Trade-Offs between Fairness and Interpretability in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

Abstract

In this work, we look at cases where we want a classifier to be both fair and interpretable, and find that it is necessary to make trade-offs between these two properties. We have theoretical results to demonstrate this tension between the two requirements. More specifically, we consider a formal framework to build simple classifiers as a means to attain interpretability, and show that simple classifiers are strictly improvable, in the sense that every simple classifier can be replaced by a more complex classifier that strictly improves both fairness and accuracy.

1 Introduction

Machine learning based decision systems have become commonplace, and are increasingly being used to make consequntial decisions. One would like these systems to be *fair*, and not discriminate against someone just because of their membership in a minority/protected group (which could be a particular race, gender, etc.). We would also like the system to be *interpretable*, what that intuitively means is that we would like to be able to understand how it works and convincingly explain any decisions it might make.

Creating models that are intuitively simple to humans is a natural strategy to increase their interpretablilty. For example, one could avoid using complex models such as deep neural networks, and instead use simple models such as linear classifiers. Another way to build simple classifiers is to reduce the number of features that are involved in the decision making process, by choosing a small number of the most informative features, or deleting unfair features [Grgić-Hlača *et al.*, 2018]. We consider the formal framework to model the construction of simple classifiers proposed in Kleinberg and Mullainathan [2019], which captures some commonly used methods of building interpretable models. We discuss the interaction between the desiderata of simplicity (interpretability), fairness and accuracy of binary classifiers in this framework.

Given a set of features, we have an optimal classifier (i.e., the most accurate classifier that can be built from the given features). One may wish to simplify the optimal classifier to increase interpretability, or to even increase fairness in some cases. Simpler models can be easier to audit, and we can possibly identify sources of unfairness and correct them with more ease in them [Doshi-Velez and Kim, 2017]. Deleting features that can be potentially viewed as unfair, has also been adopted in practice, for example, in the well known "ban the box" scenario, where the check box in hiring applications that asks if applicants have a criminal record, is removed [Doleac and Hansen, 2016].

In contrast, this work discusses the negative effects of building simple classifiers on their fairness. More specifically, we show that every simple classifier can be improved; i.e., replaced by a more complex classifier that strictly increases both fairness and accuracy with respect to the simple classifier. It is quite expected that using a simple model would result in a loss in accuracy, because imposing simplicity requirements on a classifier reduces its expressive power. The surprising finding here is that simplification leads to a loss in fairness as well, i.e., we can always find a more complex classifier that is more fair, in fact, we can always find a more complex classifier that is simultaneously more fair and accurate than the simple classifier. Hence, we see that that the properties of fairness and accuracy clash with interpretability (or simplicity).

2 Related Work

Although the ethical issues concerning algorithms that we discuss in this work have been considered widely in the now ubiquitous literature on model interpretability and algorithmic fairness, they have mostly been considered in isolation. In particular, the literature on algorithmic fairness discusses how to handle issues such as bias and discrimination [Dwork *et al.*, 2012; Kleinberg *et al.*, 2017; Feldman *et al.*, 2015], and the literature on model interpretability addresses the growing need for transparent models [Doshi-Velez and Kim, 2017; Rudin, 2019; Lipton, 2018].

Not much previous work has looked at cases where one would want to satisfy multiple of these properties simultaneously, or analysed how these properties interact. Previously, Doshi-Velez and Kim [2017] argued that increasing a model's interpretability makes the model easier to analyse, and therefore assists in (a) deciding whether the model is fair and (b) modifying the model to ensure that it is. In contrast to their work, our first main result captures the fact that the interpretability of a model could be at odds with the fairness of the model.

Our main result is similar in spirit to the main statement of Kleinberg and Mullainathan [2019]. However, our setup has some key differences and enjoys multiple advantages. We remove a restrictive assumption they make on the data distribution, and our notion of fairness is different. We will compare our work to Kleinberg and Mullainathan [2019] in greater detail later.

