A Related Work

A.1 Strategic classification

Generally, strategic behaviors can cause feature and label distribution of individuals to shift, which have long been closely related to concept drift (Lu et al., 2018), preference shift (Carroll et al., 2022), and algorithm recourse (Karimi et al., 2022). Strategic classification has been extensively studied since (Hardt et al., 2016a) formally modeled the interaction between individuals and a decision maker as a Stackelberg Game, and proposed a framework for strategic classification. While taking the individuals' best response into account, the decision maker can make the optimal decision by anticipating strategic manipulation. During recent years, more complex models on strategic classification have been proposed (Ben-Porat and Tennenholtz, 2017; Dong et al., 2018; Braverman and Garg, 2020; Jagadeesan et al., 2021; Izzo et al., 2021; Ahmadi et al., 2021; Tang et al., 2021; Zhang et al., 2020a; 2022; Eilat et al., 2022; Liu et al., 2022; Lechner and Urner, 2022; Chen et al., 2020a). Ben-Porat and Tennenholtz (2017) developed a best response linear regression predictor where two players compete and each gets a payoff depending on the proportion of the points he/she predicts more accurately than the other player. Dong et al. (2018) focused on the online version of the strategic classification algorithm. Chen et al. (2020a) developed a strategic-aware linear classifier to minimize the Stacelberg regret. Braverman and Garg (2020) modified the classifier to random ones. Moreover, Jagadeesan et al. (2021) added noise to standard strategic classification and modified the *standard microfoundations* into *alternative microfoundations* to let a portion of individuals be irrational and not have perfect knowledge about the decision maker's policy. Tang et al. (2021) considered the setting where the decision maker only knew a subset of individuals' actions. Levanon and Rosenfeld (2022) generalized strategic classification to situations where individuals and the decision maker have aligned interests. Perdomo et al. (2020); Izzo et al. (2021); Hardt et al. (2022) proposed and elaborated the concept of *performative prediction* where predictive decisions can influence the outcomes to predict. Izzo et al. (2021) proposed an algorithm *performative gradient descent* to compute performative optimal points. Hardt et al. (2022) defined *performative power* as a measure of how much a decision can change the population. This framework can formulate more general strategic classification settings. Liu et al. (2022) studied the situation where competitions between individuals are present in strategic classification. (Eilat et al., 2022) relaxed the assumption that individual best responses are independent of each other and proposed a robust learning framework based on a Graph Neural Network. Lechner and Urner (2022) proposed a novel loss function considering both the accuracy of the prediction rule and its vulnerability to strategic manipulation.

A.2 Improvement with a label change

Another line of research takes improvement into account(Liu et al., 2019; Zhang et al., 2020a; Liu et al., 2020; Rosenfeld et al., 2020; Chen et al., 2020b; Haghtalab et al., 2020; Kleinberg and Raghavan, 2020; Alon et al., 2020; Miller et al., 2020; Shavit et al., 2020; Bechavod et al., 2021; Jin et al., 2022; Barsotti et al., 2022; Ahmadi et al., 2022a; Raab and Liu, 2021; Heidari et al., 2019). Liu et al. (2019; 2020); Zhang et al. (2020a); Rosenfeld et al. (2020); Ahmadi et al. (2022b) studied the conditions under which individuals will choose to improve their qualifications. Specifically, Liu et al. (2019) investigated how different decision rules (e.g. maxutil, fair) influence population qualification. Liu et al. (2020) modeled the improvement cost as a random variable and further pointed out that a subsidizing mechanism for individual costs can be beneficial for improving behaviors. Zhang et al. (2020a) studied the dynamic of population qualification under a partially observed Markov decision problem setting, where improvement probability is given as a parameter. Rosenfeld et al. (2020) proposed a *Look-ahead regularization* to directly penalize the drop of population qualification. Ahmadi et al. (2022b) proposed a *common improvement capacity model* and a *individualized improvement capacity model* to optimize social welfare and fairness while considering individual improvement.

There are other studies considering both strategic manipulation and improvement (Chen et al., 2020b; Haghtalab et al., 2020; Kleinberg and Raghavan, 2020; Alon et al., 2020; Miller et al., 2020; Shavit et al., 2020; Bechavod et al., 2021; Jin et al., 2022; Barsotti et al., 2022; Ahmadi et al., 2022a; Harris et al., 2022; Horowitz and Rosenfeld, 2023; Yan et al., 2023). Besides the works which have been mentioned in Sec. 1, (Barsotti et al., 2022) modeled strategic manipulation and improvement similarly with costs that differ within constant factors. The paper also did simulations where manipulation and

improvement were present. (Ahmadi et al., 2022a) considered a general discrete model and a linear model where improvement and manipulation are both possible.

A.3 MACHINE LEARNING FAIRNESS

While machine learning algorithms are able to achieve high accuracy in different tasks, they are likely to be unfair to individuals from different ethnic groups. To measure the fairness of algorithms, various metrics have been proposed including *demographic parity* (Feldman et al., 2015), *equal opportunity* (Hardt et al., 2016b), *equalized odds* (Hardt et al., 2016b) and *equal resource* (Gupta et al., 2019).

More importantly, several works have studied how strategic behaviors impact fairness (Liu et al., 2019; Zhang et al., 2020a; Liu et al., 2020; Zhang et al., 2022). Specifically, Liu et al. (2019) considered one-step feedback where static fairness does not promote dynamic fairness. Zhang et al. (2020a) analyzed the long-term impact of static fairness metrics based on dynamics of population qualification. Liu et al. (2020) studied how heterogeneity across groups and the lack of realizability can destroy long-term fairness in strategic classification. Zhang et al. (2022) has proposed a probabilistic model to demonstrate strategic manipulation as well as the fairness impacts of strategic behaviors, where the individuals shift their feature distribution instead of directly changing their features. The work also assumed randomness in manipulation cost. Meanwhile, it explored influences on different fairness metrics when strategic manipulation is present(Barocas et al., 2019; Hardt et al., 2016b).

B ADDITIONAL DISCUSSIONS

B.1 THE COMPARISON BETWEEN OUR MODEL AND CAUSAL STRATEGIC LEARNING

Previous works in *causal strategic learning* model every strategic classification problem as a *structural causal model* (SCM). SCM is a graphic model depicting the causal relationships between different features and the label, where features can be classified as causal or non-causal after a causal discovery process (Miller et al., 2020). strategic manipulation means intervening in the non-causal nodes and improvement corresponds to intervening in the causal nodes. Though the model takes both behaviors into account and can accommodate complex causal structures, it has the following weaknesses: (i) The individuals can intervene in any feature node arbitrarily with a deterministic outcome to any value once their budgets permit, which is not practical as illustrated in 1; (ii) In most real-world cases, individuals are not able to intervene the observable features directly. Instead, they intervene in other unobserved features (causal or non-causal) to change the observable features. So it is sometimes meaningless to distinguish whether an observable feature is causal or non-causal, because the root causes of its value change may be diverse.

We illustrate (ii) more clearly in Fig. 5, a causal graph where U, V are unobserved. However, U is non-causal and V is causal. It is easy to see only X is observable and correlated to Y, but its change can be either "causal" or "non-causal" with respect to Y.

By contrast, our probabilistic framework does not classify X as causal or non-causal. It models both manipulation and improvement as imitating qualified individuals and incorporates the randomness of outcomes and costs. With limited control over their features, individuals can only expect a distribution shift and may even fail when they take certain actions. We believe the concise yet effective design of our model is more suitable for many practical situations nowadays, while the causal strategic models sometimes assign too much power to individuals.

