Addressing Strategic Manipulation Disparities in Fair Classification

Vijay Keswani
L. Elisa Celis
vijay.keswani@yale.edu
elisa.celis@yale.edu
Yale University
New Haven, USA

ABSTRACT

In real-world classification settings, such as loan application evaluation or content moderation on online platforms, individuals respond to classifier predictions by strategically updating their features to increase their likelihood of receiving a particular (positive) decision (at a certain cost). Yet, when different demographic groups have different feature distributions or pay different update costs, prior work has shown that individuals from minority groups often pay a higher cost to update their features. Fair classification aims to address such classifier performance disparities by constraining the classifiers to satisfy statistical fairness properties. However, we show that standard fairness constraints do not guarantee that the constrained classifier reduces the disparity in strategic manipulation cost. To address such biases in strategic settings and provide equal opportunities for strategic manipulation, we propose a constrained optimization framework that constructs classifiers that lower the strategic manipulation cost for minority groups. We develop our framework by studying theoretical connections between group-specific strategic cost disparity and standard selection rate fairness metrics (e.g., statistical rate and true positive rate). Empirically, we show the efficacy of this approach over multiple real-world datasets.

CCS CONCEPTS

• **Computing methodologies** → **Learning paradigms**.

KEYWORDS

fair classification, strategic manipulations, social burden gap

ACM Reference Format:
Vijay Keswani and L. Elisa Celis. 2023. Addressing Strategic Manipulation Disparities in Fair Classification. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3617694.3623252

EAAMO '23, October 30–November 01, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/10.
https://doi.org/10.1145/3617694.3623252

1 INTRODUCTION

In prediction/classification settings, the goal is to develop automated models that accurately predict class labels using the available demographic and task-specific features of individuals. The use of predictive models in many real-world applications, however, impacts the features of the underlying population. One direct way this happens is when individuals take steps to update their features to potentially obtain a different prediction in the future. In binary classification, where positive class labels can denote *success* for a given task, individuals who have been negatively classified will attempt to update their features in a manner that increases their likelihood of receiving a positive decision in the future.

There are numerous examples of such individual behavior in response to institutional decisions. Consider the setting of loan applications, where the features are individuals' demographics, annual income, credit history, number of dependents, the number of open credit lines, etc. The class label to be predicted is whether an individual will default on a loan or not. The number of open credit lines is a feature that is often positively correlated with the class label and individuals can increase their likelihood of positive loan application (or increase their *credit score*) by opening more credit lines[1]. However, opening credit lines requires additional investment on the part of the individuals [8]. Another example is social media websites and online platforms. Even without a complete understanding of a platform's recommendation system, users nevertheless attempt to intervene in different ways to exercise control over the platform's algorithms [40]. For example, content moderation tools used in social media platforms flag objectionable posts, which are then suppressed by the recommendation system to ensure low visibility [17] (often unfairly targeting minority voices [20, 48]). Users, in this case, curate and modify their content to work around the platform's decision [5, 40]. Beyond content moderation, strategic manipulation can allow users to avoid harassment, as seen in the case of Twitter [5]. Evidence of users' attempts to exercise control over an online platform's algorithms has similarly been observed in ride-hailing apps like Uber and Lyft [36]. A final example is the setting of college admissions, where the features are individuals' demographics, school academic records, extra-curricular records, and scores from standardized tests like GRE. The class label to be predicted is the likelihood of academic "success" to determine college admissions. In this case, while higher scores for standardized tests increase the chances of a successful college application, students

[1]For the sake of simplicity, assume all the other features are unchanged; in real-world scenarios, there will be simultaneous dependence on other variables as well here, e.g. whether the individual has been regular with their payments or not.

can take these tests multiple times and submit only the highest scores. Nevertheless, there is an additional investment required as every additional test attempt involves monetary and time expenses [42]. Feature manipulations of these kinds can also take the form of positive steps taken by individuals to improve their features (e.g., investing additional time in test preparation) [2, 30] and/or provide individuals with the agency to address model decisions [39, 41].

The above-described process involves two main players: the institution constructing a classifier and the individuals reacting to the classifier. While the institution's goal is to minimize prediction error (or maximize a certain measure of utility), individuals react to the classifier predictions by *strategically manipulating* their features to achieve a positive classification. In these strategic settings, often due to historical biases, the classifier employed by the institution can pose relatively higher costs for strategic manipulation (i.e., increased costs to improve their feature values) for individuals from minority groups (e.g., race and gender minorities). Prior work has observed such disparities in settings where the datasets used for training the classifier encode social biases or when minority groups pay larger costs to update their features [26, 34]. For example, in the case of loan applications, historical discrimination against African Americans in financial aspects often deters them from seeking new credit lines [46]. In the case of social media platforms, content moderation tools exhibit bias against minority groups, for example, by reducing the visibility of posts by advocates from minority groups [3, 20, 24] or by using biased sentiment analysis tools [10, 29]; these biases lead to greater hurdles for these groups to make their voices heard. Similarly, for graduate school admissions, Wilson [45] revealed limitations of GRE and UGPA scores in predicting graduate school success for Black students. Using these scores without considering the racial disparities can create a higher admission barrier for Black students. These biases are a result of negative stereotypes and/or historical lack of opportunities for minority groups and classifiers that inherit such biases can further propagate them. In the presence of these biases in the predictions of trained classifiers, one can ask whether an institution can construct classifiers that provide equal opportunities for strategic manipulation to all groups and address the systemic disparities in investment required to improve their outcomes. Strategic manipulation opportunities often serve as mechanisms to provide individuals with *recourse* or *agency* against biased institutional decisions [41]. As such, equalizing manipulation opportunities will ensure that majority groups do not solely take advantage of effective strategic manipulations and provide similar power to minority groups to address classifier decisions.

Fairness-constrained classification attempts to address such disparities in classifier performance by constraining the classifier to satisfy certain statistical fairness properties. For example, when constraining with respect to *statistical rate*, the classifiers are constrained to have an almost-equal selection rate for all groups [1, 6, 13, 38, 47]. Similarly, when constraining with respect to *equalized odds*, the classifiers are constrained to have equal false positive and true positive rates for all groups [6, 22, 38]. However, these fairness metrics and constraints operate in a *static* manner and do not take into account the response of the individuals to the classifier predictions or the disparity in costs that different groups pay for updating their features. Even though fairness constraints encourage the increased selection of minority group individuals, existing dataset biases or update cost disparities can still disproportionately affect the negatively-classified individuals in the minority group. *Correspondingly, the primary question we investigate is the following: Do constraints that use standard static fairness metrics lead to classifiers that reduce strategic manipulation costs for minority groups?*

Our Contributions. We first theoretically study the relationship between standard selection rate fairness metrics (like statistical rate and true positive rate disparity) and the disparity in strategic manipulation costs between majority and minority groups when only one-dimensional features and group membership of individuals are provided (Section 3). Our analysis shows that threshold-based classifiers that have an equal selection rate for all groups can still have higher strategic manipulation costs for the disadvantaged groups when feature distributions or cost functions differ across groups. (Theorem 3.3, 3.4). Prior works on strategic cost disparities only demonstrated that this disparity can be large in unconstrained settings [26, 34]. Our analysis demonstrates that even "fair" classifiers that are constrained using selection rate fairness metrics can still have large strategic cost disparities. To address this bias, we bound the strategic cost disparity using the statistical properties of the classifier and the cost function. Using these bounds as constraints, we construct classifiers that have both low selection rate disparity and low strategic cost disparity. We also extend the results to multi-dimensional settings when the classifier is linear and the cost function is linear or quadratic (Section 4). The primary technical challenge we face in proving our results is accounting for all factors that result in strategic cost disparity. As we discuss in Section 3, this disparity can arise due to multiple reasons, such as unequal group selection rates, variation in cost functions across groups, and "distance" of negatively classified individuals from classifier thresholds. Correspondingly, our results quantify the relationship between the strategic cost incurred by each group and the group's selection rate using all relevant factors, including bounds on the cost function gradient and other related empirical properties of the classifier. Using these bounds, we can construct appropriate classifiers that minimize strategic cost disparity. Our theoretical results are complemented by empirical analysis on two real-world financial datasets: the FICO credit dataset [22] and the Adult income dataset [11] (Section 5). For both datasets, we show that fair classification with our proposed constraints leads to lower manipulation costs for the minority group.

