LONG TERM IMPACT OF FAIR MACHINE LEARNING IN SEQUENTIAL DECISION MAKING: REPRESENTATION DISPARITY AND GROUP RETENTION

Xueru Zhang*
Department of Electrical and Computer Engineering
University of Michigan
xueru@umich.edu

Mohammad Mahdi Khalili*
Department of Electrical and Computer Engineering
University of Michigan
khalili@umich.edu

Cem Tekin
Department of Electrical and Electronics Engineering
Bilkent University
cemtekin@ee.bilkent.edu.tr

Mingyan Liu
Department of Electrical and Computer Engineering
University of Michigan
mingyan@umich.edu

ABSTRACT

Machine learning models trained on data from multiple demographic groups can inherit representation disparity [1] that may exist in the data: the group contributing less to the training process may suffer higher loss in model accuracy; this in turn can degrade population retention in these groups over time in terms of their contribution to the training process of future models, which then exacerbates representation disparity in the long run. In this study, we seek to understand the interplay between the model accuracy and the underlying group representation and how they evolve in a sequential decision setting over an infinite horizon, and how the use of fair machine learning plays a role in this process. Using a simple user dynamics (arrival and departure) model, we characterize the long-term property of using machine learning models under a set of fairness criteria imposed on each stage of the decision process, including the commonly used *statistical parity* and *equal opportunity* fairness. We show that under this particular arrival/departure model, both these criteria cause the representation disparity to worsen over time, resulting in groups diminishing entirely from the sample pool, while the criterion of *equalized loss* fares much better. Our results serve to highlight the fact that fairness cannot be defined outside the larger feedback loop where past actions taken by users (who are either subject to the decisions made by the algorithm or whose data are used to train the algorithm or both) will determine future observations and decisions.

1 Introduction

Most of modern machine learning techniques are developed to capture features of the data sampled from a underlying population in the training process, e.g., by minimizing overall loss over the training dataset. The premise that the data is truly representative of the underlying population [2], however, is generally incorrect, and the training dataset may lead to inaccurate machine learning models for a number of reasons such as the following.

- Bias in data samples: Certain groups (e.g., a majority group within a population) may have a (disproportionately) higher representation in the training data as compared to the other groups (e.g., a minority group). The resulting machine learning models could be biased in favor of the groups having more representation in the training dataset to the detriment of the minority group. For instance, speech recognition products recognize native speakers much better than non-native speakers [3].
- Incomplete data: It is possible that some features, which are critical to training the model, are difficult to acquire due to privacy concerns; this results in the use of sensitive/group attributes as proxies in the training which can lead to discrimination. For instance, individuals may not be willing to share personal information [4], and group attributes such as race, gender, etc. are then used as proxies.

*Equal Contribution

When machine learning algorithms inform decision making, such inaccuracies can lead to inferior and/or unfair decisions. In particular, the standard objective of minimizing overall loss allows the resulting algorithms to inherit the bias that existed in the training data, leading to high loss (and unfair decisions) for certain groups, often referred to as *disadvantaged group(s)/ minority group(s)*; this can happen in various applications, see e.g., [5, 6, 7].

Moreover, the problem does not merely stop here. Often times, decisions informed by machine learning algorithms inevitably affect future data samples, which will be collected and used to train future generations of machine learning algorithms. This closed, feedback loop becomes self-reinforcing and can lead to highly undesirable outcomes over time. A prime example is the use of credit history in lending practices. Lending models are trained using credit history (data samples), typically consisting of debt (e.g., credit card and loan) payment information. Individuals with no or short credit history are more likely to be assigned low credit ratings and as a result disqualify for a loan (decision). However, a loan rejection denies the individual the opportunity to improve his/her credit history (future data samples) and can lead to future rejections (perpetuation of the negative outcome), a phenomenon known as the *data collection circular dependency*.

To address the various fairness issues highlighted above, researchers have studied two categories of approaches. The first focused on improving data pre-processing [8, 9]. For instance, Calders *et al.* showed that by removing certain features, label modification, reweighing, and sampling, one can improve the fairness of the trained classifier, while Kamiran *et al.* proposed a method to determine a randomized mapping to transform the original dataset into a new dataset, which, when used to train a classifier, can mitigate the discrimination issue.

The second approach centers on imposing fairness constraints, in addition to minimizing loss, when training an algorithm, as a way to battle biased outcomes. A variety of fairness notions have been proposed in the literature [10]; some of these can be conflicting [11]. We review some of the most relevant studies below.

Equalized odds and equal opportunity are two criteria proposed in [12], in the context of a lending decision problem; it showed that these constraints make the outcome fairer for difference races without necessarily hurting the lender's profit. Liu *et al.* [13] considered statistical parity and equal opportunity, and showed that when the likelihood of default and its impact on a borrower's future credit is factored in, these fairness criteria may not improve the utility of the minority group and may only work under certain conditions. Both [12, 13] are based on a one-shot formulation, while Hashimoto *et al.* [1] considered a sequential decision formulation where individuals/users (of the machine learning algorithm) may choose to leave the system based on the loss/error rate they experience. To prevent this attrition, this study adopted the objective of minimizing the loss of the group with the highest loss (instead of overall or average loss); it showed that this protection of the minority group can lead to a more equitable equilibrium.

In the present study, we are interested in the impact of fairness criteria on the group representation in a sequential decision framework. We adopt a simple user dynamics (arrival and departure) model same as that proposed by [1]. However, instead of minimizing loss for the most disadvantaged group, we take the objective of minimizing the average total loss over an infinite horizon subject to an instantaneous fairness criterion (imposed and satisfied at each time step of the problem).

The goal is to understand the interplay between the model accuracy and the underlying group representation, how they evolve in a sequential setting over the long run, and how imposing fairness criterion plays a role in this process.

Toward this end, we first characterize the long-term properties induced by the use of a set of fairness criteria in the decision process, including *statistical parity* (`StatPar`) and *equal opportunity* (`EqOpt`). We show that under this particular dynamic model, both criteria can worsen the representation disparity over time, resulting in the disadvantaged group diminishing entirely from the sample pool. By contrast, the criterion of *equalized loss* (`EqLos`) is shown to be able to sustain stable group representation over time under this dynamics. Furthermore, greedy decisions in each stage subject to the equalized loss fairness criterion is shown to be optimal for the infinite horizon loss objective.

It is worth noting that equalized loss has been considered in an online learning setting [14] where data arrives sequentially and the learner wants to find the best predictor among a given set, but the predictor accuracy does not affect user participation or data arrival.

Our results also demonstrate that there can easily be a mismatch between a particular fairness definition and the underlying factors driving the system dynamics, and that this mismatch exacerbates representation disparity in the long run. So in essence, fairness cannot be defined in a one-shot problem setting without considering the long-run impact, and that long-run impact cannot be properly analyzed without understanding the underlying dynamics.

The remainder of this paper is organized as follows. Section 2 formulates the problem. The analysis of sequential problem under different fairness criteria and the equalized loss fairness are presented in Section 3. Discussions are given in Section 4 and experiments in Section 5. Section 6 concludes the paper.

2 Problem Formulation

Consider two demographic groups G_a and G_b, typically distinguished based on some sensitive/protected attributes (e.g., gender, race). An individual from either group has feature $X \in \mathbb{R}$ and label $Y \in \{0,1\}$. Denote by $G_k^j \subset G_k$ the subgroup with label j, $j \in \{0,1\}$, $k \in \{a,b\}$, $f_k^j(x)$ its feature distribution and α_k^j the size of G_k^j as a fraction of the entire population. Then $\overline{\alpha}_k := \alpha_k^0 + \alpha_k^1$ is the size of G_k as a fraction of the population. Denote by $g_k^j = \alpha_k^j/\overline{\alpha}_k$ the fraction of label $j \in \{0,1\}$ in group k. Thus the difference between $\overline{\alpha}_a$ and $\overline{\alpha}_b$ measures the representation disparity between the two groups.

The distribution of X over G_k is given by $f_k(x) = g_k^1 f_k^1(x) + g_k^0 f_k^0(x)$, and the distribution over the entire population is $f(x) = \sum_{k\in\{a,b\}, j\in\{0,1\}} \alpha_k^j f_k^j(x)$.

Consider a binary classification problem based on feature X. Let $h_\theta(x) = \mathbf{1}(x \geq \theta)$ be a decision rule parametrized by $\theta \in \mathbb{R}$ and $L(y, h_\theta(x)) = \mathbf{1}(y \neq h_\theta(x))$ the 0-1 loss. Without loss of generality, let θ_k, $k \in \{a,b\}$, be the decision parameter for group k, with a corresponding expected loss of $L_k(\theta_k)$. Then the total expected loss given θ_a and θ_b is:

$$L(\theta_a, \theta_b) = \overline{\alpha}_a L_a(\theta_a) + \overline{\alpha}_b L_b(\theta_b) ,$$

where $L_k(\theta_k) = g_k^1 \int_{-\infty}^{\theta_k} f_k^1(x)dx + g_k^0 \int_{\theta_k}^{\infty} f_k^0(x)dx$.

In practice, the distributions $f_k^j(x)$ are not directly observable, and the loss expressed above is either approximated or empirically estimated, e.g., from training or testing datasets. Note that if feature distributions of the two groups are identical, then a single parameter value $\theta_a = \theta_b = \theta$ suffices. However, as discussed earlier, bias in the training often originates from using (or being limited to using) feature X that has different distribution over different groups; subsequently the notion of fairness often hinges on applying different decisions rules for different groups. We will assume the two groups have different distributions: $f_a \neq f_b$.

Within this context, the fair machine learning problem commonly studied in the literature is the one-shot decision problem stated as follows; this type of formulation has been used in a variety of applications, such as lending [15] and hiring [16], etc.

Definition 1. *(One-shot problem)*

$$\min_{\theta_a, \theta_b} \quad L(\theta_a, \theta_b) \qquad (\star)$$
$$\textit{s.t.} \quad \Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0 ,$$

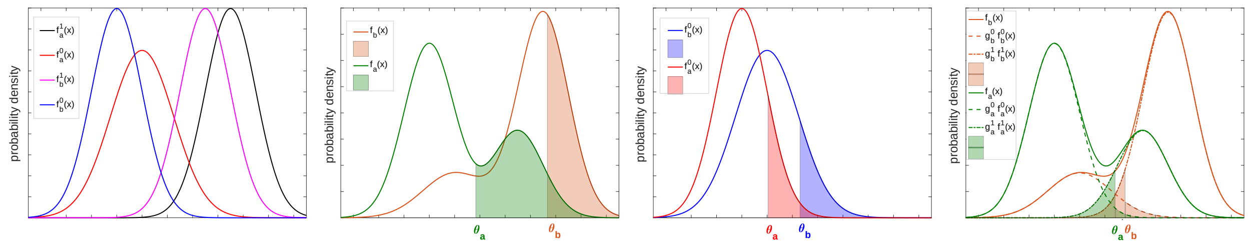
where $\Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0$ *denotes a given fairness constraint.*

The resulting solution (or decision) will be referred to as the one-shot fair decision or the *greedy fair decision* for a given $\Gamma_{\mathcal{C}}$, as the optimality only holds for a single time step.

In this paper we will study four types of fairness criteria $\Gamma_{\mathcal{C}}$, including some very commonly used. These are described as follows and illustrated in Figure 1.

1. Simple fair (`Simple`): $\Gamma_{\mathcal{C}} = \theta_a - \theta_b$. Imposing this criterion simply means we ensure the same decision parameter is used for both groups.
2. Equal opportunity (`EqOpt`): This requires the false positive rate (FPR) be the same for different groups (Fig. 1(c))[2], i.e., $\Pr(h_{\theta_a}(X) = 1 | X \in G_a^0) = \Pr(h_{\theta_b}(X) = 1 | X \in G_b^0)$. This is equivalent to: $\int_{\theta_a}^{\infty} f_a^0(x)dx = \int_{\theta_b}^{\infty} f_b^0(x)dx$. Thus the `EqOpt` fairness constraint can be written as:
$$\Gamma_{\mathcal{C}} = \int_{\theta_a}^{\infty} f_a^0(x)dx - \int_{\theta_b}^{\infty} f_b^0(x)dx.$$
3. Statistical parity (`StatPar`): This requires different groups be given equal probability of being labelled 1 (Fig. 1(b)), i.e., $\Pr(h_{\theta_a}(X) = 1 | X \in G_a) = \Pr(h_{\theta_b}(X) = 1 | X \in G_b)$. This is equivalent to: $\sum_{i\in\{0,1\}} g_a^i \int_{\theta_a}^{\infty} f_a^i(x)dx = \sum_{i\in\{0,1\}} g_b^i \int_{\theta_b}^{\infty} f_b^i(x)dx$. Thus the `StatPar` fairness constraint can be written as:
$$\Gamma_{\mathcal{C}} = \int_{\theta_a}^{\infty} f_a(x)dx - \int_{\theta_b}^{\infty} f_b(x)dx.$$

[2] Depending on the context, this criterion also refers to equal false negative rate (FNR) or true positive rate (TPR) or true negative rate (TNR), but the analysis is essentially the same.



(a) Density function for each G_k^j (b) Statistical parity (c) Equal opportunity (d) Equalized Loss

Figure 1: For G_a, G_b with subgroup proportions $\alpha_a^1 = 0.55,\ \alpha_a^0 = 0.15,\ \alpha_b^1 = 0.1,\ \alpha_b^0 = 0.2$, a pair of (θ_a, θ_b) is fair under each criterion stated in Figure 1(b)-1(d) requires the corresponding colored areas be equal.

4. Equalized loss (`EqLos`): This requires that the expected loss across different groups to be equal. Thus the `EqLos` constraint can be written as:

$$\Gamma_{\mathcal{C}} = L_a(\theta_a) - L_b(\theta_b).$$

Notice that for `Simple`, `EqOpt` and `StatPar` criteria, any (θ_a, θ_b) and (θ'_a, θ'_b) that satisfy $\Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0$ and $\Gamma_{\mathcal{C}}(\theta'_a, \theta'_b) = 0$, $\theta_a \geq \theta'_a$ if and only if $\theta_b \geq \theta'_b$.

In this study, we are interested in what happens in the long run when these fairness criteria are used in a sequential decision setting. Specifically, if loss impacts user participation, then current loss will impact future loss. In other words, past actions determine the system state which then determines future actions. To do so, we will study the following constrained optimization aimed at minimizing the average total loss over an infinite horizon.

Definition 2. *(Sequential problem)*

$$\min_{\{(\theta_a(t),\theta_b(t))\}_{t=1}^{\infty}} \quad \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} L^t(\theta_a(t), \theta_b(t)) \qquad (\star\star)$$
$$\textit{s.t.} \quad \Gamma_{\mathcal{C}}(\theta_a(t), \theta_b(t)) = 0,\ \forall t$$

where $(\theta_a(t), \theta_b(t))$ is the decision parameters for time step t, and the step loss:

$$L^t(\theta_a(t), \theta_b(t)) = \overline{\alpha}_a(t)\cdot L_a(\theta_a(t)) + \overline{\alpha}_b(t)\cdot L_b(\theta_b(t)).$$

How current decisions impact the future is driven by the following group/user retention and attrition dynamics. Denote by $N_k(t)$ the expected number of users from group k at time t:

$$N_k(t+1) = N_k(t)\cdot \nu(L_k(\theta_k(t))) + \beta_k\ , \qquad (1)$$

where $\nu(L_k(\theta_k(t)))$ is the probability of a user from group k who was in the system at time t remaining in the system at time $t+1$, and β_k is the expected number of exogenous arrivals to group k. For simplicity, here β_k is treated as a constant, but as we discuss in Section 4, our analysis and qualitative conclusions hold even when this is given by a random process provided it is independent of the decision or current state of the system. We will model $\nu(\cdot) : [0,1] \to [0,1]$ as a strictly decreasing function to capture the fact that users are less willing to stay in the system when subject to higher (perceived) loss. Accordingly, the relative group representation for time step $t+1$ is update as

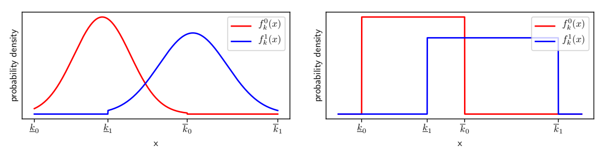
$$\overline{\alpha}_k(t+1) = \frac{N_k(t+1)}{N_a(t+1) + N_b(t+1)}.$$

Dynamics model (1) suggests that user departure is driven by model accuracy, i.e., whether a classification is made correctly (including both false negative and false positive). This applies to domains such as biometric authentication, speaker verification, and medical diagnosis.

To find the optimal sequence $(\theta_a(t), \theta_b(t))$ and solve $(\star\star)$, we need to know the user arrival and departure dynamics, i.e., observing or accurately estimating $\nu(\cdot)$ and β_k, or alternatively $N_k(t)$, which may or may not be realistic. In the next section we will take these quantities as known or observable and focus on understanding the long term property of the decisions under different fairness constraints. We will discuss the case of incomplete information in Section 4.

We end this section with a few assumptions, mostly for simplicity of exposition. We will assume that the distributions $f_a^0(x), f_a^1(x), f_b^0(x), f_b^1(x)$ have bounded support on $[\underline{a}_0, \overline{a}_0]$, $[\underline{a}_1, \overline{a}_1]$, $[\underline{b}_0, \overline{b}_0]$ and $[\underline{b}_1, \overline{b}_1]$ respectively, and that $f_k^1(x)$ and $f_k^0(x)$ overlap, i.e., $\underline{a}_0 < \underline{a}_1 < \overline{a}_0 < \overline{a}_1$ and $\underline{b}_0 < \underline{b}_1 < \overline{b}_0 < \overline{b}_1$, though this assumption does not affect our main results; two examples are shown in Figure 2. Our main technical assumption is stated in the following, although it is not required in some of our results.

Assumption 1. *Let $\mathcal{T}_a = [\underline{a}_1, \overline{a}_0]$ (resp. $\mathcal{T}_b = [\underline{b}_1, \overline{b}_0]$) be the overlapping interval between $f_a^0(x)$ and $f_a^1(x)$ (resp. $f_b^0(x)$ and $f_b^1(x)$). The distribution $f_k^1(x)$ is strictly increasing and $f_k^0(x)$ is strictly decreasing over $\mathcal{T}_k$ for $k \in \{a, b\}$.*



(a) Strict monotonic over $\mathcal{T}_k$ (b) Uniformly distributed

Figure 2: Model illustration: feature distributions of G_a, G_b

For bell-shaped feature distributions (e.g., normal, Cauchy, etc.), Assumption 1 implies that $f_k^1(x)$ and $f_k^0(x)$ are sufficiently separated. An example is shown in Fig. 2(a). The example shown in Fig. 2(b) violates the above assumption, although this case is also amenable to analysis with similar results as we discuss in Section 4.

3 Analysis of the Sequential Problem

In this section we examine the sequential problem under different fairness criteria. We begin by solving the one-shot problem under each of the first three fairness constraints. We then characterize what happens in the long run when these one-shot decisions are applied in each step, i.e., using a greedy policy for the infinite horizon problem. The last subsection studies the `EqLos` fairness and its long-term property.

3.1 Greedy, one-shot solutions

Below we find the solution to the one-shot optimization problem ($\star$) for `Simple`, `EqOpt`, and `StatPar` fairness, respectively.