B.2 MORE PRACTICAL EXAMPLES FITTING TO OUR MODEL

In Sec. 1 and Appendix B.1, we already explain the motivation of our model in detail. Here we provide more motivating examples besides *college admission*:

1. *Loan application*:
    (a) Manipulation: an unqualified applicant may "steal" the features from qualified ones by purchasing a social security card (SSN) from the hackers. The "stolen" features are still random when the applicant decides to purchase an SSN because the card is often randomly drawn from many stolen cards of qualified individuals.

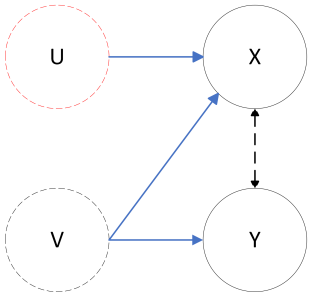

Figure 5: An example causal graph where only X is observable and U, V are unobserved.

(b) Improvement: an unqualified applicant may observe the qualified individuals' profiles and strive to imitate their behaviors. However, the applicant never knows the realization of his/her features before trying to improve. The applicant can only try their best to mimic qualified individuals and expects the successful imitation will cause his/her feature distribution to shift.

2. *Job application*:

(a) Manipulation: an unqualified applicant may "steal" the features from qualified ones by hiring an imposter to take the interview instead of him/her (especially when remote interviews are prevalent today). Similar to previous examples, the applicant does not know the exact feature realization when making the decision to manipulate.

(b) Improvement: an unqualified applicant may still observe the features of qualified ones by reading their interview preparation tips or looking at their technical portfolios. Then they may try hard to imitate the qualified individuals. Similar to previous examples, the applicant still has no idea of the exact outcome when he/she decides to improve.

B.3 The option of taking no action

The comprehensive probabilistic model can be easily extended to the setting where "manipulate", "improve" and "do nothing" are all possible. Specifically, denoting the expected utility of doing nothing as $U_N(\theta)$. We know $U_M(\theta), U_I(\theta)$ do not change, while $U_N(\theta) = 0$. Thus, we can derive manipulation probability $P_M(\theta) = Pr(U_M(\theta) > U_I(\theta)\ and\ U_M(\theta) > 0)$ and $P_I(\theta) = Pr(U_I(\theta) > U_M(\theta)\ and\ U_I(\theta) > 0)$.

With $P_M(\theta), P_I(\theta)$, we can write out the new strategic utility and study its property if necessary. However, in reality, it is more reasonable for individuals to always take an action. For instance, applicants feeling they are not qualified will at least take some measures to improve their chances to be admitted. Thus, the model in the main paper disallows "taking no action".

B.4 Discussion on adjusted preferences

Utility loss from the adjusted preferences. Although adjusting preferences is a simple yet effective way to promote fairness and disincentivize manipulation, the actual utility received by the decision-maker inevitably diminishes as k_1 or k_2 changes (as the actual utility the decision-maker receives is always determined by the original function $U(\theta)$ in Eq. equation 3). Nonetheless, such diminished utilities may still be higher than the utility under non-strategic policy $\widehat{\theta}^*$. This is illustrated empirically in Sec. 5 and Appendix C.

Adjusted preference as a regularizer to promote fairness. We have shown that adjusting weights k_1, k_2, k_3 in learning objective (Eq. equation 5) can control the individual behavior and algorithmic fairness. Indeed, we can view this adjustment mechanism as a *regularization* method: by adjusting weights, we are essentially changing the objective $U(\theta)$ by adding a regularizer, i.e.,

$$\widehat{U}(\theta) + \Phi(\theta, k_1, k_2, k_3) = U(\theta) + \underbrace{\Delta\Phi(\theta, k_1, k_2, k_3)}_{\textbf{regularizer}}$$

with the regularizer $\Delta\Phi(\theta, k_1, k_2, k_3)$ defined as follows:

$$\Phi(\theta, k_1, k_2, k_3) - \Phi(\theta, u(1-\alpha), u(1-\alpha), u(1-\alpha))$$

Weights k_1, k_2, k_3 are the regularization parameters. The analysis in Sec. 4.2 and 4.3 suggests that to learn optimal policies that satisfy certain constraints such as bounded fairness violation and/or bounded individual's manipulation, we may transform this *constrained* optimization into a *regularized unconstrained* optimization. This view, by incorporating fairness and strategic classification in a simple unified framework, may provide insights for researchers from both communities.

B.5 Estimate Model Parameters

A complete estimation procedure. With only the knowledge of conditional distribution of qualified individuals $P_{X|Y}(x|1)$ and the population's qualification rate α, we introduce a complete procedure to estimate $P_{X|Y}(x|0), q, P^I, \epsilon, P_{C_M - C_I}(x)$ sequentially. Specifically, we need to do controlled intervention experiments on an experimental population as follows.

1. Estimate $P_{X|Y}(x|0)$: Set the lowest decision threshold $\theta = 0$ to estimate $P_{X|Y}(x|0)$. Since all unqualified individuals will be accepted, the resulting distribution is the original mixture distribution $(1-\alpha) \cdot P_{X|Y}(x|0) + \alpha \cdot P_{X|Y}(x|1)$. Thus, with minor assumptions on the feature distribution families, we can estimate $P_{X|Y}(x|0)$.
2. Estimate q: Apply the strictest auditing procedures (e.g., audit everyone in [26]) to the population to disable manipulation. With manipulation disabled and arbitrary decision threshold θ applied, all unqualified people choose to improve, and the resulting qualification rate is $(1-\alpha)q + \alpha$. Thus, by examining the qualification rate after the intervention we can get the estimation of q.
3. Estimate P^I: Apply an arbitrary decision threshold θ to the population, the resulting population probability density distribution will be a mixture of $(1-\alpha)(1-q)P^I + [(1-\alpha)q + \alpha]P_{X|1}$. Similarly, with minor assumptions on the distribution family of P^I, we can estimate P^I.
4. Estimate ϵ: With q, P^I known, the decision-maker can first apply another arbitrary θ to new samples from the population and observe the resulting new population. This gives the new qualification rate α_p. Because $\alpha_p = \alpha + (1-\alpha)(1 - P_M(\theta)q)$ where $P_M(\theta)$ is the probability of manipulation under θ, we can then compute the value of $P_M(\theta)$. Note that the decision-maker also knows how many individuals (among all individuals) are discovered to manipulate (cheat), and let this proportion be ϵ_c, then we can estimate the manipulation detection probability ϵ as $\frac{\epsilon_c}{P_M(\theta)}$.
5. Finally, with all previous parameters known, we can apply different θ to the population several times to obtain data points of P_M. Then since P_M corresponds to points of $F_{C_M - C_I}$, with minor assumptions on the distribution family of P_M, we can directly fit the distribution and get $P_{C_M - C_I}$.

   It is worth noting that all the above steps can be more robust by doing multiple intervention experiments. And we also note that according to(Miller et al., 2020), to learn parameters in Strategic Classification, controlled experiments with intervention are necessary and cannot be further simplified. We will add the above discussion to the paper to improve its significance.

Robustness of results when q, ϵ are noisy.

We also present an experiment to relax the assumption that the decision-maker knows q, ϵ exactly on FICO data. Instead, they only know $q + \delta$ or $\epsilon + \delta$ where δ is a Gaussian noise. We do 100 rounds of simulations and produce plots with expectation and error bars similar to Fig. 4(Fig. 6 and Table 3 show the results with noisy q, while Fig. 7 and Table 4 show the results with noisy ϵ). The results show adjusting k still works under noisy q and ϵ although inconsistency exists.