Related Work. Studies by Milli et al. [34] and Hu et al. [26] first analyzed strategic manipulation cost disparities when feature distributions or cost functions are biased against minority groups. However, their analysis is limited to classifiers that optimize institution utility; in contrast, we also study classifiers that optimize utility subject to standard fairness constraints. Estornell et al. [14] and Braverman and Garg [4], on the other hand, assess classifier fairness in strategic settings using only selection rate fairness metrics. Estornell et al. [14] observed that statistical parity or equalized odds constrained classifiers become less "fair" (with respect to the same metrics) than unconstrained classifiers due to strategic manipulations. Braverman and Garg [4] study the impact of randomness on classifiers trained in strategic settings and propose the use of *noisy*

features to address selection rate disparities in the outputs of these classifiers. Like our work, both these papers evaluate the impact of fair classification in strategic settings; however, they analyze the fairness of final individual outcomes using only selection rate metrics and do not consider the costs disparity across groups. Similar to strategic updates, Ustun et al. [39] consider the notion of *actionable recourse* and provide tools to minimize recourse cost for linear classifiers. However, their work does not aim to address recourse cost disparities. Gupta et al. [19], von Kügelgen et al. [43] extend this line of work to study recourse disparities in classification; however, their models only handle settings where cost functions are the same for all groups. As noted in multiple prior studies [9, 41], minority group individuals often pay larger costs to update their features, which then leads to recourse disparities. Our framework is, hence, more generic (than [19, 43]) as it tackles both cost and feature disparities.

Static fairness constraints in non-strategic settings, that compare the selection rate of majority and minority groups, have been extensively studied in the context of constructing fair classifiers [1, 6, 13, 28, 31, 38, 47, 49, 50]. For non-strategic settings, Hu and Chen [25] show that selection rate-constrained classifiers may not improve the average quality of predictions received by the disadvantaged groups. We extend this direction to analyze the impact of fair classification in strategic settings. Recent work on strategic settings has also studied classifiers that are robust to strategic updates [7, 12, 21, 23, 27, 30]. The analysis in these papers is primarily from the viewpoint of an institution maximizing its utility given information about individuals' behavior and these papers do not consider the fairness goal of reducing manipulation costs disparities with respect to protected attributes. Our paper, instead, considers the individuals' perspective and addresses the cost disparities arising from group memberships. While we look at the one-step feedback models, *performative prediction* algorithms model multi-step feedback settings to construct classifiers that are stable over induced distributions [33, 37]. For ease of analysis, we limit our study to one-step feedback settings.

2 MODEL FORMALIZATION

Let $x \in \mathcal{X} \subseteq \mathbb{R}^d$ denote the features of an individual in the population, $y \in \{0, 1\}$ denote the true class label to be predicted and $z \in \mathcal{Z}$ denote the protected attribute (assumed to be binary for our current analysis). We will use $\mathcal{D}$ to denote the underlying joint distribution of features, class labels, and protected attributes, and let X, Y, Z denote the respective random variables. We will work with threshold-based classifiers $f : \mathcal{X} \to \{0, 1\}$ which set a threshold on the likelihood of any point achieving a positive class label [2].

Strategic manipulations and individual cost functions. As mentioned earlier, individuals can update or manipulate their features at a certain cost after observing a classifier prediction. Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote the cost function such that $c(x, x')$ is the cost paid by an individual to update their feature from x to x'. The subsequent utility gained by the individual from this update can be quantified as $u_x(f, x') := f(x') - c(x, x')$. In this setting, the optimal feature update for an individual (in response to a classifier f) is captured by $\Delta_f(x) := \arg\max_{x'} u_x(f, x')$. Since we only aim to model updates that lead to improved classifier prediction, we will study cost functions that have no feature update cost if the individual is already positively classified. In other words, individuals are *rational* and aim to maximize their utility (this assumption is consistent with prior work on strategic settings [21, 26]).

The institution's aim, in the unconstrained setting, is to minimize error w.r.t. a given loss function $\mathcal{L}$, i.e., find the classifier f that minimizes $\mathbb{E}_{\mathcal{D}}[\mathcal{L}(f; X, Y)]$. When using unmanipulated data, this loss will be a proxy measure for $\mathbb{P}_{\mathcal{D}}[f(X) = Y]$, while for manipulated data, this loss will be a surrogate for $\mathbb{P}_{\mathcal{D}}[f(\Delta_f(X)) = Y]$ (i.e., the standard accuracy measure in strategic classification [21, 32, 33]). Any common classification loss function can be used for $\mathcal{L}$; e.g., we use the log-loss function in some of our simulations. Other common loss functions, such as mean square loss, hinge loss, or regularized versions of these functions, can also be used with our framework. However, to incorporate fairness in this optimization program, we need additional fairness constraints.

Ground truth label y. Feature manipulations represent actions or changes that are under individual-level control, such that positively manipulating the relevant features can potentially lead to a change in the classifier decision for this individual. In a variety of real-world settings, individual actions to update features x can either change their ground truth label y or not affect their ground truth label depending on the nature of the update and the context. In all cases, it is important to ensure that equal opportunities for manipulations are available across demographic groups. However, in this paper, we primarily consider the settings where strategic manipulations are used as a recourse option to address unfair institutional decisions. While our model and theoretical analysis can handle both settings (i.e., when manipulations change ground truth and when they don't affect ground truth), our empirical analysis will primarily focus on cases where ground truth remains unchanged due to strategic updates, as these updates capture recourse strategies. This assumption is also consistent with other works on strategic or adversarial manipulations [14, 18, 21].

To see why it is important to study strategic manipulations as a recourse option we present a few examples where feature updates do not lead to a change in ground truth label y but still provide valuable agency to individuals.

– *Example (1):* On online platforms, costs associated with individuals' actions can be seen to depend on a variety of factors. Many studies have reported that Black activists face higher levels of censorship on social media platforms simply due to mentions of race-associated terms [20]. To counter this, such activists have to manipulate their posts (e.g., by changing certain words or using screenshots) to get around the automated moderation tools, paying a cost in terms of time and resources required for such manipulation. Note that, these actions do not change the ground truth label y of the post (i.e., the post continues to remain non-offensive), yet the censored individuals have to take action and pay associated costs so that the automated system aligns with their ground truth label.

– *Example (2):* In a lending situation, an individual's credit score is an important factor when evaluating their loan application. However, many studies have shown that changes in credit scores are related

[2] Any hypothesis class where the classifier output is a distribution over the labels (e.g., logistic regression, Naive Bayes and MLPs) can be represented using threshold-based classifiers and correspondingly used in our framework.

to factors beyond an individual's financial credibility. Take the following example from a CNBC article[3]: two people with equal annual income, equal credit card debt, and equal credit limit got the $1200 stimulus check. The first one used the entire amount to pay off the credit card debt while the second one used $600 towards paying off their debt and used $600 for their savings. However, due to different *credit utilization* rates, the first person's credit score will be higher than the second person's score. Both individuals in this case have the same income/resources and, if they both applied for a loan, they arguably would have a similar likelihood to pay back the loan (implying no significant changes in ground truth for default risk y). Yet, because the first person has a higher credit score, any decision-making policy that uses credit scores would prefer the first person for the loan. Hence, individual actions affect classifier decisions, sometimes independent of ground truth.