The loss for group a can be written as

$$L_a(\theta_a) = \int_{-\infty}^{\theta_a} g_a^1 f_a^1(x)dx + \int_{\theta_a}^{\infty} g_a^0 f_a^0(x)dx = \begin{cases} \int_{\theta_a}^{\overline{a}_0} g_a^0 f_a^0(x)dx, \text{ if } \theta_a \in [\underline{a}_0, \underline{a}_1] \\ \int_{\theta_a}^{\overline{a}_0} g_a^0 f_a^0(x)dx + \int_{\underline{a}_1}^{\theta_a} g_a^1 f_a^1(x)dx, \text{ if } \theta_a \in [\underline{a}_1, \overline{a}_0] \\ \int_{\underline{a}_1}^{\theta_a} g_a^1 f_a^1(x)dx, \text{ if } \theta_a \in [\overline{a}_0, \overline{a}_1] \end{cases},$$

which is decreasing in θ_a over $[\underline{a}_0, \underline{a}_1]$ and increasing over $[\overline{a}_0, \overline{a}_1]$. We have

$$\theta_a^* = \arg\min_{\theta_a} L_a(\theta_a) = \begin{cases} \underline{a}_1, \text{ if } g_a^1 f_a^1(\underline{a}_1) \geq g_a^0 f_a^0(\underline{a}_1) \\ \delta_a, \text{ if } g_a^1 f_a^1(\underline{a}_1) < g_a^0 f_a^0(\underline{a}_1) \ \& \ g_a^1 f_a^1(\overline{a}_0) > g_a^0 f_a^0(\overline{a}_0) \\ \overline{a}_0, \text{ if } g_a^1 f_a^1(\overline{a}_0) \leq g_a^0 f_a^0(\overline{a}_0) \end{cases},$$

where $\delta_a \in [\underline{a}_1, \overline{a}_0]$ is such that $\frac{dL_a(x)}{dx}|_{x=\delta_a} = g_a^1 f_a^1(\delta_a) - g_a^0 f_a^0(\delta_a) = 0$. It is easy to verify that the above three branches cover all cases under Assumption 1. For G_b, $\theta_b^* = \arg\min_{\theta_b} L_b(\theta_b) \in \{\underline{b}_1, \delta_b, \overline{b}_0\}$ can be found similarly.

Below we will focus on the case when $\theta_a^* = \delta_a$ and $\theta_b^* = \delta_b$, while noting that the results for the other cases are essentially the same.

For `Simple`, `StatPar` and `EqOpt` fairness, $\exists$ a strictly increasing function ϕ, such that $\Gamma_{\mathcal{C}}(\phi(\theta_b), \theta_b) = 0$. Denote by ϕ^{-1} the inverse of ϕ. Without loss of generality, we will assign group labels a and b such that $\phi(\delta_b) < \delta_a$ and $\phi^{-1}(\delta_a) > \delta_b$.

Lemma 1. *For `Simple`, `EqOpt` and `StatPar` fairness constraints, the one-shot fair solution to problem ($\star$) is given by $(\theta_a^*, \theta_b^*) = \arg\min_{\theta_a, \theta_b} L(\theta_a, \theta_b) \in \{(\theta_a, \theta_b) | \theta_a \in [\phi(\delta_b), \delta_a], \theta_b \in [\delta_b, \phi^{-1}(\delta_a)], \Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0\}$ regardless of group proportions $(\overline{\alpha}_a, \overline{\alpha}_b)$.*

Lemma 1 shows that although there may be many solutions to the optimization ($\star$) given different group proportions, these solutions are all bounded by the same compact intervals. The following Theorem 1 describes the conditions a solution should satisfy.

Theorem 1. *Consider the one-shot problem* ($\star$) *at some time step t, with group proportions given by* $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. *Under Assumption 1, let* $(\theta_a(t), \theta_b(t))$ *be the one-shot solution under either* `Simple`, `EqOpt` *or* `StatPar` *fairness at time step t, then* $(\theta_a(t), \theta_b(t))$ *is unique and satisfies the following:*

$$\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} = \Psi_{\mathcal{C}}(\theta_a(t), \theta_b(t)), \tag{2}$$

where $\Psi_{\mathcal{C}}$ *is a function increasing in* $\theta_a(t)$ *and* $\theta_b(t)$*, which is different under different fairness criterion* $\Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0$.

See Lemma 3 in Appendix C for details about each $\Psi_{\mathcal{C}}$. Theorem 1 illustrates the impact of $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ on the one-shot solution $(\theta_a(t), \theta_b(t))$. $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ can be used as a measure of representation disparity, the population size of two groups are more similar if it is closer to 1. The following Lemma further states the impact of this one-shot solution on the losses experienced by two groups.

Lemma 2. *Suppose* (θ_a, θ_b) *and* (θ'_a, θ'_b) *are two pairs of decisions imposed on two groups* G_a, G_b *which satisfy either* `Simple`, `EqOpt` *or* `StatPar` *fairness. If* $\theta_a \geq \theta'_a$ *and* $\theta_b \geq \theta'_b$*, then the consequent expected loss satisfy* $L_a(\theta_a) \leq L_a(\theta'_a)$ *and* $L_b(\theta_b) \geq L_b(\theta'_b)$.

Theorem 1 and Lemma 2 immediately suggest a feedback loop between one-shot decisions $(\theta_a(t), \theta_b(t))$ and the representation disparity $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$: the latter drives the former that results in certain differences in the losses experienced by the two groups, which then effects the user retention rates in two groups and drives future representation. Next we will characterize this feedback loop and understand the long-term property of applying this one-shot solutions.

3.2 Long-term property of applying one-shot solutions

Next we will characterize what will happen in the long run if applying the one-shot solution at each time step. The convergence of $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ and $(\theta_a(t), \theta_b(t))$ are concluded in Theorem 2 below.

Theorem 2. *Let* $(\theta_a(t), \theta_b(t))$ *be the solution to one-shot problem* ($\star$) *under either* `Simple`, `EqOpt` *or* `StatPar` *fairness at time step t, with group ratios* $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and expected loss* $L_a(\theta_a(t))$, $L_b(\theta_b(t))$*. If the initial states* $N_a(1), N_b(1)$ *satisfy* $\frac{N_a(1)}{N_b(1)} = \frac{\beta_a}{\beta_b}$ *and* $N_k(2) > N_k(1), k \in \{a, b\}$[3]*, then the following holds under dynamics* (1) *and Assumption 1:*

(1) If $L_a(\theta_a(1)) > L_b(\theta_b(1))$*, then* $\frac{\overline{\alpha}_a(t+1)}{\overline{\alpha}_b(t+1)} < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and* $L_a(\theta_a(t+1)) > L_a(\theta_a(t)) > L_b(\theta_b(t)) > L_b(\theta_b(t+1))$ *will hold* $\forall t$*. Moreover,* $\theta_a(t+1) < \theta_a(t)$ *and* $\theta_b(t+1) < \theta_b(t)$ *hold* $\forall t$ *and* $(\theta_a(t), \theta_b(t))$ *will converge to a constant decision* $(\theta_a^\infty, \theta_b^\infty)$ *as* $t \to \infty$*, where* $\theta_a(1) > \theta_a^\infty \geq \phi(\delta_b)$ *and* $\theta_b(1) > \theta_b^\infty \geq \delta_b$.

(2) If $L_a(\theta_a(1)) < L_b(\theta_b(1))$*, then* $\frac{\overline{\alpha}_a(t+1)}{\overline{\alpha}_b(t+1)} > \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and* $L_a(\theta_a(t+1)) < L_a(\theta_a(t)) < L_b(\theta_b(t)) < L_b(\theta_b(t+1))$ *will hold* $\forall t$*. Moreover,* $\theta_a(t+1) > \theta_a(t)$ *and* $\theta_b(t+1) > \theta_b(t)$ *hold* $\forall t$ *and* $(\theta_a(t), \theta_b(t))$ *will converge to a constant decision* $(\theta_a^\infty, \theta_b^\infty)$ *as* $t \to \infty$*, where* $\theta_a(1) < \theta_a^\infty \leq \delta_a$ *and* $\theta_b(1) < \theta_b^\infty \leq \phi^{-1}(\delta_a)$.

(3) If $L_a(\theta_a(1)) = L_b(\theta_b(1))$*, then* $\frac{\overline{\alpha}_a(t+1)}{\overline{\alpha}_b(t+1)} = \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and* $L_a(\theta_a(t+1)) = L_a(\theta_a(t)) = L_b(\theta_b(t)) = L_b(\theta_b(t+1))$ *will hold* $\forall t$*. Moreover,* $\theta_a(t+1) = \theta_a(t)$ *and* $\theta_b(t+1) = \theta_b(t)$ *hold* $\forall t$ *and* $(\theta_a(t), \theta_b(t))$ *is a constant decision.*

Theorem 2 shows the convergence of decisions $\theta_a(t)$ and $\theta_b(t)$, that if $\theta_a(t)$ moves toward to (resp. far away from) δ_a, then $\theta_b(t)$ will move far away from (resp. toward to) δ_b. It means that once a group suffer the higher loss than the other, it will always suffer the higher loss (See Fig. 3). Therefore, once $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ starts to change (either becoming smaller or larger), it will continue changing in that direction (smaller or larger). As a consequence, group representation disparity will get exacerbated. The fact that one-shot decisions converge to a constant decision immediately results in the convergence of groups proportion $\overline{\alpha}_k(t)$ and the expected total loss $L^t(\theta_a(t), \theta_b(t))$.

Corollary 1. *Consider* $(\theta_a^\infty, \theta_b^\infty)$*, the greedy decisions converge to under either the* `Simple`, `EqOpt` *or* `StatPar` *fairness, then we have the following: (1)* $\lim_{t\to\infty} N_k(t) = \frac{\beta_k}{1-\nu(L_k(\theta_k^\infty))}$, $k \in \{a, b\}$*. (2)* $\lim_{t\to\infty} \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} = \frac{\beta_a}{\beta_b} \frac{1-\nu(L_b(\theta_b^\infty))}{1-\nu(L_a(\theta_a^\infty))}$ *(3)* $\lim_{t\to\infty} L^t(\theta_a(t), \theta_b(t)) = L_b(\theta_b^\infty) + \frac{L_a(\theta_a^\infty) - L_b(\theta_b^\infty)}{1 + \frac{\beta_b}{\beta_a} \frac{1-\nu(L_a(\theta_a^\infty))}{1-\nu(L_b(\theta_b^\infty))}}$.

If the expected losses $L_a(\theta_a^\infty)$ and $L_b(\theta_b^\infty)$ are of significant difference, then there is a large gap between $\nu(L_a(\theta_a^\infty))$ and $\nu(L_b(\theta_b^\infty))$, and $\lim_{t\to\infty} \overline{\alpha}_a(t)$ will approach either 0 or 1, i.e., the disadvantaged group will eventually drop out

[3]This condition will always be satisfied when the system starts from the near empty state.

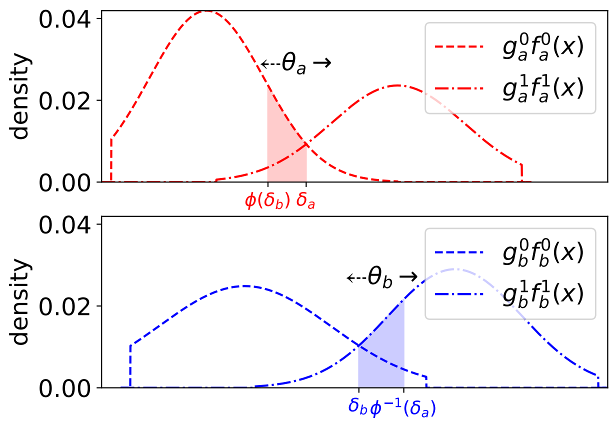


Figure 3: Convergence process illustration: let $(\theta_a(t), \theta_b(t))$ be one-shot fair solution, then $\theta_a(t) \in [\phi(\delta_b), \delta_a]$ (red shaded area) and $\theta_b(t) \in [\delta_b, \phi^{-1}(\delta_a)]$ (blue shaded area) must hold, where $(\phi(\delta_b), \delta_b)$, $(\delta_a, \phi^{-1}(\delta_a))$ are two decision pairs satisfying fairness constraints. The pair $(\theta_a(t), \theta_b(t))$ will move toward the same direction, i.e., either toward right (increase) or toward left (decrease), during the entire horizon. If both increase (resp. decrease) as t increases, then $\theta_a(t)$ becomes closer to (resp. farther away from) δ_a and $L_a(\theta_a(t))$ decreases (resp. increases); $\theta_b(t)$ becomes farther away from (resp. closer to) δ_b and $L_b(\theta_b(t))$ increases (resp. decreases). At the end, the decisions pair will be more in favor of one group and representation disparity get exacerbated.

and the system will consist mostly of one group. In short, none of the `Simple`, `EqOpt` and `StatPar` fairness criteria can prevent this type of permanent attrition under dynamics (1).

3.3 Equalized Loss Fairness

We now turn to the fourth fairness criterion, Equalized Loss (`EqLos`) fairness. By definition, this criterion seeks to guarantee that the expected loss experienced by each group to be the same: $L_a(\theta_a(t)) = L_b(\theta_b(t))$ (Figure 1(d)).

Let us again consider the greedy decision based on the one-shot solution to the optimization problem $(\star)$, where in each time step we seek to minimize the expected total loss $L^t(\theta_a(t), \theta_b(t))$ under this constraint.

Theorem 3. *Let $(\theta_a(t), \theta_b(t))$ be the `EqLos` one-shot fair solution for time step t, then $L_a(\theta_a(t)) = L_b(\theta_b(t)) = \max\{\min_\theta L_a(\theta), \min_\theta L_b(\theta)\}, \forall t$. With this solution we have $\lim_{t\to\infty} \overline{\alpha}_a(t) = \frac{\beta_a}{\beta_a+\beta_b}$ and $\lim_{t\to\infty} \overline{\alpha}_b(t) = \frac{\beta_b}{\beta_a+\beta_b}$. Furthermore, under the `EqLos` fairness, the one-shot decision is a solution to the infinite horizon problem $(\star\star)$. In other words, the sequence of $\{(\theta_a(t), \theta_b(t))\}$, where each $(\theta_a(t), \theta_b(t))$ is a solution to $(\star)$ under `EqLos` fairness, is also a solution to $(\star\star)$ under `EqLos` fairness.*

There are two main takeaways from Theorem 3. The first and most relevant to the central focus of this paper, is the fact that under dynamics (1), the `EqLos` fairness criterion can sustain group representations in accordance with the group's natural, exogenous arrivals, rather than inducing representation detached from this natural replenish rate as a consequence to users' experiences of loss, which in most cases means exacerbated disparity and in the worst case, driving one group to extinction entirely. A secondary, but equally interesting observation is that greedy one-shot decisions using the `EqLos` constraint (solution to $(\star)$) is in fact an optimal solution to the infinite horizon problem $(\star\star)$. In other words, if this is the adopted fairness, then the long-term problem has a rather simple solution. This is not the case with any of the other criterion: `Simple`, `StatPar` or `EqOpt`. In these three cases, the greedy solution (to $(\star)$) is in general not the same as that to $(\star\star)$; some examples are given in Section 5.

4 Discussion

4.1 When strict monotonicity over $\mathcal{T}_k$ doesn't hold

Similar analysis can be done when Assumption 1 does not hold. Below we show the result for the special case where $f_k^j(x) = \mathrm{U}([\underline{k}_j, \overline{k}_j])$ is uniformly distributed (Fig. 2(b)).

Theorem 4. *Consider the one-shot problem* $(\star)$ *at time step* t *with group proportions given by* $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. *The one-shot solution* $(\theta_a(t), \theta_b(t))$ *under* `Simple`, `EqOpt` *or* `StatPar` *fairness satisfies the following:*

$$(\theta_a(t), \theta_b(t)) = \begin{cases} (\theta_a^1, \theta_b^1), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} \in (0, r_1) \\ (\theta_a^2, \theta_b^2), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} \in (r_1, r_2) \\ \quad \vdots \\ (\theta_a^M, \theta_b^M), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} \in (r_{M-1}, +\infty) \end{cases} \tag{3}$$

where the finite number of decision pairs $\{(\theta_a^m, \theta_b^m)\}_{m=1}^M$ *and* $\{r_m\}_{m=1}^{M-1}$ *are different under different fairness criteria and are determined by feature distributions* $f_a(x)$, $f_b(x)$.

We refer an interested reader to Appendix H for details on this result. Theorem 4 shows that regardless of the group proportions $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$, there are only a finite number of feasible solution pairs; this is different from the case discussed in Section 3.

Definition 3. *(Visited Decision) Among all the possible one-shot solutions* $\{(\theta_a^m, \theta_b^m)\}_{m=1}^M$, *we say* (θ_a^m, θ_b^m) *is visited at time* t_0 *if* $(\theta_a(t_0), \theta_b(t_0)) = (\theta_a^m, \theta_b^m)$ *and* $(\theta_a(t_0-1), \theta_b(t_0-1)) \neq (\theta_a^m, \theta_b^m)$. *Further,* (θ_a^m, θ_b^m) *is the* k*-th visited decision if* $(\theta_a(t), \theta_b(t))$ *has changed values* $(k-1)$ *times before visiting* (θ_a^m, θ_b^m).

The long-term property of applying this one-shot solution is characterized as follows:

Theorem 5. *Let* $I_i = \{m | (\theta_a^m, \theta_b^m) \text{ is the } i\text{-th visited decision}\}$ *and* $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$ *be the ordered list of all visited decisions under the* `Simple`, `EqOpt` *or* `StatPar` *fairness. It has the following properties:*

1. *Each decision* (θ_a^m, θ_b^m), $m = 1, \cdots, M$ *can only be visited at most once, i.e., no decision is revisited.*
2. $\{I_i\}_{i=1}^K$ *is a subsequence of the consecutive numbers (either in descending order or ascending order) among* $\{1, 2, \cdots, M\}$
3. *The number of visited decisions* K *is finite, and the one-shot decision will converge to a constant decision* $(\theta_a^{I_K}, \theta_b^{I_K})$.

Algorithm 1 in Appendix J also provides a simple method for finding $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$. Theorem 5, though not the same, is consistent with the case studied in Section 3: the one-shot solutions $(\theta_a(t), \theta_b(t))$ in both cases cannot be revisited and converge to a constant decision. The main results, including the convergence of groups proportion $\overline{\alpha}_k(t)$ and the failure to prevent group attrition under the `Simple`, `EqOpt` and `StatPar` fairness, continue to hold in this uniform case.

4.2 The role of dynamic model

The current dynamic model is driven by model accuracy, which can be applied to domains such as biometric authentication, speaker verification, and medical diagnosis. Within this context, our results suggest different groups should receive the same quality of service in order to maintain presentation. The fact that `EqLos` maintains representation is precisely because of this choice of dynamic: we are essentially equalizing departure when equalizing loss. In contrast, under the other fairness criteria the factor we equalize doesnt match what drives departure, and different loss incurred to different groups over time causes significant change in group proportion.

Although this model does not capture all scenarios and may not be suitable for some applications, this is an example to illustrate the fact that there can easily be a mismatch between a particular fairness definition and the underlying factors driving the system dynamics, and that this mismatch exacerbates representation disparity in the long run. This conclusion applies to all dynamic models. For instance, consider an alternative model where user retention is driven only by true positives or false negatives. This could be more suitable for domains such as loan application and hiring process. In this case it can be shown that `EqOpt` will do a better job at maintaining group representation (See Figure 6(a) in Section 5). If consider another model where for the same group G_k ($k \in \{a, b\}$), the sub-groups with different labels, i.e., G_k^0 and G_k^1, are driven by different factors (e.g., false positives and false negatives), then none of these four fairness criteria can maintain the group representation (See Figure. 6(b) in Section 5).

In reality the true dynamic is likely a function of a mixture of factors given the application context. The model used in the paper is an example that serves to highlight the fact that fairness has to be addressed with a good understanding of how user actions are affected by their perceptions of the algorithm. In other words, fairness cannot be defined in a one-shot problem setting without considering the long-run impact, and that long-run impact cannot be properly analyzed without understanding the underlying dynamics.