Table 3: Comparison between three types of optimal thresholds (FICO data) when there is a Gaussian noise on q with standard deviation 0.1 and $k_{1,c} = k_{1,aa} = 1.25$. For utility and P_M, the left value in parenthesis is for Group a, while the right is for Group b. The fairness metric is *eqopt*.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(0.698, 0.171)$	$(0.331, 0.513)$	0.136
Original average **noisy** strategic	$(0.703, 0.201)$	$(0.212, 0.284)$	0.057
Adjusted average **noisy** strategic	$(0.700, 0.192)$	$(0.170, 0.220)$	0.043

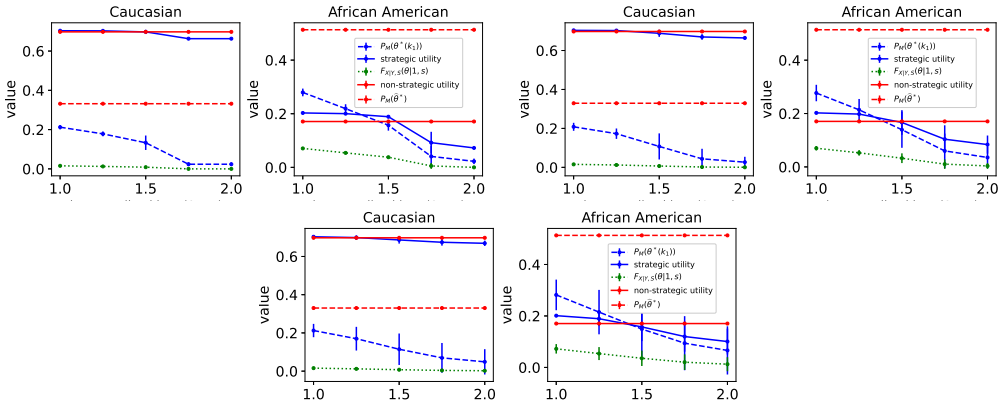

Figure 6: Impact of adjusted preferences (FICO data) when there is a Gaussian noise on q. The noises have 0 mean, and $0.05, 0.1, 0.15$ standard deviation from the left two plots to the right two plots.

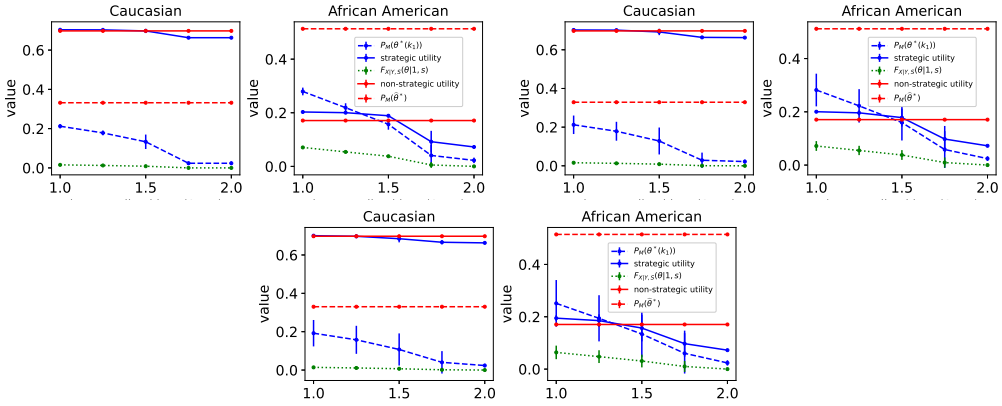

Figure 7: Impact of adjusted preferences (FICO data) when there is a Gaussian noise on ϵ. The noises have 0 mean, and $0.05, 0.1$ standard deviation from the left two plots to the right two plots.

Table 4: Comparison between three types of optimal thresholds (FICO data) when there is a Gaussian noise on ϵ with standard deviation 0.1 and other settings stay the same.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(0.698, 0.171)$	$(0.331, 0.513)$	0.136
Original average **noisy** strategic	$(0.700, 0.195)$	$(0.192, 0.251)$	0.050
Adjusted average **noisy** strategic	$(0.698, 0.185)$	$(0.158, 0.194)$	0.037

C Additional Empirical Results

C.1 Additional Results on FICO Score

Firstly, Fig. 9 shows the conditional distribution $P_{X|YS}$ and P^I of each ethnic group. Fig. 8 demonstrates Assumption 2.1 is satisfied. Fig. 23 shows the (non)-strategic optimal thresholds under different combinations of q, ϵ for each ethnic group. All four plots demonstrate the correctness of Thm. 4.1.

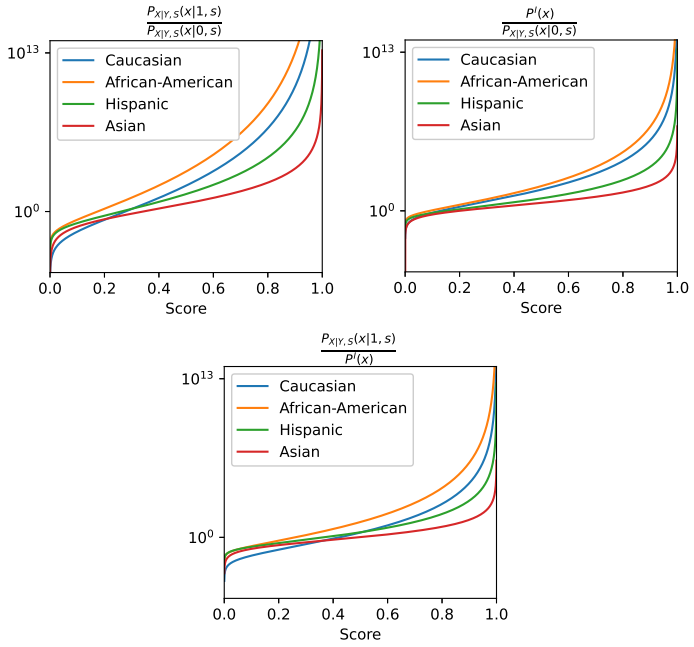

Figure 8: Illustration of Assumption 2.1 on FICO Data

Table 5: Comparison between three types of optimal thresholds for FICO data. For utility and P_M, the left value in parenthesis is for Asian, while the right is for Hispanic. The fairness metric is *eqopt*.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(0.726, 0.427)$	$(0.115, 0.322)$	0.089
Original strategic	$(0.734, 0.448)$	$(0.055, 0.161)$	0.047
Adjusted strategic	$(0.726, 0.434)$	$(0.023, 0.070)$	0.022

Moreover, besides the illustration of Thm. 4.5 and Thm. 4.6 using Caucasian and African American data. We also demonstrate the same results hold for Asian and Hispanic as in Fig. 10.

The scenario considered in Figure 12 satisfies the condition *1.(ii)* in Thm. 4.5, because the original strategic optimal threshold $\theta_s^* < \theta_{max}$ for both groups. We further conduct experiments in this setting to evaluate the impacts of adjusting preferences. We consider equal opportunity (`EqOpt`) as the fairness metric, under which $\mathbb{E}_{X\sim P_s^c}[\mathbf{1}(X \geqslant \widehat{\theta})] = F_{X|YS}(\theta|1,s)$ and the unfairness measure can be reduced to $\left|F_{X|YS}(\theta|1,a) - F_{X|YS}(\theta|1,b)\right|$.

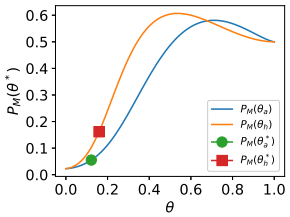

Figure 12: Manipulation probability $P_M(\theta)$ of Asian and Hispanic

The results are shown in Figure 10, where dashed red and dashed blue curves are manipulation probabilities under non-strategic $\widehat{\theta}^*$ and strategic $\theta^*(k_1)$, respectively. Solid red and solid blue curves are the actual utilities $U(\widehat{\theta}^*)$ and $U(\theta^*(k_1))$ received by the decision-maker. The difference between the two green curves measures the unfairness between Asian and Hispanic. $k_1 = 1$ corresponds to the original decision-maker while others when $k_1 > 1$ indicate the decision-maker with adjusted preferences. Results show that compared to the non-strategic $\widehat{\theta}^*$, the strategic θ^*, by taking into account strategic behavior disincentives the strategic manipulation. When condition *1(ii)* in Thm. 4.5 is satisfied, increasing k_1 can disincentivize manipulation (i.e., P_M decreases) while improving fairness. These validate Thm. 4.5 and 4.6.