The above examples are settings where individuals can use strategic manipulations as a recourse option when they believe that the institution's decision is not correct. These examples also make it clear that addressing incorrect decisions can require monetary or time investments by individuals. In these examples and numerous other settings explored in prior work [3, 10, 24, 45], strategic manipulations are used by individuals to exercise control over institutional decisions. We primarily analyze strategic manipulations in these recourse contexts because our analysis considers the perspective of the individuals and the agency provided to them to address unfair institutional decisions. The institution can use a classifier that either simply maximizes their utility/accuracy or they can use one that is fair with respect to standard selection rate fairness metrics. We evaluate the impact of such classifiers on individuals from different groups and the average cost they have to pay to positively manipulate their features based on their group membership.

Selection rate fairness metrics. The protected attribute $z \in \mathcal{Z}$ is the focus of our analysis of fairness. Standard fair classification algorithms measure fairness using group-specific selection rates, either over the entire population or over certain subpopulations [6, 35]. For any sub-population condition $\psi : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, the (conditional) selection rate of a classifier f with respect to protected attribute group z can be defined as $H_z(f, \psi) := \mathbb{P}_{\mathcal{D}}[f(X) = 1 \mid \psi(X, Y) = 1, Z = z]$. With respect to this definition, the conditional selection rate fairness of f can be quantified as

$$H(f, \psi) := H_0(f, \psi) - H_1(f, \psi).$$

If the condition is identity, i.e. $\psi(x, y) = 1$, then $H_z(f, \psi)$ simply measures the fraction of elements in group z that are positively classified and $H(f, \psi)$ in this case is the standard statistical rate metric [13]. If the condition is $\psi(x, y) = \mathbf{1}(y{=}1)$, then $H_z(f, \psi)$ measures the true positive rate for group z, and $H(f, \psi)$ is the true positive rate disparity across the protected attribute groups [22, 47]. Using the above definition of H, all standard *linear fairness metrics* considered in Celis et al. [6] can be represented in additive form. When clear from context, we will use shorthand ψ to denote $\psi(\cdot, \cdot)$.

Strategic cost disparity. The power of strategic manipulation can be different for different demographic groups, which is the primary kind of bias we tackle in this paper. As mentioned earlier, these biases can occur when the underlying distributions vary across groups due to possibly different historical evolution trajectories followed by group-specific distributions [46], or when one group pays larger update costs than others for similar updates [15].

For a classifier f, the expected cost incurred by individuals from group $Z = z$ can be measured using $\mathbb{E}_{\mathcal{D}}[c(X, \Delta_f(X)) \mid Z = z]$, a quantity referred to as the *social burden* for the group z by Milli et al. [34]. Therefore, one measure of fairness we can look at in this strategic setting is the following gap: $\mathbb{E}[c(X, \Delta_f(X)) \mid Z = 0] - \mathbb{E}[c(X, \Delta_f(X)) \mid Z = 1]$. Higher values ($> 0$) of this quantity imply that individuals from group 0, on average, pay a larger cost to strategically manipulate their features than individuals from group 1. While the above measure evaluates the cost for all individuals in each group, different contexts might require focusing on different sub-populations of individuals from each group. E.g., in the recidivism risk assessment setting [44], we may want to analyze the average cost paid by a low-risk individual from the minority group who has been deemed high-risk to overturn the classifier decision. In this case, the expected cost $\mathbb{E}[c(X, \Delta_f(X)) \mid Y = 1, Z = z]$ is more relevant ($Y = 1$ denotes low-risk). Hence, in general, for any classifier f, we can define the social burden for any group z with respect a given sub-population condition $\psi : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ as $G_z(f, \psi) := \mathbb{E}[c(X, \Delta_f(X)) \mid \psi(X, Y) = 1, Z = 0]$ and, correspondingly, define the *social burden gap* as

$$G(f, \psi) := G_0(f, \psi) - G_1(f, \psi).$$

Classifiers that equalize manipulation costs ($G(f, \cdot) = 0$) ensure that all groups have similar manipulation power. In certain cases, we might even require $G(f, \cdot)$ less than 0 to counter historical inequalities faced by disadvantaged groups. Hence, our goal is to provide an optimization framework to construct classifiers with a desired social burden gap.

3 LINKING SELECTION-RATE FAIRNESS AND SOCIAL BURDEN GAP IN ONE-DIMENSIONAL SETTING

We first look at the case when the features are one-dimensional and positive, i.e., $\mathcal{X} = \mathbb{R}_{\geq 0}$. This setting models several real-world scenarios such as the use of credit scores for loan applications or exam scores for school admissions. Furthermore, when the likelihood of positive classification ($\mathbb{P}[Y{=}1 \mid X{=}x]$) can be computed (even approximately), one can use the likelihood as the feature for classification (similar to the model of Milli et al. [34]). Secondly, we will assume outcome monotonicity of the cost function with respect to the feature: if $\tau{>}x_1{>}x_2$, then $c(x_2, \tau){>}c(x_1, \tau)$. In this case, threshold-based classifiers will classify all individuals with feature values greater than a specific threshold as positive and all individuals with feature values less than the threshold as negative. As mentioned before, we study cost functions that only have non-zero costs for the individuals classified as negative. Hence, we assume that the cost function c has the following property: $c(x_1, x_2)$ is non-zero (and positive) only when $x_1{<}x_2$; i.e., for a continuous and differentiable function $d : \mathcal{X}{\times}\mathcal{X}{\to}\mathbb{R}$, we can say that $c(x_1, x_2) = d(x_1, x_2) \cdot \mathbf{1}(x_2{>}x_1)$. Due to outcome monotonicity, the gradient of $c(x_1, x_2)$ with respect to x_1 will be negative. We note that these assumptions are similar to those considered in [26, 34].

[3] https://www.cnbc.com/select/paying-off-credit-card-debt-boosts-credit-score/

Prior work has shown that two kinds of biases can lead to manipulation cost disparities: *feature biases* and *cost function biases*. For the first part of the analysis, we focus on *feature biases* and we analyze the impact of *cost function biases* later in this section. Feature biases refer to settings where disadvantaged group individuals have scores concentrated in sub-spaces that have a lower likelihood of positive classification; e.g., credit score datasets exhibit these biases for African-Americans [22]. They can be formally defined as follows. For a sub-population condition ψ, there is feature bias against group $Z = 0$ in distribution $\mathcal{D}$ if, for all $x \in \mathcal{X}, z \in \{0, 1\}$ s.t. $\Pr[X<x \mid Z = z, \psi(X, Y)=1] \in (0, 1)$, we have that

$$\mathbb{P}_{\mathcal{D}}[X < x \mid Z = 0, \psi(X, Y) = 1] > \mathbb{P}_{\mathcal{D}}[X < x \mid Z = 1, \psi(X, Y) = 1]. \tag{1}$$

With respect to feature biases, we restate the result of Milli et al. [34] below, generalizing it to cases when cost analysis may be limited to a certain sub-population defined by condition ψ.

Proposition 3.1. *Suppose we have a sub-population condition ψ and a cost function $c(\cdot, \cdot)$. For a classifier f_τ, characterized by a single threshold τ, if there is feature bias against group 0 as defined in (1) then for all $\tau \in \mathcal{X}$, $G(f_\tau, \psi) > 0$.*

Proposition 3.1 states that if there is feature bias against group 0 then using f_τ leads to higher expected strategic cost for group 0 than group 1. For the one-dimensional setting, the above result shows that a single threshold-based classifier can be discriminatory. Classifiers that use group-specific thresholds, on the other hand, can achieve low social burden gaps as we show below. For $\tau_0, \tau_1 \in \mathcal{X}$, let f_{τ_0,τ_1} denote the classifier that uses threshold τ_0, τ_1 for group 0, 1 respectively.

