Impact of Fairness Regulations on Institutions' Policies and Population Qualifications

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we consider institutions whose primary objective is to maximize utility by selecting the most qualified individuals. We propose a regulatory framework to promote demographic parity by penalizing the disparity in the institution's selection across groups. We analyze how the shape of the penalty function and specific conditions determine the effectiveness of a discrimination penalty in reducing selection disparities. Additionally, we explore the implications of such a penalty when individual qualifications evolve over time in response to the penalization policy. We identify scenarios where the penalty could hinder the natural attainment of equity within the population. Moreover, we identify conditions that counteract this undesirable outcome and ensure fairness.

A major concern in deploying algorithmic decision-making systems is the impact of societal biases on their performance. Furthermore, there is increasing advocacy for regulations that not only mitigate biases within these systems but also leverage them to reduce broader societal biases (Madiega, 2021; Smuha, 2021). One common notion of fairness in this context is *demographic parity*. For a system that distributes a given resource (such as loans, college admissions, or employment opportunities) within a population, demographic parity requires that each protected group within the population (e.g., gender or racial groups) receives a proportionate share relative to its size (Barocas et al., 2023).

Previous studies often propose methods that fully align a system with specific fairness notions or report success in mitigating bias (e.g., Hardt et al. (2016b); Feldman et al. (2015); Calders et al. (2009); see Mehrabi et al. (2021) for a survey). In contrast, we examine demographic parity from a regulator's perspective. We ask when and how regulators can effectively enforce parity on utility-maximizing institutions through penalization. This framework introduces a trade-off between selecting individuals with higher qualifications and achieving societal parity.

First, in a single-shot static setting, we identify cases in which imposing a penalty fails to alter institutional policies. We propose a formulation for determining the requisite penalty level that would compel the institution to deviate from its default policies in the absence of any penalty. We demonstrate that the shape of the penalty function leads to qualitatively different thresholds for the required penalty level. This is illustrated by comparing linear or quadratic penalty functions. Second, we discuss when the penalty ensures complete demographic parity. Interestingly, certain penalty functions cannot guarantee complete demographic parity, regardless of how large the penalty is.

Third, we explore a dynamic setting where institutional policies can influence the qualifications of individuals over time, thereby affecting the composition of the population. We adapt the stochastic modeling of Zhang et al. (2020) and combine it with our regulatory framework. Similar to previous works (Mouzannar et al., 2019; Zhang et al., 2020; Raab & Liu, 2021) (but in a different setting), we find instances in our model where fairness regulations can prevent the natural convergence of equal qualification distribution across groups. Nevertheless, we identify conditions that prevent this undesirable effect.

To obtain the theoretical results, we identify and utilize non-trivial properties of the institution's optimal policy. Furthermore, analyzing the dynamic setting requires addressing deviations from linear systems and standard Markov models.

Finally, we empirically examine our theoretical results using the Law School Admission Council's dataset (Wightman, 1998). The experiments are consistent with the theoretical results both in the static and dynamic cases (where we use synthesized dynamics). The experiments confirm that our framework corresponds to real-world scenarios and can be utilized in the regulation of discrimination and fairness.

0.1 Related Work

In our work, we formulate fairness regulation as a penalty imposed by a regulator. In particular, we impose a penalty $\lambda \cdot g(\Delta)$ on the institution where Δ is the demographic disparity, $g(\cdot)$ is a shaping function, and λ is the penalty level. If we impose an infinite discrimination penalty by letting $\lambda \to \infty$, we recover the constraint $\Delta = 0$ as a special case. Therefore, our model in the static case is a generalization of the models proposed in previous works (Mouzannar et al., 2019; Zhang et al., 2020; Raab & Liu, 2021) which consider policies constrained to satisfy demographic parity, i.e., $\Delta = 0$. Moreover, while previous models (Mouzannar et al., 2019; Zhang et al., 2020; Raab & Liu, 2021) assume that qualifications are binary, we allow them to take values in an arbitrary finite set. Finally, Mouzannar et al. (2019) assumes that the institution makes decisions based on the group identity and qualification, and Zhang et al. (2020) and Raab & Liu (2021) consider the institution's policy as the optimal solution of an optimization problem on the features of individuals. We show (at the end of Section 1) that these two assumptions are equivalent within our framework.

In Section 3 we explore the dynamics of the population's qualification over time. The insight that fairness and bias within society evolve in interaction with algorithmic systems has garnered attention (Chaney et al., 2018; D'Amour et al., 2020). One of the first efforts (Liu et al., 2018) considered a two-stage model and analyzed the impact of decision-making based on some qualification (e.g., credit score) on the same qualification. Our dynamic modeling is based on the Markov decision process proposed by Zhang et al. (2020). With the same question of the long-term impact of fairness interventions, Mouzannar et al. (2019) considers an arbitrary and unknown dynamic for qualifications, and Raab & Liu (2021) focuses on imitation dynamics. Similar in modeling, but with a different goal is a line of work in reinforcement learning and online learning that develop algorithms to optimize a mix of utility and fairness objectives over time (Joseph et al., 2016; Jabbari et al., 2017; Wen et al., 2021).

Another important approach to the interplay between society and algorithmic systems is through game-theoretic formulations. The selecting system aims for accurate classification, and individuals can invest in their qualification (Liu et al., 2020) or try to deceive the system as in the strategic classification framework (Hardt et al., 2016a; Milli et al., 2019). Finally, a collection of works targets more specific economic settings (compared to our general framework) to analyze long-term discrimination and fairness, e.g., the effectiveness of the Rooney Rule (Celis et al., 2021) and fairness in needed effort for progress (Heidari et al., 2019). In particular, Hu & Chen (2018) considers how a short-term fairness intervention in the labor market can mitigate bias in the long term. Similarly, we ask if the fairness regulation leads to a state where an unconstrained policy selects fairly.

1 Problem Formulation in the Static Case

First, we investigate the impact of fairness regulation in a static setting. We consider a scenario in which each member of a population possesses a feature vector denoted as $x \in \mathcal{X}$, which is accessible to an institution that makes selections among them. Furthermore, individuals can be classified into distinct groups, such as racial or gender groups, and achieving fairness across these groups is a primary concern. For simplicity, we assume that there are two groups $\mathcal{A}$ and $\mathcal{B}$, and that each individual belongs to exactly one of them. This membership is represented by the variable $c \in \mathscr{C} = \{\mathcal{A}, \mathcal{B}\}$. An individual's qualification is represented by the variable $y \in \mathcal{Y} \subset \mathbb{R}$. A value of $y > 0$ indicates a positive net benefit for the institution, whereas $y < 0$ signifies that the costs of selecting the individual outweigh the benefits. Throughout, we assume $\mathcal{Y}$ is a finite set and $y \neq 0$ for all $y \in \mathcal{Y}$; thus, selecting any individual is strictly beneficial or strictly harmful. We assume an underlying joint distribution $p(c, x, y)$ for an individual selected uniformly at random from the society.

The institution employs a policy to determine whether to select an individual based on her feature x and group membership c. This decision is represented by the binary variable $D \in \{0, 1\}$. The institution's

selection policy can be modeled by a conditional distribution $p(d|c,x)$. Thus, the joint distribution of the random variables D, C, X, Y factorizes as $p(d,c,x,y) = p(d|c,x)p(c,x,y)$.

The institution's utility is defined as the expected value of the qualification of the selected individuals, denoted as $u = \mathbb{E}[YD]$. This utility can be interpreted as the expected utility for a randomly selected individual who undergoes the institution's policy or as the average utility when the policy is applied to the entire population. The randomness here arises from the distribution of features and qualifications in the population, as well as any random elements present in the selection policy itself.

To enforce demographic parity, we introduce a discrimination penalty based on selection disparity across groups, denoted by Δ:

$$\Delta = \left|\Pr(D=1|C=\mathcal{A}) - \Pr(D=1|C=\mathcal{B})\right|. \tag{1}$$

Consequently, the objective of the institution in selecting its policy can be expressed as follows:

$$\underset{p(d|c,x)}{\text{maximize}}\ \mathbb{E}[DY] - \lambda \cdot g(\Delta)$$

where $g(\cdot)$ is a non-decreasing and non-negative convex function satisfying $g(0) = 0$, and $\lambda \geq 0$. Therefore, $\lambda \cdot g(\Delta)$ determines the penalty imposed on the institution (discrimination penalty). A discrimination penalty $\lambda \cdot g(\Delta)$ can represent the fine imposed by the government on the institution to enforce demographic parity. It may also represent the negative effects of discrimination on the company, such as reduced productivity, decreased commitment, and employee demotivation, resulting in an overall cost to the institution. Thus, the true objective of the company can be written as $\mathsf{Profit} = \mathbb{E}[DY] - \lambda \cdot g(\Delta)$. As an example, consider

$$g(\Delta) = \begin{cases} 0, & 0 \leq \Delta \leq \Delta^* \\ \Delta - \Delta^*, & \Delta > \Delta^* \end{cases}.$$

This cost function sets a level Δ^* of tolerable discrimination below which no penalty is imposed.

Throughout this paper, we assume that the institution knows the underlying joint distribution $p(c,x,y)$ and can exactly solve the above optimization problem. This assumption represents an idealization of real-world scenarios, where such problems are typically addressed using approximate solutions based on the available data. Nonetheless, the insights gained from this idealized model remain applicable and valuable in practical settings.

1.1 A simplification

We claim that, without loss of generality, we may assume that the institution has access to individuals' qualification information, meaning that Y is a known function of C and X. To see this, given any x, c, we define $\hat{Y} = \mathbb{E}[Y|X,C]$. Thus, $\hat{Y}$ is a function of (X,C). Then, we have

$$\begin{aligned} \mathbb{E}[DY] - \lambda \cdot g(\Delta) &= \mathbb{E}\left[D\mathbb{E}[Y|X,C]\right] - \lambda \cdot g(\Delta) \\ &= \mathbb{E}[D\hat{Y}] - \lambda \cdot g(\Delta). \end{aligned}$$

Therefore, we obtain the same utility for $\hat{Y}$ in place of Y. This shows that we can restrict ourselves to the case of Y being a function of C and X.

Next, observe that the objective function $\mathbb{E}[DY] - \lambda \cdot g(\Delta)$ depends only on the marginal distribution on (D,C,Y) (rather than the full joint distribution of (D,C,Y,X)). Consequently, we can write the optimization problem as a maximum over $p(d|c,y)$ as follows:

$$\underset{p(d|c,y)}{\text{maximize}}\ \mathbb{E}[DY] - \lambda \cdot g(\Delta). \tag{2}$$

2 Effectiveness of the Discrimination Penalty in the Static Case

If there is no penalty for discrimination (i.e., $\lambda = 0$), the institution simply maximizes $\mathbb{E}[DY]$. Let Y_+ be the positive part of Y, i.e., $Y_+ = \max(0, Y)$. Then, since $D \in \{0,1\}$, we have $\mathbb{E}[DY] \leq \mathbb{E}[Y_+]$ and equality is

achieved by a unique policy of the institution selecting an individual when $Y > 0$. We define $u_{\mathrm{UM}} = \mathbb{E}[Y_+]$ to be the utility when $\lambda = 0$. The resulting discrimination, denoted by Δ_{UM}, simplifies as follows:

$$\begin{aligned}\Delta_{\mathrm{UM}} &= |\Pr(D = 1|C = \mathcal{A}) - \Pr(D = 1|C = \mathcal{B})| \\ &= |\Pr(Y > 0|C = \mathcal{A}) - \Pr(Y > 0|C = \mathcal{B})|. \end{aligned} \tag{3}$$

Thus, Δ_{UM}, the discrimination for the optimal policy when $\lambda = 0$, can be computed explicitly.

Note that if $\Pr(Y > 0|C = \mathcal{A}) = \Pr(Y > 0|C = \mathcal{B})$, we have $\Delta_{\mathrm{UM}} = 0$ and there is no tension between maximizing $\mathbb{E}[DY]$ and minimizing Δ. Thus, the more interesting case is when $\Delta_{\mathrm{UM}} > 0$, *i.e.,* $\Pr(Y > 0|C = \mathcal{A}) \neq \Pr(Y > 0|C = \mathcal{B})$. Throughout the paper, when discussing the static case, we assume (without loss of generality) that individuals in group $\mathcal{A}$ are relatively more beneficial to the institutions than individuals in group $\mathcal{B}$, i.e.

$$\Pr(Y > 0|C = \mathcal{A}) > \Pr(Y > 0|C = \mathcal{B}) \tag{4}$$

If this condition does not hold, we can swap the names for groups $\mathcal{A}$ and $\mathcal{B}$.

Unlike $\lambda = 0$, the optimal policy is not necessarily unique when $\lambda > 0$. This property, along with several other basic properties of optimal policies of the institution, is discussed in Lemma 2 in Appendix A. One important property to highlight here is that for any λ and $g(\cdot)$, under any optimal policy, the advantaged group never has a smaller selection rate, i.e.,

$$\Pr(D = 1|C = \mathcal{A}) \geq \Pr(D = 1|C = \mathcal{B}).$$

Thus, under any optimal policy, equation 1 can be rewritten as

$$\Delta = \Pr(D = 1|C = \mathcal{A}) - \Pr(D = 1|C = \mathcal{B}). \tag{5}$$

We begin by considering when the penalty is effective, meaning it will force the institution to abandon the utility-maximizing policy and reduce disparity in selection across groups.

Definition 1. *As above, assume $\Delta_{UM} > 0$. We call a penalty, given by function $g(\cdot)$ and scalar λ, effective when every optimal solution of the problem in equation 2 satisfies*

$$\begin{aligned}\Delta &= |\Pr(D = 1|C = \mathcal{A}) - \Pr(D = 1|C = \mathcal{B})| \\ &< \Delta_{UM} = \Pr(Y > 0|C = \mathcal{A}) - \Pr(Y > 0|C = \mathcal{B}).\end{aligned}$$

The next theorem states the condition to ensure the effectiveness of the penalty, which depends on the size of the groups, possible qualifications $\mathcal{Y}$, and the bias present in the population.