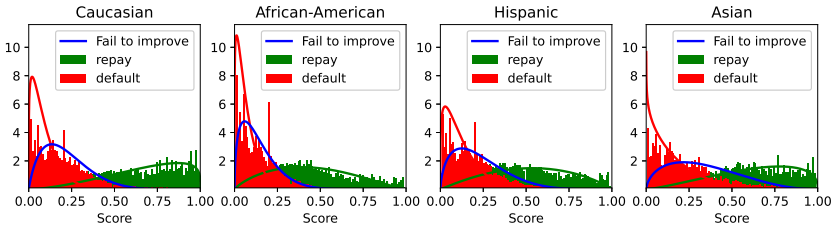

Figure 9: Manipulation curve and manipulation probability for both groups under optimal non-strategic thresholds

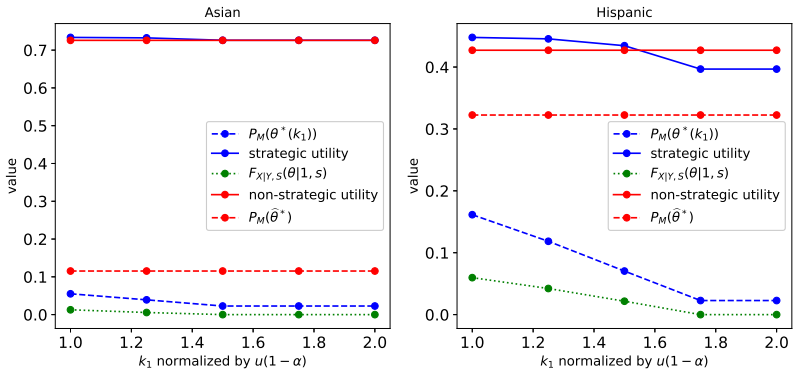

Figure 10: Illustration of Thm. 4.5 and Thm. 4.6 on FICO Data (Asian vs Hispanic)

In Table 5, We summarize the comparison between non-strategic $\widehat{\theta}^*$, original strategic θ^*, and adjusted strategic $\theta^*(k_1)$ (when $k_{1,a} = k_{1,h} = 1.5$). It shows that decision-makers by adjusting preferences can significantly mitigate unfairness and disincentivize manipulation, with only slight decreases in utilities.

Experiments with demographic parity as new fairness metric

We also reconducted the above experiments with demographic parity (`DP`) as the new fairness metric. As illustrated in Sec. 4.3, $P_s^{\text{DP}}(x) = P_{X|S}(x|s)$. Similar to Fig. 4 and the bottom plot of Fig. 10, we produce Fig. 13 based on `DP`, which demonstrate the same patterns as the figures based on `Eqopt`.

C.2 Results for Gaussian Data

Assume there are two groups For $s \in \{a, b\}$, we both have:

$$\begin{aligned} P_{X|YS}(x|0,s) &\sim N(0,1) \\ P_s^I &\sim N(0.5,1) \\ P_{X|YS}(x|1,s) &\sim N(1,1) \\ C_M - C_I &\sim N(0,0.25) \end{aligned} \tag{6}$$

We first illustrate the conditional feature distributions for Gaussian data in Fig. 11. With these parameters pre-determined, we still need to vary α, ϵ, q to obtain $\widehat{\theta}^*, \theta^*$ under different parameter combinations.

(Non)-strategic optimal threshold and utility

To illustrate the complex nature under different permutations of parameters, with the pre-determined parameters in equation 6 and $\alpha = 0.6$, we vary q and plot both non-strategic optimal thresholds and regular strategic ones with respect to different ϵ as shown in the bottom plot of Fig. 14, where the lower graphs illustrate Thm. 4.1, i.e. the red line is always under the blue line.

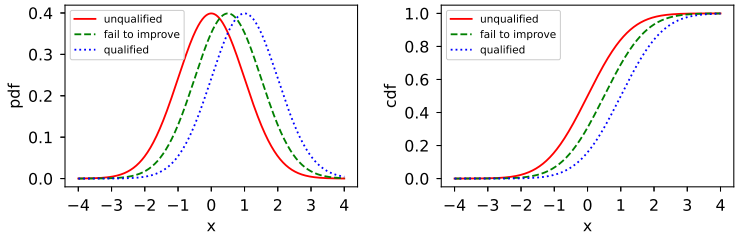

Figure 11: Illustration of equation 6

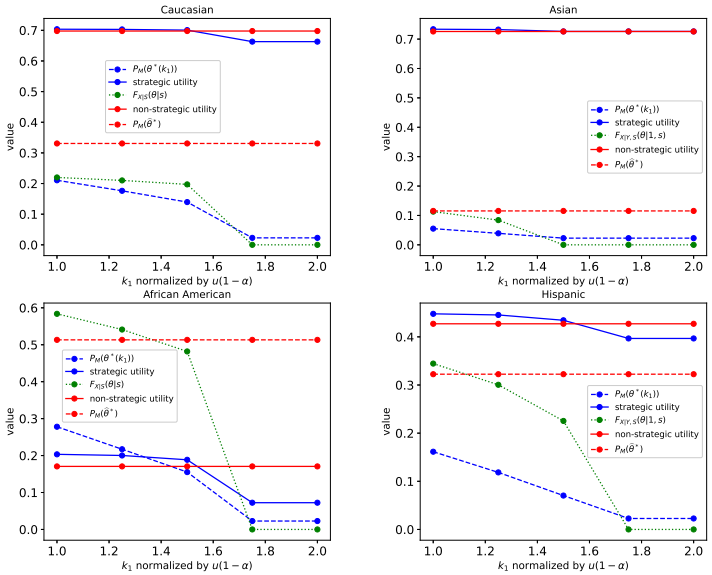

Figure 13: Illustration of Thm. 4.5 and Thm. 4.6 in FICO Data with fairness metric DP. Left figure is for Caucasian and African American, while the right is for Asian and Hispanic

We also demonstrate the strategic utility under different combinations of q, ϵ with pre-determined parameters in equation 6 and $\alpha = 0.3$ or $\alpha = 0.6$. Fig. 15 and 16 suggest the complicated nature of regular strategic utility under different parameter combinations. It is possible to have 0,1 or 2 extreme points.

Illustration of threshold shifts while adjusting k

To illustrate 4.5, we demonstrate the effects of adjusting each of k_1, k_2, k_3. According to Fig. 17 and 18, we can see when k_1 is large enough, the optimal strategic threshold is definitely lower than the optimal non-strategic ones. However, when α is small, we need larger k_1 to pull θ^* downward. According to Fig. 19 and 20, we can see when the population is majority qualified, adjusting k_2 is not guaranteed to shift θ^* upward (Fig. 20).

