# Fairness Interventions as (Dis)Incentives for Strategic Manipulation

**Xueru Zhang** [1] **Mohammad Mahdi Khalili** [2] **Kun Jin** [3] **Parinaz Naghizadeh** [4] **Mingyan Liu** [3]

## Abstract

Although machine learning (ML) algorithms are widely used to make decisions about individuals in various domains, concerns have arisen that (1) these algorithms are vulnerable to strategic manipulation and "gaming the algorithm"; and (2) ML decisions may exhibit bias against certain social groups. Existing works have largely examined these as two separate issues, e.g., by focusing on building ML algorithms robust to strategic manipulation, or on training a fair ML algorithm. In this study, we set out to understand the impact they each have on the other, and examine how to characterize fair policies in the presence of strategic behavior. The strategic interaction between a decision maker and individuals (as decision takers) is modeled as a two-stage (Stackelberg) game; when designing an algorithm, the former anticipates the latter may manipulate their features in order to receive more favorable decisions. We analytically characterize the equilibrium strategies of both, and examine how the algorithms and their resulting fairness properties are affected when the decision maker is strategic (anticipates manipulation), as well as the impact of fairness interventions on equilibrium strategies. In particular, we identify conditions under which anticipation of strategic behavior may mitigate/exacerbate unfairness, and conditions under which fairness interventions can serve as (dis)incentives for strategic manipulation.

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. [2]Yahoo! Inc., the author is also with The Ohio State University. [3]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [4]Integrated Systems Engineering & Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA.. Correspondence to: Xueru Zhang <zhang.12807@osu.edu>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

## 1. Introduction

As machine learning (ML) algorithms are increasingly being used to make high-stake decisions in domains such as hiring, lending, criminal justice, and college admissions, the need for transparency increases in terms of how decisions are reached given input. However, given (partial) information about an algorithm, individuals subject to its decisions can adapt their behavior by strategically manipulating their data in order to obtain favorable decisions. This strategic behavior in turn hurts the performance of ML models and diminishes their utility. Such a phenomenon has been widely observed in real-world applications, and is known as *Goodhart's law*, which states "once a measure becomes a target, it ceases to be a good measure" (Strathern, 1997). For instance, a hiring or admissions practice that heavily depends on GPA might motivate students to cheat on exams; not accounting for such manipulation may result in disproportionate hiring of under-qualified individuals. A strategic decision maker is one who anticipates such behavior and thus aims to make its models robust to strategic manipulation.

A second challenge facing ML algorithms is the growing concern over bias in their decisions, and various notions of fairness (e.g., demographic parity (Barocas et al., 2019), equal opportunity (Hardt et al., 2016b)) have been proposed to measure and remedy biases. These measures typically impose an (approximate) equality constraint over certain statistical measures (e.g., positive rate, true positive rate, etc.) across different groups when building ML algorithms.

In this paper, we study (fair) machine learning in the presence of strategic manipulation. Specifically, we consider a decision maker whose goal is to select individuals that are *qualified* for certain tasks based on a given set of features. Given knowledge of the selection policy, individuals can tailor their behavior and manipulate their features in order to receive favorable decisions. We shall assume that this feature manipulation does not affect an individual's true qualification state. We say the decision maker (and its policy) is *strategic* if it anticipates such manipulation; it is *non-strategic* if it does not take into account individuals' manipulation in its policies.

We adopt a two-stage (Stackelberg) game setting where the decision maker commits to its policies, following which individuals best-respond. A crucial difference between this

study and existing models of strategic interaction is that existing models typically assume features and their manipulation are *deterministic* so that the manipulation cost can be modeled as *a function of the change in features* (Hardt et al., 2016a; Dong et al., 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020; Brückner & Scheffer, 2011; Haghtalab et al., 2020; Kleinberg & Raghavan, 2019; Chen et al., 2020; Miller et al., 2020); by contrast, in our setting features are *random variables* whose realizations are *unknown* prior to an individual's manipulation decision. In fact, this is the case in many important applications, motivating examples are presented in Sec. 3.

Moreover, among these existing works, only (Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020) studied the disparate impact of ML decisions on different social groups, where the disparity stems from different manipulation costs and different feature distributions. No fairness intervention was considered in these works. In contrast, we study the impact of fairness intervention on different groups in the presence of strategic manipulative behavior, and explore the role of fairness intervention in (dis)incentivizing such manipulation. We aim to answer the following questions: how does the anticipation of individuals' strategic behavior impact a decision maker's utility, and the resulting policies' fairness properties? How is the Stackelberg equilibrium affected when fairness constraints are imposed? Can fairness intervention serve as (dis)incentives for individuals' manipulation?

Our main contributions and findings are as follows.

1. We formulate a Stackelberg game to model the interaction between a decision maker and strategic individuals (Sec. 3). We characterize both strategic (fair) and non-strategic (fair) optimal policies of the decision maker, and individuals' best response (Sec. 4, Lemmas 4.2-4.7).
2. We study the impact of the decision maker's anticipation of individuals' strategic manipulation by comparing non-strategic with strategic policies (Sec. 5):
   - We show that compared to non-strategic policy, strategic policy always disincentivizes manipulation; it over (resp. under) selects when a population is majority-qualified (resp. majority-unqualified)[1](Thm. 5.3).
   - We show that the anticipation of manipulation can *worsen* the fairness of a strategic policy: when one group is majority-qualified while the other is majority-unqualified (Thm. 5.4); on the other hand, when both groups are majority-unqualified, we show the possibility of using a strategic policy to *mitigate unfairness* and even flip the disadvantaged group (Thm. 5.5).
3. We study the impact of fairness interventions on policies and individuals' manipulation (Sec. 6).
   - If a decision maker lacks information to anticipate manipulative behavior (but which in fact exists), we identify conditions under which such non-strategic decision maker benefits from using fairness constrained policies rather than unconstrained policies (Thm. 6.1).
   - By comparing individuals' responses to strategic policy with and without fairness intervention, we identify scenarios under which a strategic fair policy can (dis)incentivize manipulation compared to an unconstrained strategic policy (Theorems 6.2-6.4).
4. We examine our theoretical findings using both synthetic and real-world datasets (Sec. 8).

[1]A group is majority-(un)qualified if a decision maker would receive positive(negative) utility upon accepting all individuals in that group.

## 2. Related Work

Our work closely connects to the literature on classification problems in the presence of strategic manipulation. Hardt et al. (2016a) formulated such problem as a Stackelberg competition between the decision maker and individuals, where the decision maker publishes the classifier first, and individuals after observing the classifier can manipulate their features at costs to maximize their utilities. Different from the Stackelberg formulation in our work, manipulation cost in (Hardt et al., 2016a) is modeled as a deterministic function of change in features before and after manipulation. The decision maker aims to find an optimal classifier such that the classification accuracy is maximized when individuals best respond, and the learning algorithms are developed in (Hardt et al., 2016a). Dong et al. (2018) extended this strategic classification to an online setting, where data arrives sequentially and only the manipulated data is revealed. An online convex classification learning algorithm was designed such that the averaged regret diminishes in the long run. (Milli et al., 2019; Hu et al., 2019) extended (Hardt et al., 2016a) by assuming individuals from the different social groups have different costs in manipulation, and the disparate impacts on different groups were studied. Braverman & Garg (2020) explored the role of randomness in strategic classification and focused on randomized classifiers. They showed that randomness can improve classification accuracy and mitigate the disparate effects incurred by manipulation costs across different groups in strategic settings. Sundaram et al. (2021) generalized (Hardt et al., 2016a) by allowing individuals' heterogeneous preferences over classification outcomes and studied PAC-learnability of strategic classification.

Note that the manipulation does not affect an individual's underlying label in the works mentioned above, i.e., strategic manipulation is viewed as gaming. In contrast, another line of research (Haghtalab et al., 2020; Kleinberg & Raghavan, 2019; Coate & Loury, 1993; Alon et al., 2020) considers a setting where the individual's label (qualification) changes

in accordance with the strategic behavior. Specifically, the goal of the decision maker is to design a mechanism such that individuals are incentivized to behave toward directions that improve the underlying qualifications (Haghtalab et al., 2020; Kleinberg & Raghavan, 2019; Alon et al., 2020). (Chen et al., 2020; Miller et al., 2020; Shavit et al., 2020; Bechavod et al., 2021) consider both types of strategic behavior: gaming without changing labels and improvement. Specifically, Chen et al. (2020) developed classifiers that disincentivize manipulation while incentivizing improvement. Miller et al. (2020) proposed a causal framework for distinguishing between gaming and improvement. Shavit et al. (2020) considered a strategic regression problem and suggested improving individuals' underlying states using causal interventions through decision rules. Bechavod et al. (2021) studied strategic interactions in an online regression setting and showed that the decision-maker can identify meaningful variables (i.e., features whose values affact the true label) from a sequence of strategic interactions.

In Stackelberg game formulations, the decision maker always moves first and individuals respond after decision maker's action has been disclosed. Instead, (Coate & Loury, 1993; Liu et al., 2020; Brückner et al., 2012) consider scenarios where both individuals and the decision maker act simultaneously. They formulate the strategic interaction between individuals and decision maker as a game and study the Nash equilibria of the game. In particular, Coate & Loury (1993) considered a setting where individuals are from two social groups which are identical in nature but one group suffers from the negative stereotype. They showed that such stereotype results in different equilibria of two groups. The impact of *demographic parity* fairness is also examined in (Coate & Loury, 1993). Liu et al. (2020) studied a similar game, but assumed two groups can be different in feature distributions and manipulation costs.

## 3. Problem Formulation

Consider two groups $\mathcal{G}_a$, $\mathcal{G}_b$ distinguished by a sensitive attribute $S \in \{a, b\}$ (e.g., gender), with fractions $n_s = \Pr(S = s)$ of the population. An individual from either group has observable features $X \in \mathbb{R}^d$ and a hidden qualification state $Y \in \{0, 1\}$. Let $\alpha_s = P_{Y|S}(1|s)$ be the qualification rate of $\mathcal{G}_s$. A decision maker makes a decision $D \in \{0, 1\}$ ( "0" being negative/reject and "1" positive/accept) for an individual using a group-dependent policy $\pi_s(x) = P_{D|XS}(1|x, s)$[2]. An individual's action is denoted by $M \in \{0, 1\}$, with $M = 1$ indicating manipulation and $M = 0$ otherwise. Note that in our context manipulation does not change the true qualification state $Y$. It is the qualification state $Y$, sensitive attribute $S$, and manipulation action $M$ together that drive the realizations of features $X$.

**Best response.** An individual in $\mathcal{G}_s$ incurs a random cost $C_s \geq 0$ when manipulating its features, with probability density function (PDF) $f_s(c)$ and cumulative density function (CDF) $\mathbb{F}_{C_s}(c) = \int_0^c f_s(z)dz$. The realization of this random cost is known to an individual when determining its action $M$; while the decision maker only knows the overall cost distribution of each group. Thus the response that the decision maker anticipates (from the group as a whole or from a randomly selected individual) is expressed as follows, whereby given policy $\pi_s$, an individual in $\mathcal{G}_s$ will manipulate its features if doing so increases its utility:

$$\overline{w}P_{D|YMS}(1|y, 1, s) - C_s \geq \overline{w}P_{D|YMS}(1|y, 0, s).$$

Here $\overline{w} > 0$ is a fixed benefit to the individual associated with a positive decision $D = 1$ (the benefit is 0 otherwise); without loss of generality we let $\overline{w} = 1$. In other words, the best response the decision maker expects from individuals of $\mathcal{G}_s$ with qualification $y$ is their probability of manipulation, denoted by $p_s^y := P_{M|YS}(1|y, s)$ and written as:

$$p_s^y(\pi_s) = \Pr\big(C_s \leq P_{D|YMS}(1|y, 1, s) - P_{D|YMS}(1|y, 0, s)\big).$$

We assume that individuals manipulate by imitating the features of those qualified, e.g., students cheat on exams by hiring a qualified person to take exams (or copying answers of those qualified), job applicants manipulate resumes by mimicking those of the skilled employees, loan applicants fool the lender by using/stealing identities of qualified people, etc. This is inspired by the *imitative learning* behavior observed in social learning, whereby new behaviors are acquired by copying social models' actions (Ganos et al., 2012; Gergely & Csibra, 2006). Under this assumption, the qualified individuals will not have incentives to manipulate (as manipulation brings no additional benefit but only cost) and only those unqualified may choose to manipulate, i.e., $p_s^1(\pi_s) = 0$. To simplify the notation, we will use $P_{X|YS}(x|y, s)$ to denote the distributions *before* manipulation. The feature distribution of those unqualified after manipulation becomes $(1 - p_s^0(\pi_s))P_{X|YS}(x|0, s) + p_s^0(\pi_s)P_{X|YS}(x|1, s)$.

*Motivating Example:* The above formulation is fundamentally different from existing literature: 1) the manipulated outcomes are not deterministic where individuals only have probabilistic knowledge of how features may change upon manipulation; 2) the manipulation cost is fixed and known, as opposed to being a function of the actual change in the feature before and after manipulation. A prime example is cheating on an exam by paying for someone else to take it, where the exam score is treated as feature (in admissions or employment decisions): (i) here individual's own

[2]We use group-dependent policies to ensure that *perfect* fairness can be attained under fairness intervention; this allows us to study the impact of fairness constraints more precisely.

feature (unrealized score) and the manipulated feature outcome (actual score received by an imposter) are random, but individuals have a good idea from past experience what those score distributions would be like; (ii) the cost of hiring someone is more or less fixed, determined by the (expected) outcome (the fake score) rather than the difference in score improvement. As the true test score was never realized (those who hire someone do not take the exam themselves), it can be hard to compute precisely how much the feature has improved and put a price on it even after the fact.

In addition to the above, there are many other real-world scenarios where manipulation outcome is not deterministic: athletes may choose to dope but how much performance (feature) improvement they get is not guaranteed; purchasing a stolen credit card number (or SSN) from hackers where the improved feature (e.g., purchase/cash limit) is random as the card is drawn from many stolen cards. Existing models do not capture such inherent randomness.

**Optimal (fair) policy.** The decision maker receives a true-positive (resp. false-positive) benefit (resp. penalty) $u_+$ (resp. $u_-$) when accepting a qualified (resp. unqualified) individual. Its utility, denoted by $R(D, Y)$, is $R(1,1) = u_+$, $R(1,0) = u_-$, $R(0,0) = R(0,1) = 0$. The decision maker aims to find optimal policies for the two groups such that its expected total utility $\mathbb{E}[R(D, Y)]$ is maximized.

There are two types of decision makers, strategic and non-strategic: A *strategic decision maker* anticipates strategic manipulation, has perfect information on the manipulation cost distribution and accounts for this in determining policies, while a *non-strategic decision maker* ignores manipulative behavior in determining its policies. Either type may further impose a fairness constraint $\mathcal{C}$, to ensure that $\pi_a$ and $\pi_b$ satisfy the following:

$$\mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\pi_a(X)] = \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\pi_b(X)]\,, \tag{1}$$

where $\mathcal{P}_s^{\mathcal{C}}$ is some probability distribution over $X$ associated with fairness constraint $\mathcal{C}$. Many fairness notions can be written in this form (Zhang et al., 2019; 2020a; Zhang & Liu, 2021; Khalili et al., 2021a), e.g., equal opportunity (`EqOpt`) (Hardt et al., 2016b) where $\mathcal{P}_s^{\texttt{EqOpt}}(x) = P_{X|YS}(x|1,s)$, or demographic parity (`DP`) (Barocas et al., 2019) where $\mathcal{P}_s^{\texttt{DP}}(x) = P_{X|S}(x|s)$.

The above leads to four types of optimal policies a decision maker can use, which we consider in this paper: 1) non-strategic policy; 2) non-strategic fair policy; 3) strategic policy; 4) strategic fair policy. These are detailed in Sec. 4.

**The Stackelberg game.** The interaction between the decision maker and individuals consists of the following two stages in sequence: (i) The former publishes its policies $(\pi_a, \pi_b)$, which may be strategic or non-strategic, and may or may not satisfy a fairness constraint, and (ii) the latter, while observing the published policies and their realized costs, decide whether to manipulate their features.

## 4. Four types of (non-)strategic (fair) policies

**Non-strategic policy.** A decision maker who does not account for individuals' strategic manipulation maximizes the expected utility $\widehat{U}_s(\pi_s)$ over $\mathcal{G}_s$ defined as follows:

$$\int_X \Big[u_+\alpha_s P_{X|YS}(x|1,s) - u_-(1-\alpha_s)P_{X|YS}(x|0,s)\Big]\pi_s(x)dx.$$

Define $\mathcal{G}_s$'s *qualification profile* as $\gamma_s(x) = P_{Y|XS}(1|x,s)$. Then, we can show that the non-strategic policy $\widehat{\pi}_s^{\texttt{UN}} = \text{argmax}_{\pi_s}\widehat{U}_s(\pi_s)$ is in the form of a threshold policy, i.e., $\widehat{\pi}_s^{\texttt{UN}}(x) = \mathbf{1}\big(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-}\big)$ (Appendix F.1). Throughout the paper, we will present results in the one dimensional feature space. Generalization to high dimensional spaces is discussed in Appendix A.

**Assumption 4.1.** $P_{X|YS}(x|1,s)$, $P_{X|YS}(x|0,s)$ are continuous and satisfy the strict monotone likelihood ratio property, i.e., $\frac{P_{X|YS}(x|1,s)}{P_{X|YS}(x|0,s)}$ is increasing in $x \in \mathbb{R}$.

Assumption 4.1 is relatively mild and can be satisfied by distributions such as exponential and Gaussian, and has been widely used (Zhang et al., 2020b; Jung et al., 2020; Barman & Rathi, 2020; Khalili et al., 2021b; Coate & Loury, 1993). It implies that an individual is more likely to be qualified as their feature value increases. Under Assumption 4.1, the threshold policy can be written as $\pi_s(x) = \mathbf{1}(x \geq \theta_s)$ for some $\theta_s \in \mathbb{R}$. Throughout the paper, we assume Assumption 4.1 holds and focus on threshold policies. We will frequently use $\theta_s$ to denote policy $\pi_s$. Then, the thresholds for non-strategic policies are characterized as follows.

**Lemma 4.2.** *Let $(\widehat{\theta}_a^{\texttt{UN}}, \widehat{\theta}_b^{\texttt{UN}})$ be the non-strategic optimal thresholds. Then $\frac{P_{X|YS}(\widehat{\theta}_s^{\texttt{UN}}|1,s)}{P_{X|YS}(\widehat{\theta}_s^{\texttt{UN}}|0,s)} = \frac{u_-(1-\alpha_s)}{u_+\alpha_s}$.*

**Non-strategic fair policy.** Denoted as $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}})$, this is found by maximizing the total utility subject to fairness constraint $\mathcal{C}$, i.e., $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}}) = \text{argmax}_{(\pi_a,\pi_b)} n_a\widehat{U}_a(\pi_a) + n_b\widehat{U}_b(\pi_b)$ such that Eqn (1) holds. It can be shown that for `EqOpt` and `DP` fairness, the optimal fair policies are also threshold policies and can be characterized by the following.