Theorem 1. *The penalty function $g(\cdot)$ and scalar λ are effective if and only if $\beta_e < \lambda \cdot g'_-(\Delta_{UM})$, where $g'_-(\cdot)$ is the left-hand derivative of g*

$$g'_-(x) = \lim_{h \to 0^-} \frac{g(x + h) - g(x)}{h}$$

and $\beta_e = \min_{(y,c)\in\mathcal{T}_e} p(c) \cdot |y|$ where $\mathcal{T}_e \subset \mathcal{Y} \times \{\mathcal{A}, \mathcal{B}\}$ is defined as

$$\mathcal{T}_e = \{(y, \mathcal{A}) : y > 0, p(y|C = \mathcal{A}) > 0\} \cup \{(y, \mathcal{B}) : y < 0, p(y|C = \mathcal{B}) > 0\}.$$

Note that equation 4 guarantees that the set $\mathcal{T}_e$ is non-empty. The proof of this and the rest of the theorems and lemmas is deferred to Appendix A.2.

Remark 1. *Observe that β_e depends on the distribution of qualifications across groups $p(y|c)$ only through its support set, i.e., only on the set of y where $p(y|c) > 0$ is important and not on the exact values of $p(y|c)$. Next, note that effectiveness only depends on the derivative of the discrimination penalty at Δ_{UM}. The choice of $g(\cdot)$ is also important. For instance, consider two functions $g_1(x) = x$ and $g_2(x) = x^2$. As $g_1(x)$ has a*

constant derivative, the threshold of λ for effectiveness is found by $\beta_e < \lambda$, which is independent of Δ_{UM}. But in the case of the square function $g_2(x)$, the threshold will be $\beta_e/(2\Delta_{UM}) < \lambda$, so the current level of bias in the population affects the effectiveness of the penalty. Therefore, assessing the effectiveness of the penalty becomes more complex. Also, in an evolving population (as in the next section), the effectiveness may change over time.

Next, we discuss the possibility of reaching complete demographic parity, $\Pr(D = 1|C = \mathcal{A}) = \Pr(D = 1|C = \mathcal{B})$. Note that the institution always has the option of adopting a policy that satisfies complete demographic parity, i.e.,

$$\max_{p(d|c,y)} \mathbb{E}[DY] - \lambda \cdot g(\Delta) \geq \max_{\substack{p(d|c,y):\\ p(D=1|\mathcal{A})=p(D=1|\mathcal{B})}} \mathbb{E}[DY],$$

where, for simplicity of notation, we use $p(D = 1|\mathcal{A})$ and $p(D = 1|\mathcal{B})$ to denote $p(D = 1|C = \mathcal{A})$ and $p(D = 1|C = \mathcal{B})$, respectively.

We are interested in when the discrimination penalty is strong enough to ensure complete demographic parity in the institution's policy. This concept is formalized below:

Definition 2. *We call a penalty, given by the function $g(\cdot)$ and scalar λ, fully satisfactory when at least one optimal solution of the problem in equation 2 satisfies*

$$\Delta = |\Pr(D = 1|C = \mathcal{A}) - \Pr(D = 1|C = \mathcal{B})| = 0.$$

Equivalently,

$$\max_{p(d|c,y)} \mathbb{E}[DY] - \lambda \cdot g(\Delta) = \max_{\substack{p(d|c,y):\\ p(D=1|\mathcal{A})=p(D=1|\mathcal{B})}} \mathbb{E}[DY].$$

We have the following result:

Theorem 2. *For a given distribution $p(y, c)$, if $\Delta_{UM} > 0$ then, the penalty function $g(\cdot)$ and scalar λ are fully satisfactory if and only if $\beta_s \leq \lambda \cdot g'_+(0)$, where $g'_+(\cdot)$ is the right-hand derivative of g:*

$$g'_+(x) = \lim_{h \to 0^+} \frac{g(x + h) - g(x)}{h}$$

and β_s is defined as follows: Consider the problem of finding the best policy that satisfies demographic parity:

$$\max_{\substack{p(d|c,y):\\ p(D=1|\mathcal{A})=p(D=1|\mathcal{B})}} \mathbb{E}[DY]. \tag{6}$$

Let $p^*(d|c, y)$ be any optimal solution to the above optimization (if there is more than one optimal policy, we can choose any one of them). Then, let $\mathcal{T}_s \subset \mathcal{Y} \times \{\mathcal{A}, \mathcal{B}\}$ be defined as

$$\begin{aligned}\mathcal{T}_s =& \{(y, \mathcal{A}) : y > 0, p(y|\mathcal{A}) > 0, p^*(D = 1|y, \mathcal{A}) < 1\} \\ &\cup \{(y, \mathcal{B}) : y < 0, p(y|\mathcal{B}) > 0, p^*(D = 1|y, \mathcal{B}) > 0\}.\end{aligned}$$

Finally, set $\beta_s = \max_{(y,c)\in\mathcal{T}_s} p(c) \cdot |y|$. It turns out that the value of β_s will be the same if we start from different optimal solutions of equation 6.

In this case, the condition depends on the policy with a demographic parity constraint. Therefore, employing this theorem to assess a discrimination penalty is not as straightforward as Theorem 1. Nevertheless, it underscores important implications in selecting the penalty function $g(\cdot)$. Following our comparison of the functions $g_1(x) = x$ and $g_2(x) = x^2$ in the context of Theorem 1, note that for the square function, we have $g'_+(0) = 0$. Since $g'_+(0) = 0$, the condition $\beta_s \leq \lambda \cdot g'_+(0)$ fails whenever $\beta_s > 0$, so no finite λ can guarantee complete demographic parity. However, when considering $g_1(x) = x$, the condition for achieving complete demographic parity simplifies to $\beta_s \leq \lambda$, indicating the existence of some finite level of penalty sufficient to achieve demographic parity. Once again, this highlights a distinction between these two penalty functions.

3 Dynamics of the Population

After discussing fairness regulations at a static, single point in time, we now shift our focus towards understanding how the population may dynamically evolve over time in response to the selection policy. Real-world examples of such transformations vary widely, ranging from individuals investing in their attributes for better future outcomes, to scenarios where receiving a loan based on a credit score further enhances that score.

We assume that individuals' group membership does not change over time. However, qualifications are states that individuals can acquire and may change over time. Extending our notation from the previous section, we denote a randomly chosen individual's qualification at time t by Y_t. We assume that the institution makes a decision regarding the individual at every time step t, denoted by the random variable D_t. The decision at time t (D_t) has the potential to alter the qualifications of individuals in the next time step. We assume the following Markov property governs these changes:

$$p(y_{t+1}|y_1, \cdots, y_t, d_1, \cdots, d_t, c) = p(y_{t+1}|y_t, d_t) \tag{7}$$

We adopt this assumption from Mouzannar et al. (2019); Zhang et al. (2020) and Raab & Liu (2021). Observe that this assumption implies that the new qualification Y_{t+1} depends only on the current qualification Y_t and the current decision by the institution D_t, but not on group C or past history. Additionally, following the same literature, we assume that the conditional probability $\Pr(Y_{t+1} = y|Y_t = y', D_t = d)$ does not change over time, i.e.,

$$\Pr(Y_{t+1} = y|Y_t = y', D_t = d) = q(y|y', d)$$

for some conditional distribution $q(y|y', d)$ that does not depend on time t.

Note that the decision variable D_t depends not only on the individual's qualification Y_t, but possibly also on the population *distribution* of qualifications across the society, i.e., the joint distribution of (C, Y_t). In other words, $p(d_t|c, y_t)$ may depend on the distribution $p(c, y_t)$ in order to ensure fairness across the society. Consequently, the qualification of a randomly selected individual over time $\{Y_t\}_{t=1,2,\cdots}$ does not form a Markov chain despite the Markov assumption in equation 7.

In this section, we examine the institution's adoption of myopic policies, which maximize the instantaneous objective function discussed in the previous sections at every time step. In other words, we assume that the institution solves the following optimization problem in each step t:

$$\underset{p(d_t|c,y_t)}{\text{maximize}} \ \mathbb{E}[D_t Y_t] - \lambda \cdot g(\Delta_t) \tag{8}$$

where $\Delta_t = |\Pr(D_t = 1|C = \mathcal{A}) - \Pr(D_t = 1|C = \mathcal{B})|$.

If there are multiple maximizers $p(d_t|c, y_t)$, the institution may choose any of them. Note that, as in the previous section, the institution's policy at time t depends only on the marginal distribution of C and Y_t. So, the state of the system at each time step t is determined by the distribution of qualifications of each group, denoted by $p_{Y_t|\mathcal{A}}$ and $p_{Y_t|\mathcal{B}}$, as well as the choice of maximizer if the optimization problem in equation 8 has more than one maximizer. We call a state stationary when it does not change in the next time step.

Definition 3. *For an instance of the system, given by group sizes* $\Pr(C = \mathcal{A})$ *and* $\Pr(C = \mathcal{B})$*, discrimination penalty* $g(\cdot)$ *and* λ *and the conditional distribution* $q(y|y', d)$*, a state* $(p_{Y|\mathcal{A}}, p_{Y|\mathcal{B}}, p_{D|C,Y})$ *is stationary if, starting from* $p_{Y_t|\mathcal{A}} = p_{Y|\mathcal{A}}, p_{Y_t|\mathcal{B}} = p_{Y|\mathcal{B}}$*, we have that* $p_{D|C,Y}$ *is a maximizer for the optimization problem in equation 8, and the state does not change in the next time step:*

$$p_{Y_t|\mathcal{A}} = p_{Y_{t+1}|\mathcal{A}}, \qquad p_{Y_t|\mathcal{B}} = p_{Y_{t+1}|\mathcal{B}}.$$

Stationary states are interesting because if the system reaches them at some point in time, it will not move to any other state thereafter.

Within the literature exploring societal dynamics and algorithmic decision-making, social equality between groups within a population is often characterized by the condition $\mathbb{E}[Y_t|C = \mathcal{A}] = \mathbb{E}[Y_t|C = \mathcal{B}]$ (Mouzannar

et al., 2019; Zhang et al., 2020; Raab & Liu, 2021). This condition alone does not guarantee equal selection outcomes under a utility-maximizing policy (unless Y_t is a binary random variable). More generally, in this paper, we study the joint distribution of (C, Y_t). The statements we prove can be used to study the difference between $\mathbb{E}[Y_t|C = \mathcal{A}]$ and $\mathbb{E}[Y_t|C = \mathcal{B}]$.

Our goal is to study the impact of λ and penalty function $g(\cdot)$ on the limit of disparity in selection, i.e. $\lim_{t\to\infty} \Delta_t$.

First, we show that under some mild assumptions if $\lambda = 0$ in equation 8, the amount of discrimination Δ_t vanishes over time. In other words, even when we do not apply any discrimination penalty, both groups will end up with the same qualification distribution in the long run. This is because we use the same update rule $q(y|y', d)$ for qualifications across different groups (the new qualification Y_{t+1} depends only on the current qualification Y_t and the current decision by the institution D_t, *but not on the individual's group identity* C). In other words, equal opportunities for growth or decline are provided to different groups. In this scenario, Theorem 3 below informs us that even if there is initially a disparity in qualifications, which implies uneven selection under a policy that maximizes utility, equal opportunities for change will eventually result in identical qualification distributions across the groups in the long run.

On the other hand, Theorem 3 below shows that the amount of discrimination Δ_t may not vanish over time when $\lambda > 0$. This is rather surprising because by choosing a positive λ, we penalize discrimination at each step. However, we observe that adopting such a myopic policy in some cases might result in non-vanishing discrimination in the long run, i.e., $\lim_{t\to\infty} \Delta_t > 0$. Furthermore, it is possible that $\lim_{t\to\infty} \Delta_t > \Delta_0$, i.e., the limiting discrimination value is higher than the original discrimination in the society at time $t = 0$.

Theorem 3. *Assume that $q(y|y', d) > 0$ for all $y, y' \in \mathcal{Y}$ and $\mathcal{Y}$ is a finite set. Then, if $\lambda = 0$, we have $\lim_{t\to\infty} \Delta_t = 0$. More generally, C and Y_t will become asymptotically independent: $\lim_{t\to\infty} \|p_{C,Y_t} - p_C p_{Y_t}\|_1 = 0$ where $\|\cdot\|_1$ is the L^1 distance. However, one can find examples for $\lambda > 0$ in which $\lim_{t\to\infty} \Delta_t > \Delta_0 > 0$.*

In summary, while penalizing discrimination is often driven by moral imperatives or the desire to mitigate population biases, Theorem 3 suggests this approach may inadvertently prevent the long-term achievement of equality among groups. We note that this negative impact is similar to what has been described (in different settings) by Mouzannar et al. (2019); Zhang et al. (2020) and Raab & Liu (2021), where enforcing certain fairness criteria in decision-making can prevent society from naturally attaining equality.

Figure 1b shows an example of the negative impact (discussed in Theorem 3) for the penalty function $g(x) = x$. The figure depicts convergence to equality assuming no penalty, as well as the impact of an always fully satisfactory penalty ($\lambda = 1.5$). It also shows how an effective penalty may converge to a persistent disparity ($\lambda = 0.7$). Details are given after the proof of Theorem 3 in Appendix A.4.

Observing the unintended consequences of the discrimination penalty, it becomes important to identify conditions that can prevent it. In the following theorem, we establish that if the transition probabilities $q(y|y', d)$ are sufficiently high, the disparity between groups will vanish regardless of the level of penalty imposed.

Theorem 4. *Define $\alpha = \min_{y,y',d}\{q(y|y', d)\}$, then for any $\lambda \geq 0$ and convex function $g(\cdot)$ we have*

$$\|p_{Y_{t+1}|\mathcal{A}} - p_{Y_{t+1}|\mathcal{B}}\|_1 \leq 2\,(1 - \alpha|\mathcal{Y}|)\,\|p_{Y_t|\mathcal{A}} - p_{Y_t|\mathcal{B}}\|_1$$

Specifically, $\alpha > 1/(2|\mathcal{Y}|)$ implies $\lim_{t\to\infty} \|p_{Y_{t+1}|\mathcal{A}} - p_{Y_{t+1}|\mathcal{B}}\|_1 = 0$ and $\lim_{t\to\infty} \Delta_t = 0$.