Illustration of condition 1.(i), Thm. 4.5

We first show a parameter setting satisfying condition *1.(i)* in Thm. 4.5. With pre-determined parameters in equation 6 , we set $q = \epsilon = 0.5$ and $\alpha_a = 0.2, \alpha_b = 0.25$. This matches the notation tradition in Sec. 4.3 where group a is the disadvantaged group with a lower qualified percentage. Also, because $q + \epsilon \geqslant 1$, the setting satisfies condition *1.(i)* in Thm. 4.5. We first illustrate the manipulation probability under optimal original strategic threshold θ_s^* as in Fig. 11. From Fig. 22, we can set $k_{1a} = k_{2a} = 1.25$ to let the strategic utility still be larger than the one under non-strategic optimal threshold(i.e. the solid blue line is above the solid red line), while lower the cumulative

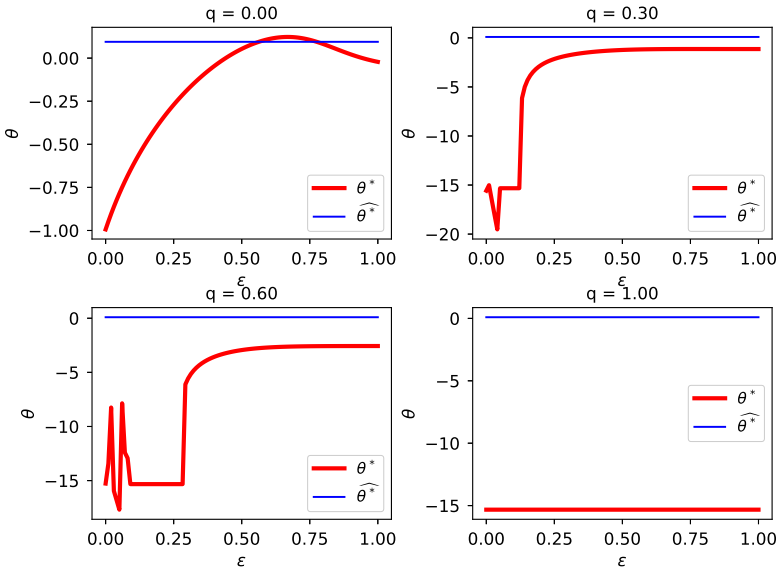

Figure 14: (Non)-strategic optimal threshold

density dramatically (i.e. the dotted green line) to admit more qualified individuals and disincentivize manipulation (i.e. the dashed blue line). The details of comparisons are shown in Table 7.

Illustration of condition 1.(ii), Thm. 4.5

With pre-determined parameters in equation 6 , we set $q = \epsilon = 0.25$ and $\alpha_a = 0.4, \alpha_b = 0.6$. This matches the notation tradition in Sec. 4.3 where group a is the disadvantaged group with a lower qualified percentage. We first illustrate the manipulation probability under optimal original strategic threshold θ_s^* as in Fig. 21. Fig. 21 reveals that *1.(ii)* in 4.5 is satisfied because the orange and green points are both located before the extreme large point of $P_M(\theta)$. Thus, we could increase k_{1s} to disincentivize manipulation while improving fairness as shown in Fig. 22. From Fig. 22, we can set $k_{1a} = k_{2a} = 1.25$ to let the strategic utility still be larger than the one under non-strategic optimal threshold(i.e. the solid blue line is above the solid red line), while lower the cumulative density dramatically (i.e. the dotted green line) to admit more qualified individuals and disincentivize manipulation (i.e. the dashed blue line). In Table 6, We summarize the comparison between non-strategic $\widehat{\theta}^*$, original strategic θ^*, and adjusted strategic $\theta^*(k_1)$ (when $k_{1,c} = k_{1,aa} = 1.25$). It shows that decision-makers by adjusting preferences can significantly mitigate unfairness and disincentivize manipulation, with only slight decreases in utilities.

Illustration of condition 2, Thm. 4.5

Besides, we also show one more parameter setting satisfying condition 2 in Thm. 4.5. With pre-determined parameters in equation 6 , we also set $q = \epsilon = 0.2$ and $\alpha_a = 0.3, \alpha_b = 0.35$. This matches the notation tradition in Sec. 4.3 where group a is the disadvantaged group with a lower qualified percentage. Also, based on Fig. 21, $q + \epsilon < 1$ and $\alpha_a, \alpha_b < 0.5$, the setting satisfies condition 2 in Thm. 4.5. We first illustrates the manipulation probability under optimal original strategic threshold θ_s^* and non-strategic threshold $\widehat{\theta_s^*}$ as in Fig. 21. As shown in Fig. 22, for both groups, we demonstrate the manipulation probability for $\widehat{\theta}^*, \theta^*$ and $\theta(k_1)$ when k_1 varies, (non)-strategic utility and cumulative density conditioned on $Y = 1$ (i.e. this measures the unfairness based on *equal opportunity*). This plot suggests we can find suitable k_{2a} and k_{2b} to disincentivize manipulation and promote fairness, while also making the utility higher than the one under non-strategic optimal threshold. From Fig. 22, we can set both k_{2a} and k_{2b} at 1.25 to let the strategic utility still be larger than the utility under non-strategic optimal threshold (i.e. the solid blue line is above the solid red line), while keeping the cumulative density function closer (i.e. the green dotted line) to mitigate unfairness, and also disincentivize manipulation (i.e. the blue dashed line). The details of comparisons are shown in Table 8.

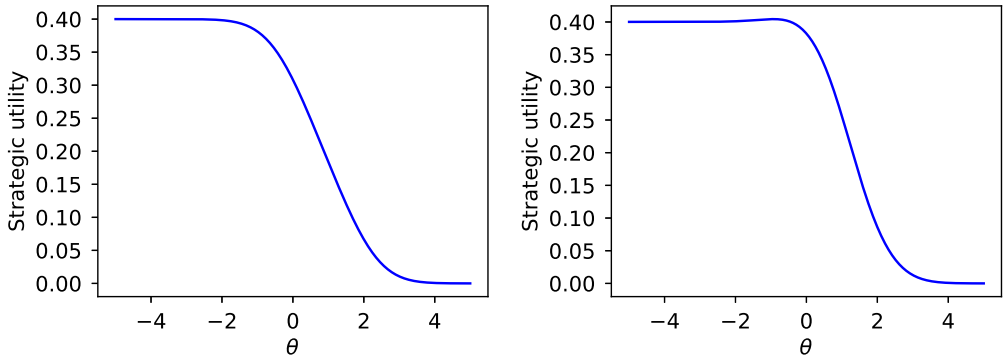

Figure 15: Regular strategic utility when $\alpha = 0.6$. The left figure has $\epsilon = 0, q = 0.5$ and the right has $\epsilon = 0.75, q = 0.25$

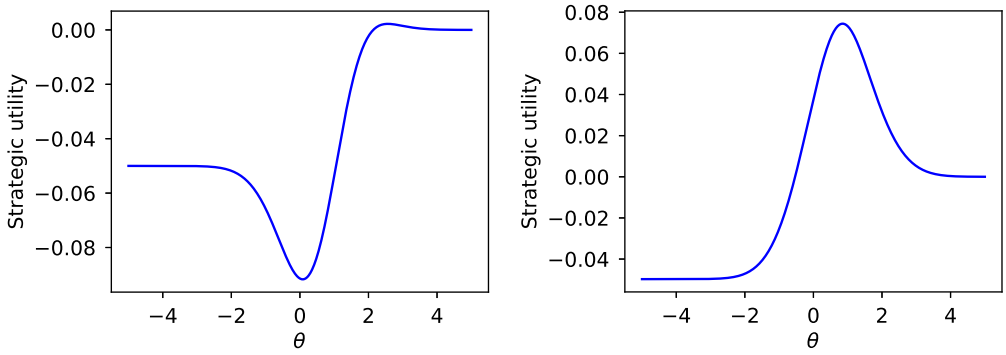

Figure 16: Regular strategic utility when $\alpha = 0.3$. The left figure has $\epsilon = 0, q = 0.5$ and the right has $\epsilon = 0.75, q = 0.25$