Proposition 3.2. *Suppose we are given a sub-population condition ψ and a cost function $c(x_1, x_2)$. Say there is feature bias against group 0 in distribution $\mathcal{D}$ as defined in (1). Then there exist $\tau_0, \tau_1 \in \mathcal{X}^2$ such that $G(f_{\tau_0,\tau_1}, \psi) < 0$.*

Proposition 3.2 shows that appropriately selected group-specific thresholds can lead to a relatively lower social burden for disadvantaged groups; strategies for efficiently searching for these appropriate thresholds are discussed later in this section. The proofs of both propositions are presented in Appendix A. We next study whether group-specific classifiers that are fair with respect to selection rate fairness metric $H(f, \psi)$ also have low social burden gap $G(f, \psi)$.

As mentioned earlier, one of the goals of fair classification is to provide equal opportunities for all demographic groups. By equalizing selection rates across groups, prior work forces the classifier to select more disadvantaged group individuals who otherwise would not be selected in the unconstrained case due to dataset biases. However, we show below that achieving fairness w.r.t. selection rate-based metrics (like statistical rate) may not lead to reduced strategic manipulation costs for the disadvantaged group. This is because the average strategic cost incurred by a group depends on the distance between the classifier's threshold for the group and the features of negatively classified individuals from this group; the greater this distance, the greater the strategic cost. Even when a classifier f with fair w.r.t. selection rate fairness metric $H(f, \cdot)$, it may not be fair w.r.t. social burden gap $G(f, \cdot)$ since the distance between classifier threshold and features of negatively-classified individuals of a minority group can still be large (Section 5.1 presents simulations on this point). The following theorem quantifies this issue and shows that the social burden gap depends not just on selection rate disparity, but also on cost function and classifier properties.

Theorem 3.3. *Suppose we are given group-specific thresholds τ_0, τ_1 and a sub-population condition ψ, and the cost function $c(x_1, x_2) = d(x_1, x_2)\mathbf{1}(x_2 > x_1)$. For a fixed x_2, suppose that the gradient of d with respect to x_1 at any point in $(0, x_2)$ is in the range $[g_l, g_u]$, for some $g_l \le g_u \le 0$. Let $P_z(\tau) = \mathbb{P}[X \in (0, \tau) \mid Z = z, \psi]$ and $E_{z,\tau} = \mathbb{E}[X \mid X \in [0, \tau], Z = z, \psi]P_z(\tau)$. Then, we can bound the social burden gap of classifier f_{τ_0,τ_1} as follows*

$$G(f_{\tau_0,\tau_1}, \psi) \le g_u\tau_1 H(f_{\tau_0,\tau_1}, \psi) + (g_u\tau_1 - g_l\tau_0)P_0(\tau_0) - g_u E_{1,\tau_1} + g_l E_{0,\tau_0},$$

$$G(f_{\tau_0,\tau_1}, \psi) \ge g_l\tau_1 H(f_{\tau_0,\tau_1}, \psi) + (g_l\tau_1 - g_u\tau_0)P_0(\tau_0) - g_l E_{1,\tau_1} + g_u E_{0,\tau_0}.$$

The proof is presented in Appendix A. Note that Theorem 3.3 does not assume feature bias (Eq (1)) to be explicitly present and can handle generic feature distributions. To interpret the above theorem, consider the impact of different $H(f_{\tau_0,\tau_1}, \psi)$ values. For simplicity, suppose $\psi(x, y)=1$ for all (x, y) (i.e., $H(f_{\tau_0,\tau_1}, \psi)$ is the statistical rate). When $H(f_{\tau_0,\tau_1}, \psi)=0$, the classifier has an equal selection rate for group 0 and group 1. Consider the setting when d is linear, i.e., $d(x_1, x_2) = x_2 - x_1$. In this case, the upper and lower bounds are equal and the social burden gap is $(\tau_0 - \tau_1)P_0(\tau_0) - E_{0,\tau_0} + E_{1,\tau_1}$. Since $H(f_{\tau_0,\tau_1}, \psi) = 0$, we can simplify $G(f_{\tau_0,\tau_1}, \psi)$ to be $(\tau_0 - \mathbb{E}[X \mid X \in [0, \tau_0], Z = 0] - \tau_1 + \mathbb{E}[X \mid X \in [0, \tau_1], Z = 1])P_0(\tau_0)$.

In this equation, $(\tau_z - \mathbb{E}[X \mid X \in [0, \tau_z], Z = z])$ is the average distance of feature values of negative-classified individuals of group z from the decision boundary. The difference between these distances for group 0 and group 1 depends on the choice of τ_0, τ_1 and group distributions. To intuitively understand this dependence, we provide simulations over datasets generated using Gaussian distributions in Section 5.1. The simulations show that as the distribution variance increases, the social burden gap of classifiers, constrained to have $H(f_{\tau_0,\tau_1}, \psi) \approx 0$, can also dramatically increase. This is why simply constraining $H(f_{\tau_0,\tau_1}, \psi)$ is not sufficient to obtain a classifier with a low social burden gap. However, when $H(f_{\tau_0,\tau_1}, \psi)<0$, the difference between the above distances is unlikely to be small since (a) low H implies that τ_0 is higher or similar to τ_1 and (b) due to feature bias the feature values of group 0 are lower than group 1. Hence, $H(f_{\tau_0,\tau_1}, \psi)$ being greater than or equal to 0 is necessary (but not sufficient) to have low social burden gap.

Extension to group-specific cost functions. In many settings, strategic cost disparity can arise due to *cost function biases*, i.e., from different groups having different cost functions. Due to these biases, for the same unit of a feature update, the disadvantaged group would pay a larger cost than the advantaged group; e.g., African Americans face larger access barriers to credit than White Americans [9]. To account for group-specific costs, Theorem 3.3 can be extended to use group-specific gradient bounds for the cost function; incorporating them leads to the following bounds.

Theorem 3.4. *Suppose we are given group-specific thresholds τ_0, τ_1 and a sub-population condition ψ. Let $c_z(x_1, x_2) = d_z(x_1, x_2)\mathbf{1}(x_2 > x_1)$ denote the cost for group z individuals. For a fixed x_2, suppose that the gradient of d_z with respect to x_1 at any point in $(0, x_2)$ is in the range $[g_{l,z}, g_{u,z}]$, for some $g_{l,z} \le g_{u,z} \le 0$ for all $z \in \{0, 1\}$. Let*

$P_z(\tau) := \mathbb{P}[X \in (0,\tau) \mid Z = z, \psi]$ *and* $E_{z,\tau} = \mathbb{E}[X \mid X \in [0,\tau], Z = z, \psi]P_z(\tau)$. *Then, the social burden gap* $G(f_{\tau_0,\tau_1}, \psi)$ *is upper-bounded by*

$$g_{u,1}\tau_1 H(f_{\tau_0,\tau_1}, \psi) + (g_{u,1}\tau_1 - g_{l,0}\tau_0)P_0(\tau_0) - g_{u,1}E_{1,\tau_1} + g_{l,0}E_{0,\tau_0},$$

and lower bounded by

$$g_{l,1}\tau_1 H(f_{\tau_0,\tau_1}, \psi) + (g_{l,1}\tau_1 - g_{u,0}\tau_0)P_0(\tau_0) - g_{l,1}E_{1,\tau_1} + g_{u,0}E_{0,\tau_0}.$$

The proof of the theorem is presented in Appendix A. Cost function biases can be inherently captured using the gradient of the group-specific cost functions. This is because these biases imply that the rate of increase in cost for the disadvantaged group is greater than that of the advantaged group. Hence, if group 0 faces higher manipulation costs to move to the same point than group 1, then the absolute gradient of c_0 will be larger than the absolute gradient of c_1; this can result in higher social burden gaps than the case when cost function is same for both groups. For instance, suppose $d_0(x_1, x_2)$=$a(x_2-x_1)$ and $d_1(x_1, x_2)$=(x_2-x_1) and $H(f, \psi)$=0. If $a > 1$, then the social burden gap can be shown to be $(a(\tau_0 - \mathbb{E}[X \mid X \in [0,\tau_0], Z = 0]) - (\tau_1 - \mathbb{E}[X \mid X \in [0,\tau_1], Z = 1]))P_0(\tau_0)$ which is larger than the corresponding social burden gap when cost function is same for both groups. The above theorem can account for both cost function and feature biases and provides a succinct relationship between standard fairness metrics and social burden gap.