**Lemma 4.3** ((Zhang et al., 2020b)). *Let $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ be thresholds in non-strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s \left(\frac{u_+\alpha_s P_{X|YS}(\widehat{\theta}_s^{\mathcal{C}}|1,s) - u_-(1-\alpha_s)P_{X|YS}(\widehat{\theta}_s^{\mathcal{C}}|0,s)}{\mathcal{P}_s^{\mathcal{C}}(\widehat{\theta}_s^{\mathcal{C}})}\right) = 0.$$

**Strategic policy.** Let $p_s^0 := P_{M|YS}(1|0,s)$ be the probability that unqualified individuals in $\mathcal{G}_s$ manipulate.[3] Under

[3] Because individuals manipulate by imitating those qualified (Sec.3), qualified individuals do not have incentives to manipulate.

a policy $\pi_s(x) = \mathbf{1}(x \geq \theta)$, the strategic decision maker's expected utility $U_s(\theta)$ over $\mathcal{G}_s$ is as follows:

$$\widehat{U}_s(\theta) - u_-(1-\alpha_s)\Big(\mathbb{F}_{X|YS}(\theta|0,s) - \mathbb{F}_{X|YS}(\theta|1,s)\Big)p_s^0$$

where $\widehat{U}_s(\theta)$ is the expected utility under non-strategic policy, $\mathbb{F}_{X|YS}(x|y,s) = \int_{-\infty}^{x} P_{X|YS}(z|y,s)dz$ denotes CDF.

**Definition 4.4.** Define *manipulation benefit* as the benefit an individual gains from manipulation, i.e.,

$$\Delta_s(\theta) := \mathbb{F}_{X|YS}(\theta|0,s) - \mathbb{F}_{X|YS}(\theta|1,s).$$

Then, unqualified individuals' best-response (i.e., manipulation probability introduced in Sec. 3) to a policy $\pi_s(x) = \mathbf{1}(x \geq \theta)$ can be equivalently written as

$$p_s^0(\theta) := p_s^0(\pi_s) = \mathbb{F}_{C_s}(\Delta_s(\theta)).$$

The detailed derivation is in Appendix F.1. Let $x_s^*$ be such that $P_{X|YS}(x_s^*|1,s) = P_{X|YS}(x_s^*|0,s)$, which is unique under Assumption 4.1. Then the manipulation probability $p_s^0(\theta)$ is single-peaked with maximum occurring at $x_s^*$, and $\lim_{\theta\to-\infty} p_s^0(\theta) = \lim_{\theta\to+\infty} p_s^0(\theta) = 0$, meaning that when the threshold is sufficiently low or high, unqualified individuals are less likely to manipulate their features. Plugging this in the decision maker's utility, we have

$$U_s(\theta) = \widehat{U}_s(\theta) - \underbrace{u_-(1-\alpha_s)\Delta_s(\theta)\mathbb{F}_{C_s}(\Delta_s(\theta))}_{\textbf{term 2}:=\Psi_s(\Delta_s(\theta))}. \quad (2)$$

Define a function $\Psi_s(z) := u_-(1-\alpha_s)\mathbb{F}_{C_s}(z)z$, then **term 2** in Eqn. (2) can be written as $\Psi_s(\Delta_s(\theta))$, and can be interpreted as the additional loss incurred by the decision maker due to manipulation (equivalently, the average manipulation gain by group $\mathcal{G}_s$). Further, let $\Psi_s'(z)$ be denoted as the first order derivative of $\Psi_s(z)$.

**Definition 4.5.** $\Psi_s'(\Delta_s(\theta))$ indicates the decision maker's *marginal loss* caused by strategic manipulation (equivalently, the *marginal manipulation gain* of $\mathcal{G}_s$).

The thresholds for strategic policies are as follows.

**Lemma 4.6.** *For $(\theta_a^{UN}, \theta_b^{UN})$, the strategic optimal thresholds, $\frac{P_{X|YS}(\theta_s^{UN}|1,s)}{P_{X|YS}(\theta_s^{UN}|0,s)} = \frac{u_-(1-\alpha_s)-\Psi_s'(\Delta_s(\theta_s^{UN}))}{u_+\alpha_s-\Psi_s'(\Delta_s(\theta_s^{UN}))}$.*

**Strategic fair policy.** Strategic fair thresholds $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ are found by maximizing the total expected utility subject to fairness constraint $\mathcal{C}$, i.e., $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}}) = \text{argmax}_{(\theta_a,\theta_b)} n_a U_a(\theta_a) + n_b U_b(\theta_b)$ such that Eqn. (1) holds. They can be characterized by the following.

**Lemma 4.7.** *Let $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ be the thresholds in strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s\Big(\frac{P_{X|YS}(\theta_s^{\mathcal{C}}|0,s) - P_{X|YS}(\theta_s^{\mathcal{C}}|1,s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})}\Psi_s'(\Delta_s(\theta_s^{\mathcal{C}})) + \frac{u_+\alpha_s P_{X|YS}(\theta_s^{\mathcal{C}}|1,s) - u_-(1-\alpha_s)P_{X|YS}(\theta_s^{\mathcal{C}}|0,s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})}\Big) = 0.$$

Note that besides $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$, equations in Lemmas 4.6 and 4.7 may be satisfied by other threshold pairs that are not optimal. We will discuss this in the next section.

## 5. Impact of anticipating manipulations

**Impact on the optimal policy & utility function.** We first compare strategic policy $\theta_s^{\text{UN}}$ with non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, and examine how the policy and the decision maker's expected utility differ. Let $\overline{\Delta}_s := \max_\theta \Delta_s(\theta)$ be defined as the ***maximum*** *manipulation benefit* an individual in $\mathcal{G}_s$ may gain from all possible policies.

**Assumption 5.1.** $\forall s \in \{a,b\}$, the marginal manipulation gain of $\mathcal{G}_s$, $\Psi_s'(z) < \infty$, is non-decreasing over $[0, \overline{\Delta}_s]$.

Assumption 5.1 says that a group's *marginal* manipulation gain does not decrease as manipulation benefit increases. It implies that when each individual benefits more from manipulation (increased $\Delta_s(\theta)$), more are incentivized to manipulate and the total loss of the decision maker caused by manipulation (or group's total manipulation gain) increases *faster*. This a very natural assumption: the incentives for manipulation increases when manipulation benefit increases. Examples (e.g., beta/uniformly distributed cost) satisfying this can also be found in Appendix B. We assume it holds in Sections 5 and 6.

*Remark* 5.2. $\Psi_s'(0) = 0$ and $\Psi_s'(\Delta_s(\theta))$ is single-peaked with maximum occurring at $x_s^*$.

For simplicity, let $\delta_u = \frac{u_-}{u_-+u_+}$, $\nu_s = \max\{u_+\alpha_s, u_-(1-\alpha_s)\}$. Define set $\mathcal{Z} = \{z|\Psi_s'(\Delta_s(z)) = \nu_s\}$ whose cardinality $|\mathcal{Z}| \in \{0,1,2\}$ depends on the maximum marginal manipulation gain $\Psi_s'(\Delta_s(x_s^*))$: if $\Psi_s'(\Delta_s(x_s^*)) > \nu_s$, then $|\mathcal{Z}| = 2$ and denote $\mathcal{Z} = \{\underline{z_s}, \overline{z_s}\}$ where $\underline{z_s} < x_s^* < \overline{z_s}$ (see Fig. 1 (top) for an illustration).

**Theorem 5.3.** *1. If $\alpha_s = \delta_u$, then $\theta_s^{UN} = \widehat{\theta}_s^{UN} = x_s^*$ when $|\mathcal{Z}| \leq 1$, otherwise $\theta_s^{UN} \in \mathcal{Z}$.*

*2. If $\alpha_s < \delta_u$ (resp. $\alpha_s > \delta_u$), then $\theta_s^{UN} > \widehat{\theta}_s^{UN} > x_s^*$ (resp. $\theta_s^{UN} < \widehat{\theta}_s^{UN} < x_s^*$). Moreover, if $|\mathcal{Z}| = 2$, then $\widehat{\theta}_s^{UN} > \overline{z_s}$ (resp. $\widehat{\theta}_s^{UN} < \underline{z_s}$) and $U_s(\theta)$ may have additional extreme points in $(\underline{z_s}, x_s^*)$ (resp. $(x_s^*, \overline{z_s})$); otherwise $\widehat{\theta}_s^{UN}$ is the unique extreme point of $U_s(\theta)$.*

Thm. 5.3 shows that as compared to a non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, a strategic policy $\theta_s^{\text{UN}}$ over(under) selects when a group is majority-(un)qualified[4]. In either case, as shown by Thm. 5.3, this means $\widehat{\theta}_s^{\text{UN}}$ is always closer to $x_s^*$ (the single

[4]We say $\mathcal{G}_s$ is majority-unqualified (resp. majority-qualified) if $\alpha_s < \delta_u$ (resp. $\alpha_s > \delta_u$): if $\mathcal{G}_s$ is majority-qualified, then the total utility the decision maker receives by accepting all individuals in $\mathcal{G}_s$ is $\alpha_s u_+ - (1-\alpha_s)u_- > 0$. Intuitively, it measures a group's overall qualification level in terms of potential benefit it could bring to the decision maker.

peak of $p_s^0(\theta)$) compared to $\theta_s^{\text{UN}}$. Therefore, the strategic policy always disincentivizes manipulative behavior, i.e., manipulation probability $p_s^0(\theta_s^{\text{UN}}) < p_s^0(\widehat{\theta}_s^{\text{UN}})$.

Moreover, Thm. 5.3 states that $U_s(\theta)$ has multiple extreme points if maximum marginal manipulation gain $\Psi_s'(\Delta_s(x_s^*))$ is sufficiently large, and it also specifies the range of those extreme points. In other words, although both $\widehat{U}_s(\theta)$ (non-strategic utility) and $\Psi_s(\Delta_s(\theta))$ are single-peaked with unique extreme points, their difference $U_s(\theta)$ (Eqn. (2)) may have multiple extreme points. As we will see later, this results in strategic and non-strategic policies having different properties in many aspects.

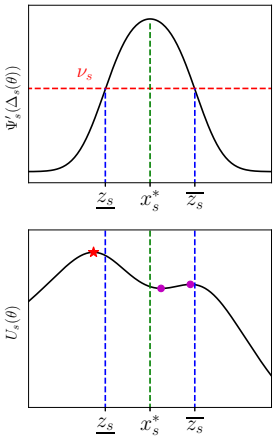



*Figure 1.* Illustration of functions: $\Psi_s'(\Delta_s(\theta)), U_s(\theta)$

An example of $U_s(\theta)$ is shown in Fig. 1 (bottom), where $X|Y=0, S=s$ and $X|Y=1, S=s$ are both Gaussian distributed with the same variance, and manipulation cost $C_s$ is beta distributed, $\alpha_s > \delta_u$. The red star is the optimal threshold $\theta_s^{\text{UN}} < \underline{z_s}$; two magenta dots are other extreme points of $U_s(\theta)$, which are in $(x_s^*, \overline{z_s})$.

**Impact on fairness.** The characterization of a strategic policy $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and a non-strategic policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ allows us to further compare them against a given fairness criterion $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$. Suppose we define the *unfairness* of a threshold policy $(\theta_a, \theta_b)$ as $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b) = \mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\mathbf{1}(x \geq \theta_a)] - \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\mathbf{1}(x \geq \theta_b)] = \mathbb{F}_b^{\mathcal{C}}(\theta_b) - \mathbb{F}_a^{\mathcal{C}}(\theta_a)$, where the CDF $\mathbb{F}_s^{\mathcal{C}}(\theta) = \int_{-\infty}^{\theta} \mathcal{P}_s^{\mathcal{C}}(x)dx$. Define the *disadvantaged group* under a policy $(\theta_a, \theta_b)$ as the group with the larger $\mathbb{F}_s^{\mathcal{C}}(\theta_s)$, i.e., the group with the smaller selection rate (DP) or the smaller true positive rate (EqOpt). Define group index $-i := \{a, b\} \setminus i$. Note that we measure unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b)$ over the original feature distributions $P_{X|YS}(x|y, s)$ before manipulation.

We first identify distributional conditions under which the strategic optimal policy worsens unfairness.

**Theorem 5.4.** *If $\alpha_i > \delta_u > \alpha_{-i}$ (i.e., $\mathcal{G}_i(\mathcal{G}_{-i})$ is majority-(un)qualified) and $\mathcal{G}_{-i}$ is disadvantaged under the non-strategic policy, then the strategic policy $(\theta_a^{UN}, \theta_b^{UN})$ has worse unfairness than the non-strategic $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$, i.e., $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{UN}, \theta_b^{UN})\right| > \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})\right|$, $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$. Moreover, the disadvantaged group under $(\theta_a^{UN}, \theta_b^{UN})$ and $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$ is the same.*

Thm. 5.4 is applicable when one group is majority-qualified while the other majority-unqualified and disadvantaged under a non-strategic policy, a common occurrence in the real world as the less qualified group is typically less selected and disadvantaged. Because the majority-(un)qualified $\mathcal{G}_i(\mathcal{G}_{-i})$ is over(under)-selected under a strategic policy (by Thm.5.3), $\mathcal{G}_{-i}$ becomes more disadvantaged while $\mathcal{G}_i$ becomes more advantaged, i.e., the unfairness gap is wider under the strategic policy.

We next identify conditions on the manipulation cost, under which a strategic policy $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ can lead to a more equitable outcome or flip the (dis)advantaged group compared to a non-strategic $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

**Theorem 5.5.** *If $\alpha_a, \alpha_b < \delta_u$ (i.e., both groups are majority-unqualified) and $\mathcal{G}_{-i}$ is disadvantaged under non-strategic policy, then given any $\mathcal{G}_{-i}$, there always exists cost $C_i$ for group $\mathcal{G}_i$ such that its maximum marginal manipulation gain $\Psi_i'(\Delta_i(x_i^*))$ is sufficiently large and*

1. *$(\theta_a^{UN}, \theta_b^{UN})$ mitigates the unfairness; or*
2. *$(\theta_a^{UN}, \theta_b^{UN})$ flips the disadvantaged group from $\mathcal{G}_{-i}$ to $\mathcal{G}_i$.*

Intuitively, as $\mathcal{G}_i$'s manipulation cost decreases, more individuals from $\mathcal{G}_i$ can afford manipulation; thus a strategic decision maker disincentivizes manipulation by increasing the threshold $\theta_i^{\text{UN}}$. For any $\mathcal{G}_{-i}$, as $\mathbb{F}_i^{\mathcal{C}}(\theta_i^{\text{UN}})$ increases, either the unfairness gets mitigated or $\mathbb{F}_i^{\mathcal{C}}(\theta_i^{\text{UN}})$ becomes larger than $\mathbb{F}_{-i}^{\mathcal{C}}(\theta_{-i}^{\text{UN}})$. Proposition D.1 in Appendix D considers a special case when $P_{X|YS}(x|y, a) = P_{X|YS}(x|y, b)$, and gives conditions on $\Psi_s'(\cdot)$ under which $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ mitigates the unfairness or flips the disadvantaged group.

## 6. Impact of fairness interventions

In this section, we study how non-strategic and strategic policies are affected by fairness interventions $\mathcal{C} \in \{\texttt{DP}, \texttt{EqOpt}\}$.

**Impact of fairness intervention on non-strategic policy.** First, we consider a non-strategic decision maker and compare $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ with $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$, both ignoring strategic manipulation but the latter imposing a fairness criterion. Thm. 6.1 identifies conditions under which a fairness constrained $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ yields *higher* utility from both groups compared to unconstrained $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$. It is worth noting because had strategic manipulation been absent, $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ by definition would attain the optimal/highest utility for decision maker.

**Theorem 6.1.** *Let $\nu_s = \max\{u_+\alpha_s, u_-(1-\alpha_s)\}$. Suppose $\Psi_s'(\Delta_s(\widehat{\theta}_s^{\mathcal{C}})) > \nu_s, \forall s \in \{a, b\}$ and $\mathcal{G}_{-i}$ is disadvantaged under the non-strategic optimal policy, then any of the following three cases could result in $U_s(\widehat{\theta}_s^{\mathcal{C}}) > U_s(\widehat{\theta}_s^{\text{UN}}), \forall s \in \{a, b\}$:*

1. *$\alpha_i < \delta_u < \alpha_{-i}$*
2. *$\alpha_a, \alpha_b > \delta_u$ and $\alpha_i \to \delta_u$*
3. *$\alpha_a, \alpha_b < \delta_u$ and $\alpha_{-i} \to \delta_u$*

Condition $\alpha_i \to \delta_u$ (resp. $\alpha_{-i} \to \delta_u$) means that qualification rate $\alpha_i$ (resp. $\alpha_{-i}$) is sufficiently close to $\delta_u$. Thm. 6.1 says that when the marginal manipulation gain of both groups under a non-strategic fair policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ are sufficiently large, $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ may outperform $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ in terms of both fairness and utility due to the misalignment of $U_s(\theta)$ and $\widehat{U}_s(\theta)$ caused by manipulation. This means that if the decision maker lacks information or awareness to anticipate manipulative behavior (but which in fact exists), then it would benefit from using a fairness constrained policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ rather than $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

**Impact of fairness intervention on the strategic policy.** We now compare $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ with $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$. We also explore their respective subsequent impact on individuals' manipulative behavior by comparing manipulation probabilities $\big(p_a^0(\theta_a^{\text{UN}}), p_b^0(\theta_b^{\text{UN}})\big)$ and $\big(p_a^0(\theta_a^{\mathcal{C}}), p_b^0(\theta_b^{\mathcal{C}})\big)$. The goal here is to understand whether fairness intervention can serve as incentives or disincentives for strategic manipulation. According to Thm. 5.3, $U_s(\theta)$ may have multiple extreme points under strategic manipulation if the group's marginal manipulation gain is sufficiently large. Depending on whether $U_s(\theta)$ has multiple extreme points, different conclusions result as outlined in Theorems 6.2 and 6.3 below, which identifies conditions under which fairness intervention may increase the manipulation incentive for one group while disincentivizing the other, or it may serve as incentives for both groups. Denote $p_s^{\mathcal{C}} := p_s^0(\theta_s^{\mathcal{C}})$ and $p_s^{\text{UN}} := p_s^0(\theta_s^{\text{UN}})$.

**Theorem 6.2.** *When at least one of $U_a(\theta)$, $U_b(\theta)$ has multiple extreme points, then it is possible that $\forall s \in \{a, b\}$, $\theta_s^{UN} > \theta_s^{\mathcal{C}}$ or $\theta_s^{UN} < \theta_s^{\mathcal{C}}$, i.e., both groups are over/under selected under fair policies. In this case,*

1. *If $\alpha_i > \delta_u > \alpha_{-i}$, then $(p_i^{UN} - p_i^{\mathcal{C}})(p_{-i}^{UN} - p_{-i}^{\mathcal{C}}) < 0$.*
2. *If $\alpha_a, \alpha_b > \delta_u$ (or $\alpha_a, \alpha_b < \delta_u$), then either $p_a^{UN} < p_a^{\mathcal{C}}$, $p_b^{UN} < p_b^{\mathcal{C}}$ or $(p_a^{UN} - p_a^{\mathcal{C}})(p_b^{UN} - p_b^{\mathcal{C}}) < 0$.*

When not accounting for strategic manipulation, $\widehat{U}_s(\theta)$ has a unique extreme point, and imposing a fairness constraint results in one group getting under-selected and the other over-selected. In contrast, when the decision maker anticipates strategic manipulation, $U_s(\theta)$ may have multiple extreme points. One consequence of this difference is that both $\mathcal{G}_a$ and $\mathcal{G}_b$ may be over- or under-selected when fairness is imposed, resulting in more complex incentive relationships. Specifically, if one group is majority-qualified while the other is majority-unqualified (i.e., $\alpha_i > \delta_u > \alpha_{-i}$), then under-selecting (resp. over-selecting) both groups under fair policies will increase (resp. decrease) the incentives of the former to manipulate, while disincentivizing (resp. incentivizing) the latter (by *Case 1*); if both groups are majority-(un)qualified (i.e., $\alpha_a, \alpha_b \lessgtr \delta_u$), then a fair policy may incentivize both to manipulate (by *Case 2*).