3 Formalising the Framework

We denote the domain set by $\mathcal{X}$. There is an underlying distribution $\mathcal{D}$ over $\mathcal{X}$. We assume the existence of a *Ground Truth* function, that assigns a label to each point in the domain set, that is,

$$\mathcal{G} : \mathcal{X} \to \{0, 1\}.$$

For example, in the case where a bank needs to classify loan applicants, a person in the domain set is assigned the label 1 if they would return the loan, and 0 otherwise. In general, we refer to an instance labeled 1 as *good*, and *bad* otherwise.[1]

Features Each instance in $\mathcal{X}$ is represented by the set of features $\mathcal{F} = \{f_1, \ldots, f_k\}$. For example, in the bank loans case, the features could be things like credit score, income, and so on. Each instance also belongs to one of two groups - A or D. A stands for the advantaged group, whereas D stands for the disadvantaged group. D can be thought of as the minority group that we wish to protect from discrimination. The group membership feature $f_m : \mathcal{X} \to \{A, D\}$ maps an instance to their group. For simplicity, we assume that each $f_i : \mathcal{X} \to \{0, 1\}$ is a binary feature.[2]

Task Given an unlabeled set of applicants generated by the underlying distribution, we want to build a classifier to admit a fixed fraction r (known as admission rate) of them, such that we are as accurate as possible (i.e., admit as many good applicants as possible).

Partitions and Cells We can partition the domain set $\mathcal{X}$ into different parts, and we call each part a *cell*. A natural way to create cells is based on their feature vectors. That is, two instances are part of the same cell if and only if they have the same feature vector representation.

Recall, we are given access to a set of features $\mathcal{F} = \{f_1, \ldots, f_k\}$. We also had the group membership feature f_m and if we append that to the feature set $\mathcal{F}$, we denote the resultant feature set by $\mathcal{F}'$. The partition induced by $\mathcal{F}$ is denoted by f, and we denote the cells of f by $\mathcal{C}_1, \ldots, \mathcal{C}_n$, (where $n = 2^k$, because each feature is binary). The partition induced by $\mathcal{F}'$ is denoted by f', and consists of 2^{k+1} cells, as there are $k+1$ binary features. The cells in f' are obtained by splitting each cell in f into two parts, according to the group membership feature f_m. For e.g., $\mathcal{C}_1$ is split into $\mathcal{C}_1^A$ and $\mathcal{C}_1^D$, which represent the advantaged and disadvantaged people in the cell $\mathcal{C}_1$ respectively.

Score Function We say that the probability of a random instance sampled according to $\mathcal{D}$ being good (given that it lies in some cell $\mathcal{C}$) is the *score* of $\mathcal{C}$. We denoted the score of $\mathcal{C}$ by $\mathcal{S}(\mathcal{C})$, i.e.,

$$\mathcal{S}(C) = \Pr_{x \sim \mathcal{D}}[\mathcal{G}(x) = 1 \mid x \in \mathcal{C}]$$

By score of an instance $x \in \mathcal{X}$, we mean the score of the cell it belongs to in the partition f'. Given the feature set, and the fact that we only have access to the features of any instance, the score of an instance is the most accurate estimate we can have of the probability of the instance being good.

4 Classifiers

A classifier assigns every point in the domain set a label from $\{0, 1\}$. Because each point in the domain set is represented by its feature vector, the classifier is essentially a function from the space of all feature vectors to the label set., i.e., from $\{0, 1\}^{k+1} \to \{0, 1\}$. A given partition h of the domain set and admission rate r induces a threshold classifier that we denote by h_r. The classifier h_r sorts the cells of h in descending order of their scores (after merging together cells with the same score). We then admit applicants in this order until we admit the desired fraction r. We use the terms classifier and partition interchangeably.

Recall that we had discussed the partition f' above, which is the partition induced by all the features we have. Given the feature set we have, the most accurate classifier we can construct is the one induced by the partition f'.

4.1 Formal Description of Classifier

A given partition h of the domain set and admission rate r induces a threshold classifier that we denote by h_r. Consider an arbitrary partition h which partitions $\mathcal{X}$ into the cells $\mathcal{C}_1^{\wedge}, \mathcal{C}_2^{\wedge}, \ldots \mathcal{C}_t^{\wedge}, \ldots, \mathcal{C}_d^{\wedge}$. We sort the cells of h in descending order of their scores. Without loss of generality assume that h partitions $\mathcal{X}$ into cells $\mathcal{C}_1^{\wedge}, \mathcal{C}_2^{\wedge}, \ldots, \mathcal{C}_d^{\wedge}$ with decreasing (not necessarily strict) order of scores. We merge cells with the same scores to form a new partition h^* with cells $\mathcal{C}_1^*, \mathcal{C}_2^*, \ldots, \mathcal{C}_{d'}^*$, in strictly decreasing order of scores. Now, start admitting applicants in order as follows until you admit a fraction r of them. Let r_j be the fraction of the first j cells of h^* in the order they are represented. If $j(r)$ is the unique index j such that $r_{j-1} \leq r < r_j$, then the instances admitted consist of all the applicants in the cells $\mathcal{C}_1^*, \mathcal{C}_2^*, \ldots, \mathcal{C}_{j(r)-1}^*$, together with a subset of $\mathcal{C}_{j(r)}^*$ of fraction $(r - r_{j-1})$. The instances in $\mathcal{C}_{j(r)}^*$ that have to be admitted will be picked randomly.