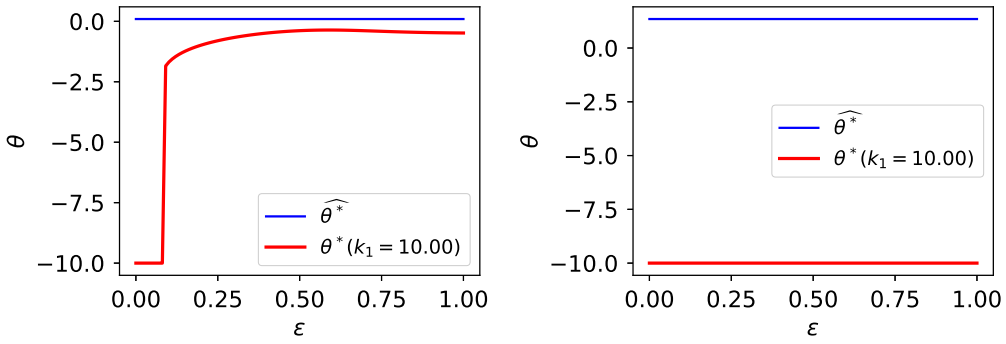

Figure 17: Strategic optimal threshold $\theta^*(k_1)$ after increasing k_1 while keeping k_2, k_3 fixed. Left figure has $q = 0.01$ and right figure has $q = 0.99$, while both figures have $\alpha = 0.6$

Table 6: Comparison between three types of optimal thresholds for Gaussian data satisfying condition *1.(i)*. For utility and P_M, the left value in parenthesis is for Group a, while the right is for Group b. The fairness metric is *eqopt*.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(0.054, 0.327)$	$(0.519, 0.368)$	0.280
Original strategic	$(0.081, 0.384)$	$(0.266, 0.168)$	0.073
Adjusted strategic	$(0.088, 0.385)$	$(0.176, 0.159)$	0.008

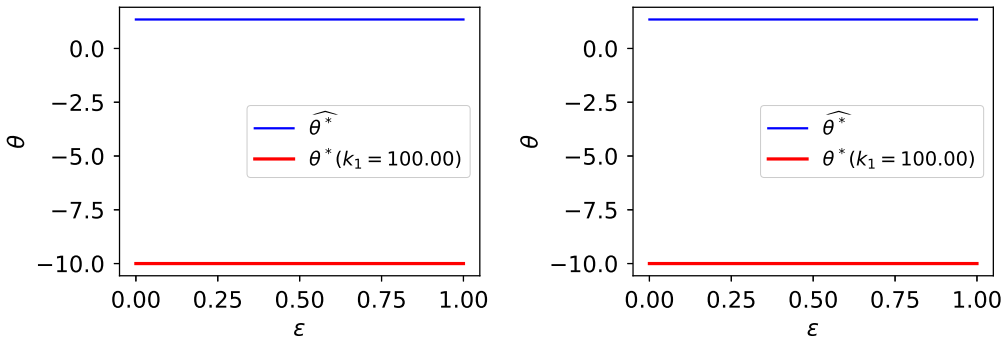

Figure 18: Strategic optimal threshold $\theta^*(k_1)$ after increasing k_1 while keeping k_2, k_3 fixed. Left figure has $q = 0.01$ and right figure has $q = 0.99$, while both figures have $\alpha = 0.3$

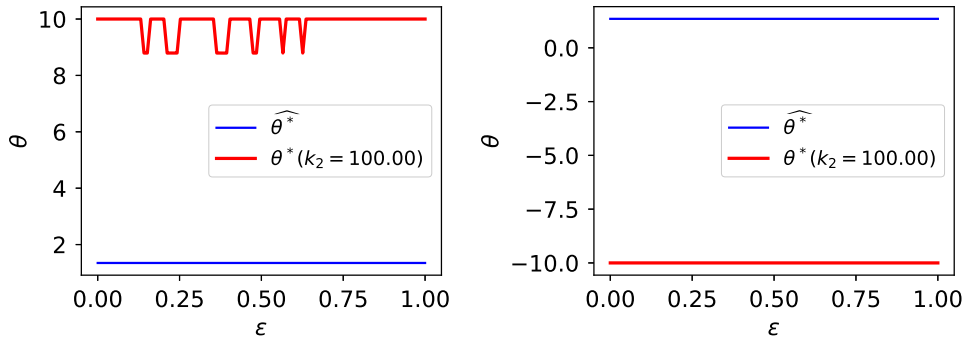

Figure 19: Strategic optimal threshold $\theta^*(k_2)$ after increasing k_2 while keeping k_1, k_3 fixed. Left figure has $q = 0.01$ and right figure has $q = 0.99$, while both figures have $\alpha = 0.3$

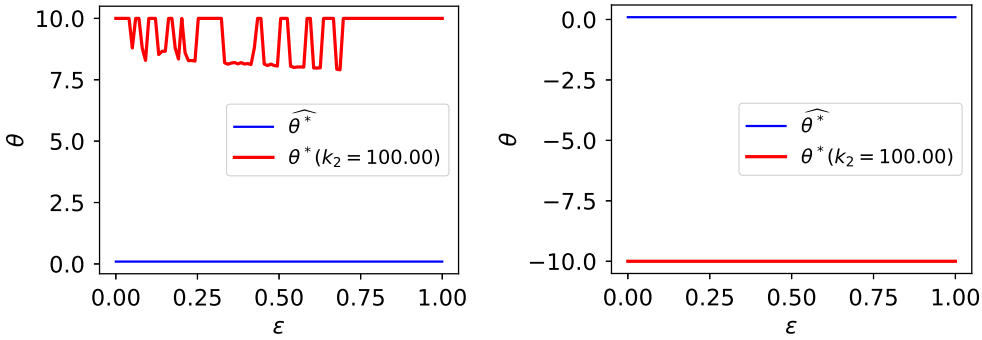

Figure 20: Strategic optimal threshold $\theta^*(k_2)$ after increasing k_2 while keeping k_1, k_3 fixed. Left figure has $q = 0.01$ and right figure has $q = 0.99$, while both figures have $\alpha = 0.6$

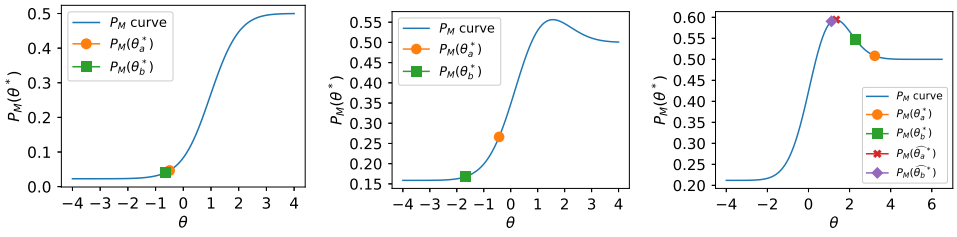

Figure 21: Manipulation probability $P_M(\theta)$: from left to right are plots for condition *1.(i), 1.(ii), 2*

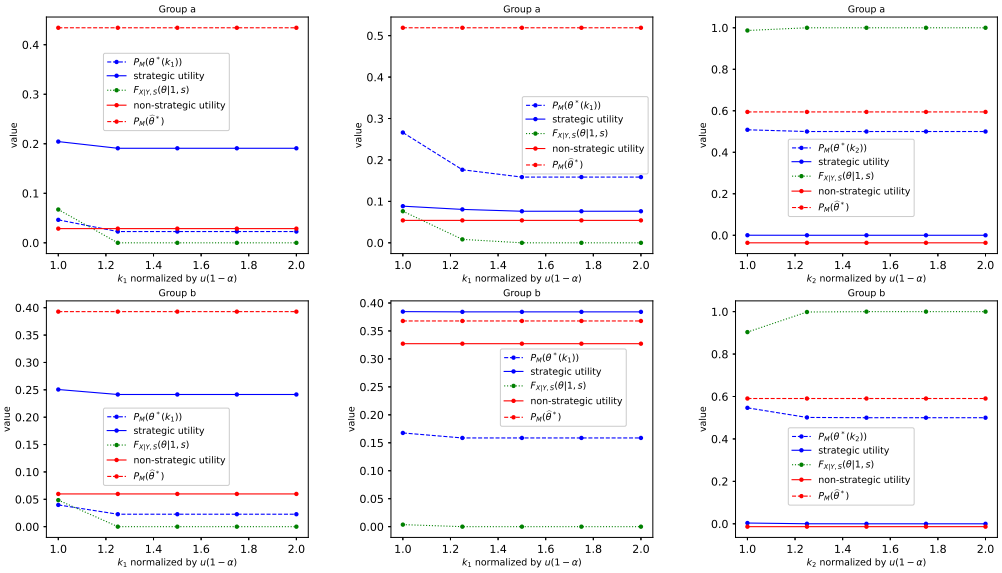

Figure 22: Illustration of Thm. 4.5 and Thm. 4.6. From left to right are illustrations for condition *1.(i),1.(ii),2*