If the marginal manipulation gain of both groups are not sufficiently large, i.e., both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points, then fairness intervention always results in one group getting over-selected and the other under-selected. However, its subsequent impact on incentives may vary depending on $P_{X|YS}(x|y,s)$, $n_s$, as shown in Thm. 6.3.

**Theorem 6.3.** *When both $U_a(\theta)$, $U_b(\theta)$ have unique extreme points, we have $\theta_i^{UN} > \theta_i^{\mathcal{C}}$ and $\theta_{-i}^{UN} < \theta_{-i}^{\mathcal{C}}$. Moreover,*

1. *If $\alpha_i > \delta_u > \alpha_{-i}$, then $\forall \alpha_{-i}$, $\exists \kappa, \tau \in (0,1)$ such that $\forall \alpha_i > \kappa$ and $\forall n_i > \tau$, we have $p_i^{UN} < p_i^{\mathcal{C}}$, $p_{-i}^{UN} > p_{-i}^{\mathcal{C}}$.*
2. *If $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$), then $\forall \alpha_{-i}$, there exists $\kappa \in (\delta_u, 1)$ (resp. $\kappa \in (0, \delta_u)$) such that $\forall\ \alpha_i > \kappa$ (resp. $\alpha_i < \kappa$), we have $(p_a^{UN} - p_a^{\mathcal{C}})(p_b^{UN} - p_b^{\mathcal{C}}) < 0$.*

Thm. 6.3 identifies two scenarios under which fair policies incentivize one group (say $\mathcal{G}_i$) while disincentivizing the other ($\mathcal{G}_{-i}$): when $\mathcal{G}_i$ is majority-qualified, $\mathcal{G}_{-i}$ majority-unqualified, and $\mathcal{G}_i$ sufficiently qualified $(\alpha_i > \kappa)$ and represented in the entire population $(n_i > \tau)$ (by *Case 1*); or, when both are majority-(un)qualified and one group sufficiently (un)qualified (by *Case 2*).

Next, we identify conditions under which fairness intervention can *disincentivize* both groups. Let $x_s^{\text{UN}}$ be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$. Note that $x_s^{\text{UN}}$ is the point at which $p_s^0(x_s^{\text{UN}}) = p_s^0(\theta_s^{\text{UN}})$. Because manipulation probability is single-peaked, fairness intervention incentivizes manipulative behavior of $\mathcal{G}_s$ if $\theta_s^{\mathcal{C}}$ falls between $x_s^{\text{UN}}$ and $\theta_s^{\text{UN}}$.

**Theorem 6.4** (Disincentives for both groups)**.** *When both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points. If $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$) and $\mathbb{F}_{-i}^{\mathcal{C}}(x_{-i}^{UN}) < \mathbb{F}_i^{\mathcal{C}}(x_i^*)$ (resp. $\mathbb{F}_{-i}^{\mathcal{C}}(x_{-i}^{UN}) > \mathbb{F}_i^{\mathcal{C}}(x_i^*)$), then $\exists \kappa, \tau \in (0,1)$ s.t. $\forall \alpha_i \in (\delta_u, \kappa)$ (resp. $\alpha_i \in (\kappa, \delta_u)$) and $\forall n_i > \tau$, we have $p_a^{UN} > p_a^{\mathcal{C}}$ and $p_b^{UN} > p_b^{\mathcal{C}}$.*

Note that $x_i^*$ depends on $P_{X|YS}(x|y,i)$ and $x_{-i}^{\text{UN}}$ is determined by $u_-, u_+$, $P_{X|YS}(x|y,-i)$ and $\alpha_{-i}$. Thm. 6.4 says that when both groups are majority-(un)qualified, for certain population distributions and $\mathcal{G}_{-i}$, fair policies disincentivize both groups if $\mathcal{G}_i$ is sufficiently unqualified(qualified) and sufficiently represented in the population. For a special Gaussian case, conditions for satisfying $\mathbb{F}_{-i}^{\mathcal{C}}(x_{-i}^{\text{UN}}) \lessgtr \mathbb{F}_i^{\mathcal{C}}(x_i^*)$ in Thm. 6.4 are given in Proposition D.2 in Appendix D.

Theorems 6.2, 6.3 and 6.4 suggest that the impact of fairness intervention on the individuals' manipulative behavior highly depends on manipulation costs, feature distributions, group qualification and representation. This complexity stems from the misalignment in manipulation probability $p_s^0(\theta)$, utility $U_s(\theta)$, and fairness $\mathcal{C}$. In particular, the manipulation probability of $\mathcal{G}_s$ is single-peaked with maximum at $x_s^*$, which does not depend on group qualification and representation, but on which the decision maker's total utility depends, as varying $\alpha_s$ and $n_s$ will affect the policies.

As a result, depending on which region $\theta_s^{\text{UN}}$ falls into, i.e., smaller or larger than $x_s^*$, and how it may change under constraint $\mathcal{C}$, fairness intervention will have different impacts on incentives.

Although Theorems 6.2, 6.3 and 6.4 hold for both `EqOpt` and `DP` fairness, there are scenarios under which they have different impact on incentives. Proposition D.3 in Appendix D considers a special case when $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$ and one group is majority-qualified while the other majority-unqualified, in which `EqOpt` never disincentivize both groups while `DP` can disincentivize both.

## 7. Discussion

*More complicated model to capture strategic behavior.* In practice, individual strategic behavior can be much more complicated than modeled here: those considered qualified may also have an incentive to manipulate, and manipulation may only lead to partial improvement in features. The latter can be modeled by introducing a sequence of progressively "better" distributions (each with a different manipulation cost), and the goal of manipulation is to imitate/acquire a distribution better than one's own. The model studied in this paper is essentially the two-distribution (one for the unqualified, one for the qualified) version of this more general model. Even in this simplified model, there exists a complex relationship between fairness intervention and incentives for strategic manipulation as we have shown. Our results provide insights and build a foundation for analyzing more complicated models in future work.

*Beyond binary settings.* Our present model is limited to scenarios where individual qualification states and manipulation actions are binary. In reality, qualification states can be on a continuous spectrum, and individuals may face more complex manipulation decisions such as what features to manipulate, what types of actions to take, etc., than a binary decision of whether to manipulate or not. Going beyond the binary settings is also a direction of future research.

*Societal impact.* The paper aims to understand the impact of being able to anticipate manipulative behavior on policy/fairness, and the relationship between fairness interventions and incentives for manipulation. We identified conditions for each possible outcome. Our findings could help guide decisions on when to use a strategic policy and whether to impose fairness interventions, e.g., avoid using a strategic policy (or fairness intervention) if it exacerbates unfairness or incentivizes manipulation.

## 8. Experiments

We conduct experiments on both a Gaussian synthetic dataset, and the FICO scores dataset (Reserve, 2007). Due to the lack of real-world data on manipulation cost, we consider manipulation costs following either uniform ($C_s \sim U[0,\overline{c}]$) or beta distributions ($C_s \sim \text{Beta}(v,w)$, smaller $w$ and larger $v$ lead to larger manipulation costs, see Fig. 8 in Appendix E).[5] Note that these are examples for illustration, our results do not rely on these choices.

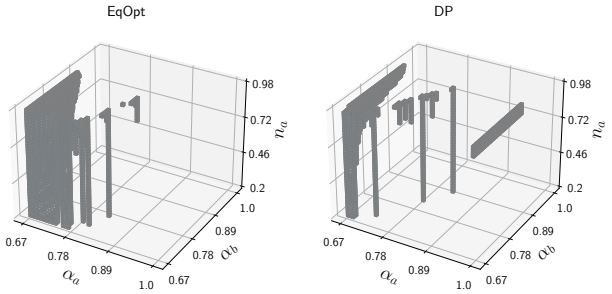



*Figure 2.* $\alpha_s, \alpha_b > \delta_u$, $C_a = C_b \sim \text{Beta}(10,1)$, $\frac{u_+}{u_-} = \frac{1}{2}$. Grey region is $(\alpha_a, \alpha_b, n_a)$ satisfying $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$ in Thm. 6.4; meanwhile both groups are disincentivized under $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$.

**Gaussian data.** Suppose $X|Y,S$ is Gaussian distributed. Fig. 2 shows an example where fairness intervention can serve as disincentive for manipulation for both groups. In particular, gray dots indicate scenarios (specified by $n_a$, $\alpha_a, \alpha_b$) under which strategic fair decisions discourage both groups to manipulate; these scenarios match the conditions in Thm 6.4, i.e., if $\alpha_a, \alpha_b > \delta_u$, for $\mathcal{G}_b$ that satisfies $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$, both groups are disincentivized when $\mathcal{G}_a$ is sufficiently represented ($n_a \to 1$) and unqualified ($\alpha_a \to \delta_u$). Detailed parameters and more experiments (e.g., verification of Theorems 5.4, 5.5, and 6.3) on Gaussian data can be found in Appendix E.

**FICO scores (Reserve, 2007).** FICO scores are widely used in the US to assess an individual's creditworthiness. This is a dataset pre-processed by (Hardt et al., 2016b) to generate CDF of scores $\mathbb{F}_{X|S}(x|s)$ and qualification profile $P_{Y|XS}(1|x,s)$ for different social groups (Caucasian, African-American, Hispanic, Asian). We use these to estimate the conditional feature distribution $P_{X|YS}(x|y,s)$ by fitting the simulated data to a Beta distribution. This allows us to derive various equilibrium strategies studied in this paper. We also calculate repayment rates $\alpha_s$ and proportions $n_s$. These are summarized in Figures 15-16 and Table 3 in Appendix E. Here we focus on beta distributed costs, results for the uniformly distributed $C_a, C_b$ are in Appendix E.

We first compare strategic $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and non-strategic policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ in terms of their fairness. Let $\mathcal{G}_a$ denote Caucasian, Hispanic or Asian, and $\mathcal{G}_b$ denote African-American.

[5] Uniformly distributed $C_s$ has been adopted in (Liu et al., 2020). In economics, a choice of *generalized beta distribution* is common to model costs (e.g., healthcare costs (Jones et al., 2014)).

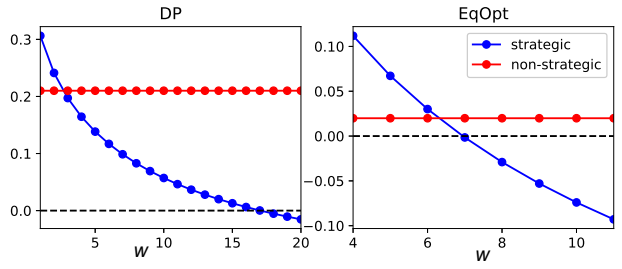



*Figure 3.* Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$, $\frac{u_+}{u_-} = \frac{1}{2}$, $\alpha_a, \alpha_b < \delta_u$. Perfect fairness is indicated by the black dashed line. $C_b \sim \text{Beta}(10, 5)$, $C_a \sim \text{Beta}(10, w)$, where larger $w$ indicates smaller costs.

*Table 1.* Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ for $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$: $\mathcal{G}_b$ = African-American, $u_+ = u_-$, $C_a \sim \text{Beta}(10, 2)$. When cost $C_a \neq C_b$, $C_b \sim \text{Beta}(10, 6)$.

| | $\mathcal{G}_a$ | strategic | | non-strategic |
|---|---|---|---|---|
| | | $C_a = C_b$ | $C_a \neq C_b$ | |
| EqOpt | Caucasian | 0.355 | 0.556 | 0.136 |
| | Hispanic | 0.292 | 0.493 | 0.034 |
| | Asian | 0.333 | 0.533 | 0.123 |
| DP | Caucasian | 0.611 | 0.680 | 0.449 |
| | Hispanic | 0.421 | 0.490 | 0.242 |
| | Asian | 0.634 | 0.703 | 0.522 |

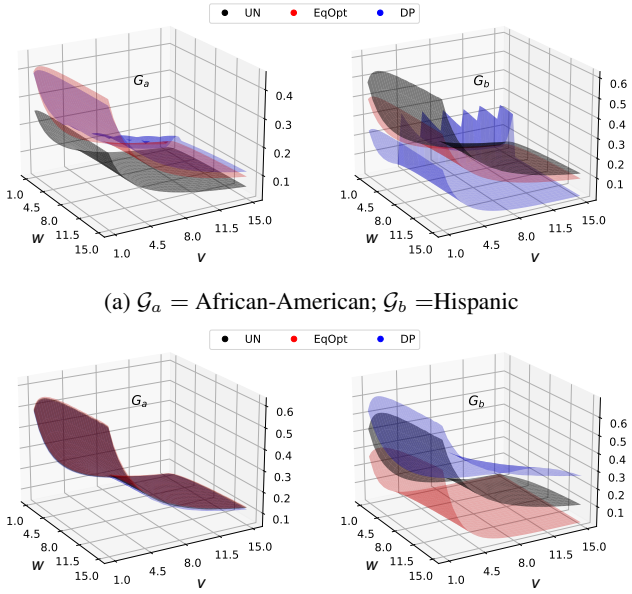
As shown in Table 1, $\mathcal{G}_b$ is always disadvantaged compared to other groups, and strategic policy worsens unfairness. When $C_a \neq C_b$, the manipulation cost of $\mathcal{G}_b$ is shifted lower. It further shows that this gets worse when it is less costly for those in $\mathcal{G}_b$ to manipulate their features. Since $\alpha_a > \delta_u > \alpha_b$, this is consistent with Thm. 5.4.

Fig. 3 illustrates how unfairness can be mitigated and how the disadvantaged group can gain advantage under strategic policy. Let $\mathcal{G}_a, \mathcal{G}_b$ be Hispanic and African-American respectively. We fix $\mathcal{G}_b$ and vary $\mathcal{G}_a$'s manipulation cost. It shows while $\mathcal{G}_b$ is disadvantaged under non-strategic policy ($\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}}) > 0$), unfairness can be mitigated under strategic policy as $\mathcal{G}_a$'s manipulation cost decreases, and the disadvantaged group may gain an advantage in the process ($\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) < 0$). This is an example of Thm. 5.5.

*Table 2.* $\mathcal{G}_a$ = Caucasian($\alpha_a = 0.758$), $\mathcal{G}_b$ = Asian($\alpha_b = 0.804$), $\mathcal{C} = \texttt{EqOpt}$. $C_b \sim \text{Beta}(10, 10)$. The first (resp. second) row corresponds to case 1 (resp. case 2) in Thm. 6.1.

| $\delta_u$ | $C_a$ | $U_a(\widehat{\theta}_a^{\text{UN}})$ | $U_a(\widehat{\theta}_a^{\mathcal{C}})$ | $U_b(\widehat{\theta}_b^{\text{UN}})$ | $U_b(\widehat{\theta}_b^{\mathcal{C}})$ |
|---|---|---|---|---|---|
| 0.8 | Beta(10, 10) | -0.190 | -0.189 | 0.024 | 0.034 |
| 0.756 | Beta(10, 1) | 0.396 | 0.397 | 0.181 | 0.201 |

According to Thm. 6.1, under strategic manipulation, a non-strategic fair policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ may yield higher utilities from both groups compared to $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$. We verify this in




(a) $\mathcal{G}_a$ = African-American; $\mathcal{G}_b$ =Hispanic

(b) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Asian

*Figure 4.* Manipulation probabilities under strategic (fair) policy: $C_a = C_b \sim \text{Beta}(v, w)$, $v, w \in [1, 15]$.

Table 2, in which $\mathcal{G}_a, \mathcal{G}_b$ denote Caucasian and Asian respectively, with EqOpt as the fairness constraint. It illustrates two cases (Case 1 & 2) in Thm. 6.1, and $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$, $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$ hold in both cases, i.e., $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ satisfies fairness and attains higher utility than $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

Lastly, we examine how fairness intervention acts as incentives for manipulation. Manipulation probabilities $p_s^0(\theta_s^{\text{UN}})$, $p_s^0(\theta_s^{\texttt{EqOpt}})$, and $p_s^0(\theta_s^{\texttt{DP}})$ are compared under different manipulation costs in Fig. 4. Here groups have the same manipulation costs $C_a = C_b \sim \text{Beta}(v, w)$ and $u_- = u_+$. Experiments on different manipulation costs ($C_a \sim U[0, \bar{c}_a]$, $C_b \sim U[0, \bar{c}_b]$) are shown in Appendix E. Black, red and blue surfaces indicate the manipulation probabilities $p_s^0(\theta_s)$ under $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$, $(\theta_a^{\texttt{EqOpt}}, \theta_b^{\texttt{EqOpt}})$ and $(\theta_a^{\texttt{DP}}, \theta_b^{\texttt{DP}})$ policies as manipulation costs change. It shows the complex impact fairness constraints have on (dis)incentives: in general, fair polices encourage one group to manipulate while disincentivizing the other (blue/red surface is above black for one but below for the other). However, when manipulation is very costly, African-American and Hispanic groups can both be incentivized under DP (in Fig. 4a, when $C_s \sim \text{Beta}(15, 1)$, blue surface is above black for both groups). Fig. 4b also shows the contrarian impact DP, EqOpt can have: one serves as incentive while the other disincentive. More experiments on other group pairs are in Appendix E.

## Acknowledgments

This work is supported by the NSF under grants IIS-2040800 and IIS-2112471.

## References


Alon, T., Dobson, M., Procaccia, A., Talgam-Cohen, I., and Tucker-Foltz, J. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1774–1781, 2020.

Barman, S. and Rathi, N. Fair cake division under monotone likelihood ratios. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 401–437, 2020.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 1234–1242. PMLR, 2021.

Braverman, M. and Garg, S. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Brückner, M. and Scheffer, T. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555, 2011.

Brückner, M., Kanzow, C., and Scheffer, T. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

Chen, Y., Wang, J., and Liu, Y. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 2020.

Coate, S. and Loury, G. C. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pp. 1220–1240, 1993.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Ganos, C., Ogrzal, T., Schnitzler, A., and Münchau, A. The pathophysiology of echopraxia/echolalia: relevance to gilles de la tourette syndrome. *Movement Disorders*, 27 (10):1222–1229, 2012.

Gergely, G. and Csibra, G. Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. *Roots of human sociality: Culture, cognition, and human interaction*, pp. 229–255, 2006.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 160–166, 2020.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016a.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016b.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268, 2019.

Jones, A. M., Lomas, J., and Rice, N. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, 29(4):649–670, 2014.

Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., and Vohra, R. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 677–678, 2020.

Khalili, M. M., Zhang, X., and Abroshan, M. Fair sequential selection using supervised learning models. *Advances in Neural Information Processing Systems*, 34, 2021a.

Khalili, M. M., Zhang, X., Abroshan, M., and Sojoudi, S. Improving fairness and privacy in selection problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021b.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 825–844, 2019.

Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6917–6926. PMLR, 13–18 Jul 2020.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019.

Reserve, U. F. 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*, 2007.

Shavit, Y., Edelman, B., and Axelrod, B. Causal strategic linear regression. In *International Conference on Machine Learning*, pp. 8676–8686. PMLR, 2020.

Strathern, M. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

Sundaram, R., Vullikanti, A., Xu, H., and Yao, F. Pac-learning for strategic classification. In *International Conference on Machine Learning*, pp. 9978–9988. PMLR, 2021.

Zhang, X. and Liu, M. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pp. 525–555. Springer, 2021.

Zhang, X., Khaliligarekani, M., Tekin, C., and Liu, M. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, X., Khalili, M. M., and Liu, M. Long-term impacts of fair machine learning. *ergonomics in design*, 28(3): 7–11, 2020a.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18457–18469, 2020b.