4.2 Modeling Simple Classifiers

We use the framework introduced in Kleinberg and Mullainathan [2019] to model the construction of simple classifiers. Two particular approaches to build simple classifiers that this framework captures are (i) shallow decision trees, and (ii) using a small number of informative features (feature

[1] Our results and proofs also go through for the case where the ground truth function $\mathcal{G}$ is non-deterministic, that is, instead of being labeled 0 or 1, a particular instance might be labeled 0 with probability 0.6, and 1 with probability 0.4. However, for simplicity, we assume that the ground truth function is deterministic. If we allow $\mathcal{G}$ to be non-deterministic, the underlying distribution $\mathcal{D}$ would be over $\mathcal{X} \times \{0, 1\}$, not $\mathcal{X}$.

[2] However, the results, and pretty much the same proofs also hold for the case when each feature can take finitely many values.

selection). Both these approaches follow a common principle: they simplify the underlying model by combining distinguishable instances (applicants with different feature vector representations) together into larger sets and making a common decision at the level of each set. Previously, the instances were in different sets, and were therefore potentially treated differently. What that means in our framework, is that we would simplify f' (or in general, any partition) by combining multiple cells of it together into one larger cell, to result in a simpler partition, with fewer cells. We define a simplification formally below.

Definition (Simplification)**.** A partition w of a set $\mathcal{X}$ is a simplification of partition v of $\mathcal{X}$ if every cell of v is a subset of some cell of w, and $w \neq v$.

Definition (Non-trivial simplification)**.** A simplification w of f' is non-trivial if it contains at least one cell $\mathcal{C}$ such that $\mathcal{C}$ contains at least two cells of f' with different scores. If a simplification of f' only combines together cells with the same scores, the partition-induced classifier remains unchanged, and hence, such a simplification is not very meaningful.

Definition (Non-trivial cell)**.** We say that such a cell $\mathcal{C}$ as above is a non-trivial cell.

Structured Simplifications

The approaches to building simple classifiers that the framework captures, which are (i) shallow decision trees, and (ii) feature selection, do not combine cells at random, but they do it in a constrained way. For example, f is the simplification of f' associated with deleting the group membership feature f_m. Deleting a feature (a way to implement feature selection) is a specific form of simplification that halves the number of cells. We define two other forms of structured simplification in this work, namely, group agnostic simplifications (defined below) and graded simplifications (defined later).

Definition (Group Agnostic Simplification)**.** A simplification of f' such that instances differing only in the group membership feature are mapped to the same cell. This basically means that as a simplification step, the classifier is constrained at the very least to completely ignore/delete the group membership feature. There may or may not be further simplification steps on top of this.

4.3 Evaluating Classifiers

Fairness Most notions of fairness just tell us when a classifier is unfair, but do not quantify the amount of unfairness. The few notions that have been proposed previously to quantify unfairness have essentially been of the following form: Unfairness = $|R(A) - R(D)|$, which is the absolute value in the difference of some quantity R (such as error rate, or false positive/negative rates), between the advantaged group A and disadvantaged group D. Such notions support the viewpoint that false positive rates, say 1 and 0 for A and D respectively is as unfair as false positive rates 0 and 1 for A and D respectively. We take a different viewpoint, and believe that the former should be considered more unfair than the latter.

