Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents

Anonymous Author(s)
Submission Id: (10)

ABSTRACT

The use of algorithmic decision making systems in domains which impact the financial, social, and political well-being of people has created a demand for these decision making systems to be "fair" under some accepted notion of equity. This demand has in turn inspired a large body of work focused on the development of fair learning algorithms which are then used in lieu of their conventional counterparts. Most analysis of such fair algorithms proceeds from the assumption that the people affected by the algorithmic decisions are represented as immutable feature vectors. However, strategic agents may possess both the ability and the incentive to manipulate this observed feature vector in order to attain a more favorable outcome. We explore the impact that strategic agent behavior can have on group-fair classification. We find that in many settings strategic behavior can lead to *fairness reversal*, with a conventional classifier exhibiting higher fairness than a classifier trained to satisfy group fairness. Further, we show that fairness reversal occurs as a result of a group-fair classifier becoming more *selective*, achieving fairness largely by excluding individuals from the advantaged group. In contrast, if group fairness is achieved by the classifier becoming more *inclusive*, fairness reversal does not occur.

KEYWORDS

Group Fairness, Strategic Classification, Machine Learning

ACM Reference Format:
Anonymous Author(s). 2022. Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 24 pages.

1 INTRODUCTION

The increasing deployment of algorithmic decision making systems in social, political, and economic domains has brought with it a demand that fairness of decisions be a central part of algorithm design. While the specific notion of fairness appropriate to a domain is often a matter of debate, several have come to be commonly used in prior literature, such as positive (or selection) rate and false positive rate. A common goal in the design of fairness-aware (*group-fair*) algorithms is to balance predictive efficacy (such as accuracy) with achieving near-equality on a chosen fairness measure among demographic categories, such as race or gender. A question that arises in many domains where such "fair" algorithms could be used is whether they are susceptible to, and create incentives for, manipulation by agents who may misrepresent themselves in order to achieve better outcomes. For example, in selection of individuals to receive assistance from social service programs, or in admission to selective educational programs, it may be possible for applicants to misreport features like the number of dependents, income, or other self-reported characteristics.

We investigate the effects of such strategic manipulation of a binary *group-fair* classifier. In the context of the social services example, the classifier's job is to determine if an applicant should, or should not, be granted assistance, and the fairness guarantee of this classifier could be approximate equality of false positive rate between male and female applicants. Our first observation is that the ability of individuals to manipulate the features a classifier uses can lead to *fairness reversal*, with the conventional (accuracy-maximizing) classifier exhibiting greater fairness than a group-fair classifier. We observe this phenomenon on a number of standard benchmark datasets commonly used in evaluating group-fair classifiers. Next, we theoretically investigate the conditions under which such fairness reversal occurs. We prove that the key characteristic that leads to fairness reversal is that the group fair classifier becomes more selective, excluding some of the individuals in the advantaged group from being selected. Moreover, we show that this condition is sufficient for fairness reversal for several classes of functions measuring the costs of misreporting features. In contrast, we experimentally demonstrate that when a group-fair classifier exhibits inclusiveness instead by selecting additional individuals from the disadvantaged group, fairness reversal does not occur.

Summary of results: We begin by observing empirically the phenomenon of fairness reversal, exhibited on a number of datasets commonly used in benchmarking group-fair classification efficacy. The key factor that results in fairness reversal is the extent to which group fairness is achieved through increased selectivity (the fair classifier f_F positively classifies fewer inputs than the conventional classifier f_C) as opposed to increased inclusiveness (f_F positively classifies more inputs than f_C). Next, we examine this issue theoretically, and prove that selectivity is a sufficient condition for fairness reversal. Further, we show that, under some additional conditions, selectivity is also a necessary condition. These results obtain for two common classes of functions measuring the cost of misreporting attributes, and explain our empirical observations.

Related Work: Our work is closely related to two major strands in the literature: algorithmic fairness (in particular, approaches for group-fair classification) and adversarial machine learning (also called strategic classification).

The algorithmic fairness literature aims to study the extent to which algorithmic decisions are perceived as unfair, for example, by being inequitable to historically disadvantaged groups [2, 4, 5, 8]. Many approaches have been introduced, particularly in machine learning, that investigate how to balance fairness and task-related efficacy, such as accuracy [1, 10, 13, 17, 25–27]. Many of these

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.

impose hard constraints to ensure that pre-defined groups are near-equitable on some exogenously specified metric, e.g., selection (positive) rate [1, 17, 26], although alternative means, such as modifying the data to eliminate disparities, have also been proposed [7, 10].

The adversarial machine learning literature was initially motivated by security considerations, such as spam and malware detection [15, 19, 24]. The primary issue of concern is that as we use machine learning techniques to identify malicious behavior, malicious actors change behavior characteristics to evade detection. It has, however, come to have a far broader scope, encompassing robustness of machine learning techniques in computer vision as well as social applications [3, 6, 9, 11, 12]. In the latter context, this is known as *strategic classification*, to indicate the concern that individuals impacted by algorithmic decisions change their features (e.g., by misreporting their household characteristics on surveys used to allocate housing to the homeless) and thereby undermine algorithms' efficacy. The intersection between strategic classification and fairness is particularly salient to our work, and has featured studies that highlight the inequity that results from strategic behavior by individuals [14], as well as inequity (social cost) resulting from making classifiers robust to strategic behavior [21, 25]. Our goal, however, is quite distinct: we investigate the extent to which *group-fair* classification itself leads to greater inequity compared to baseline approaches that do not include group-fairness constraints as a result of strategic behavior by individuals.

2 PRELIMINARIES

We consider a setting with a population of agents, with each characterized by 1) a feature vector $\mathbf{x} \in \mathcal{X}$, 2) a group $g \in G \equiv \{0, 1\}$ to which it belongs (as is common in much prior literature, we treat it as binary here), and 3) a (true) binary label $y \in \mathcal{Y} \equiv \{0, 1\}$, denoting, for example, the agent's qualification (for a service, employment, bail, etc). Let $\mathcal{D}$ be the joint distribution over $G \times \mathcal{X} \times \mathcal{Y}$. We define $h(\mathbf{x})$ as the *marginal* pdf of $\mathbf{x}$, and assume that $h(\mathbf{x}) > 0$ for each $\mathbf{x} \in \mathcal{X}$.

Since using the sensitive group membership feature may pose a legal challenge, we assume that neither the conventional nor the group-fair classifier do so at prediction time (but may at training time), results relating to group aware classifiers (those that use group membership at prediction time) are provided in the supplement. We denote the conventional classifier by f_C, while the group-fair classifier is denoted by f_F, and both map feature vectors $\mathbf{x}$ into a binary label $y \in \mathcal{Y}$. We assume that the conventional classifier aims to maximize accuracy, i.e., $f_C \in \arg\max_f \mathbb{P}_{(\mathbf{x},y)}(f(\mathbf{x}) = y)$, while f_F aims to balance accuracy and fairness, solving

$$\begin{aligned} f_F = \operatorname{argmax}_f \; & (1-\alpha)\mathbb{P}_{(\mathbf{x},y)}(f(\mathbf{x}) = y) \\ & - \alpha\big|\mathcal{M}(f; g=0) - \mathcal{M}(f; g=1)\big|, \end{aligned}$$

where $\alpha \in [0, 1]$ specifies the relative weight of accuracy and fairness terms, while $\mathcal{M}(f; g)$ is a measure of efficacy (e.g., positive rate) of f restricted to a group g.

In the literature fairness is sometimes defined with hard constraints, rather than the soft constraints of α-fairness, for example

$$\begin{aligned} f_F = \operatorname{argmax}_f \; & \mathbb{P}_{(\mathbf{x},y)}(f(\mathbf{x}) = y) \\ & \text{subj. to } \big|\mathcal{M}(f; g=0) - \mathcal{M}(f; g=1)\big| \le \beta \end{aligned}$$

With hard constraints, decreasing β can never increase the unfairness of f_F. In general soft constraints do not have the propriety that increasing α will never increase the unfairness of f_F. However, in the settings we study this is not an issue as there is a direct correspondence between α and β fairness (Lemma A.2 in the appendix).

We consider the impact of strategic behavior of agents when they face a classifier f (whether conventional or group-fair). Specifically, we suppose that each agent with features $\mathbf{x}$ can modify these, transforming them into another feature vector $\mathbf{x}'$ that is reported to the classifier. In doing so, the agent incurs a cost, captured by a manipulation cost function $c(\mathbf{x}, \mathbf{x}') \ge 0$ [11, 12, 19].

We study two families of manipulation cost functions:

Feature-monotonic costs: $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $||\mathbf{x} - \mathbf{x}'||$, (larger manipulations are more costly).

Outcome-monotonic costs: $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $\mathbb{P}(y = 1|\mathbf{x}') - \mathbb{P}(y = 1|\mathbf{x})$ and $c(\mathbf{x}, \mathbf{x}') = 0$ for all $\mathbf{x}'$ such that $\mathbb{P}(y = 1|\mathbf{x}) > \mathbb{P}(y = 1|\mathbf{x}')$, (manipulations leading to better outcomes are more costly).

We define the agent's utility as

$$u(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}') - f(\mathbf{x}) - \frac{1}{B}c(\mathbf{x}, \mathbf{x}'),$$

where B is a parameter trading off costs and benefits of manipulation. Following the standard setting in strategic classification or adversarial machine learning, we assume any misreporting behavior would not change the true nature of $\mathbf{x}$'s label y. We assume that all agents are rational utility maximizers. Thus, since $f(\mathbf{x}') - f(\mathbf{x}) \le 1$, the agent will misreport its features only when $c(\mathbf{x}, \mathbf{x}') \le B$. Additionally, the agent will not misreport if $f(\mathbf{x}) = 1$ (they are selected even with true values of features). Consequently, we can equivalently view B as an upper bound on the costs that agents are willing to incur from misreporting their features, that is, the *manipulation budget*.

3 FAIRNESS REVERSAL

Our central goal is to understand the conditions under which *fairness reversal* occurs in strategic settings, that is, when a fair classifier f_F becomes less fair than its conventional counterpart f_C if agents act strategically. Fairness reversal occurs when there is a range of strategic manipulation budgets B for which the conventional classifier f_C exhibits greater fairness than the group-fair model f_F. In this section, we study this phenomenon empirically, demonstrating that it is commonly observed for several benchmark datasets.

For our empirical study, we use five datasets used as common benchmarks for group-fair classification:

Adult: Dataset of working professionals where the goal is to predict high or low income (protected feature: gender).

Community Crime: Dataset of communities where the objective is to predict if the community has high crime (protected feature: race).

Law School: Dataset of law students where the objective is to predict bar-exam passage (protected feature: race).

Student: Dataset of students where the objective is to predict a student receiving high math grades (protected feature: race).

Credit: Dataset of people applying for credit where the objective is to predict creditworthiness (protected feature: age).

All five datasets have binary outcomes, and we label the more desirable outcome for the individuals by $y = 1$ (e.g., having a high

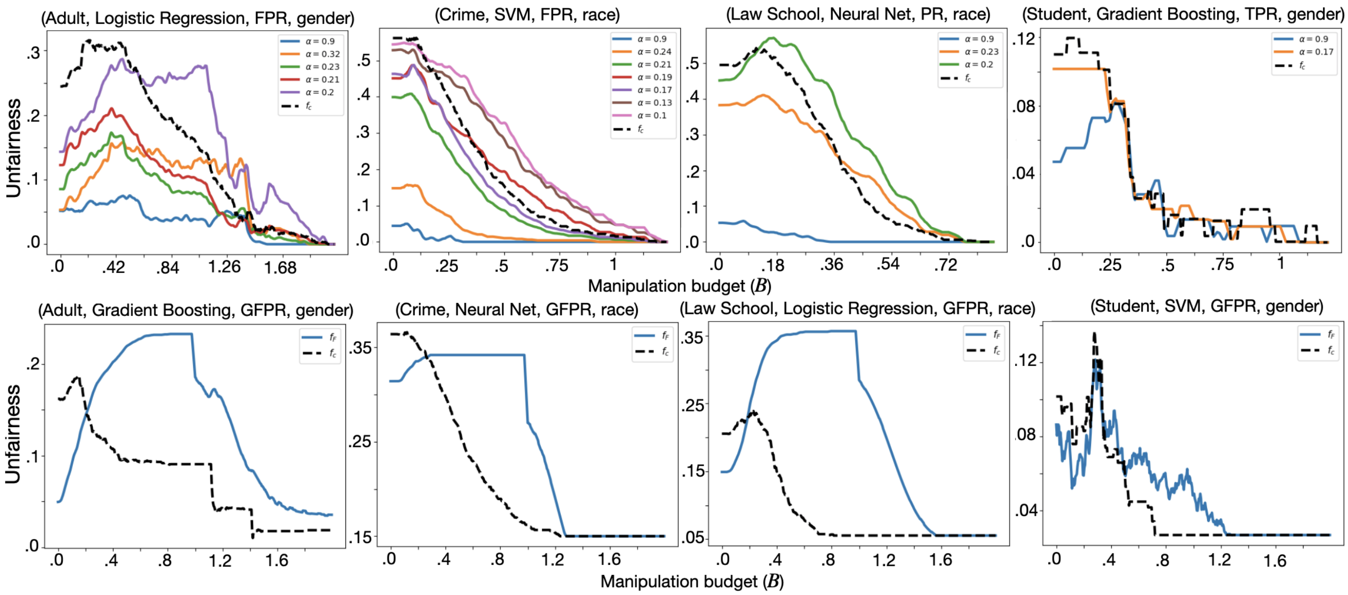

Figure 1: Difference in unfairness between groups on several datasets as a function of the manipulation budget B when manipulation costs are feature-monotonic. The dashed black lines correspond to f_C and colored lines correspond to f_F. Fairness reversal occurs when one of the colored lines is above the black line. The top row displays results when f_F is learned via the Reductions algorithm, with fairness defined in terms of PR, TPR, or FPR, for several different values of α. The bottom row displays results when f_F is learned via the EqOdds algorithm, with fairness defined in terms of GFPR. Reductions is group-agnostic, and EqOdds is group-aware.

income in the Adult data), with the less desirable outcome labeled by $y = 0$. Consequently, higher *positive rate (PR)*, *true positive rate (TPR)*, or *false positive rate (FPR)* is more desirable for individuals. Group membership in each dataset is determined by race, gender, or age which in these datasets corresponds to a binary feature (as in [16] the age feature is made binary by considering those older than 25 as Old, and those 25 or younger as Young). A detailed breakdown of the datasets can be found in the Appendix. In all cases, we refer to the "advantaged" group (e.g. the group with higher *PR* for *PR* based fairness) as group 1, or G_1, while the disadvantaged group is referred to as 0 or G_0. In our experiments, we only consider features that can potentially be manipulated (see the Appendix for further details). We use four classifiers as *conventional* f_C, namely logistic regression (LGR), support vector machines with an RBF kernel (SVM), neural networks (NN), and gradient boosting trees (GB), and three group-fair approaches to obtain f_F, *Reductions* [1], *GerryFair* [17], and *EqOdds* [22] The first two impose hard group-fairness constraints with a specified tolerance level β, while the third remedies unfairness through post processing. To study strategic manipulation, we use a mix of local search for categorical features [18, 23] and projected gradient descent (PGD) for continuous features [20]; further details are provided in the Appendix.

We investigate fairness reversals on four of the datasets and for Reductions and EqOdds fairness methods in Figure 1; additional experiments in section E appendix show that this illustration is representative. Consider first Figure 1 (top), which considers settings where predictions do not take the sensitive features as an input (we call these *group-agnostic* classifiers). In these four plots, the dashed line corresponds to f_C, and the rest are group-fair classifiers f_F for different values of α (recall that higher α entails greater importance of group fairness). What we observe is that in many cases, particularly when α is not very high, there is a range of budget values B for which f_F becomes less fair than f_C. Moreover, in many cases, this range is considerable. In Figure 1 (bottom plots), where group-fair classifiers are *group-aware*, including the sensitive feature as an input, the fairness reversal phenomenon is even more dramatic.

Figure 1 exhibits several additional phenomena. Note, in particular, that in many cases the unfairness (i.e., FPR difference between the groups) initially *increases* as the budget increases, but in all cases as budgets B keep increasing, eventually unfairness vanishes *as a result of strategic behavior by agents*. Furthermore, much as we observe this initial unfairness increase for both f_C and f_F, it appears *amplified* for some of the group fair classifiers f_F.

What causes fairness reversal? As we formally prove below, the essential condition is *selectivity* of fair classifier f_F compared to f_C. Specifically, in binary classification, there are, roughly, two ways one can improve fairness on a given dataset (that is, without any consideration of strategic behavior): *inclusiveness* (selecting additional agents from the disadvantaged group by changing their predicted class to 1) and *selectivity* (excluding some of the members of the advantaged group by changing their predicted class to 0). **Our key observation is that *selectivity* leads to fairness reversals, while *inclusiveness* does not.**

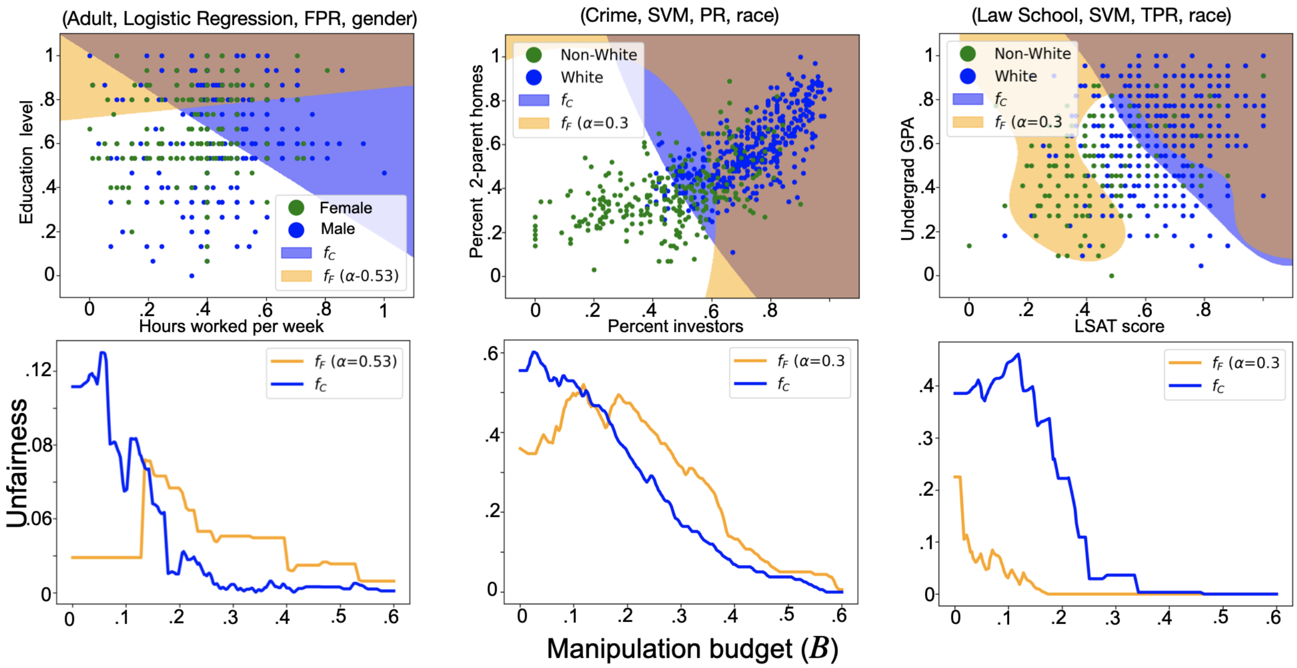

Figure 2: Fairness reversals and selectivity of classifiers on two ordinal features. The top row shows regions with positive predictions using two features (corresponding to the axes), and dot colors correspond to the sensitive demographics. The bottom row shows the relative unfairness between demographic groups (for the classifiers shown in the top row) as a function of strategic manipulation budget B (lower means more fair).