3.1 Dynamics of the institution's objective

In this section, we study the dynamics of the institution's objective $\mathsf{Profit}(t) = \mathbb{E}[D_t Y_t] - \lambda \cdot g(\Delta_t)$ as time progresses. Figure 1a considers the setting in Example 1 and plots the function $t \mapsto \mathsf{Profit}(t)$ for different values of λ. One can observe that $\mathsf{Profit}(t)$ might be increasing or decreasing over time, depending on the value of λ. The following theorem studies the behavior of $\mathsf{Profit}(t)$ as a function of time for a different natural setting in which the selection of an individual has a positive average impact on one's qualification.

(a) Institution's objective
$\mathsf{Profit}(t) = \mathbb{E}[D_t Y_t] - \lambda \cdot g(\Delta_t)$

(b)
Disparity Δ

Figure 1: Evolution of the institution's objective and disparity for different levels of discrimination penalty λ.

Theorem 5. *Assume that selecting an individual would cause her average qualification to grow on average (for instance due to the job experience), i.e.,*

$$\mathbb{E}[Y_{t+1}|D_t = 1, Y_t = y_t] \geq y_t, \qquad \forall y_t.$$

We do not make any assumption about the way qualifications of unselected individuals change over time. Then, for $\lambda \geq 0$, any penalty function $g(\cdot)$ and any initial distribution $p(c, y_0)$, the sequence of objective values at time t

$$\mathsf{Profit}(t) = \underset{p(d_t|c,y_t)}{\text{maximize}} \; [\mathbb{E}[D_t Y_t] - \lambda \cdot g(\Delta_t)] \tag{9}$$

will be a non-decreasing function of t.

Note that the objective function has two components $\mathbb{E}[D_t Y_t]$ and $-\lambda \cdot g(\Delta_t)$. The fact that their sum will be a non-decreasing function of t indicates that at least one of the components should not decrease in each time slot. However, this may not lead to social parity. Observe that if we have $\Pr(Y_0 > 0|C = \mathcal{B}) = 0$ and λ is not large enough for the penalty to become effective, the institution may never select from group $\mathcal{B}$ and one may never achieve social parity, even though the institution's total income (after paying the discrimination penalty) might still increase over time.

3.2 Stationary states

In this section, we are interested in stationary states (as defined in Definition 3) in which groups have the same qualification distribution and, therefore, enjoy equal selection rates. Such stationary states always exist, as shown in the following lemma:

Lemma 1. *Consider a finite Markov chain with the transition matrix $T_{\mathcal{Y}\times\mathcal{Y}}$*

$$T(y'|y) = \begin{cases} q(y'|y, 1) & \text{if } y > 0 \\ q(y'|y, 0) & \text{if } y < 0 \end{cases}, \qquad \forall y, y' \in \mathcal{Y}$$

Let r_Y be any stationary distribution of the above Markov chain. Then, $(p_{Y_t|\mathcal{A}}, p_{Y_t|\mathcal{B}}) = (r_Y, r_Y)$ along with

$$D = \begin{cases} 0, & Y < 0 \\ 1, & Y > 0 \end{cases}$$

is a stationary state of our dynamics as defined in Definition 3. This stationary distribution satisfies $p_{Y_t|\mathcal{A}} = p_{Y_t|\mathcal{B}}$. Thus, in stationary states of this type, we have

$$\Pr(D_t = 1|C = \mathcal{A}) = \Pr(D_t = 1|C = \mathcal{B}).$$

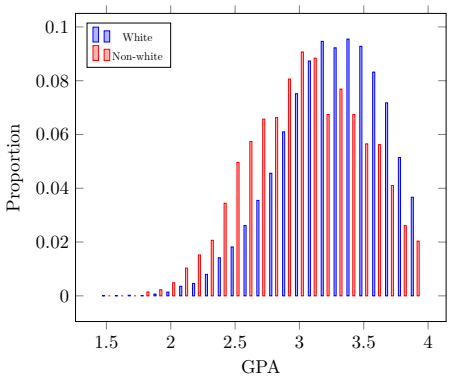


Figure 2: Distribution of GPA across the two race groups in LSAC dataset

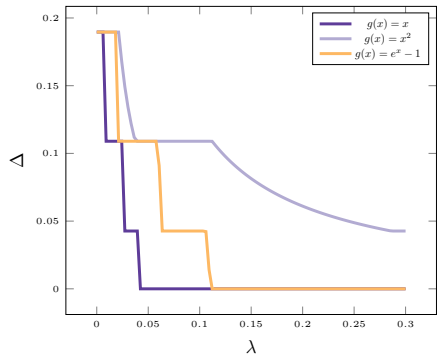


Figure 3: Impact of different penalty functions on disparity on LSAC dataset

We did not define $T(y'|y)$ for $y = 0$, as we assumed $y \neq 0$ for all $y \in \mathcal{Y}$ throughout this paper.

Note that, in general, the Markov chain T does not necessarily have a unique stationary distribution. Therefore, our dynamics will not always have a unique stationary state with equal selection rates. However, it is desirable to find cases where the stationary state is unique.

Next, we examine the case where qualifications change slowly over time, meaning that an individual can only move to the next higher or lower level of qualification or stay put. One preliminary observation in this scenario is that stationary states of the type described in Lemma 1 are unique in this case because the Markov chain in Lemma 1 will be irreducible and aperiodic (Ross, 2024). However, apart from this unique stationary state with equality across groups, there may exist other stationary states that can persist disparities. But in the following theorem, we prove that if the selection increases the chance of qualification improvement and decreases the chance of degrading, then the stationary state from Lemma 1 is the unique stationary state of the system.

Theorem 6. *Assume that* $\mathcal{Y} = \{y^{(1)}, y^{(2)}, \cdots, y^{(n)}\}$ *where* $y^{(1)} < y^{(2)} < \cdots < y^{(n)}$. *Moreover, assume that*

$$|i - j| \leq 1 \iff q(y^{(i)}|y^{(j)}, d) > 0 \tag{10}$$

and

$$q(y^{(i+1)}|y^{(i)}, 1) \geq q(y^{(i+1)}|y^{(i)}, 0), \qquad \forall 1 \leq i < n \tag{11}$$

$$q(y^{(i-1)}|y^{(i)}, 1) \leq q(y^{(i-1)}|y^{(i)}, 0), \qquad \forall 1 < i \leq n \tag{12}$$

Then, for any discrimination penalty $g(\cdot)$ *and* λ *the system has a unique stationary state. In this stationary state (which can be found using the procedure given in Lemma 1), the groups have the same qualification distribution.*

4 Experiments

We conducted experiments using real-world data from the Law School Admission Council (LSAC) dataset (Wightman, 1998). This dataset tracks the academic performance of 21,406 law students, encompassing sensitive attributes such as race. We consider admission to law school as a selection problem, with individual qualifications defined by their cumulative GPA. To determine an individual's qualification score, we subtracted 2.95 from their GPA. Consequently, a GPA below 3.0 represents a negative qualification, while a GPA of 3.0 or above is considered positive. Because GPAs in the dataset are reported in increments of 0.1, subtracting 2.95 rather than 3.0 ensures that no individual receives a qualification score of exactly zero. We categorized individuals into two distinct race groups, consisting of 17921 white students and 3485 non-white students. The normalized histogram of the data in each group is depicted in Figure 2.

Figure 3 illustrates the impact of three different penalty functions $g(x) = x$, $g(x) = x^2$, and $g(x) = e^x - 1$ on the selection disparity. As predicted by Theorem 1 and confirmed by the experiments, we observe that for each penalty function, there is a threshold level of λ, below which the penalty is ineffective in decreasing disparity in selection. Furthermore, since only $g(x) = x$ and $g(x) = e^x - 1$ have non-zero derivatives at $x = 0$, only these two functions can guarantee complete demographic parity ($\Delta = 0$) for sufficiently large λ in accordance with Theorem 2 and as depicted in the figure.

4.1 Dynamics

As real-world datasets only capture a snapshot in time, we synthesize the dynamics for the experiments and build the system's initial state based on the real-world data. First, we consider dynamics satisfying the conditions of Theorem 6. Moreover, we assume that for each decision, the probabilities of growth and degradation are the same for all qualifications. We generated 100 random transition probabilities $q(y|y', d)$ with these conditions. Assuming the conditions from Theorem 6 and that the probabilities of growth and degradation are the same for all qualifications, implies that for some $q^{1+}, q^{1-}, q^{0+}, q^{0-} \in [0, 1]$ satisfying $q^{0+} + q^{0-} \leq 1$, $q^{1+} + q^{1-} \leq 1$ and $q^{0-} \geq q^{1-}$, we have that for all $1 \leq i < n$:

$$
\begin{aligned}
q(y^{(i+1)}|y^{(i)}, 1) &= q^{1+}, \\
q(y^{(i-1)}|y^{(i)}, 1) &= q^{1-}, \\
q(y^{(i)}|y^{(i)}, 1) &= 1 - q^{1+} - q^{1-}, \\
q(y^{(i+1)}|y^{(i)}, 0) &= q^{0+}, \\
q(y^{(i-1)}|y^{(i)}, 0) &= q^{0-}, \\
q(y^{(i)}|y^{(i)}, 0) &= 1 - q^{0+} - q^{0-}
\end{aligned}
$$

To generate a random dynamic, we sample $q^{1+}, q^{1-}, q^{0+}, q^{0-} \in [0, 1]$ as follows:

$$
\begin{aligned}
q^{1+} &\sim \mathit{Uniform}(0, 1) \\
q^{1-} &\sim \mathit{Uniform}(0, 1 - q^{1+}) \\
q^{0+} &\sim \mathit{Uniform}(0, q^{1+}) \\
q^{0-} &\sim \mathit{Uniform}(q^{1-}, 1 - q^{0+}).
\end{aligned}
$$

Using $g(x) = x$ and a random $\lambda \sim \mathit{Uniform}(0, 1)$ for each dynamic, Figure 4 depicts the quartiles of the L^1 distance between qualification distributions across the groups over time, showcasing the convergence $\lim_{t\to\infty} ||p_{Y_t|\mathcal{A}} - p_{Y_t|\mathcal{B}}||_1 = 0$ (which also implies $\lim_{t\to\infty} \Delta_t = 0$) in all samples. We conjecture that this convergence holds more generally for the dynamics considered in Theorem 6, leaving it for future work.

Furthermore, in Theorem 3 we constructed an example for $\lambda > 0$ in which $\lim_{t\to\infty} \Delta_t > \Delta_0$. This raises the question of how common such examples are. To answer this question, we designed two simulations. We choose $g(x) = x$. To choose λ, we refer to Figure 3. We consider two choices of $\lambda = 0.03$ and $\lambda = 0.05$. Note that the penalty is fully satisfactory at $\lambda = 0.05$, so in this case at time $t = 0$, the penalty forces $\Delta_0 = 0$. However, for $\lambda = 0.03$, we will have $\Delta_0 > 0$ at time $t = 0$.

For $\lambda = 0.03$, we first generated 20 random conditional distributions $q(y|y', d)$ and illustrated their dynamics in Figure 6a (Figure 6b depicts a zoomed version of the same curves to show the initial dynamics). We observe that persistent disparity ($\lim_{t\to\infty} \Delta_t > \Delta_0$) occurs in 7 of the 20 samples. The same experiment is run for $\lambda = 0.05$ in Figure 5, where at time $t = 0$, the penalty is fully satisfactory. Interestingly, we observe a form of delayed impact (Liu et al., 2018). While at time $t = 0$, the penalty is fully satisfactory, as the underlying distribution of the qualifications of the two groups changes, the penalty may no longer be fully satisfactory and the disparity may emerge.

Next, as another experiment, considering the initial configuration derived from the dataset, we first manually identified a dynamic $q(y|y', d)$ leading to persistent disparities ($\lim_{t\to\infty} \Delta_t > \Delta_0$) when $g(x) = x$ and $\lambda = 0.03$. Figure 7 depicts the final distribution of qualifications across groups under this dynamic. Subsequently,

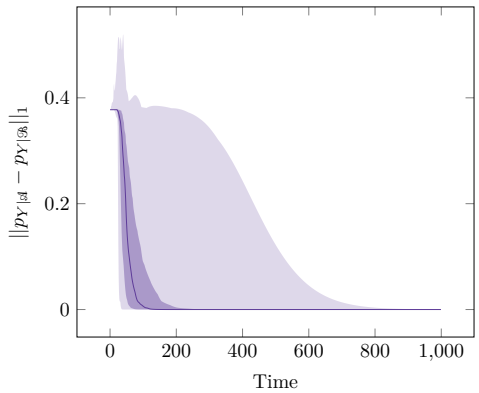


Figure 4: Quartiles of $||p_{Y_t|\mathcal{A}} - p_{Y_t|\mathcal{B}}||_1$ for 100 randomly generated dynamics satisfying conditions of Theorem 6

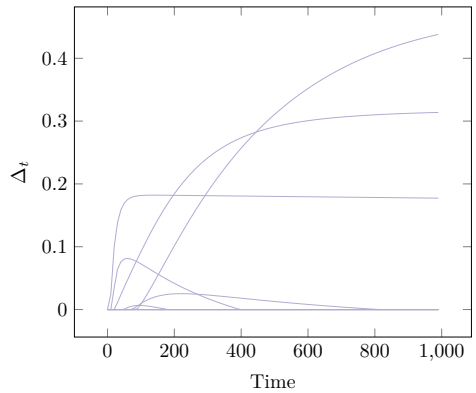
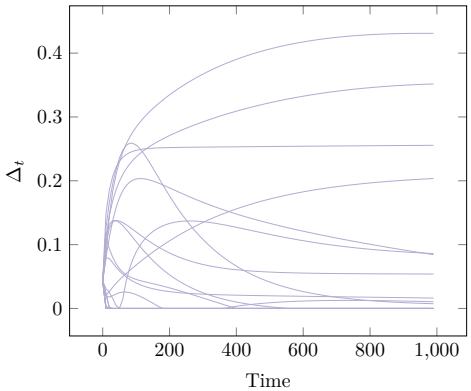
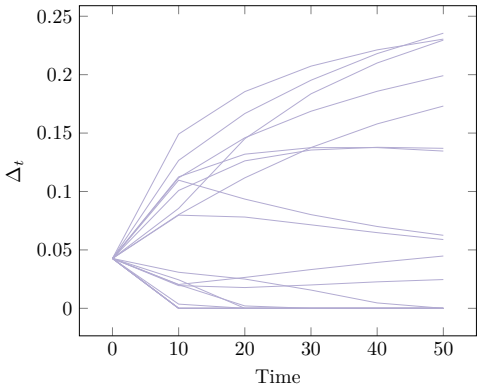