We propose a notion more in line with *affirmative action*, that actively supports the disadvantaged group. We penalise the decisions that unfairly hurt the disadvantaged group (false negatives of the disadvantaged group, and false positives of the advantaged group) with a high weight. We also penalise the decisions that unfairly hurt the advantaged group (false negatives of the advantaged group, and false positives of the disadvantaged group), but with a lower weight than the decisions that hurt the disavantaged group. Our fairness objective function penalises FP_A (False Positives for group A), FN_D (False Negatives for group D), FP_D (False Positives for group D), and FN_A (False Negatives for group A), and aims to minimise a weighted sum of them.

$$FP_A(h_r) = \mathbb{E}(\text{Fraction of bad instances in } A \text{ that } h_r \text{ accepts})$$
$$FN_D(h_r) = \mathbb{E}(\text{Fraction of good instances in } D \text{ that } h_r \text{ rejects})$$
$$FP_D(h_r) = \mathbb{E}(\text{Fraction of bad instances in } D \text{ that } h_r \text{ accepts})$$
$$FN_A(h_r) = \mathbb{E}(\text{Fraction of good instances in } A \text{ that } h_r \text{ rejects})$$

$$\begin{aligned}\text{Unfairness}(h_r) = \mathbb{E}[&\alpha(FN_D(h_r)) + \beta(FP_A(h_r)) \\ &+ \gamma(FN_A(h_r)) + \delta(FP_D(h_r))]\end{aligned}$$

where $\gamma \leq \alpha$, and $\delta \leq \beta$.

$$\text{Fairness}(h_r) = -\text{Unfairness}(h_r)$$

When $\gamma = \alpha = \delta = \beta$, optimising for fairness is the same as optimising for accuracy. On the other extreme, When $\gamma = 0$, and $\delta = 0$, the notion of fairness basically only penalises the decisions that hurt the disadvantaged group, and hence can be seen as an extreme case of affirmative action. Between these two extremes, the fairness notion aims to balance accuracy with affirmative action, and one could tweak the weights as per the given situation.

Accuracy

$$\text{Accuracy}(h_r) = \frac{\mathbb{E}(\text{Fraction of good instances } h_r \text{ accepts})}{\text{Total fraction of instances } h_r \text{ accepts (i.e., } r)}$$

Comparing Two Classifiers

Consider two partitions of $\mathcal{X}$, say h and g. We say that a partition h *improves* on partition g in criteria Q (e.g., accuracy) if for every $r \in [0, 1]$, $Q(h_r)$ is at least $Q(g_r)$. We say that a partition h *strictly improves* on partition g in criteria Q (e.g., accuracy) if for every $r \in [0, 1]$, $Q(h_r)$ is at least $Q(g_r)$, and there exists an $r' \in [0, 1]$ such that $Q(h_{r'})$ is strictly more than $Q(g_{r'})$.

5 Result

5.1 Group Agnostic Case

We first consider the case where we restrict simplifications to group agnostic ones. We informally explain the result of this section. Recall that the classifier resulting from partition f' is the most accurate classifier we can build with the features we have. If we choose to use a simpler classifier than f', say w, it might lead to an increase in interpretability, or fairness, but we lose accuracy. That might have been a good tradeoff, but we show that the simple classifier w is not optimal if we ignore the requirement of interpretability, as there exists a partition h (achievable by the features we have)[3] that is simultaneously more fair, and accurate, than w. Therefore, we would strictly prefer h over w, if we ignore interpretability requirements, and therefore we see that interpretability clashes with the desiderata of fairness, and accuracy.

Theorem 1. *For every non-trivial group-agnostic simplification of f', say w, there exists a classifier h that simultaneously strictly improves both accuracy and fairness over w.*

Assumptions on the Data Distribution

Before moving on to the proof, we list the assumptions we use (also used by Kleinberg and Mullainathan [2019]).

1. *Equality assumption:* For every cell $\mathcal{C}_i \in f$, if we split it by group membership, both resultant cells $\mathcal{C}_i^A$ and $\mathcal{C}_i^D$ have the same score. This intuitively means that if we have enough informative features about a person, their membership in a protected group does not affect their performance.
2. *Denseness assumption:* We denote the cells of f' by $\mathcal{C}'_1, \ldots, \mathcal{C}'_{2n}$. $\mathcal{C}^A$ denotes the instances of cell $\mathcal{C}$ that are advantaged. Similarly, $\mathcal{C}^D$ denotes the disadvantaged instances of cell $\mathcal{C}$. For every cell $\mathcal{C}_i \in f$, if we split it by group membership, both resultant cells $\mathcal{C}_i^A$ and $\mathcal{C}_i^D$ have positive measure (The measure of a cell $\mathcal{C}$, denoted by $\mu(C)$, is the mass of the probability distribution $\mathcal{D}$ in cell $\mathcal{C}$). This intuitively means that there exist people in both groups A and D exhibiting every feature vector.
3. *Genericity assumption:* Let $R, T \subseteq f'$ be two distinct sets of cells such that if $R = \mathcal{C}_i^A$ then $T \neq \mathcal{C}_i^D$. We then assume that $\mathcal{S}(R) \neq \mathcal{S}(T)$ (For a set of cells $R \subseteq f'$, use $\mathcal{S}(R)$ to denote the weighted average value of S in the cells of R).