We illustrate this in Figure 2, which shows the decision boundaries of f_F and f_C (top row), as well as associated fairness as a function of budget (bottom row) for several combinations of dataset, classifier, and fairness definition. On the Adult and Crime datasets (first two columns), fairness is achieved predominantly through selectivity, as the orange region (f_C) includes few additional green points (disadvantaged group) compared to the blue region (f_C), but excludes many blue points (advantaged group). This, in turn, leads to instances of fairness reversal (bottom row first column). In the Law School dataset (third column), in contrast, fairness is achieved primarily through inclusiveness, and f_F remains more fair than f_C over a broad range of strategic manipulation budgets B. The reason that selectivity leads to fairness reversal is that those from the advantaged group who are excluded tend as a result to be closer to the decision boundary than those from the disadvantaged group. In the Appendix we provide further results linking selectivity of the fair classifier to fairness reversals. We also observe in the Appendix that when strategic agent behavior results in a fairness reversal between f_F and f_C, the relative accuracy of the classifiers is also reversed, implying a fundamental relationship between fairness and accuracy when agents are strategic.

In the next section, we study the phenomenon of fairness as well as accuracy reversal in strategic classification settings theoretically, demonstrating that selectivity is indeed a sufficient (and, under some additional qualifications, necessary) condition for fairness reversal.

4 THEORETICAL ANALYSIS

In this section we provide theoretical explanations of the empirical observations made in the previous section. We start with single-variable classifiers. We then proceed to generalize our observations to multi-feature classifiers. Our key observation is that selectivity is in fact a sufficient condition for fairness reversal, providing a theoretical underpinning for the empirical observations above. Additionally, we investigate the underlying *causes* of fair classifiers become more selective, and provide conditions on the underlying distribution for this to be the case. In the case of single variable classifiers with feature-monotonic costs and multivariable classifiers with outcome-monotonic costs, we further show that selectivity also leads to accuracy reversals and outline conditions on the underlying distribution such that selectivity is also a necessary condition for both of these phenomena.

To begin, we now formally define fairness reversal.

Definition 4.1. ***(Fairness Reversal)*** *Let M be a fairness metric (e.g. FPR), f_F be a classifier which is* group-fair *with respect to M, and f_C be a conventional accuracy-maximizing classifier. Define $U(f) = \left|M(f|g=1) - M(f|g=0)\right|$. Suppose that $U(f_F) < U(f_C)$. Let $f_C^{(c,B)}, f_F^{(c,B)}$ be the induced classifiers when agents best respond to f_C and f_F respectively with manipulation cost $c(\mathbf{x}, \mathbf{x}')$ and budget B. We say that a* budget B leads to fairness reversal *between f_C and f_F if $U(f_F^{(c,B)}) \geq U(f_C^{(c,B)})$.*

We will then say that fairness reversal between f_F and f_C occurs if there is some strategic manipulation budget B which leads to fairness reversal, that is, for this budget, f_C becomes more fair than f_F after manipulation. Note that if the budget B is 0, f_F will be more fair than f_C by construction, whereas if the budget is infinite, as long as any input is classified as the positive class, all individuals can misreport their features to be this class, and consequently either classifier is fair, in the sense that every input is predicted as 1. As a result, our analysis in the sequel is focused solely on the intermediate cases between these extremes.

4.1 Single Variable Classifier

We begin our theoretical exploration of fairness reversals with an exemplar case: a single variable threshold classifier. In this setting agents possess a single ordinal feature x. For simplicity we demonstrate our results for a continuous feature $x \in [0, 1]$, but the results hold for any ordinal feature (discrete or continuous). Both the conventional classifier (selected for maximal accuracy) and fair classifier (selected for weighted combination of accuracy and fairness with respect to a fairness metric M) can be expressed as a single parameter $\theta_C, \theta_F \in [0, 1]$ respectively where $f(x) = \mathbb{I}[x \geq \theta]$.

Our first result is that in single-feature classification, higher selectivity of the group-fair classifier is sufficient for fairness reversal.

Theorem 4.2. *Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x'-x|$, and θ_C is the most accurate, and θ_F the optimal α-fair, threshold. If $\theta_C < \theta_F$, then there exists a budget B that leads to fairness reversal between f_F and f_C.*

Proof Sketch. The full proof is provided in the Appendix. The unfairness of threshold θ w.r.t. to the distribution $\mathcal{D}$ and fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$ is expressed as,

$$U_{\mathcal{D}}(\theta) = \big|\mathcal{M}_{\mathcal{D}}(\theta|g = 1) - \mathcal{M}_{\mathcal{D}}(\theta|g = 0)\big|,$$

For a given threshold θ and manipulation budget B the best response of an agent with true feature x is

$$x^{(B)}_{\theta} = \text{argmax}_{x'} \big(\mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]\big) \;\text{ s.t. } c(x, x') \leq B,$$

When agents from $\mathcal{D}$ play this optimal response, let the resulting distribution be $\mathcal{D}^{(c,B)}_{\theta}$. The difference in unfairness between classifiers when agents are strategic is $U_{\mathcal{D}^{(c,B)}_{\theta_C}}(\theta_C) - U_{\mathcal{D}^{(c,B)}_{\theta_F}}(\theta_F)$. Since both f_C and f_F are thresholds and c is feature-monotonic, we can express the decisions of both on the modified distribution $\mathcal{D}^{(c,B)}_{\theta}$ as modified thresholds on the original distribution $\mathcal{D}$:

$$U_{\mathcal{D}^{(c,B)}_{\theta_C}}(\theta_C) - U_{\mathcal{D}^{(c,B)}_{\theta_F}}(\theta_F) = U_{\mathcal{D}}(\theta^{(c,B)}_C) - U_{\mathcal{D}}(\theta^{(c,B)}_F)$$
$$\text{where } \theta^{(c,B)}_C = \text{argmin}_x x \text{ s.t. } c(x, \theta_C) \leq B \text{ and,}$$
$$\theta^{(c,B)}_F = \text{argmin}_x x \text{ s.t. } c(x, \theta_F) \leq B$$

By the monotonicity of c both θ_C, θ_F are monotonically decreasing w.r.t. B and $\theta^{(c,B)}_C \leq \theta^{(c,B)}_F$. When $\theta^{(c,B)}_C = 0$, the unfairness of $\theta^{(c,B)}_C$ is trivially 0, i.e. $U_{\mathcal{D}}(0) = 0$. Let B' be the largest B for which $U_{\mathcal{D}}(\theta^{(c,B')}_C) > 0$. The manipulated threshold $\theta^{(c,B')}_C$ is continuous in B and thus, $U_{\mathcal{D}}(\theta^{(c,B)}_C)$ is also continuous in B, implying that $U_{\mathcal{D}}(\theta^{(c,B)}_C)$ takes on all values between $U_{\mathcal{D}}(\theta^{(c,B')}_C)$ and $U_{\mathcal{D}}(\theta_C)$. If at B', we have $U_{\mathcal{D}}(\theta^{(c,B')}_C) = U_{\mathcal{D}}(\theta^{(c,B')}_C)$ the proof is complete since both classifiers have equal unfairness. If this is not the case, then there exists $\varepsilon > 0$ such that $U_{\mathcal{D}}(\theta^{(c,B'+\varepsilon)}_C) = 0$, but $\theta^{(c,B'+\varepsilon)}_F \geq \theta^{(c,B'+\varepsilon)}_C$, implying that $U_{\mathcal{D}}(\theta^{(c,B')}_F) > U_{\mathcal{D}}(\theta^{(c,B')}_C)$. Thus a fairness reversal must occur at $B = B'$. □

We now turn our attention to a complementary observation: fairness reversal is accompanied by *accuracy reversal*, that is, strategic behavior leads to f_F having higher accuracy than f_C. This is primarily due to the fact that f_F becomes more selective and therefore more resilient to manipulation.

Theorem 4.3. *Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x'-x|$, and θ_C is the most accurate, and θ_F the optimal α-fair, threshold. If $\theta_C < \theta_F$, then there exists a budget B such that f_F is more accurate than f_C.*

Proof. Let $\mathcal{L}_{\mathcal{D}}(\theta)$ denote the error of classifier f with threshold θ on distribution $\mathcal{D}$. Note that by definition $\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta)$ for all $\theta \in [0, 1]$, and $\mathcal{L}_{\mathcal{D}}(0) = \mathbb{P}(y = 0)$. Thus, for $B \in [0, \infty)$, by continuity and the fact that a threshold θ on the manipulated distribution $\mathcal{D}^{(c,B)}_{\theta}$ can be expressed as a threshold $\theta^{(c,B)}$ on the original distribution $\mathcal{D}$, we have

$$\mathcal{L}_{\mathcal{D}^{(c,B)}}(\theta_C) = \mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_C) \in [\mathcal{L}_{\mathcal{D}}(\theta_C), \mathbb{P}(y = 0)].$$

Moreover, since $\theta_C \leq \theta_F$, any $B \geq 0$ implies $\theta^{(c,B)}_C \leq \theta^{(c,B)}_F$. If $\theta^{(c,B)}_C = \theta^{(c,B)}_F$ then $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_C) = \mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_F)$. If not then $\theta^{(c,B)}_C = \theta^{(c,B+\varepsilon)}_F$ for some $\varepsilon > 0$. That is, for any budget B either an accuracy reversal occurs, or $\theta^{(c,B)}_C$ is strictly more selective than $\theta^{(c,B)}_F$. This, combined with the fact that $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_C)$ takes on all values in $[\mathcal{L}_{\mathcal{D}}(\theta_C), \mathbb{P}(y = 0)]$ indicates that $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_C)$ and $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B+\varepsilon)}_F)$ is arbitrarily close to $\mathbb{P}(y = 0)$. When this happens, it must be the case that $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_C) = \mathcal{L}_{\mathcal{D}}(\theta^{(c,B+\varepsilon)}_F) \geq \mathcal{L}_{\mathcal{D}}(\theta^{(c,B)}_F)$. □

We showed thus far that selectivity is *sufficient* for fairness and accuracy reversal. Under what conditions is it also *necessary*? Loosely speaking, when a feature x is a good predictor of both y and g, both the error and unfairness f_C and f_F are *unimodal* (formally defined in the Appendix) with respect to the manipulation budget B. As documented in the Appendix, most ordinal features produce threshold classifiers which have (approximately) unimodal error and unfairness. When this occurs, the selectivity of f_F is not only sufficient for fairness and accuracy reversals, but also necessary. This is stated informally in the following theorem, and is formally stated and proved in the Appendix.

Definition 4.4. ***(Unimodal):*** *A function $H : [a, b] \to \mathbb{R}$ is* unimodal *if there exists a point $r \in [a, b]$ such that H is monotone decreasing (increasing) on $[a, r]$ and monotone increasing (decreasing) on $[r, b]$.*
(All convex and concave functions are unimodal.)

Theorem 4.5 (Informal). *Let f_C and f_F be threshold classifiers and $c(x, x')$ be feature monotonic. When error and unfairness are unimodal with respect to the manipulation budget B, then both fairness and accuracy reversal will occur between f_F and f_C if and only if $\theta_F > \theta_C$.*

In the Appendix we also discuss why error and unfairness tend to be unimodal on the benchmark datasets. The upshot, however, is that in many natural settings, selectivity is both necessary and sufficient for fairness reversal.

Now that we have established the critical role of selectivity in fairness reversal, we next analyze *why* that is. As mentioned previously, there are roughly two ways to achieve fairness: inclusiveness (classifying more examples as positive) or selectivity (classifying fewer examples as positive). Which of these will be the outcome of training f_F depends intimately on the data distribution. In the case of single feature classification, Theorem B.2 in the Appendix provides conditions on the underlying distribution such that the optimal fair classifier will achieve its fairness via selectivity (thus resulting in a fairness reversal if agents are strategic). This condition can be intuitively interpreted as follows. Suppose that S is the set of individuals selected (i.e., classified as 1) by f_C who are also near the decision boundary of f_C. If the advantaged group is overrepresented in S, there is a range of parameters α such that the optimal α-fair classifier is more selective than f_C (recall that higher α places greater importance on group-fairness in learning).

4.2 General Classifiers

Next we discuss general multi-variate classifiers, generalizing several of the results from Section 4.1. First we show that when f_F is more selective than f_C, fairness reversal occurs for both feature-monotonic and outcome-monotonic cost functions. Second, we give conditions which lead to f_F being more selective than f_C. For outcome-monotonic costs, we provide two additional results: 1) greater selectivity of f_F also leads to accuracy reversal, and 2) unimodality of each classifier's error and unfairness causes selectivity to be both necessary and sufficient for fairness and accuracy reversal.

4.2.1 Outcome-Monotonic Costs. Recall that a manipulation cost function $c(\mathbf{x}, \mathbf{x}')$ is outcome-monotonic if it is monotonic in $\mathbb{P}(y = 1|\mathbf{x}') - \mathbb{P}(y = 1|\mathbf{x})$ and 0 for any $\mathbb{P}(y = 1|\mathbf{x}') \leq \mathbb{P}(y = 1|\mathbf{x})$, i.e., reporting "better" features is more expensive. With respect to outcome-monotonic costs, we define the selectivity of a classifier f to be the minimum value of $\mathbb{P}(y = 1|\mathbf{x})$ such that $f(\mathbf{x}) = 1$, (i.e., what is the example with the lowest true probability to have $y = 1$ classified as positive by f).

As shown in [21] these manipulation costs result in the following best response for classifier f. Let

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x}) \quad \text{s.t. } f(\mathbf{x}) = 1.$$

Then for any agent with feature $\mathbf{x}$ such that $f(x) = 0$ the optimal strategy is to play $\mathbf{x}'$ where

$$\mathbf{x}' = \begin{cases} \mathbf{x} & \text{if } c(\mathbf{x}, \mathbf{x}^*) > B \\ \mathbf{x}^* & \text{otherwise.} \end{cases}$$

With this best response in hand we show that f_F having greater selectivity than f_C leads to fairness reversal.

Theorem 4.6. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic. Let $p_C = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_C(\mathbf{x}) = 1$ and $p_F = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_F(\mathbf{x}) = 1$. If $p_C < p_F$, then there exists a budget B that leads to fairness reversal between f_C and f_F.*

Proof sketch. The full proof is deferred to the Appendix. We can express agents' best responses in terms of a modified classifier, rather than a modified distribution. For classifier f, let

$$p' = \min_{\mathbf{x}: f(\mathbf{x})=1} \mathbb{P}(y = 1|\mathbf{x}),$$

and $\mathbf{x}'$ be the feature corresponding to p'. Then when agents best respond to f the resulting manipulated classifier can be expressed as a threshold on the underlying probabilities $\mathbb{P}(y = 1|\mathbf{x})$, i.e., let

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x}) \quad \text{s.t. } c(\mathbf{x}^*, \mathbf{x}') \leq B,$$

then $f^{(c,B)}$ classifies any $\mathbf{x}$ with $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1|\mathbf{x}^*)$ positively, and all others negatively. Similarly to Theorem 4.2, the monotonicity of $\mathbb{P}(y = 1|\mathbf{x}^*)$, as a function of B, implies the existence of a budget interval over which the unfairness of $f_C^{(c,B)}$ decreases below $f_F^{(c,B)}$, thus resulting in a fairness reversal. □

Similar to the single-variable, fairness reversals in the outcome-monotonic case also result in accuracy reversal:

Theorem 4.7. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome-monotonic. Let $p_C = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_C(\mathbf{x}) = 1$ and $p_F = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_F(\mathbf{x}) = 1$. If $p_C < p_F$, then there exists a budget B under which f_F becomes more accurate than f_C.*

The full proof is deferred to the Appendix.

Empirically we observe that when costs are outcome-monotonic, the majority of classifiers have error and unfairness which is unimodal with respect to the manipulation budget B. When this occurs, selectivity of f_F becomes both necessary and sufficient.

Theorem 4.8 (Informal). *Let f_C and f_F the optimal conventional and fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(x, x')$ is outcome-monotonic. If error and unfairness are unimodal with respect to the manipulation budget B, then there exists a budget B which leads to fairness and accuracy reversal if and only if f_F is more selective than f_C.*

The formal statement and full proof of this theorem is presented in the Appendix.

4.2.2 Feature-Monotonic Costs. Finally, we demonstrate that selectivity remains sufficient for fairness reversal in general when costs are feature-monotonic. Recall that in the single-variable case relative selectivity of f_F and f_C selectivity could be simply defined in terms of the classification thresholds θ_C and θ_F. Outcome-monotonic cost functions similarly admitted a relatively straightforward definition of selectivity, with f_F being more selective than f_C if $\min_{\mathbf{x}: f_C(\mathbf{x})=1} \left(\mathbb{P}(y = 1|\mathbf{x},)\right) < \min_{\mathbf{x}: f_F(x)=1} \left(\mathbb{P}(y = 1|\mathbf{x})\right)$ (i.e., f_F only accepts examples with higher true probability than that those accepted by f_C). In the general case when cost functions are feature-monotonic, we define selectivity in terms of the *sets of examples* that each classifier classifies positively. Specifically we say f_F is more selective than f_C if the set of examples positively classified by f_F constitutes a subset of those positively classified by f_C.

Definition 4.9. *Let* $\mathcal{X}_{f_C} = \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$ *and* $\mathcal{X}_{f_F} = \{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\}$*. We say that* f_F is more selective than f_C *if* $\mathcal{X}_{f_F} \subset \mathcal{X}_{f_C}$.

Note that this definition of selectivity generalizes the earlier definitions. In single-variable classification, $\mathcal{X}_{f_F} \subset \mathcal{X}_{f_C}$ if and only if $\theta_C < \theta_F$, and in the case of outcome monotonic costs $\mathcal{X}_{f_F} \subset \mathcal{X}_{f_C}$ implies

$$\min_{\mathbf{x}: f_C(\mathbf{x})=1} \left(\mathbb{P}(y = 1|\mathbf{x})\right) < \min_{\mathbf{x}: f_F(\mathbf{x})=1} \left(\mathbb{P}(y = 1|\mathbf{x})\right).$$

Of course, in reality selectivity is an *approximation*: in general, we will rarely have instances in which f_F is strictly more selective than f_C in the above sense. The practical upshot of our results here, therefore, is that they provide an explanation for the empirically observed phenomenon that ties approximate selectivity to fairness reversal.