Fair classification with low social burden gap and low selection rate disparity. To construct a classifier that has high institution utility and low social burden gap, we can employ Theorem 3.3, 3.4. Recall that loss function $\mathcal{L}$ measures the expected risk of any classifier. Standard fair classification algorithms already optimize error rate of f w.r.t. $\mathcal{L}$ and subject to selection rate constraints (i.e., $\mathbb{E}_{\mathcal{D}}[\mathcal{L}(f; X, Y)]$) subject to constraints on $H(f, \psi)$.

To construct low social burden gaps, we can alternately use a modified constraint on the upper bound in Theorem 3.3 or Theorem 3.4 to obtain a classifier with low social burden gap. For example, in the setting of just feature bias, we can use the upper bound from Theorem 3.3 as a constraint; i.e., minimize $\mathbb{E}_{\mathcal{D}}[\mathcal{L}(f; X, Y)]$ subject to $(g_u\tau_1 - g_l\tau_0)P_0(\tau_0) - g_uE_{1,\tau_1} + g_lE_{0,\tau_0} \leq 0$. Quantities $P_z(\tau_z)$ and E_{z,τ_z} can be computed empirically for a given dataset: $P_z(\tau_z), E_{z,\tau_z}$ are the fraction and empirical mean, respectively, of group z individuals who are negatively classified. In Section 5, we empirically show that using this modified optimization program results in classifiers that low social burden gaps.

4 EXTENSION TO MULTIPLE DIMENSIONS

Suppose that features are n-dimensional, for n>1. For this multi-dimensional setting, we consider classifiers that threshold over a linear combination of the features of the individuals. We again assume that all features are outcome monotonic, i.e., increasing each feature value results in increase in likelihood of positive classification. Hence, we only consider manipulations from x_1 to x_2 when $x_2 \geq x_1$, i.e., for all $i \in [n]$, $x_2^{(i)} \geq x_1^{(i)}$. Finally, the cost function is assumed to be linear, i.e, for an individual from group $z \in \{0,1\}$, the cost function is $c_z(x_1, x_2) = d_z^\top(x_2 - x_1)$ if $x_2 \geq x_1$ and $c(x_1, x_2) = 0$ if $x_2 \leq x_1$, given $d_0, d_1 \in \mathbb{R}^n$ (we study the quadratic cost function setting in Appendix A). Linear cost functions have been used in prior work to approximately model real-world strategic settings [14, 21, 26]. Vector d_z can encode the different costs paid for updating different features; e.g., in the credit score setting, opening new credit lines has lower costs than increasing annual income. In this multi-dimensional setting, we can prove the following result.

Theorem 4.1. *Suppose we have a linear classifier* f *such that for an individual with* x *and group* z, $f(x) = 1$ *if and only if* $u^\top x \geq v_z$ *and 0 otherwise. For an individual from group* z *with unmanipulated datapoint* x_1, *the cost to move to point* x_2 *is defined as* $c_z(x_1, x_2) = d_z^\top(x_2 - x_1)$ *if* $x_2 \geq x_1$ *and 0 otherwise, for* $d_0, d_1 \in \mathbb{R}^n$. *Let* $w_z^\star := \max_{i\in[n]} u_i/d_{z,i}$. *Then,*

$$G(f, \psi) = -\frac{1}{w_1^\star}(v_1 H(f, \psi)) - \delta,$$

where $\delta = \left(\frac{v_1}{w_1^\star} - \frac{v_0}{w_0^\star}\right)P_0 - \frac{1}{w_1^\star}E_{1,v_1} + \frac{1}{w_0^\star}E_{0,v_0}$, $P_z = \mathbb{P}[f(X) = 0 \mid Z = z, \psi]$, $E_{z,\tau} = \mathbb{E}[(u^\top X) \mid f(X) = 0, Z = z, \psi]P_0$.

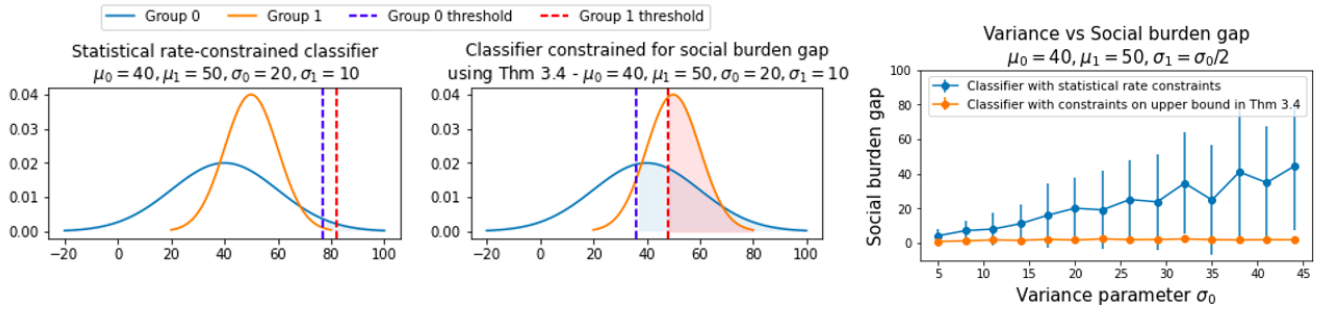
We obtain an equality relation here since the cost function is linear. The proof (presented in Appendix A) follows by reducing this case to the single-dimensional setting. This is possible since in the case of linear classifiers the distance of negatively-classified individuals from the classifier threshold can be captured using $u^\top x$. While limiting the analysis to linear classifiers might seem restrictive, Theorem 4.1 still provides evidence that social burden gap $G(\cdot)$ can be large for classifiers in multi-dimensional settings even when selection rate disparity $H(\cdot)$ is small, due to additional factors captured by δ. Furthermore, empirical analysis of real-world fairness benchmark datasets (Section 5) shows that linear classifiers achieve close to state-of-the-art performance for these datasets. Thus, it is important to study constraints on linear classifiers that can ensure low manipulation costs for all groups.

5 EMPIRICAL ANALYSIS

5.1 Synthetic Simulation

To intuitively explain different components of Theorem 3.3, 3.4, we design a simulation using a synthetic data generation process. Suppose that features of group $z \in \{0,1\}$ are sampled from $N(\mu_z, \sigma_z)$, where $\mu_0, \mu_1, \sigma_0 \in \mathbb{R}_{>0}$ and $\sigma_1 = \sigma_0/2$. Let X_z denote the features of group z. Suppose that for element x_i from group z, class label y_i is 1 with probability $(x_i + \min(X_z))/(\max(X_z) + \min(X_z))$. We sample 500 elements for each group. When $\mu_0 < \mu_1$, there will likely be feature bias in this dataset and any classifier f trained over this dataset will have to use group-specific thresholds to achieve statistical parity. Suppose the cost function for manipulation is linear. Consider two classifiers trained using the above data. The first classifier is trained to achieve maximum accuracy subject to $|H(f, \psi)| \leq 0.4$; here ψ is the identity function. The second classifier is trained to achieve maximum accuracy subject to the constraint that $|G(f, \psi)| \leq 4$. In other words, the first classifier uses constraints on the statistical rate while the second classifier uses constraints on the social burden gap. From Figure 1a,b, we can see that these classifiers can have different group thresholds.