   Remark. This in particular implies that the cells of f can be arranged in strictly descending order of scores. Without loss of generality, we assume that $\mathcal{S}(\mathcal{C}_1) > \mathcal{S}(\mathcal{C}_2) > \cdots > \mathcal{S}(\mathcal{C}_n)$.

Proof of Theorem 1. Consider non-trivial group-agnostic simplification w of f'. It partitions $\mathcal{X}$ into the cells

$$\mathcal{C}_1^\wedge, \mathcal{C}_2^\wedge, \cdots, , \mathcal{C}_j^\wedge, \mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge, \ldots, \mathcal{C}_d^\wedge$$

with descending order of scores. Take a non trivial cell of w, say $\mathcal{C}_t^\wedge$. The non trivial cell $\mathcal{C}_t^\wedge$ consists of two or more cells of f with different scores. Say $\mathcal{C}_t^\wedge$ is the union of $\mathcal{C}_a, \mathcal{C}_b, \ldots, \mathcal{C}_z \in f$. Let the cell of f in $\mathcal{C}_t^\wedge$ with the highest score be $\mathcal{C}_b$.

Construct h as follows: Remove $\epsilon > 0$ measure of $\mathcal{X}$ from $\mathcal{C}_b^D$ to create a separate cell $\mathcal{C}'$. This is the new partition h. Denote the remainder of $\mathcal{C}_t^\wedge$ by $\mathcal{C}''$. Observe that $\mathcal{S}(\mathcal{C}') > \mathcal{S}(\mathcal{C}_t^\wedge) > \mathcal{S}(\mathcal{C}'')$. Take ϵ small enough to not change order of $\mathcal{C}''$ in the partition w. It should be in the same position as $\mathcal{C}_t^\wedge$ was before. (we can do this because of the genericity assumption) The only change in the order is that $\mathcal{C}'$ jumps to some position ahead of $\mathcal{C}''$. The new partition h is

$$\mathcal{C}_1^\wedge, \mathcal{C}_2^\wedge, \cdots, \mathcal{C}_j^\wedge, \mathcal{C}', \mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_{t-1}^\wedge, \mathcal{C}'', \mathcal{C}_{t+1}^\wedge \cdots, \mathcal{C}_d^\wedge$$

with descending order of scores.

Remark. Removing $\epsilon > 0$ measure of a cell to create a separate new cell can be viewed as randomising over instances in that cell. Each instance goes to the new cell with probability ϵ, and stays in the old cell with probability $1 - \epsilon$.

We can show that for all rates r, the fairness, and accuracy of h is at least as good as w, and for at least one value of r, strictly better in both criteria. Let r_j be the fraction of the first j cells of a partition in the order they are represented.

Case 1

$r \geq r_t$ or $r \leq r_j$:

We note that in h, the measure of all cells uptil $\mathcal{C}''$ is r_t. The classifiers resulting from w and h with admission rate r as above classify all cells the same way. Therefore, h_r has the same accuracy, and fairness as w_r.

Case 2

$r_j + \mu(\mathcal{C}') \geq r > r_j$:

Both h_r and w_r classify all instances of $\mathcal{C}_1^\wedge, \ldots \mathcal{C}_j^\wedge$ as 1. The admission rule h_r classifies instances of $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_{t-1}^\wedge$ as 0 and some mass $\mu = r - r_j$ of $\mathcal{C}'$ as 1, while the admission rule w_r classifies some mass μ of $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge$ as 1, and the remaining as 0 (we start by classifying instances from from $\mathcal{C}_{j+1}^\wedge$ as 1, if $\mu(\mathcal{C}_{j+1}^\wedge) < \mu$, then we move on to $\mathcal{C}_{j+2}$, and so on). Since the score of $\mathcal{C}'$ is greater than the score of each cell $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge$, the mass μ of $\mathcal{C}'$ that h_r classifies as 1 has a higher measure of expected true 1's than the mass μ of $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge$ that w_r classifies as 1. Therefore, h_r is in expectation more accurate than w_r.