Theorem 4.10. *Let* f_C *and* f_F *be the most accurate and the optimal* α*-fair classifier, respectively. Suppose fairness is defined by PR, FPR, or TPR and* $c(\mathbf{x}, \mathbf{x}')$ *is feature-monotonic. If* f_F *is more selective than* f_C*, then there exists a budget* B *that leads to fairness reversal between* f_F *and* f_C*.*

Proof Sketch. The full proof is deferred to the Appendix. The key idea is that trivial classifiers (i.e., those that predict $f(\mathbf{x}) = 1$ for all $\mathbf{x}$) have 0 unfairness (in terms of PR, FPR, and TPR fairness). As B increases, both $f_C^{(c,B)}$ and $f_F^{(c,B)}$ (the classifiers resulting from agents best responding to either classifier with budget B and cost function c) will approach 0 unfairness (not necessarily monotonically) as they become more like trivial classifiers. At some point prior to reaching trivial classification, the relative fairness of the classifiers will be flipped. This is due to the fact that

$$\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$$

and feature-monotonic costs result in the manipulated classifiers maintaining the subset propriety, i.e.,

$$\{\mathbf{x} \in \mathcal{X} : f_F^{(c,B)}(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C^{(c,B)}(\mathbf{x}) = 1\} \text{ for any } B.$$

Thus implying that $f_F^{(c,B)}$ approaches a trivial classifier more "slowly" than $f_C^{(c,B)}$ with respect to B. Moreover, prior to approaching triviality $f_F^{(c,B)}$ will effectively approach f_C, thus absorbing some of the original unfairness of f_C, resulting in a fairness reversal. □

Next, we give a condition that leads f_F to be more selective than f_C. Here, we provide this condition for the PR fairness metric; analogous results for TPR and FPR are given in the Appendix. For this result, we define the following notation $P_{G_z} = \mathbb{P}(g = z)$, $g(\mathbf{x}) = P(g = 1|\mathbf{x})$, and $\mathcal{X}_0 = \{\mathbf{x} \in \mathcal{X} : g(x) < P_{G_1} \text{ and } \mathbb{P}(y = 1|\mathbf{x}) < 1/2\}$, (i.e. those who are less likely than chance to have $g = 0$ and $y = 0$).

Theorem 4.11. *Suppose* f_C *and* f_F *are the most accurate and optimal* α*-fair classifier, respectively, and we use the PR fairness metric. Then* f_F *is more selective than* f_C *if and only if* $0 < \alpha \leq \alpha^*$*, where*

$$\alpha^* = \min_{\mathbf{x} \in \mathcal{X}_0} \frac{P_{G_0} P_{G_1} (2\mathbb{P}(y{=}1|\mathbf{x}){-}1)}{g(\mathbf{x}){+}P_{G_1}\left(P_{G_1}{-}2g(\mathbf{x}){-}2P_{G_1}\mathbb{P}(y{=}1|\mathbf{x})\right)}.$$

Proof Sketch. The full proof is deferred to the Appendix. Here we provide a proof sketch for PR fairness and discrete features. Both the conventional and fair objectives can be written as follows:

$$\begin{aligned} f_C &= \text{argmin}_f \mathbb{P}(f(\mathbf{x}) \neq y) \\ f_F &= \text{argmin}_f (1 - \alpha)\mathbb{P}(f(\mathbf{x}) \neq y) \\ &\quad + \alpha \left|\mathbb{P}(f(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f(x) = 1|g = 0)\right| \end{aligned}$$

Assuming the optimal f_F has higher positive rate for group 1, the argument of the fair objective can be simplified to

$$\begin{aligned} &(1 - \alpha) \sum_{\mathbf{x} \in \mathcal{X}} \left((1 - f(\mathbf{x}))\mathbb{P}(y = 1|\mathbf{x}) + f(\mathbf{x})\mathbb{P}(y = 0|\mathbf{x})\right)\mathbb{P}(\mathbf{x}) \\ &+\alpha \sum_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \left(\frac{\mathbb{P}(g = 1|\mathbf{x})}{\mathbb{P}(g = 1)} - \frac{\mathbb{P}(g = 0|\mathbf{x})}{\mathbb{P}(g = 0)}\right) \mathbb{P}(\mathbf{x}) \end{aligned}$$

Thus $f_F(\mathbf{x}) = 1$ is optimal if

$$\begin{aligned} &\alpha \frac{(\mathbb{P}(g = 1|\mathbf{x}) + (\mathbb{P}(g = 1) - 2)\mathbb{P}(g = 1))}{(1 - \mathbb{P}(g = 1))\mathbb{P}(g = 1)} \\ &\quad - (1 - \alpha)2\mathbb{P}(y = 1|\mathbf{x}) + 1 \\ &\geq 0, \end{aligned} \tag{1}$$

and $f_C(\mathbf{x}) = 1$ is optimal if $\mathbb{P}(y = 1|\mathbf{x}) \geq 1/2$. Thus, f_F will positively classify an example $\mathbf{x}$, which is negatively classified by f_C (i.e., $f_f(\mathbf{x}) = 1 \neq f_C(\mathbf{x}) = 0$) , only if Equation 1 is nonnegative and $\mathbb{P}(y = 1|\mathbf{x}) \geq 1/2$. Simplifying this condition yields α^*. □

Remark 4.12. The condition on α^* given in Theorem 4.11 is also sufficient for f_F to be more selective than f_C, when selectivity is defined in terms outcome-monotonic costs.

The key observation from Theorem 4.11 is that fairness reversal is a *small-*α phenomenon. This may seem surprising, since f_F is likely to be most similar to f_C for smaller values of α (in particular, the two are identical when $\alpha = 0$). However, when α is high, the fairness term is sufficiently dominant that reversals are unlikely. Consequently, it is precisely the intermediate values of α, where we aspire to preserve high accuracy while improving group-fairness that are most susceptible to fairness reversal. And, indeed, this is consistent with our empirical observations in Section 3. Further, note that for some distributions, $\alpha^* \leq 0$, which means that fairness reversal cannot be guaranteed.

5 CONCLUSION

We demonstrate a fairness-reversal phenomenon, where a trained-to-be fair classifier exhibits more unfairness than the conventional accuracy-maximizing one if human agents can strategically respond to a classifier. We show that a sufficient condition for observing fairness reversal is "selectivity", that is, a group-fair classifier making fewer positive predictions than its conventional counterpart. Our results caution against a naive expectation of fairness guarantees when a fair classifier sees real-world deployment. A more nuanced understanding of when fair classifiers may suffer from such problems and how to mitigate them is an important direction for future work.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[2] Ifeoma Ajunwa, Sorelle A. Friedler, C. Scheidegger, and S. Venkatasubramanian. Hiring by algorithm: Predicting and preventing disparate impact, 2016.

[3] Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-proof machine learning. *arXiv preprint*, 2020.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[6] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

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

[8] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning, 2018. arXiv preprint.

[9] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[12] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Innovations in Theoretical Computer Science*, 2016.

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[14] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic classification. In *Conference on Fairness, Accountability, and Transparency*, 2019.

[15] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *ACM Workshop on Security and Artificial Intelligence*, pages 43–58, 2011.

[16] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, 2009. doi: 10.1109/IC4.2009.4909197.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[18] Bo Li and Yevgeniy Vorobeychik. Evasion-robust classification on binary domains. *ACM Transactions on Knowledge Discovery from Data*, 12(4):1–32, 2018.

[19] Daniel Lowd and Christopher Meek. Adversarial learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 641–647, 2005.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[21] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Conference on Fairness, Accountability, and Transparency*, page 230–239, 2019.

[22] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *CoRR*, abs/1709.02012, 2017.

[23] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. Improving robustness of ML classifiers against realizable evasion attacks using conserved features. In *USENIX Security Symposium*, pages 285–302, 2019.

[24] Yevgeniy Vorobeychik and Murat Kantarcioglu. *Adversarial Machine Learning*. Morgan & Claypool Publishers, 2018.

[25] Han Xu, Xiaorui Liu, Yaxin Li, Anil K Jain, and Jiliang Tang. To be robust or to be fair: towards fairness in adversarial training. In *International Conference on Machine Learning*, 2021.

[26] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[27] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

APPENDIX

In the appendix we provide full proofs for our theoretical claims as well as additional experimental results.

A GENERAL DEFINITIONS AND LEMMAS

Definition A.1. *(PR, TPR, FPR): Postive Rate (PR), True Positive Rate (TPR), and False Positive Rate (FPR) are defined, for classifier f (with predicted probabilities h) and distribution $\mathcal{D}$ over $G \times \mathcal{X} \times \mathcal{Y}$, as*

$$\begin{aligned} PR_{\mathcal{D}}(f) &= \mathbb{P}\big(f(x) = 1\big) \\ TPR_{\mathcal{D}}(f) &= \mathbb{P}\big(f(x) = 1 | y = 1\big) \\ FPR_{\mathcal{D}}(f) &= \mathbb{P}\big(f(x) = 1 | y = 0\big) \\ GTPR_{\mathcal{D}}(f) &= \mathbb{E}[h(x) | y = 0] \\ GFPR_{\mathcal{D}}(f) &= \mathbb{E}[(1 - h(x)) | y = 1] \end{aligned}$$

Lemma A.2. *Consider the objective*

$$\begin{aligned} f_F = \arg\min_f\ & (1-\alpha)\mathbb{P}_{(\mathbf{x},y)}(f(\mathbf{x}) = y) \\ & - \alpha\big|\mathcal{M}(f; g = 0) - \mathcal{M}(f; g = 1)\big| \end{aligned}$$

When $\mathcal{M}$ is defined in terms of PR, TPR, FPR, increasing α could never in a less fair classifier.

Proof: (Lemma A.2). Let $U(f_F) = |\mathcal{M}(f_F; g = 0) - \mathcal{M}(f_F; g = 1)|$ and $\mathcal{L}(f) = 1 - \mathbb{P}(f(\mathbf{x}) = y)$. First observe that $f = 1$ causes $U(f) = 0$, implying that for α arbitrarily close to 1 the $U(f)$ approaches 0. Next consider two classifier f_1, f_2, which are solutions to the soft constrained objective for α_1, α_2 respectively, where $\alpha_1 > \alpha_2$.

$$\begin{aligned} &(1-\alpha_1)\mathcal{L}(f_1) + \alpha_1 U(f_1) < (1-\alpha_1)\mathcal{L}(f_2) + \alpha_1 U(f_2) \\ &\quad \text{and} \\ &(1-\alpha_2)\mathcal{L}(f_2) + \alpha_2 U(f_2) < (1-\alpha_2)\mathcal{L}(f_1) + \alpha_2 U(f_1) \end{aligned}$$

If for all such pairs of f_1, f_2 it was the case that $U(f_1) \geq U(f_2)$, then the proof is complete, so assume by way of contradiction that $U(f_1) < U(f_2)$. That is, increasing the fairness coefficient α (from α_1 to α_2) has resulted in a less fair classifier. In this case, we then know that $\mathcal{L}(f_1) > \mathcal{L}(f_2)$ since otherwise f_1 would be the optimal classifier for both α_1 and α_2. Thus for some $\varepsilon, \gamma > 0$, we can express $\mathcal{L}(f_1) = \mathcal{L}(f_2) + \epsilon$ and $U(f_1) = U(f_2) - \gamma$. Hence

$$\begin{aligned} &(1-\alpha_1)\mathcal{L}(f_1) + \alpha_1 U(f_1) < (1-\alpha_1)\mathcal{L}(f_2) + \alpha_1 U(f_2) \\ \Longrightarrow\ &(1-\alpha_1)(\mathcal{L}(f_2) + \varepsilon) + \alpha_1(U(f_1) - \gamma) \\ &\quad < (1-\alpha_1)\mathcal{L}(f_2) + \alpha_1 U(f_2) \\ \Longrightarrow\ &(1-\alpha_1)\varepsilon < \alpha_1\gamma \end{aligned}$$

and

$$\begin{aligned} &(1-\alpha_2)\mathcal{L}(f_1) + \alpha_2 U(f_1) > (1-\alpha_2)\mathcal{L}(f_2) + \alpha_2 U(f_2) \\ \Longrightarrow\ &(1-\alpha_2)(\mathcal{L}(f_2) + \varepsilon) + \alpha_2(U(f_2) - \gamma) \\ &\quad < (1-\alpha_2)\mathcal{L}(f_2) + \alpha_2 U(f_2) \\ \Longrightarrow\ &(1-\alpha_2)\varepsilon > \alpha_2\gamma \end{aligned}$$

Since $\alpha_1, \alpha_2, \gamma, \varepsilon > 0$

$$\begin{aligned} &(1-\alpha_1)\varepsilon > (1-\alpha_2)\varepsilon > \alpha_2\gamma > \alpha_1\gamma \\ \Longrightarrow\ &\alpha_2 > \alpha_1. \end{aligned}$$

which is a contradiction. Hence increasing α could not increase classifier unfairness. □

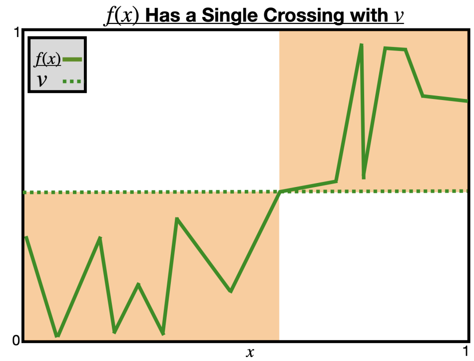

Figure 3: Example of a a function $f(x)$ (solid green) which has a *single crossing* (Def A.4) with the constant function v (dotted green). $f(x)$ can take on any values within the orange regions and maintaining the single crossing condition with v. That is, so long as $f(x)$ is upper bounded by v prior to crossing v, and lower bounded by v after crossing v, the single crossing condition holds.

Definition A.3. *(Unimodal): A function $H : [a, b] \to \mathbb{R}$ is negatively unimodal (*positively unimodal*) on the interval $[a, b]$ if there exists a point $r \in [a, b]$ such that H is monotone decreasing (increasing) on $[a, r]$ and monotone increasing (decreasing) on $[r, b]$.*
(All convex functions are negatively unimodal *and all concave functions are* positively unimodal.*)*

Definition A.4. *(Single Crossing): A function f is said to have a single crossing with the function g if there exists z s.t.*

$$\forall x \leq z : f(x) \geq g(x) \quad and \quad \forall x \geq z : f(x) \geq g(x)$$

This *single crossing* property is relevant to our work as we look at conditional distributions which have a single crossing with their unconditioned counterpart, e.g.. the functions $\mathbb{P}(y = 1|x)$ and $\mathbb{P}(y = 1)$ have a single crossing w.r.t. x. Note that any monotone function has a single crossing with *every* constant function and thus monotone conditionals trivially satisfy this condition.

Lemma A.5. *A once-differentiable function is unimodal if its derivative has a single crossing with the constant function 0.*

Proof. Let $f(x) : \mathbb{R} \to \mathbb{R}$ be a once-differentiable function. Suppose that $f'(x)$ has a single crossing with 0. Then there exists a point z such that $x \leq z$ implies $f'(x) \leq 0$ and $z \leq x$ implies $0 \leq f'(x)$. Thus f is monotonically decreasing on the interval $(-\infty, z]$ and monotonically increasing on the interval $[z, \infty)$. Implying that f is unimodal. □

B SINGLE VARIABLE CLASSIFIERS

Here we provide full proofs for the results in Section 4.1. Before proving the main results we first provide several helping lemmas and definition.

First we show that threshold classifiers predicting on manipulated distributions, can be expressed as thresholds on the original unmanipulated distribution.

Lemma B.1. *Suppose f is of the form $f(x) = \mathbb{I}[x \geq \theta]$, and the cost of manipulating a feature from x to x' is given as $c(x, x')$, where an agent with true feature x can submit any x' subject to $c(x, x') \leq B$. Then for any cost function c which is monotone in $|x' - x|$ there exists a classifier $f^{(c,B)}(x) = \mathbb{I}[x \geq \theta^{(c,B)}]$ which makes identical predictions on the true distribution $\mathcal{D}$ as f makes on manipulated data $\mathcal{D}_f^{(c,B)}$, i.e. when agents behave strategically $f(x') = f^{(c,B)}(x)$ for all $x \in \mathcal{X}$ and*

$$\theta^{(c,B)} = argmin_x x \ \ s.t. \ \ c(x, \theta) \leq B.$$

Lemma B.1 implies that strategic agent behavior can be examined through both the perspective of the original classifier f making predictions on the modified distribution $\mathcal{D}_f^{(c,B)}$ or a modified classifier $f^{(c,B)}$ on the original distribution $\mathcal{D}$. Since our investigation involves comparing two classifiers, f_C and f_F, the latter perspective is of particular usefulness given that the distribution $\mathcal{D}$ remains invariant between classifiers which will useful for our proofs.

Proof: (Lemma B.1). When all agents prefer positive predictions to negative predictions, their manipulations will change the classifier in only a single direction, namely manipulations cause negatively predicted examples to become positively predicted. Thus, only agents with feature x, where $f(x) = 0$ need be considered.

Suppose f is a threshold classifier with threshold θ, then the agent's best response to f is,

$$x^* = \text{argmax}_x \mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]$$
$$\text{s.t.} \ \ c(x, x') \leq B$$

Since the cost function $c(x, x')$ is monotone w.r.t. $|x' - x|$ the above best response has solution

$$x^* = \begin{cases} \theta & \text{if } c(x, \theta) \leq B \ \text{ and } x < \theta \\ x & \text{otherwise} \end{cases}$$

Moreover, the monotonicity of $c(x, x')$ also implies that if an agent with feature x has best response $x^* = \theta$, then so will any other agent with x_1 where $x \leq x_1 < \theta$.

Thus, the distribution shift of $\mathcal{D}$ caused by strategic behavior, can be quantified in terms of the agent with the smallest feature which is able to report a value of θ, i.e. the feature

$$x_{\min} = \text{argmin}_x x$$
$$\text{s.t. } c(x, \theta) \leq B$$

Thus when agents are strategic, any agent with feature $x \geq x_{\min}$ will be positively classified by f. Therefore, the threshold $\theta' = x_{\min}$ makes the same classifications on the unmanipulated distribution $\mathcal{D}$ as θ makes on the manipulated distribution $\mathcal{D}_\theta^{(c,B)}$. □

Next we restart the three theorems provided in Section 4.1 (Theorems 4.2, 4.3, 4.5) and provide their full proofs.

Theorem (4.2). Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x' - x|$, and θ_C, θ_F are respectively the most accurate and optimal α-fair thresholds. If $\theta_C < \theta_F$ then here exists a budget B such that strategic behavior agent behavior, leads to f_F becoming less fair than f_C if $\theta_C < \theta_F$, (i.e. fair classifiers which are more *selective* than their baseline counterpart become less fair under strategic manipulation).