Figure 5: Evolution of Δ_t for random transitions with $\lambda = 0.05$



(a) $0 \leq t \leq 1000$



(b) $0 \leq t \leq 50$

Figure 6: Evolution of Δ_t for random transitions with $\lambda = 0.03$

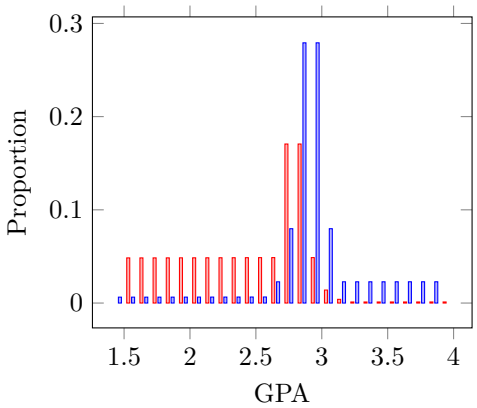
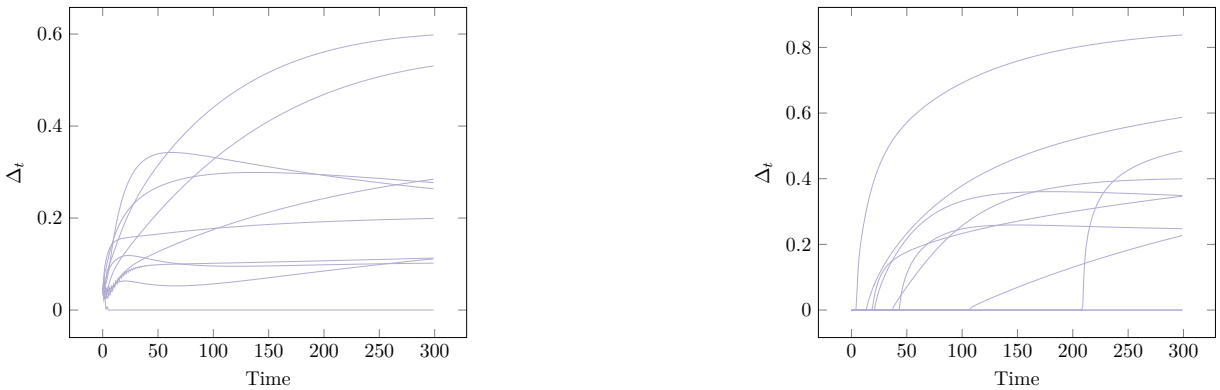


Figure 7: Final distribution of GPA across the groups in a dynamic leading to persisting disparity



(a) $\lambda = 0.03$

(b) $\lambda = 0.05$

Figure 8: Evolution of Δ_t for 10 random perturbations of a dynamic resulting in persistent disparity

we generated 10 random perturbations of $q(y|y', d)$ by adding Gaussian noise of $\mathcal{N}(0, \sigma^2 = 0.1)$ to each transition probability, then taking the absolute value and normalizing to obtain a valid $\tilde{q}(y|y', d)$ again. Note that the noise standard deviation $\sigma \approx 0.316...$, so the perturbations are not small. Figure 8a reveals that, aside from one instance, all systems tend towards persistent disparities over time, i.e., $\lim_{t\to\infty} \Delta_t > \Delta_0$. We repeated the experiment of Figure 8b for $\lambda = 0.05$.

5 Conclusion

In addressing the impact of regulating AI discrimination within society, it is essential to undertake a thorough assessment of the efficiency of potential interventions. Interventions driven solely by intuition may prove ineffective or, worse, produce unintended consequences.

In this paper, we introduced a simple model that encapsulates the strategy of penalizing discrimination among various groups in the selection process of utility-maximizing institutions. Initially, we established the conditions necessary for the effectiveness of such penalties (i.e., compelling the institution to reduce the disparity from its initial state) and the conditions for them to be fully satisfactory (i.e., reducing the disparity to zero). These conditions highlight scenarios in which penalties may be ineffective and different penalty functions may produce markedly different outcomes. Subsequently, we examined the evolution of individuals' qualifications with respect to institutions' selection through a stochastic process model. We demonstrated that equal opportunities for change across groups naturally lead to parity in qualifications and selection over time. However, we highlighted counterintuitive cases where penalizing discrimination

may prevent this convergence. Next, we identified some conditions that prevent such scenarios, including sufficiently high transition probabilities and a positive impact of selection in cases of gradual change.

Despite the insights gained from this study, certain limitations must be acknowledged. This study assumes that the institution is aware of the distribution of qualifications. Designing robust fairness regulation policies under imperfect knowledge remains an important direction for future work. Next, we studied fairness under the assumption of equal opportunity for change across groups. However, many real-world settings may not satisfy this property, and further study in this direction is required.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*, pp. 76–103. MIT Press, 2023.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. The effect of the rooney rule on implicit bias in the long term. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 678–689, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445930. URL https://doi.org/10.1145/3442188.3445930.

Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 224–232, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240370. URL https://doi.org/10.1145/3240323.3240370.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 525–534, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372878. URL https://doi.org/10.1145/3351095.3372878.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Robert G. Gallager. *Discrete Stochastic Processes*, pp. 149–186. Springer US, Boston, MA, 1996. ISBN 978-1-4615-2329-1. doi: 10.1007/978-1-4615-2329-1_5. URL https://doi.org/10.1007/978-1-4615-2329-1_5.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pp. 111–122, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, Red Hook, NY, USA, 2016b. Curran Associates Inc. ISBN 9781510838819.

Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2692–2701. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/heidari19a.html.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pp. 1389–1398, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186044. URL https://doi.org/10.1145/3178876.3186044.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1617–1626. JMLR.org, 2017.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3150–3158. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 381–391, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372861. URL https://doi.org/10.1145/3351095.3372861.

Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021. URL https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 230–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287576. URL https://doi.org/10.1145/3287560.3287576.

Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 359–368, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287599. URL https://doi.org/10.1145/3287560.3287599.

Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26053–26065. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/db00f1b7fdf48fd26b5fb5f309e9afaf-Abstract.html.

Sheldon M. Ross. 4 - markov chains. In Sheldon M. Ross (ed.), *Introduction to Probability Models (Thirteenth Edition)*, pp. 201–299. Academic Press, thirteenth edition edition, 2024. ISBN 978-0-443-18761-2. doi: https://doi.org/10.1016/B978-0-44-318761-2.00009-9. URL https://www.sciencedirect.com/science/article/pii/B9780443187612000099.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

Nathalie A. Smuha. From a 'race to ai' to a 'race to ai regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1):57–84, 2021. doi: 10.1080/17579961.2021.1898300. URL https://doi.org/10.1080/17579961.2021.1898300.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1144–1152. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wen21a.html.

Linda F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, volume 33, pp. 18457–18469. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d6d231705f96d5a35aeb3a76402e49a3-Abstract.html.

A Proofs

A.1 A useful lemma

Lemma 2. *If $\lambda > 0$, the optimal policy $p(d|c, y)$ maximizing equation 2 is not necessarily unique. Moreover, each optimal policy $p(d|c, y)$ satisfies the following properties:*

1. *The selection probability within a group is a non-decreasing function of the qualification level (y). More precisely, if $\Pr(Y = y, C = c) > 0$ and $\Pr(Y = y', C = c) > 0$ for some $y > y'$ and $c \in \mathscr{C}$, then*

$$\Pr(D = 1|Y = y, C = c) \geq \Pr(D = 1|Y = y', C = c)$$

2. *Within each group, at most one qualification subgroup has a fractional selection probability, i.e.,*

$$\begin{aligned} &\forall c\ \exists y' \in \mathcal{Y}\ \forall y \neq y' : \\ &\Pr(Y = y, C = c) > 0 \\ &\qquad \implies \Pr(D = 1|Y = y, C = c) \in \{0, 1\}. \end{aligned}$$

3. *Assume $\Pr(Y > 0|C = \mathscr{A}) \geq \Pr(Y > 0|C = \mathscr{B})$.[1] The advantaged group, $\mathscr{A}$, never has a smaller selection rate:*

$$\Pr(D = 1|C = \mathscr{A}) \geq \Pr(D = 1|C = \mathscr{B})$$

4. *Positively qualified individuals from group $\mathscr{B}$ are always selected, and negatively qualified individuals from group $\mathscr{A}$ are never selected; that is,*

$$\begin{aligned} &\Pr(Y = y, C = \mathscr{A}) > 0, \quad y < 0 \\ &\qquad \implies \Pr(D = 1|Y = y, C = \mathscr{A}) = 0, \\ &\Pr(Y = y, C = \mathscr{B}) > 0, \quad y > 0 \\ &\qquad \implies \Pr(D = 1|Y = y, C = \mathscr{B}) = 1. \end{aligned}$$

Proof.

1. If the inequality

$$\Pr(D = 1|y, C = c) \geq \Pr(D = 1|y', C = c)$$

does not hold for some group c and qualifications $y > y'$, we can increase $\Pr(D = 1|y, C = c)$ and decrease $\Pr(D = 1|y', C = c)$ while keeping $\Pr(D = 1|C = c)$ fixed. By doing so, Δ remains unchanged, while the utility $u = \mathbb{E}[DY]$ strictly increases.

2. Assume that

$$\begin{aligned} 0 &< \Pr(D = 1|y, C = c) < 1, \\ 0 &< \Pr(D = 1|y', C = c) < 1 \end{aligned}$$

for some $y > y'$. We can increase $\Pr(D = 1|y, C = c)$ and decrease $\Pr(D = 1|y', C = c)$ while keeping $\Pr(D = 1|C = c)$ fixed. In this way, Δ remains unchanged, while the utility $u = \mathbb{E}[DY]$ will increase.

3. Assume

$$\Pr(D = 1|C = \mathscr{A}) < \Pr(D = 1|C = \mathscr{B}).$$

[1] This assumption can be made without loss of generality in the static case. See equation 4.

Since for $c \in \mathscr{C}$ we have

$$\begin{aligned}
&\Pr(D = 1|C = c) \\
&= \Pr(Y > 0|C = c) \cdot \Pr(D = 1|Y > 0, C = c) \\
&\quad + \Pr(Y \leq 0|C = c) \cdot \Pr(D = 1|Y \leq 0, C = c)
\end{aligned}$$

then $\Pr(D = 1|C = \mathscr{A}) < \Pr(D = 1|C = \mathscr{B})$ implies either

$$\begin{aligned}
&\Pr(Y > 0|C = \mathscr{A}) \cdot \Pr(D = 1|Y > 0, C = \mathscr{A}) \\
&< \Pr(Y > 0|C = \mathscr{B}) \cdot \Pr(D = 1|Y > 0, C = \mathscr{B})
\end{aligned} \tag{13}$$

or

$$\begin{aligned}
&\Pr(Y \leq 0|C = \mathscr{A}) \cdot \Pr(D = 1|Y \leq 0, C = \mathscr{A}) \\
&< \Pr(Y \leq 0|C = \mathscr{B}) \cdot \Pr(D = 1|Y \leq 0, C = \mathscr{B}).
\end{aligned} \tag{14}$$

Consider the first case. The assumption

$$\Pr(Y > 0|C = \mathscr{A}) \geq \Pr(Y > 0|C = \mathscr{B})$$

in conjunction with equation 13 implies

$$\begin{aligned}
&\Pr(D = 1|Y > 0, C = \mathscr{A}) \\
&< \Pr(D = 1|Y > 0, C = \mathscr{B}).
\end{aligned}$$

Thus, $\Pr(D = 1|Y > 0, C = \mathscr{A}) < 1$. Therefore, for at least one $y > 0$, we have $\Pr(D = 1|y, C = \mathscr{A}) < 1$ and $\Pr(Y = y|C = \mathscr{A}) > 0$. Select $\epsilon \in \mathbb{R}$ such that

$$0 < \epsilon < \min\left\{\frac{\Pr(D = 1|\mathscr{B}) - \Pr(D = 1|\mathscr{A})}{\Pr(Y = y|\mathscr{A})}, 1 - \Pr(D = 1|y, \mathscr{A})\right\}$$

If the institution increases $\Pr(D = 1|y, C = \mathscr{A})$ by ϵ, the inequality

$$\Pr(D = 1|C = \mathscr{A}) < \Pr(D = 1|C = \mathscr{B})$$

will remain satisfied, and the discrimination

$$\Pr(D = 1|C = \mathscr{B}) - \Pr(D = 1|C = \mathscr{A})$$

will decrease. Moreover, $u = \mathbb{E}[DY]$ increases. This contradicts the assumed optimality of the policy.

The second case is similar. Assume that equation 14 holds. Then, we deduce $0 < \Pr(D = 1|Y \leq 0, C = \mathscr{B})$. Thus, there exists a $y < 0$ such that $\Pr(D = 1|y, C = \mathscr{B}) > 0$ and $\Pr(Y = y|C = \mathscr{B}) > 0$. Select $\epsilon \in \mathbb{R}$ such that

$$0 < \epsilon < \min\left\{\frac{\Pr(D = 1|\mathscr{B}) - \Pr(D = 1|\mathscr{A})}{\Pr(Y = y|\mathscr{B})}, \Pr(D = 1|y, \mathscr{B})\right\}$$

Consequently, the institution can increase its utility and decrease the penalty by decreasing $\Pr(D = 1|y, C = \mathscr{B})$ by ϵ.