The mass μ of $\mathcal{C}'$ that h_r classifies as 1 has a higher measure of disadvantaged instances than the mass μ of $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge$ that w_r classifies as 1 because $\mathcal{C}'$ only consists of disadvantaged instances, while each cell in $\mathcal{C}_{j+1}^\wedge, \cdots \mathcal{C}_t^\wedge$ consists of both disadvantaged and advantaged instances (because of the denseness assumption). It is easy to see that h_r on expectation has lower FP_A and FN_D values than w_r. The increase in FP_D and FN_A values is not more than the amount of decrease in FP_A and FN_D values respectively. Hence, h_r has higher fairness than w_r.

Case 3

$r_t > r \geq r_j + \mu(\mathcal{C}')$:

Both h_r and w_r classify all instances of $\mathcal{C}_1^\wedge, \ldots \mathcal{C}_j^\wedge$ as 1 and all instances of $\mathcal{C}_{t+1}^\wedge, \ldots \mathcal{C}_d^\wedge$ as 0. h_r classifies all instances of $\mathcal{C}'$ as 1, while w_r classifies some mass μ of them as 0 and instead classifies some mass μ from $\mathcal{C}_{j+1}^\wedge, \ldots \mathcal{C}''$ with score lower than that of $\mathcal{C}'$ as 1. This is where the two classifiers differ. Cells $\mathcal{C}_{j+1}^\wedge, \ldots \mathcal{C}''$ have a lower score and lesser proportion of disadvantaged instances than $\mathcal{C}'$. Reasoning similarly as Case 2, we observe that w_r is less fair, and less accurate than h_r.

□

[3] If we do not require partition h to be achievable with the features we have, it is trivial to find an h that strictly improves in fairness and accuracy over any w (where w is a non-trivial simplification of f'). For example, the following partition would work: $h = \mathcal{C}_1^\wedge, \mathcal{C}_2^\wedge, \mathcal{C}_3^\wedge, \mathcal{C}_4^\wedge$, where $\mathcal{C}_1^\wedge =$ Good instances in D, $\mathcal{C}_2^\wedge =$ Good instances in A, $\mathcal{C}_3^\wedge =$ Bad instances in D, $\mathcal{C}_4^\wedge =$ Bad instances in A. Here we ensure to not merge any cells in h while admitting instances.

5.2 Differences with Respect to Previous Work

As mentioned before, our setup enjoys multiple advantages over Kleinberg and Mullainathan [2019].

1. The following assumption on the data distribution below, which is quite restrictive, is used by Kleinberg and Mullainathan [2019], but we do not use it for our results.

   Disadvantage assumption: Given cells $\mathcal{C}_i, \mathcal{C}_j \in f$ such that $\mathcal{S}(\mathcal{C}_i) < \mathcal{S}(\mathcal{C}_j)$, then

   $$\frac{\mu(\mathcal{C}_i^A)}{\mu(\mathcal{C}_i^D)} < \frac{\mu(\mathcal{C}_j^A)}{\mu(\mathcal{C}_j^D)}.$$

   This condition intuitively means that for every two feature vectors a and b such that instances having feature vector representation a have a higher chance of success than instances having feature vector representation b, instances having feature vector representation a have a higher chance of belonging to the advantaged group than instances having feature vector representation b.

2. They use the notion of equity (defined below) to quantify the fairness of a classifier, which essentially involves maximizing the number of minority group applicants the classifier labels positively.

   $$\text{Equity}(h_r) = \frac{\mathbb{E}(\text{Fraction of instances in } D \text{ that } h_r \text{ accepts})}{\text{Total fraction of instances } h_r \text{ accepts (i.e., } r)}$$

   Our notion of fairness is more aligned with accuracy. We believe that a desirable property of any notion of fairness is that a classifier that is perfectly accurate is also perfectly fair, which is something our notion satisfies but theirs does not. In addition, we also prove similar tradeoff results for equity (Theorems 2, 3, 4).

Adding Equity to Theorem 1

If we additionally consider the notion of equity in the scenario of group agnostic simplifications as in Theorem 1, we get the result below.

Theorem 2. *For every non-trivial group-agnostic simplification of f', say w, there exists a classifier h that simultaneously strictly improves accuracy, fairness, and equity over w.*

Proof. The same classifier h as constructed in the proof of Theorem 1 works for this as well.

□

5.3 General Case

Now we move on from group-agnostic simplifications to a more general notion of simplification, called graded simplification. Note that in group-agnostic simplifications, we constrained the classifier to always ignore/delete the protected group membership feature. Graded simplifications are more general, and do not suffer from this constraint.