Proof: (Theorem 4.2). The unfairness of threshold θ w.r.t. to the distribution $\mathcal{D}$ and fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$ is expressed as,

$$U_{\mathcal{D}}(\theta) = \big|\mathcal{M}_{\mathcal{D}}(\theta; g = 1) - \mathcal{M}_{\mathcal{D}}(\theta; g = 0)\big|,$$

For a given threshold θ and manipulation budget B the best response of an agent with true feature x is

$$x_\theta^{(B)} = \text{argmax}_{x'} \big(\mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]\big) \ \text{ s.t. } c(x, x') \leq B,$$

When agents from $\mathcal{D}$ play this optimal response, let the resulting distribution be $\mathcal{D}_\theta^{(c,B)}$. The difference in unfairness, between classifiers, when agents are strategic is $U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F)$. By lemma B.1 this unfairness can be expressed in terms of the true distribution $\mathcal{D}$, namely

$$U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F) = U_{\mathcal{D}}(\theta_C^{(c,B)}) - U_{\mathcal{D}}(\theta_F^{(c,B)})$$
$$\text{where } \theta_C^{(c,B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_C) \leq B \text{ and,}$$
$$\theta_F^{(c,B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_F) \leq B$$

By the monotonicity of c both θ_C, θ_F are monotonically decreasing w.r.t. B and $\theta_C^{(c,B)} \leq \theta_F^{(c,B)} \leq \theta_C$. When $\theta_C^{(c,B)} = 0$, the unfairness of $\theta_C^{(c,B)}$ is trivially 0, i.e. $U_{\mathcal{D}}(0) = 0$ Let B' be the largest B for which $U_{\mathcal{D}}(\theta_C^{(c,B')} > 0$. Since c is continuous, $\theta_C^{(c,B')}$ is continuous in B and thus, $U_{\mathcal{D}}(\theta_C^{(c,B)}$ is also continuous in B. Hence $U_{\mathcal{D}}$ takes on all values in the interval $[U_{\mathcal{D}}(\theta_C^{(c,B')}, U_{\mathcal{D}}(\theta_C)]$. Now if at B', we have $U_{\mathcal{D}}(\theta_C^{(c,B')} = U_{\mathcal{D}}(\theta_C^{(c,B')})$ the proof is complete since both classifiers have equal unfairness. If this is not the case, then $\theta_C^{(c,B')} < \theta_F^{(c,B')}$, and there must exits some $\varepsilon > 0$ such that $U_{\mathcal{D}}(\theta_C^{(c,B'+\varepsilon)}) = 0$, but $\theta_F^{(c,B'+\varepsilon)} > \theta_C^{(c,B'+\varepsilon)}$. If no such epsilon existed then for any $B > B'$ it would have to be the case that

$$U_{\mathcal{D}}(\theta_C^{(c,B)}) > U_{\mathcal{D}}(\theta_F^{(c,B)})$$
or
$$U_{\mathcal{D}}(\theta_C^{(c,B)}) = U_{\mathcal{D}}(\theta_F^{(c,B)}) = 0$$

However, this would imply that $\theta_F^{(c,B)} < \theta_C^{(c,B)}$, a contradiction. Thus $U_{\mathcal{D}}(\theta_F^{(c,B'+\varepsilon)}) > U_{\mathcal{D}}(\theta_C^{(c,B'+\varepsilon)} = 0$, which in turn implies that $U_{\mathcal{D}}(\theta_F^{(c,B')}) > U_{\mathcal{D}}(\theta_C^{(c,B')}$. That is a fairness reversal must occur at $B = B'$. □

Theorem (4.3). Suppose fairness is defined by PR, TPR, or FPR. $c(x, x')$ is monotone in $|x' - x|$, and θ_C, θ_F are respectively the most accurate and optimal α-fair thresholds. If $\theta_C < \theta_F$ then here exists a budget B such that strategic behavior agent behavior, leads to f_F becoming more accurate than f_C if $\theta_C < \theta_F$.

PROOF: (THEOREM 4.3). Let $\mathcal{L}_{\mathcal{D}}(\theta)$ denote the error of classifier f with threshold θ on distribution $\mathcal{D}$. By definition $\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta)$ for all $\theta \in [0,1]$, and $\mathcal{L}_{\mathcal{D}}(0) = \mathbb{P}(y=1)$. Thus, for $B \in [0,\infty)$, by continuity and Lemma B.1 we have $\mathcal{L}_{\mathcal{D}^{(c,B)}}(\theta_C) = \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B)}) \in [\mathcal{L}_{\mathcal{D}}(\theta_C), \mathbb{P}(y=1)]$. Moreover, since $\theta_C \leq \theta_F$, it must be the case that for all B the relationship $\theta_C^{(c,B)} \leq \theta_B^{(c,B)}$ holds. Thus we can express $\theta_C^{(c,B)} = \theta_F^{(c,B)} - \varepsilon$ for some $\varepsilon \geq 0$. By lemma B.1 we can express $\theta_C^{(c,B)} = \theta_F^{(c,B)} - \varepsilon = \theta_F^{(c,B')}$ for some B' with $B \leq B'$.

That is, for any budget B we may express $\theta_C^{(c,B)}$ as $\theta_F^{(c,B')}$ for some larger budget B'. This combine with the fact that $\mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B)}$ takes on all values in $[\mathcal{L}_{\mathcal{D}}(\theta_C), \mathbb{P}(y=0)]$ indicates that there must be a region $[B_1, B2] \subset [0,\infty)$ over which $\mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B)})$ is decreasing for $B \in [B_1, B_2]$ such that when $B' = B_1$, we have $\mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B')} \geq \mathcal{L}_{\mathcal{D}}(\theta_F^{(c,B')} > \mathbb{P}(y=0)$. Thus for budget B' an accuracy reversal occurs. ◻

Selectivity. First we provide the sufficient condition under which an optimal fair classifier f_F is indeed more selective than its conventional counterpart f_C. This condition can be interpreted as saying that if the advantaged group is overrepresented close to, but on the beneficial side of f_C, then there exists a range of fairness coefficients $[\alpha_1, \alpha_2]$ such that for $\alpha \in [\alpha_1, \alpha_2]$ the optimal α-fair classifier is more selective than the most accurate classifier f_C.

Theorem B.2. *Suppose fairness is defined by PR, TPR, or FPR. Suppose further that $\mathbb{P}(y=1|x)$ has a single crossing with $\mathbb{P}(y=1)$, and that $\mathbb{P}(g=1|x)$ has a single crossing with the respective value given in Lemmas B.4 and B.5, call this value p_g. Let x_y and x_g be defined by*

$$\mathbb{P}(y=1|x_y) = \mathbb{P}(y=1) \;\; \textit{and} \;\; \mathbb{P}(g=1|x_g) = p_g$$

If $x_g < x_y$, then there exists a nonempty interval $[\alpha_0, \alpha_1]$ s.t. for any $\alpha \in [\alpha_0, \alpha_1]$ the optimal α-fair classifier f_F, has the propriety that $\theta_C < \theta_F$ (implying that strategic agent behavior leads to f_F becoming less fair than f_C as outlined by Theorem 4.2).

PROOF: (THEOREM B.2). Given $\alpha \in (0,1)$, fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$, and data distribution $\mathcal{D}$, the objective of the fair learning scheme is to find θ_F such that

$$\theta_F = \text{argmin}_\theta (1-\alpha)\mathbb{P}(\mathbb{I}[x \geq \theta] \neq y) + \alpha U_{\mathcal{D}}(\theta) \tag{2}$$

where

$$U_{\mathcal{D}}(\theta) = \left|\mathcal{M}(\theta|g=1) - \mathcal{M}(\theta|g=0)\right|$$

By Lemma B.3 the error term $\mathbb{P}(\mathbb{I}[x \leq \theta] = y)$ is negatively unimodal in θ and achieves a minimum at θ_C where $\mathbb{P}(y=1|x=\theta_C) = \mathbb{P}(y=1)$. Similarly, by Lemmas B.4, B.5 the unfairness term $U_{\mathcal{D}}(\theta)$ is positively unimodal in θ and achieves a maximum at θ_U where $\mathbb{P}(g=1|x=\theta_U) = \mathbb{P}(g=1)$. Thus for any α the fair learning objective (Equation 2) is monotonically increasing, implying $\theta_F \notin [\theta_U, \theta_C]$. So either $\theta_F \in [0, \theta_U)$ or $\theta_F \in (\theta_C, 1]$. By the continuity of Equation 2 w.r.t. to θ. For some $\varepsilon > 0$, any $\varepsilon' < \varepsilon$ yields $\mathbb{P}(\mathbb{I}[x \geq \theta_C + \varepsilon'] \neq y) \leq \mathbb{P}(\mathbb{I}[x \geq \theta_U] \neq y)$ and $U_{\mathcal{D}}(\theta_C) \leq U_{\mathcal{D}}(\theta_U - \varepsilon')$. Thus implying that for small enough since both $|\theta_C - \theta_F|$ and $|\theta_U - \theta_F|$ are monotonic w.r.t. to α, it must be the case that that for α small enough $\theta_F = \theta_C + \varepsilon'$ is the optimal α-fair classifier, thus implying the existence of of fairness coefficients $[\alpha_1, \alpha_2]$ such that $\alpha \in [\alpha_1, \alpha_2]$ implies $\theta_F > \theta_C$. ◻

Next we discuss conditions under which selectivity of the fair classifier is not only sufficient, but also necessary for fairness and accuracy reversals. These conditions are motivated by our empirical findings shown in Figures 21-24. These figures display error and unfairness of threshold classifiers. In each of figure we see that almost all ordinal features found in the datasets we study, produce error and unfairness curves (red and orange respectively) which achieve a single extrema and are monotonic on either side of said extrema. Functions which exhibit this propriety are know as *unimodal functions* (Definition A.3).

Empirically we observe that most ordinal features exhibit unimodality of error and unfairness, or are simply not predictive features (i.e. the most accurate classifier f_C achieves roughly accuracy of 0.5). When error and unfairness are unimodal, selectivity of f_C is both necessary and sufficient for a fairness reversal between f_C and f_F.

Naturally, the next step is to investigate *why* error and unfairness would be unimodal. Ultimately the unimodality $\mathcal{L}(\theta)$ and $U(\theta)$ is a function of the underlying distribution, specifically the functions $\mathbb{P}(g=g', y=' |x)$, $\mathbb{P}(y=1|x)$, and $\mathbb{P}(g=1|x)$ for $g', y' \in \{0,1\}$. We theoretically show that when these functions have a *single-crossing* (Definition A.4) with there respective unconditioned value (e.g. $\mathbb{P}(g=1|x)$ has a single crossing with $\mathbb{P}(g=1)$), error and unfairness will be unimodal. In Figure 20 we see that each of the features used in Figures 21-24 approximately exhibits this single-crossing propriety, which causes unimodality, which in turn causes selectivity to be both necessary and sufficient for fairness reversals to occur.

Lemma B.3. *Suppose that $\mathbb{P}(y=1|x)$ has a single crossing with $\mathbb{P}(y=1)$. Then error is negatively unimodal w.r.t. θ and the optimal base threshold is θ_C s.t. $\mathbb{P}(y=1|\theta_C) = \mathbb{P}(y=1)$.*

PROOF: (LEMMA B.3). The error of a classifier $f(x) = \mathbb{I}[x \geq \theta]$ is given by,

$$\begin{aligned}
&1 - \mathbb{P}(\mathbb{I}[x \geq \theta] = y) \\
&= 1 - \mathbb{P}(x \geq \theta, y=1) - \mathbb{P}(x \leq \theta, y=0) \\
&= \mathbb{P}(y=0) + \mathbb{P}(x \leq \theta, y=1) - \mathbb{P}(x \leq \theta, y=0)
\end{aligned}$$

Since x is a continuous random variable and the terms involving θ are joint CDFs with well defined conditional PDFs, the derivative of the above expression w.r.t. to θ, exists and is equal to

$$\begin{aligned}
&h_{y,x}(y=1, x=\theta) - h_{y,x}(y=0, x=\theta) \\
&= h_x(x=\theta)\big(\mathbb{P}(y=1|x=\theta) - \mathbb{P}(y=0|x=\theta)\big) \\
&= h_x(x=\theta)\big(2\mathbb{P}(y=1|x=\theta) - 1\big)
\end{aligned}$$

Since $\mathbb{P}(y=1|x=\theta)$ has a single crossing with $\mathbb{P}(y=1)$, the above derivative is *split* by the value 0, thus by Lemma A.5 error is *negatively unimodal* with global minima at any θ_C s.t. $\mathbb{P}(y=1|x=\theta_C) = \mathbb{P}(y=1)$. ◻

Lemma B.4. *Suppose that fairness is defined in terms of Positive Rate (PR) and that $\mathbb{P}(g=1|x)$ has a single crossing with $\mathbb{P}(g=1)$, then*

(1) $PR_{\mathcal{D}}(\theta|g=1) \geq PR_{\mathcal{D}}(\theta|g=0)$ *for any* $\theta \in [0,1]$, *(i.e. group 1 is advantaged under any threshold classifier), and*
(2) *the unfairness term* $\left|PR_{\mathcal{D}}(\theta|g=1) - PR_{\mathcal{D}}(\theta|g=0)\right|$ *is positively unimodal w.r.t.* θ *and is maximized at any* θ_U *s.t.* $\mathbb{P}(g=1|x=\theta_U) = \mathbb{P}(g=1)$.

Proof: (Lemma B.4). For a classifier $f(x) = \mathbb{I}[x \geq \theta]$, we begin by demonstrating (1) the *unimodality* of $\mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=0)$ and then use this propriety to show (2) the equivalence between $\mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=0)$ and the unfairness term

$$\left|\mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=0)\right|.$$

First, note that

$$\begin{aligned}
&= \mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=1) \\
&= \frac{\mathbb{P}(g=1, x \geq \theta)}{\mathbb{P}(g=1)} - \frac{\mathbb{P}(g=0, x \geq \theta)}{\mathbb{P}(g=0)} \\
&= \frac{\mathbb{P}(g=1) - \mathbb{P}(g=1, x \leq \theta)}{\mathbb{P}(g=1)} \\
&\quad - \frac{\mathbb{P}(g=0) - \mathbb{P}(g=0, x \leq \theta)}{\mathbb{P}(g=0)} \\
&= 1 - \frac{\mathbb{P}(g=1, x \leq \theta)}{\mathbb{P}(g=1)} - 1 + \frac{\mathbb{P}(g=0)\mathbb{P}(g=0, x \leq \theta)}{\mathbb{P}(g=0)} \\
&= -\frac{\mathbb{P}(g=1, x \leq \theta)}{\mathbb{P}(g=1)} + \frac{\mathbb{P}(g=0, x \leq \theta)}{\mathbb{P}(g=0)}
\end{aligned}$$

Since each term involving θ is a joint CDF, the derivative of this term w.r.t to θ exists and is equal to

$$\begin{aligned}
&\frac{h_{g,x}(g=0, x=\theta)}{\mathbb{P}(g=0)} - \frac{h_{g,x}(g=1, x=\theta)}{\mathbb{P}(g=1)} \\
&= \frac{\mathbb{P}(g=0|x=\theta)h_x(x=\theta)}{\mathbb{P}(g=0)} - \frac{\mathbb{P}(g=1|x=\theta)h_x(x=\theta)}{\mathbb{P}(g=1)} \\
&= \frac{\left(1 - \mathbb{P}(g=1|x=\theta)\right)h_x(x=\theta)}{\mathbb{P}(g=0)} \\
&\quad - \frac{\mathbb{P}(g=1|x=\theta)h_x(x=\theta)}{\mathbb{P}(g=1)} \\
&= h_x(x=\theta)\frac{\mathbb{P}(g=1) - \mathbb{P}(g=1|x=\theta)}{\mathbb{P}(g=1)\mathbb{P}(g=0)}
\end{aligned}$$

Since $\mathbb{P}(g=1|x)$ is *split* by the value $\mathbb{P}(g=1)$ the above term is *split* by the value 0, thus by Lemma the term $\mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=0)$ is *positively unimodal*, and is maximized at any θ_U s.t.

$$h_x(x=\theta_U)\frac{\mathbb{P}(g=1) - \mathbb{P}(g=1|x=\theta_U)}{\mathbb{P}(g=1)\mathbb{P}(g=0)} = 0$$

Since $h_x(x=\theta) > 0$ any such θ_U has the propriety that $\mathbb{P}(g=1|x=\theta_U) = \mathbb{P}(g=1)$. Thus concluding the proof of (2).

We now use (2) to show that (1) immediately follows. Note that for $\theta \in \{0,1\}$ we have $\mathbb{P}(x \geq \theta|g=1) = \mathbb{P}(x \geq \theta|g=0)$. Since the function is *positively unimodal* and $\mathbb{P}(g=1) > 0$ neither $\theta = 0$ nor $\theta = 1$ can be points corresponding to local maximums, hence for any θ we have

$$\begin{aligned}
&\mathbb{P}(x \geq \theta|g=1) - \mathbb{P}(x \geq \theta|g=0) \\
&\geq \mathbb{P}(x \geq 1|g=1) - \mathbb{P}(x \geq 1|g=0) \\
&= 0
\end{aligned}$$

□

Lemma B.5. *Suppose that fairness is defined by either True Positive Rate or False Positive Rate and that* g, y *are conditionally independent given* x. *Suppose further that* $\mathbb{P}(g=1|x)$ *has a single crossing with* $\mathbb{P}(g=1|y=1)$ *in the TPR case and by* $\mathbb{P}(g=1|y=0)$ *in the FPR case. Then when* $\mathcal{M}$ *is TPR or FPR,*

(1) $\mathcal{M}_{\mathcal{D}}(\theta|g=1) \geq \mathcal{M}_{\mathcal{D}}(\theta|g=0)$ *for any* $\theta \in [0,1]$, *(i.e. group 1 is advantaged under any threshold classifier), and*
(2) *the unfairness term* $\left|\mathcal{M}_{\mathcal{D}}(\theta|g=1) - \mathcal{M}_{\mathcal{D}}(\theta|g=0)\right|$ *is positively unimodal w.r.t.* θ *and is maximized at any* θ_U *s.t.* $\mathbb{P}(g=1|x=\theta_U) = \mathbb{P}(g=1|y=1)$ *in the TPR case and* $\mathbb{P}(g=1|x=\theta_U) = \mathbb{P}(g=1|y=0)$ *in the FPR case.*

Proof: (Lemma B.5). This proof follows identically from Lemma B.4. □

Theorem B.6. *Suppose fairness is defined by PR, TPR, or FPR. Suppose further that* $\mathbb{P}(y=1|x)$ *has a single crossing (Def A.4) with* $\mathbb{P}(y=1)$, $\mathbb{P}(g=1|x)$ *has a single crossing with value given in Lemmas B.4 and B.5,* $c(x,x')$ *is monotone in* $|x'-x|$, *and* θ_C, θ_F *are respectively the most accurate and optimal* α*-fair thresholds. Then there exists a range of budgets* $[B_1, B_2]$ *such that strategic behavior agent behavior, with budget* $B \in [B_1, B_2]$, *leads to* f_F *being less fair than* f_C *if and only if* $\theta_C < \theta_F$, *(i.e. fair classifiers which are more selective than their baseline counterpart become less fair under strategic manipulation)*

Proof: (Theorem B.6). We first show that $\theta_C < \theta_F$ implies the existence of a budget interval $[B_1, B_2]$ s.t. strategic agent behavior under any $B \in [B_1, B_2]$ leads to f_F being less fair than f_C. We then show that if $\theta_F < \theta_C$, no such budget interval exists.