The importance of dynamics in evaluating fairness is also pointed out by [13], which focuses on two consecutive time steps and constructs a one-step feedback model that characterizes the impact of fairness criteria on changing each individual's feature and reshaping the entire population. Specifically, it considers a lending scenario and shows that under certain conditions, imposing StatPar and EqOpt can potentially be harmful to minority group and reshape the average feature (e.g., credit score). Different from [13], this work regards each individual's feature as fixed and the users' feedback is described by a retention/attrition model. Instead of considering the impact of fairness criteria on each group's average feature (e.g., credit score of a disadvantaged group decreases), we focus on their long-term impact on group representation over infinite horizon (e.g., the extinction of one group in the system).

4.3 Impact of classifier's quality

As users leave the system, another potential problem is the negative impact on the quality of the classifier, e.g., if users are also contributors to the training dataset, which the current model does not take into account. What we have modeled is the change in total loss induced by different group proportions, where the overall objective is dominated by the larger group, and the model for minority group can deviate more from its own optimum. We show that even with these Bayes optimal decisions, the group representation disparity can be exacerbated when imposing inappropriate fairness constraints.

If we further model the impact of classifier quality, this exacerbation could potentially get more severe, especially once a group starts to go down in size (See Fig. 7 in Section 5).

4.4 Non-constant or unknown arrival β_k

We have assumed that the system dynamics follows (1), where β_k denotes constant amount of arrival in each time step. Below we discuss two relaxations to this.

i) No information on arrival dynamics: Even if we do not know the retention/attrition dynamics, we still are able to find the optimal decision parameter $(\theta_a(t), \theta_b(t))$ under equalized loss constraint. The EqLos one-shot fair solution for time step t is given by,

$$\begin{aligned} \min_{\theta_a(t),\theta_b(t)} \quad & \overline{\alpha}_a(t)L_a(\theta_a(t)) + \overline{\alpha}_b(t)L_b(\theta_b(t)) \\ s.t., \quad & L_a(\theta_a(t)) = L_b(\theta_b(t)). \end{aligned}$$

As $L_a(\theta_a(t) = L_b(\theta_b(t))$, the above optimization is equivalent to the following problem.

$$\begin{aligned} \min_{\theta_a(t),\theta_b(t)} \quad & L_a(\theta_a(t)) \text{ or } L_b(\theta_b(t)) \\ s.t., \quad & L_a(\theta_a(t)) = L_b(\theta_b(t)), \end{aligned}$$

As the EqLos one-shot fair solution for time step t does not depend on the group proportion and dynamics model, $(\theta_a^*(t), \theta_b^*(t))$ is the solution to the infinite horizon problem $(\star\star)$ under EqLos constraint for any arbitrary dynamics. Interestingly, the optimal EqLos fair decision $(\theta_a^*(t), \theta_b^*(t))$ can be found by solving unconstrained optimization problems and $L_a(\theta_a^*(t)) = L_b(\theta_b^*(t)) = \max\{\min_\theta L_a(\theta), \min_\theta L_b(\theta)\}$.

ii) Random arrival β_k: Now consider the arrival model (1) with $\beta_k(t)$ being a positive random variable with mean value $\mu_k, k \in \{a, b\}$. It is easy to see that $N_k(t)$ under equalized loss criterion is given by,

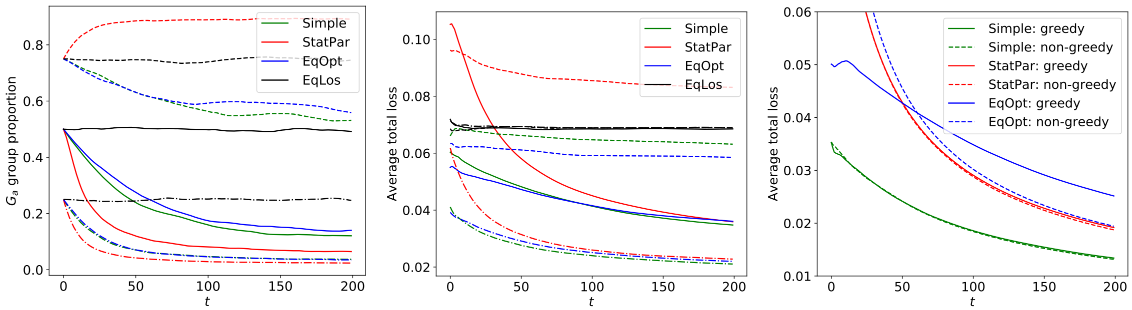
$$N_k(t) = N_k(1)\nu(L_k(\theta_k^*))^{t-1} + \sum_{i=1}^{t-1} \beta_k(i)\nu(L_k(\theta_k^*))^{t-1-i}$$

$$\lim_{t\to\infty} E(N_k(t)) = \frac{\mu_k}{1-\nu(L_k(\theta_k^*))}, \lim_{t\to\infty} \frac{E(N_a(t))}{E(N_b(t))} = \frac{\mu_a}{\mu_b}$$

Therefore, we arrive at a similar result as in Theorem 3 under the EqLos constraint: 1) one-shot fair solution is the solution to the infinite horizon problem, and 2) EqLos fairness is able to sustain group representation in accordance with expected arrivals.

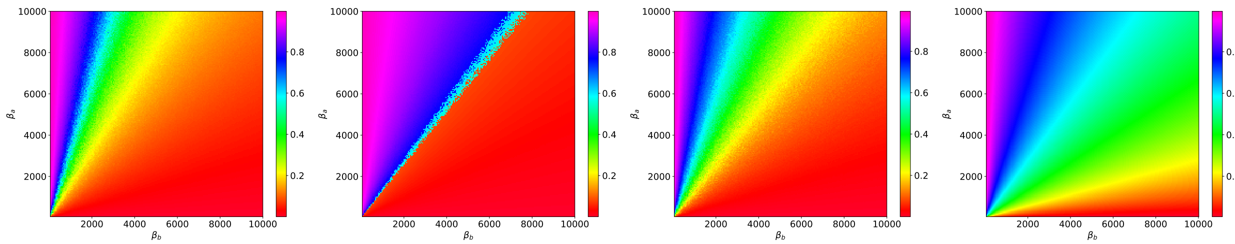
4.5 Case of incomplete information

So far we have assumed that key parameters of the system are known or observable to the decision maker, thereby resulting in a deterministic problem. This requirement may be relaxed for some parameters as discussed above. In general, if these parameters, including the arrival dynamics, instantaneous group proportion, etc., are not observable, but the underlying random processes driving these quantities have known distributions, then a Markov decision process (MDP) may be formulated, turning the deterministic sequential problem into a stochastic sequential decision problem. The resulting problem is likely to be analytically intractable unless one imposes strong assumptions. It is however possible to use approximation techniques as well as value/policy iteration techniques [17] to look for solutions.



(a) Group proportion (b) Average total loss (c) greedy decision is suboptimal

Figure 4: Sample paths for truncated normal example under different fairness criteria when $\beta_a + \beta_b = 20000$. Group proportion $\overline{\alpha}_a(t)$ and average total loss are shown in Figure 4(a)4(b) respectively: solid lines are for the case $\beta_a = \beta_b$, dashed lines for $\beta_a = 3\beta_b$, and dotted dashed lines for $\beta_a = \beta_b/3$. Figure 4(c) shows the existence of a solution (non-greedy) that can achieve lower average total loss under `Simple`, `StatPar` and `EqOpt` fairness criteria.



(a) `Simple` fair (b) `StatPar` fair (c) `EqOpt` fair (d) `EqLos` fair

Figure 5: Each dot in Figure 5(a)-5(d) represents the final group proportion $\lim_{t\to\infty} \overline{\alpha}_a(t)$ of one sample path under a pair of arriving rates (β_a, β_b). If the group representation is sustained, then $\lim_{t\to\infty} \overline{\alpha}_a(t) = \frac{1}{1+\beta_b/\beta_a}$ for each pair of (β_a, β_b), as shown in Figure 5(d) under `EqLos` fairness. Representation disparity is illustrated under `Simple`, `StatPar` and `EqOpt` fairness, where $\lim_{t\to\infty} \overline{\alpha}_a(t) = \frac{1}{1+\frac{\beta_b(1-\nu(L_a(\theta_a^\infty)))}{\beta_a(1-\nu(L_b(\theta_b^\infty)))}}$ depends on the converged decision $(\theta_a^\infty, \theta_b^\infty)$ and the group's retention degrades if it suffers the higher loss.

5 Experiments

We performed a set of experiments on synthetic data where every G_k^j, $k \in \{a, b\}, j \in \{0, 1\}$ follows either the uniform (Fig 2(b)) or the truncated normal (Fig 2(a)) distributions. We only present results on the latter in this section; similar results for uniform distributions can be found in Appendix K.2. A sequence of greedy decisions under different fairness constraints are used and group population size changes over time according to (1). For detailed information about parameters and experimental setup, please refer to Appendix K.1.

Figure 4 shows sample paths of the group proportion and average total loss using greedy decisions under various fairness criteria and different combinations of β_a, β_b. As our analysis predicted, in all cases convergence is reached (we did not include the decisions $\theta_k(t)$ but convergence holds there as well). In particular, under `EqLos` fairness, the group representation is sustained throughout the horizon, with an average total loss comparable to using other fairness constraints (Fig. 4(b))[4].

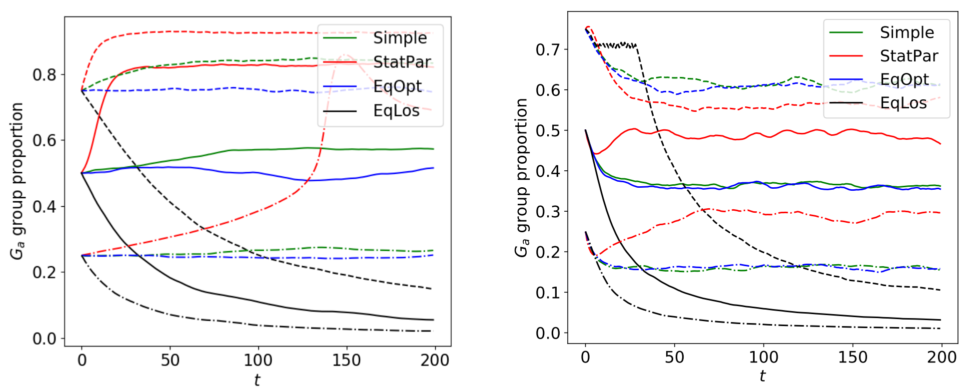
By contrast, under other fairness constraints, even a "major" group (one with a larger arrival β_k) can overtime be significantly marginalized (blue/green dashed line in Fig. 4(a)). This occurs when the loss of the minor group happens to be smaller than that of the major group, which is determined by feature distributions of the two groups (see Appendix K.3). Whenever this is the case, the greedy decision will seek to increase the minor group's proportion in order to drive down the average loss.

[4]For the uniform distribution example shown in Appendix K.2, the average total loss of `EqLos` fairness can be even smaller.

In addition to these sample paths, we also illustrate the sub-optimality of the greedy decisions under the `Simple`, `EqOpt` and `StatPar` fairness criteria for the infinite horizon problem ($\star\star$). In Theorem 3 we established such optimality for the `EqLos` greedy fair decision, but this does not in general hold for all the other fairness criteria. Figure 4(c) shows examples where a non-greedy policy can perform better over the long horizon.

Figure 5 further illustrates the final group proportion (the converged state) as a function of the exogenous arrival sizes β_a and β_b under different fairness criteria. As seen, with the exception of `EqLos` fairness, the group representations are severely skewed in the long run, with the system consisting mostly of G_b, even for scenarios when G_a has larger arrival, i.e., $\beta_a > \beta_b$. A particularly interesting observation is in Fig. 5(b), which approximately splits into three regions, which means that a minor change in β_a and β_b can result in very different representation in the long run; in other words, poor robustness. The same phenomenon could be observed under `Simple` and `EqOpt` fairness as well given a different choice of feature distributions; see Appendix K.2 for more examples.

As discussed in Section 4.2, if adopt different dynamic models, different fairness criteria should be adopted to maintain group representation. The key point is that the fairness definition should match the factors that drive user departure and arrival. Two examples with different dynamics and the performance of four fairness criteria are demonstrated in Fig. 6.



(a) Users from G_k are driven by false negative rate (b) Users from G_k^j are driven by their own perceived loss

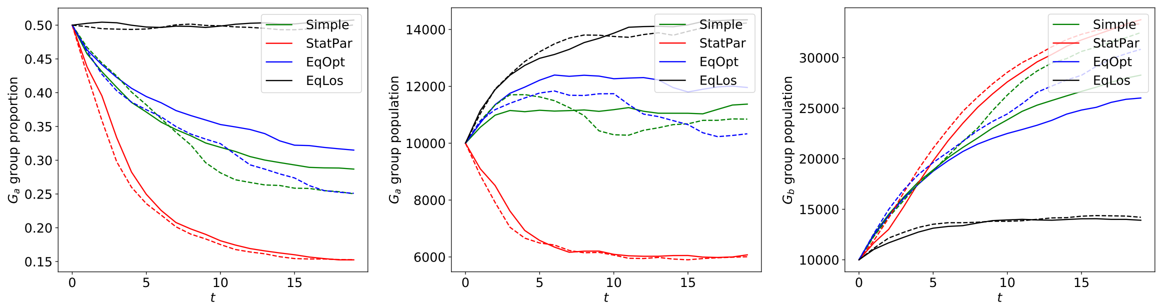
Figure 6: Sample paths under different dynamic models: Three cases are demonstrated including $\beta_a = \beta_b$ (solid curves); $\beta_a = 3\beta_b$ (dashed curves); $\beta_a = \beta_b/3$ (dotted dash curves). Fig. 6(a) illustrates the model where the user departure is driven by false negative rate: $N_k(t+1) = N_k(t)\nu(\text{FN}_k(\theta_k(t))) + \beta_k$, with $\text{FN}_k(\theta_k(t)) = \int_{\theta_k(t)}^{\infty} f_k^0(x)dx$. Under this model `EqOpt` is better at maintaining representation. Fig. 6(b) illustrates the model where the users from each sub-group G_k^j are driven by their own perceived loss: $N_k^j(t+1) = N_k^j(t)\nu(L_k^j(\theta_k(t))) + g_k^j\beta_k$, with $L_k^j(\theta_k)$ being false positives for $j = 0$ and false negatives for $j = 1$. Under this model none of four criteria can maintain group representation.

As discussed in Section 4.3, if $f_k^j(x)$ is unknown to decision maker and the decision is learned from users in the system, then as users leave the system the decision can be more inaccurate and the exacerbation could potentially get more severe. In order to illustrate it, we first modify the dynamic model such that the users' arrivals are also effected by the model accuracy[5], i.e.,

$$N_k(t+1) = (N_k(t) + \beta_k)\nu(L_k(\theta_k(t))) \tag{4}$$

Under dynamic (4), we compare the performance of two cases: *(i)* the Bayes optimal decisions are applied in every round; and *(ii)* decisions in $(t+1)$-th round are learned from the remaining users in t-th round. The empirical results are shown in Fig. 7 where each solid curve (resp. dashed curve) is a sample path of case *(i)* (resp. case *(ii)*). Although $\beta_a = \beta_b$, G_b suffers the less loss at the beginning and starts to dominant the overall objective gradually. It results in the less and less users from group G_a than G_b in the sample pool and the model trained from minority group G_a suffers an additional loss due to its insufficient samples. In contrast, as G_b dominants more in the objective and its loss may be decreased compared with the case *(i)* (See Fig. 7(c)). As a consequent, the exacerbation in group representation disparity gets more severe (See Fig. 7(a)).

[5]The size of one group can decrease in this case, while the size of two groups is always increasing for the dynamic (1).



(a) Group proportion (b) G_a's total population (c) G_b's total population

Figure 7: Impact of the classifier's quality: dashed curves represent the results for decisions learned from users (case *(ii)*), solid curves represent the results for Bayes optimal decisions (case *(i)*). It shows the exacerbation of group disparity get more severe under case *(ii)* for `Simple`, `EqOpt` and `StatPar` criteria.

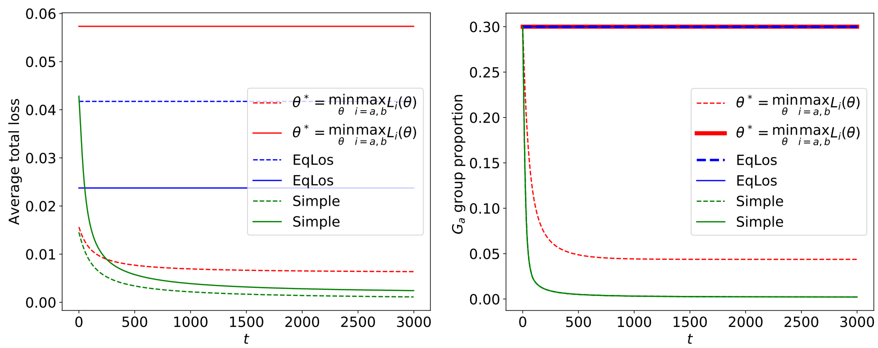


Figure 8: Trade-off between average total loss and fairness under `Simple`, `EqLos` and MinMax loss: solid and dashed lines denote two examples with different feature distributions.

We end this section by comparing `EqLos` fair decision with the policy suggested in [1], minimizing maximum loss (MinMax): $\theta^*(t) = \arg\min_\theta \max\{L_a(\theta), L_b(\theta)\}$. Fig. 8 shows in general `Simple` can achieve the lowest loss but G_a eventually disappears entirely. `EqLos` sustains group representation but may result in higher loss compared to MinMax in this example, but presents a better tradeoff: at the same group representation level MinMax comes with higher loss than `Eqlos`.

6 Conclusion

This paper characterizes the long-term property of fair machine learning in a sequential setting. We show the convergence of the decision under a particular dynamic model, and that both statistical parity and equal opportunity fairness can exacerbate group representation disparity over time. In contrast, Equalized loss fairness can maintain the group representation. Our results highlight the fact that the fairness cannot be tested in a one-shot problem, it must be defined with the good understanding of dynamics, i.e., how users react according to the decisions.

References

[1] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[2] Toon Calders and Indrė Žliobaitė. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, pages 43–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[3] Drew Harwell. Amazons alexa and google home show accent bias, with chinese and spanish hardest to understand. 2018.

[4] Paul D Allison. *Missing data*, volume 136. Sage publications, 2001.

[5] Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010.

[6] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

[7] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, 2017.

[8] F. d. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, Oct 2018.

[9] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct 2012.

[10] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[11] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[12] Moritz Hardt, Eric Price, , and Nati Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

[13] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[14] Avrim Blum, Suriya Gunasekar, Thodoris Lykouris, and Nati Srebro. On preserving non-discrimination when combining expert advice. In *Advances in Neural Information Processing Systems*, pages 8386–8397, 2018.

[15] Parikshit Ghosh and Debraj Ray. Information and enforcement in informal credit markets. *Economica*, 83(329):59–90, 2016.

[16] Nurhuda Syed. Recruiters blasted for myopic hiring process. 2018.

[17] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1133–1142, 2018.