Definition (Graded-simplification)**.** Consider cell partition f' of $\mathcal{X}$: $\mathcal{C}'_1, \mathcal{C}'_2, \ldots, \mathcal{C}'_{2n}$. Consider simplification w of f' that partitions $\mathcal{X}$ into the cells $\mathcal{C}_1^{\wedge}, \mathcal{C}_2^{\wedge}, \ldots, \mathcal{C}_t^{\wedge}, \ldots, \mathcal{C}_d^{\wedge}$ with descending order of scores. Each cell $\mathcal{C}_i^{\wedge} \in w$ can be written as $\mathcal{C}_i^{\wedge} = \cup_{j=1}^{k} \mathcal{C}'_{i_j}$ (i.e., the union of some cells $\mathcal{C}'_{i_1}, \mathcal{C}'_{i_2}, \ldots, \mathcal{C}'_{i_k} \in f'$). We denote the set of indices $\{i_1, i_2 ... i_k\}$ corresponding to $\mathcal{C}_i^{\wedge}$ as $V(\mathcal{C}_i^{\wedge})$.

A graded simplification w of f' is one where each cell $\mathcal{C}_i^{\wedge} \in w$ has the property that either $V(\mathcal{C}_i^{\wedge^A}) \subseteq V(\mathcal{C}_i^{\wedge^D})$ or $V(\mathcal{C}_i^{\wedge^D}) \subseteq V(\mathcal{C}_i^{\wedge^A})$.

Result

We first informally explain the result of this section. If we use a simpler classifier than f', say w, it might lead to an increase in fairness, interpretability, or equity, but we lose accuracy. We show that the simple classifier w is not optimal if we ignore the requirement of interpretability, as there exists a partition h (achievable by the features we have) that is simultaneously both more fair and accurate than w, while also improving equity. Therefore, we would strictly prefer h over w, if we ignore interpretability requirements, and therefore we see that interpretability clashes with the desiderata of fairness, accuracy, and equity.

Remark. Unlike Theorem 2, the partition h does not guarantee an increase in equity. This makes sense, as we are now considering a more general notion of simplification.

Theorem 3. *For every non-trivial graded-simplification, say w, there exists a partition h that simultaneously strictly improves accuracy and fairness, while also improving equity, with respect to w.*

Proof. Consider simplification w. It partitions $\mathcal{X}$ into the cells $\mathcal{C}_1^{\wedge}, \mathcal{C}_2^{\wedge}, \ldots \mathcal{C}_t^{\wedge}, \ldots, \mathcal{C}_d^{\wedge}$ with descending order of scores. Take a non trivial cell of w, say $\mathcal{C}_t^{\wedge}$. Say $\mathcal{C}_t^{\wedge}$ is the union of $\mathcal{C}_a, \mathcal{C}_b, \ldots, \mathcal{C}_z \in f'$.

Case 1: $V(\mathcal{C}_t^{\wedge^A}) \subseteq V(\mathcal{C}_t^{\wedge^D})$

There exists a cell $\mathcal{C}_a$ such that $\mathcal{C}_a \in f'$, $\mathcal{C}_a \subset \mathcal{C}_t^{\wedge}$, such that $\mathcal{C}_a$ has the highest score amongst all cells $\mathcal{C}_a, \mathcal{C}_b, \ldots, \mathcal{C}_z \subset \mathcal{C}_t^{\wedge}$ and only consists of disadvantaged instances.

Construct h as follows: Remove $\epsilon > 0$ mass of $\mathcal{X}$ from $\mathcal{C}_a$ to create a separate cell $\mathcal{C}'$. Denote the remainder of $\mathcal{C}_t^{\wedge}$ by $\mathcal{C}''$. Observe that $\mathcal{S}(\mathcal{C}') > \mathcal{S}(\mathcal{C}_t^{\wedge}) > \mathcal{S}(\mathcal{C}'')$. Take ϵ small enough to not change order of $\mathcal{C}''$ in the partition w (we can do this because of the genericity assumption). It should be in the same position as $\mathcal{C}_t^{\wedge}$ was before. The only change in the order is that $\mathcal{C}'$ jumps to some position ahead of $\mathcal{C}''$.