# Appendix

## A. Generalization to high dimensional feature space

All analysis and conclusions can be generalized to high dimensional feature space $X \in \mathbb{R}^d$. In this case, high dimensional features are first mapped to one dimensional qualification profile $\gamma_s(x) = P_{Y|XS}(1|x,s)$, based on which the decision maker makes decisions about individuals. A threshold policy is in the form of $\pi_s(x) = \mathbf{1}(\gamma_s(x) \geq \phi_s)$ with threshold $\phi_s \in [0,1]$.

Let $\gamma_s^{-1}(l_s) \subset \mathbb{R}^b$ be defined as the preimage of $l_s$ under qualification profile $\gamma_s$, then we can adjust all analysis using $\gamma_s^{-1}(\cdot)$. For example, the strict monotone likelihood ratio property in Assumption 4.1 can be adjusted as follows: *for any two likelihoods* $0 \leq \underline{l_s} < \overline{l_s} \leq 1$*, we have* $\gamma_s^{-1}([\overline{l_s}, 1]) \subset \gamma_s^{-1}([\underline{l_s}, 1])$, i.e., any individual who can get accepted under threshold $\overline{l_s}$ can also be accepted under any lower threshold $\underline{l_s}$.

Because $\gamma_s(x) = P_{Y|XS}(1|x,s) = \frac{1}{1+\frac{P_{X|YS}(x|0,s)}{P_{X|YS}(x|1,s)}\frac{(1-\alpha_s)}{\alpha_s}}$, (non-)strategic (fair) threshold $\phi_s$ in the space of qualification profile can be found based on $\frac{P_{X|YS}(\theta_s|1,s)}{P_{X|YS}(\theta_s|0,s)}$ given in Lemmas 4.2-4.7. Specifically, replace $\frac{P_{X|YS}(\theta_s|1,s)}{P_{X|YS}(\theta_s|0,s)}$ with $\frac{1-\alpha_s}{\alpha_s}\frac{\phi_s}{1-\phi_s}$, and $\Delta_s(\theta_s)$ with $\int_{x\in\gamma_s^{-1}([\phi_s,1])} P_{X|YS}(x|1,s) - P_{X|YS}(x|0,s)dx$ in Lemmas 4.2-4.7. Then the consequent policy $\pi_s(x) = \mathbf{1}(\gamma_s(x) \geq \phi_s)$ is the optimal policy.

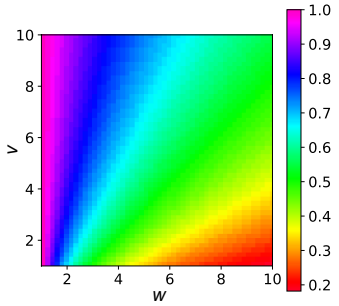
In multi-dimensional space, manipulation cost $C_s$ should be regarded as the sum total of effort/investment an individual makes to imitate features of a qualified individual. Specifically, an individual needs to manipulate multiple features to mimic a qualified individual's features; manipulation of each feature can induce some cost (which may or may not be correlated) and the overall effect is captured by the sum of all component costs, which is the total manipulation cost in our model.

## B. Assumption 5.1: $\Psi_s'(z) < \infty$ is non-decreasing over $[0, \overline{\Delta}_s]$

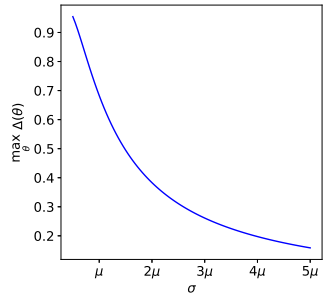
For simplicity, we drop subscript $s$ in the following.

**Example 1:** cost $C \sim U[0, \bar{c}]$. In this case, $\Psi'(z) = u_-(1-\alpha)\frac{2}{\bar{c}}z$ is non-decreasing.

**Example 2:** cost $C \sim \text{Beta}(v, w)$ with $v \in [1, 10]$, $w \in [1, 10]$.



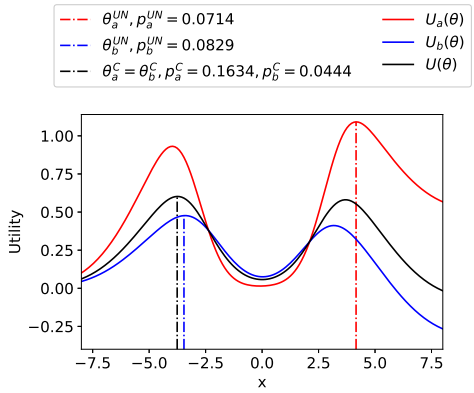

(a) $\tilde{\Delta}$ that ensures $\Psi'(z)$ to be non-decreasing over $[0, \tilde{\Delta}]$ when $C \sim \text{Beta}(v, w)$, $v \in [1, 10]$, $w \in [1, 10]$



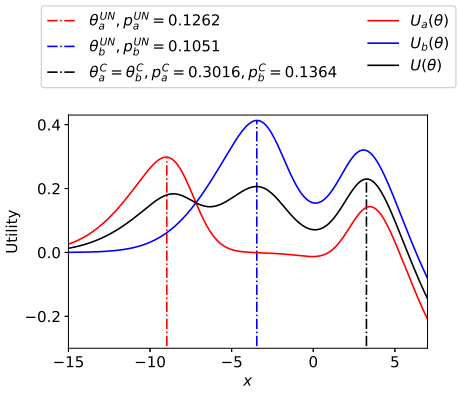

(b) $\overline{\Delta}$ for Gaussian distributed feature where $X|Y=1 \sim \mathcal{N}(\mu, \sigma^2)$, $X|Y=0 \sim \mathcal{N}(-\mu, \sigma^2)$, and $\mu > 0$.

For Beta distributed cost and Gaussian distributed features, the following figures show that the condition "$\Psi'(z)$ is non-decreasing over $[0, \overline{\Delta}]$" is relatively mild. For example, when $C \sim \text{Beta}(8, 3)$, the left plot shows that $\Psi'(z)$ is non-decreasing over $[0, 0.82]$. For features that follow Gaussian distributions $X|Y=1 \sim \mathcal{N}(\mu, \sigma^2)$ and $X|Y=0 \sim \mathcal{N}(-\mu, \sigma^2)$, the condition is satisfied as long as $\sigma > 0.72\mu$.

**Other examples:** There are many other probability density distributions with support $[0, 1]$ or $[0, \infty)$ that could satisfy this condition, such as beta prime distribution, gamma distribution, chi distribution, chi-squared distribution, etc.




(a) $X|Y=1, S=s \sim \mathcal{N}(5,4)$, $X|Y=0, S=s \sim \mathcal{N}(-5,4), \forall s \in \{a,b\}$, $u_- = u_+$, $n_a = 0.3$, $\alpha_a = 0.4, \alpha_b = 0.6$, $C_a \sim \text{Beta}(10,2), C_b \sim \text{Beta}(10,1)$, and fairness constraint $\mathcal{C} = \texttt{EqOpt}$. It shows that $\theta_s^{\mathcal{C}} < \theta_s^{\text{UN}}, \forall s \in \{a,b\}$ and $p_a^{\mathcal{C}} > p_a^{\text{UN}}, p_b^{\mathcal{C}} < p_b^{\text{UN}}$.




(b) $X|Y=1, S=s \sim \mathcal{N}(5,9), \forall s \in \{a,b\}$, $X|Y=0, S=b \sim \mathcal{N}(-5,9)$, $X|Y=0, S=a \sim \mathcal{N}(-10,9)$, $u_- = u_+$, $n_a = 0.5$, $\alpha_a = 0.65, \alpha_b = 0.6$, $C_a \sim \text{Beta}(10,3), C_b \sim \text{Beta}(10,2)$, and fairness constraint $\mathcal{C} = \texttt{EqOpt}$. It shows that $\theta_s^{\mathcal{C}} > \theta_s^{\text{UN}}$ and $p_s^{\mathcal{C}} > p_s^{\text{UN}}, \forall s \in \{a,b\}$.

## C. An example when both $U_a(\theta)$ and $U_b(\theta)$ have multiple extreme points

Because $P_{X|YS}(x|1,a) = P_{X|YS}(x|1,b)$, under EqOpt fairness, $\theta_a^{\mathcal{C}} = \theta_b^{\mathcal{C}}$ and the total utility $n_a U_a(\theta_a^{\mathcal{C}}) + n_b U_b(\theta_b^{\mathcal{C}})$ can be expressed as a function of $\theta = \theta_a^{\mathcal{C}} = \theta_b^{\mathcal{C}}$.

Two examples in above figures show that when $U_a(\theta)$ and $U_b(\theta)$ have multiple extreme points, it's possible that both groups are over (left)/under (right) selected under strategic fair policies. When $\alpha_b > \delta_u > \alpha_a$ (left), fairness intervention incentivizes $\mathcal{G}_a$ while disincentivizing $\mathcal{G}_b$; when $\alpha_a, \alpha_b > \delta_u$ (right), fairness intervention incentivizes both groups to manipulate. These results are consistent with Thm. 6.2.

## D. Additional Results

**Proposition D.1.** *Suppose $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, $\alpha_{-s} < \alpha_s < \delta_u$, then $\mathbb{F}_{-s}^{\mathcal{C}}(\widehat{\theta}_{-s}^{UN}) > \mathbb{F}_s^{\mathcal{C}}(\widehat{\theta}_s^{UN})$, i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy. Denote $\Delta(\cdot) = \Delta_a(\cdot) = \Delta_b(\cdot)$. Given any $\mathcal{G}_{-s}$, always there exists manipulation cost $C_s$ for $\mathcal{G}_s$ s.t. $\Psi_s'(\cdot)$ satisfies the followings:*

1. $\frac{\Psi_s'(\Delta(\theta_{-s}^{UN}))-u_+\alpha_s}{\Psi_{-s}'(\Delta(\theta_{-s}^{UN}))-u_+\alpha_{-s}} = \frac{u_-(1-\alpha_s)-u_+\alpha_s}{u_-(1-\alpha_{-s})-u_+\alpha_{-s}}$, *then $\theta_a^{UN} = \theta_b^{UN}$ and $(\theta_a^{UN}, \theta_b^{UN})$ mitigates unfairness.*

2. $\Psi_s'(\Delta(\eta^{\mathcal{C}}(\theta_{-s}^{UN}))) \geq u_-(1-\alpha_s)$, *then $(\theta_a^{UN}, \theta_b^{UN})$ flips the disadvantaged group.*

*where $\left(\eta^{\mathcal{C}}(\theta_{-s}^{UN}), \theta_{-s}^{UN}\right)$ satisfies fairness $\mathcal{C}$, i.e., $\eta^{EqOpt}(\theta_{-s}^{UN}) = \theta_{-s}^{UN}$, $\eta^{DP}(\theta_{-s}^{UN}) = (\mathbb{F}_s^{DP})^{-1}\mathbb{F}_{-s}^{DP}(\theta_{-s}^{UN})$.*

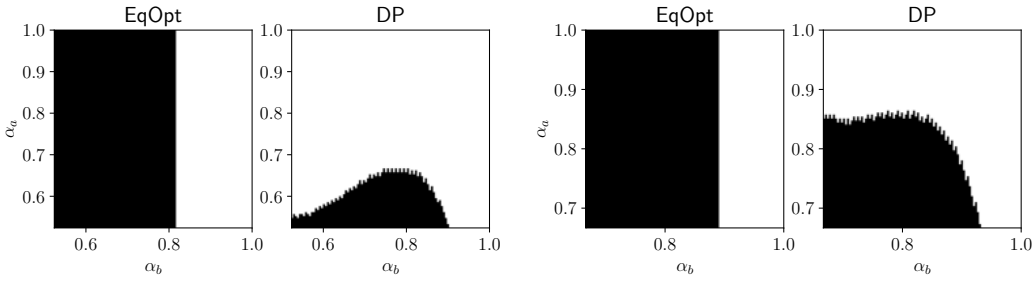
Proposition D.1 explicitly states the conditions on $\Psi_s'(\cdot)$ under which strategic policy mitigates the unfairness or flips the disadvantaged group. Note that these conditions are sufficient, especially for *Case 1*, where the perfect EqOpt fairness is attained (i.e., $\mathcal{E}^{\text{EqOpt}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) = 0$) and DP fairness is improved (i.e., $\left|\mathcal{E}^{\text{DP}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})\right| < \left|\mathcal{E}^{\text{DP}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})\right|$).

**Proposition D.2.** *Suppose $X|Y=y, S=s \sim \mathcal{N}(\mu_s^y, \sigma^2)$ with $0 < \mu_s^1 - \mu_s^0 < \mu_{-s}^1 - \mu_{-s}^0$, i.e., qualified and unqualified individuals from $\mathcal{G}_s$ are less distinguishable than those from $\mathcal{G}_{-s}$, then*

- *$\mathcal{C} = \texttt{EqOpt}$: $\forall \alpha_s > \delta_u$ (resp. $\alpha_s < \delta_u$), there exists $\omega > \delta_u$ (resp. $\omega < \delta_u$) such that $\forall \alpha_{-s} \in [\delta_u, \omega]$ (resp. $\alpha_{-s} \in [\omega, \delta_u]$), conditions $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Thm. 6.4 hold.*
- *$\mathcal{C} = \texttt{DP}$: if $u_+ < u_-$ (resp. $u_+ > u_-$ ), then there exist $\omega_1, \omega_2 > \delta_u$ (resp. $\omega_1, \omega_2 < \delta_u$) such that $\forall \alpha_b \in [\delta_u, \omega_1]$ (resp. $\forall \alpha_b \in [\omega_1, \delta_u]$) and $\forall \alpha_a \in [\delta_u, \omega_2]$ (resp. $\forall \alpha_a \in [\omega_2, \delta_u]$), conditions $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Thm. 6.4 hold.*

**Proposition D.3.** *Suppose $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, if $\alpha_s > \delta_u > \alpha_{-s}$, then*

- *$\forall P_{X|YS}(x|y,s)$, $p_a^{EqOpt} < p_a^{UN}$, $p_b^{EqOpt} < p_b^{UN}$ is unattainable, i.e., EqOpt never disincentivize both groups.*
- *$\exists P_{X|YS}(x|y,s)$, $(\alpha_a, \alpha_b)$, and $n_a$ under which $p_a^{DP} < p_a^{UN}$, $p_b^{DP} < p_b^{UN}$ , i.e., DP may disincentivize both groups.*




(a) $u_+ : u_- = 1 : 1.1$ (b) $u_+ : u_- = 1 : 2$

*Figure 7.* Examples validating Proposition D.2: black region indicates $(\alpha_a, \alpha_b)$ satisfying condition $\mathbb{F}^{\mathcal{C}}_{-s}(x^{\text{UN}}_{-s}) < \mathbb{F}^{\mathcal{C}}_{s}(x^*_s)$ in Thm.6.4: $\alpha_a, \alpha_b > \delta_u$, $C_a, C_b \sim \text{Beta}(10, 1)$, $X|Y = y, S = s \sim \mathcal{N}(\mu^y_s, \sigma^2)$ with $[\mu^0_a, \mu^1_a, \mu^0_b, \mu^1_b] = [-2, 2, -5, 5]$, $\sigma = 4.5$.

## E. Additional Experiments

In the experiments, we assume manipulation costs $C_a, C_b$ follow beta distributions $\text{Beta}(v, w)$ or uniform distributions $U[0, \overline{c}]$. For a beta distributed cost $C_s$, Fig. 8 illustrates examples of probability density function $f_s(z)$ and scaled marginal manipulation gain $\frac{\Psi'_s(z)}{u_-(1-\alpha_s)} = \mathbb{F}_{C_s}(z) + zf(z)$.

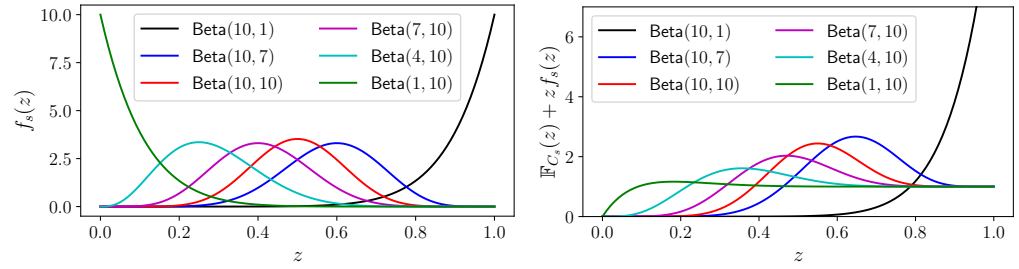



*Figure 8.* Illustration of $f_s(z)$ and $\mathbb{F}_{C_s}(z) + zf_s(z)$: $C_s \sim \text{Beta}(v, w)$, $v \in \{1, 4, 7, 10\}, w \in \{1, 7, 10\}$.

**Gaussian data.** We verify Thm. 5.4 by conducting 40 rounds of experiment independently. In each round of experiment, $(\alpha_a, \alpha_b)$ is randomly generated with $\alpha_a > \delta_u > \alpha_b$. We consider `EqOpt` (red) or `DP` (blue) as fairness measure. In Fig. 10, circles and stars represent the unfairness $\mathcal{E}^{\mathcal{C}}(\theta^{\text{UN}}_a, \theta^{\text{UN}}_b)$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$ respectively. It shows that the strategic policy (circles) always worsens the unfairness (both `EqOpt` and `DP`) compared to non-strategic policy (stars), and $\mathcal{G}_b$ is disadvantaged in all scenarios. Varying costs $C_s$, distributions $P_{X|YS}(x|y, s)$, and $u_+, u_-$, we observe the similar results.

Similarly, we verify Thm. 5.5 by running 40 rounds of experiments independently. In each round, $(\alpha_a, \alpha_b)$ is randomly generated with $\delta_u > \alpha_a > \alpha_b$. In Fig. 11, circles that fall below the black dashed line indicate the disadvantaged group being flipped under strategic policy. It shows as $\mathcal{G}_a$'s manipulation cost decreases, unfairness can be mitigated (circles fall below stars) and disadvantaged group can be flipped (circles fall below black dashed line).

Figures 9 and 12 illustrate the manipulation probabilities of two groups under strategic policy (`UN`) and strategic fair policy (`EqOpt`, `DP`), where $u_+ = u_-$, $C_s \sim \text{Beta}(10, 1)$, $X|Y = 1, S = b \sim \mathcal{N}(5, 5^2)$, $X|Y = 0, S = b \sim \mathcal{N}(-5, 5^2)$, $X|Y = 1, S = a \sim \mathcal{N}(5, 4^2)$, $X|Y = 0, S = a \sim \mathcal{N}(-5, 4^2)$. Black, blue, red surfaces correspond to $p^0_s(\theta^{\text{UN}}_s) := p^{\text{UN}}_s$, $p^0_s(\theta^{\text{DP}}_s) := p^{\text{DP}}_s$, $p^0_s(\theta^{\text{EqOpt}}_s) := p^{\text{EqOpt}}_s$ respectively. Fig. 9 shows that when $n_a$ and $\alpha_a$ are sufficiently large, $p^{\text{UN}}_a < p^{\mathcal{C}}_a$ and $p^{\text{UN}}_b > p^{\mathcal{C}}_b$ hold, $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$. Fig. 12 shows when two groups are majority-(un)qualified, $p^{\text{UN}}_a < p^{\mathcal{C}}_a$, $p^{\text{UN}}_b > p^{\mathcal{C}}_b$ or $p^{\text{UN}}_a > p^{\mathcal{C}}_a$, $p^{\text{UN}}_b < p^{\mathcal{C}}_b$ holds as long as one of $\alpha_a$, $\alpha_b$ is sufficiently large (small). These are consistent with Thm. 6.3.