To understand Theorem 3.3, 3.4 using this example, note that the bound in both theorems depend on the difference between group-specific quantities (τ_z–$\mathbb{E}[X|X\in[0,\tau_z], Z$=$z]$): the average distance of feature values of negative-classified individuals from group z



(a) Stat. rate constrained classifier **(b) Thm 3.4 constrained classifier** **(c) Variance vs social burden gap**

Figure 1: Performance of statistical rate constrained classifier and classifier constrained using our method on a synthetic dataset. Plots (a), (b) show the distribution and classifier thresholds for one random iteration. Plot (c) shows the mean and deviation (over 50 repetitions) of social burden gap of classifiers that are statistical rate-constrained and classifiers that are constrained using Thm. 3.4.

from the decision boundary. This difference will increase as σ_0 increases since within-group variances will increase and the group 0 variance grows faster than group 1 variance (since $\sigma_1 = \sigma_0/2$). Hence, as σ_0, increases, the social burden gap of a classifier can increase even when the statistical rate remains small. We empirically observe this phenomenon in Figure 1c, where we see that increasing σ_0 leads to an increase in the social burden gap of the statistical rate-constrained classifier. However, we also observe that classifiers constrained using Theorem 3.4 have almost-zero social burden gap, demonstrating that using Theorem 3.4 for forming fairness constraints leads to classifiers with low social burden gaps for all σ_0 values.

5.2 FICO Credit Dataset

Dataset. We use the FICO credit data [22] for preliminary real-world data analysis of classifiers that are fair with respect to standard fairness metrics and classifiers that are fair with respect to the social burden gap. This dataset contains 116k credit scores corresponding to White individuals and 16k credit scores corresponding to Black/African-American individuals and a binary class label for loan default for each individual (pre-processing details are provided in Appendix B). As shown by prior work [34], this dataset exhibits feature bias against African-American individuals.

Methodology. Around 20k random samples from the dataset are removed to create a test partition. Each classifier is composed of two thresholds (τ_0, τ_1). Threshold τ_0 is for credit scores of African-American individuals and threshold τ_1 is for credit scores of White individuals. A classifier assigns a positive class label to an individual if the individual's credit score is larger than the classifier's threshold for the individual's group. Since the credit scores lie in the range from 1 to 100, we evaluate all possible classifiers, with τ_0, τ_1 in the set $\{1, 2, \ldots, 100\} \times \{1, 2, \ldots, 100\}$, and record their properties. We use the linear cost function $c(x, x') := \mathbf{1}(x > x') \cdot (x - x')$ for this section and provide results for the quadratic separable cost function in Appendix C. We analyze classifier performance for two sub-population conditions: (a) ψ_{sr} which is always 1, i.e., $\psi_{sr}(x, y)$=1, for all x, y, and (b) ψ_{tpr} which is 1 if true class label is 1, i.e., $\psi_{sr}(x, y)$=$\mathbf{1}(y$=1). $H(f, \psi_{sr})$ measures statistical rate and $H(f, \psi_{tpr})$ measures true positive rate disparity.

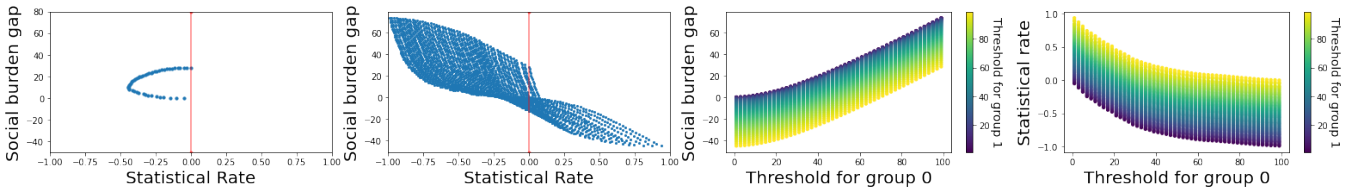
Results. *Statistical rate $H(\cdot, \psi_{sr})$ vs social burden gap $G(\cdot, \psi_{sr})$.* Plot 2a presents the results for classifiers that use the same threshold for both groups. As discussed in Proposition 3.1, these classifiers always have social burden gap ≥0 and lead to higher strategic manipulation cost for African-American individuals. Furthermore, even the statistical rate of these classifiers is low implying that all classifiers using single thresholds select White individuals at a higher rate. Plot 2b presents fairness metrics for classifiers that use group-specific thresholds. Here, we observe that the range of values achieved for statistical rate and low social burden gap is much larger. A classifier with a high statistical rate favoring the disadvantaged group (>0.5) also has a low social burden gap (<−10) for this dataset. However, almost equal group selection rates do not imply parity with respect to social burden. For classifiers with statistical rate close to 0, the social burden gap ranges from $[-12, 30]$.

True positive rate $H(\cdot, \psi_{tpr})$ vs social burden gap $G(\cdot, \psi_{tpr})$. With ψ_{tpr}, we compute costs for individuals who are incorrectly negatively classified. From Plot 3a, we again see that single-threshold classifiers have social burden gap ≥0. Plot 3b, however, shows that there exist group-specific thresholds that result in low social burden gap and high true positive rate for African-American individuals. Plots 2c, d, 3c, d present the relationship between group thresholds and fairness metrics. Increasing group 0 threshold increases the social burden for group 0 and decreases the statistical/true positive rate as positive classifications decrease.

Constructing classifiers with low social burden gap. We next empirically analyze the inequalities in Theorem 3.3. Suppose the goal of the institution is to maximize accuracy subject to the constraint that social burden gap is $\leq g$, for some $g \in \mathbb{R}$. Since group 0 is the marginalized one, the institution aims to achieve a non-positive social burden gap to address the manipulation cost disparities. As shown in 3.3, social burden gap is upper bounded by $g_u\tau_1 H(f, \psi) + (g_u\tau_1 - g_l\tau_0)P_0(\tau_0) - g_u E_{1,\tau_1} + g_l E_{0,\tau_0}$. For the linear cost function, $g_u = g_l = 1$. Hence, we use the burden gap constraint

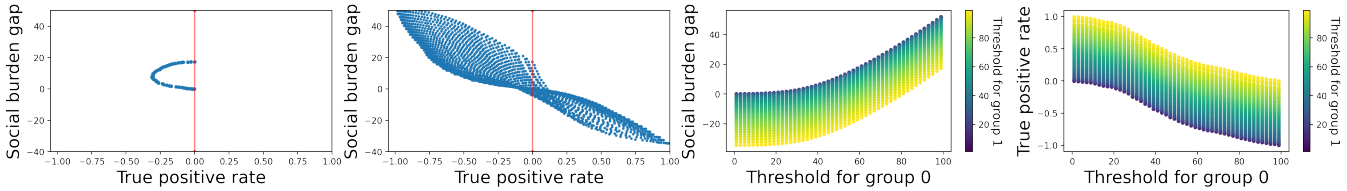
$$\tau_1 H(f, \psi) - (\tau_1 - \tau_0)P_0(\tau_0) + E_{1,\tau_1} - E_{0,\tau_0} \leq g. \quad (2)$$

For g=0, Figure 4 plots the accuracy and social burden gap of all classifiers that satisfy the above conditions for ψ_{sr} and ψ_{tpr}. The



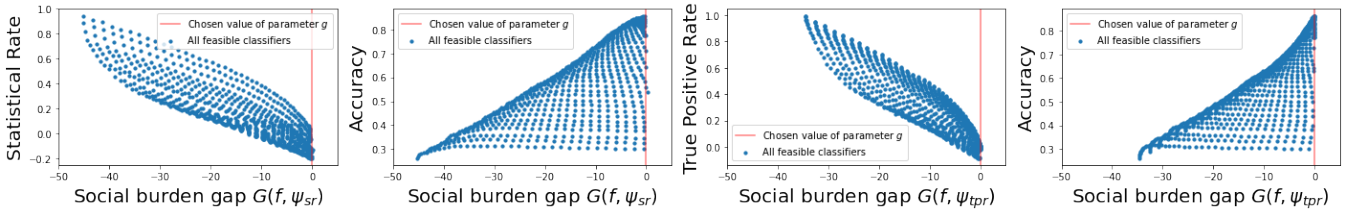
(a) $\tau_0{=}\tau_1{\in}\{1,\ldots,100\}$ (b) $(\tau_0,\tau_1){\in}\{1,\ldots,100\}^2$ (c) Social burden gap vs τ_0 (d) Statistical rate vs τ_0

Figure 2: Statistical rate and social burden gap of all classifiers for FICO dataset. Each point represents a classifier and the axes plot different properties of these classifiers. Plot (a) presents social burden gap $G(f,\psi_{sr})$ vs statistical rate $H(f,\psi_{sr})$ for classifiers that use the same threshold for both groups. Plots (b), (c), (d) present social burden gap $G(f,\psi_{sr})$ vs statistical rate $H(f,\psi_{sr})$ for classifiers that can use group-specific threshold. The range of statistical rate and social burden gap values achieved for these classifiers is larger.