4. Assume $\Pr(D = 1|Y = y, C = \mathscr{A}) > 0$ for some $y < 0$. From the previous part, we know that

$$\Pr(D = 1|C = \mathscr{A}) \geq \Pr(D = 1|C = \mathscr{B}).$$

If

$$\Pr(D = 1|C = \mathcal{A}) > \Pr(D = 1|C = \mathcal{B})$$

we can slightly decrease $\Pr(D = 1|Y = y, C = \mathcal{A})$ such that $u = \mathbb{E}[DY]$ increases and Δ decreases. Therefore, the assumed policy is not optimal. Next, consider the case of

$$\Pr(D = 1|C = \mathcal{A}) = \Pr(D = 1|C = \mathcal{B}).$$

The assumption $\Pr(D = 1|Y = y, C = \mathcal{A}) > 0$ and parts 1 and 2 of the lemma imply that for any $y' > y$ such that $\Pr(Y = y'|C = \mathcal{A}) > 0$, we have

$$\Pr(D = 1|y', C = \mathcal{A}) = 1$$

So,

$$\begin{aligned}\Pr(D = 1|C = \mathcal{B}) &= \Pr(D = 1|C = \mathcal{A}) \\ &\geq \Pr(Y > 0|C = \mathcal{A}) \\ &> \Pr(Y > 0|C = \mathcal{B}).\end{aligned}$$

This, in turn, implies the existence of some $y' < 0$ such that

$$\Pr(D = 1|Y = y', C = \mathcal{B}) > 0.$$

By decreasing $\Pr(D = 1|Y = y, C = \mathcal{A})$ and $\Pr(D = 1|Y = y', C = \mathcal{B})$ appropriately, we can increase $u = \mathbb{E}[DY]$ while keeping Δ unchanged.

Next, assume $\Pr(D = 1|Y = y, C = \mathcal{B}) < 1$ for some $y > 0$. In this case, if

$$\Pr(D = 1|C = \mathcal{A}) > \Pr(D = 1|C = \mathcal{B})$$

we can slightly increase $\Pr(D = 1|Y = y, C = \mathcal{B})$ to improve the institution's objective. If

$$\Pr(D = 1|C = \mathcal{A}) = \Pr(D = 1|C = \mathcal{B}),$$

parts 1 and 2 of Lemma 2 imply that for all $y' < y$ we have $\Pr(D = 1|Y = y', C = \mathcal{B}) = 0$. In other words, $\Pr(D = 1|Y = y', C = \mathcal{B}) > 0$ implies that $y' > 0$. This implies that

$$\Pr(D = 1|C = \mathcal{B}) \leq \Pr(Y > 0|C = \mathcal{B}).$$

Then, using the assumption

$$\Pr(Y > 0|C = \mathcal{A}) > \Pr(Y > 0|C = \mathcal{B})$$

in equation 4 we have

$$\begin{aligned}\Pr(D = 1|C = \mathcal{A}) &= \Pr(D = 1|C = \mathcal{B}) \\ &\leq \Pr(Y > 0|C = \mathcal{B}) \\ &< \Pr(Y > 0|C = \mathcal{A}).\end{aligned}$$

From

$$\Pr(D = 1|C = \mathcal{A}) < \Pr(Y > 0|C = \mathcal{A})$$

we conclude that there is some $y'' > 0$ such that $\Pr(D = 1|Y = y'', C = \mathcal{A}) < 1$. We can increase $\Pr(D = 1|Y = y'', C = \mathcal{A})$ and $\Pr(D = 1|Y = y, C = \mathcal{B})$ simultaneously to improve the objective while keeping Δ unchanged.

□

Next, we introduce an equivalent formulation for the main optimization problems in equation 2 and equation 6, which simplifies the proofs of Theorems 1 and 2. It relates the optimization problem for any $\lambda > 0$ to $u_{\text{UM}} = \mathbb{E}[Y_+]$, which is the value of the optimization problem for $\lambda = 0$.

Lemma 3. *The optimization problem in equation 2 has the following equivalent form:*

$$\begin{aligned}&\max_{p(d|c,y)} \ \mathbb{E}[DY] - \lambda \cdot g(\Delta) \\ &= \max_{z_{yc}} \ \left[\mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\Big(\Delta_{UM} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\Big)\right]\end{aligned} \tag{15}$$

where Δ_{UM} is defined in equation 3, the set $\mathcal{T}_e$ is defined as follows

$$\mathcal{T}_e = \{(y, \mathcal{A}) : y > 0, p(y|C = \mathcal{A}) > 0\} \cup \{(y, \mathcal{B}) : y < 0, p(y|C = \mathcal{B}) > 0\}, \tag{16}$$

and the maximum is over z_{yc} (for $c \in \mathcal{C}$) satisfying

$$\sum_{(y,c)\in\mathcal{T}_e} z_{yc} \leq \Delta_{UM}, \tag{17}$$

$$0 \leq z_{yc} \leq p(y|c) \qquad \forall y, c. \tag{18}$$

Similarly, the optimization problem in equation 6 has the following equivalent form:

$$\max_{\substack{p(d|c,y):\\ p(D=1|\mathcal{A})=p(D=1|\mathcal{B})}} \mathbb{E}[DY] = \max_{z_{yc}} \ \mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| \tag{19}$$

where the maximum is over z_{yc} satisfying

$$\sum_{(y,c)\in\mathcal{T}_e} z_{yc} = \Delta_{UM}, \tag{20}$$

$$0 \leq z_{yc} \leq p(y|c), \qquad \forall y, c. \tag{21}$$

Proof. Let $p^*(d|y, c)$ be any optimal solution of the optimization problem in equation 2. Define

$$\begin{aligned}z_{y\mathcal{A}} &= p(y|\mathcal{A})\left(1 - p^*(D = 1|Y = y, C = \mathcal{A})\right), \\ z_{y\mathcal{B}} &= p(y|\mathcal{B})p^*(D = 1|Y = y, C = \mathcal{B}).\end{aligned} \tag{22}$$

Applying the fourth part of Lemma 2, we can state the institution's utility under this solution as

$$\begin{aligned}\mathbb{E}_{p^*}[YD] &= \sum_{y>0} y \cdot p(\mathcal{A}) \cdot p(y|\mathcal{A}) - \sum_{y>0} |y| \cdot p(\mathcal{A}) \cdot z_{y\mathcal{A}} \\ &\quad + \sum_{y>0} y \cdot p(\mathcal{B}) \cdot p(y|\mathcal{B}) - \sum_{y<0} |y| \cdot p(\mathcal{B}) \cdot z_{y\mathcal{B}} \\ &= \mathbb{E}[Y_+] - \sum_{y<0} |y| \cdot p(\mathcal{B}) \cdot z_{y\mathcal{B}} - \sum_{y>0} |y| \cdot p(\mathcal{A}) \cdot z_{y\mathcal{A}}\end{aligned}$$

Using equation 5 and the fourth part of Lemma 2, the disparity can be written as

$$\begin{aligned}\Delta_{p^*} &= \sum_{y>0} p(y|\mathcal{A}) - \sum_{y>0} z_{y\mathcal{A}} - \sum_{y>0} p(y|\mathcal{B}) - \sum_{y<0} z_{y\mathcal{B}} \\ &= \Delta_{\text{UM}} - \sum_{y>0} z_{y\mathcal{A}} - \sum_{y<0} z_{y\mathcal{B}}.\end{aligned}$$

Note that the above inequality implies that

$$\sum_{y>0} z_{y\mathcal{A}} + \sum_{y<0} z_{y\mathcal{B}} \leq \Delta_{\text{UM}}. \tag{23}$$

We would like to apply a change of variables and write the optimization problem in equation 2 in terms of $z_{y\mathcal{A}}$ for $y > 0$, and $z_{y\mathcal{B}}$ for $y < 0$. More specifically, we use variables z_{yc} where (y, c) is in $\mathcal{T}_e$:

$$\mathcal{T}_e = \{(y, \mathcal{A}) : y > 0, p(y|C = \mathcal{A}) > 0\} \cup \{(y, \mathcal{B}) : y < 0, p(y|C = \mathcal{B}) > 0\}.$$

We claim that any z_{yc} (for $(y, c) \in \mathcal{T}_e$) satisfying equation 17 and equation 18 corresponds to some decision rule $p(d|y, c)$. Conversely, any *optimal* decision rule $p^*(d|y, c)$ corresponds to some z_{yc} satisfying equation 17 and equation 18. The latter can be verified from equation 23 and the definition given in equation 22.

Take some z_{yc} satisfying equation 17 and equation 18, and define

$$\begin{aligned}
&\forall y > 0 \text{ with } p(y|C = \mathcal{A}) > 0 : \\
&\qquad p(D = 1|Y = y, C = \mathcal{A}) = 1 - \frac{z_{y\mathcal{A}}}{p(y|\mathcal{A})}, \\
&\forall y < 0 \text{ with } p(y|C = \mathcal{B}) > 0 : \\
&\qquad p(D = 1|Y = y, C = \mathcal{B}) = \frac{z_{y\mathcal{B}}}{p(y|\mathcal{B})}.
\end{aligned} \tag{24}$$

Moreover, set $p(D = 1|Y = y, C = \mathcal{A}) = 0$ for $y < 0$ and $p(D = 1|Y = y, C = \mathcal{B}) = 1$ for $y > 0$. Since for all d, y and c we have $0 \leq p(D = d|Y = y, C = c) \leq 1$, this constitutes a valid decision rule.

Thus, the optimization problem in equation 2 is equivalent to maximizing

$$u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right)$$

where the maximum is over z_{yc} satisfying equation 17 and equation 18.

Next, consider the second part of the theorem. We have

$$\max_{\substack{p(d|c,y):\\ p(D=1|\mathcal{A})=p(D=1|\mathcal{B})}} \mathbb{E}[DY] = \max_{p(d|c,y)} \inf_{\lambda\geq 0} \mathbb{E}[DY] - \lambda \cdot g(\Delta)$$

Note that $\mathbb{E}[DY] - \lambda \cdot g(\Delta)$ is concave in $p(d|c, y)$ for any fixed λ. This follows from the fact that $x \mapsto |x|$ is a convex function and $g(\cdot)$ is a non-decreasing convex function. Next, $\mathbb{E}[DY] - \lambda \cdot g(\Delta)$ is linear in λ for any fixed $p(d|c, y)$. Moreover, the domain of $p(d|c, y)$ is a compact set. Therefore, Sion's minimax theorem Sion (1958) implies that

$$\max_{p(d|c,y)} \inf_{\lambda\geq 0} \mathbb{E}[DY] - \lambda \cdot g(\Delta) = \inf_{\lambda\geq 0} \max_{p(d|c,y)} \mathbb{E}[DY] - \lambda \cdot g(\Delta).$$

We can now utilize the first part of the theorem to deduce that

$$\begin{aligned}
&\inf_{\lambda\geq 0} \max_{p(d|c,y)} \mathbb{E}[DY] - \lambda \cdot g(\Delta) \\
&= \inf_{\lambda\geq 0} \max_{z_{yc}} \left[\mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\Big(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\Big)\right]
\end{aligned}$$

where the maximum is over z_{yc} satisfying equation 17 and equation 18. Using a similar minimax exchange, we have

$$\begin{aligned}
&\inf_{\lambda\geq 0} \max_{z_{yc}} \mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) \\
&= \max_{z_{yc}} \inf_{\lambda\geq 0} \mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) && (25) \\
&= \max_{z_{yc}:\Delta_{\text{UM}}=\sum_{(y,c)\in\mathcal{T}_e} z_{yc}} \mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| && (26)
\end{aligned}$$

This completes the proof. □

Lemma 4. *Consider the equivalent form in equation 19 for the problem of finding the best policy that satisfies demographic parity:*

$$\max_{z_{yc}} \ \mathbb{E}[Y_+] - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y|$$

where the maximum is over z_{yc} satisfying

$$\sum_{(y,c)\in\mathcal{T}_e} z_{yc} = \Delta_{UM},$$
$$0 \leq z_{yc} \leq p(y|c), \qquad \forall y, c.$$

where $\mathcal{T}_e$ is defined in equation 16. Then, there is some constant β_s such that any maximizer satisfies

$$\max_{(c,y)\in\mathcal{T}_e : z_{yc}>0} p(c) \cdot |y| = \beta_s.$$

Moreover, any maximizer z_{yc} satisfies the following property:

$$p(c) \cdot |y| > \beta_s \implies z_{yc} = 0,$$
$$p(c) \cdot |y| < \beta_s \implies z_{yc} = p(y|c),$$

implying that two maximizers can differ only on pairs (c, y) where $p(c) \cdot |y| = \beta_s$.

Proof. We begin by showing that for any arbitrary (y, c) and (y', c') in $\mathcal{T}_e$ such that $z_{yc} > 0$ and $p(c') \cdot |y'| < p(c) \cdot |y|$ we have $z_{y'c'} = p(y'|c')$. This property must hold; otherwise, one could choose some $0 < \epsilon < z_{yc}$ to decrease z_{yc} by ϵ while increasing $z_{y'c'}$ by the same value. This move will keep Δ unchanged but decreases $\sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y|$, thereby violating the optimality assumption.