The new partition h is

$$\mathcal{C}_1^{\wedge}, \mathcal{C}_2^{\wedge}, \cdots, \mathcal{C}_j^{\wedge}, \mathcal{C}', \mathcal{C}_{j+1}^{\wedge}, \cdots \mathcal{C}_{t-1}^{\wedge}, \mathcal{C}'', \mathcal{C}_{t+1}^{\wedge} \cdots, \mathcal{C}_d^{\wedge}$$

with descending order of scores.

Similar to the proof of Theorem 1, it is easy to check that for all rates r, the fairness and accuracy of h is at least as good as w, and for at least one value of r, strictly better in both criteria. We also see that the equity does not reduce.

Case 2: $V(\mathcal{C}_t^{\wedge^D}) \subseteq V(\mathcal{C}_t^{\wedge^A})$

There exists a cell $\mathcal{C}_a$ such that $\mathcal{C}_a \in f'$, $\mathcal{C}_a \subset \mathcal{C}_t^{\wedge}$, such that $\mathcal{C}_a$ has the lowest score amongst all cells $\mathcal{C}_a, \mathcal{C}_b, \ldots, \mathcal{C}_z \subset \mathcal{C}_t^{\wedge}$ and only consists of advantaged instances.

Construct h as follows: Remove $\epsilon > 0$ mass of $\mathcal{X}$ from $\mathcal{C}_a$ to create a separate cell $\mathcal{C}'$. Denote the remainder of $\mathcal{C}_t^{\wedge}$ by $\mathcal{C}''$. Observe that $\mathcal{S}(\mathcal{C}') < \mathcal{S}(\mathcal{C}_t^{\wedge}) < \mathcal{S}(\mathcal{C}'')$. Take $\epsilon > 0$ small enough to not change order of $\mathcal{C}''$ in the partition w. It should be in the same position as $\mathcal{C}_t^\wedge$ was before (We can do this because of the genericity assumption). The only change in the order is that $\mathcal{C}'$ jumps to some position behind $\mathcal{C}''$.

The new partition h is

$$\mathcal{C}_1^\wedge, \mathcal{C}_2^\wedge, \cdots, \mathcal{C}_{t-1}^\wedge, \mathcal{C}'', \mathcal{C}_{t+1}^\wedge \cdots, \mathcal{C}_v^\wedge, \mathcal{C}', \mathcal{C}_{v+1}^\wedge, \cdots \mathcal{C}_d^\wedge$$

with descending order of scores.

Similar to the proof of Theorem 1, it is easy to check that for all rates r, the fairness and accuracy of h is at least as good as w, and for at least one value of r, strictly better in both criteria. We also see that the equity does not reduce.

□

Adding the Disadvantage Condition

In Theorem 3, if we make the disadvantage assumption, we can find a partition h that simultaneously guarantees a strict increase in equity as well. That is, we get the following statement below.

Theorem 4. *For every non-trivial graded-simplification, say w, there exists a partition h that simultaneously strictly improves accuracy, fairness, and equity with respect to w.*

Proof. We omit the proof because essentially the same construction as in the result of Kleinberg and Mullainathan [2019] works for this result as well.

□

6 Conclusion and Future Work

Through this work, we see that in machine learning based decision systems, the desiderata of fairness and interpretability may be at odds with each other and it is often necessary to make trade-offs between them, if we want to maintain accuracy. We prove theoretical results to demonstrate this. We consider a formal framework to build simple classifiers as a means to achieve interpretability, and show that if we restrict our classifier to be simple within this framework, it can be replaced by a more complex classifier that strictly improves both fairness and accuracy. Therefore, we see that simplicity/interpretability clashes with the properties of fairness and accuracy.

There are many variants of the setup that we could investigate for further work. While this result talks about the tradeoffs between fairness and simplicity, it is important to note that not all forms of building simple classifiers (for e.g., linear classifiers) are captured by this framework. It would be interesting to investigate the compatibility between fairness and other notions of simplicity. Also, we deploy a particular objective function to quantify unfairness, and it might be worth looking into the interplay between interpretability and fairness for other fairness objectives.

Acknowledgments

We would like to thank Nivasini Ananthakrishnan, Shai Ben-David, Gautam Kamath, Tosca Lechner, Yaoliang Yu, and anonymous reviewers for useful comments.

References

Jennifer L Doleac and Benjamin Hansen. Does “ban the box” help or hurt low-skilled workers? statistical discrimination and employment outcomes when criminal histories are hidden. Technical report, National Bureau of Economic Research, 2016.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

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

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.