The unfairness of threshold θ w.r.t. to the distribution $\mathcal{D}$ and fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$ is expressed as,

$$U(\theta, \mathcal{D}) = \left|\mathcal{M}_{\mathcal{D}}(\theta|g=1) - \mathcal{M}_{\mathcal{D}}(\theta|g=0)\right|,$$

For a given threshold θ and manipulation budget B the best response of an agent with true type $a = (g, x)$ is

$$x_\theta^{(B)} = \text{argmax}_{x'}\left(\mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]\right) \ \text{ s.t. } c(x,x') \leq B,$$

When agents, originally distributed in accordance with $\mathcal{D}$, play this optimal responses w.r.t. θ and B, let the resulting distribution be $\mathcal{D}_\theta^{(B)}$. The difference in unfairness, between classifiers, when agents are strategic, can then be expressed as $U(\theta_C, \mathcal{D}_{\theta_C}^{(B)}) - U(\theta_F, \mathcal{D}_{\theta_F}^{(B)})$. Lemma B.1 gives a way to express this difference in terms of the original distribution $\mathcal{D}$ by changing the thresholds, namely

$$\begin{aligned}
&U(\theta_C, \mathcal{D}_{\theta_C}^{(B)}) - U(\theta_F, \mathcal{D}_{\theta_F}^{(B)}) \\
&= U(\theta_C^{(B)}, \mathcal{D}) - U(\theta_F^{(B)}, \mathcal{D})
\end{aligned}$$

where

$$\theta_C^{(B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_C) \leq B \text{ and,}$$
$$\theta_F^{(B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_F) \leq B$$

By the monotonicity of $c(x, x')$ w.r.t. $x' - x$ we have that

$$\theta_C < \theta_F \implies \theta_C^{(B)} \leq \theta_F^{(B)} \quad \forall\, B \geq 0$$
$$\theta_C > \theta_F \implies \theta_C^{(B)} \geq \theta_F^{(B)} \quad \forall\, B \geq 0$$

i.e. the relative ordering of the thresholds is preserved under strategic behavior for any manipulation budget B, this fact will be of use later.

Let $\theta_U = \text{argmax}_\theta U(\theta, \mathcal{D})$, we now proceed to prove the forward direction of our claim by three cases of the relative order of the thresholds $\theta_C, \theta_F, \theta_U$:

$$1.)\ \theta_C < \theta_F \leq \theta_U$$
$$2.)\ \theta_C \leq \theta_U \leq \theta_F$$
$$3.)\ \theta_U \leq \theta_C < \theta_F$$

First note that case (1) is infeasible. By Lemmas B.4 and B.5, we know that $U(\theta, \mathcal{D})$ is *positively unimodal* and maximized at θ_U. Therefore, for $\theta \in [\theta_C, \theta_U]$ we have that $U(\theta, \mathcal{D})$ is monotonically increasing. Thus in case (1) we have $U(\theta_F, \mathcal{D}) \geq U(\theta_C, \mathcal{D})$, which is impossible since f_F is assumed to be strictly more fair than f_C.

To prove that the claim holds in cases (2) and (3) we use the fact that the unfairness term $U(\theta, \mathcal{D})$, being *positively unimodal* implies that the term is also monotonically increasing on the interval $[0, \theta_U]$ and monotonically decreasing on $[\theta_U, 1]$. Hence, as stated previously, for any $\theta_1 \leq \theta_2 \leq \theta_U$ it must also be the case that $U(\theta_1, \mathcal{D}) \leq U(\theta_2, \mathcal{D}) \leq U(\theta_U, \mathcal{D})$. Therefore it suffices to show that there exists a budget interval $[B_1, B_2]$ s.t. for any $B \in [B_1, B_2]$ we have $\theta_C^{(B)} \leq \theta_F^{(B)} \theta_U$, and as stated previously, for any $B \geq 0$ $\theta_C < \theta_F$ implies that $\theta_C^{(B)} \leq \theta_F^{(B)}$. Hence we need only show that in cases (2), (3) having $B \in [B_1, B_2]$ implies $\theta_F^{(B)} \leq \theta_U$.

In both cases, (2) and (3), this follows immediately form Lemma B.1, which gives the existence of a budget B_U, such that $\theta_F^{(B_U)} = \theta_U$, and implies that $\theta_F(B)$ is monotonically decreasing w.r.t. to B. Therefore, for any $B \in [B_U, \infty)$ it must be the case that $U(\theta_C^{(B)}, \mathcal{D}) \leq U(\theta_F^{(B)}, \mathcal{D})$.

To prove the reverse direction, we need to show that when $\theta_F < \theta_C$ it is the case that for any $B \geq 0$ we have $U(\theta_F^{(B)}, \mathcal{D}) \leq U(\theta_C^{(B)}, \mathcal{D})$. We again show this by three cases on the relative order of the thresholds $\theta_F, \theta_C, \theta_U$:

$$1.)\ \theta_F < \theta_C \leq \theta_U$$
$$2.)\ \theta_F \leq \theta_U \leq \theta_C$$
$$3.)\ \theta_U \leq \theta_F < \theta_C$$

Similar to the forward direction of the proof, one case is infeasible, namely case (3). This can be seen be a symmetric argument to the previous one, specifically that on the interval $[\theta_U, \theta_C]$ both error and unfairness are monotonically decreasing, and thus θ_F could not be an optimal fair threshold.

As shown previously, when $\theta_F < \theta_C$ it is also the case that for any $B \geq 0$ we have $\theta_F^{(B)} \leq \theta_C^{(B)}$, and if $\theta_C^{(B)} \geq \theta_U$ then $U(\theta_F^{(B)}, \mathcal{D}) \leq U(\theta_C^{(B)}, \mathcal{D})$. Thus the claim holds for case (1), leaving only case (2) left to prove.

In case (2) we have $\theta_F \leq \theta_U \leq \theta_C$. Let B_U the budget s.t. $\theta_C(B_U) = \theta_U$, then for $B \in [0, B_U]$ the term $U(\theta_C^{(B)}, \mathcal{D})$ is monotone increasing, while $U(\theta_F^{(B)}, \mathcal{D})$ is monotone decreasing, and thus $U(\theta_F^{(B)}, \mathcal{D}) \leq U(\theta_C^{(B)}, \mathcal{D})$. Moreover for $B \in [B_U, \infty)$ we have already have show that $U(\theta_F^{(B)}, \mathcal{D}) \leq U(\theta_C^{(B)}, \mathcal{D})$.

Therefore the reverse direction of the claim holds, and thus there exists an interval $[B_1, B_2]$ s.t. $U(\theta_C^{(B)}, \mathcal{D}) \geq U(\theta_F^{(B)}, \mathcal{D})$ for $B \in [B_1, B_2]$ iff $\theta_C < \theta_F$. □

C GENERAL CLASSIFIERS AND OUTCOME-MONOTONIC MANIPULATION COSTS

Here we restate and provide full proofs for the results in Section 4.2.1.

Theorem (4.6). Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose that $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic. Let $p_C = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_C(\mathbf{x}) = 1$ and $p_F = \min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x})$ such that $f_F(\mathbf{x}) = 1$. If $p_C < p_F$, then there exists a budget B such that a fairness reversal occurs between f_C and f_F.

Proof: (Theorem 4.6). When agents best respond to classifier f with budget B and outcome-monotonic cost function $c(\mathbf{x}, \mathbf{x}')$, the resulting classifier $f^{(c,B)}$ can be expressed in terms of the underlying distribution of true label y conditioned on feature $\mathbf{x}$. Namely, $f^{(c,B)}$ can be expressed in terms of two quantities: the example $\mathbf{x}'$ with lowest true probability of being positive (but is positively classified by f), and the example $\mathbf{x}^*$ who has the highest cost to report $\mathbf{x}'$ (with $c(\mathbf{x}^*, \mathbf{x}') \leq B$). Note that the values $\mathbf{x}^*$ and $\mathbf{x}'$ need not be distinct, but ultimately we will care about $\mathbb{P}(y = 1|\mathbf{x}')$ and $\mathbb{P}(y = 1|\mathbf{x}^*)$, not specific features themselves. That is, let

$$\mathbf{x}'_{f^{(c,B)}} = \arg\min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x}) \quad \text{s.t. } f(\mathbf{x}) = 1$$

and

$$\mathbf{x}^*_{f^{(c,B)}} = \arg\max c(\mathbf{x}, \mathbf{x}'_{f^{(c,B)}}) \quad \text{s.t. } c(\mathbf{x}, \mathbf{x}'_{f^{(c,B)}}) \leq B$$

Then the classifier $f^{(c,B)}$ can be written as

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1|\mathbf{x}^*_{f^{(c,B)}}) \\ 0 & \text{otherwise} \end{cases}$$

When viewing f in this way, we see that strategic agent behavior results in classifiers which are thresholds on the underlying conditional distribution given by $\mathbb{P}(y = 1|\mathbf{x})$. Moreover as B increases, $\mathbb{P}(y = 1|\mathbf{x}^*_{f^{(c,B)}})$ decreases.

Now, given that f_C and f_F are respectively the most accurate and optimal α-fair classifiers, with $p_C < p_F$, we can express $f_C^{(c,B)}$ and $f_F^{(c,B)}$ as a thresholds on $\mathbb{P}(y = 1|\mathbf{x}^*_{f_C^{(c,B)}})$ and $\mathbb{P}(y = 1|\mathbf{x}^*_{f_F^{(c,B)}})$ respectively. Since $p_C < p_F$, the monotonicity of $\mathbb{P}(y = 1|\mathbf{x}^*_{f^{(c,B)}})$

with respect to B, implies that $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) \leq \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$ for all B. Thus, for any B, there exists a corresponding B' (with $B < B'$) such that $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) = \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B')}})$.

Similar to the single variable case, unfairness $U_{\mathcal{D}}(f^{(c,B)})$ is continuous with respect to B. For sufficiently large B, we have $f_C^{(c,B)}(\mathbf{x}) = 1$ for all $\mathbf{x}$, (i.e., all agents are capable of constructing manipulations which result in positive classification). For such a budget, the resulting unfairness $U_{\mathcal{D}}(f_C^{(c,B^*)})$ is 0 for PR, TPR, and FPR based fairness. Suppose that B^* is the largest budget such that $U_{\mathcal{D}}(f_C^{(c,B^*)}) > 0$. If $U_{\mathcal{D}}(f_C^{(c,B^*)}) \leq U_{\mathcal{D}}(f_F^{(c,B^*)})$, the proof is complete. If not then $U_{\mathcal{D}}(f_C^{(c,B^*)}) > U_{\mathcal{D}}(f_F^{(c,B^*)} \geq 0$. However, for some $B' < B^*$ we have $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B')}}) = \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B^*)}})$, implying $U_{\mathcal{D}}(f_C^{(c,B')}) = U_{\mathcal{D}}(f_F^{(c,B^*)}$. Thus $U_{\mathcal{D}}(f_C^{(c,B')}) \leq U_{\mathcal{D}}(f_F^{(c,B')})$, resulting in a fairness reversal at B'. ◻

Theorem (4.7). Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose that $c(\mathbf{x},\mathbf{x}')$ is outcome-monotonic. Let $p_C = \min_{\mathbf{x}} \mathbb{P}(y=1|\mathbf{x})$ such that $f_C(\mathbf{x}) = 1$ and $p_F = \min_{\mathbf{x}} \mathbb{P}(y=1|\mathbf{x})$ such that $f_F(\mathbf{x}) = 1$. If $p_C < p_F$, then there exists a budget B that f_F becomes more accurate than f_C.

(Proof: Theorem 4.7). Let $\mathcal{L}_{\mathcal{D}}(f)$ be the error of classifier f on distribution $\mathcal{D}$. As shown in the proof of Theorem 4.6, when agents best respond to f with outcome-monotonic cost $c(\mathbf{x},\mathbf{x}')$ and budget B, the resulting classifier $f^{(c,B)}$ can be expressed as

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|\mathbf{x}) \geq \mathbb{P}(y=1|\mathbf{x}^*_{f^{(c,B)}}) \\ 0 & \text{otherwise} \end{cases}$$

where

$$\mathbf{x}'_{f^{(c,B)}} = \arg\min_{\mathbf{x}} \mathbb{P}(y=1|\mathbf{x}) \quad \text{s.t. } f(\mathbf{x}) = 1$$

and

$$\mathbf{x}^*_{f^{(c,B)}} = \arg\max c(\mathbf{x},\mathbf{x}'_{f^{(c,B)}}) \quad \text{s.t. } c(\mathbf{x},\mathbf{x}'_{f^{(c,B)}}) \leq B$$

Moreover, for any classifier f_1 where $f_1(\mathbf{x}) = 1$ for all $\mathbf{x}$, the error of f_1 is $\mathcal{L}(f_1) = \mathbb{P}(y=0)$. For sufficiently large budget, all agents will be capable of manipulating their features and being positively classified, implying that for large enough B results in $\mathcal{L}(f^{(c,B)}) = \mathbb{P}(y=0)$. Let B^* be the largest B for which $\mathcal{L}(f^{(c,B)}) < \mathbb{P}(y=0)$. Since $p_C < p_F$, it must be the case that

$$\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) \leq \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}}) \text{ for all } B.$$

If for some B, we have $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) = \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$, then $\mathcal{L}(f_C^{(c,B)}) = \mathcal{L}(f_F^{(c,B)}$, i.e. both classifiers have equal accuracy and the theorem holds. Thus assume $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) < \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$, then by the monotonicity of $\mathbb{P}(y=1|\mathbf{x}^*_{f^{(c,B)}})$ as a function of B, there exists $B' > B$ such that $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) = \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B')}})$. When this B' corresponds to B^* we get, $\mathcal{L}(f_F^{(c,B')}) = \mathcal{L}(f_C^{(c,B^*}})$. Again if $\mathcal{L}(f_F^{(c,B^*)}) \leq \mathcal{L}(f_C^{(c,B^*})$ the proof is complete. If not then $\mathcal{L}(f_F^{(c,B')}) < \mathcal{L}(f_F^{(c,B^*)})$ which implies that $\mathcal{L}(f_F^{(c,B')}) \leq \mathcal{L}(f_F^{(c,B')}$ implying that either the accuracy reversal occurs at B^* or occurs at B'. ◻

Theorem (4.8). Let f_C and f_F the optimal conventional and fair classifiers respectively. Suppose fairness is defined in terms of PR, TPR, or FPR fairness, and $c(x,x')$ is outcome-monotonic. When error and unfairness are unimodal with respect to the manipulation budget B, a fairness and accuracy reversal will occur between f_F and f_C if and only if f_F is more selective than f_C.

Proof: Theorem 4.8. Let f_C and f_F be respectively the most accurate and α-fair classifiers. Suppose that for $f \in \{f_C, f_F\}$ the error term $\mathcal{L}(f^{(c,B)})$ is negatively unimodal w.r.t. to B and the unfairness term $U(f^{(c,B)})$ is positively unimodal w.r.t. to B. As shown previously when $p_C < p_F$ (i.e. f_F is more selective than f_C), both a fairness and accuracy reversal occurs.

Now, suppose that a fairness and accuracy reversal does occur between f_C and f_F. Then, for some B, we have $U(f_F^{(c,B)}) \geq U(f_C^{(c,B)})$ and $\mathcal{L}(f_F^{(c,B)}) \leq \mathcal{L}(f_F^{(c,B)}$. First we focus on the unfairness term. As per our assumption $U(f_C^{(c,B)})$ and $U(f_F^{(c,B)}$ are both unimodal w.r.t. B and $U(f_C) > U(f_F)$. Assume by way of contradiction that $p_C > p_F$ (i.e., f_C is more selective than f_F). As show previously, this implies $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) > \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$. Moreover, since both quantities of the inequality are monotonic w.r.t. B and unfairness is positively unimodal w.r.t. to B, both $U(f_C^{(c,B)})$ and $U(f_F^{(c,B)})$ are respectively unimodal w.r.t $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}})$ and $\mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$. That is $\mathbb{P}(y=1|\mathbf{x}^*_{f_C^{(c,B)}}) > \mathbb{P}(y=1|\mathbf{x}^*_{f_F^{(c,B)}})$ for all B implies $U(f_C^{(c,B)}) \geq U(f_F^{(c,B)})$ for all B, and no fairness reversals occur. Thus a fairness reversal occurs if and only if $p_C < p_F$ when unfairness is positively unimodal in B. A symmetric argument holds for error. ◻

D MULTIVARIABLE CLASSIFIERS AND FEATURE-MONOTONIC COSTS

Theorem (4.10). Let f_C and f_F be the most accurate and optimal α-fair classifiers respectively, $c(\mathbf{x},\mathbf{x}')$ be feature monotonic. Suppose fairness is defined by PR, FPR, or TPR. If $\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$ (i.e., the set of positively predicted examples of f_F is a subset of the positively predicted examples of f_C), then there exists a budget B for which a fairness reversal occurs between f_F and f_C.

Proof. Note that for PR, TPR, and FPR fairness, the constant function $f = 1$ achieves 0 unfairness. Let f_C and f_F be the most accurate and optimal α-fair classifiers respectively. This proof follows similar ideas to that of the previous proofs involving fairness reversals. However, there is a key difference: in the single variable case or outcome-monotonic cost case, both classifiers "share" and error and unfairness curve with respect to B. Meaning that when f_F is more selective than f_C, we are able to express the decisions of the conventional classifier $f_C^{(c,B)}$ in terms of fair classifier $f_F^{(c,B')}$

for some $B' > B$. In the case of multivariate classifiers with feature-monotonic costs, this is no longer the case. In stead we will make use of the fact that for some range of budgets the unfairness of $f_C^{(c,B)}$ can be expressed as the unfairness of $f_F^{(c,B')}$ for $B' > B$.