Appendix

A Notation table

Notation	Description
G_k, $k \in \{a,b\}$	two demographic groups
G_k^j, $j \in \{0,1\}$	subgroup with label j in G_k
α_k^j	size of G_k^j as a fraction of entire population
$\overline{\alpha}_k$	size of G_k as a fraction of entire population, i.e., $\alpha_a^0 + \alpha_k^1$
g_k^j	fraction of subgroup with label j in G_k, i.e., $\alpha_k^j/\overline{\alpha}_k$, i.e., $\Pr(Y=j\|K=k)$
$f_k^j(x)$	feature distribution of G_k^j, i.e., $\Pr(X=x\|K=k, Y=j)$
$g_k^j f_k^j(x)$	feature distribution of G_k that has label j, i.e., $\Pr(X=x, Y=j\|K=k)$
$f_k(x)$	feature distribution of G_k, i.e., $\Pr(X=x\|K=k)$ and $f_k(x) = g_k^1 f_k^1(x) + g_k^0 f_k^0(x)$
$f(x)$	feature distribution over the entire population, i.e., $\sum_{k\in\{a,b\}, j\in\{0,1\}} \alpha_k^j f_k^j(x)$
$h_\theta(x)$	decision rule parameterized by θ
θ_k	decision parameter for G_k
$L_k(\theta_k)$	expected loss incurred to G_k by taking decision θ_k
$L(\theta_a, \theta_b)$	total expected loss, i.e., $\sum_{k\in\{a,b\}} \overline{\alpha}_k L_k(\theta_k)$
$\Gamma_{\mathcal{C}}(\theta_a, \theta_b)$	a fairness constraint imposed on θ_a and θ_b for two groups
$N_k(t)$	expected number of users from G_k at time t
$\nu(L_k(\theta_k(t)))$	retention rate of G_k at time t when imposing decision $\theta_k(t)$
β_k	number of exogenous arrivals to G_k at every time step
$[\underline{k}_j, \overline{k}_j]$	bounded support of distribution $f_k^j(x)$
δ_k	optimal decision for G_k such that $\delta_k = \arg\min_\theta L_k(\theta)$ and satisfies $g_k^1 f_k^1(\delta_k) = g_k^0 f_k^0(\delta_k)$
$\phi(\cdot)$	a increasing function determined by constraint $\Gamma_{\mathcal{C}}(\theta_a, \theta_b)$ mapping θ_b to θ_a, i.e., $\Gamma_{\mathcal{C}}(\phi(\theta_b), \theta_b)$

B Proof of Lemma 1

Proof by contradiction.

Let $\mathcal{V} = \{(\theta_a, \theta_b) \| \theta_a \in [\phi(\delta_b), \delta_a], \theta_b \in [\delta_b, \phi^{-1}(\delta_a)], h_{\mathcal{C}}(\theta_a, \theta_b) = 0\}$.

Since for `Simple`, `EqOpt`, `StatPar` fairness, any (θ_a, θ_b) and (θ_a', θ_b') that satisfy constraints $h_{\mathcal{C}}(\theta_a, \theta_b) = 0$ and $h_{\mathcal{C}}(\theta_a', \theta_b') = 0$, $\theta_a \geq \theta_a'$ if and only if $\theta_b \geq \theta_b'$. Suppose $(\check{\theta}_a, \check{\theta}_b)$ satisfies $h_{\mathcal{C}}(\check{\theta}_a, \check{\theta}_b) = 0$ and $(\check{\theta}_a, \check{\theta}_b) = \arg\min_{\theta_a, \theta_b} L(\theta_a, \theta_b) \notin \mathcal{V}$, then one of the followings must hold: (1) $\check{\theta}_a < \phi(\delta_b)$, $\check{\theta}_b < \delta_b$; (2) $\check{\theta}_a > \delta_a$, $\check{\theta}_b > \phi^{-1}(\delta_a)$. Consider two cases separately.

(1) $\check{\theta}_a < \phi(\delta_b)$, $\check{\theta}_b < \delta_b$

Since $L_b(\check{\theta}_b) > L_b(\delta_b)$, to satisfy $L(\check{\theta}_a, \check{\theta}_b) < L(\phi(\delta_b), \delta_b)$, $L_a(\check{\theta}_a) < L_a(\phi(\delta_b))$ must hold. However, under the Assumption 1, $L_a(\theta_a)$ is strictly decreasing on $[\underline{a}_0, \delta_a]$ and strictly increasing on $[\delta_a, \overline{a}_1]$. Since $\check{\theta}_a < \phi(\delta_b) < \delta_a$, which implies $L_a(\check{\theta}_a) > L_a(\phi(\delta_b))$. Therefore, $(\check{\theta}_a, \check{\theta}_b)$ cannot be the optimal pair.

(2) $\check{\theta}_a > \delta_a$, $\check{\theta}_b > \phi^{-1}(\delta_a)$

Since $L_a(\check{\theta}_a) > L_a(\delta_a)$, to satisfy $L(\check{\theta}_a, \check{\theta}_b) < L(\delta_a, \phi^{-1}(\delta_a))$, $L_b(\check{\theta}_b) < L_b(\phi^{-1}(\delta_a))$ must hold. However, under the Assumption 1, $L_b(\theta_b)$ is strictly decreasing on $[\underline{b}_0, \delta_b]$ and strictly increasing on $[\delta_b, \overline{b}_1]$. Since $\check{\theta}_b > \phi^{-1}(\delta_a) > \delta_b$, which implies $L_b(\check{\theta}_b) > L_b(\phi^{-1}(\delta_a))$. Therefore, $(\check{\theta}_a, \check{\theta}_b)$ cannot be the optimal pair.

C Proof of Theorem 1

Proof of Theorem 1 is based on the following Lemma.

Lemma 3. *Consider the one-shot problem ($\star$) at some time step t, with group proportions given by $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. Under Assumption 1 the greedy decision $(\theta_a(t), \theta_b(t))$ for this time step is unique and satisfies the following:*

(1) Under `EqOpt` fairness:

- *If* $\theta_b(t) \in [\underline{b}_1, \min\{\overline{b}_0, \phi^{-1}(\underline{a}_1)\}]$, *then* $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1} + \frac{\overline{\alpha}_a(t) g_a^0}{\overline{\alpha}_b(t) g_b^1}$.
- *If* $\theta_b(t) \in [\max\{\underline{b}_1, \phi^{-1}(\underline{a}_1)\}, \min\{\overline{b}_0, \phi^{-1}(\overline{a}_0)\}]$, *then* $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1} + \frac{\overline{\alpha}_a(t) g_a^0}{\overline{\alpha}_b(t) g_b^1} - \frac{g_a^1 \overline{\alpha}_a(t) f_a^1(\phi(\theta_b(t)))}{g_b^1 \overline{\alpha}_b(t) f_a^0(\phi(\theta_b(t)))}$.

(2) Under `StatPar` *fairness:*

- *If* $\theta_b(t) \in [\underline{b}_1, \min\{\overline{b}_0, \phi^{-1}(\underline{a}_1)\}]$, *then* $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1}\left(\frac{1+\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}}{1-\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}}\right)$.
- *If* $\theta_b(t) \in [\max\{\underline{b}_1, \phi^{-1}(\underline{a}_1)\}, \min\{\overline{b}_0, \phi^{-1}(\overline{a}_0)\}]$, *then* $\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)} = \frac{\left(\frac{g_b^1 f_b^1(\theta_b(t))}{g_b^0 f_b^0(\theta_b(t))}+1\right)\left(1-\frac{g_a^1 f_a^1(\phi(\theta_b(t)))}{g_a^0 f_a^0(\phi(\theta_b(t)))}\right)}{\left(\frac{g_b^1 f_b^1(\theta_b(t))}{g_b^0 f_b^0(\theta_b(t))}-1\right)\left(\frac{g_a^1 f_a^1(\phi(\theta_b(t)))}{g_a^0 f_a^0(\phi(\theta_b(t)))}+1\right)}$.
- *If* $\theta_b(t) \in [\max\{\underline{a}_1, \phi^{-1}(\overline{b}_0)\}, \overline{a}_0]$, *then* $\frac{f_a^1(\phi(\theta_b(t)))}{f_a^0(\phi(\theta_b(t)))} = \frac{g_a^0}{g_a^1}\left(\frac{1-\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}}{1+\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}}\right)$.

(3) Under `Simple` *fairness:*

- *If we further assume* $\delta_a, \delta_b \in \mathcal{T}_a \cap \mathcal{T}_b$[6], *then* $\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)} = \frac{g_a^0 f_a^0(\theta_b(t)) - g_a^1 f_a^1(\theta_b(t))}{g_b^1 f_b^1(\theta_b(t)) - g_b^0 f_b^0(\theta_b(t))}$.

Proof. We focus on the case when $g_a^1 f_a^1(\underline{a}_1) < g_a^0 f_a^0(\underline{a}_1)$ & $g_a^1 f_a^1(\overline{a}_0) > g_a^0 f_a^0(\overline{a}_0)$ and $g_b^1 f_b^1(\underline{b}_1) < g_b^0 f_b^0(\underline{b}_1)$ & $g_b^1 f_b^1(\overline{b}_0) > g_b^0 f_b^0(\overline{b}_0)$. That is, $\theta_k^* = \text{armin}_\theta L_k(\theta) = \delta_k$ holds for $k = a, b$

The constraint $h_{\mathcal{C}}(\theta_a, \theta_b) = 0$ can be rewritten as $\theta_a = \phi(\theta_b)$ for some strictly increasing function ϕ. Then $\frac{d\phi(\theta_b)}{d\theta_b} = -\frac{\frac{\partial h_{\mathcal{C}}(\theta_a,\theta_b)}{\partial \theta_b}}{\frac{\partial h_{\mathcal{C}}(\theta_a,\theta_b)}{\partial \theta_a}}|_{\theta_a=\phi(\theta_b)}$. For the `EqOpt` fairness, $\frac{d\phi(\theta_b)}{d\theta_b} = \frac{f_b^0(\theta_b)}{f_a^0(\phi(\theta_b))}$, for `StatPar` fairness, $\frac{d\phi(\theta_b)}{d\theta_b} = \frac{g_b^0 f_b^0(\theta_b) + g_b^1 f_b^1(\theta_b)}{g_a^0 f_a^0(\phi(\theta_b)) + g_a^1 f_a^1(\phi(\theta_b))}$, for `Simple` fairness, $\frac{d\phi(\theta_b)}{d\theta_b} = 1$.

The one-shot problem can be expressed with only one variable, either θ_a or θ_b, here we express in θ_b. At each round, decision maker find $\theta_b(t) = \arg\min_{\theta_b} L^t(\theta_b) = \overline{\alpha}_a L_a(\phi(\theta_b)) + \overline{\alpha}_b L_b(\theta_b)$ and $\theta_a(t) = \phi(\theta_b(t))$. Since $\phi(\delta_b) < \delta_a$ ($\phi^{-1}(\delta_a) > \delta_b$), then each $(\theta_a(t), \theta_b(t))$ can have 3 possibilities: (1) $\theta_a(t) \in [\underline{a}_0, \underline{a}_1]$, $\theta_b(t) \in [\underline{b}_1, \overline{b}_0]$; (2) $\theta_a(t) \in [\underline{a}_1, \overline{a}_0]$, $\theta_b(t) \in [\underline{b}_1, \overline{b}_0]$; (3) $\theta_a(t) \in [\underline{a}_1, \overline{a}_0]$, $\theta_b(t) \in [\overline{b}_0, \overline{b}_1]$. For `EqOpt` and `StatPar`, consider each case separately.

Case 1: $\theta_a(t) \in [\underline{a}_0, \underline{a}_1]$, $\theta_b(t) \in [\underline{b}_1, \overline{b}_0]$

Let $\theta_b^{\max} = \min\{\overline{b}_0, \phi^{-1}(\underline{a}_1)\}$, which is the maximum value θ_b can take for case 1.

$L^t(\theta_b) = \overline{\alpha}_b(t) \int_{\underline{b}_1}^{\theta_b} g_b^1 f_b^1(x) - g_b^0 f_b^0(x) dx - \overline{\alpha}_a(t) \int_{\underline{a}_0}^{\phi(\theta_b)} g_a^0 f_a^0(x) dx + \overline{\alpha}_a(t) g_a^0 + \overline{\alpha}_b(t) \int_{\underline{b}_1}^{\overline{b}_0} g_b^0 f_b^0(x) dx$

Take derivative w.r.t. θ_b gives $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) - \overline{\alpha}_a(t) g_a^0 f_a^0(\phi(\theta_b)) \frac{d\phi(\theta_b)}{d\theta_b}$.

1. `EqOpt` greedy fair decision:

   $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) - \overline{\alpha}_a(t) g_a^0 f_b^0(\theta_b)$, since $g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)$ is increasing from negative to positive and $f_b^0(\theta_b)$ is decreasing over $[\underline{b}_1, \overline{b}_0]$. $\frac{dL^t(\theta_b)}{d\theta_b}$ is increasing over $[\underline{b}_1, \overline{b}_0]$. There are two possibilities: (i) $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < \frac{g_b^1 f_b^1(\theta_b^{\max}) - g_b^0 f_b^0(\theta_b^{\max})}{g_a^0 f_b^0(\theta_b^{\max})}$, i.e., $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b^{\max}} > 0 \rightarrow \exists \theta_b(t)$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b(t)} = 0$, then $\theta_b(t)$ satisfies: $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1} + \frac{\overline{\alpha}_a(t) g_a^0}{\overline{\alpha}_b(t) g_b^1}$ and it is unique given $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. (ii) $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} \geq \frac{g_b^1 f_b^1(\theta_b^{\max}) - g_b^0 f_b^0(\theta_b^{\max})}{g_a^0 f_b^0(\theta_b^{\max})}$, i.e., $\frac{dL^t(\theta_b)}{d\theta_b} \leq 0, \forall \theta_b \in [\underline{b}_1, \theta_b^{\max}] \rightarrow \theta_b(t) \geq \theta_b^{\max}$.

2. `StatPar` greedy fair decision:

   $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) - \overline{\alpha}_a(t) \frac{g_b^1 f_b^1(\theta_b) + g_b^0 f_b^0(\theta_b)}{1 + \frac{g_a^1 f_a^1(\phi(\theta_b))}{g_a^0 f_a^0(\phi(\theta_b))}} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) - \overline{\alpha}_a(t)(g_b^1 f_b^1(\theta_b) + g_b^0 f_b^0(\theta_b))$, where the last equality is because $f_a^1(\phi(\theta_b)) = 0$ over $[\underline{a}_0, \underline{a}_1]$. Since $f_b^1(x)$ is

[6] This extra assumption on simple case applies to Theorem 2

increasing and $f_b^0(x)$ is decreasing over $[\underline{b}_1, \overline{b}_0]$. (i) If $\exists\theta_b(t)$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b(t)} = 0$, then $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1}(\frac{1+\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}}{1-\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}})$ and $\theta_b(t)$ is unique given $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. (ii) If $\frac{dL^t(\theta_b)}{d\theta_b} \leq 0, \forall\theta_b \in [\underline{b}_1, \theta_b^{\max}] \rightarrow \theta_b(t) \geq \theta_b^{\max}$.

Case 2: $\theta_a(t) \in [\underline{a}_1, \overline{a}_0], \theta_b(t) \in [\underline{b}_1, \overline{b}_0]$

Let $\theta_b^{\max} = \min\{\overline{b}_0, \phi^{-1}(\overline{a}_0)\}$ and $\theta_b^{\min} = \max\{\underline{b}_1, \phi^{-1}(\underline{a}_1)\}$ be the maximum and minimum value respectively that θ_b can take for case 2. $L^t(\theta_b) = \overline{\alpha}_b(t)\int_{\underline{b}_1}^{\theta_b} g_b^1 f_b^1(x) - g_b^0 f_b^0(x)dx + \overline{\alpha}_a(t)\int_{\underline{a}_1}^{\phi(\theta_b)} g_a^1 f_a^1(x) - g_a^0 f_a^0(x)dx + \overline{\alpha}_b(t)\int_{\underline{b}_1}^{\overline{b}_0} g_b^0 f_b^0(x)dx + \overline{\alpha}_a(t)\int_{\underline{a}_1}^{\overline{a}_0} g_a^0 f_a^0(x)dx$

Take derivative w.r.t. θ_b gives $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) + \overline{\alpha}_a(t)(g_a^1 f_a^1(\phi(\theta_b)) - g_a^0 f_a^0(\phi(\theta_b)))\frac{d\phi(\theta_b)}{d\theta_b}$.

1. `EqOpt` greedy fair decision:
   $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) + \overline{\alpha}_a(t)(g_a^1\frac{f_a^1(\phi(\theta_b))}{f_a^0(\phi(\theta_b))} - g_a^0)f_b^0(\theta_b) = ((g_a^1\frac{f_a^1(\phi(\theta_b))}{f_a^0(\phi(\theta_b))} - g_a^0)\overline{\alpha}_a(t) - g_b^0\overline{\alpha}_b(t))f_b^0(\theta_b) + g_b^1 f_b^1(\theta_b)\overline{\alpha}_b(t)$, since $g_a^1\frac{f_a^1(\phi(\theta_b))}{f_a^0(\phi(\theta_b))} - g_a^0$ is increasing and $g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)$ is increasing over $[\theta_b^{\min}, \theta_b^{\max}]$. Since $f_a^0(\phi(\theta_b^{\max})) = f_b^0(\theta_b^{\max}) = 0$ under `EqOpt` fairness, $\frac{dL^t(\theta_b)}{d\theta_b}$ is increasing to positive over $[\theta_b^{\min}, \theta_b^{\max}]$. There are two possibilities: (i) If $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b^{\min}} < 0 \rightarrow \exists\theta_b(t)$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b(t)} = 0$, then $\theta_b(t)$ satisfies: $\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} = \frac{g_b^0}{g_b^1} + \frac{\overline{\alpha}_a(t)g_a^0}{\overline{\alpha}_b(t)g_b^1} - \frac{g_a^1\overline{\alpha}_a(t)f_a^1(\phi(\theta_b(t)))}{g_b^1\overline{\alpha}_b(t)f_a^0(\phi(\theta_b(t)))}$ and it is unique given $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. (ii) If $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b^{\min}} \geq 0 \rightarrow \frac{dL^t(\theta_b)}{d\theta_b} \geq 0, \forall\theta_b \in [\theta_b^{\min}, \theta_b^{\max}] \rightarrow \theta_b(t) \leq \theta_b^{\min}$.
2. `StatPar` greedy fair decision:
   $\frac{dL^t(\theta_b)}{d\theta_b} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta_b) - g_b^0 f_b^0(\theta_b)) + \overline{\alpha}_a(t)(g_b^0 f_b^0(\theta_b) + g_b^1 f_b^1(\theta_b))\frac{g_a^1 f_a^1(\phi(\theta_b)) - g_a^0 f_a^0(\phi(\theta_b))}{g_a^1 f_a^1(\phi(\theta_b)) + g_a^0 f_a^0(\phi(\theta_b))}$. $\frac{dL^t(\theta_b)}{d\theta_b}$ is increasing over $[\theta_b^{\min}, \theta_b^{\max}]$. There are three possibilities: (i) If $\exists\theta_b(t)$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b(t)} = 0$, then $\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)} = \frac{(\frac{g_b^1 f_b^1(\theta_b(t))}{g_b^0 f_b^0(\theta_b(t))}+1)(1-\frac{g_a^1 f_a^1(\phi(\theta_b(t)))}{g_a^0 f_a^0(\phi(\theta_b(t)))})}{(\frac{g_b^1 f_b^1(\theta_b(t))}{g_b^0 f_b^0(\theta_b(t))}-1)(\frac{g_a^1 f_a^1(\phi(\theta_b(t)))}{g_a^0 f_a^0(\phi(\theta_b(t)))}+1)}$ and such $\theta_b(t)$ is unique given $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. (ii) If $\frac{dL^t(\theta_b)}{d\theta_b} \geq 0, \forall\theta_b \in [\theta_b^{\min}, \theta_b^{\max}] \rightarrow \theta_b(t) \leq \theta_b^{\min}$. (iii) If $\frac{dL^t(\theta_b)}{d\theta_b} \leq 0, \forall\theta_b \in [\theta_b^{\min}, \theta_b^{\max}] \rightarrow \theta_b(t) \geq \theta_b^{\max}$.