(a) $\tau_0{=}\tau_1{\in}\{1,\ldots,100\}$ (b) $(\tau_0,\tau_1){\in}\{1,\ldots,100\}^2$ (c) Social burden gap vs τ_0 (d) True positive rate vs τ_0

Figure 3: True positive rate and social burden gap of all classifiers for the FICO dataset. Plot (a) presents social burden gap $G(f,\psi_{tpr})$ vs true positive rate $H(f,\psi_{tpr})$ for single-threshold classifiers. Once again, the true positive rate and social burden gap for these classifiers favor the majority group. Plot (b), (c), (d) present $G(f,\psi_{tpr})$ vs $H(f,\psi_{tpr})$ for classifiers that use group-specific thresholds.



(a) Stat. rate vs $G(f,\psi_{sr})$ (b) Accuracy vs $G(f,\psi_{sr})$ (c) TPR vs $G(f,\psi_{tpr})$ (d) Accuracy vs $G(f,\psi_{tpr})$

Figure 4: Performance of classifiers that satisfy the modified fairness constraints. Plots (a),(b) present the statistical rate and accuracy vs social burden gap $G(f,\psi_{sr})$ for all classifiers for which condition (2) is satisfied with $\psi=\psi_{sr}$. Plots (c),(d) present the true positive rate (TPR) and accuracy vs social burden gap $G(f,\psi_{tpr})$ for all classifiers for which condition (2) is satisfied with $\psi=\psi_{tpr}$.

plots show that all classifiers that satisfy the constraint on the upper bound from Theorem 3.3 satisfy the condition $G(f_{\tau_0,\tau_1},\psi_{sr}){<}0$. Furthermore, even in this case, the classifier that optimizes this constrained problem has high accuracy. For ψ_{sr} the accuracy of the optimal constrained classifier is 0.86 and $G(f_{\tau_0,\tau_1},\psi_{sr}){=}{-}0.21$ and for ψ_{tpr} the accuracy of the optimal constrained classifier is 0.86 and $G(f_{\tau_0,\tau_1},\psi_{tpr}){=}{-}0.03$. In comparison, the accuracy of the optimal unconstrained classifier is 0.88, showing minimal loss in accuracy due to the constraints.

5.3 Adult Income Dataset

Dataset. For analysis of multi-dimensional data, we use the Adult Income dataset. We use the new version of this dataset developed and preprocessed by Ding et al. [11]. It contains information on around 251k individuals from the state of California surveyed in 2019. The classification task is to predict whether the income of an individual is above $50k or not. The strategic features available are "class of worker", "occupation", and "hours worked per week" (the other five features are listed in Appendix B). We use race as the protected attribute, limiting the dataset to White (93% of the dataset; $z=1$) and Black/African-American (7% of the dataset; $z=0$) individuals.

Methodology. The cost function used is linear and group-specific. Let $d'{\in}\mathbb{R}^9$ be the underlying cost vector such that $d'_i{=}100$ if i represents "class of worker" feature, $d'_i{=}10$ if i represents "occupation", $d'_i{=}1$ if i represents "hours per week", $d'_i{=}\infty$ for other non-strategic features. d' assigns a higher cost factor depending on the difficulty of updating a feature value. The cost function for group 0 is $c_0(x,x'){:=}2{\cdot}d'^{\top}(x'{-}x)$ and cost function for group 1 is $c_1(x,x'):=d'^{\top}(x'{-}x)$. In this case, African-American individuals pay twice the cost that White individuals pay for the same feature update. The dataset is partitioned into 80-20 random train-test splits. Once again, suppose $\psi_{sr}(x,y){=}1$ for all (x,y).

Table 1: Performance of the unconstrained classifier f_{uncons}, classifier constrained to achieve statistical rate ≥ 0 f_{sr}, fair classifier from Rezaei et al. f_{RFMZ}, and classifier constrained to achieve low social burden gap using our method f_{strat} on the Adult dataset.

	Classifier	Accuracy	Stat. Rate	Burden gap
Baselines	f_{uncons}	**0.84** ± 0.00	-0.11 ± 0.01	2.23 ± 0.04
	f_{sr}	0.83 ± 0.01	0.01 ± 0.03	1.12 ± 0.15
	f_{RFMZ}	0.83 ± 0.00	-0.08 ± 0.01	1.88 ± 0.08
Our method	f_{strat}	0.82 ± 0.00	**0.24** ± 0.01	**0.04** ± 0.01

We will restrict the classifiers to be from the linear family and use the logistic log-loss function $\mathcal{L}(f;x,y) := -y\log\sigma(f(x)) - (1-y)\log(1-\sigma(f(x)))$ to measure prediction error of f (here $\sigma(\cdot)$ is the standard sigmoid function). We analyze the following different classifiers. Classifier $f_{uncons} := \arg\min_f \mathbb{E}[L(f;X,Y)]$ will denote the unconstrained classifier. For pre-specified desired statistical rate $\epsilon \in [-1,1]$, classifier $f_{sr} := \arg\min_f \mathbb{E}[L(f;X,Y)]$ subject to $H(f,\psi_{sr}) \geq \epsilon$. Finally, for a pre-specified $g \in \mathbb{R}$, we construct classifiers with social burden gap $G(f,\psi_{sr}) \leq g$. To do so, we use the result from Section 4 and construct classifier $f_{strat} := \arg\min_f \mathbb{E}[L(f;X,Y)]$ subject to $-\frac{1}{w_1^\star}(v_1 H(f,\psi)) - \delta \leq g$ (quantities $w_1^\star, v_1, \delta$ are defined in Theorem 4.1). We will set ϵ=0 and g=0 to analyze classifiers with equal selection rate and zero social burden gap in this section, and present variation of performance with these parameters in Appendix D. To compare with another fair classification baseline, we also implement the fair logistic regression algorithm of Rezaei et al. [38] with statistical rate constraints; we will call this classifier f_{RFMZ}. We report the mean and standard error of accuracy, statistical rate, and social burden gap of all classifiers over 100 random train-test splits. Implementation details of all methods are provided in Appendix B.

Results. Table 1 presents the performance of classifiers f_{uncons}, f_{sr}, f_{RFMZ}, f_{strat}. First note that the unconstrained classifier f_{uncons} has an average statistical rate of -0.11 (i.e., the selection rate of African-American individuals is much lower than the selection rate of White individuals) and the average social burden gap is 2.23 (i.e, cost of strategic manipulation is higher for African-Americans). Hence, fairness interventions are necessary in this case to achieve equal performance. For classifier f_{sr}, the statistical rate is close to 0; however, the social burden gap is still greater than 0 in this case, showing that an almost equal selection rate does not necessarily imply a low social burden gap. Similarly, baseline f_{RFMZ} has a high social burden gap despite having a better statistical rate than the unconstrained classifier. In comparison, our classifier f_{strat} has a statistical rate of 0.24, i.e., the selection rate for African-Americans is higher. Furthermore, classifier f_{strat} achieves a social burden gap close to 0 on average. Hence, both groups pay almost equal cost for strategic manipulation. In terms of accuracy, f_{uncons} achieves the highest accuracy (0.84) and the accuracy of f_{strat} is only slightly lower (0.82), implying a minimal loss in accuracy due to fairness constraints.

