in \mathcal{Y}} \left\{ u(x_i, y) + w_y^\top x_j - w_{s_i}^\top x_j \right\} ,$$

where s_i is as defined in the main body[5]. On plugging in the convex surrogates of all three terms in Equation (7), we obtain the optimization problem (3).

[5]Note that, instead of using s_i, an alternative to use in this equation is b_j. In particular, for a pair (i, j), using s_i encourages the assignment to give i their favorite outcome while j the outcome that i likes the least (and hence causing i to envy j as less as possible), while using b_j encourages the assignment to give both i and j their favorite outcomes (pushing the assignment to just give everyone their favorite outcomes).