For every optimizer $z_{y,c}$, let us define:

$$\beta_s(z_{y,c}) = \max_{(c,y)\in\mathcal{T}_e : z_{yc}>0} p(c) \cdot |y|.$$

It follows from the above property and the definition of $\beta_s(z_{y,c})$ that

$$p(c) \cdot |y| > \beta_s(z_{y,c}) \implies z_{yc} = 0,$$
$$p(c) \cdot |y| < \beta_s(z_{y,c}) \implies z_{yc} = p(y|c),$$

It remains to show that $\beta_s(z_{y,c}) = \beta_s(z'_{y,c})$ for any two maximizers $z_{y,c}$ and $z'_{y,c}$. Assume, for the sake of contradiction, that $\beta_s(z'_{y,c}) > \beta_s(z_{y,c})$. Then, we obtain that

$$\begin{aligned} z_{yc} > 0 &\implies p(c) \cdot |y| \leq \beta_s(z_{y,c}) \\ &\implies p(c) \cdot |y| < \beta_s(z'_{y,c}) \\ &\implies z'_{yc} = p(y|c) \geq z_{yc}. \end{aligned}$$

Moreover, $z'_{yc} \geq z_{yc}$ also holds when $z_{yc} = 0$. Thus, $z'_{yc} \geq z_{yc}$ holds regardless of the value of z_{yc}. On the other hand

$$\sum_{(y,c)\in\mathcal{T}_e} z'_{yc} = \sum_{(y,c)\in\mathcal{T}_e} z_{yc} = \Delta_{\text{UM}}.$$

Therefore, we must have $z'_{yc} = z_{yc}$ for all $(y, c) \in \mathcal{T}_e$, which is in contradiction with $\beta_s(z'_{y,c}) > \beta_s(z_{y,c})$. □

A.2 Proof of Theorem 1

Proof of Theorem 1. Take an arbitrary

$$(y^*, c^*) \in \arg\min_{(y,c)\in\mathcal{T}_e} p(c) \cdot |y|.$$

Using the assumption $\Delta_{\text{UM}} > 0$, if $\lambda \cdot g'_-(\Delta_{\text{UM}}) > \beta_e$ we conclude that there exists $0 < \epsilon < \min\{\Delta_{\text{UM}}, p(y^*|c^*)\}$ such that

$$\lambda(g(\Delta_{\text{UM}}) - g(\Delta_{\text{UM}} - \epsilon)) > \epsilon \cdot \beta_e.$$

Thus,

$$u_{\text{UM}} - \epsilon \cdot \beta_e - \lambda g(\Delta_{\text{UM}} - \epsilon) > u_{\text{UM}} - \lambda g(\Delta_{\text{UM}}). \tag{27}$$

Consider the representation in Lemma 3 and the optimization of maximizing the following expression:

$$u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) \tag{28}$$

where the maximum is over z_{yc} satisfying equation 17 and equation 18. Let z'_{yc} be defined by

$$z'_{yc} = \begin{cases} \epsilon & \text{if } (y,c) = (y^*, c^*) \\ 0 & \text{otherwise.} \end{cases}$$

One can verify by inspection that the variables z'_{yc} satisfy the constraints equation 17 and equation 18. Thus, the maximum value of equation 28 is greater than or equal to the expression evaluated at z'_{yc}. In other words, if $\tilde{z}_{yc}$ is any maximizer in equation 28, we have:

$$\begin{aligned} & u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} \tilde{z}_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} \tilde{z}_{yc}\right) \\ & \geq \; u_{\text{UM}} - z'_{y^*c^*} \cdot \beta_e - \lambda \cdot g\left(\Delta_{\text{UM}} - z'_{y^*c^*}\right) \\ & > \; u_{\text{UM}} - \lambda \cdot g(\Delta_{\text{UM}}) \end{aligned} \tag{29}$$

where equation 29 follows from equation 27. This in turn implies $g(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} \tilde{z}_{yc}) < g(\Delta_{\text{UM}})$. Therefore, for any optimal solution, we have $\Delta < \Delta_{\text{UM}}$, and the penalty is effective.

On the other hand, assume that $\lambda \cdot g'_-(\Delta_{\text{UM}}) \leq \beta_e$. We show that the utility-maximizing solution is optimal for the problem in equation 28. We need to show that for any arbitrary z_{yc}, the following holds:

$$\begin{aligned} & u_{\text{UM}} - \lambda \cdot g(\Delta_{\text{UM}}) \\ & \geq u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) \end{aligned}$$

Note that

$$\begin{aligned} & [u_{\text{UM}} - \lambda \cdot g(\Delta_{\text{UM}})] - \left[u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right)\right] \\ & = \lambda \left[g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) - g(\Delta_{\text{UM}})\right] + \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| \\ & \geq \lambda \left[g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right) - g(\Delta_{\text{UM}})\right] + \beta_e \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \\ & \geq -\lambda \cdot g'_-(\Delta_{\text{UM}}) \sum_{(y,c)\in\mathcal{T}_e} z_{yc} + \beta_e \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \qquad (30) \\ & = \left(\beta_e - \lambda \cdot g'_-(\Delta_{\text{UM}})\right) \sum_{(y,c)\in\mathcal{T}_e} z_{yc} \\ & \geq 0 \qquad (31) \end{aligned}$$

where equation 30 follows from the property $g(x) - g(y) \geq g'(y)(x-y)$ of convex functions, and equation 31 follows from $\sum_{(y,c)\in\mathcal{T}_e} z_{yc} \geq 0$ and the assumption $\lambda \cdot g'_-(\Delta_{\text{UM}}) \leq \beta_e$. Thus, we obtain that the utility-maximizing solution is optimal in this case.

□

A.3 Proof of Theorem 2

Proof of Theorem 2. Assume that $\beta_s \leq \lambda \cdot g'_+(0)$. Take an arbitrary optimal solution to equation 19 and denote it by z_{yc}. Note that z_{yc} is in the domain of the optimization problem in equation 15. We would like to show that z_{yc} is also a maximizer for the problem in equation 15. This will prove that the penalty is fully satisfactory, meaning there is a policy satisfying demographic parity that is also optimal.

From the identification in equation 24, the value β_s introduced in the theorem can be stated as

$$\beta_s = \max_{(c,y)\in\mathcal{T}_e : z_{yc}>0} p(c)\cdot|y|.$$

Note that Lemma 4 shows that this quantity is the same for all maximizers of the equation 19. Take any maximizer for equation 15 denoted by z'_{yc}. We need to prove that the objective function of equation 15 at z_{yc} is not smaller than that at z'_{yc}, i.e.,

$$\begin{aligned} & u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\cdot p(c)\cdot|y| - \lambda\cdot g(0) \\ & \geq u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\cdot p(c)\cdot|y| - \lambda\cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right) \end{aligned}$$

This would show that z_{yc} is a maximizer for equation 15.

First, we note that the following holds:

$$\begin{aligned} & \left[u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\cdot p(c)\cdot|y| - \lambda\cdot g(0)\right] - \left[u_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\cdot p(c)\cdot|y| - \lambda\cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right)\right] \\ & = \left[\sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\cdot p(c)\cdot|y| - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\cdot p(c)\cdot|y|\right] + \lambda\left[g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right) - g(0)\right] && (32) \\ & \geq \left[\sum_{(y,c)\in\mathcal{T}_e} z'_{yc} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right]\beta_s + \lambda\left[g\left(\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right) - g(0)\right] && (33) \\ & \geq \left[\sum_{(y,c)\in\mathcal{T}_e} z'_{yc} - \sum_{(y,c)\in\mathcal{T}_e} z_{yc}\right]\beta_s + \lambda\left[\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right] g'_+(0) && (34) \\ & = \left[\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc}\right](\lambda\cdot g'_+(0) - \beta_s) && (35) \\ & \geq 0 && (36) \end{aligned}$$

The inequality equation 34 follows from the convexity of $g(\cdot)$ and equation 36 follows from $\beta_s \leq \lambda\cdot g'_+(0)$ and $\Delta_{\text{UM}} - \sum_{(y,c)\in\mathcal{T}_e} z'_{yc} \geq 0$. The inequality equation 33 follows from

$$(z'_{yc} - z_{yc})\cdot p(c)\cdot|y| \geq (z'_{yc} - z_{yc})\cdot\beta_s, \qquad \forall (c,y)\in\mathcal{T}_e$$

which can be argued as follows: Lemma 4 shows that

$$\begin{aligned} p(c)\cdot|y| > \beta_s &\implies z_{yc} = 0, \\ p(c)\cdot|y| < \beta_s &\implies z_{yc} = p(y|c). \end{aligned}$$

Take some $(c, y) \in \mathcal{T}_e$. If $p(c) \cdot |y| > \beta_s$, we have $z_{yc} = 0$ and hence

$$z'_{yc} \cdot p(c) \cdot |y| - z_{yc} \cdot p(c) \cdot |y| \geq \left[z'_{yc} - z_{yc}\right] \beta_s$$

Next, if $p(c) \cdot |y| = \beta_s$, we have

$$z'_{yc} \cdot p(c) \cdot |y| - z_{yc} \cdot p(c) \cdot |y| = \left[z'_{yc} - z_{yc}\right] \beta_s.$$

Finally, if $p(c) \cdot |y| < \beta_s$, we have $z_{yc} = p(y|c)$ and hence $z'_{yc} - z_{yc} \leq 0$ which implies that

$$(z'_{yc} - z_{yc}) \cdot p(c) \cdot |y| \geq (z'_{yc} - z_{yc}) \cdot \beta_s.$$

This completes the proof for one direction.

Next, assume that $\beta_s > \lambda \cdot g'_+(0)$. Take an optimal solution z_{yc} to the problem equation 19. Let

$$(c^*, y^*) \in \arg\max_{y,c} \{p(c) \cdot |y| : (c, y) \in \mathcal{T}_e \wedge z_{yc} > 0\}.$$

Find $\epsilon > 0$ such that $z_{y^*c^*} > \epsilon$ and

$$\beta_s \cdot \epsilon > \lambda(g(\epsilon) - g(0)). \tag{37}$$

Define

$$z'_{yc} = \begin{cases} z_{yc} - \epsilon & \text{if } (y, c) = (y^*, c^*) \\ z_{yc} & \text{otherwise.} \end{cases}$$

The assignment z'_{yc} satisfies equation 17 and equation 18 and belongs to the domain of the optimization problem in equation 15. Note that equation 37 implies that

$$-z'_{y^*c^*} \cdot p(c^*) \cdot |y^*| - \lambda \cdot g(\epsilon) > -z_{y^*c^*} \cdot p(c^*) \cdot |y^*| - \lambda \cdot g(0).$$

Therefore,

$$\begin{aligned} & u_{\text{UM}} - \sum_{(y,c) \in \mathcal{T}_e} z'_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c) \in \mathcal{T}_e} z'_{yc}\right) \\ & > u_{\text{UM}} - \sum_{(y,c) \in \mathcal{T}_e} z_{yc} \cdot p(c) \cdot |y| - \lambda \cdot g\left(\Delta_{\text{UM}} - \sum_{(y,c) \in \mathcal{T}_e} z_{yc}\right). \end{aligned}$$

Thus, z_{yc} is not a maximizer for equation 15 and the penalty is not fully satisfactory.

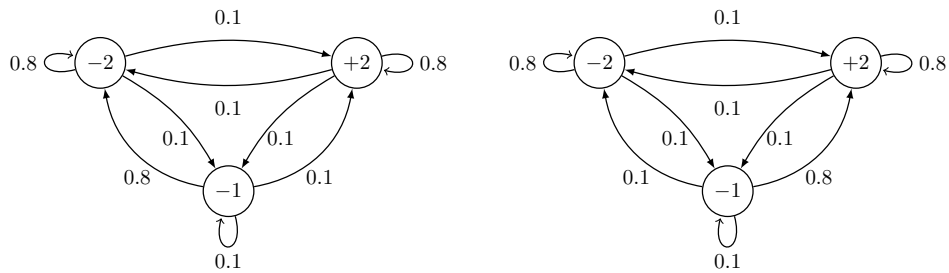
□

A.4 Proof of Theorem 3

Proof of Theorem 3. Note that when $\lambda = 0$, the optimal decision at time t will be to select an individual if and only if $Y_t > 0$ (recall the assumption that $0 \notin \mathcal{Y}$, meaning hiring any individual is strictly beneficial or harmful). [2] Consequently, for all $c \in \mathcal{C}$ and $y \in \mathcal{Y}$, we can write

$$\begin{aligned} & p_{t+1}(y|c) \\ & = \sum_{y' \in \mathcal{Y}} \sum_{d \in \{0,1\}} p_t(y'|c) \cdot \Pr(d|Y = y', C = c) \cdot q(y|y', d) \\ & = \sum_{y' < 0} p_t(y'|c) \cdot q(y|y', 0) + \sum_{y' > 0} p_t(y'|c) \cdot q(y|y', 1). \end{aligned}$$

[2] Note that the optimal decision in this case only depends on Y_t and not the distribution of $p(c, y_t)$ (as in the general case of $\lambda > 0$).



(a) $D = 1$ (b) $D = 0$

Figure 9: Transition probabilities when a) getting selected and b) when not

So, if we define

$$q^*(y|y') = \begin{cases} q(y|y', 1) & \text{if } y' > 0, \\ q(y|y', 0) & \text{if } y' < 0, \end{cases}$$

we have

$$\forall y \in \mathcal{Y}: \; p_{t+1}(y|c) = \sum_{y'} p_t(y'|c) \cdot q^*(y|y').$$

Therefore, the updates of qualifications within each group form a discrete-time, finite support, time-homogeneous Markov chain in which their transition probabilities are identical, given by $\Pr(Y_{t+1} = y|Y_t = y') = q^*(y|y')$. The assumption $q(y|y', d) > 0$ for all y, y' and d implies $q^*(y|y') > 0$ for all y and y'. This, in turn, implies that these Markov chains are aperiodic and irreducible and have an identical, unique stationary distribution, and the chain converges to it from any initial distribution. So, regardless of the initial distributions $p_{Y_0|A}$ and $p_{Y_0|B}$, we have

$$\lim_{t\to\infty} \|p_{Y_t|A} - p_{Y_t|B}\|_1 = 0.$$

Next, for the case of $\lambda > 0$, we consider the following example. This example is based on the concept of "lack of motivation" identified by Zhang et al. (2020), which refers to situations where institutional selection reduces the likelihood of qualification growth.

Example 1. *We consider three possible qualifications: -2, -1 and $+2$, and their corresponding transition probabilities $q(y|y', d)$, are illustrated in Figure 9. Notably, the difference between the two decisions lies in the transition from state -1 to other states, where being selected increases the probability of experiencing a decline to qualification -2 and decreases the probability of growth.*

We examine the groups $\mathcal{A}$ and $\mathcal{B}$ of equal size. Assuming a penalty function $g(x) = x$, we observe the impact of different levels of λ on the disparity Δ over time, as illustrated in Figure 1b. The initial distribution of the two groups is as follows: $p(Y_0 = -2|\mathcal{A}) = 0.3, p(Y_0 = -1|\mathcal{A}) = 0.1$, $p(Y_0 = +2|\mathcal{A}) = 0.6$ and $p(Y_0 = -2|\mathcal{B}) = 0.5, p(Y_0 = -1|\mathcal{B}) = 0.1$, $p(Y_0 = +2|\mathcal{B}) = 0.4$ which we can show by the vectors

$$p_{Y_0|\mathcal{A}} = [0.3, 0.1, 0.6], p_{Y_0|\mathcal{B}} = [0.5, 0.1, 0.4].$$

When λ falls within the range of 0 to 0.5 the penalty is not effective, as stated by Theorem 1, and the disparity tends to decrease over time, approaching zero. On the other hand, when λ exceeds 1, the penalty is fully satisfactory, resulting in a constant disparity of zero. However, in the interval between 0.5 and 1, the penalty is effective but prevents equalization of qualifications and also increases the selection disparity.