Fig. 2 shows a scenario where fairness intervention can serve as disincentives for both groups, where $\frac{u_+}{u_-} = \frac{1}{2}$. In Fig. 13, we illustrate the case when $\frac{u_+}{u_-} = \frac{1}{1.1}$. In both cases, $X|Y = y, S = s \sim \mathcal{N}(\mu^y_s, 4.5^2)$ with $[\mu^0_a, \mu^1_a, \mu^0_b, \mu^1_b] = [-2, 2, -5, 5]$.

**FICO scores data.** The data pre-processed by (Hardt et al., 2016b) is publicly available which gives group proportions $n_s$, CDF of scores $\mathbb{F}_{X|S}(x|s)$ and qualification profiles $P_{Y|XS}(1|x, s)$ for four groups. It doesn't contain personally identifiable

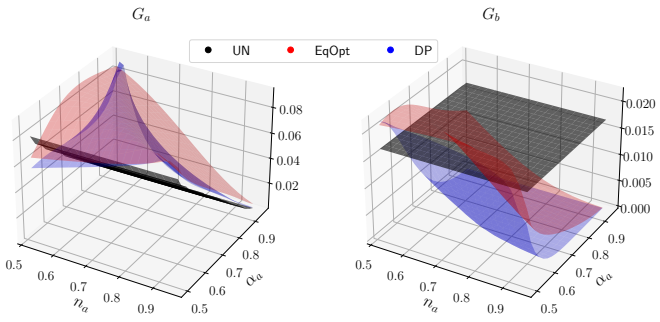



*Figure 9.* Verification of *Case 1* in Thm. 6.3: $\alpha_b = 0.4$. Varying $\mathcal{G}_a$'s qualification $\alpha_a \in [0.5, 1]$ and representation $n_a \in [0.5, 1]$, resulting manipulation probabilities of two groups are shown in plots.

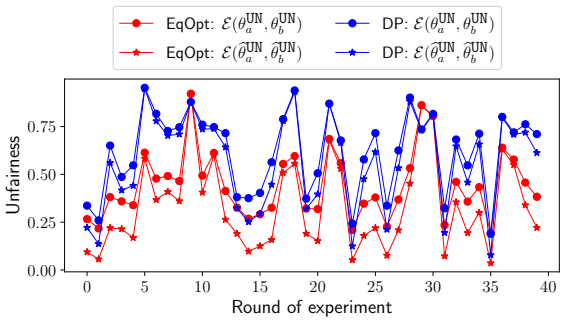
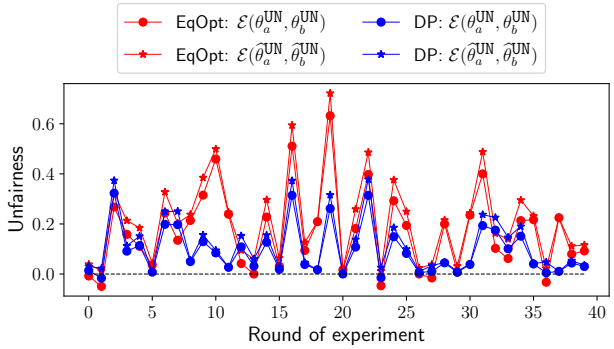
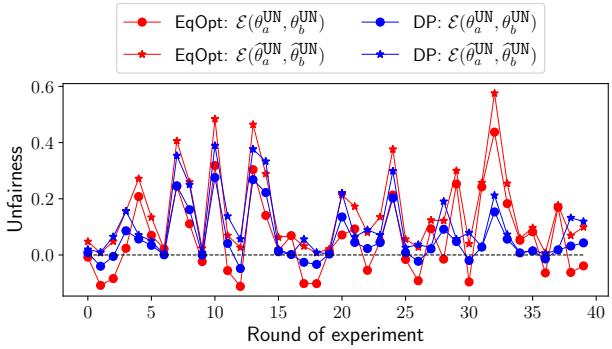


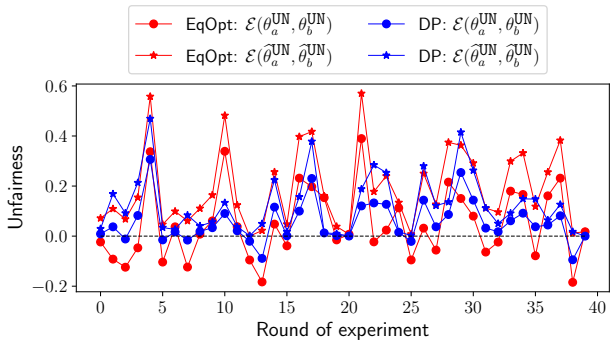

*Figure 10.* Verification of Thm. 5.4: $C_a \sim \text{Beta}(10, 1)$, $C_b \sim \text{Beta}(10, 3)$, $u_- = u_+$, $X|Y = 1, S = s \sim \mathcal{N}(5, 5^2)$ and $X|Y = 0, S = s \sim \mathcal{N}(-5, 5^2)$ for $s = a, b$.



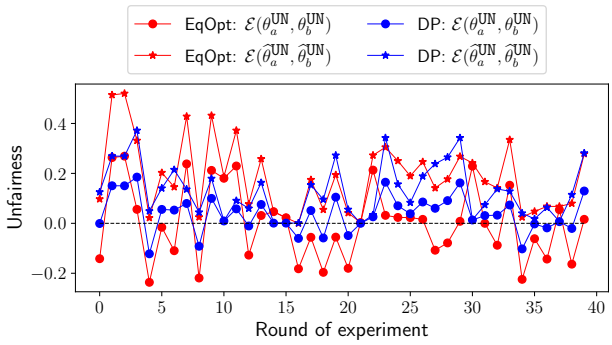

(a) $C_a \sim \text{Beta}(10, 2)$




(b) $C_a \sim \text{Beta}(10, 3)$




(c) $C_a \sim \text{Beta}(10, 4)$



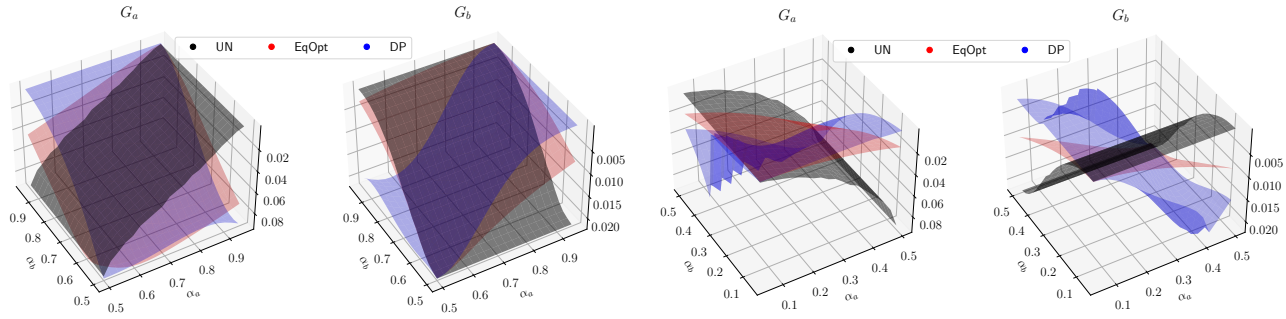

(d) $C_a \sim \text{Beta}(10, 5)$

*Figure 11.* $C_b \sim \text{Beta}(10, 1)$, $X|Y = 1, S = s \sim \mathcal{N}(5, 5^2)$, $X|Y = 0, S = s \sim \mathcal{N}(-5, 5^2)$, $s = a, b$.

information or offensive content.

From these, we can obtain PDF of scores $P_{X|S}(x|s)$ of each group. These are shown in Fig. 15. We then simulate a dataset based on joint probability distribution $P_{XY|S}(x, 1|s) = P_{Y|XS}(1|x, s) \cdot P_{X|S}(x|s)$. By fitting Beta distribution to the simulated data, we can obtain conditional feature distribution $P_{X|YS}(x|y, s)$ as shown in Fig. 16. We can see from Fig. 15 that $\frac{P_{X|YS}(x|1,s)}{P_{X|YS}(x|0,s)}$ is strictly increasing, it implies that Assumption 4.1 holds for FICO scores data. The details about qualification rate $\alpha_s = P_{Y|S}(1|s)$, conditional feature distributions $P_{X|YS}(x|y, s)$, and group proportions $n_s$ of four groups are summarized in Table 3.

Table 1 and Fig. 3 validate Thm. 5.4 and Thm. 5.5 respectively when manipulation costs follow Beta distributions. In Table 4 and Fig. 14, we present the similar results for uniformly distributed costs. These are still consistent with Thm. 5.4 and




(a) $\alpha_a \in [0.5, 1]$ and $\alpha_b \in [0.5, 1]$

(b) $\alpha_a \in [0, 0.5]$ and $\alpha_b \in [0, 0.5]$

*Figure 12.* Verification of *Case 2* in Thm. 6.3: $n_a = 0.5$. In the left (resp. right), varying two groups' qualification $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$), the resulting manipulation probabilities of two groups are shown in the plots.

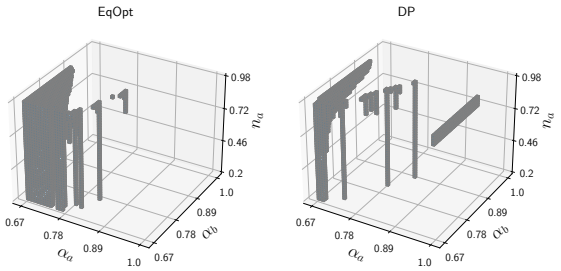



*Figure 13.* $\alpha_a, \alpha_b > \delta_u$, $C_a = C_b \sim \text{Beta}(10, 1)$, $\frac{u_+}{u_-} = \frac{1}{1.1}$. Grey region is $(\alpha_a, \alpha_b, n_a)$ satisfying $\mathbb{F}^{\mathcal{C}}_{-s}(x^{\text{UN}}_{-s}) < \mathbb{F}^{\mathcal{C}}_{s}(x^*_s)$ in Thm. 6.4; meanwhile both groups are disincentivized under $(\theta^{\mathcal{C}}_a, \theta^{\mathcal{C}}_b)$.

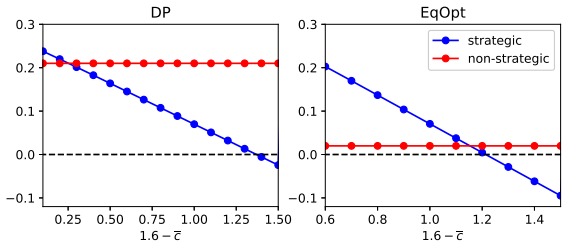



*Figure 14.* Unfairness $\mathcal{E}^{\mathcal{C}}(\theta^{\text{UN}}_a, \theta^{\text{UN}}_b)$, $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$: $\frac{u_+}{u_-} = \frac{1}{2}$, $\alpha_a, \alpha_b < \delta_u$. Perfect equity is indicated by the black dashed line. $C_b \sim U[0, 1]$, $C_a \sim U[0, \overline{c}]$.

Thm. 5.5.

Lastly, we provide additional experiments to examine the impact of fairness intervention on incentives for strategic manipulation. First, we consider the case when both groups have the same manipulation costs $C_a = C_b \sim \text{Beta}(v, w)$. Fig. 4 has shown the comparison of manipulation probabilities under strategic policy and strategic fair policy for two pairs of groups, i.e., when $(\mathcal{G}_a, \mathcal{G}_b)$ are (African-American, Hispanic) or (Caucasian, Asian). The results for other four group pairs, i.e., (Asian, African-American), (Asian, Hispanic), (Caucasian, African-American), (Caucasian, Hispanic), are shown in Fig. 17. For the case when manipulation costs are uniformly distributed and $C_a \neq C_b$, the comparisons for all group pairs are shown in Fig. 18. These results show that when there is a significant gap in the two groups' manipulation costs, fairness intervention incentivizes the group with a low manipulation cost while disincentivizing the group with a high manipulation cost.

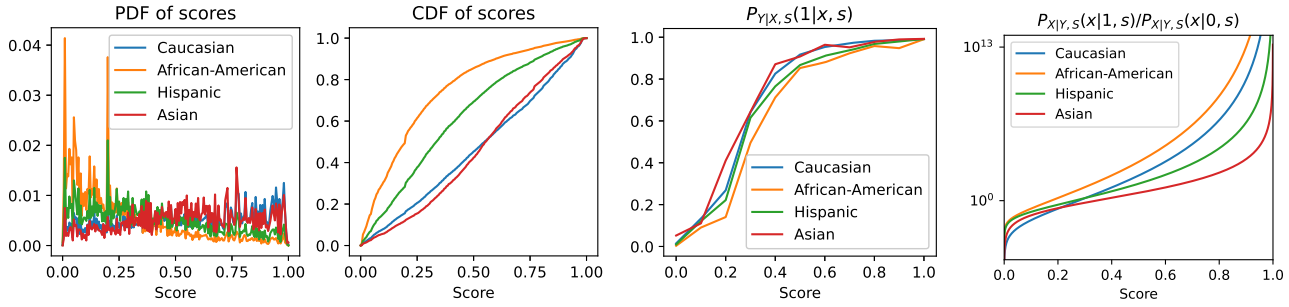



*Figure 15.* Illustration of PDF/CDF of scores, qualification profiles, and validation of Assumption 4.1.

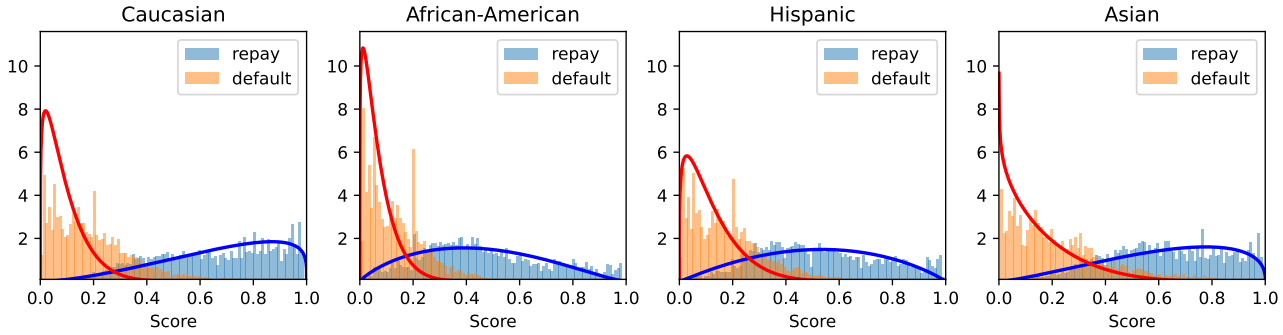



*Figure 16.* Fit Beta distributions to the simulated data to get $P_{X|Y,S}(x|y,s)$.

*Table 3.* Qualification rate $\alpha_a = P_{Y|S}(1|s)$, conditional feature distributions $P_{X|YS}(x|y,s)$, group proportions $n_s$ of four social groups. $x_s^*$ satisfies $P_{X|YS}(x_s^*|1,s) = P_{X|YS}(x_s^*|0,s)$.

| $\mathcal{G}_s$ | $\alpha_s$ | $P_{X\|YS}(x\|0,s)$ | $P_{X\|YS}(x\|1,s)$ | $n_s$ | $x_s^*$ |
|---|---|---|---|---|---|
| Caucasian | 0.758 | Beta$(1.23, 12.34)$ | Beta$(2.57, 1.24)$ | 0.7651 | 0.277 |
| African-American | 0.338 | Beta$(1.18, 15.99)$ | Beta$(1.84, 2.32)$ | 0.1050 | 0.174 |
| Hispanic | 0.570 | Beta$(1.23, 9.02)$ | Beta$(2.03, 1.90)$ | 0.0845 | 0.262 |
| Asian | 0.804 | Beta$(0.89, 4.94)$ | Beta$(2.31, 1.38)$ | 0.0454 | 0.342 |

*Table 4.* Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ for $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$: $\mathcal{G}_b$ = African-American, $u_+ = u_-$, $C_a \sim U[0,1]$. When cost $C_a \neq C_b$, $C_b \sim U[0, 0.5]$.

| | | EqOpt | | | DP | | |
|---|---|---|---|---|---|---|---|
| | $\mathcal{G}_a$ | strategic $C_a = C_b$ | strategic $C_a \neq C_b$ | non-strategic | strategic $C_a = C_b$ | strategic $C_a \neq C_b$ | non-strategic |
| Uniform | Caucasian | 0.743 | 0.871 | 0.136 | 0.794 | 0.838 | 0.449 |
| | Hispanic | 0.722 | 0.850 | 0.034 | 0.684 | 0.727 | 0.242 |
| | Asian | 0.738 | 0.866 | 0.123 | 0.825 | 0.868 | 0.522 |

# F. Proofs

## F.1. Proofs for Section 4

**Non-strategic policy.**

*Claim* F.1. The non-strategic optimal policy $\widehat{\pi}_s^{\text{UN}} = \arg\max_{\pi_s} \widehat{U}_s(\pi_s)$ is a threshold policy $\widehat{\pi}_s^{\text{UN}}(x) = \mathbf{1}\big(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-}\big)$.

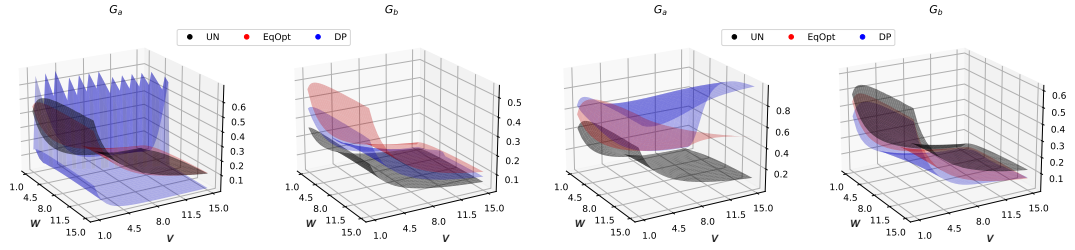
*Proof.* The non-strategic optimal policy $\widehat{\pi}_s^{\text{UN}} = \arg\max_{\pi_s} \widehat{U}_s(\pi_s)$ is given by

$$\widehat{\pi}_s^{\text{UN}}(x) = \begin{cases} 1, \text{ if } \frac{P_{X|YS}(x|1,s)}{P_{X|YS}(x|0,s)} \geq \frac{u_-(1-\alpha_s)}{u_+ \alpha_s} \\ 0, \text{ o.w.} \end{cases} \tag{3}$$

Re-writing based on qualification the profile $\gamma_s(x) = \dfrac{1}{\frac{P_{X|YS}(x|0,s)}{P_{X|YS}(x|1,s)} \frac{(1-\alpha_s)}{\alpha_s} + 1}$, (3) is reduced to

$$\widehat{\pi}_s^{\text{UN}}(x) = \mathbf{1}\Big(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-}\Big).$$

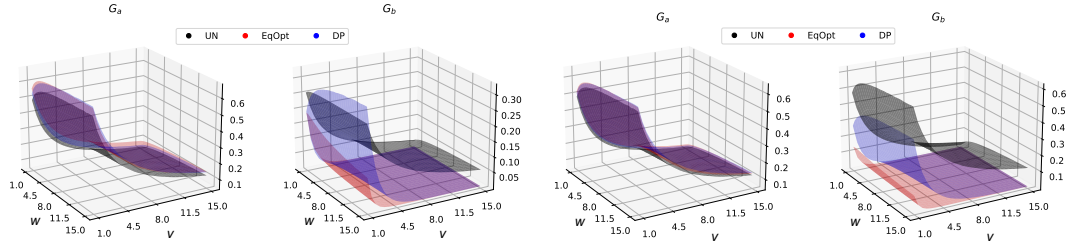
□




(a) $\mathcal{G}_a$ = Asian; $\mathcal{G}_b$ =African-American




(b) $\mathcal{G}_a$ = Asian; $\mathcal{G}_b$ =Hispanic

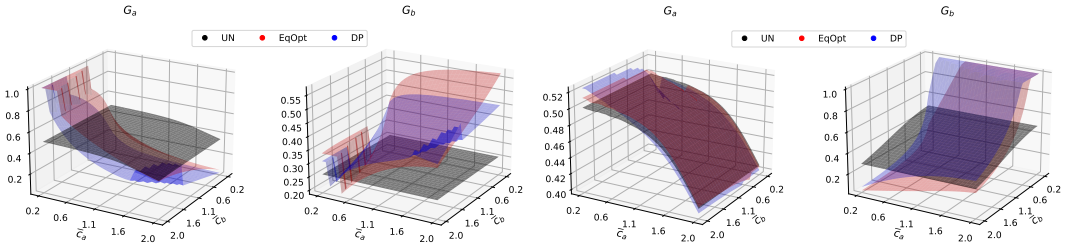



(c) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =African-American




(d) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Hispanic

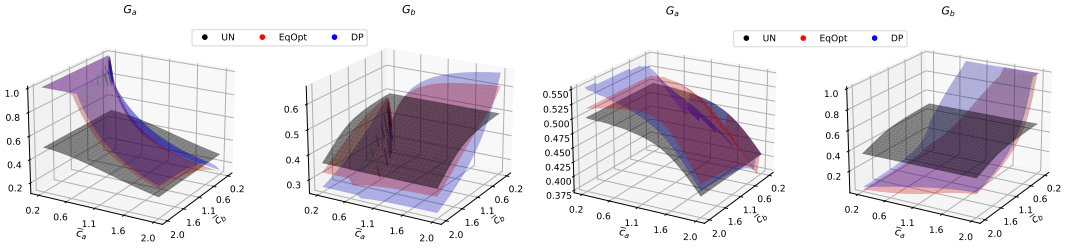
*Figure 17.* Manipulation probabilities under strategic (fair) policy: $C_a = C_b \sim \text{Beta}(v, w)$, $v, w \in [1, 15]$.