Let the set of positive examples for both classifiers be $\mathcal{X}_F = \{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\}$ and $\mathcal{X}_C = \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$, and for the manipulated classifiers let $\mathcal{X}_F^{(c,B)} = \{\mathbf{x} \in \mathcal{X} : f_F^{(c,B)}(\mathbf{x}) = 1\}$ and $\mathcal{X}_C^{(c,B)} = \{\mathbf{x} \in \mathcal{X} : f_C^{(c,B)}(\mathbf{x}) = 1\}$. In the case that $\mathcal{X}_F \subset \mathcal{X}_C$ and cost functions are feature monotonic, $\mathcal{X}_F^{(c,B)} \subset \mathcal{X}_C^{(c,B)}$ for all $B \geq 0$. To see this, take any $\mathbf{x}' \in \mathcal{X}_F^{(c,B)}$, then there exists $\mathbf{x} \in \mathcal{X}_F$ with $c(\mathbf{x}, \mathbf{x}') \leq B$ and $f_F^{(c,B)}(\mathbf{x}') = 1$. Since $\mathbf{x} \in \mathcal{X}_F$ we also have $\mathbf{x} \in \mathcal{X}_C$, implying $f_C(\mathbf{x}) = 1$ and thus for any $\mathbf{x}'$ with $c(\mathbf{x}, \mathbf{x}') \leq B$ $f_C^{(c,B)}(\mathbf{x}') = 1$ and $\mathbf{x}' \in \mathcal{X}_C^{(c,B)}$. That is, for any budget B the resulting manipulated classifiers will maintain this subset propriety: $\{\mathbf{x} \in \mathcal{X} : f_F^{(c,B)}(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C^{(c,B)}(\mathbf{x}) = 1\}$.

Since $\mathcal{X}_F^{(c,B)} \subset \mathcal{X}_C^{(c,B)}$ for all $B \geq 0$ we can write $\mathcal{X}_C^{(c,B)} = \mathcal{X}_F^{(c,B)} \cup \mathcal{X}_r^{(c,B)}$ where $\mathcal{X}_F^{(c,B)} \cap \mathcal{X}_r^{(c,B)} = \emptyset$. If $\mathcal{X}_r^{(c,B)} = \emptyset$ then $\mathcal{X}_F^{(c,B)} = \mathcal{X}_C^{(c,B)}$ and thus $U(f_F^{(c,B)}) = U(f_C^{(c,B)})$ and the theorem holds, so assume $\mathcal{X}_r^{(c,B)} \neq \emptyset$. For PR fairness (TPR and FPR follow an identical argument) the difference in unfairness between $f_F^{(c,B)}$ and $f_C^{(c,B)}$ as

$$
\begin{aligned}
&U(f_F^{(c,B)}) - U(f_C^{(c,B)}) \\
&= \left|\mathbb{P}(f_F^{(c,B)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_F^{(c,B)}(\mathbf{x}) = 1|g = 0)\right| \\
&\quad - \left|\mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 0)\right|
\end{aligned}
$$

Since $g = 1$ is the advantaged group

$$
\begin{aligned}
&\mathbb{P}(f_C^{(c,0)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C^{(c,0)}(\mathbf{x}) = 1|g = 0) \\
&= \mathbb{P}(f_C(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C(\mathbf{x}) = 1|g = 0) \\
&> 0
\end{aligned}
$$

Moreover, since $0 \leq U(f_F^{(c,B)})$, if

$$
\mathbb{P}(f_C^{(c,B')}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C^{(c,B')}(\mathbf{x}) = 1|g = 0) \leq 0
$$

a fairness reversal must have occurred for some $B \leq B'$. So assume

$$
\mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 0) > 0
$$

In this case we can write the difference in fairness as

$$
\begin{aligned}
&\mathbb{P}(f_F^{(c,B)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_F^{(c,B)}(\mathbf{x}) = 1|g = 0) \\
&\quad - \mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f_C^{(c,B)}(\mathbf{x}) = 1|g = 0) \\
&= \mathbb{P}(\mathbf{x} \in \mathcal{X}_F^{(c,B)}|g = 1) - \mathbb{P}(\mathbf{x} \in \mathcal{X}_F^{(c,B)}|g = 0) \\
&\quad - \mathbb{P}(\mathbf{x} \in \mathcal{X}_C^{(c,B)}|g = 1) - \mathbb{P}(\mathbf{x} \in \mathcal{X}_C^{(c,B)}|g = 0) \\
&= \mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 1) - \mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 0)
\end{aligned}
$$

Thus if $\mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 1) - \mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 0) \geq 0$ then $U(f_F^{(c,B)}) \geq U(f_C^{(c,B)}) \geq 0$. As the size of $\mathcal{X}_r^{(c,B)}$ decreases, (i.e. $f_F^{(c,B)}$ is closer to $f_C^{(c,B)}$ both $\mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 1)$ and $\mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 0)$ tend towards 0. Thus we ultimately care about $\mathcal{X}_r^{(c,B)}$ as a function of B. For notational simplicity denote $U(\mathcal{X}_r^{(c,B)}) = \mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 1) - \mathbb{P}(\mathbf{x} \in \mathcal{X}_r^{(c,B)}|g = 0)$, then $U(\mathcal{X}_r^{(c,B)}) \geq 0$ indicates a fairness reversal. For large enough B we have $U(f_C^{(c,B)}) = 0$. Let B' be such that $U(f_C^{(c,B')})$ but for small $\varepsilon > 0$ $U(f_C^{(c,B'+\varepsilon)}) = 0$. Then if $U(\mathcal{X}_r^{(c,B'+\varepsilon)}) = 0$ the proof is complete, otherwise $U(\mathcal{X}_r^{(c,B'+\varepsilon)}) < 0$. In this case it must be the case that $U(\mathcal{X}_r^{(c,B')}) \geq 0$, resulting in a fairness reversal. □

Theorem (4.11). For distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y} \times G$ let f_C and f_F be the most accurate and optimal α-fair classifier respectively. Then for PR, FPR, and TPR fairness there exists α^* such that, the set of examples positively predicted by f_F will be a subset of the set of those positively predicted by f_C if and only if $0 \leq \alpha \leq \alpha^*$. The value of α^* will depend on the fairness metric, for PR fairness (for notational simplicity, let $\mathbb{P}(g = g') = P_{G_{g'}}, P(g = 1|\mathbf{x}) = g(\mathbf{x})$),

$$
\alpha^* = \min_{\mathbf{x}} \frac{P_{G_0} P_{G_1} (2\mathbb{P}(y{=}1|\mathbf{x}){-}1)}{g(\mathbf{x}){+}P_{G_1} \left(P_{G_1}{-}2g(\mathbf{x}){-}2P_{G_1}\mathbb{P}(y{=}1|\mathbf{x})\right)}
$$

(Proof: Theorem 4.11). Both the conventional and fair objectives can be written as follows:

$$
\begin{aligned}
f_C &= \text{argmin}_f \mathbb{P}(f(\mathbf{x}) \neq y) \\
f_F &= \text{argmin}_f (1 - \alpha)\mathbb{P}(f(\mathbf{x}) \neq y) \\
&\quad + \alpha\left|\mathbb{P}(f(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f(\mathbf{x}) = 1|g = 0)\right|
\end{aligned}
$$

Assuming the optimal f_F has higher positive rate for group 1, the argument of the fair objective can be simplified to,

$$
\begin{aligned}
&(1 - \alpha) \sum_{\mathbf{x} \in \mathcal{X}} \big((1 - f(\mathbf{x}))\mathbb{P}(y = 1|\mathbf{x}) + f(\mathbf{x})\mathbb{P}(y = 0|\mathbf{x})\big)\mathbb{P}(\mathbf{x}) \\
&+\alpha \sum_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \left(\frac{\mathbb{P}(g = 1|\mathbf{x})}{\mathbb{P}(g = 1)} - \frac{\mathbb{P}(g = 0|\mathbf{x})}{\mathbb{P}(g = 0)}\right) \mathbb{P}(\mathbf{x})
\end{aligned}
$$

Thus $f_F(\mathbf{x}) = 1$ is optimal if

$$
\begin{aligned}
&\alpha \frac{(\mathbb{P}(g = 1|\mathbf{x}) + (\mathbb{P}(g = 1) - 2)\mathbb{P}(g = 1))}{(1 - \mathbb{P}(g = 1))\mathbb{P}(g = 1)} \\
&\quad - (1 - \alpha)2\mathbb{P}(y = 1|\mathbf{x}) + 1 \\
&\geq 0,
\end{aligned} \tag{3}
$$

and $f_C(\mathbf{x}) = 1$ is optimal if $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1)$. Thus, f_F will positively classify an example $\mathbf{x}$, which is negatively classified by f_C (i.e., $f_f(\mathbf{x}) = 1 \neq f_C(\mathcal{X}) = 0$) , only if Equation 3 is nonnegative and $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1)$. Simplifying the condition in Equation 3 yields α^*. □

E EXPERIMENTS

E.1 Selectivity and fairness reversals

Recall that for feature monotonic costs the fair classifier f_F is said to be more selective if the set of examples positively predicted by f_F is a subset of those positively predicted by f_C. That is $\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$. While this strict definition of selectivity rarely holds in practice. Selectivity can be more generally defined by $\mathbb{P}_{\mathbf{x}}\big(f_C(\mathbf{x}) = 1, f_F(\mathbf{x}) = 0\big) - \mathbb{P}_{\mathbf{x}}\big(f_C(\mathbf{x}) = 0, f_F(\mathbf{x}) = 1\big)$, i.e. for any example $\mathbf{x}$, how much more likely is $\mathbf{x}$ to be positively classified by f_C and not f_F, compared to positively classified by f_F and not f_C.

Under this definition of selectivity Figure 4 shows the magnitude of the maximum fairness reversal (y-axis) as a function of selectivity (x-axis). The maximum fairness reversal is defined as

$$\max_B \left|\mathcal{M}(f_F; g = 0) - \mathcal{M}(f_F; g = 1)\right| - \left|\mathcal{M}(f_C; g = 0) - \mathcal{M}(f_C; g = 1)\right|$$

for $\mathcal{M}$ defined as ther PR, TPR, or FPR. Each plot show the selectivity and and fairness reversal for a difference combination of classifier type, α and fairness definition for $\alpha \in \{0.05, 0.1, 0.15 \ldots 0.95\}$. There is a general positive correlation between selectivity and fairness reversals. This relationship is is more prominent in datasets such as Law School and Community Crime. We postulate that this is due to these two datasets possessing features which have higher correlation to both group and label, than the other datasets.

E.2 Accuracy reversals

We also observe that when strategic agent behavior results in a fairness reversal between f_F and f_C, the relative accuracy of the classifiers is also reversed. Figure 5 shows the error and unfairness of the fair (f_F) and conventional (f_C) classifier. The shaded part indicates the region (and magnitude) of fairness reversal. We see that in cases where a fairness reversal occurs, there is also an accuracy reversal. In the cases corresponding to the Adult and Credit datasets, we see that f_F is always the more fair classifier and f_C is always the more accurate classifier. In contrast, the cases corresponding to the Crime and Law School datasets we see that for a range of budgets, f_C becomes more fair than f_F, and over a range of budgets, f_F becomes more accurate than f_C. In essence the functionality of the two classifiers has been swapped. These observations suggest that there is a fundamental trade-off between a classifiers accuracy and fairness in the presence of strategic manipulation. This phenomenon is theoretically explained in Theorems 4.3, 4.7.

E.3 Group aware classifiers

Figures 15-18 show the relative unfairness of f_C and f_F on each dataset when EqOdds is used as the fair learning scheme and fairness is defined in terms of GFPR. We observe that fairness reversals are common across datasets and classifier type. Moreover, we observe that unfairness is unimodal.

Figures 17 and 19 show unfairness in terms of GFPR and GFTPR respectively on the Law School dataset. While the GFPR case tends to lead to fairness reversals, the GTPR case does not. This discrepancy is due to the way that EqOdds remedies fairness in either case. Specifically: in the case GFPR fairness EqOdds achieves fairness by specifically decreasing the predicted positive probabilities on the advantaged group, while in the case of GTPR by increasing the predicted positive probabilities of the disadvantaged group. That is, when fairness is defined in terms of GTPR it is typically more desirable to be classified as a member of group 0, compared to the GFPR case. This also ties into our observations of selectivity, namely that the classifier decreasing positive predictions (the GFPR case) incurs a higher rate of fairness reversals than the classifier increasing positive predictions (the GTPR case).

E.4 Single crossing and unimodality

Figure 20 show the single crossing conditions between $\mathbb{P}(y = 1|x)$, and $\mathbb{P}(g = 1|x)$, and their respective constant functions given in Lemmas B.3, B.4, B.5. We see that in all three datasets the single crossing conditions approximately holds in the sense that when the condition is violated, (i.e. crossing the respective horizontal line more than once) the violation is small in magnitude. Recall that the single crossing propriety implies the unimodality of the error and unfairness terms. Small violations (both in magnitude and duration) of the single crossing condition amount to small changes in the derivative of error or unfairness, which in term does not consequentially impact the unimodality of either term from an empirical perspective.

E.5 Fair learning schemes

We make use of three fair learning algorithms to generate the fair models (denoted as f_F), namely GerryFair, Reductions, and EqOdds. Each algorithm takes as input a base-learner (not to be confused with the conventional classifier which we denote as f_C). This base-learner is used solve the fair learning objective through cost sensitive learning. Each algorithm uses their respective base learner in a unique way, and the fair models produced by each learning scheme different considerably in terms of their structure. Reductions uses the base-learner to perform traditional cost sensitive learning and outputs a model which is of the same type of the base-learner. For example if the base-learner is Logistic Regression, then f_F is also a Logistic Regression model. Thus the Projected Gradient Decent attack (PDG) is effective at computing an agents best response to f_F when f_F is learned via reductions and a differentiable base-learner (e.g. Logistic Regression, SVM, and Neural Networks). In the case of GerryFair the returned fair model f_F has a different structure from the base learner, namely f_F is an ensemble of models produced from the base learner. Thus the resulting model may not be smooth and PGD will not work to compute agents best response. In this case, we use the same local search attack used against Gradient Boosted Trees. In the case of EqOdds the resulting classifier is stochastic and predicts using the base-leaner with probability p_g and uses a trivial classifier (i.e. one that predicts the base rate) with probability $(1-p_g)$, for $g \in \{0, 1\}$. For each agent, it is always optimal to submit the features constituting an optimal response to the base-learner. The one difference when using EqOdds is that group membership now factors into classification. Since the agents utility is linear with respect to group selective, the best group in expectation is trivially computable.

E.6 Costs and agent best responses

For feature-monotonic costs we use $c(\mathbf{x}, \mathbf{x}')= ||\mathbf{x} - \mathbf{x}'||_2$ and for outcome-monotonic costs we use $c(\mathbf{x}, \mathbf{x}') = \max\left\{0, \frac{\mathbb{P}(y=1|\mathbf{x}')-\mathbb{P}(y=1|\mathbf{x})}{2}\right\}$.

All features are scaled to have range $[0, 1]$ and non-ordinal features are one-hot encoding. When computing the cost of a manipulation for feature-monotonic costs, we scale one-hot encoded and binary features by a factor of 0.2 (that is when computing the norm $x_i \in \{0, 0.2\}$ if x_i corresponds to a binary or one-hot feature. This scaling is intended to make manipulating categorical variables feasible. We noticed when categorical variables where not scaled

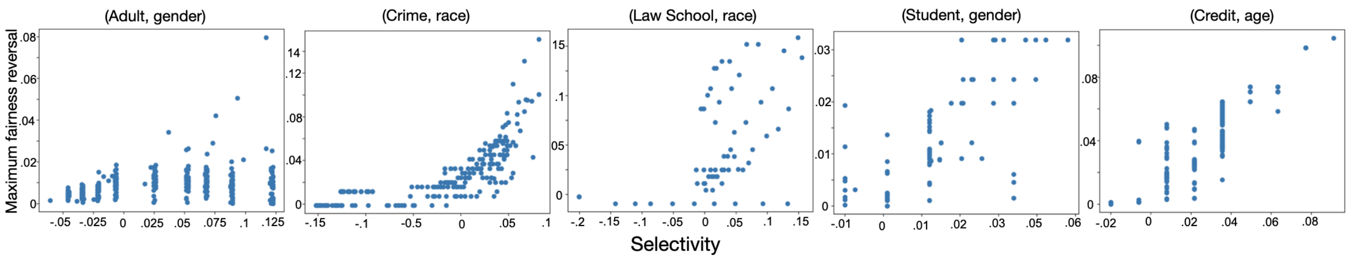

Figure 4: Maximum fairness reversal (y-axis) as a function of fair classifier selectivity (x-axis). Each point in each figure corresponds to a comparison between f_C and f_F (for a particular choice of α and fairness definition). Selectivity is defined as $\mathbb{P}(f_C(\mathbf{x}) = 1, f_F(\mathbf{x}) = 0) - \mathbb{P}(f_C(\mathbf{x}) = 0, f_F(\mathbf{x}) = 1)$, (i.e. selectivity shows the difference in the fraction of individuals who are positively classified by f_C but negatively classified by f_F, and the fraction of individuals who are positively included classified by f_F but negatively classified by f_C).

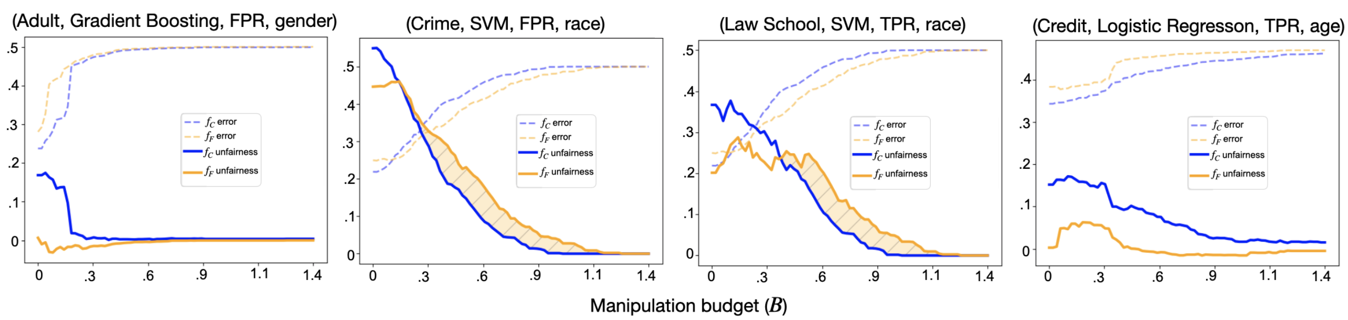

Figure 5: Unfairness (solid line) and error (dashed line) of the conventional classifier (blue) and fair classifier (orange), when the fair classifier is learned via the GerryFair algorithm. For both error and unfairness lower values are better. The shaded orange region indicates range of the manipulation budget B such that the relative fairness and accurate of the classifiers has swapped.