Case 3: $\theta_a(t) \in [\underline{a}_1, \overline{a}_0], \theta_b(t) \in [\overline{b}_0, \overline{b}_1]$

Express $L^t(\theta_a, \theta_b)$ as function of θ_a, the analysis will be similar to case 1.

Let $\theta_a^{\min} = \max\{\underline{a}_1, \phi(\overline{b}_0)\}$, which is the minimum value θ_a can take for case 3.

$L^t(\theta_a) = \overline{\alpha}_a(t)\int_{\underline{a}_1}^{\theta_a} g_a^1 f_a^1(x) - g_a^0 f_a^0(x)dx + \overline{\alpha}_b(t)\int_{\underline{b}_1}^{\phi^{-1}(\theta_a)} g_b^1 f_b^1(x)dx + \overline{\alpha}_a(t)\int_{\underline{a}_1}^{\overline{a}_0} g_a^0 f_a^0(x)dx$

Take derivative w.r.t. θ_a gives $\frac{dL^t(\theta_a)}{d\theta_a} = \overline{\alpha}_a(t)(g_a^1 f_a^1(\theta_a) - g_a^0 f_a^0(\theta_a)) + \overline{\alpha}_b(t)g_b^1 f_b^1(\phi^{-1}(\theta_a))\frac{d\phi^{-1}(\theta_a)}{d\theta_a}$.

Since $\theta_b(t) \in [\overline{b}_0, \overline{b}_1]$ is not possible for `EqOpt` greedy fair decision, Case 3 only considers the `StatPar` greedy fair decision.

$\frac{dL^t(\theta_a)}{d\theta_a} = \overline{\alpha}_a(t)(g_a^1 f_a^1(\theta_a) - g_a^0 f_a^0(\theta_a)) + \overline{\alpha}_b(t)\frac{g_a^1 f_a^1(\theta_a) + g_a^0 f_a^0(\theta_a)}{1+\frac{g_b^0 f_b^0(\phi^{-1}(\theta_a))}{g_b^0 f_b^0(\phi^{-1}(\theta_a))}} = \overline{\alpha}_a(t)(g_a^1 f_a^1(\theta_a) - g_a^0 f_a^0(\theta_a)) + \overline{\alpha}_b(t)(g_a^1 f_a^1(\theta_a) + g_a^0 f_a^0(\theta_a))$, where the last equality is because $f_b^0(\phi^{-1}(\theta_a)) = 0$ over $[\overline{b}_0, \overline{b}_1]$. Since $f_a^1(x)$ is increasing and $f_a^0(x)$ is decreasing over $[\underline{a}_1, \overline{a}_0]$. (i) If $\exists\theta_a(t)$ such that $\frac{dL^t(\theta_a)}{d\theta_a}|_{\theta_a=\theta_a(t)} = 0$, then $\frac{f_a^1(\theta_a(t))}{f_a^0(\theta_a(t))} = \frac{g_a^0}{g_a^1}(\frac{1-\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}}{1+\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}})$ and $\theta_a(t)$ is unique given $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$. (ii) If $\frac{dL^t(\theta_a)}{d\theta_a} \geq 0, \forall\theta_a \in [\underline{b}_1, \theta_a^{\min}] \rightarrow \theta_a(t) \leq \theta_a^{\min}$.

Now consider `Simple` greedy decision, where $\theta_a(t) = \theta_b(t) = \theta(t)$. Since $\delta_a > \delta_b$, under the condition that both $\delta_a, \delta_b \in \mathcal{T}_a \cap \mathcal{T}_b$ and according to Lemma 1, there could be only one case: $\theta(t) \in [\underline{a}_1, \overline{b}_0]$.

Take derivative w.r.t. θ gives $\frac{dL^t(\theta)}{d\theta} = \overline{\alpha}_b(t)(g_b^1 f_b^1(\theta) - g_b^0 f_b^0(\theta)) + \overline{\alpha}_a(t)(g_a^1 f_a^1(\theta) - g_a^0 f_a^0(\theta))$. $\frac{dL^t(\theta)}{d\theta}$ is increasing from negative to positive over $[\delta_b, \delta_a]$, $\exists\theta(t)$ such that $\frac{dL^t(\theta)}{d\theta}|_{\theta=\theta(t)} = 0$, and it satisfies $\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)} = \frac{g_a^0 f_a^0(\theta(t)) - g_a^1 f_a^1(\theta(t))}{g_b^1 f_b^1(\theta(t)) - g_b^0 f_b^0(\theta(t))}$.

□

Based on Lemma 3 and under Assumption 1, for all three notions of fairness, the larger $\theta_b(t)$ and $\theta_a(t)$ will result in the larger $\frac{\alpha_a(t)}{\alpha_b(t)}$. Therefore, $\frac{\alpha_a(t)}{\alpha_b(t)} = \Psi_{\mathcal{C}}(\theta_a(t), \theta_b(t))$ for some increasing function $\Psi_{\mathcal{C}}$, which may not have the analytic expression.

D Proof of Lemma 2

Group labels a, b are assigned to two groups such that $\phi(\delta_b) < \delta_a$ and $\delta_b < \phi^{-1}(\delta_a)$. According to Lemma 1, $\theta_a(t) \in [\phi(\delta_b), \delta_a]$ and $[\delta_b, \phi^{-1}(\delta_a)]$. Since $L_k(\theta_k)$ is decreasing over $[\underline{k}_0, \delta_k]$ and increasing over $[\delta_k, \overline{k}_1]$, for $k \in \{a, b\}$. If $\theta_a \geq \theta_a'$ and $\theta_b \geq \theta_b'$, then $L_a(\theta_a) \leq L_a(\theta_a')$ and $L_b(\theta_b) \geq L_b(\theta_b')$. Lemma 2 is proved.

E Proof of Theorem 2

Theorem 2 is proved based on the following lemmas.

Lemma 4. *For the function $f(t) = \frac{1-a^t}{1-b^t}, t \in \mathbb{Z}_+$, where $a, b \in (0,1)$ are two constants. If $0 < a < b < 1$, then $f(t)$ is strictly decreasing in t and $f(t) > 1, \forall t \in \mathbb{Z}_+$; if $0 < b < a < 1$, then $f(t)$ is strictly increasing in t and $f(t) < 1, \forall t \in \mathbb{Z}_+$.*

Proof. Consider

$$f(t) - f(t+1) = \frac{1-a^t}{1-b^t} - \frac{1-a^{t+1}}{1-b^{t+1}} = \underbrace{\frac{1}{(1-b^t)(1-b^{t+1})}}_{>0} \underbrace{((1-a^t)(1-b^{t+1}) - (1-a^{t+1})(1-b^t))}_{g(t)}$$

If $0 < a < b < 1$, re-organize $g(t)$ as

$$\begin{aligned} g(t) &= b^t(1-b) - a^t(1-a) + a^t b^t(b-a) = b^t(1-b) - a^t(1-b+b-a) + a^t b^t(b-a) \\ &= (b^t - a^t)(1-b) - a^t(1-b^t)(b-a) = a^t(1-b)(((\frac{b}{a})^t - 1) - a(\frac{b}{a} - 1)\frac{1-b^t}{1-b}) \\ &= \underbrace{a^t(1-b)(\frac{b}{a} - 1)}_{>0}(\frac{1-(b/a)^t}{1-b/a} - a\frac{1-b^t}{1-b}) \end{aligned}$$

and there is:

$$\frac{1-(b/a)^t}{1-b/a} = \sum_{i=0}^{t-1}(\frac{b}{a})^i > \sum_{i=0}^{t-1} 1 > a\sum_{i=0}^{t-1} b^i = a\frac{1-b^t}{1-b}$$

Thus, $g(t) > 0, \forall t \in \mathbb{Z}_+$, implies $f(t) > f(t+1), \forall t \geq 0$, i.e., $f(t)$ is a strictly increasing function. Furthermore, $f(t) < \lim_{t\to\infty} f(t) = 1, \forall t \in \mathbb{Z}_+$.

Similarly, if $0 < b < a < 1$, by the above analysis, $\frac{1}{f(t)} = \frac{1-b^t}{1-a^t}$ is strictly increasing function, implies $f(t)$ is strictly decreasing function. Furthermore, $f(t) > \lim_{t\to\infty} f(t) = 1, \forall t \in \mathbb{Z}_+$.

□

Lemma 5. *For the function $g(t) = \frac{\sum_{i=0}^{t-1} a^i + a^t K_a}{\sum_{i=0}^{t-1} b^i + b^t K_b}, t \in \mathbb{Z}_+$, where $a, b \in (0,1)$ and K_a, K_b are constants. Let $C_1 \geq C_2 > 0$ be constants. If $0 < a < b < 1$, $\frac{K_a}{K_b} > C_1$ and $K_b < \frac{1}{1-b}$, then $g(t)$ is strictly decreasing in t and $\exists t_0$ such that $g(t_0) < C_2$ if and only if $\frac{1-b}{1-a} < C_2$; if $0 < b < a < 1$, $\frac{K_a}{K_b} < C_2$ and $K_a < \frac{1}{1-a}$, then $g(t)$ is strictly increasing in t and $\exists t_0$ such that $g(t_0) > C_1$ if and only if $\frac{1-b}{1-a} > C_1$.*

Proof. First consider the case $0 < a < b < 1$.

To compare $g(t)$ and C_2, consider:

$$g(t) - C_2 = \frac{\sum_{i=0}^{t-1} a^i + a^t K_a - C_2(\sum_{i=0}^{t-1} b^i + b^t K_b)}{\sum_{i=0}^{t-1} b^i + b^t K_b} = \frac{\frac{1}{1-a} + \frac{a^t}{1-a^t} K_a - C_2(\frac{1-b^t}{1-a^t}\frac{1}{1-b} + \frac{b^t}{1-a^t} K_b)}{\frac{1}{1-a^t}(\sum_{i=0}^{t-1} b^i + b^t K_b)} \tag{5}$$

Let the numerator of (5) be $h(t)$ and re-organize it gives:

$$\begin{aligned} h(t) &= \frac{1}{1-a} + \frac{a^t}{1-a^t}K_a - \frac{1-b^t}{1-a^t}\frac{C_2}{1-b} - \frac{b^t}{1-a^t}C_2 K_b \\ &= \frac{1}{1-a} + \frac{1}{1-a^t}(a^t K_a - K_a + K_a - C_2 K_b + C_2 K_b - b^t C_2 K_b) - \frac{1-b^t}{1-a^t}\frac{C_2}{1-b} \\ &= \frac{1}{1-a} - K_a + \frac{K_a - C_2 K_b}{1-a^t} + C_2(K_b - \frac{1}{1-b})\frac{1-b^t}{1-a^t} \end{aligned}$$

Since $\frac{1}{1-a^t}$ is strictly decreasing in t and $\frac{1-b^t}{1-a^t}$ is strictly increasing in t by Lemma 4, if $\frac{K_a}{K_b} > C_1 \geq C_2$ and $K_b < \frac{1}{1-b}$ also hold, then $h(t)$ is strictly decreasing in t. Since $\sum_{i=0}^{t-1} b^i + b^t K_b$ is increasing in t when $K_b < \frac{1}{1-b}$. $g(t)$ is strictly decreasing in t . If $\exists t_0$ such that $g(t_0) < C_2$, i.e., $\exists t_0$ such that $h(t_0) < 0$, then $0 > h(t_0) > \lim_{t\to\infty} h(t) = \frac{1}{1-a} - \frac{C_2}{1-b}$ must also hold. Therefore, If $\frac{1-b}{1-a} < C_2$, then $\exists t_0$ such that $g(t_0) < C_2$; if $\frac{1-b}{1-a} \geq C_2$, then $g(t) \geq C_2, \forall t$.

Similarly, if $0 < b < a < 1$, using the above analysis, it can be concluded that if $\frac{K_b}{K_a} > \frac{1}{C_2}$ and $K_a < \frac{1}{1-a}$, then $\frac{1}{g(t)}$ is strictly decreasing in t and $\exists t_0$ such that $\frac{1}{g(t_0)} < \frac{1}{C_1}$ if and only if $\frac{1-a}{1-b} < \frac{1}{C_1}$. Re-written it becomes: if $\frac{K_a}{K_b} < C_2$ and $K_a < \frac{1}{1-a}$, then $g(t)$ is strictly increasing in t and $\exists t_0$ such that $g(t_0) > C_1$ if and only if $\frac{1-b}{1-a} > C_1$.

□

Prove Theorem 2 by induction. To simplify the notation, denote $\nu_k^s = \nu(L_k(\theta_k(s)))$. Since $N_k(2) = N_k(1)\nu_k^1 + \beta_k > N_k(1), k \in \{a, b\}$, it implies that $N_k(1) < \frac{\beta_k}{1-\nu_k^1}$.

Base case:

(1) If $L_a(\theta_a(1)) > L_b(\theta_b(1)) \to \nu_a^1 < \nu_b^1$, then $\frac{\overline{\alpha}_a(2)}{\overline{\alpha}_b(2)} = \frac{N_a(1)\nu_a^1 + \beta_a}{N_b(1)\nu_b^1 + \beta_b} < \frac{N_a(1)}{N_b(1)} = \frac{\overline{\alpha}_a(1)}{\overline{\alpha}_b(1)}$. By Lemma 2, it results in $\theta_a(2) < \theta_a(1) \to L_a(\theta_a(2)) > L_a(\theta_a(1)) \to \nu_a^2 < \nu_a^1$ and $\theta_b(2) < \theta_b(1) \to L_b(\theta_b(2)) < L_b(\theta_b(1)) \to \nu_b^2 > \nu_b^1$. Therefore, $L_a(\theta_a(2)) > L_a(\theta_a(1)) > L_b(\theta_b(1)) > L_b(\theta_b(2)) \to \nu_b^2 > \nu_b^1 > \nu_a^1 > \nu_a^2$ and $N_b(2) = N_b(1)\nu_b^1 + \beta_b < \frac{\beta_b}{1-\nu_b^1} < \frac{\beta_b}{1-\nu_b^2}$.

(2) If $L_a(\theta_a(1)) < L_b(\theta_b(1)) \to \nu_a^1 > \nu_b^1$, then $\frac{\overline{\alpha}_a(2)}{\overline{\alpha}_b(2)} = \frac{N_a(1)\nu_a^1 + \beta_a}{N_b(1)\nu_b^1 + \beta_b} > \frac{N_a(1)}{N_b(1)} = \frac{\overline{\alpha}_a(1)}{\overline{\alpha}_b(1)}$. By Lemma 2, it results in $\theta_a(2) > \theta_a(1) \to L_a(\theta_a(2)) < L_a(\theta_a(1)) \to \nu_a^2 > \nu_a^1$ and $\theta_b(2) > \theta_b(1) \to L_b(\theta_b(2)) > L_b(\theta_b(1)) \to \nu_b^2 < \nu_b^1$. Therefore, $L_a(\theta_a(2)) < L_a(\theta_a(1)) < L_b(\theta_b(1)) < L_b(\theta_b(2)) \to \nu_b^2 < \nu_b^1 < \nu_a^1 < \nu_a^2$ and $N_a(2) = N_a(1)\nu_a^1 + \beta_a < \frac{\beta_a}{1-\nu_a^1} < \frac{\beta_a}{1-\nu_a^2}$.

(3) If $L_a(\theta_a(1)) = L_b(\theta_b(1)) \to \nu_a^1 = \nu_b^1$, then $\frac{\overline{\alpha}_a(2)}{\overline{\alpha}_b(2)} = \frac{N_a(1)\nu_a^1 + \beta_a}{N_b(1)\nu_b^1 + \beta_b} = \frac{N_a(1)}{N_b(1)} = \frac{\overline{\alpha}_a(1)}{\overline{\alpha}_b(1)}$ and $\theta_a(2) = \theta_a(1)$, $\theta_b(2) = \theta_b(1) \to L_a(\theta_a(2)) = L_a(\theta_a(1)) = L_b(\theta_b(1)) = L_b(\theta_b(2))$.

Induction Step:

(1) $L_a(\theta_a(1)) > L_b(\theta_b(1))$

Suppose Theorem 2 is true for time step n, i.e., $\frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)} < \frac{\overline{\alpha}_a(n)}{\overline{\alpha}_b(n)}$, $N_b(n+1) < \frac{\beta_b}{1-\nu(L_b(\theta_b(n+1)))}$ and $L_a(\theta_a(n+1)) > L_a(\theta_a(n)) > L_b(\theta_b(n)) > L_b(\theta_b(n+1))$. Show that for time step $n+1$, Theorem 2 also holds.

Define function $\eta(s) = \frac{\frac{N_a(n)}{\beta_a}(\nu_a^n)^s + \sum_{i=0}^{s-1}(\nu_a^n)^i}{\frac{N_b(n)}{\beta_b}(\nu_b^n)^s + \sum_{i=0}^{s-1}(\nu_b^n)^i}$, since $\frac{N_b(n)}{\beta_b} < \frac{1}{1-\nu_b^n}$, $\frac{N_a(n)\beta_b}{N_b(n)\beta_a} > \frac{N_a(n+1)\beta_b}{N_b(n+1)\beta_a}$ and $\nu_a^1 < \nu_b^1$, by Lemma 5, $\eta(s)$ is strictly decreasing in s. Therefore, $\eta(2) < \eta(1) = \frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)}$. Since $\nu_a^{n+1} < \nu_a^n$ and $\nu_b^{n+1} > \nu_b^n$, it implies $\frac{\overline{\alpha}_a(n+2)}{\overline{\alpha}_b(n+2)} = \frac{(N_a(n)\nu_a^n + \beta_a)\nu_a^{n+1} + \beta_a}{(N_b(n)\nu_b^n + \beta_b)\nu_b^{n+1} + \beta_b} < \eta(2) < \frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)}$. By Lemma 2, it results in $\theta_a(n+2) < \theta_a(n+1) \to L_a(\theta_a(n+2)) > L_a(\theta_a(n+1)) \to \nu_a^{n+2} < \nu_a^{n+1}$ and $\theta_b(n+2) < \theta_b(n+1) \to L_b(\theta_b(n+2)) < L_b(\theta_b(n+1)) \to \nu_b^{n+2} > \nu_b^{n+1}$.

Therefore, $L_a(\theta_a(n+2)) > L_a(\theta_a(n+1)) > L_b(\theta_b(n+1)) > L_b(\theta_b(n+2))$ and $N_b(n+2) = N_b(n+1)\nu_b^{n+1}+\beta_b < \frac{\beta_b}{1-\nu_b^{n+1}} < \frac{\beta_b}{1-\nu_b^{n+2}}$.

(2) $L_a(\theta_a(1)) < L_b(\theta_b(1))$

Suppose Theorem 2 is true for time step n, i.e., $\frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)} > \frac{\overline{\alpha}_a(n)}{\overline{\alpha}_b(n)}$, $N_a(n+1) < \frac{\beta_a}{1-\nu(L_a(\theta_a(n+1)))}$ and $L_a(\theta_a(n+1)) < L_a(\theta_a(n)) < L_b(\theta_b(n)) < L_b(\theta_b(n+1))$. Show that for time step $n+1$, Theorem 2 also holds.