6 DISCUSSION AND LIMITATIONS

Our paper untangles the relationship between manipulation cost disparities and standard fairness metrics and provides a framework to construct classifiers with low costs for minority groups. The proposed framework can be useful for real-world classification settings where individuals repeatedly interact with an institution and its classifier (e.g., applying for loans after rejection with updated features). In these settings, the institution can employ our framework to ensure that minority individuals do not pay disparately higher costs to exercise their available recourse options. In this section, we discuss philosophical grounding, practical advantages, and certain limitations of our framework.

Philosophical grounding. Venkatasubramanian and Alfano [41] argued how recourse options can be systematically helpful to groups that have been historically oppressed. In particular, they distinguish between "*token acts of exercising recourse (reversing a single harmful decision) and the general state of enjoying systematic access to the power to reverse harmful decisions (knowing that if a harmful decision were to be made, one would be able to get it reversed)*." Recognizing this distinction at an institutional level motivates the construction of classifiers that equalize manipulation costs across groups to ensure systematic access to recourse for all. This way, the institution can acknowledge biases in data and costs and provide redressal mechanisms that account for structural inequalities.

Practical advantages and information required by the institution to address manipulation disparities. Beyond the presented analysis, our framework has advantages that allow for easy use in applications. For instance, the bounds on the social burden gap require minimal information about cost functions and can handle settings where cost functions vary across individuals but belong to the same class of gradient-bounded or Lipschitz functions. Hence, an institution aiming to construct classifiers with a low social burden gap would not need to model the complete update behaviors of every individual. Nevertheless, some information about the cost function gradient is required for implementing our framework in practice. While this can be quantified by observing the past and current behavior of individuals who have been negatively classified, errors in cost function gradient measurement can affect the performance of our proposed fairness intervention. Future work can additionally explore methods to reduce strategic manipulation disparities while ensuring that the method is robust to gradient measurement errors.

A social burden gap of 0 implies equal manipulation costs for all groups; however, in some cases, a gap of less than 0 may be necessary. In multi-feedback settings, the predictions at one time step affect the classifier training at the next time step [37] and biases can be amplified across feedback iterations. In these cases, our framework can also be used to construct a classifier with a negative social burden gap to tackle these biases.

Negative manipulations. Strategic manipulations can potentially be used to "game" a classifier [21] and, by reducing manipulation costs, our methods can potentially lead to increased "gamification". The issue of gamification primarily arises due to noisy features that are only superficially predictive of the class label. In the absence of other robust features, a classifier will use these noisy features for prediction. However, the presence of noisy features does not warrant disparity in manipulation costs, especially if these features favor the majority group. Nevertheless, recent papers have also suggested classifier designs that incentivize *positive manipulations*

to improve individuals' task-related qualities [30]; using our framework with these classifiers can ensure that all groups have equal opportunities for improvement.

Model limitations. For multi-dimensional settings, our framework considers linear and quadratic cost functions. Extensions of our framework for generic cost functions in multi-dimensional settings can be further studied as part of future work. Secondly, while we consider binary protected attributes in our analysis, our results can be extended to non-binary attributes. This is because the non-binary setting can be reduced to the binary setting by considering the pairwise comparison of measures for different protected attribute values. However, due to multiple comparisons, the bounds for $G(\cdot, \cdot)$ will be weaker, and future work can explore ways to improve these bounds for non-binary attributes.

Additional limitations of our framework are related to the accuracy of information about the individuals available to the institution. As mentioned earlier, if the institution does not have accurate information about the costs associated with feature updates, then our proposed framework might not be completely effective in addressing manipulation disparities. Another information-based limitation is the assumption that the classifier used by the institution is known. This may not be true in real-world settings and only partial information about decision rules may be publicly available. Recent work by Ghalme et al. [16] aims to address this problem in general strategic classification settings and our framework can potentially be extended in the future along similar lines.

7 CONCLUSION

We study the impact of fair classifiers on individuals' ability to positively manipulate their features based on their group membership. In settings where feature distributions or cost functions are biased against minority groups, we observe that classifiers can have (almost) equal selection rates for all groups but can still have relatively higher strategic manipulation costs for individuals from minority groups. We propose modified fairness constraints to construct classifiers that reduce this disparity and show its efficacy over the FICO Credit and Adult Income datasets. Our work demonstrates the necessity of analyzing the impact of fair classifiers in dynamic settings and developing approaches that provide recourse opportunities that are independent of their group memberships.

REFERENCES

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

[2] Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. 2020. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1774–1781.

[3] Umberto Bacchi. 2020. TikTok apologises for censoring LGBT+ content. https://www.reuters.com/article/britain-tech-lgbt-idUSL5N2GJ459

[4] M Braverman and S Garg. 2020. The Role of Randomness and Noise in Strategic Classification. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, Vol. 156.

[5] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–20.

[6] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*. 319–328.

[7] Yiling Chen, Yang Liu, and Chara Podimata. 2020. Learning Strategy-Aware Linear Classifiers. *Advances in Neural Information Processing Systems* 33 (2020), 15265–15276.

[8] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

[9] Ethan Cohen-Cole. 2011. Credit card redlining. *Review of Economics and Statistics* 93, 2 (2011), 700–713.

[10] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).

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

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

[14] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 2021. Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents. *arXiv preprint arXiv:2112.02746* (2021).

[15] Board of Governors FRS. 2007. *Report to the congress on credit scoring and its effects on the availability and affordability of credit.* Board of Governors of the Federal Reserve System.

[16] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. 2021. Strategic classification in the dark. In *International Conference on Machine Learning*. PMLR, 3672–3681.

[17] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

[18] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. *Commun. ACM* 61, 7 (2018), 56–66.

[19] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166* (2019).

[20] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[21] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.

[22] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

[23] Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. 2023. Recommending to Strategic Users. *arXiv preprint arXiv:2302.06559* (2023).

[24] Julian Van Horne. 2020. Shadowbanning is a Thing — and It's Hurting Trans and Disabled Advocates. https://saltyworld.net/shadowbanning-is-a-thing-and-its-hurting-trans-and-disabled-advocates/

[25] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.

[26] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.

[27] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. 2021. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*. PMLR, 4687–4697.

[28] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[29] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 43–53.

[30] Jon Kleinberg and Manish Raghavan. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.

[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[32] Sagi Levanon and Nir Rosenfeld. 2021. Strategic classification made practical. In *International Conference on Machine Learning*. PMLR, 6243–6253.

[33] John Miller, Juan C Perdomo, and Tijana Zrnic. 2021. Outside the Echo Chamber: Optimizing the Performative Risk. *arXiv preprint arXiv:2102.08570* (2021).

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

[35] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[36] Marieke Möhlmann and Lior Zalmanson. 2017. Hands on the wheel: Navigating algorithmic management and Uber drivers'. In *Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea*. 10–13.
[37] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.
[38] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5511–5518.
[39] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.
[40] Emily van der Nagel. 2018. 'Networks that work too well': intervening in algorithmic connections. *Media International Australia* 168, 1 (2018), 81–92.
[41] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.
[42] Jacob L Vigdor and Charles T Clotfelter. 2003. Retaking the SAT. *Journal of Human resources* 38, 1 (2003), 1–33.
[43] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2022. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9584–9594.
[44] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.
[45] Raeshanda Wilson. 2020. Predicting graduate school success: A critical race analysis of the Graduate Record Examination. (2020).
[46] Lori Teresa Yearwood. 2019. Many minorities avoid seeking credit due to generations of discrimination. Why that keeps them back. https://www.cnbc.com/2019/09/01/many-minorities-avoid-seeking-credit-due-to-decades-of-discrimination.html
[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.
[48] Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95.
[49] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[50] Wenbin Zhang and Eirini Ntoutsi. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 1480–1486.