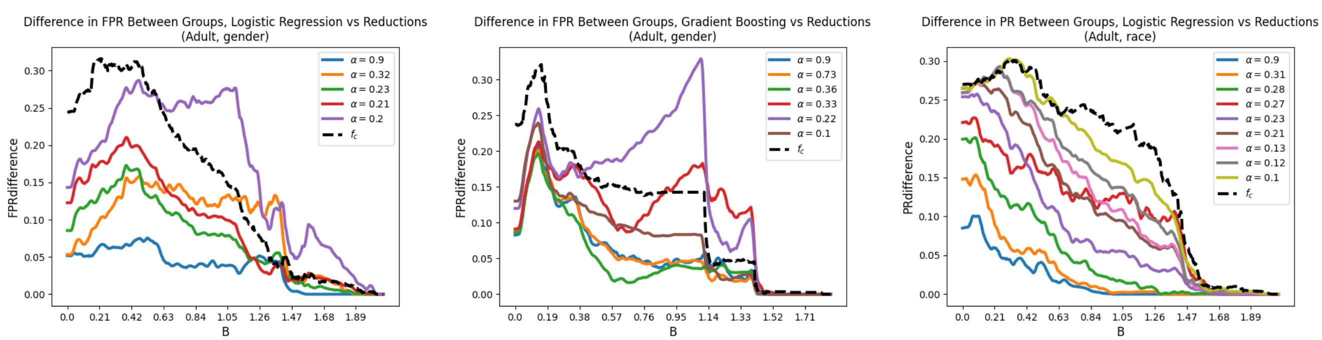

Figure 6: Fairness reversals on the Adult dataset with Reductions Classifiers. The y-axis displays unfairness between groups (lower is better). A fairness reversal occurs when a colored line (corresponding to f_F) is above the black dotted line (corresponding to f_C). Costs are feature-monotonic and Reductions is used as f_F. Only values of α leading to sufficiently distinct classifiers, compared to other values of α, are shown. The unfairness of most classifiers is approximately unimodal.

in this manner, the vast majority of agents manipulated only their ordinal features.

For outcome-monotonic costs we exclusively use the Community Crime dataset. Of the datasets we study, this is the only dataset for which computing the distribution of true labels $\mathbb{P}(y = 1|\mathbf{x})$, is feasible since the dataset has originally contains continuous labels (crimes per capita) which where made binary by thresholding on the 70^{th} percentile. As in [21] we normalize this value to lie between 0 and 1 and treat the value as a probability. Results relating to this outcome-monotonic costs are shown in 25.

E.7 Datasets

The following datasets are used in our experiments:

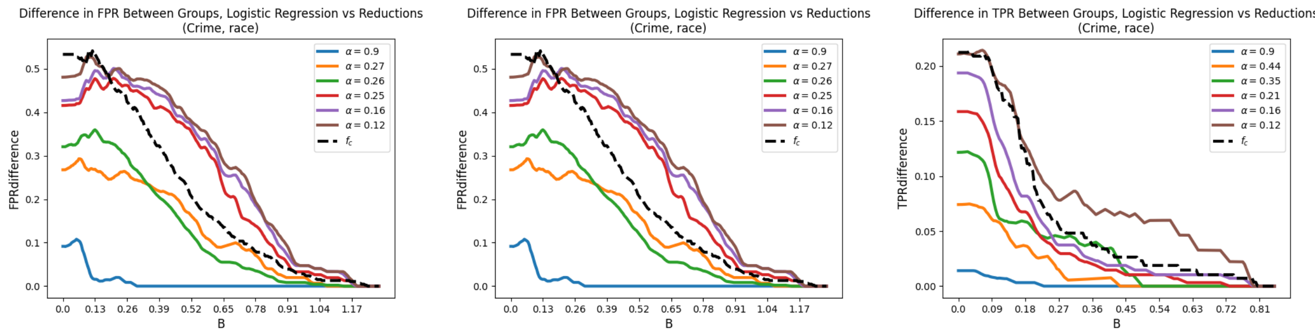

Figure 7: Fairness reversals on the Community Crime dataset with Reductions Classifiers.

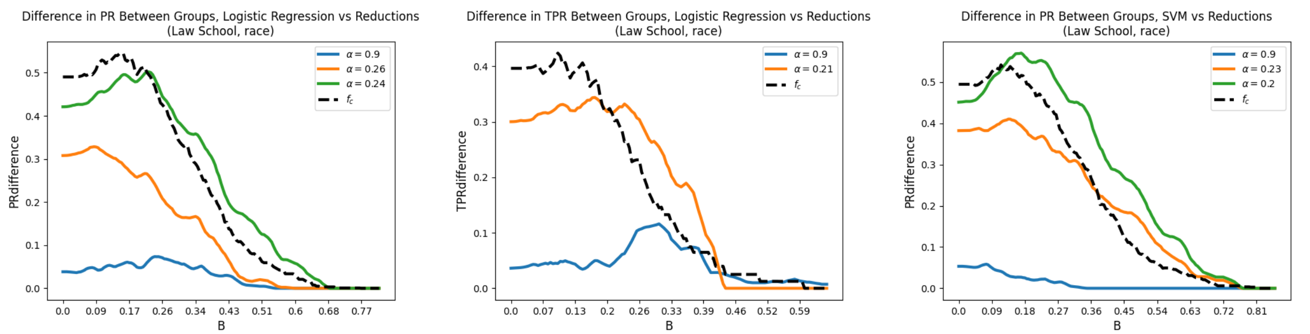

Figure 8: Fairness reversals on the Law School dataset with Reductions Classifiers.

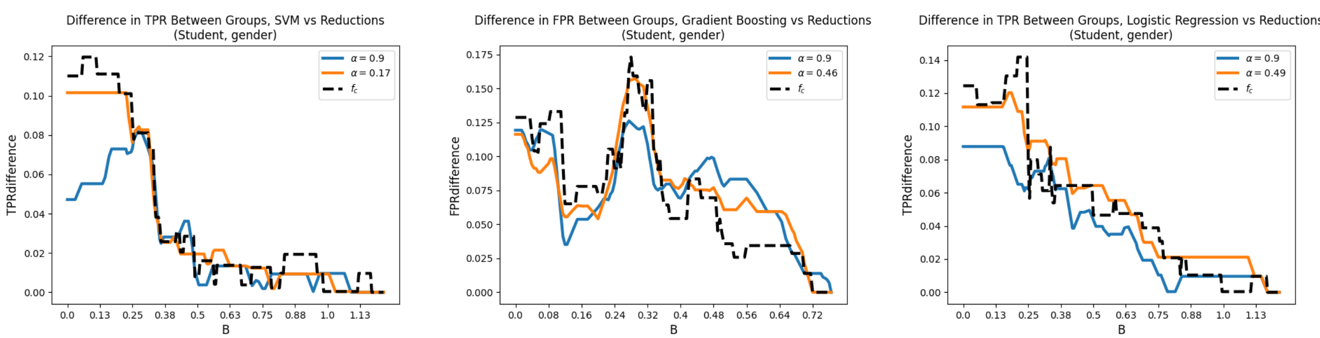

Figure 9: Fairness reversals on the Student dataset with Reductions Classifiers.

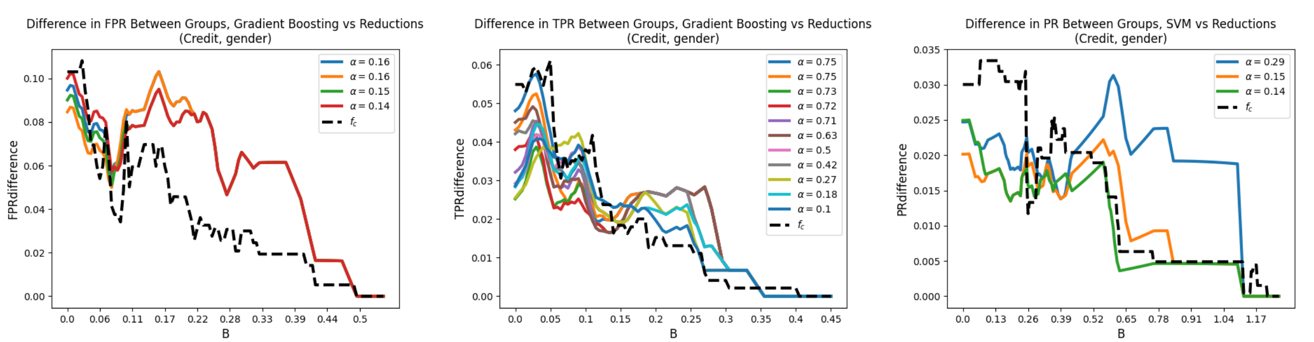

Figure 10: Fairness reversals on the Credit dataset with Reductions Classifiers.

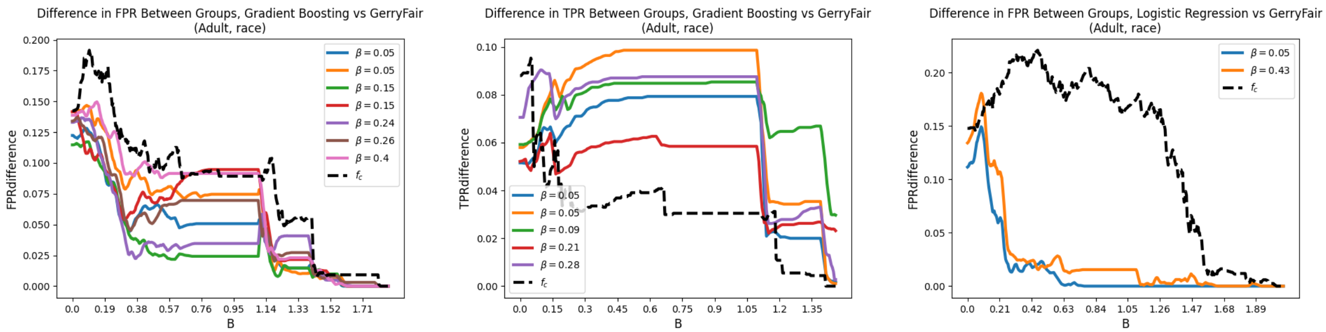

Figure 11: Fairness reversals on the Adult dataset with GerryFair. Costs are feature-monotonic. Only values of α resulting in sufficiently distinct classifiers are displayed.

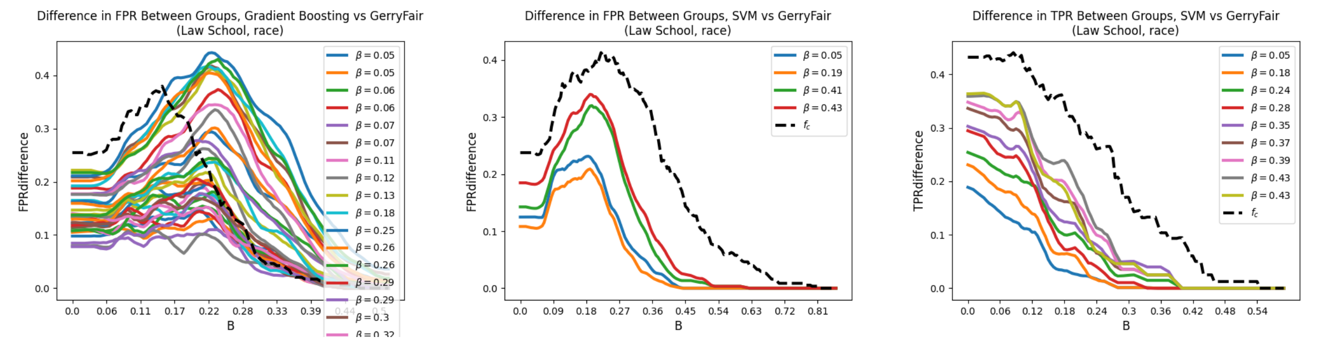

Figure 12: Fairness reversals on the Law School dataset with GerryFair.

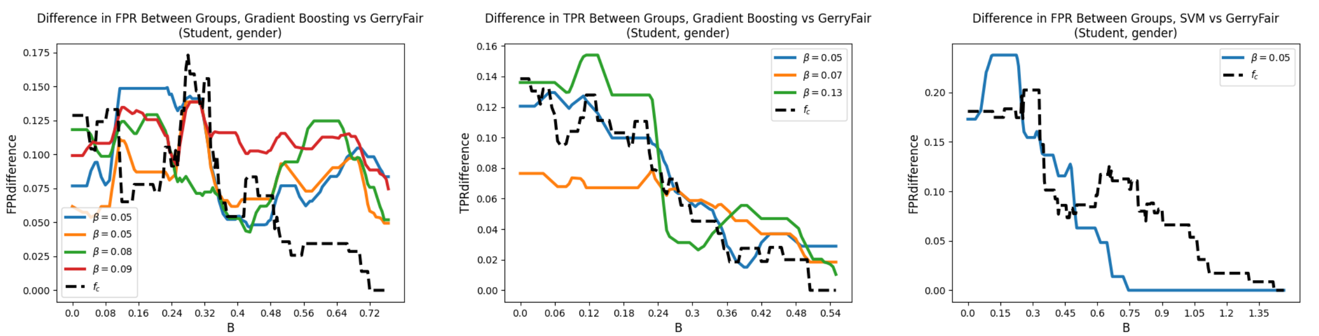

Figure 13: Fairness reversals on the Student dataset with GerryFair

Adult: Dataset of working professionals where the goal is to predict high or low income (protected feature: gender).
Community Crime: Dataset of communities where the objective is to predict if the community has high crime (protected feature: race).
Law School: Dataset of law students where the objective is to predict bar-exam passage (protected feature: race).
Student: Dataset of students where the objective is to predict a student receiving high math grades (protected feature: race).
Credit: Dataset of people applying for credit where the objective is to predict creditworthiness (protected feature: age).

Each dataset is prepossessed in the following manner. All sensitive features are removed from X, this includes age, race, gender, ethnicity, and other, (the feature which defines groups is saved, but included in X). If a dataset has class imbalance, such as the Law School or Crime datasets, the dataset is down-sampled to have $\mathbb{P}(y = 1) = 0.5$. All ordinal features are normalized and then scaled to have range $[0, 1]$, all non-ordinal categorical features are one-hot encoded.

In both the Community Crime and Credit dataset, the protected features (race and age respectively) is real valued. These are made binary by threshold on a particular value. In the case Community

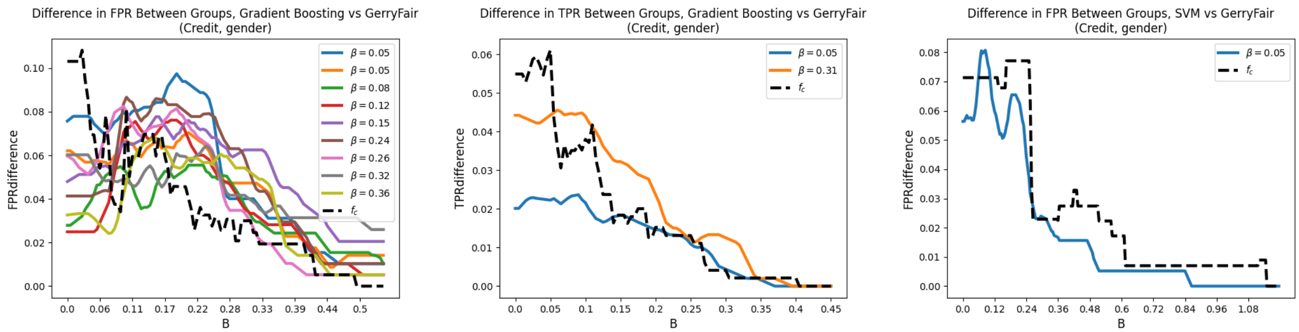

Figure 14: Fairness reversals on the Credit dataset with GerryFair

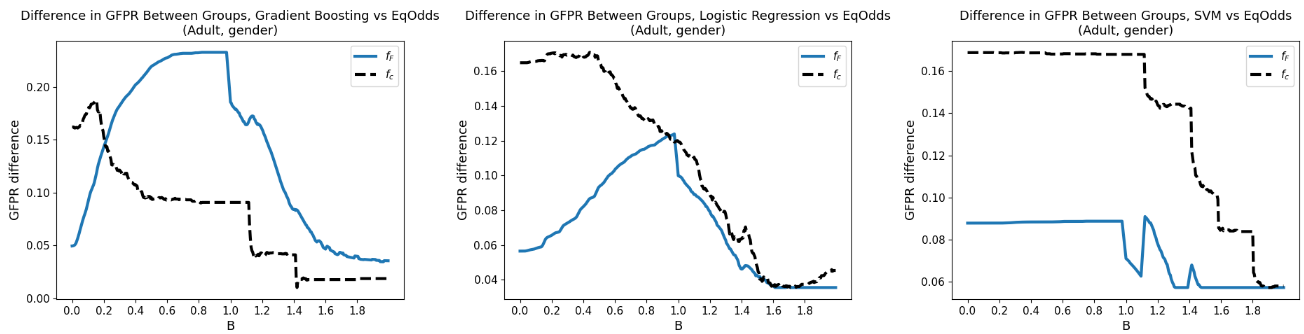

Figure 15: Fairness reversals on the Adult dataset when EqOdds (with GFPR fairness) is used as the fair learning scheme and costs are feature-monotonic. Misreporting group membership carries a flat cost of 1.

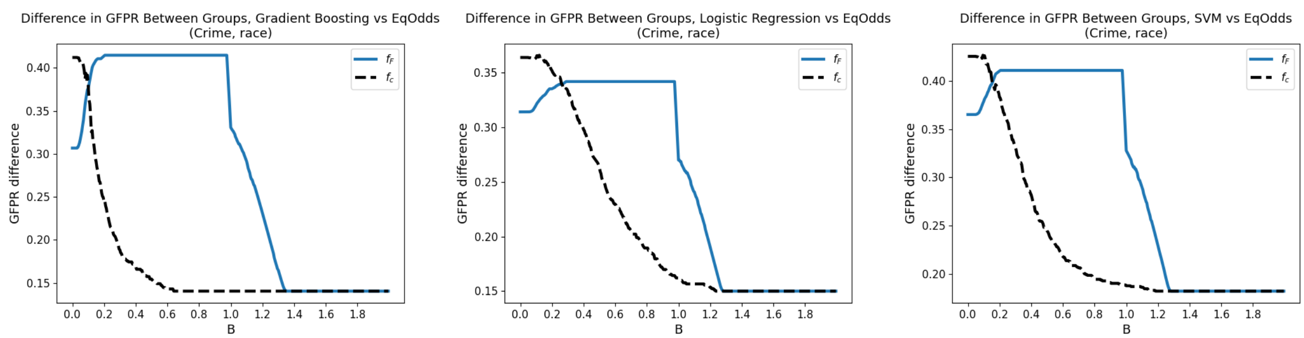

Figure 16: Fairness reversals on the Crime dataset when EqOdds (with GFPR fairness) is used for f_F.

Crime a community is said to be White if more than 70% of the population is classified as White. In the Credit dataset an applicant is considered to be Young if they are 25 or younger and Old otherwise.

Some datasets posses features which would be unrealistic to manipulate, such as information reported by law enforcement in the Community Crime dataset. We remove each non-manipulable feature from the dataset. Adult dataset, all features are considered manipulable. Community Crime: crime statistics, and police statistics are removed. Law School: which law school the student is attending (given in terms of school cluster) is removed. Student: number of filers is removed. Credit: all features are considered manipulable.