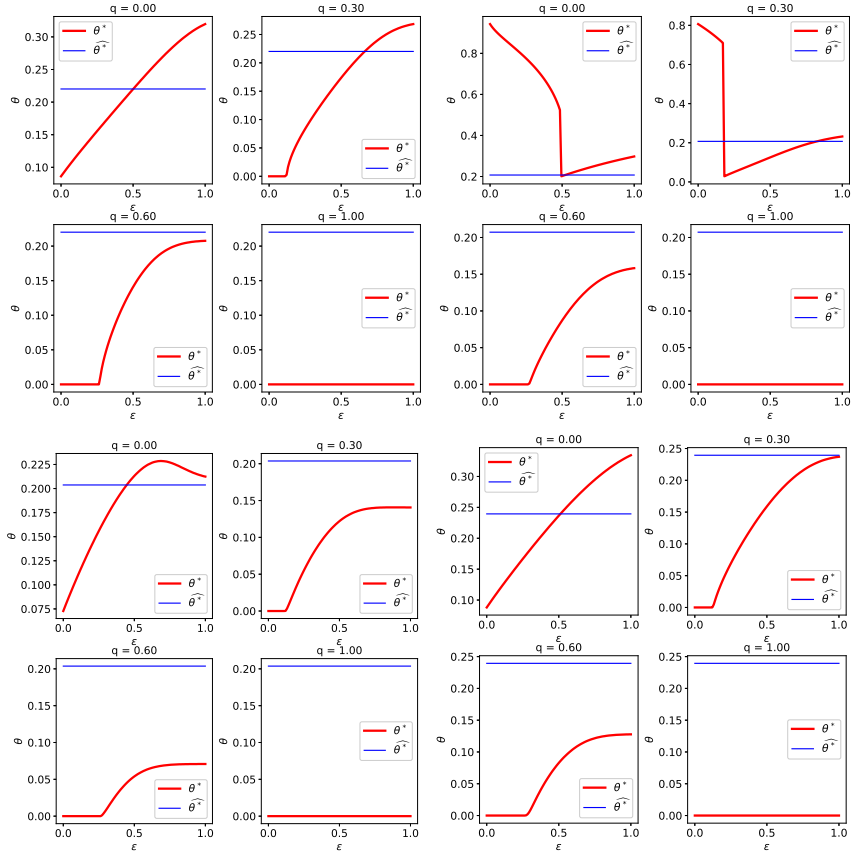

Figure 23: (Non)-strategic optimal thresholds under different q, ϵ for different ethnic groups (top left: Caucasian; top right: African American; bottom left: Asian; bottom right: Hispanic)

Table 7: Comparison between three types of optimal thresholds for Gaussian data satisfying condition *1.(ii)*. For utility and P_M, the left value in parenthesis is for Group a, while the right is for Group b. The fairness metric is *eqopt*.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(0.029, 0.060)$	$(0.434, 0.393)$	0.086
Original strategic	$(0.204, 0.251)$	$(0.046, 0.040)$	0.019
Adjusted strategic	$(0.191, 0.241)$	$(0.023, 0.023)$	0

Table 8: Comparison between three types of optimal thresholds for Gaussian data satisfying condition *2*. For utility and P_M, the left value in parenthesis is for Group a, while the right is for Group b. The fairness metric is *eqopt*.

Threshold category	Utility	P_M	Unfairness
Non-strategic	$(-0.036, -0.014)$	$(0.674, 0.686)$	0.088
Original strategic	$(0.001, 0.004)$	$(0.508, 0.547)$	0.084
Adjusted strategic	$(0, 0)$	$(0.500, 0.500)$	0.002

D DERIVATIONS AND PROOFS

D.1 DERIVATIONS OF EQ. EQUATION 1

$U_M(\theta)$ is the expected utility gain of an unqualified agent if choosing to manipulate: i. If the manipulation is not exposed, the probability of admission is $1 - F_{X|Y}(\theta|1)$ because the manipulation leads the agents to get his/her new feature from $P(X|1)$, which happens at a probability $1-\epsilon$; ii. If the manipulation is exposed, the probability of admission is 0, which happens at a probability ϵ; iii. If the agent does not manipulate, the probability of admission is $1 - F_{X|Y}(\theta|0)$ because now his/her feature is from the unqualified population, and keep in mind that the agents will never know the exact values of his/her feature when he/she makes decisions; Then according to the total probability theorem, the expectation of utility gain $U_M(\theta) = (1-\epsilon)\cdot(1 - F_{X|Y}(\theta|1)) + \epsilon \cdot 0 - (1 - F_{X|Y}(\theta|0)) - C_M$.

$U_I(\theta)$ is the expected utility gain of an unqualified agent if choosing to improve: i. If the improvement succeeds, the probability of admission is $1 - F_{X|Y}(\theta|1)$ because the improvement leads the agents to get his/her new feature from $P(X|1)$, which happens at a probability q; ii. If the manipulation is exposed, the probability of admission is $1 - F^I(\theta)$, which happens at a probability $1 - q$; iii. If the agent does not manipulate, the probability of admission is $1 - F_{X|Y}(\theta|0)$.

Then according to the total probability theorem, we can derive $U_I(\theta)$ as well. Finally, substitute above two terms into $P_M(\theta) = \Pr\left(U_M(\theta) > U_I(\theta)\right)$ and we get Eq. equation 1.

D.2 PROOF OF THM. 2.3

Assumption 2.2 ensures that $P_{C_M - C_I} > 0$ when θ in its domain. Thus, we can directly take the derivative inside equation 1, we can get $(1-q)\cdot P^I(\theta) - (1-q-\epsilon)P_{X|Y}(\theta|1)$. To get its sign, we only need to consider $(1-q) - (1-q-\epsilon)\frac{P_{X|Y}(\theta|1)}{P^I(\theta)}$.

Thus, if $1 - q - \epsilon \leqslant 0$, the derivative is always larger than 0 (since $q < 1$). So under this situation, P_M is always increasing. Otherwise, since $\frac{P_{X|Y}(\theta|1)}{P^I(\theta)}$ is increasing according to Assumption 2.1, it will first increase and then decrease, with $\frac{P_{X|Y}(\theta_{max}|1)}{P^I(\theta_{max})} = \frac{1-q}{1-q-\epsilon}$.

Since $\frac{P_{X|Y}(\theta_{max}|1)}{P^I(\theta_{max})}$ is monotonically increasing and $\frac{1-q}{1-q-\epsilon} = 1 + \frac{\epsilon}{1-q-\epsilon}$, when q increases $1 + \frac{\epsilon}{1-q-\epsilon}$ increases, making θ_{max} increases. The same also holds when ϵ increases. Note that while q or ϵ increases, we still need $q + \epsilon \leqslant 1$.

D.3 PROOF OF THM. 4.1

Assume $\theta \in (a, b)$. When $q \to 1$, improvement will always succeed. Also, Thm. 2.3 reveals $P_M(\theta)$ reaches its minimum when $\theta \to a$, so $P_M(a) < 0.5$. Thus, improvement will always bring a benefit that is larger than manipulation to the strategic decision-maker (since improvement always succeeds). Thus, the decision maker may set a threshold as low as possible ($\to a$) to maximize its utility, which will always be lower than the non-strategic optimal threshold.