Define function $\eta(s) = \frac{\frac{N_a(n)}{\beta_a}(\nu_a^n)^s+\sum_{i=0}^{s-1}(\nu_a^n)^i}{\frac{N_b(n)}{\beta_b}(\nu_b^n)^s+\sum_{i=0}^{s-1}(\nu_b^n)^i}$, since $\frac{N_a(n)}{\beta_a} < \frac{1}{1-\nu_a^n}$, $\frac{N_a(n)\beta_b}{N_b(n)\beta_a} < \frac{N_a(n+1)\beta_b}{N_b(n+1)\beta_a}$ and $\nu_a^1 > \nu_b^1$, by Lemma 5, $\eta(s)$ is strictly increasing in s. Therefore, $\eta(2) > \eta(1) = \frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)}$. Since $\nu_a^{n+1} > \nu_a^n$ and $\nu_b^{n+1} < \nu_b^n$, it implies $\frac{\overline{\alpha}_a(n+2)}{\overline{\alpha}_b(n+2)} = \frac{(N_a(n)\nu_a^n+\beta_a)\nu_a^{n+1}+\beta_a}{(N_b(n)\nu_b^n+\beta_b)\nu_b^{n+1}+\beta_b} > \eta(2) > \frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)}$. By Lemma 2, it results in $\theta_a(n+2) > \theta_a(n+1) \rightarrow L_a(\theta_a(n+2)) < L_a(\theta_a(n+1)) \rightarrow \nu_a^{n+2} > \nu_a^{n+1}$ and $\theta_b(n+2) > \theta_b(n+1) \rightarrow L_b(\theta_b(n+2)) > L_b(\theta_b(n+1)) \rightarrow \nu_b^{n+2} < \nu_b^{n+1}$. Therefore, $L_a(\theta_a(n+2)) < L_a(\theta_a(n+1)) < L_b(\theta_b(n+1)) < L_b(\theta_b(n+2))$ and $N_a(n+2) = N_a(n+1)\nu_a^{n+1}+\beta_a < \frac{\beta_a}{1-\nu_a^{n+1}} < \frac{\beta_a}{1-\nu_a^{n+2}}$.

(3) $L_a(\theta_a(1)) = L_b(\theta_b(1))$

Suppose Theorem 2 is true for time step n, i.e., $\frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)} = \frac{\overline{\alpha}_a(n)}{\overline{\alpha}_b(n)}$ and $\theta_k(n+1) = \theta_k(n), k \in \{a,b\}$. Show that for time step $n+1$, Theorem 2 also holds.

Since $\theta_k(n+1) = \theta_k(n), k \in \{a,b\} \rightarrow \nu_k^{n+1} = \nu_k^n, k \in \{a,b\}$. $\frac{\overline{\alpha}_a(n+2)}{\overline{\alpha}_b(n+2)} = \frac{N_a(n+1)\nu_a^{n+1}+\beta_a}{N_b(n+1)\nu_b^{n+1}+\beta_b} = \frac{\overline{\alpha}_a(n+1)}{\overline{\alpha}_b(n+1)}$ and $L_a(\theta_a(n+1)) = L_a(\theta_a(n)) = L_b(\theta_b(n)) = L_b(\theta_b(n+1))$.

Finally, combine with Lemma 1 that $(\theta_a(t), \theta_b(t))$ is bounded by a compact interval, Theorem 2 is complete.

F Proof of Corollary 1

To simplify the notation, denote $\nu_k^s = \nu(L_k(\theta_k(s))), k \in \{a,b\}$.

Since $\lim_{t\to\infty}(\theta_a(t),\theta_b(t)) = (\theta_a^\infty, \theta_b^\infty)$, $\forall \epsilon > 0$, $\exists t_0$ such that $\forall t > t_0$, $|\nu_k^t - \nu(L_k(\theta_k^\infty))| < \epsilon$ for both $k = a$ and $k = b$.

For $t > t_0$, $N_k(t) = N_k(t_0)\prod_{i=0}^{t-t_0-1}\nu_k^{t_0+i} + \beta_k(1+\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}\nu_k^{t_0+i}) = N_k(t_0)\prod_{i=0}^{t-t_0-1}(\nu_k^{t_0}+\nu_k^{t_0+i}-\nu_k^{t_0}) + \beta_k(1+\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}(\nu_k^{t_0}+\nu_k^{t_0+i}-\nu_k^{t_0}))$, the following holds $\forall \epsilon > 0$:

$$N_k(t_0)\prod_{i=0}^{t-t_0-1}(\nu_k^{t_0}-\epsilon)+\beta_k(1+\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}(\nu_k^{t_0}-\epsilon)) < N_k(t) < N_k(t_0)\prod_{i=0}^{t-t_0-1}(\nu_k^{t_0}+\epsilon)+\beta_k(1+\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}(\nu_k^{t_0}+\epsilon))$$

$$\beta_k(1+\lim_{t\to\infty}\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}(\nu_k^{t_0}-\epsilon)) < \lim_{t\to\infty}N_k(t) < \beta_k(1+\lim_{t\to\infty}\sum_{j=1}^{t-t_0-1}\prod_{i=j}^{t-t_0-1}(\nu_k^{t_0}+\epsilon))$$

Therefore,

$\lim_{t\to\infty}N_k(t) = \beta_k(1+\sum_{j=1}^{\infty}(\nu(L(\theta_k^\infty)))^j) = \frac{\beta_k}{1-\nu(L_k(\theta_k^\infty))}$

$\lim_{t\to\infty}\overline{\alpha}_a(t) = \frac{1}{1+\frac{\beta_b}{\beta_a}\frac{1-\nu(L_a(\theta_a^\infty))}{1-\nu(L_b(\theta_b^\infty))}}$, $\quad \lim_{t\to\infty}\overline{\alpha}_b(t) = 1 - \lim_{t\to\infty}\overline{\alpha}_a(t)$

$\lim_{t\to\infty}L^t(\theta_a(t),\theta_b(t)) = \lim_{t\to\infty}\overline{\alpha}_a(t)L_a(\theta_a(t)) + \lim_{t\to\infty}\overline{\alpha}_b(t)L_b(\theta_b(t)) = L_b(\theta_b^\infty) + \frac{L_a(\theta_a^\infty)-L_b(\theta_b^\infty)}{1+\frac{\beta_b}{\beta_a}\frac{1-\nu(L_a(\theta_a^\infty))}{1-\nu(L_b(\theta_b^\infty))}}$

G Proof of Theorem 3

The `EqLos` one-shot fair solution for time step t is given by,

$$\begin{aligned} \min_{\theta_a(t),\theta_b(t)} \quad & \overline{\alpha}_a(t)L_a(\theta_a(t)) + \overline{\alpha}_b(t)L_b(\theta_b(t)) \\ s.t., \quad & L_a(\theta_a(t)) = L_b(\theta_b(t)), \end{aligned} \tag{6}$$

Since $L_a(\theta_a(t) = L_b(\theta_b(t))$, we can re-write above optimization problem as follows

$$\begin{array}{rl} \min_{\theta_a(t),\theta_b(t)} & L_k(\theta_k(t)) \\ s.t., & L_a(\theta_a(t)) = L_b(\theta_b(t)) \end{array} \tag{7}$$

where k can be either a or b and has the solution (θ_a^*, θ_b^*) satisfy $L_a(\theta_a^*(t)) = L_a(\theta_b^*(t)) = \max\{\min_\theta L_a(\theta), \min_\theta L_b(\theta)\}$. Since the optimization (7) doesn't depend on $\overline{\alpha}_a(t), \overline{\alpha}_b(t)$ and its solution is the same $\forall t$. The sequential problem ($\star\star$) $\min_{\{(\theta_a(t),\theta_b(t))\}} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T L(\theta_a(t), \theta_b(t))$ has the solution $(\theta_a(t), \theta_b(t)) = (\theta_a^*, \theta_b^*), \forall t$.

Note that retention rate for both groups is equal to $\nu(L_a(\theta_a^*))$ under `Eqlos`. Therefore, $N_k(t+1) = \beta_k(1 + \sum_{j=1}^{t-1}(\nu(L_a(\theta_a^*)))^i) + N_k(1)\nu(L_a(\theta_a^*))^{t-1}$, $\forall t \geq 1$. Therefore, as t goes to infinity, $N_k(t)$ goes to $\frac{\beta_k}{1-\nu(L_a(\theta_a^*))}$, implying $\lim_{t\to\infty} \frac{N_a(t)}{N_b(t)} = \frac{\beta_a}{\beta_b}$.

Therefore, $\lim_{t\to\infty} \overline{\alpha}_a(t) = \frac{\beta_a}{\beta_a+\beta_b}$ and $\lim_{t\to\infty} \overline{\alpha}_b(t) = \frac{\beta_b}{\beta_a+\beta_b}$ $\forall t$.

H Proof of Theorem 4

Theorem 4 is proved by showing that the `Simple` , `EqOpt` and `StatPar` greedy decision at each round all have the form of (3). Which are presented in Lemma 6, 7, 8 respectively.

• `Simple` **greedy fair decision**

For the simple decision, $\theta_a = \theta_b = \theta$ should be satisfied.

Lemma 6. *Let* $A = \frac{\frac{g_b^0}{\overline{b}_0-\underline{b}_0}}{\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}}$, $B = \frac{\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1}}{\frac{g_a^1}{\overline{a}_1-\underline{a}_1}}$, $C = \frac{\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1}}{\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}}$, *then the* `Simple` *greedy decision maker at each time t selects parameter* $\theta_a(t) = \theta_b(t) = \theta(t)$ *based on* $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and A, B, C according to the following:*

(i) If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $0 < B < C < A$ *and* $\theta(t) = \begin{cases} \overline{b}_0, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < B \\ \overline{a}_0, \text{ if } B < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < C \\ \underline{b}_1, \text{ if } C < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A \\ \underline{a}_1, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > A \end{cases}$

(ii) If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $C < B < 0 < A$ *and* $\theta(t) = \begin{cases} \underline{b}_1, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A \\ \underline{a}_1, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > A \end{cases}$

(iii) If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $A < C < 0 < B$ *and* $\theta(t) = \begin{cases} \overline{b}_0, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < B \\ \overline{a}_0, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > B \end{cases}$

(iv) If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $A < B < 0 < C$ *and* $\theta(t) = \begin{cases} \underline{b}_1, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < C \\ \overline{a}_0, \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > C \end{cases}$

Proof. There are 3 possibilities for an optimal θ.

(1) $\theta \in [\underline{a}_1, \underline{b}_1]$

$L^t(\theta_a, \theta_b) = (\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} - \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} - \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0})\theta + (\frac{\alpha_a^0(t)\overline{a}_0}{\overline{a}_0-\underline{a}_0} - \frac{\alpha_a^1(t)\underline{a}_1}{\overline{a}_1-\underline{a}_1} + \frac{\alpha_b^0(t)\overline{b}_0}{\overline{b}_0-\underline{b}_0})$ and $\theta(t) = \begin{cases} \underline{a}_1, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} - \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} > \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \\ \underline{b}_1, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} - \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} < \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \end{cases}$

(2) $\theta \in [\underline{b}_1, \overline{a}_0]$

$L^t(\theta_a, \theta_b) = (\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} + \frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} - \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} - \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0})\theta + (\frac{\alpha_a^0(t)\overline{a}_0}{\overline{a}_0-\underline{a}_0} - \frac{\alpha_a^1(t)\underline{a}_1}{\overline{a}_1-\underline{a}_1} - \frac{\alpha_b^1(t)\underline{b}_1}{\overline{b}_1-\underline{b}_1} + \frac{\alpha_b^0(t)\overline{b}_0}{\overline{b}_0-\underline{b}_0})$ and

$$\theta(t) = \begin{cases} \underline{b}_1, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} + \frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} > \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} + \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \\ \overline{a}_0, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} + \frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} < \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} + \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \end{cases}$$

(3) $\theta \in [\overline{a}_0, \overline{b}_0]$

$L^t(\theta_a,\theta_b) = (\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1}-\frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0})\theta+(\frac{\alpha_b^0(t)\overline{b}_0}{\overline{b}_0-\underline{b}_0}-\frac{\alpha_a^1(t)\underline{a}_1}{\overline{a}_1-\underline{a}_1}-\frac{\alpha_b^1(t)\underline{b}_1}{\overline{b}_1-\underline{b}_1})$ and $\theta(t) = \begin{cases}\overline{a}_0, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} > \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0}\\ \overline{b}_0, \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} < \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0}\end{cases}$

Therefore, replacing $\alpha_k^j(t)$ as $\alpha_k(t)g_k^j$ for $k\in\{a,b\}, j\in\{0,1\}$. combining and re-organizing the above conditions implies:

$$\theta(t) = \begin{cases}\underline{a}_1, \text{ if } \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}) > \overline{\alpha}_b(t)\frac{g_b^0}{\overline{b}_0-\underline{b}_0}\\ \overline{b}_0, \text{ if } \overline{\alpha}_a(t)\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \overline{\alpha}_b(t)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1})\\ \underline{b}_1, \text{ if } \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}) < \overline{\alpha}_b(t)\frac{g_b^0}{\overline{b}_0-\underline{b}_0} \;\&\; \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}) > \overline{\alpha}_b(t)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1})\\ \overline{a}_0, \text{ if } \overline{\alpha}_a(t)\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \overline{\alpha}_b(t)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1}) \;\&\; \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}) < \overline{\alpha}_b(t)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1})\end{cases}$$

Substitute with A, B, C and reorganize gives Lemma 6.

□

• `EqOpt` **greedy fair decision**

For the `EqOpt` decision, $\int_{\theta_a}^{\overline{a}_0}\frac{1}{\overline{a}_0-\underline{a}_0}dx = \int_{\theta_b}^{\overline{b}_0}\frac{1}{\overline{b}_0-\underline{b}_0}dx$ should be satisfied. Let $\overline{\theta}_a$ and $\overline{\theta}_b$ be defined such that $(\overline{\theta}_a,\underline{b}_1)$, $(\underline{a}_1,\overline{\theta}_b)$ are two pairs satisfying the fairness constraint, i.e., $\frac{\overline{a}_0-\overline{\theta}_a}{\overline{a}_0-\underline{a}_0} = \frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}$ and $\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0} = \frac{\overline{b}_0-\overline{\theta}_b}{\overline{b}_0-\underline{b}_0}$. Without loss of generality, assuming $\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0} < \frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}$, which implies $\overline{\theta}_b < \underline{b}_1$ and $\overline{\theta}_a > \underline{a}_1$.

Lemma 7. *Let* $A = \frac{(\overline{b}_0-\underline{b}_0)}{(\overline{a}_0-\underline{a}_0)}\frac{\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1}}{\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}}$ *and* $B = \frac{\frac{g_b^0}{\overline{a}_0-\underline{a}_0}}{\frac{g_a^1}{\overline{a}_1-\underline{a}_1}-\frac{g_a^0}{\overline{a}_0-\underline{a}_0}}$, *then the* `EqOpt` *greedy one-shot decision maker at each time* t *selects parameter pair* $(\theta_a(t),\theta_b(t))$ *based on* $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ *and A, B according to the following:*

(i) If $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $A<B (B>0)$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases}(\overline{a}_0,\overline{b}_0), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A\\ (\overline{\theta}_a,\underline{b}_1), \text{ if } A < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < B\\ (\underline{a}_1,\overline{\theta}_b), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > B\end{cases}$

(ii) If $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $A>B (B<0)$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases}(\overline{\theta}_a,\underline{b}_1), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A\\ (\overline{a}_0,\overline{b}_0), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > A\end{cases}$

Proof. There are 2 possibilities for an optimal (θ_a,θ_b) pair:

(1) $\theta_a\in[\underline{a}_1,\overline{\theta}_a], \theta_b\in[\overline{\theta}_b,\underline{b}_1]$

$L^t(\theta_a,\theta_b) = (\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}-\frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0})\theta_a+\frac{\overline{a}_0(\alpha_a^0(t)+\alpha_b^0(t))}{\overline{a}_0-\underline{a}_0}-\frac{\alpha_a^1(t)\underline{a}_1}{\overline{a}_1-\underline{a}_1}$

$(\theta_a(t),\theta_b(t)) = \begin{cases}(\underline{a}_1,\overline{\theta}_b), \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} > \frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0}\\ (\overline{\theta}_a,\underline{b}_1), \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} < \frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0}\end{cases}$

(2) $\theta_a\in[\overline{\theta}_a,\overline{a}_0], \theta_b\in[\underline{b}_1,\overline{b}_0]$.

$L^t(\theta_a,\theta_b) = (\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{a}_0-\underline{a}_0}\frac{\overline{b}_0-\underline{b}_0}{\overline{b}_1-\underline{b}_1}-\frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0})\theta_a+\frac{\overline{a}_0(\alpha_a^0(t)+\alpha_b^0(t))}{\overline{a}_0-\underline{a}_0}-\frac{\alpha_a^1(t)\underline{a}_1}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)(\overline{b}_0-\underline{b}_1)}{\overline{b}_1-\underline{b}_1}-\frac{\alpha_b^1(t)\overline{a}_0}{\overline{a}_0-\underline{a}_0}\frac{\overline{b}_0-\underline{b}_0}{\overline{b}_1-\underline{b}_1}$

$(\theta_a(t),\theta_b(t)) = \begin{cases}(\overline{\theta}_a,\underline{b}_1), \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{a}_0-\underline{a}_0}\frac{\overline{b}_0-\underline{b}_0}{\overline{b}_1-\underline{b}_1} > \frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0}\\ (\overline{a}_0,\overline{b}_0), \text{ if } \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}+\frac{\alpha_b^1(t)}{\overline{a}_0-\underline{a}_0}\frac{\overline{b}_0-\underline{b}_0}{\overline{b}_1-\underline{b}_1} < \frac{\alpha_a^0(t)+\alpha_b^0(t)}{\overline{a}_0-\underline{a}_0}\end{cases}$

Therefore, replacing $\alpha_k^j(t)$ as $\alpha_k(t)g_k^j$ for $k \in \{a,b\}, j \in \{0,1\}$. combining and re-organizing the above conditions implies:

$$(\theta_a(t),\theta_b(t)) = \begin{cases} (\underline{a}_1, \overline{\theta}_b), \text{ if } \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}) > \overline{\alpha}_b(t)\frac{g_b^0}{\overline{a}_0-\underline{a}_0} \\ (\overline{\theta}_a, \underline{b}_1), \text{ if } \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}) > \overline{\alpha}_b(t)\frac{\overline{b}_0-\underline{b}_0}{\overline{a}_0-\underline{a}_0}(\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1}) \;\&\; \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}) < \overline{\alpha}_b(t)\frac{g_b^0}{\overline{a}_0-\underline{a}_0} \\ (\overline{a}_0, \overline{b}_0), \text{ if } \overline{\alpha}_a(t)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0}) < \overline{\alpha}_b(t)\frac{\overline{b}_0-\underline{b}_0}{\overline{a}_0-\underline{a}_0}(\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1}) \end{cases}$$

Substitute with A, B gives Lemma 7. $\square$

• `StatPar` **greedy fair decision**

Denote $P_k(\theta_k) = \sum_{i\in\{0,1\}} g_k^i \int_{\theta_k}^{\infty} f_k^i(x)dx$, then for the `StatPar` decision, $P_a(\theta_a) = P_b(\theta_b)$ should be satisfied.