(a) $\mathcal{G}_a$ =Hispanic; $\mathcal{G}_b$ =African-American



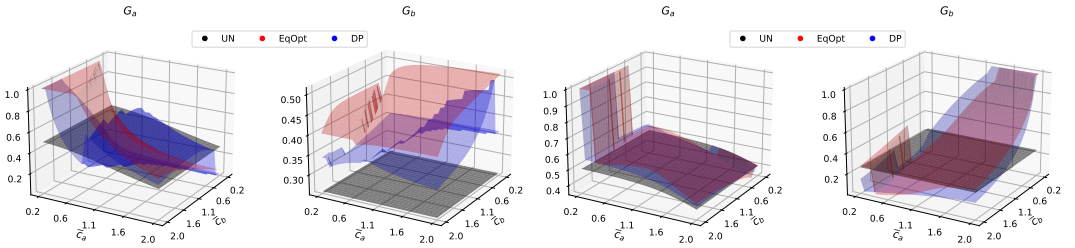

(b) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Asian




(c) $\mathcal{G}_a$ =Asian; $\mathcal{G}_b$ =Hispanic




(d) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Hispanic




(e) $\mathcal{G}_a$ =Asian; $\mathcal{G}_b$ =African-American




(f) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =African-American

*Figure 18.* Manipulation probabilities under strategic (fair) policy: $C_s \sim U[0, \bar{c}_s]$, $s = a, b$, $\bar{c}_a, \bar{c}_b \in [0.2, 2]$.

**Proof of Lemma 4.2.**

*Proof.* Let $\pi_s(x) = \mathbf{1}(x \geq \theta)$, then $\widehat{U}_s(\pi_s) := \widehat{U}_s(\theta)$ can be written as

$$\begin{aligned}
\widehat{U}_s(\theta) &= u_+\alpha_s(1 - \mathbb{F}_{X|YS}(\theta|1,s)) - u_-(1-\alpha_s)(1 - \mathbb{F}_{X|YS}(\theta|0,s)) \\
&= u_+\alpha_s - u_-(1-\alpha_s) + u_-(1-\alpha_s)\mathbb{F}_{X|YS}(\theta|0,s) - u_+\alpha_s\mathbb{F}_{X|YS}(\theta|1,s) \\
\frac{\partial \widehat{U}_s(\theta)}{\partial \theta} &= u_-(1-\alpha_s)P_{X|YS}(\theta|0,s) - u_+\alpha_s P_{X|YS}(\theta|1,s)
\end{aligned}$$

Under Assumption 4.1, $\widehat{U}_s(\theta)$ increases over $\theta \leq \widehat{\theta}_s^{\text{UN}}$ and decreases over $\theta \geq \widehat{\theta}_s^{\text{UN}}$. $\widehat{\theta}_s^{\text{UN}}$ is the optimal threshold and is the unique extreme point of $\widehat{U}_s(\theta)$. □

**Manipulation Probability.**

*Claim* F.2. The probability of manipulation under a threshold policy $\pi_s(x) = \mathbf{1}(x \geq \theta)$ is given by $p_s^0(\pi_s) = \mathbb{F}_{C_s}\Big(\mathbb{F}_{X|YS}(\theta|0,s) - \mathbb{F}_{X|YS}(\theta|1,s)\Big)$.

*Proof.* When $\pi_s(x) = \mathbf{1}(x \geq \theta)$ is a threshold policy, we have

$$\begin{aligned}
P_{D|YMS}(1|y,m,s) &= \int_X P_{DX|YMS}(1,x|y,m,s)dx \\
&= \int_X P_{D|XYMS}(1|x,y,m,s)P_{X|YMS}(x|y,m,s)dx \\
&= \int_X \pi_s(x)P_{X|YMS}(x|y,m,s)dx = 1 - \mathbb{F}_{X|YMS}(\theta|y,m,s)
\end{aligned}$$

Therefore,

$$\begin{aligned}
p_s^0(\pi_s) &= \mathbb{F}_{C_s}\Big(P_{D|YMS}(1|0,1,s) - P_{D|YMS}(1|0,0,s)\Big) \\
&= \mathbb{F}_{C_s}\Big(\mathbb{F}_{X|YMS}(\theta|0,0,s) - \mathbb{F}_{X|YMS}(\theta|0,1,s)\Big) \\
&= \mathbb{F}_{C_s}\Big(\mathbb{F}_{X|YS}(\theta|0,s) - \mathbb{F}_{X|YS}(\theta|1,s)\Big).
\end{aligned}$$

□

**Proof of Lemma 4.6.**

*Proof.* Take derivative of $U_s(\theta)$ w.r.t. $\theta$, we have

$$\begin{aligned}
\frac{\partial U_s(\theta)}{\partial \theta} &= \Big(P_{X|YS}(\theta|0,s)\big(u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))\big) + P_{X|YS}(\theta|1,s)\Psi_s'(\Delta_s(\theta))\Big) - u_+\alpha_s P_{X|YS}(\theta|1,s) \\
&\propto \Big(\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}\big(u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))\big) + \Psi_s'(\Delta_s(\theta))\Big) - u_+\alpha_s
\end{aligned}$$

As $\theta \to \pm\infty$, $\Delta_s(\theta) \to 0$, $\Psi_s'(\Delta_s(\theta)) \to 0$ and $\frac{\partial U_s(\theta)}{\partial \theta} \propto u_-(1-\alpha_s)\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)} - u_+\alpha_s$. Therefore, $\frac{\partial U_s(\theta)}{\partial \theta} > 0$ as $\theta \to -\infty$ and $\frac{\partial U_s(\theta)}{\partial \theta} < 0$ as $\theta \to +\infty$.

The strategic optimal threshold $\theta_s^{\text{UN}}$ satisfies

$$\frac{P_{X|YS}(\theta_s^{\text{UN}}|0,s)}{P_{X|YS}(\theta_s^{\text{UN}}|1,s)} = \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}.$$

□

**Proof of Lemma 4.7.**

*Proof.* To satisfy fairness constraint $\mathcal{C}$, $\int_{\theta_a}^{\infty} \mathcal{P}_a^{\mathcal{C}}(x)dx = \int_{\theta_b}^{\infty} \mathcal{P}_b^{\mathcal{C}}(x)dx$ should hold. Denote CDF $\mathbb{F}_s^{\mathcal{C}}(\theta_s) = \int_{-\infty}^{\theta_s} \mathcal{P}_s^{\mathcal{C}}(x)dx$, then for any pair $(\theta_a, \theta_b)$ that is fair, we have $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b) = \eta^{\mathcal{C}}(\theta_b)$ hold for some strictly increasing function $\eta^{\mathcal{C}}(\cdot)$. Denote $u = \mathbb{F}_b^{\mathcal{C}}(\theta_b)$ and $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}(u)$, the following holds:

$$\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}(u)}{du}\frac{du}{d\theta_b} = \frac{1}{(\mathbb{F}_a^{\mathcal{C}})'((\mathbb{F}_a^{\mathcal{C}})^{-1}(\theta_b))}\frac{du}{d\theta_b} = \frac{(\mathbb{F}_b^{\mathcal{C}})'(\theta_b)}{(\mathbb{F}_a^{\mathcal{C}})'(\theta_a)} = \frac{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}{\mathcal{P}_a^{\mathcal{C}}(\theta_a)}$$

The total utility can be written as a function of $\theta_b$, take the derivative of $n_a U_a(\eta^{\mathcal{C}}(\theta_b)) + n_b U_b(\theta_b)$ w.r.t. $\theta_b$, the optimal $\theta_b^{\mathcal{C}}$ satisfies the following,

$$\begin{aligned} & n_a \frac{dU_a(\eta^{\mathcal{C}}(\theta_b))}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} \frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} + n_b \frac{dU_b(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} = 0 \\ \iff & n_a \frac{dU_a(\eta^{\mathcal{C}}(\theta_b))}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} \frac{\mathcal{P}_b^{\mathcal{C}}(\theta_b^{\mathcal{C}})}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b^{\mathcal{C}}))} + n_b \frac{dU_b(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} = 0 \end{aligned}$$

Simplifying above equation gives the result. □

### F.2. Proofs for Section 5

**Proof of Theorem 5.3.**

*Proof.* According to Lemma 4.6, $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ satisfies

$$\frac{P_{X|YS}(\theta_s^{\text{UN}}|0,s)}{P_{X|YS}(\theta_s^{\text{UN}}|1,s)} = \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))} := \Omega_s(\theta_s^{\text{UN}})$$

Under Assumption 4.1, $\Delta_s(\theta)$ is single-peaked with maximum occurring at $x_s^*$. Define function $\Omega_s(\theta) := \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))}$ and denote $\overline{\Psi_s'} := \max_\theta \Psi_s'(\Delta_s(\theta)) = \Psi_s'(\Delta_s(x_s^*))$.

1. If $\alpha_s = \delta_u$, then

$$\frac{\partial U_s(\theta)}{\partial \theta} \propto \Big(\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)} - 1\Big)(u_+\alpha_s - \Psi_s'(\Delta_s(\theta)))$$

   consider two cases:

   - $\overline{\Psi_s'} \le u_-(1-\alpha_s)$
     $\theta_s^{\text{UN}} = \widehat{\theta}_s^{\text{UN}} = x_s^*$ is unique optimal solution.
   - $\overline{\Psi_s'} > u_-(1-\alpha_s)$
     $U_s(\theta)$ has three extreme points where both $\theta_s^{\text{UN}} = \overline{z_s}$, $\theta_s^{\text{UN}} = \underline{z_s}$ are optimal, and $x_s^*$ is the other extreme point that is not optimal.

2. If $\alpha_s < \delta_u$, then consider two cases:

   - $\overline{\Psi_s'} \le u_-(1-\alpha_s)$
     $\Omega_s(\theta)$ decreases over $\theta < x_s^*$ and increases over $\theta > x_s^*$. $\Omega_s(\theta) \to \frac{u_+\alpha_s}{u_-(1-\alpha_s)} < 1$ as $\theta \to \pm\infty$. Under Assumption 4.1, $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}$ intersects with $\Omega_s(\theta)$ at one unique point, i.e., $\theta_s^{\text{UN}}$ is unique and satisfies $\theta_s^{\text{UN}} > \widehat{\theta}_s^{\text{UN}} > x_s^*$.
   - $\overline{\Psi_s'} > u_-(1-\alpha_s)$
     $\Omega_s(\theta)$ decreases from $\frac{u_+\alpha_s}{u_-(1-\alpha_s)}$ to $-\infty$ over $\theta < \underline{z_s}$; increases from $-\infty$ to $\frac{u_+\alpha_s}{u_-(1-\alpha_s)}$ over $\theta > \overline{z_s}$; decreases over $\theta \in (\underline{z_s}, x_s^*)$ and increases over $\theta \in (x_s^*, \overline{z_s})$.

Because $\frac{P_{X|YS}(x_s^*|0,s)}{P_{X|YS}(x_s^*|1,s)} = 1$ and $\Omega_s(x_s^*) = 1 + \frac{u_-(1-\alpha_s)-u_+\alpha_s}{\overline{\Psi'_s}-u_-(1-\alpha_s)} > 1$, under Assumption 4.1, there exists a unique $\theta_s^{\text{UN}} > \widehat{\theta}_s^{\text{UN}} > x_s^*$ at which $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}$ intersects with $\Omega_s(\theta)$, and $\theta_s^{\text{UN}} > \overline{z_s}$.

Moreover, if $\exists\ \theta$ s.t. $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)} > \Omega_s(\theta)$, then $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}$ will also intersects with $\Omega_s(\theta)$ at least two more points over $(\underline{z_s}, x_s^*)$.

Next, we show that among all the extreme points, the one satisfying $\theta_s^{\text{UN}} > x_s^*$ is the optimal.

Re-organize $U_s(\theta)$, we have

$$\arg\max_\theta U_s(\theta) = \arg\max_\theta \underbrace{\Delta_s(\theta)\big(1 - \mathbb{F}_{C_s}(\Delta_s(\theta))\big)}_{:=h_1(\theta)} + \underbrace{\mathbb{F}_{X|YS}(\theta|1,s)\Big(1 - \frac{u_+\alpha_s}{u_-(1-\alpha_s)}\Big)}_{:=h_2(\theta)}$$

For any extreme point $\theta' \in (\underline{z_s}, x_s^*)$, always there exists a point $x' > x_s^*$ satisfying $\Delta_s(x') = \Delta_s(\theta')$, so that $h_1(x') = h_1(\theta')$. Since $x' > \overline{\theta'}$, $h_2(x') > h_2(\theta')$ holds so that $U_s(x') > U_s(\theta')$. In other words, $\exists$ a point over $(x_s^*, \overline{z_s})$ whose utility is higher than those of extreme points in $(\underline{z_s}, x_s^*)$. Since $\theta_s^{\text{UN}}$ is the optimal over $(x_s^*, \overline{z_s})$. It implies that $\theta_s^{\text{UN}}$ is optimal.

3. If $\alpha_s > \delta_u$, then consider two cases:

- $\overline{\Psi'_s} \le u_+\alpha_s$

  $\frac{1}{\Omega_s(\theta)}$ decreases over $\theta < x_s^*$ and increases over $\theta > x_s^*$. $\frac{1}{\Omega_s(\theta)} \to \frac{u_-(1-\alpha_s)}{u_+\alpha_s} < 1$ as $\theta \to \pm\infty$. Under Assumption 4.1, $\frac{P_{X|YS}(\theta|1,s)}{P_{X|YS}(\theta|0,s)}$ intersects with $\frac{1}{\Omega_s(\theta)}$ at one unique point, i.e., $\theta_s^{\text{UN}}$ is unique and satisfies $\theta_s^{\text{UN}} < \widehat{\theta}_s^{\text{UN}} < x_s^*$.
- $\overline{\Psi'_s} > u_+\alpha_s$

  $\frac{1}{\Omega_s(\theta)}$ decreases from $\frac{u_-(1-\alpha_s)}{u_+\alpha_s}$ to $-\infty$ over $\theta < \underline{z_s}$; increases from $-\infty$ to $\frac{u_-(1-\alpha_s)}{u_+\alpha_s}$ over $\theta > \overline{z_s}$; decreases over $\theta \in (\underline{z_s}, x_s^*)$ and increases over $\theta \in (x_s^*, \overline{z_s})$.

  Because $\frac{P_{X|YS}(x_s^*|1,s)}{P_{X|YS}(x_s^*|0,s)} = 1$ and $\frac{1}{\Omega_s(\theta)} = 1 + \frac{u_+\alpha_s - u_-(1-\alpha_s)}{\overline{\Psi'_s}-u_+\alpha_s} > 1$, under Assumption 4.1, there exists a unique $\theta_s^{\text{UN}} < \widehat{\theta}_s^{\text{UN}} < x_s^*$ at which $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}$ intersects with $\Omega_s(\theta)$, and $\theta_s^{\text{UN}} < \underline{z_s}$.

  Moreover, if $\exists\ \theta$ s.t. $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)} < \Omega_s(\theta)$, then $\frac{P_{X|YS}(\theta|0,s)}{P_{X|YS}(\theta|1,s)}$ will also intersect with $\Omega_s(\theta)$ at least two more points over $(x_s^*, \overline{z_s})$.

  We show that among all the extreme points, the one satisfying $\theta_s^{\text{UN}} < x_s^*$ is the optimal.

  For any extreme point $\theta' \in (x_s^*, \overline{z_s})$, always there exists a point $x' < x_s^*$ satisfying $\Delta_s(x') = \Delta_s(\theta')$, so that $h_1(x') = h_1(\theta')$. Since $x' < \theta'$ and $1 < \frac{u_+\alpha_s}{u_-(1-\alpha_s)}$, $h_2(x') > h_2(\theta')$ holds so that $U_s(x') > U_s(\theta')$. In other words, $\exists$ a point over $(\underline{z_s}, x_s^*)$ whose utility is higher than those of extreme points in $(x_s^*, \overline{z_s})$. Since $\theta_s^{\text{UN}}$ is optimal over $(\underline{z_s}, x_s^*)$, it implies that $\theta_s^{\text{UN}}$ is optimal.

□

**Proof of Theorem 5.4.**

*Proof.* WLOG, let $i := a$ and $-i := b$.

Because $\alpha_a > \delta_u > \alpha_b$, according to Thm. 5.3, we have $x_b^* < \widehat{\theta}_b^{\text{UN}} < \theta_b^{\text{UN}}$ and $x_a^* > \widehat{\theta}_a^{\text{UN}} > \theta_a^{\text{UN}}$. It implies that $\mathbb{F}_a^{\mathcal{C}}(x_a^*) > \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) > \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$ and $\mathbb{F}_b^{\mathcal{C}}(x_b^*) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}})$.

Therefore, we have $\mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}})$, so that $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) > \mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}}) > 0$.

□

**Proof of Theorem 5.5.**

*Proof.* WLOG, let $s := i$ and $-i := b$.

By Thm. 5.3, $\theta_a^{\text{UN}} > \widehat{\theta}_a^{\text{UN}}$ always hold. If marginal manipulation gain of $\mathcal{G}_a$ is sufficiently small such that $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\text{UN}})) \to 0$, then $\theta_a^{\text{UN}} \to \widehat{\theta}_a^{\text{UN}}$; If marginal manipulation gain of $\mathcal{G}_a$ is sufficiently large such that $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\text{UN}})) \to u_-(1-\alpha_a)$, then $\theta_a^{\text{UN}} \gg \widehat{\theta}_a^{\text{UN}}$.