More precisely, assume $\lambda = 0.7$. If we index the elements of $\mathcal{Y}$ in increasing order, a stationary state of the system is as follows:

$$p_{Y|\mathcal{A}} = \left[\frac{1}{3}, \frac{1}{10}, \frac{17}{30}\right], \quad p_{Y|\mathcal{B}} = \left[\frac{17}{30}, \frac{1}{10}, \frac{1}{3}\right]. \tag{38}$$

In this stationary state, the institution's selection policy will be

$$p_{D=1|Y_t,\mathcal{A}} = [0, 0, 1], \tag{39}$$
$$p_{D=1|Y_t,\mathcal{B}} = [0, 1, 1], \tag{40}$$

which will result in $\Delta_t = \left|\frac{17}{30} - \frac{1}{10} - \frac{1}{3}\right| = \frac{4}{30} \approx 0.133....$ *Now, consider the initial state as*

$$p_{Y_0|\mathcal{A}} = [0.3, 0.1, 0.6],$$
$$p_{Y_0|\mathcal{B}} = [0.5, 0.1, 0.4].$$

In this case, the initial selection disparity equals $\Delta_0 = 0.1 < \frac{4}{30}$. *The selection policy, in this case, is again the one given in equation 39 and equation 40 for all* $t > 0$, *and the qualification distributions converge to the stationary state equation 38. Therefore,* $\lim_{t\to\infty} \Delta_t = \frac{4}{30} > \Delta_0 = 0.1$ *meaning that the selection disparity will increase over time. This is illustrated in Figure 1b for the choice of* $\lambda = 0.7$.

□

A.5 Proof of Theorem 4

Proof of Theorem 4. Fix some time instance t. Without loss of generality, assume

$$\Pr(Y_t > 0|C = \mathcal{A}) \geq \Pr(Y_t > 0|C = \mathcal{B}). \tag{41}$$

Otherwise, if the above assumption does not hold, we can just swap the names of the two groups.

Define $\alpha = \min_{y,y',d}\{q(y|y', d)\}$ and

$$f(y', d) = p_t(y'|\mathcal{A}) \cdot \Pr(D_t = d|Y_t = y', C = \mathcal{A}) - p_t(y'|\mathcal{B}) \cdot \Pr(D_t = d|Y_t = y', C = \mathcal{B}).$$

First, consider the trivial case of $\alpha = 1/|\mathcal{Y}|$, which means both groups undergo the uniform transition $q(y|y') = 1/|\mathcal{Y}|$ independent of the selection decision. In this case, for all $t > 1$, $y \in \mathcal{Y}$, and c, we have $\Pr(Y_t = y|C = c) = 1/|\mathcal{Y}|$ and $\sum_{y\in\mathcal{Y}} |p_t(y|\mathcal{A}) - p_t(y|\mathcal{B})| = 0$.

For the rest, we assume $\alpha < 1/|\mathcal{Y}|$. Using simple algebraic (in)equalities, we show that

$$\sum_{y\in\mathcal{Y}} |p_{t+1}(y|\mathcal{A}) - p_{t+1}(y|\mathcal{B})| \leq (1 - |\mathcal{Y}|\alpha) \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} |f(y', d)|\,. \tag{42}$$

Then, by utilizing Lemma 2 in the institution's policy at time t we deduce

$$\sum_{y'\in\mathcal{Y}} |f(y', 1)| \leq \sum_{y'} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| \tag{43}$$

and similarly

$$\sum_{y'\in\mathcal{Y}} |f(y', 0)| \leq \sum_{y'} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})|\,. \tag{44}$$

Combining them, we get the desired result

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} |p_{t+1}(y|\mathcal{A}) - p_{t+1}(y|\mathcal{B})| \\
&\leq (1 - |\mathcal{Y}|\alpha) \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} |f(y', d)| \\
&\leq 2(1 - |\mathcal{Y}|\alpha) \sum_{y\in\mathcal{Y}} |p_t(y|\mathcal{A}) - p_t(y|\mathcal{B})|.
\end{aligned}$$

Consequently, if $2(1-|\mathcal{Y}|\alpha) < 1$, then

$$\lim_{t\to\infty} \sum_{y\in\mathcal{Y}} |p_t(y|\mathcal{A}) - p_t(y|\mathcal{B})| = 0.$$

To show equation 42, let us denote

$$q^*(y|y',d) = \frac{q(y|y',d)-\alpha}{1-|\mathcal{Y}|\alpha}.$$

Then, we have

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} |p_{t+1}(y|\mathcal{A}) - p_{t+1}(y|\mathcal{B})| \\
&= \sum_{y\in\mathcal{Y}} \Big| \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} \big[p_t(y'|\mathcal{A})\cdot \Pr(D_t=d|Y_t=y',\mathcal{A})\cdot q(y|y',d) - p_t(y'|\mathcal{B})\cdot \Pr(D_t=d|Y_t=y',\mathcal{B})\cdot q(y|y',d)\big]\Big| && (45)\\
&= \sum_{y\in\mathcal{Y}} \left| \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} q(y|y',d) f(y',d) \right| && (46)\\
&= \sum_{y\in\mathcal{Y}} \left| \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} (1-|\mathcal{Y}|\alpha) q^*(y|y',d) f(y',d) \right| && (47)\\
&\le (1-|\mathcal{Y}|\alpha) \sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} q^*(y|y',d)\, |f(y',d)| && (48)\\
&= (1-|\mathcal{Y}|\alpha) \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} |f(y',d)| \sum_{y\in\mathcal{Y}} q^*(y|y',d) && (49)\\
&= (1-|\mathcal{Y}|\alpha) \sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} |f(y',d)| && (50)
\end{aligned}$$

where equation 47 follows from

$$\sum_{y'\in\mathcal{Y}} \sum_{d\in\{0,1\}} f(y',d) = \sum_{y'\in\mathcal{Y}} [p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})] = 0$$

and equation 50 follows from

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} q^*(y|y',d) \\
&= \sum_{y\in\mathcal{Y}} \frac{q(y|y',d)-\alpha}{1-|\mathcal{Y}|\alpha} \\
&= \frac{1}{1-|\mathcal{Y}|\alpha} - \frac{|\mathcal{Y}|\alpha}{1-|\mathcal{Y}|\alpha} \\
&= 1.
\end{aligned}$$

Next, we show equation 43. Define

$$y_{\mathcal{A}} = \max_y \left\{ y : \ \forall y' < y \quad p_t(Y_t = y'|C=\mathcal{A}) > 0 \implies \Pr(D_t = 1|Y_t = y', C=\mathcal{A}) = 0 \right\}$$

for the decision rule of the institution at time t. The 4th statement in Lemma 2 along with the assumption that $0 \notin \mathcal{Y}$ implies $y_{\mathcal{A}} > 0$. Furthermore, the lemma yields

$$
\begin{aligned}
&p_t(y_{\mathcal{A}}|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, C = \mathcal{B}) = p_t(y_{\mathcal{A}}|\mathcal{B}), \\
&\forall y > y_{\mathcal{A}}, c \in \{\mathcal{A}, \mathcal{B}\} : \\
&p_t(y|c) \cdot \Pr(D_t = 1|Y_t = y, C = c) = p_t(y|c), \\
&\forall y < y_{\mathcal{A}} : \\
&p_t(y|\mathcal{A}) \Pr(D_t = 1|Y_t = y, C = \mathcal{A}) = 0.
\end{aligned}
$$

Using these facts we can write

$$
\begin{aligned}
&\sum_{y' \in \mathcal{Y}} |f(y', 1)| \\
&= \sum_{y' \in \mathcal{Y}} \Big| p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y', \mathcal{A}) - p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', \mathcal{B}) \Big| \\
&= \sum_{y' > y_{\mathcal{A}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| + \sum_{y' < y_{\mathcal{A}}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', \mathcal{B}) \\
&\quad + |p_t(y_{\mathcal{A}}|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, \mathcal{A}) - p_t(y_{\mathcal{A}}|\mathcal{B})| \\
&\leq \sum_{y' > y_{\mathcal{A}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| + p_t(y_{\mathcal{A}}|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, \mathcal{A}) \\
&\quad + \sum_{y' < y_{\mathcal{A}}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', \mathcal{B}) + |p_t(y_{\mathcal{A}}|\mathcal{A}) - p_t(y_{\mathcal{A}}|\mathcal{B})| \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, \mathcal{A}) \qquad (51)
\end{aligned}
$$

where equation 51 follows from the triangle inequality.

On the other hand, part 3 of Lemma 2 along with the assumption in equation 41 shows that

$$
\begin{aligned}
&\sum_{y' \in \mathcal{Y}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y', C = \mathcal{A}) \\
&\geq \sum_{y' \in \mathcal{Y}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', C = \mathcal{B}).
\end{aligned}
$$

We can expand it as

$$
\begin{aligned}
&p_t(y_{\mathcal{A}}|\mathcal{B}) + \sum_{y' < y_{\mathcal{A}}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', C = \mathcal{B}) + \sum_{y' > y_{\mathcal{A}}} p_t(y'|\mathcal{B}) \\
&\leq p_t(y_{\mathcal{A}}|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, C = \mathcal{A}) + \sum_{y' > y_{\mathcal{A}}} p_t(y'|\mathcal{A})
\end{aligned}
$$

which implies

$$
\begin{aligned}
& p_t(y_{\mathcal{A}}|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, C = \mathcal{A}) \\
& \quad + \sum_{y' < y_{\mathcal{A}}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 1|Y_t = y', C = \mathcal{B}) \\
& \leq (p_t(y_{\mathcal{A}}|\mathcal{A}) - p_t(y_{\mathcal{A}}|\mathcal{B})) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, C = \mathcal{A}) \\
& \quad + \sum_{y' > y_{\mathcal{A}}} [p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})] \\
& = (p_t(y_{\mathcal{A}}|\mathcal{B}) - p_t(y_{\mathcal{A}}|\mathcal{A})) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, C = \mathcal{A}) \\
& \quad + \sum_{y' < y_{\mathcal{A}}} (p_t(y'|\mathcal{B}) - p_t(y'|\mathcal{A})) \qquad (52) \\
& \leq |p_t(y_{\mathcal{A}}|\mathcal{B}) - p_t(y_{\mathcal{A}}|\mathcal{A})| \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, C = \mathcal{A}) \\
& \quad + \sum_{y' < y_{\mathcal{A}}} |p_t(y'|\mathcal{B}) - p_t(y'|\mathcal{A})| \qquad (53)
\end{aligned}
$$

where equation 52 follows from the fact that for $c \in \mathcal{C}$ we have

$$
\begin{aligned}
& p_t(y_{\mathcal{A}}|c) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, C = c) + \sum_{y' > y_{\mathcal{A}}} p_t(y'|c) \\
& = 1 - p_t(y_{\mathcal{A}}|c) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, C = c) - \sum_{y' < y_{\mathcal{A}}} p_t(y'|c).
\end{aligned}
$$

From equation 53 and equation 51 we get

$$
\begin{aligned}
& \sum_{y' \in \mathcal{Y}} |f(y', 1)| \\
& \leq \sum_{y' > y_{\mathcal{A}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| \\
& \quad + |p_t(y_{\mathcal{A}}|\mathcal{B}) - p_t(y_{\mathcal{A}}|\mathcal{A})| \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{A}}, \mathcal{A}) \\
& \quad + \sum_{y' < y_{\mathcal{A}}} |p_t(y'|\mathcal{B}) - p_t(y'|\mathcal{A})| \\
& \quad + |p_t(y_{\mathcal{A}}|\mathcal{A}) - p_t(y_{\mathcal{A}}|\mathcal{B})| \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{A}}, \mathcal{A}) \\
& = \sum_{y'} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| .
\end{aligned}
$$

This completes the proof for equation 43.

It remains to prove equation 44, i.e., to show that

$$
\sum_{y' \in \mathcal{Y}} |f(y', 0)| \leq \sum_{y'} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| .
$$

We follow a similar approach. Define

$$
y_{\mathcal{B}} = \min_y \Big\{ y : \ \forall y' > y \quad p_t(Y_t = y'|C = \mathcal{B}) > 0 \implies \Pr(D_t = 1|Y_t = y', C = \mathcal{B}) = 1 \Big\}.
$$

The 4th statement in Lemma 2 along with the assumption that $0 \notin \mathcal{Y}$ implies $y_{\mathcal{B}} < 0$. Furthermore, the lemma yields

$$
\begin{aligned}
&p_t(y_{\mathcal{B}}|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, C = \mathcal{A}) = p_t(y_{\mathcal{B}}|\mathcal{A}), \\
&\forall y < y_{\mathcal{B}}, c \in \{\mathcal{A}, \mathcal{B}\} : \\
&p_t(y|c) \cdot \Pr(D_t = 0|Y_t = y, C = c) = p_t(y|c), \\
&\forall y > y_{\mathcal{B}} : \\
&p_t(y|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y, C = \mathcal{B}) = 0.
\end{aligned}
$$

Now, these imply:

$$
\begin{aligned}
&\sum_{y' \in \mathcal{Y}} |f(y', 0)| \\
&= \sum_{y' \in \mathcal{Y}} \Big| p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', \mathcal{A}) - p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y', \mathcal{B}) \Big| \\
&= \sum_{y' < y_{\mathcal{B}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| + \sum_{y' > y_{\mathcal{B}}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', \mathcal{A}) \\
&\quad + |p_t(y_{\mathcal{B}}|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, \mathcal{B}) - p_t(y_{\mathcal{B}}|\mathcal{A})| \\
&\leq \sum_{y' < y_{\mathcal{B}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| + p_t(y_{\mathcal{B}}|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, \mathcal{B}) \\
&\quad + \sum_{y' > y_{\mathcal{B}}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', \mathcal{A}) + |p_t(y_{\mathcal{B}}|\mathcal{B}) - p_t(y_{\mathcal{B}}|\mathcal{A})| \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, \mathcal{B}). \qquad (54)
\end{aligned}
$$