Let $\overline{\theta}_a$, $\overline{\theta}_b$, $\tilde{\theta}_a$, $\tilde{\theta}_b$ be defined such that $(\overline{\theta}_a, \overline{b}_0)$, $(\overline{a}_0, \overline{\theta}_b)$, $(\tilde{\theta}_a, \underline{b}_1)$, $(\underline{a}_1, \tilde{\theta}_b)$ are pairs satisfying the statistical parity fairness constraint, i.e., $P_a(\overline{\theta}_a) = P_b(\overline{b}_0)$, $P_a(\overline{a}_0) = P_b(\overline{\theta}_b)$, $P_a(\tilde{\theta}_a) = P_b(\underline{b}_1)$ and $P_a(\underline{a}_1) = P_b(\tilde{\theta}_b)$. Without loss of generality, assuming $\overline{\theta}_b \in [\overline{b}_0, \overline{b}_1]$, which implies $\overline{\theta}_a < \overline{a}_0$. There are three cases and under which $\overline{\theta}_a$, $\overline{\theta}_b$, $\tilde{\theta}_a$, $\tilde{\theta}_b$ satisfy: (1) $\overline{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_b \in [\overline{b}_0, \overline{b}_1]$; (2) $\overline{\theta}_a \in [\underline{a}_1, \overline{a}_0]$, $\tilde{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_b \in [\underline{b}_1, \overline{b}_0]$; (3) $\overline{\theta}_a \in [\underline{a}_1, \overline{a}_0]$, $\tilde{\theta}_a \in [\underline{a}_1, \overline{a}_0]$, $\tilde{\theta}_b \in [\underline{b}_0, \underline{b}_1]$

Lemma 8. *Let* $A = \frac{\frac{g_b^1}{\overline{b}_1-\underline{b}_1} - \frac{g_b^0}{\overline{b}_0-\underline{b}_0}}{\frac{g_b^0}{\overline{b}_0-\underline{b}_0} + \frac{g_b^1}{\overline{b}_1-\underline{b}_1}}$ *and* $B = \frac{\frac{g_a^1}{\overline{a}_1-\underline{a}_1} + \frac{g_a^0}{\overline{a}_0-\underline{a}_0}}{\frac{g_a^0}{\overline{a}_0-\underline{a}_0} - \frac{g_a^1}{\overline{a}_1-\underline{a}_1}}$, *then the* `StatPar` *greedy fair decision maker at each time t selects parameter pair $(\theta_a(t), \theta_b(t))$ based on $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$, and A, B according to the following:*

For the case when $\overline{\theta}_b \in [\overline{b}_0, \overline{b}_1]$, $\overline{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_b \in [\overline{b}_0, \overline{b}_1]$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $0 < A < 1 < B$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases} (\tilde{\theta}_a, \underline{b}_1), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A \\ (\overline{\theta}_a, \overline{b}_0), \text{ if } A < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < 1 \\ (\underline{a}_1, \tilde{\theta}_b), \text{ if } 1 < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < B \\ (\overline{a}_0, \overline{\theta}_b), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > B \end{cases}$

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $B < 0 < A < 1$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases} (\tilde{\theta}_a, \underline{b}_1), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < A \\ (\overline{\theta}_a, \overline{b}_0), \text{ if } A < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < 1 \\ (\underline{a}_1, \tilde{\theta}_b), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > 1 \end{cases}$

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, *then* $A < 0 < 1 < B$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases} (\overline{\theta}_a, \overline{b}_0), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < 1 \\ (\underline{a}_1, \tilde{\theta}_b), \text{ if } 1 < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < B \\ (\overline{a}_0, \overline{\theta}_b), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > B \end{cases}$

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$ *then* $A < 0$, $B < 0$ *and* $(\theta_a(t),\theta_b(t)) = \begin{cases} (\overline{\theta}_a, \overline{b}_0), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} < 1 \\ (\underline{a}_1, \tilde{\theta}_b), \text{ if } \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} > 1 \end{cases}$

The conclusions for the other two cases are similar.

Proof. When $\overline{\theta}_b \in [\overline{b}_0, \overline{b}_1]$, $\overline{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_a \in [\underline{a}_0, \underline{a}_1]$, $\tilde{\theta}_b \in [\overline{b}_0, \overline{b}_1]$. There are 3 possibilities for an optimal (θ_a, θ_b) pair:

(1) $\theta_a \in [\underline{a}_1, \overline{a}_0]$, $\theta_b \in [\tilde{\theta}_b, \overline{\theta}_b]$

$$L^t(\theta_a,\theta_b) = ((1+\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} - (1-\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})\frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0})\theta_a + \alpha_b^1(t) - \frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1}(\underline{a}_1 + \overline{a}_1\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}) + \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0}(\overline{a}_0 - \overline{a}_0\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})$$

$$(\theta_a(t),\theta_b(t)) = \begin{cases} (\underline{a}_1, \tilde{\theta}_b), \text{ if } (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} > (\overline{\alpha}_a(t)-\overline{\alpha}_b(t))\frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} \\ (\overline{a}_0, \overline{\theta}_b), \text{ if } (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{\alpha_a^1(t)}{\overline{a}_1-\underline{a}_1} < (\overline{\alpha}_a(t)-\overline{\alpha}_b(t))\frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0} \end{cases}$$

(2) $\theta_a \in [\overline{\theta}_a, \underline{a}_1]$, $\theta_b \in [\overline{b}_0, \tilde{\theta}_b]$

$L^t(\theta_a,\theta_b) = \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0}(\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}-1)\theta_a + \alpha_b^1(t) + \frac{\alpha_a^0(t)}{\overline{a}_0-\underline{a}_0}(\overline{a}_0 - \overline{a}_0\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}) - \alpha_a^1\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}$

$(\theta_a(t),\theta_b(t)) = \begin{cases} (\overline{\theta}_a,\overline{b}_0), \text{ if } \overline{\alpha}_a(t) > \overline{\alpha}_b(t) \\ (\underline{a}_1,\tilde{\theta}_b), \text{ if } \overline{\alpha}_a(t) < \overline{\alpha}_b(t) \end{cases}$

(3) $\theta_a \in [\tilde{\theta}_a,\overline{\theta}_a]$, $\theta_b \in [\underline{b}_1,\overline{b}_0]$

$L^t(\theta_a,\theta_b) = ((1-\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} - (1+\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})\frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0})\theta_b - \alpha_a^1(t) - \frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1}(\underline{b}_1 - \overline{b}_1\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)}) + \frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0}(\overline{b}_0 + \overline{b}_0\frac{\overline{\alpha}_b(t)}{\overline{\alpha}_a(t)})$

$(\theta_a(t),\theta_b(t)) = \begin{cases} (\tilde{\theta}_a,\underline{b}_1), \text{ if } (\overline{\alpha}_b(t)-\overline{\alpha}_a(t))\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} > (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \\ (\overline{\theta}_a,\overline{b}_0), \text{ if } (\overline{\alpha}_b(t)-\overline{\alpha}_a(t))\frac{\alpha_b^1(t)}{\overline{b}_1-\underline{b}_1} < (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{\alpha_b^0(t)}{\overline{b}_0-\underline{b}_0} \end{cases}$

Therefore, replacing $\alpha_k^j(t)$ as $\alpha_k(t)g_k^j$ for $k \in \{a,b\}, j \in \{0,1\}$. combining and re-organizing the above conditions implies:

$$(\theta_a^t,\theta_b^t) = \begin{cases} (\tilde{\theta}_a,\underline{b}_1), \text{ if } (\overline{\alpha}_b(t)-\overline{\alpha}_a(t))\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{g_b^0}{\overline{b}_0-\underline{b}_0} \\ (\overline{a}_0,\overline{\theta}_b), \text{ if } (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < (\overline{\alpha}_a(t)-\overline{\alpha}_b(t))\frac{g_a^0}{\overline{a}_0-\underline{a}_0} \\ (\overline{\theta}_a,\overline{b}_0), \text{ if } (\overline{\alpha}_b(t)-\overline{\alpha}_a(t))\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{g_b^0}{\overline{b}_0-\underline{b}_0} \ \& \ \overline{\alpha}_a(t) < \overline{\alpha}_b(t) \\ (\underline{a}_1,\tilde{\theta}_b), \text{ if } (\overline{\alpha}_a(t)+\overline{\alpha}_b(t))\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > (\overline{\alpha}_a(t)-\overline{\alpha}_b(t))\frac{g_a^0}{\overline{a}_0-\underline{a}_0} \ \& \ \overline{\alpha}_a(t) > \overline{\alpha}_b(t) \end{cases}$$

Substitute with A, B gives Lemma 8. □

I Proof of Theorem 5

Lemma 9. *For the* `Simple`*,* `EqOpt` *and* `StatPar` *greedy fair decision as given in* (3), $\{\nu(L_a(\theta_a^m))\}_{m=1}^M$ *is monotonic increasing in* m *and* $\{\nu(L_b(\theta_b^m))\}_{m=1}^M$ *is monotomic decreasing in* m.

Proof. It is proved by showing each case separately.

• `Simple` **greedy fair decision**

$$\underline{a}_1 : \begin{cases} L_a(\underline{a}_1) = g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0} \\ L_b(\underline{a}_1) = g_b^0\frac{\overline{b}_0-\underline{a}_1}{\overline{b}_0-\underline{b}_0} \end{cases} \quad \overline{b}_0 : \begin{cases} L_a(\overline{b}_0) = g_a^1\frac{\overline{b}_0-\underline{a}_1}{\overline{a}_1-\underline{a}_1} \\ L_b(\overline{b}_0) = g_b^1\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_1-\underline{b}_1} \end{cases}$$

$$\underline{b}_1 : \begin{cases} L_a(\underline{b}_1) = g_a^1\frac{\underline{b}_1-\underline{a}_1}{\overline{a}_1-\underline{a}_1} + g_a^0\frac{\overline{a}_0-\underline{b}_1}{\overline{a}_0-\underline{a}_0} \\ L_b(\underline{b}_1) = g_b^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0} \end{cases} \quad \overline{a}_0 : \begin{cases} L_a(\overline{a}_0) = g_a^1\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_1-\underline{a}_1} \\ L_b(\overline{a}_0) = g_b^1\frac{\overline{a}_0-\underline{b}_1}{\overline{b}_1-\underline{b}_1} + g_b^0\frac{\overline{b}_0-\overline{a}_0}{\overline{b}_0-\underline{b}_0} \end{cases}$$

Re-organize it gives: $L_a(\overline{b}_0) - L_a(\overline{a}_0) = g_a^1\frac{\overline{b}_0-\overline{a}_0}{\overline{a}_1-\underline{a}_1} > 0$, $L_a(\overline{a}_0) - L_a(\underline{b}_1) = (\overline{a}_0 - \underline{b}_1)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0})$, $L_a(\underline{b}_1) - L_a(\underline{a}_1) = (\underline{b}_1-\underline{a}_1)(\frac{g_a^1}{\overline{a}_1-\underline{a}_1} - \frac{g_a^0}{\overline{a}_0-\underline{a}_0})$, $L_b(\underline{a}_1) - L_b(\underline{b}_1) = g_b^0\frac{\underline{b}_1-\underline{a}_1}{\overline{b}_0-\underline{b}_0} > 0$, $L_b(\underline{b}_1) - L_b(\overline{a}_0) = (\overline{a}_0-\underline{b}_1)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1})$, $L_b(\overline{a}_0) - L_b(\overline{b}_0) = (\overline{b}_0 - \overline{a}_0)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0} - \frac{g_b^1}{\overline{b}_1-\underline{b}_1})$

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{b}_0) > L_a(\overline{a}_0) > L_a(\underline{b}_1) > L_a(\underline{a}_1)$ and $L_b(\underline{a}_1) > L_b(\underline{b}_1) > L_b(\overline{a}_0) > L_b(\overline{b}_0)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} > \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\underline{b}_1) > L_a(\underline{a}_1)$ and $L_b(\underline{a}_1) > L_b(\underline{b}_1)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} < \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{b}_0) > L_a(\overline{a}_0)$ and $L_b(\overline{a}_0) > L_b(\overline{b}_0)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1} > \frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1} < \frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{a}_0) < L_a(\underline{b}_1)$ and $L_b(\underline{b}_1) < L_b(\overline{a}_0)$.

• `EqOpt` **greedy fair decision**

$$(\underline{a}_1,\overline{\theta}_b):\begin{cases}L_a(\underline{a}_1)=g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}\\ L_b(\overline{\theta}_b)=g_b^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}\end{cases}\quad(\overline{\theta}_a,\underline{b}_1):\begin{cases}L_a(\overline{\theta}_a)=g_a^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}+g_a^1\frac{\overline{\theta}_a-\underline{a}_1}{\overline{a}_1-\underline{a}_1}\\ L_b(\underline{b}_1)=g_b^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}\end{cases}\quad(\overline{a}_0,\overline{b}_0):\begin{cases}L_a(\overline{a}_0)=g_a^1\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_1-\underline{a}_1}\\ L_b(\overline{b}_0)=g_b^1\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_1-\underline{b}_1}\end{cases}$$

Re-organize it gives: $L_a(\overline{\theta}_a)-L_a(\overline{a}_0)=\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}(g_a^0-g_a^1\frac{\overline{a}_0-\underline{a}_0}{\overline{a}_1-\underline{a}_1})$, $L_a(\overline{\theta}_a)-L_a(\underline{a}_1)=(g_a^0-g_a^1\frac{\overline{a}_0-\underline{a}_0}{\overline{a}_1-\underline{a}_1})(\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}-\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0})$, $L_b(\overline{\theta}_b)-L_b(\underline{b}_1)=g_b^0(\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}-\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0})$, $L_b(\underline{b}_1)-L_b(\overline{b}_0)=(\overline{b}_0-\underline{b}_1)(\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1})$

Notice that we have assumed that $\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}<\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}$.

If $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}>\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\underline{a}_1)<L_a(\overline{\theta}_a)<L_a(\overline{a}_0)$.

If $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}<\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{\theta}_a)>L_a(\overline{a}_0)$.

Furthermore, In order to take $(\overline{a}_0,\overline{b}_0)$, $A>\frac{\eta_a(t)}{\eta_b(t)}>0$ should be satisfied. Therefore, if $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}>\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1}>0$ should hold and $L_b(\overline{\theta}_b)>L_b(\underline{b}_1)>L_b(\overline{b}_0)$.

if $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}<\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $\frac{g_b^0}{\overline{b}_0-\underline{b}_0}-\frac{g_b^1}{\overline{b}_1-\underline{b}_1}<0$ should hold and $L_b(\underline{b}_1)<L_b(\overline{b}_0)$.

• `StatPar` **greedy fair decision** When $\overline{\theta}_b\in[\overline{b}_0,\overline{b}_1]$, $\overline{\theta}_a<\overline{a}_0$, $\overline{\theta}_a\in[\underline{a}_0,\underline{a}_1]$, $\tilde{\theta}_a\in[\underline{a}_0,\underline{a}_1]$, $\tilde{\theta}_b\in[\overline{b}_0,\overline{b}_1]$.

$$(\tilde{\theta}_a,\underline{b}_1):\begin{cases}L_a(\tilde{\theta}_a)=g_b^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}+g_b^1-g_a^1\\ L_b(\underline{b}_1)=g_b^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}\end{cases}\quad(\overline{a}_0,\overline{\theta}_b):\begin{cases}L_a(\overline{a}_0)=g_a^1\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_1-\underline{a}_1}\\ L_b(\overline{\theta}_b)=g_b^1-g_a^1\frac{\overline{a}_1-\overline{a}_0}{\overline{a}_1-\underline{a}_1}\end{cases}$$

$$(\overline{\theta}_a,\overline{b}_0):\begin{cases}L_a(\overline{\theta}_a)=g_b^1\frac{\overline{b}_1-\overline{b}_0}{\overline{b}_1-\underline{b}_1}-g_a^1\\ L_b(\overline{b}_0)=g_b^1\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_1-\underline{b}_1}\end{cases}\quad(\underline{a}_1,\tilde{\theta}_b):\begin{cases}L_a(\underline{a}_1)=g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}\\ L_b(\tilde{\theta}_b)=g_b^1-g_a^1-g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}\end{cases}$$

Re-organize it gives: $L_a(\tilde{\theta}_a)-L_a(\overline{\theta}_a)=g_b^0\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_0-\underline{b}_0}+g_b^1\frac{\overline{b}_0-\underline{b}_1}{\overline{b}_1-\underline{b}_1}>0$, $L_a(\overline{\theta}_a)-L_a(\underline{a}_1)=g_b^1\frac{\overline{b}_1-\overline{b}_0}{\overline{b}_1-\underline{b}_1}-g_a^1-g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}>0$, $L_a(\underline{a}_1)-L_a(\overline{a}_0)=(\overline{a}_0-\underline{a}_1)(\frac{g_a^0}{\overline{a}_0-\underline{a}_0}-\frac{g_a^1}{\overline{a}_1-\underline{a}_1})$, $L_b(\overline{b}_0)-L_b(\underline{b}_1)=(\overline{b}_0-\underline{b}_1)(\frac{g_b^1}{\overline{b}_1-\underline{b}_1}-\frac{g_b^0}{\overline{b}_0-\underline{b}_0})$, $L_b(\tilde{\theta}_b)-L_b(\overline{b}_0)=g_b^1\frac{\overline{b}_1-\overline{b}_0}{\overline{b}_1-\underline{b}_1}-g_a^1-g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}>0$, $L_b(\overline{\theta}_b)-L_b(\tilde{\theta}_b)=g_a^0\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_0-\underline{a}_0}+g_a^1\frac{\overline{a}_0-\underline{a}_1}{\overline{a}_1-\underline{a}_1}>0$

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1}>\frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}<\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\tilde{\theta}_a)>L_a(\overline{\theta}_a)>L_a(\underline{a}_1)>L_a(\overline{a}_0)$ and $L_b(\overline{\theta}_b)>L_b(\tilde{\theta}_b)>L_b(\overline{b}_0)>L_b(\underline{b}_1)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1}>\frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}>\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\tilde{\theta}_a)>L_a(\overline{\theta}_a)>L_a(\underline{a}_1)$ and $L_b(\tilde{\theta}_b)>L_b(\overline{b}_0)>L_b(\underline{b}_1)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1}<\frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}<\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{\theta}_a)>L_a(\underline{a}_1)>L_a(\overline{a}_0)$ and $L_b(\overline{\theta}_b)>L_b(\tilde{\theta}_b)>L_b(\overline{b}_0)$.

If $\frac{g_b^1}{\overline{b}_1-\underline{b}_1}<\frac{g_b^0}{\overline{b}_0-\underline{b}_0}$ & $\frac{g_a^1}{\overline{a}_1-\underline{a}_1}>\frac{g_a^0}{\overline{a}_0-\underline{a}_0}$, then $L_a(\overline{\theta}_a)>L_a(\underline{a}_1)$ and $L_b(\tilde{\theta}_b)>L_b(\overline{b}_0)$.