D.4 PROOF OF PROP. 4.2

Assume $\theta \in (a, b)$. Consider the situation when k_2, k_3 both stay fixed and $k_1 \to \infty$, $U = \Phi + \hat{U}$ is dominated by $k_1\phi_1$. Noticing ϕ_1 reaches its maximum when $\theta \to a$, we will also have the new optimal $\theta^*(k_1) \to a$. Since a is the minimum possible value of the threshold, the optimal threshold when k_a is large enough will definitely be smaller than the optimal non-strategic threshold as well as the original optimal strategic threshold.

D.5 PROOF OF PROP. 4.3

Assume $\theta \in (a, b)$. Consider the situation when k_1, k_3 both stay fixed and $k_2 \to \infty$, $U = \Phi + \hat{U}$ is dominated by $-k_2\phi_2$. $\phi_2 \to 0$ both when $\theta \to b$ *or* a (i.e. ϕ_2 reaches its minimum). However, the non-strategic utility should be 0 when $\theta \to b$ but smaller than 0 when $\theta \to a$ if not majority of people are qualified. This will make the new optimal $\theta^*(k_2) \to b$. Since b is the maximum possible value of

the threshold, the optimal threshold when k_2 is large enough will definitely be larger than the optimal non-strategic threshold as well as the original optimal strategic threshold.

D.6 PROOF OF PROP. 4.4

Assume $\theta \in (a, b)$. Consider the situation when k_1, k_2 both stay fixed and $k_3 \to b$, $U = \Phi + \hat{U}$ is dominated by $-k_3\phi_3$. Take the derivative of $(1-\epsilon)\cdot(1-F_{X|Y}(\theta|1)) - (1-F_{X|Y}(\theta|0))$ (the term multiplied by P_M in ϕ_3), we get $1-(1-\epsilon)\frac{P_{X|Y}(X|1)}{P_{X|Y}(X|0)}$. This suggests the term will first increase and then decrease. Thus, the maximizer of $-k_3 \cdot \phi_3 = -k_3 \cdot P_M \cdot (1-(1-\epsilon)\frac{P_{X|Y}(X|1)}{P_{X|Y}(X|0)})$ will locate before the root of $(1-\epsilon)\cdot(1-F_{X|Y}(\theta|1)) - (1-F_{X|Y}(\theta|0))$. Then noticing that increasing ϵ will lower the value of the root, we can confirm the existence of $\bar{\epsilon}$ to make the root small enough, thereby making the maximizer of $-k_3 \cdot \phi_3$ smaller enough. Then because U is dominated by $-k_3\phi_3$, $\theta^*(k_3)$ will also be small enough.

D.7 PROOF OF THM. 4.5

Assume $\theta \in (a, b)$.

1. Under condition *1.(i)*, Thm. 2.3 shows $P_M(\theta)$ strictly increases. Because increasing k_1 will cause $\theta^*(k_1)$ to left shift until approaching a, $P_M(\theta^*(k_1))$ will keep decreasing to its minimum value.

2. Under condition *1.(ii)*, Thm. 2.3 shows $P_M(\theta)$ strictly increases before θ_{max}, where $\frac{P_{X|Y}(\theta_{max}|1)}{P^I(\theta_{max})} = \frac{1-q}{1-q-\epsilon}$. Since $\frac{P_{X|Y}(\theta|1)}{P^I(\theta)}$ is increasing, we would know $\theta^* < \theta_{max}$. Because increasing k_1 will cause $\theta^*(k_1)$ to left shift until approaching a, $P_M(\theta^*(k_1))$ will keep decreasing to its minimum value.

3. Under condition *2*, Thm. 2.3 shows $P_M(\theta)$ strictly decreases after θ_{max}, where $\frac{P_{X|Y}(\theta_{max}|1)}{P^I(\theta_{max})} = \frac{1-q}{1-q-\epsilon}$. Since $\frac{P_{X|Y}(\theta|1)}{P^I(\theta)}$ is increasing, we would know $\theta^* > \theta_{max}$. Because increasing k_2 when $\alpha \leqslant 0.5$ will cause $\theta^*(k_2)$ to right shift until approaching a, $P_M(\theta^*(k_2))$ will keep decreasing to $F_{C_M - C_I}(0)$, which is smaller than $P_M(\widehat{\theta}^*)$.

D.8 PROOF OF THM. 4.6

Define $\mathbb{F}_s^c$ as some cumulative density function (CDF) associated with fairness metric C. The unfairness $\left|\mathbb{E}_{X\sim P_a^c}[\mathbf{1}(x \geqslant \theta_a)] - \mathbb{E}_{X\sim P_b^c}[\mathbf{1}(x \geqslant \theta_b)]\right|$ can also be written as $\mathbb{F}_a^c(\theta_a) - \mathbb{F}_a^c(\theta_b)$.

1. Under situation 1, Thm. 4.5 already reveals increasing k_1 can disincentivize strategic manipulation. Meanwhile, $\mathbb{F}_s^c(\theta_s^*(k_1))$ will decrease for both groups because $\theta_s^*(k_1)$ decreases for both group. Thus, there must exist k_{1a}, k_{1b} to mitigate the difference between $\mathbb{F}_a^c(\theta_a^*(k_1))$ and $\mathbb{F}_b^c(\theta_b^*(k_1))$, which is promoting the fairness at the same time of disincentivizing manipulation.

2. Under situation 2, Thm. 4.5 already reveals increasing k_2 can disincentivize strategic manipulation. Meanwhile, $\mathbb{F}_s^c(\theta_s^*(k_2))$ will increase for both groups because $\theta_s^*(k_2)$ increases for both group. Thus, there must exist k_{2a}, k_{2b} to mitigate the difference between $\mathbb{F}_a^c(\theta_a^*(k_2))$ and $\mathbb{F}_b^c(\theta_b^*(k_2))$, which is promoting the fairness at the same time of disincentivizing manipulation.

3.Under situation 3, Thm. 4.5 already reveals increasing k_1 for group a and increasing k_2 for group b can disincentivize strategic manipulation. Meanwhile, $\mathbb{F}_s^c(\theta_a^*(k_1))$ will decrease for a and $\mathbb{F}_s^c(\theta_b^*(k_2))$ increase for b. Thus, because a is already the disadvantaged group, the difference between $\mathbb{F}_s^c(\theta_a^*(k_1))$ and $\mathbb{F}_s^c(\theta_b^*(k_2))$ will be mitigated, which is promoting the fairness at the same time of disincentivizing manipulation.

D.9 PROOF OF COROLLARY 4.7

Corollary 4.7 can be derived directly from Thm. 4.5 and Thm. 4.6. To recap, Thm. 4.5 identifies all scenarios under which manipulation is guaranteed to be disincentivized via adjusting preferences; Theorem reftheorem:fairness finds all scenarios when promoting fairness and disincentivizing manip-

ulation can be attained simultaneously; Corollary 4.7 emphasizes all scenarios where disincentivizing manipulation does not guarantee fairness improvement.

In Corollary 4.7, to ensure the manipulation to always be disincentivized, both groups a, b should satisfy **either** scenario identified in Thm. 4.5. This results in four possible combinations, and three out of these four are the scenarios found in Thm. 4.6. The left one situation is the case in Corollary 4.7 (group a satisfies condition 2 and group b satisfies condition 1). In this case, group a can be disincentivized only by increasing k_2. However, increasing k_2 can only make the decision threshold $\widehat{\theta}_a^*$ higher, which will exacerbate the unfairness (since group a has $\alpha_a < 0.5$, by condition 2.(i)).