On the other hand, part 3 of Lemma 2 along with the assumption in equation 41 shows that

$$
\begin{aligned}
&\sum_{y' \in \mathcal{Y}} p_t(y'|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y', C = \mathcal{B}) \\
&\geq \sum_{y' \in \mathcal{Y}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', C = \mathcal{A}).
\end{aligned}
$$

We can expand it as

$$
\begin{aligned}
&p_t(y_{\mathcal{B}}|\mathcal{A}) + \sum_{y' > y_{\mathcal{B}}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', C = \mathcal{A}) + \sum_{y' < y_{\mathcal{B}}} p_t(y'|\mathcal{A}) \\
&\leq p_t(y_{\mathcal{B}}|\mathcal{B}) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, C = \mathcal{B}) + \sum_{y' < y_{\mathcal{B}}} p_t(y'|\mathcal{B})
\end{aligned}
$$

which implies

$$
\begin{aligned}
& p_t(y_{\mathcal{B}}|\mathcal{A}) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, C = \mathcal{B}) \\
& \quad + \sum_{y' > y_{\mathcal{B}}} p_t(y'|\mathcal{A}) \cdot \Pr(D_t = 0|Y_t = y', C = \mathcal{A}) \\
& \leq (p_t(y_{\mathcal{B}}|\mathcal{B}) - p_t(y_{\mathcal{B}}|\mathcal{A})) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, C = \mathcal{B}) \\
& \quad + \sum_{y' < y_{\mathcal{B}}} [p_t(y'|\mathcal{B}) - p_t(y'|\mathcal{A})] \\
& = (p_t(y_{\mathcal{B}}|\mathcal{A}) - p_t(y_{\mathcal{B}}|\mathcal{B})) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, C = \mathcal{B}) \\
& \quad + \sum_{y' > y_{\mathcal{B}}} (p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})) \qquad (55) \\
& \leq |p_t(y_{\mathcal{B}}|\mathcal{A}) - p_t(y_{\mathcal{B}}|\mathcal{B})| \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, C = \mathcal{B}) \\
& \quad + \sum_{y' > y_{\mathcal{B}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| \qquad (56)
\end{aligned}
$$

where equation 55 follows from the fact that for $c \in \mathcal{C}$ we have

$$
\begin{aligned}
& p_t(y_{\mathcal{B}}|c) \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, C = c) + \sum_{y' < y_{\mathcal{B}}} p_t(y'|c) \\
& = 1 - p_t(y_{\mathcal{B}}|c) \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, C = c) - \sum_{y' > y_{\mathcal{B}}} p_t(y'|c).
\end{aligned}
$$

From equation 56 and equation 54 we get

$$
\begin{aligned}
& \sum_{y' \in \mathcal{Y}} |f(y', 0)| \\
& \leq \sum_{y' < y_{\mathcal{B}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| \\
& \quad + |p_t(y_{\mathcal{B}}|\mathcal{A}) - p_t(y_{\mathcal{B}}|\mathcal{B})| \cdot \Pr(D_t = 1|Y_t = y_{\mathcal{B}}, \mathcal{B}) \\
& \quad + \sum_{y' > y_{\mathcal{B}}} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| \\
& \quad + |p_t(y_{\mathcal{B}}|\mathcal{B}) - p_t(y_{\mathcal{B}}|\mathcal{A})| \cdot \Pr(D_t = 0|Y_t = y_{\mathcal{B}}, \mathcal{A}) \\
& = \sum_{y'} |p_t(y'|\mathcal{A}) - p_t(y'|\mathcal{B})| .
\end{aligned}
$$

This concludes the proof of equation 44.

□

A.6 Proof of Theorem 5

Proof of Theorem 5. Note that

$$
\begin{aligned}
\mathbb{E}[D_t Y_{t+1}] &= \mathbb{E}_{D_t, Y_t}[\mathbb{E}[D_t Y_{t+1}|D_t, Y_t]] \\
&= p(D_t = 1)\mathbb{E}_{Y_t|D_t=1}[\mathbb{E}[Y_{t+1}|D_t = 1, Y_t]] \\
&\geq p(D_t = 1)\mathbb{E}_{Y_t|D_t=1}[Y_t] \\
&= \mathbb{E}[D_t Y_t].
\end{aligned}
$$

Thus, $\mathbb{E}[D_t Y_t] \leq \mathbb{E}[D_t Y_{t+1}]$. Now, for step $t+1$, let us consider the specific policy

$$
p(D_{t+1} = d|c, Y_{t+1} = y_{t+1}) \triangleq p(D_t = d|c, Y_{t+1} = y_{t+1}). \qquad (57)
$$

Then, observe that

$$p(D_{t+1} = d, C = c) = p(D_t = d, C = c)$$

and therefore $\Delta_t = \Delta_{t+1}$. Moreover,

$$\mathbb{E}[D_t Y_t] \leq \mathbb{E}[D_{t+1} Y_{t+1}]$$

This implies that if we use the policy equation 57, the objective function at time $t+1$, $\mathbb{E}[D_{t+1}Y_{t+1}] - \lambda \cdot g(\Delta_{t+1})$ will be greater than or equal to the objective function at time t. Consequently, if at time $t+1$, we maximize over all possible $p(d_{t+1}|c, y_{t+1})$, the maximum value of the objective function will also be higher than that at time t. In other words,

$$\max_{p(d_t|c,y_t)} \mathbb{E}[D_t Y_t] - \lambda \cdot g(\Delta_t)$$

will be a non-decreasing function of t. □

A.7 Proof of Lemma 1 and Theorem 6

Proof of Lemma 1. Assume the joint distribution $p(c, y_t)$ is such that

$$p(Y_t = y|C = \mathcal{A}) = p(Y_t = y|C = \mathcal{B}), \qquad \forall y \in \mathcal{Y}. \tag{58}$$

In particular, this implies that $\Pr(Y_t > 0|C = \mathcal{A}) = \Pr(Y_t > 0|C = \mathcal{B})$ and there is no tension between maximizing $\mathbb{E}[D_t Y_t]$ and minimizing Δ_t. In this case, the optimal policy is the utility-maximizing policy and $D_t = \mathbf{1}[Y_t > 0]$ (see the beginning of Section 2 for a discussion of the utility-maximizing policy). As discussed in proof of Theorem 4, the evolution of qualifications within each group follows the same Markov chain with the transition matrix T given in the statement of the lemma. Consequently, equation 58 implies that

$$p(Y_{t+1} = y|C = \mathcal{A}) = p(Y_{t+1} = y|C = \mathcal{B}), \qquad \forall y \in \mathcal{Y}.$$

If we take a stationary distribution r_Y of the Markov chain T and assume that the initial qualifications within each group have distribution r_Y:

$$p(Y_0 = y|C = \mathcal{A}) = p(Y_0 = y|C = \mathcal{B}) = r_Y(y),$$

we get a stationary state of our dynamic, i.e., for all $t \geq 0$ we will have

$$p(Y_t = y|C = \mathcal{A}) = p(Y_t = y|C = \mathcal{B}) = r_Y(y).$$

□

Proof of Theorem 6. First, note that in a stationary state, if $\Pr(Y > 0|\mathcal{A}) = \Pr(Y > 0|\mathcal{B})$, then, under any discrimination penalty, the decision policy for both groups is the same: selecting all positively qualified individuals. Therefore, the transition probabilities of the groups and the stationary state will be the ones described in Lemma 1.

So, assume there exists a stationary state with societal disparity. Without loss of generality, assume that $\mathcal{A}$ is the advantaged group in the stationary state:

$$\Pr(Y > 0|\mathcal{A}) > \Pr(Y > 0|\mathcal{B}). \tag{59}$$

Denote the qualification distribution of the groups in this stationary state by

$$\pi_{\mathcal{A}}(y) = p(Y = y|C = \mathcal{A})$$

and

$$\pi_{\mathcal{B}}(y) = p(Y = y|C = \mathcal{B})$$

and the corresponding decision policy $p^*(d|y,c)$. We denote the transition matrices in this state for groups $\mathcal{A}$ and $\mathcal{B}$ by $T_{\mathcal{A}}$ and $T_{\mathcal{B}}$ respectively, which can be computed as follows:

$$\begin{aligned}
T_{\mathcal{A}}(i,j) =& q(y^{(j)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{A}) \\
&+ q(y^{(j)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{A}), \\
T_{\mathcal{B}}(i,j) =& q(y^{(j)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{B}) \\
&+ q(y^{(j)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{B}).
\end{aligned}$$

Note that the assumption equation 10 implies that $T_{\mathcal{A}}$ and $T_{\mathcal{B}}$ are aperiodic and irreducible.

The last part of Lemma 2 and the assumptions equation 11 and equation 12 imply

$$\begin{aligned}
&y^{(i)} < 0, i < n: \\
&\quad T_{\mathcal{A}}(i,i+1) = q(y^{(i+1)}|y^{(i)},0) \\
&\qquad\qquad \leq\ q(y^{(i+1)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{B}) + q(y^{(i+1)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{B}) \\
&\qquad\qquad = T_{\mathcal{B}}(i,i+1) \qquad (60) \\
&y^{(i)} < 0, i > 0: \\
&\quad T_{\mathcal{A}}(i,i-1) = q(y^{(i-1)}|y^{(i)},0) \\
&\qquad\qquad \geq\ q(y^{(i-1)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{B}) + q(y^{(i-1)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{B}) \\
&\qquad\qquad = T_{\mathcal{B}}(i,i-1) \qquad (61) \\
&y^{(i)} > 0, i < n: \\
&\quad T_{\mathcal{A}}(i,i+1) = q(y^{(i+1)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{A}) + q(y^{(i+1)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{A}) \\
&\qquad\qquad \leq\ q(y^{(i+1)}|y^{(i)},1) \\
&\qquad\qquad = T_{\mathcal{B}}(i,i+1) \qquad (62) \\
&y^{(i)} > 0, i > 0: \\
&\quad T_{\mathcal{A}}(i,i-1) = q(y^{(i-1)}|y^{(i)},1)\cdot p^*(D=1|Y=y^{(i)},C=\mathcal{A}) + q(y^{(i-1)}|y^{(i)},0)\cdot p^*(D=0|Y=y^{(i)},C=\mathcal{A}) \\
&\qquad\qquad \geq\ q(y^{(i-1)}|y^{(i)},1) \\
&\qquad\qquad = T_{\mathcal{B}}(i,i-1) \qquad (63)
\end{aligned}$$

Therefore, Lemma 5 (given below) yields

$$\sum_{y^{(i)}>0} \pi_{\mathcal{A}}(y^{(i)}) \leq \sum_{y^{(i)}>0} \pi_{\mathcal{B}}(y^{(i)}),$$

or $\Pr(Y>0|\mathcal{A}) \leq \Pr(Y>0|\mathcal{B})$. But this contradicts the assumption in equation 59 that $\mathcal{A}$ is the advantaged group. □

Lemma 5. *Consider two (discrete-time) Markov process $\{A_t\}$ and $\{B_t\}$ sharing the state support $\mathcal{Y} = \{y^{(1)},\cdots,y^{(n)}\}$ such that $y^{(1)} \leq \cdots \leq y^{(n)}$. Assume these two processes are identified by transition matrices P and Q such that for $i,j \in \{1,2,\cdots,n\}$ we have*

$$P_{i,j} = \Pr(A_t = y^{(j)}|A_{t-1} = y^{(i)}),$$

$$Q_{i,j} = \Pr(B_t = y^{(j)}|B_{t-1} = y^{(i)}).$$

Assume that A_t and B_t are birth-death Markov processes and their transition matrices satisfy

$$\begin{aligned}
|i-j| \leq 1 \iff P_{i,j} > 0, \\
|i-j| \leq 1 \iff Q_{i,j} > 0.
\end{aligned}$$

Assume P and Q satisfy the following for every $1 \leq k \leq n-1$:

$$P_{k,k+1} \geq Q_{k,k+1}, \tag{64}$$
$$P_{k+1,k} \leq Q_{k+1,k}. \tag{65}$$

Then, for every s

$$\lim_{t\to\infty} \Pr(A_t \geq s) \geq \lim_{t\to\infty} \Pr(B_t \geq s).$$

and

$$\lim_{t\to\infty} \Pr(A_t \leq s) \leq \lim_{t\to\infty} \Pr(B_t \leq s).$$

In other words, the stationary distribution of A_t stochastically dominates the stationary distribution of B_t.

Proof. Let $\alpha_0 = \beta_0 = 1$ and

$$\alpha_k = \frac{P_{k,k+1}}{P_{k+1,k}}, \qquad k \in \{1, 2, \cdots, n-1\}.$$

$$\beta_k = \frac{Q_{k,k+1}}{Q_{k+1,k}}, \qquad k \in \{1, 2, \cdots, n-1\}.$$

The assumptions in equation 64 and equation 65 imply that $\alpha_k \geq \beta_k$ for all k. Since the birth-death Markov chain is aperiodic and irreducible, the stationary distribution is unique and given by Gallager (1996)

$$\lim_{t\to\infty} \Pr(A_t = i) = \frac{\alpha_0\alpha_1\alpha_2\cdots\alpha_{i-1}}{\sum_{j=1}^{n} \alpha_0\alpha_1\alpha_2\cdots\alpha_{j-1}},$$
$$\lim_{t\to\infty} \Pr(B_t = i) = \frac{\beta_0\beta_1\beta_2\cdots\beta_{i-1}}{\sum_{j=1}^{n} \beta_0\beta_1\beta_2\cdots\beta_{j-1}}$$

where $i = 1, \cdots, n$. Let

$$g(x_0, x_1, \cdots, x_{n-1}) = \frac{\sum_{i=s}^{n} x_0x_1x_2\cdots x_{i-1}}{\sum_{j=1}^{n} x_0x_1x_2\cdots x_{j-1}}.$$

Then,

$$\lim_{t\to\infty} \Pr(A_t \geq s) = g(\alpha_0, \alpha_1, \cdots, \alpha_{n-1})$$
$$\lim_{t\to\infty} \Pr(B_t \geq s) = g(\beta_0, \beta_1, \cdots, \beta_{n-1})$$

Since $g(x_0, x_1, \cdots, x_{n-1})$ is increasing in x_k for every k and using the fact that $\alpha_k \geq \beta_k$ for all k, we get the desired result.

□