For any given $\mathcal{G}_b$, $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) > \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) > \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}})$, since any $\mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}}) \in (\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}), 1)$ is attainable by controlling manipulation cost $C_a$, it implies that there exists $C_a$ s.t. $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})\right| < \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})\right|$ or $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$. □

### F.3. Proofs for Section 6

**Proof of Theorem 6.1.**

*Proof.* WLOG, let $i := a$ and $-i := b$.

1. $\alpha_a < \delta_u < \alpha_b$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$.

   Since $\alpha_a < \delta_u < \alpha_b$, we have $\widehat{\theta}_a^{\text{UN}} > x_a^*$ and $\widehat{\theta}_b^{\text{UN}} < x_b^*$.

   Under Assumption 4.1, $\widehat{U}_s(\theta)$ is non-decreasing over $(-\infty, \widehat{\theta}_s^{\text{UN}})$ and non-increasing over $(\widehat{\theta}_s^{\text{UN}}, +\infty)$. One of the followings must hold: (1) $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ (2) $\widehat{\theta}_a^{\mathcal{C}} < \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} > \widehat{\theta}_b^{\text{UN}}$. Because if $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} > \widehat{\theta}_b^{\text{UN}}$ or $\widehat{\theta}_a^{\mathcal{C}} < \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ holds, we can always find another pair of thresholds satisfying fairness $\mathcal{C}$ but achieves a higher utility $\sum_{s=a,b} n_s \widehat{U}_s(\theta_s)$ so that $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ cannot be non-strategic optimal fair policy.

   Because $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\mathcal{C}}) = \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\mathcal{C}})$, $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}} > x_a^*, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}} < x_b^*$ must hold.

   If $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\mathcal{C}})) > u_-(1-\alpha_a)$ and $\Psi_b'(\Delta_b(\widehat{\theta}_b^{\mathcal{C}})) > u_+\alpha_b$, then we have $\widehat{\theta}_a^{\mathcal{C}} < \overline{z_a}$ and $\widehat{\theta}_b^{\mathcal{C}} > \underline{z_b}$, where $\overline{z_a}, \underline{z_b}$ are defined s.t. $\Psi_a'(\Delta_a(\overline{z_a})) = u_-(1-\alpha_a)$ and $\Psi_b'(\Delta_b(\underline{z_b})) = u_+\alpha_b$. By Thm. 5.3, $U_a(\theta)$ is increasing over $(x_a^*, \overline{z_a})$ and $U_b(\theta)$ is decreasing over $(\underline{z_b}, x_b^*)$. It implies that $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$ and $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$.

2. $\alpha_a, \alpha_b > \delta_u$, $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$, and $\alpha_a \to \delta_u$.

   Since $\alpha_a, \alpha_b > \delta_u$, we have $\widehat{\theta}_a^{\text{UN}} < x_a^*$ and $\widehat{\theta}_b^{\text{UN}} < x_b^*$.

   Because $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\mathcal{C}}) = \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\mathcal{C}})$, $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ must hold.

   If $\alpha_a \to \delta_u$, then $\widehat{\theta}_a^{\text{UN}} \to x_a^*$ and $\widehat{\theta}_a^{\text{UN}} < x_a^* < \widehat{\theta}_a^{\mathcal{C}}$ hold.

   If $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\mathcal{C}})) > u_+\alpha_a$, $\Psi_b'(\Delta_b(\widehat{\theta}_b^{\mathcal{C}})) > u_+\alpha_b$, then we have $\widehat{\theta}_a^{\mathcal{C}} < \overline{z_a}$ and $\widehat{\theta}_b^{\mathcal{C}} > \underline{z_b}$. By Thm. 5.3, $U_b(\theta)$ is decreasing over $(\underline{z_b}, x_b^*)$ implying $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$, and $U_a(\theta)$ may have additional extreme points over $(x_a^*, \overline{z_a})$. Specifically, as $\alpha_a \to \delta_u$, there are two extreme points $x_1, x_2$ with $x_1 \to x_a^*$, $x_2 \to \overline{z_a}$ (by Thm. 5.3), Because $U_a(\theta)$ is increasing over $[x_1, x_2]$, $U_a(x_2) \to U_a(\theta_a^{\text{UN}}) = \max_\theta U_a(\theta)$, and $U_a(x_1) \to U_a(x_a^*), U_a(\widehat{\theta}_a^{\text{UN}}) \to U_a(x_a^*)$, we have $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$.

3. $\alpha_a, \alpha_b < \delta_u$, $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$, and $\alpha_b \to \delta_u$.

   It can be proved similarly as *Case 2* and is omitted.

□

**Proof of Theorems 6.2-6.3.**

*Proof.* For any pair $(\theta_a, \theta_b)$ satisfying fairness $\mathcal{C}$, $\mathbb{F}_a^{\mathcal{C}}(\theta_a) = \mathbb{F}_b^{\mathcal{C}}(\theta_b)$ should hold. We have $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b) = \eta^{\mathcal{C}}(\theta_b)$ for some strictly increasing function $\eta^{\mathcal{C}}(\cdot)$.

1. (Thm. 6.2) At least one of $U_a(\theta)$, $U_b(\theta)$ has multiple extreme points. WLOG, let $i := a$ and $-i := b$.

   - $\alpha_a > \delta_u > \alpha_b$

     (i) $U_a(\theta)$ has multiple extreme points while $U_b(\theta)$ has a unique extreme point.
     Let $x_1, x_2$ be two extreme points over $(x_a^*, \overline{z_a})$ with $x_2$ being the optimal extreme point over $(x_a^*, \overline{z_a})$ and $x_1$ the largest extreme point satisfying $x_1 < x_2$. By Thm. 5.3, $\theta_a^{\text{UN}} < x_a^*$.
     As $n_b \to 1$, $\theta_b^{\mathcal{C}} \to \theta_b^{\text{UN}}$ and $\theta_a^{\mathcal{C}} \to \eta^{\mathcal{C}}(\theta_b^{\text{UN}})$. If $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) \in (x_1, x_2)$ happens to be satisfied under groups' feature distributions and manipulation costs, then it's possible that there exists a sufficiently large $n_b$ such that the a fair

threshold pair $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ results in a higher total utility than that of $(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), \theta_b^{\text{UN}})$. In this case, $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ and $\theta_a^{\mathcal{C}} \in (\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), x_2)$ must hold.
Because $\theta_a^{\text{UN}} < \underline{z_s}$, $\theta_a^{\mathcal{C}} < \overline{z_a}$, we have $\Delta_a(\theta_a^{\text{UN}}) < \Delta_a(\theta_a^{\mathcal{C}})$ and $p_a^{\mathcal{C}} > p_a^{\text{UN}}$.
Because $\alpha_b < \delta_u$, we have $\theta_b^{\text{UN}} > x_b^*$. Since $\theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$, it holds that $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

(ii) $U_a(\theta)$ has a unique extreme point while $U_b(\theta)$ has multiple extreme points.
Similar to the reasoning in (i), let $x_1, x_2$ be two extreme points over $(\underline{z_b}, x_b^*)$ with $x_1$ being the optimal extreme point over $(\underline{z_b}, x_b^*)$ and $x_2$ the smallest extreme point satisfying $x_1 < x_2$.
If $(\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}) \in (x_1, x_2)$ happens to be satisfied, then it's possible to find a sufficiently large $n_a$ such that the fair pair $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ results in a higher utility than that of $(\theta_a^{\text{UN}}, (\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}))$. In this case, $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ and $\theta_b^{\mathcal{C}} \in (x_1, (\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}))$ must hold.
Because $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}} < x_a^*$ and $\theta_b^{\text{UN}} > \overline{z_b}$, $\theta_b^{\mathcal{C}} > \underline{z_b}$, we have $\Delta_a(\theta_a^{\text{UN}}) > \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$. As such, $p_a^{\mathcal{C}} < p_a^{\text{UN}}, p_b^{\mathcal{C}} > p_b^{\text{UN}}$.

(iii) Both $U_a(\theta), U_b(\theta)$ have multiple extreme points.
In this case, $\theta_a^{\text{UN}} < x_a^*$ and all other extreme points of $U_a(\theta)$ fall in $(x_a^*, \overline{z_a})$ with $\underline{z_a} > \theta_a^{\text{UN}}$; $\theta_b^{\text{UN}} > x_b^*$ and all other extreme points of $U_b(\theta)$ fall in $(\underline{z_b}, x_b^*)$ with $\overline{z_b} < \theta_b^{\text{UN}}$.
If $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ happens to be satisfied, then $\theta_b^{\mathcal{C}} \in (\underline{z_b}, x_b^*)$ must hold. It implies that $\Delta_a(\theta_a^{\text{UN}}) > \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$. As such, $p_a^{\mathcal{C}} < p_a^{\text{UN}}, p_b^{\mathcal{C}} > p_b^{\text{UN}}$.
Similarly, if $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ happens to be satisfied, then $\theta_a^{\mathcal{C}} \in (x_a^*, \overline{z_a})$ must hold. It implies that $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

- $\alpha_a, \alpha_b > \delta_u$
  In this case, $\theta_a^{\text{UN}} < x_a^*, \theta_b^{\text{UN}} < x_b^*$ and $U_a(\theta)$ (or $U_b(\theta)$) increases over $\theta < \theta_a^{\text{UN}}$ (or $\theta < \theta_b^{\text{UN}}$). WLOG, let $\mathcal{G}_a$ has multiple extreme points, while $\mathcal{G}_b$ may or may not have multiple extreme points.
  Note that $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ cannot hold, otherwise always there exists a fair threshold pair $(\theta'_a, \theta'_b)$ with $\theta'_a \in (\theta_a^{\mathcal{C}}, \theta_a^{\text{UN}})$ and $\theta'_b \in (\theta_b^{\mathcal{C}}, \theta_b^{\text{UN}})$ whose utility is higher than that of $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$.
  In contrast, $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ may hold. In this case, $\theta_a^{\mathcal{C}} \in (x_a^*, \overline{z_a})$ must hold, while either $\theta_b^{\mathcal{C}} < x_b^{\text{UN}}$ or $\theta_b^{\mathcal{C}} > x_b^{\text{UN}}$ holds.
  Therefore, $\Delta_a(\theta_a^{\text{UN}}) < \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$ (or $\Delta_b(\theta_b^{\text{UN}}) > \Delta_b(\theta_b^{\mathcal{C}})$) must hold so that $p_a^{\mathcal{C}} > p_a^{\text{UN}}, p_b^{\mathcal{C}} > p_b^{\text{UN}}$ (or $p_b^{\mathcal{C}} < p_b^{\text{UN}}$).
  We can prove in a similar way for the case when $\alpha_a, \alpha_b < \delta_u$.

2. (Thm. 6.3) Both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points.

   Prove $\theta_a^{\text{UN}} > \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} < \theta_b^{\mathcal{C}}$ or $\theta_a^{\text{UN}} < \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} > \theta_b^{\mathcal{C}}$ by contradiction. Suppose $\theta_a^{\text{UN}} > \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} > \theta_b^{\mathcal{C}}$, then we can always find another pair of thresholds $(\theta'_a, \theta'_b)$ that satisfies $\mathcal{C}$ with $\theta_a^{\mathcal{C}} < \theta'_a \leq \theta_a^{\text{UN}}$ and $\theta_b^{\mathcal{C}} < \theta'_b \leq \theta_b^{\text{UN}}$. Because $U_s(\theta)$ has unique extreme point and it increases over $\theta < \theta_s^{\text{UN}}$, $U_s(\theta_s^{\mathcal{C}}) < U_s(\theta'_s), \forall s \in \{a, b\}$ holds, i.e., $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ can not be the optimal pair that satisfies the fairness. Similarly, we can show that $\theta_a^{\text{UN}} < \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} < \theta_b^{\mathcal{C}}$ cannot hold.

   Let $x_s^{\text{UN}}$ be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$. Note that $x_s^{\text{UN}}$ is the point at which $p_s^0(x_s^{\text{UN}}) = p_s^0(\theta_s^{\text{UN}})$. WLOG, let $i := a$ and $-i := b$.

   Let $x_a^{\mathcal{C}} := \eta^{\mathcal{C}}(x_b^{\text{UN}})$, i.e., $(x_a^{\mathcal{C}}, x_b^{\text{UN}})$ satisfies fairness constraint $\mathcal{C}$. Given any fixed $\alpha_b$, as $\alpha_a$ changes, $x_a^{\text{UN}}$, $x_a^{\mathcal{C}}$, and $\theta_a^{\text{UN}}$ also change. Rewrite them as functions of $\alpha_a$, i.e., $x_a^{\text{UN}}(\alpha_a)$, $x_a^{\mathcal{C}}(\alpha_a) := \eta^{\mathcal{C}}(x_b^{\text{UN}}; \alpha_a)$, and $\theta_a^{\text{UN}}(\alpha_a)$.

   - $\alpha_a > \delta_u > \alpha_b$
     $x_a^{\text{UN}}(\alpha_a)$ increases in $\alpha_a \in (\delta_u, 1)$

     $$\lim_{\alpha_a \to \delta_u} x_a^{\text{UN}}(\alpha_a) = x_a^*, \quad \lim_{\alpha_a \to 1} x_a^{\text{UN}}(\alpha_a) = +\infty$$

     $\theta_a^{\text{UN}}(\alpha_a)$ decreases in $\alpha_a \in (\delta_u, 1)$

     $$\lim_{\alpha_a \to \delta_u} \theta_a^{\text{UN}}(\alpha_a) = x_a^*, \quad \lim_{\alpha_a \to 1} \theta_a^{\text{UN}}(\alpha_a) = -\infty$$

     $x_a^{\mathcal{C}}(\alpha_a)$ is non-decreasing in $\alpha_a$

     $$\lim_{\alpha_a \to \delta_u} x_a^{\mathcal{C}}(\alpha_a) = \eta^{\mathcal{C}}(x_b^{\text{UN}}; \delta_u) < +\infty, \quad \lim_{\alpha_a \to 1} x_a^{\mathcal{C}}(\alpha_a) = \eta^{\mathcal{C}}(x_b^{\text{UN}}; 1) < +\infty$$

Therefore, $\exists\ \kappa > \delta_u$ s.t. for any $\alpha_a > \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (\theta_a^{\text{UN}}(\alpha_a), x_a^{\text{UN}}(\alpha_a))$.
As $n_a \to 1$, $\theta_a^{\mathcal{C}} \to \theta_a^{\text{UN}}$. Therefore, $\forall \alpha_a \in (\kappa, 1)$, there exists $\tau \in (0,1)$ s.t. $\forall n_a > \tau$, we have $\theta_a^{\mathcal{C}} \in (\theta_a^{\text{UN}}, x_a^{\mathcal{C}})$ and $\theta_b^{\mathcal{C}} < x_b^{\text{UN}}$. It implies that $\Delta_a(\theta_a^{\mathcal{C}}) > \Delta_a(\theta_a^{\text{UN}})$ and $\Delta_b(\theta_b^{\mathcal{C}}) < \Delta_a(\theta_b^{\text{UN}})$ so that $p_a^{\mathcal{C}} > p_a^{\text{UN}}$ and $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

- $\alpha_a, \alpha_b > \delta_u$
  From the above, $\exists \kappa > \delta_u$ s.t. $\forall \alpha_a > \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (\theta_a^{\text{UN}}(\alpha_a), x_a^{\text{UN}}(\alpha_a))$.
  Since $U_a(\theta)$, $U_b(\theta)$ have unique extreme points, neither $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ nor $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ hold. When $\alpha_a > \kappa$, either of the followings holds: (1) $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}$, $\theta_b^{\mathcal{C}} \in (\theta_b^{\text{UN}}, x_b^{\text{UN}})$; (2) $\theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$, $\theta_a^{\mathcal{C}} \in (\theta_a^{\text{UN}}, x_a^{\mathcal{C}})$. It implies $p_b^{\mathcal{C}} > p_b^{\text{UN}}$, $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, or $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.
- $\alpha_a, \alpha_b < \delta_u$
  Prove in the similar way. $\exists \kappa < \delta_u$ s.t. $\forall \alpha_a < \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (x_a^{\text{UN}}(\alpha_a), \theta_a^{\text{UN}}(\alpha_a))$.
  Since $U_a(\theta)$, $U_b(\theta)$ have unique extreme points, either of the followings holds when $\alpha_a < \kappa$: (1) $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}$, $\theta_b^{\mathcal{C}} \in (x_b^{\text{UN}}, \theta_b^{\text{UN}})$; (2) $\theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$, $\theta_a^{\mathcal{C}} \in (x_a^{\mathcal{C}}, \theta_a^{\text{UN}})$. It implies $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, $p_b^{\mathcal{C}} > p_b^{\text{UN}}$, or $p_b^{\mathcal{C}} < p_b^{\text{UN}}$, $p_a^{\mathcal{C}} > p_a^{\text{UN}}$.

□

**Proof of Theorem 6.4.**

*Proof.* First consider case when $\alpha_a, \alpha_b > \delta_u$.

WLOG, let $i := a$ and $-i := b$.

Define function $\eta^{\mathcal{C}}(\cdot) := (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\cdot)$. If $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$, then $\eta^{\mathcal{C}}(x_b^{\text{UN}}) < x_a^*$.

As $\alpha_a \to \delta_u$, $\theta_a^{\text{UN}} \to x_a^*$. As $\alpha_a$ decreases, $\eta^{\mathcal{C}}(x_b^{\text{UN}})$ is non-increasing (constant w.r.t. $\alpha_a$ for `EqOpt` and decreases for `DP`). $\exists\ \kappa > \delta_u$ s.t. when $\alpha_a = \kappa$, $\theta_a^{\text{UN}} = \eta^{\mathcal{C}}(x_b^{\text{UN}})$. Then $\forall \alpha_a < \kappa$, $\eta^{\mathcal{C}}(x_b^{\text{UN}}) < \theta_a^{\text{UN}}$.

As $n_a \to 1$, $\theta_a^{\mathcal{C}} \to \theta_a^{\text{UN}}$ and $\lim_{n_a \to 1} \theta_b^{\mathcal{C}} > x_b^{\text{UN}}$. Therefore, $\exists \tau \in (0,1)$ s.t. for any $n_a > \tau$, we have $\theta_a^{\mathcal{C}} \in (\eta^{\mathcal{C}}(x_b^{\text{UN}}), \theta_a^{\text{UN}})$ and $\theta_b^{\mathcal{C}} > x_b^{\text{UN}}$. It implies that $p_a^{\mathcal{C}} < p_a^{\text{UN}}, p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

For the case when $\alpha_a, \alpha_b < \delta_u$, it can be proved in a similar way and is omitted. □

### F.4. Proofs for Appendix D

**Proof of Proposition D.1.**

*Proof.* WLOG, let $i := a$, $-i := b$.

Since $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, denote $\Delta(\cdot) = \Delta_a(\cdot) = \Delta_b(\cdot)$.

By Lemma 4.6, for $s \in \{a,b\}$, $\widehat{\theta}_s^{\text{UN}}$ satisfies $\frac{P_{X|YS}(\widehat{\theta}_s^{\text{UN}}|1,s)}{P_{X|YS}(\widehat{\theta}_s^{\text{UN}}|0,s)} = \frac{u_-(1-\alpha_s)}{u_+\alpha_s}$. Since $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, $\alpha_b < \alpha_a < \delta_u$, $\frac{u_-(1-\alpha_b)}{u_+\alpha_b} > \frac{u_-(1-\alpha_a)}{u_+\alpha_a}$. Under Assumption 4.1, we have $\widehat{\theta}_a^{\text{UN}} < \widehat{\theta}_b^{\text{UN}}$.