E.8 Fairness Reversal

Recall that in the single variable case, strategic manipulation leads to a fairness reversal between the base and fair thresholds θ_C and θ_F respectively, if and only if $\theta_C < \theta_F$. Figures 21-24 show the relationship between θ_C and θ_F for each of the variables, dataset, and fairness metrics we study. In these figures we see that $\theta_C < \theta_F$ is a common. Moreover, we see that the cases where this relationship does not hold are cases in which either $\theta_C < \theta_U$ (meaning the

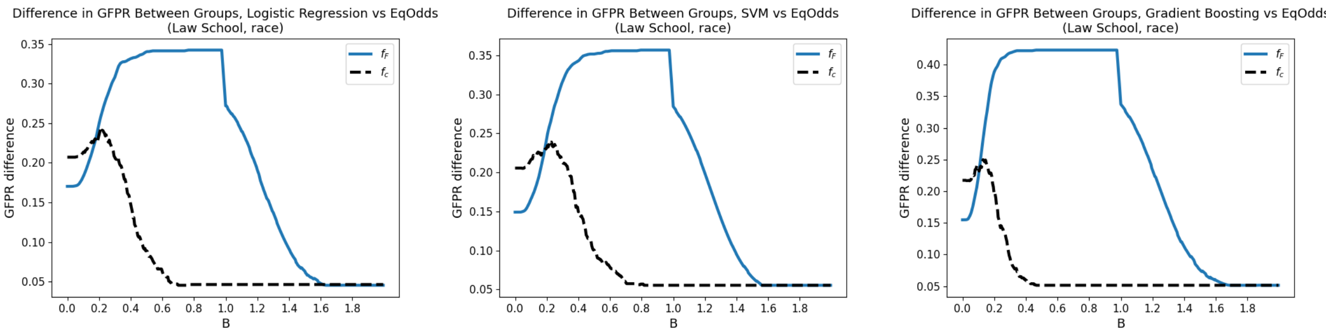

Figure 17: Fairness reversals on the Law School dataset when EqOdds (with GFPR fairness) is used for f_F.

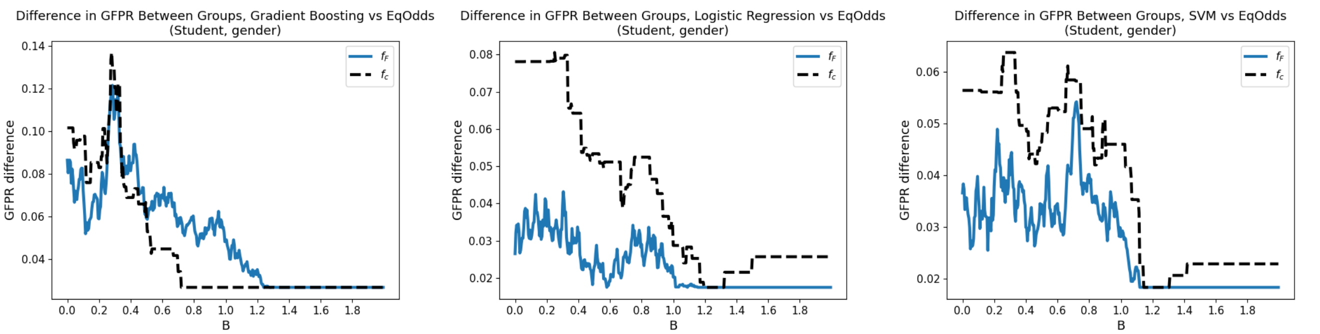

Figure 18: Fairness reversals on the Student dataset when EqOdds (with GFPR fairness) is used for f_F.

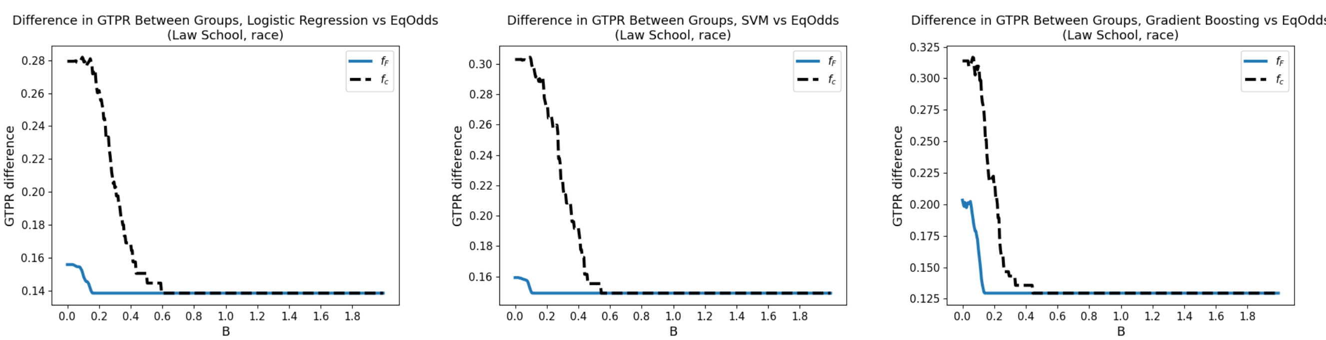

Figure 19: Fairness reversals on the Law School dataset when EqOdds (with GTPR fairness) is used for f_F.

sufficient condition of Theorem B.2 does not hold), the fair classifier is trivial (i.e. $\theta_F = 0$), or there is negligible unfairness regardless of the value selected for θ. Moreover, we see that both error and unfairness are unimodal w.r.t. θ, thus Lemma B.1 implies that error and unfairness will remain unimodal w.r.t. the manipulation budget B for *any* manipulation cost function $c(x, x')$ which is monotone in $|x' - x|$.

With respect to Figures 20-24, agent manipulation amounts to "moving" each threshold to the left. We can see that when $\theta_C < \theta_F$, moving θ_C to the left decreases unfairness, while moving θ_F to left increases unfairness, until the manipulated θ_F has been moved all the way to θ_U (the most unfair threshold). Additionally in these figures we see that not only does this leftward shift increase the unfairness of θ_F, but also increased the accuracy of θ_F: a phenomenon outlined by Theorem 4.3. That is, in the cases where $\theta_C < \theta_F$, strategic manipulation leads to both a fairness, and an accuracy, reversal between θ_C and θ_F.

In the multivariate case, Figures 6-18, show that again the fairness reversal is common. Moreover, as was the case in the single variable case, we see that in the multivariate case both error and unfairness exhibit unimodality w.r.t. to the budget B.

In the single variable case, we would expect that once f_C and f_F respectively hit the point with maximum unfairness (as a function of B) their unfairness would decrease at an equal rate from that

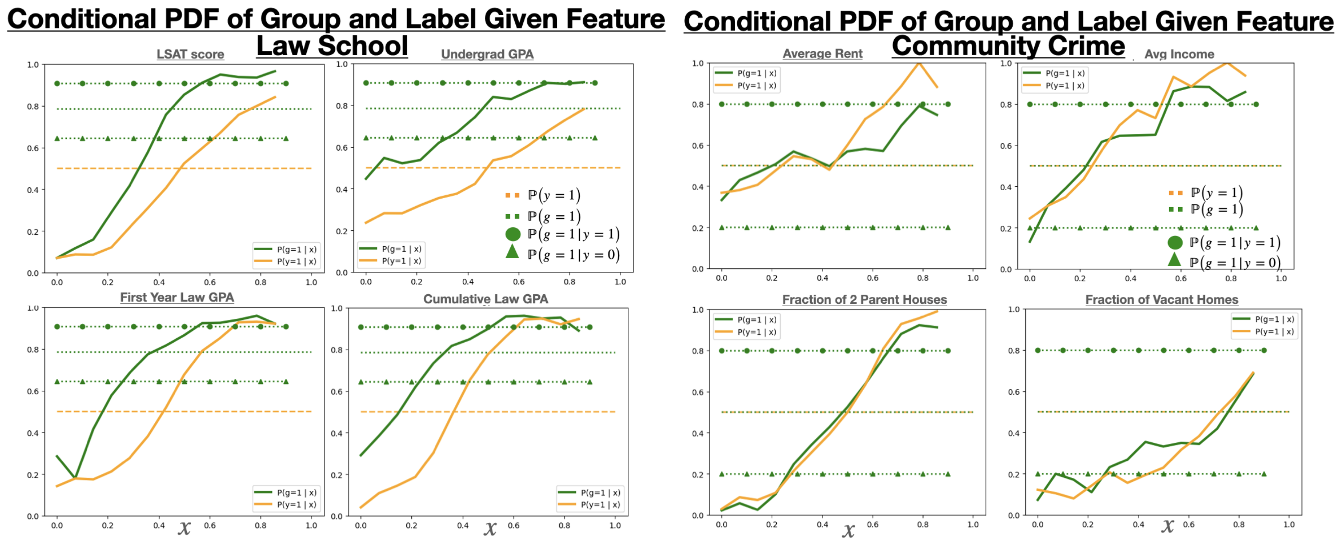

Figure 20: Probabilities of group membership g (green) and true label y (orange). Probabilities conditioned on the feature x are given as solid lines, while those unconditioned are given as dotted, or dashed, lines. Recall that if the conditioned probabilities $\mathbb{P}(g=1|x)$ and $\mathbb{P}(y=1|x)$ having a single crossing with the respective unconditioned value (outlined in Lemmas B.3, B.4, B.5) then error and unfairness will be unimodal w.r.t. to the threshold θ. For example, in the case of PRfairness, if $\mathbb{P}(g=1|x)$ has a single crossing with $\mathbb{P}(g=1)$ and $\mathbb{P}(y=1|x)$ has a single crossing with $\mathbb{P}(y=1)$ then error and unfairness are unimodal w.r.t. to θ.

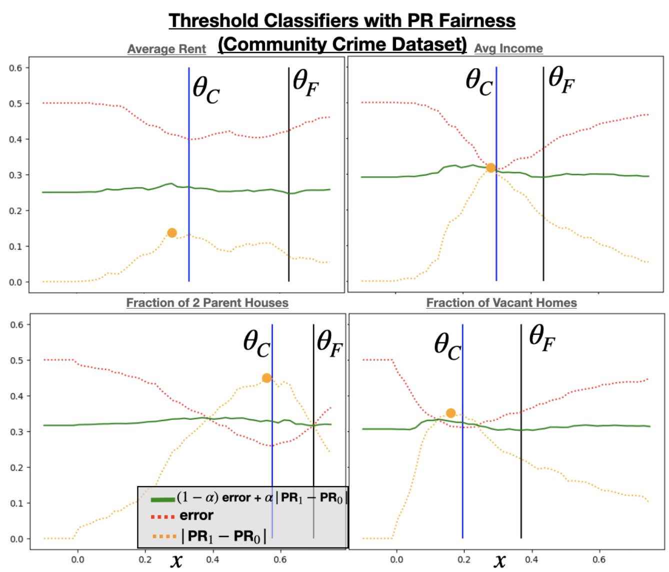

Figure 21: Unfairness and error of threshold classifiers. Both error and unfairness are approximately unimodal w.r.t. threshold $\theta = x$. Thus error and unfairness are also unimodal w.r.t. the manipulation budget B for any manipulation cost function $c(x, x')$ which is monotone in $|x' - x|$. When this unimodality holds $\theta_C < \theta_F$ implies that strategic manipulation will lead to θ_C becoming more fair than θ_F. This fairness reversal is due to the fact that strategic manipulation amounts to lowering (shifting to the left) the threshold. In this figure, as well as the subsequent figures, we see that $\theta_C < \theta_F$ is a common occurrence (namely 30 our of the 36 combinations of variable, fairness metric, and dataset studied).

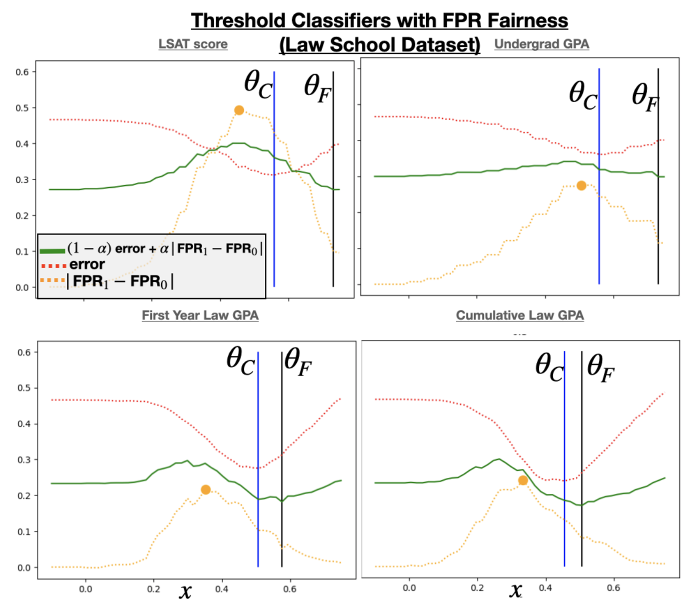

Figure 22

point onward since both classifiers are effectively sharing the same unfairness curve, but sit at different points. In the multivariate case, we make this same observation. After reaching the most unfair B, both classifiers decreases at similar rates. However, f_C requires a

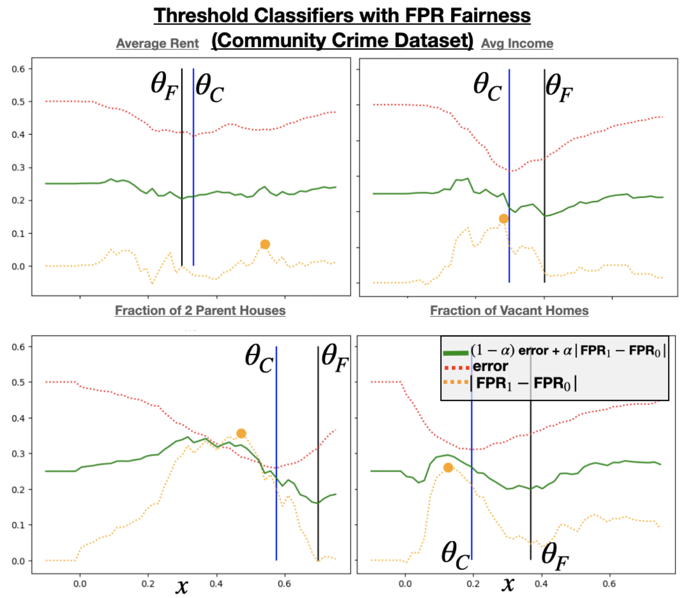

Figure 23

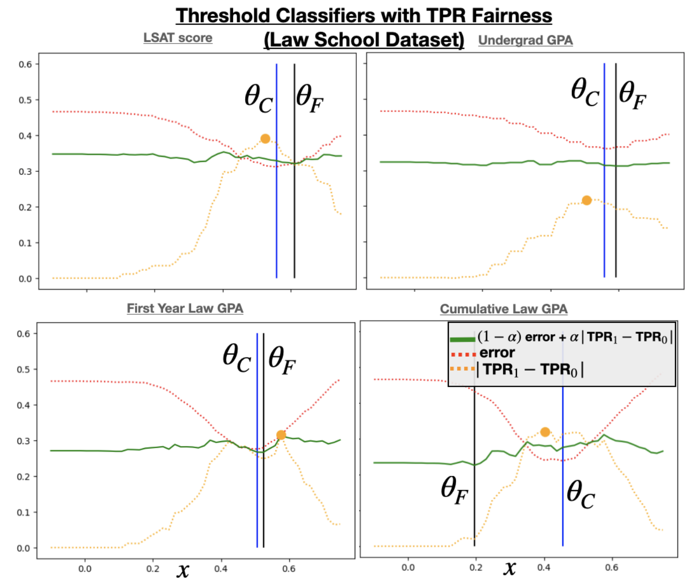

Figure 24

larger B, than f_C, to reach this point. Which ultimately leads to f_F becoming less fair, since the unfairness of f_F is still increasing while the unfairness of f_C has already begun to fall.

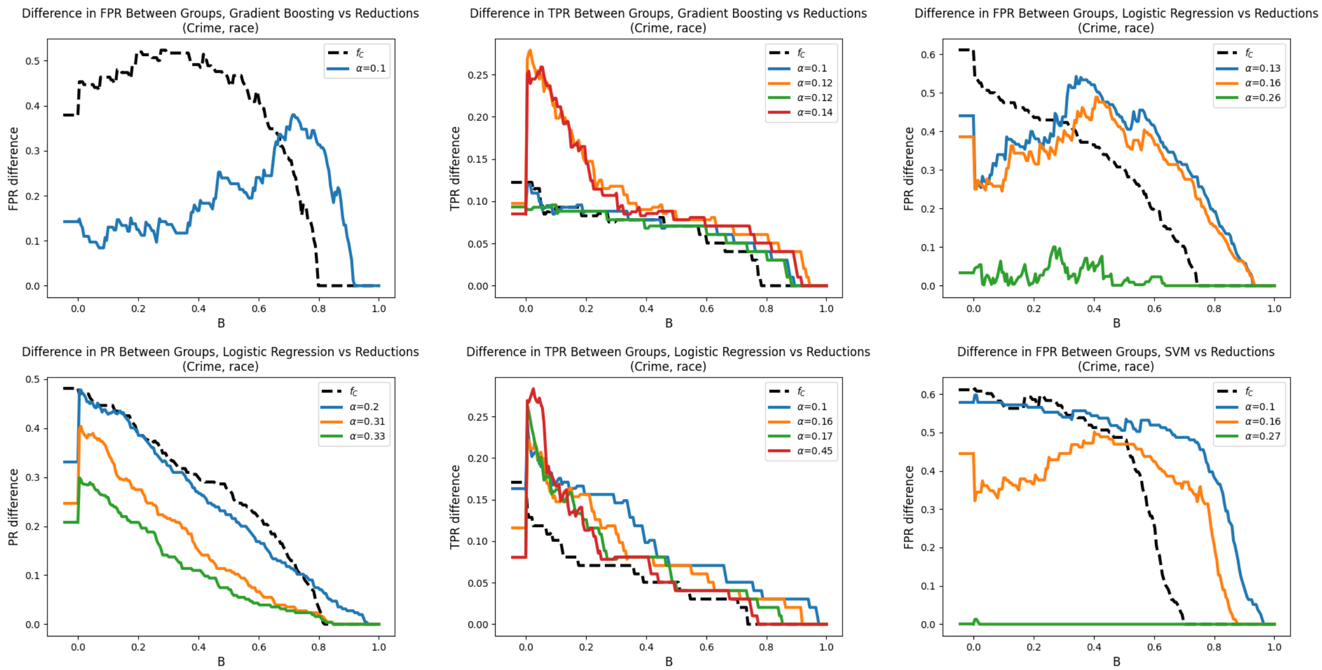

Figure 25: Fairness reversals on the Community Crime dataset when costs are outcome-monotonic. Each line represents the the difference in PR, FPR, or TPR between groups (defined by race) as a function of the manipulation budget B. Costs are outcome-monotonic. The y-intercept of each plots shows the respective unfairness of each classifier with no strategic behavior.