It implies that $\mathbb{F}_{X|YS}(\widehat{\theta}_a^{\text{UN}}|1,a) < \mathbb{F}_{X|YS}(\widehat{\theta}_b^{\text{UN}}|1,b)$, so that $\mathbb{F}_a^{\texttt{EqOpt}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\texttt{EqOpt}}(\widehat{\theta}_b^{\text{UN}})$.

Note that $\mathbb{F}_s^{\texttt{DP}}(\widehat{\theta}_s^{\text{UN}}) = \alpha_s \mathbb{F}_{X|YS}(\widehat{\theta}_s^{\text{UN}}|1,s) + (1-\alpha_s)\mathbb{F}_{X|YS}(\widehat{\theta}_s^{\text{UN}}|0,s)$. Since $\mathbb{F}_{X|YS}(\widehat{\theta}_a^{\text{UN}}|0,a) < \mathbb{F}_{X|YS}(\widehat{\theta}_b^{\text{UN}}|0,b)$ and $\alpha_b < \alpha_a$, we have $\mathbb{F}_a^{\texttt{DP}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\texttt{DP}}(\widehat{\theta}_b^{\text{UN}})$.

First, we show that the unfairness can be mitigated under some cost random variable $C_a$.

Given $\alpha_b, C_b$, $\theta_b^{\text{UN}}$ is determined and satisfies $\frac{P_{X|YS}(\theta_b^{\text{UN}}|0,b)}{P_{X|YS}(\theta_b^{\text{UN}}|1,b)} = \frac{u_+\alpha_b - \Psi_b'(\Delta(\theta_b^{\text{UN}}))}{u_-(1-\alpha_b) - \Psi_b'(\Delta(\theta_b^{\text{UN}}))}$ (by Lemma 4.6), where $\Delta(\theta) = \mathbb{F}_{X|YS}(\theta|0,b) - \mathbb{F}_{X|YS}(\theta|1,b) = \mathbb{F}_{X|YS}(\theta|0,a) - \mathbb{F}_{X|YS}(\theta|1,a)$.

Given any $\alpha_a \in (\alpha_b, \delta_u)$, if $\mathcal{G}_a$'s cost $C_a$ satisfies $\frac{u_+\alpha_a - \Psi_a'(\Delta(\theta_b^{\text{UN}}))}{u_-(1-\alpha_a) - \Psi_a'(\Delta(\theta_b^{\text{UN}}))} = \frac{u_+\alpha_b - \Psi_b'(\Delta(\theta_b^{\text{UN}}))}{u_-(1-\alpha_b) - \Psi_b'(\Delta(\theta_b^{\text{UN}}))}$, i.e.,

$$\Psi_a'(\Delta(\theta_b^{\text{UN}})) = \underbrace{\frac{u_-(1-\alpha_a) - u_+\alpha_a}{u_-(1-\alpha_b) - u_+\alpha_b}}_{>0 \text{ (since } \alpha_a, \alpha_b < \delta_u)} \cdot \underbrace{\left(\Psi_b'(\Delta(\theta_b^{\text{UN}})) - u_+\alpha_b\right)}_{<0 \text{ (by Thm. 5.3)}} + u_+\alpha_a < u_+\alpha_a < u_-(1-\alpha_a) \tag{4}$$

then $\frac{P_{X|YS}(\theta_b^{\text{UN}}|0,a)}{P_{X|YS}(\theta_b^{\text{UN}}|1,a)} = \frac{u_+\alpha_a - \Psi_a'(\Delta(\theta_b^{\text{UN}}))}{u_-(1-\alpha_a)-\Psi_a'(\Delta(\theta_b^{\text{UN}}))}$ holds and $\theta_a^{\text{UN}} = \theta_b^{\text{UN}}$.

Therefore, $\mathbb{F}_a^{\texttt{EqOpt}}(\theta_a^{\text{UN}}) = \mathbb{F}_{X|YS}(\theta_a^{\text{UN}}|1,a) = \mathbb{F}_{X|YS}(\theta_b^{\text{UN}}|1,b) = \mathbb{F}_b^{\texttt{EqOpt}}(\theta_b^{\text{UN}})$.

Because $\mathbb{F}_{X|YS}(\theta_a^{\text{UN}}|0,a) = \mathbb{F}_{X|YS}(\theta_b^{\text{UN}}|0,b)$ also holds,

$$\begin{aligned}
\left|\mathbb{F}_a^{\texttt{DP}}(\theta_a^{\text{UN}}) - \mathbb{F}_b^{\texttt{DP}}(\theta_b^{\text{UN}})\right| &= (\alpha_a - \alpha_b)\Delta(\theta_b^{\text{UN}}) \\
\left|\mathbb{F}_a^{\texttt{DP}}(\widehat{\theta}_a^{\text{UN}}) - \mathbb{F}_b^{\texttt{DP}}(\widehat{\theta}_b^{\text{UN}})\right| &= (\alpha_a - \alpha_b)\Delta(\widehat{\theta}_b^{\text{UN}}) + \alpha_a(\mathbb{F}_{X|YS}(\widehat{\theta}_b^{\text{UN}}|1,b) - \mathbb{F}_{X|YS}(\widehat{\theta}_a^{\text{UN}}|1,a)) \\
&\quad +(1-\alpha_a)(\mathbb{F}_{X|YS}(\widehat{\theta}_b^{\text{UN}}|0,b) - \mathbb{F}_{X|YS}(\widehat{\theta}_a^{\text{UN}}|0,a)) \\
&> (\alpha_a - \alpha_b)\Delta(\widehat{\theta}_b^{\text{UN}})
\end{aligned}$$

Since $\theta_b^{\text{UN}} > \widehat{\theta}_b^{\text{UN}} > x_b^*$ (by Thm. 5.3), $\Delta(\widehat{\theta}_b^{\text{UN}}) > \Delta(\theta_b^{\text{UN}})$.

Therefore, $\left|\mathbb{F}_a^{\texttt{DP}}(\theta_a^{\text{UN}}) - \mathbb{F}_b^{\texttt{DP}}(\theta_b^{\text{UN}})\right| < \left|\mathbb{F}_a^{\texttt{DP}}(\widehat{\theta}_a^{\text{UN}}) - \mathbb{F}_b^{\texttt{DP}}(\widehat{\theta}_b^{\text{UN}})\right|$.

Next, we show that the disadvantaged group can be flipped under some cost random variable $C_a$.

Given any $\alpha_a \in (\alpha_b, \delta_u)$, let $\left(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), \theta_b^{\text{UN}}\right)$ be a pair of thresholds satisfying fairness $\mathcal{C}$, then if $\Psi_a'(\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))) \geq u_-(1-\alpha_a) = \Psi_a'(\Delta(\overline{z_a}))$, we have $\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}})) \geq \Delta(\overline{z_a})$ implying $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) \leq \overline{z_a}$. Since $\theta_a^{\text{UN}} > \overline{z_a}$, $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) < \theta_a^{\text{UN}}$ must hold.

Therefore, $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) = \mathbb{F}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b^{\text{UN}})) < \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$.

Lastly, we show that cost $C_a$ mentioned above always exists.

Since $\Psi_a'(z) = u_-(1-\alpha_a)\big(\mathbb{F}_{C_a}(z) + zf_a(z)\big)$, condition $\Psi_a'(\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))) \geq u_-(1-\alpha_a)$ is equivalent to $\mathbb{F}_{C_a}(z) + zf_a(z) \geq 1$ with $z = \Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))$, which is attainable. Similarly, the condition in Eqn. (4) is equivalent to $\mathbb{F}_{C_a}(z) + zf_a(z) = c$ for some $c < 1$ with $z = \Delta(\theta_b^{\text{UN}})$, which is also attainable.

□

**Proof of Proposition D.2.**

*Proof.* Consider the case when $\alpha_a, \alpha_b > \delta_u$. WLOG, let $i := a$, $-i := b$.

1. $\mathcal{C} = \texttt{EqOpt}$: $\mathcal{P}_s^{\texttt{EqOpt}}(x) = P_{X|YS}(x|1,s)$

   Because $X|Y = y, S = s, y = \{0,1\}, s = \{a,b\}$ have the same variance $\sigma^2$, and $\mu_a^1 - \mu_a^0 < \mu_b^1 - \mu_b^0$, we have $x_s^* = \frac{\mu_s^1 + \mu_s^0}{2}$ and $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) > \mathbb{F}_b^{\texttt{EqOpt}}(x_b^*)$.

   When $\alpha_b > \delta_u$, we have $\theta_b^{\text{UN}} < x_b^*$ and $x_b^{\text{UN}} > x_b^*$. As $\alpha_b$ increases, $x_b^{\text{UN}}$ and $\mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$ increase; as $\alpha_b \to \delta_u$, $x_b^{\text{UN}} \to x_b^*$. Therefore, $\exists \omega > \delta_u$ s.t. when $\alpha_b = \omega$, the consequent $x_b^{\text{UN}}$ satisfies $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) = \mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$. For any $\alpha_b < \omega$, $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) > \mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$ holds.

2. $\mathcal{C} = \texttt{DP}$: $\mathcal{P}_s^{\texttt{DP}}(x) = P_{X|S}(x|s) = \alpha_s P_{X|YS}(x|1,s) + (1-\alpha_s)P_{X|YS}(x|0,s)$.

   Since $\mathbb{F}_{X|YS}(x|1,s) < \mathbb{F}_{X|YS}(x|0,s), \forall x$, as $\alpha_a$ increases, $\mathbb{F}_a^{\texttt{DP}}(x_a^*)$ decreases.

   Because $X|Y = y, S = s, y = \{0,1\}, s = \{a,b\}$ have the same variance $\sigma^2$, we have $\frac{\mathbb{F}_{X|YS}(x_a^*|1,a) - \mathbb{F}_{X|YS}(x_b^*|1,b)}{\mathbb{F}_{X|YS}(x_b^*|0,b) - \mathbb{F}_{X|YS}(x_a^*|0,a)} = 1$.

   If $\frac{u_+}{u_-} < 1$, $\frac{u_+}{u_-} < \frac{\mathbb{F}_{X|YS}(x_a^*|1,a) - \mathbb{F}_{X|YS}(x_b^*|1,b)}{\mathbb{F}_{X|YS}(x_b^*|0,b) - \mathbb{F}_{X|YS}(x_a^*|0,a)}$, which implies that $\delta_u \mathbb{F}_{X|YS}(x_a^*|1,a) + (1-\delta_u)\mathbb{F}_{X|YS}(x_a^*|0,a) > \delta_u \mathbb{F}_{X|YS}(x_b^*|1,b) + (1-\delta_u)\mathbb{F}_{X|YS}(x_b^*|0,b)$, i.e., $\mathbb{F}_a^{\texttt{DP}}(x_a^*) > \mathbb{F}_b^{\texttt{DP}}(x_b^*)$ when $\alpha_a = \alpha_b = \delta_u$.

   As $\alpha_b \to \delta_u$, $x_b^{\text{UN}} \to x_b^*$. As such, there exist $\omega_1, \omega_2 > \delta_u$ such that $\forall \alpha_b < \omega_1$ and $\forall \alpha_a < \omega_2$, we have $\mathbb{F}_a^{\texttt{DP}}(x_a^*) > \mathbb{F}_b^{\texttt{DP}}(x_b^{\text{UN}})$.

The case when $\alpha_a, \alpha_b < \delta_u$ can be proved similarly and is omitted. □

**Proof of Proposition D.3.**

*Proof.* WLOG, let $i := a$ and $-i := b$. Let $x_s^{\text{UN}}$ be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$,

Since $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, $x_a^* = x_b^*$ holds. If $U_s(\theta)$ has multiple extreme points, then according to Thm. 5.3, all extreme points fall between $x_s^{\texttt{UN}}$ and $\theta_s^{\texttt{UN}}$.

Since $\alpha_a > \delta_u > \alpha_b$, $U_a(\theta)$ is increasing over $(-\infty, \theta_a^{\texttt{UN}})$ and decreasing over $(x_a^{\texttt{UN}}, +\infty)$, while $U_b(\theta)$ is increasing over $(-\infty, x_b^{\texttt{UN}})$ and decreasing over $(\theta_b^{\texttt{UN}}, +\infty)$.

- $\mathcal{C} = \texttt{EqOpt}$

  Since $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, $\theta_a^{\texttt{EqOpt}} = \theta_b^{\texttt{EqOpt}}$. To disincentivize under EqOpt fairness, one of the following four possibilities must hold: (1) $\theta_a^{\texttt{EqOpt}} > x_a^{\texttt{UN}}$, $\theta_b^{\texttt{EqOpt}} < x_b^{\texttt{UN}}$ (2) $\theta_a^{\texttt{EqOpt}} < \theta_a^{\texttt{UN}}$, $\theta_b^{\texttt{EqOpt}} > \theta_b^{\texttt{UN}}$ (3) $\theta_a^{\texttt{EqOpt}} < \theta_a^{\texttt{UN}}$, $\theta_b^{\texttt{EqOpt}} < x_b^{\texttt{UN}}$ (4) $\theta_a^{\texttt{EqOpt}} > x_a^{\texttt{UN}}$, $\theta_b^{\texttt{EqOpt}} > \theta_b^{\texttt{UN}}$.

  Note that (3) and (4) never hold.

  Suppose (3) (resp. (4)) holds, then always $\exists(\theta'_a, \theta'_b)$ satisfying EqOpt with $\theta'_a > \theta_a^{\texttt{EqOpt}}, \theta'_b > \theta_b^{\texttt{EqOpt}}$ (resp. $\theta'_a < \theta_a^{\texttt{EqOpt}}, \theta'_b < \theta_b^{\texttt{EqOpt}}$) s.t. $(\theta'_a, \theta'_b)$ attains a higher utility. In other words, $(\theta_a^{\texttt{EqOpt}}, \theta_b^{\texttt{EqOpt}})$ cannot be optimal fair policies. It concludes that (3) and (4) cannot hold.

  Note that (1) and (2) cannot be satisfied, because $x_b^{\texttt{UN}} < x_b^* = x_a^* < x_a^{\texttt{UN}}$, $\theta_b^{\texttt{UN}} > x_b^* = x_a^* > \theta_a^{\texttt{UN}}$, and $\theta_a^{\texttt{EqOpt}} = \theta_b^{\texttt{EqOpt}}$ must hold.

  Therefore, none of four cases can be satisfied. EqOpt cannot disincentivize both groups.

- $\mathcal{C} = \texttt{DP}$

  To disincentivize under DP fairness, one of the following four possibilities must hold: (1) $\theta_a^{\texttt{DP}} > x_a^{\texttt{UN}}$, $\theta_b^{\texttt{DP}} < x_b^{\texttt{UN}}$ (2) $\theta_a^{\texttt{DP}} < \theta_a^{\texttt{UN}}$, $\theta_b^{\texttt{DP}} > \theta_b^{\texttt{UN}}$ (3) $\theta_a^{\texttt{DP}} < \theta_a^{\texttt{UN}}$, $\theta_b^{\texttt{DP}} < x_b^{\texttt{UN}}$ (4) $\theta_a^{\texttt{DP}} > x_a^{\texttt{UN}}$, $\theta_b^{\texttt{DP}} > \theta_b^{\texttt{UN}}$.

  Similar as the case when $\mathcal{C} = \texttt{EqOpt}$, (3) and (4) never hold.

  Note that in order to satisfy DP, it is impossible for (2) to hold. Because $\alpha_a > \alpha_b$ and $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, $\theta_b^{\texttt{DP}} < \theta_a^{\texttt{DP}}$ must hold under DP. Moreover, $\theta_a^{\texttt{UN}} < x_a^* = x_b^* < \theta_b^{\texttt{UN}}$. Therefore, (2) never hold.

  However, (1) is likely to be satisfied.

  When $U_a(\theta)$, $U_b(\theta)$ have unique extreme point.

  Re-write $x_s^{\texttt{UN}}$ as a function of $\alpha_s$: $x_s^{\texttt{UN}}(\alpha_s)$, take derivative of $\mathbb{F}_b^{\texttt{DP}}(x_s^{\texttt{UN}}(\alpha_s))$ w.r.t. $\alpha_s$, we have

$$\frac{d\mathbb{F}_s^{\texttt{DP}}(x_s^{\texttt{UN}}(\alpha_s))}{d\alpha_s} = \underbrace{\mathbb{F}_{X|YS}(x_s^{\texttt{UN}}(\alpha_s)|1,s) - \mathbb{F}_{X|YS}(x_s^{\texttt{UN}}(\alpha_s)|0,s)}_{\textbf{term 1} = -\Delta_s(x_s^{\texttt{UN}}(\alpha_s))} + \underbrace{P_{X|S}(x_s^{\texttt{UN}}(\alpha_s)|s) \cdot \frac{dx_s^{\texttt{UN}}(\alpha_s)}{d\alpha_s}}_{\textbf{term 2}}$$

  Note that $\lim_{\alpha_a \to 1} \mathbb{F}_a^{\texttt{UN}}(x_a^{\texttt{UN}}(\alpha_a)) = \mathbb{F}_a^{\texttt{UN}}(+\infty) = 1$, $\lim_{\alpha_b \to 0} \mathbb{F}_b^{\texttt{UN}}(x_b^{\texttt{UN}}(\alpha_b)) = \mathbb{F}_b^{\texttt{UN}}(-\infty) = 0$,
  $\lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\texttt{UN}}(x_a^{\texttt{UN}}(\alpha_a)) = \delta_u \mathbb{F}_{X|YS}(x_a^*|1,a) + (1-\delta_u)\mathbb{F}_{X|YS}(x_a^*|0,a)$,
  $\lim_{\alpha_b \to \delta_u} \mathbb{F}_b^{\texttt{UN}}(x_b^{\texttt{UN}}(\alpha_b)) = \delta_u \mathbb{F}_{X|YS}(x_b^*|1,b) + (1-\delta_u)\mathbb{F}_{X|YS}(x_b^*|0,b)$.
  Since $x_a^* = x_b^*$, $\lim_{\alpha_b \to \delta_u} \mathbb{F}_b^{\texttt{UN}}(x_b^{\texttt{UN}}(\alpha_b)) = \lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\texttt{UN}}(x_a^{\texttt{UN}}(\alpha_a))$.
  If $\Delta_b(x_b^*) > P_{X|S}(x_b^*|b) \cdot \frac{dx_b^{\texttt{UN}}(\alpha_b)}{d\alpha_b}\Big|_{\alpha_b = \delta_u}$ (for a special case where $X|Y=y, S=s$ is Gaussian distributed, it can be satisfied if $X|Y=1, S=s$ and $X|Y=0, S=s$ are sufficiently separable),
  then $\frac{d\mathbb{F}_b^{\texttt{DP}}(x_b^{\texttt{UN}}(\alpha_b))}{d\alpha_b}\Big|_{\alpha_b = \delta_u} < 0$, and $\exists \mathcal{I} \subset (0, \delta_u)$ such that $\forall \alpha_b \in \mathcal{I}$, we have $\mathbb{F}_b^{\texttt{DP}}(x_b^{\texttt{UN}}(\alpha_b)) > \lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\texttt{UN}}(x_a^{\texttt{UN}}(\alpha_a))$
  Therefore, $\exists(\alpha_a, \alpha_b)$ with $\alpha_a \to \delta_u$ and $\alpha_b \in \mathcal{I}$ s.t. $\mathbb{F}_b^{\texttt{DP}}(x_b^{\texttt{UN}}(\alpha_b)) > \mathbb{F}_a^{\texttt{UN}}(x_a^{\texttt{UN}}(\alpha_a))$.
  In this case, if $n_a$ is sufficiently large, we have $\theta_a^{\texttt{DP}} > x_a^{\texttt{UN}}$ and $\theta_b^{\texttt{DP}} < x_b^{\texttt{UN}}$.

□