Since $\nu(\cdot)$ is a strictly decreasing function, Lemma 9 is proved. □

Lemma 10. *Suppose the greedy fair decision $(\theta_a(t),\theta_b(t))$ at round t follows (3), in which $\{\nu(L_a(\theta_a^m))\}_{m=1}^M$ is monotonic increasing in m and $\{\nu(L_b(\theta_b^m))\}_{m=1}^M$ is monotonic decreasing in m. Let $I_i=\{m|(\theta_a^m,\theta_b^m)$ is the i-th visited decision by decision maker$\}$ and $N_k(1)$ be the population of G_k at the first round. If $N_k(2)>N_k(1), k\in\{a,b\}$, then the following holds for each visited decision $(\theta_a^{I_i},\theta_b^{I_i})$, $i=1,2,\cdots$*

(1) If $r_{I_i}>\frac{N_a(1)}{N_b(1)}$, then $\exists$ only two possibilities for greedy decision:

- *If $\frac{\beta_a(1-\nu(L_b(\theta_b^{I_i})))}{\beta_b(1-\nu(L_a(\theta_a^{I_i})))}<r_{I_i}$, then greedy decision will be $(\theta_a^{I_i},\theta_b^{I_i})$, $\forall t$*

- *If* $\frac{\beta_a(1-\nu(L_b(\theta_b^{I_i})))}{\beta_b(1-\nu(L_a(\theta_a^{I_i})))} > r_{I_i}$, *then greedy decision will change from* $(\theta_a^{I_i}, \theta_b^{I_i})$ *to* $(\theta_a^{I_i+1}, \theta_b^{I_i+1})$ *at some* t_0, *i.e.,* $I_{i+1} = I_i + 1$. *Moreover,* $N_a(t_0) < \frac{\beta_a}{1-\nu(L_a(\theta_a^{I_{i+1}}))}$ *holds.*

(2) If $r_{I_i-1} < \frac{N_a(1)}{N_b(1)}$, *then* $\exists$ *only two possibilities for greedy decision:*

- *If* $\frac{\beta_a(1-\nu(L_b(\theta_b^{I_i})))}{\beta_b(1-\nu(L_a(\theta_a^{I_i})))} > r_{I_i-1}$, *then greedy decision will be* $(\theta_a^{I_i}, \theta_b^{I_i})$, $\forall t$
- *If* $\frac{\beta_a(1-\nu(L_b(\theta_b^{I_i})))}{\beta_b(1-\nu(L_a(\theta_a^{I_i})))} < r_{I_i-1}$, *then greedy decision will change from* $(\theta_a^{I_i}, \theta_b^{I_i})$ *to* $(\theta_a^{I_i-1}, \theta_b^{I_i-1})$ *at some* t_0, *i.e.,* $I_{i+1} = I_i - 1$. *Moreover,* $N_b(t_0) < \frac{\beta_b}{1-\nu(L_b(\theta_b^{I_{i+1}}))}$ *holds.*

Proof. We denote $r_M = +\infty$ and $r_0 = 0$ in the following analysis. Let $N_k(1)$ be the population of G_k at the first round, then according to dynamic model, $N_k(t+1) = \beta_k(1+\sum_{j=1}^{t-1}\prod_{i=j}^{t-1}\nu(L_k(\theta_k(i)))) + N_k(1)\prod_{i=0}^{t}\nu(L_k(\theta_k(i)))$, $\forall t \geq 1$. To simplify the notations, denote $\nu_k^m = \nu(L_k(\theta_k^m))$ for $k \in \{a, b\}$ and $m = 1, \cdots, M$.

Prove Lemma 10 by induction.

Base case: For the first visited decision, assume $I_1 = j$, i.e., $(\theta_a(1), \theta_b(1)) = (\theta_a^j, \theta_b^j)$, then $\frac{N_a(1)}{N_b(1)} \in (r_{j-1}, r_j)$. For $t \geq 1$, $N_k(t) = (\nu_k^j)^{t-1}N_k(1) + \beta_k\sum_{s=0}^{t-2}(\nu_k^j)^s$ and $\frac{N_a(t)}{N_b(t)} = \frac{(\nu_a^j)^{t-1}N_a(1)+\beta_a\sum_{s=0}^{t-2}(\nu_a^j)^s}{(\nu_b^j)^{t-1}N_b(1)+\beta_b\sum_{s=0}^{t-2}(\nu_b^j)^s}$.

(1) If $\nu_a^j > \nu_b^j$. Since $N_a(2) > N_a(1) \to N_a(1) < \frac{\beta_a}{1-\nu_a^j}$ and $\frac{\beta_b N_a(1)}{\beta_a N_b(1)} < r_j\frac{\beta_b}{\beta_a}$, according to Lemma 5: (i) If $\frac{\beta_a(1-\nu_b^j)}{\beta_b(1-\nu_a^j)} < r_j$, then $\frac{N_a(t)}{N_b(t)} \in (r_{j-1}, r_j)$ and $(\theta_a(t), \theta_b(t)) = (\theta_a^j, \theta_b^j)$, $\forall t$. greedy decision will stay at (θ_a^j, θ_b^j). (ii) If $\frac{\beta_a(1-\nu_b^j)}{\beta_b(1-\nu_a^j)} > r_j$, then $\exists t_0$ such that $\frac{N_a(t_0)}{N_b(t_0)} > r_j$ and greedy decision starts to change to $(\theta_a^{j+1}, \theta_b^{j+1})$, i.e., $I_2 = I_1 + 1$. Moreover,$N_a(t_0) = (\nu_a^j)^{t_0-1}N_a(1) + \beta_a\sum_{s=0}^{t_0-2}(\nu_a^j)^s < \frac{\beta_a}{1-\nu_a^j} < \frac{\beta_a}{1-\nu_a^{j+1}} = \frac{\beta_a}{1-\nu_a^{I_2}}$, where the second inequality is because $\nu_a^{j+1} > \nu_a^j$.

(2) If $\nu_a^j < \nu_b^j$. Since $N_b(2) > N_b(1) \to N_b(1) < \frac{\beta_b}{1-\nu_b^j}$ and $\frac{\beta_b N_a(1)}{\beta_a N_b(1)} > r_{j-1}\frac{\beta_b}{\beta_a}$, according to Lemma 5: (i) If $\frac{\beta_a(1-\nu_b^j)}{\beta_b(1-\nu_a^j)} > r_{j-1}$, then $\frac{N_a(t)}{N_b(t)} \in (r_{j-1}, r_j)$ and $(\theta_a(t), \theta_b(t)) = (\theta_a^j, \theta_b^j)$, $\forall t$. greedy decision will stay at (θ_a^j, θ_b^j). (ii) If $\frac{\beta_a(1-\nu_b^j)}{\beta_b(1-\nu_a^j)} < r_{j-1}$, then $\exists t_0$ such that $\frac{N_a(t_0)}{N_b(t_0)} < r_{j-1}$ and greedy decision starts to change to $(\theta_a^{j-1}, \theta_b^{j-1})$, i.e., $I_2 = I_1 - 1$. Moreover,$N_b(t_0) = (\nu_b^j)^{t_0-1}N_b(1) + \beta_b\sum_{s=0}^{t_0-2}(\nu_b^j)^s < \frac{\beta_b}{1-\nu_b^j} < \frac{\beta_b}{1-\nu_b^{j-1}} = \frac{\beta_b}{1-\nu_b^{I_2}}$, where the second inequality is because $\nu_b^{j-1} > \nu_b^j$.

Induction Step: Suppose Lemma 10 is true for i-th visited decision ($i > 1$), let $I_i = m$. Then for the $(i+1)$-th visited decision that is visited at t_0, there could be two possibilities:

(1) $I_{i+1} = I_i + 1 = m + 1$, where $\frac{\beta_a(1-\nu_b^m)}{\beta_b(1-\nu_a^m)} > r_m$ and $N_a(t_0) < \frac{\beta_a}{1-\nu_a^{m+1}}$ and $r_m < \frac{N_a(t_0)}{N_b(t_0)} < r_{m+1}$ hold.

For $t > t_0$, $N_k(t) = \beta_k\sum_{s=0}^{t-t_0-1}(\nu_k^{m+1})^s + (\nu_k^{m+1})^{t-t_0}N_k(t_0)$ Since $\nu_a^{m+1} > \nu_a^m > \nu_b^m > \nu_b^{m+1}$, $\frac{N_a(t_0)}{N_b(t_0)} < r_{m+1}$ and $N_a(t_0) < \frac{\beta_a}{1-\nu_a^{m+1}}$ hold. By Lemma 5, (i) If $\frac{\beta_a(1-\nu_b^{m+1})}{\beta_b(1-\nu_a^{m+1})} < r_{m+1}$, then $\frac{N_a(t)}{N_a(t)} < r_{m+1}$ holds $\forall t \geq t_0$. greedy decision will stay at $(\theta_a^{m+1}, \theta_b^{m+1})$. (ii) if $\frac{\beta_a(1-\nu_b^{m+1})}{\beta_b(1-\nu_a^{m+1})} > r_{m+1}$, then $\exists t_1$ such that $\frac{N_a(t_1)}{N_b(t_1)} > r_{m+1}$ and greedy decision starts to change to $(\theta_a^{m+2}, \theta_b^{m+2})$, i.e., $I_{i+2} = I_{i+1} + 1$. Moreover, $N_a(t_1) = \sum_{s=0}^{t_1-t_0-1}(\nu_a^{m+1})^s + (\nu_a^{m+1})^{t_1-t_0}N_a(t_0) < \frac{\beta_a}{1-\nu_a^{m+1}} < \frac{\beta_a}{1-\nu_a^{m+2}} = \frac{\beta_a}{1-\nu_a^{I_{i+2}}}$.

(2) $I_{i+1} = I_i - 1 = m - 1$, where $\frac{\beta_a(1-\nu_b^m)}{\beta_b(1-\nu_a^m)} < r_{m-1}$ and $N_b(t_0) < \frac{\beta_b}{1-\nu_a^{m-1}}$ and $r_{m-1} < \frac{N_a(t_0)}{N_b(t_0)} < r_m$ hold.

For $t > t_0$, $N_k(t) = \beta_k\sum_{s=0}^{t-t_0-1}(\nu_k^{m-1})^s + (\nu_k^{m-1})^{t-t_0}N_k(t_0)$ Since $\nu_a^{m-1} < \nu_a^m < \nu_b^m < \nu_b^{m-1}$, $\frac{N_a(t_0)}{N_b(t_0)} > r_{m-1}$ and $N_b(t_0) < \frac{\beta_b}{1-\nu_b^{m-1}}$ hold. By Lemma 5, (i) If $\frac{\beta_a(1-\nu_b^{m+1})}{\beta_b(1-\nu_a^{m+1})} > r_{m-1}$, then $\frac{N_a(t)}{N_a(t)} > r_{m-1}$ holds $\forall t \geq t_0$. greedy decision will stay at $(\theta_a^{m-1}, \theta_b^{m-1})$. (ii) if $\frac{\beta_a(1-\nu_b^{m-1})}{\beta_b(1-\nu_a^{m-1})} < r_{m-1}$, then $\exists t_1$ such that $\frac{N_a(t_1)}{N_b(t_1)} < r_{m-1}$ and greedy decision starts to change to $(\theta_a^{m-2}, \theta_b^{m-2})$, i.e., $I_{i+2} = I_{i+1} - 1$. Moreover, $N_b(t_1) = \sum_{s=0}^{t_1-t_0-1}(\nu_a^{m-1})^s + (\nu_a^{m-1})^{t_1-t_0} N_b(t_0) < \frac{\beta_b}{1-\nu_b^{m-1}} < \frac{\beta_b}{1-\nu_b^{m-2}} = \frac{\beta_b}{1-\nu_b^{I_{i+2}}}$.

Therefore, Lemma 10 also holds for $(i+1)$-th visited decision, and the proof of the induction step is complete. □

Then Theorem 5 can be concluded from Lemma 10 and the fact that there are only finite number of actions $\{(\theta_a^m, \theta_b^m)\}_{m=1}^M$ directly.

J Algorithm to find visited decisions for uniform case

Lemma 10 also provides a simple method for finding the visited decision list $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$, the complete procedure is given in Algorithm 1. Input $\{(\theta_a^m, \theta_b^m), r_m\}_{m=1}^M$ fully describe one-shot decisions at each time step, and $1-\nu(L_k(\theta_k^m))$, $k \in \{a, b\}$ is the fraction of users from group G_k that leave the system if taking decision θ_k^m.

Algorithm 1 Finding the visited decisions list

Input: $\{(\theta_a^m, \theta_b^m), r_m, \frac{\beta_a(1-\nu(L_b(\theta_b^m)))}{\beta_b(1-\nu(L_a(\theta_a^m)))}\}_{m=1}^M$
Output: $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$
$k = \{m | \frac{N_a(1)}{N_b(1)} \in (r_{m-1}, r_m)\}$;
$i = 1$;
while $true$ **do**
  $(\theta_a^{I_i}, \theta_b^{I_i}) = (\theta_a^k, \theta_b^k)$;
  if $\frac{\beta_a(1-\nu(L_b(\theta_b^k)))}{\beta_b(1-\nu(L_a(\theta_a^k)))} \in (r_{k-1}, r_k)$ **then**
    break;
  else if $\frac{\beta_a(1-\nu(L_b(\theta_b^k)))}{\beta_b(1-\nu(L_a(\theta_a^k)))} < r_{k-1}$ **then**
    $k \leftarrow k-1$;
  else if $\frac{\beta_a(1-\nu(L_b(\theta_b^k)))}{\beta_b(1-\nu(L_a(\theta_a^k)))} > r_k$ **then**
    $k \leftarrow k+1$;
  end if
  $i = i + 1$;
end while
return $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$

K Experiments Supplementary

K.1 Parameter settings

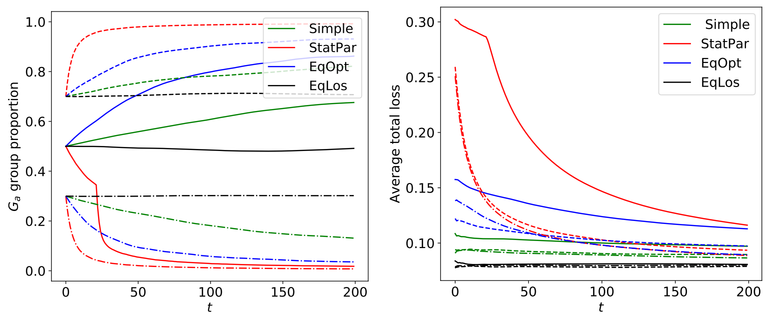
For the truncated normal case, the supports of $f_k^j(x), k \in \{a, b\}, j \in \{0, 1\}$ are $[\underline{a}_0, \underline{a}_1, \overline{a}_0, \overline{a}_1] = [-8, 5, 19, 35]$, $[\underline{b}_0, \underline{b}_1, \overline{b}_0, \overline{b}_1] = [-6, 25, 9, 43]$, with the means $[\mu_a^0, \mu_a^1, \mu_b^0, \mu_b^1] = [4, 20, 8, 27]$ and standard deviations $[\sigma_a^0, \sigma_a^1, \sigma_b^0, \sigma_b^1] = [5, 6, 3, 6]$. The label proportions are $g_a^0 = 0.4$, $g_b^0 = 0.6$. The dynamics (1) uses $\nu(x) = 1 - x$

For the uniform case, the supports of $f_k^j(x), k \in \{a, b\}, j \in \{0, 1\}$ are $[\underline{a}_0, \underline{a}_1, \overline{a}_0, \overline{a}_1] = [-5, 10, 20, 35]$, $[\underline{b}_0, \underline{b}_1, \overline{b}_0, \overline{b}_1] = [3, 17, 25, 45]$. The label proportions are $g_a^0 = 0.8$, $g_b^0 = 0.2$. The dynamics (1) uses $\nu(x) = 1 - x^2$

K.2 Results on uniformly distributed case

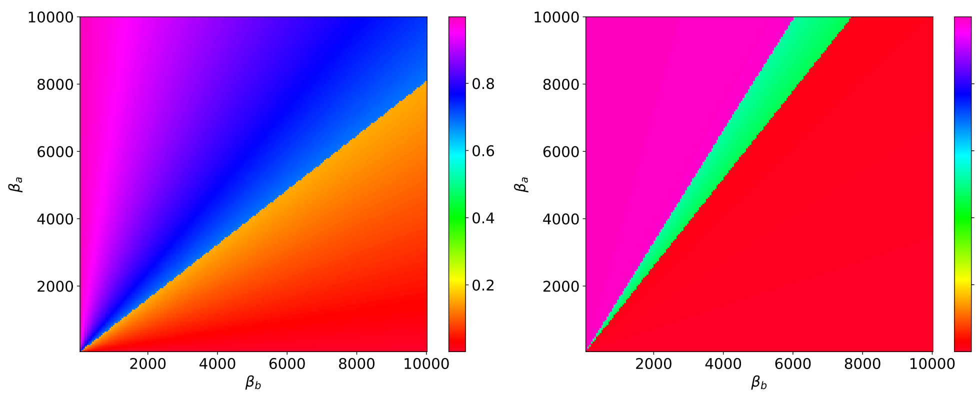
	Simple fair	StatPar fair	EqOpt fair
$\beta_a = 3000; \beta_b = 7000$	$[(17, 17)]$	$[(-1.02, 17)]$	$[(10.91, 17)]$
$\beta_a = 7000; \beta_b = 3000$	$[(20, 20)]$	$[(20, 40.80)]$	$[(20, 25)]$
$\beta_a = 5000; \beta_b = 5000$	$[(20, 20)]$	$[(10, 26.8), (8.39, 25), (-1.02, 17)]$	$[(20, 25)]$

Table 1: Visited decisions $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$

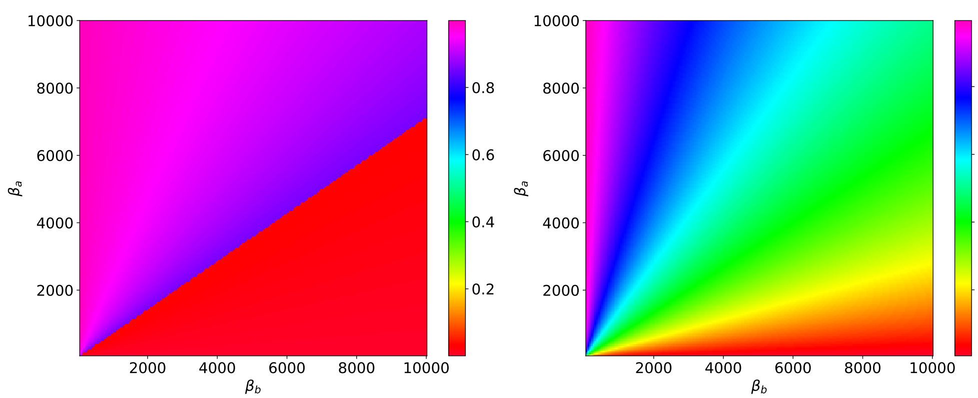


(a) Group proportion (b) Average total loss

Figure 9: Sample paths for a scenario with uniform feature distributions and $\beta_a + \beta_b = 10000$. Table. 1 illustrates all the visited decisions $\{(\theta_a^{I_i}, \theta_b^{I_i})\}_{i=1}^K$ under different fairness criteria, the corresponding $\overline{\alpha}_a(t)$ and the average total loss at each time step t are shown in Figure 9(a)9(b) respectively: the solid lines are for the case when $\beta_a = \beta_b$, dashed lines are for $3\beta_a = 7\beta_b$, dot dashed lines are for $7\beta_a = 3\beta_b$. Notice that the `EqLos` greedy fair decision can achieve the lowest averaged total loss in this example.



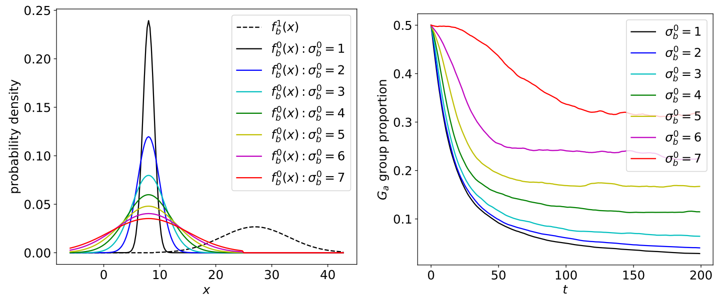
(a) `Simple` fair (b) `StatPar` fair



(c) `EqOpt` fair (d) `EqLos` fair

Figure 10: Each dot in Figure 10(a)-10(d) represents the final group proportion $\lim_{t\to\infty} \overline{\alpha}_a(t)$ of one sample path under a pair of arriving rates (β_a, β_b). By comparing this figure with Figure. 5, we notice that group representation is more unstable under uniform distribution (small variation in arrival rate can change the group representation significantly) as compared to an example with truncated normal feature distribution.

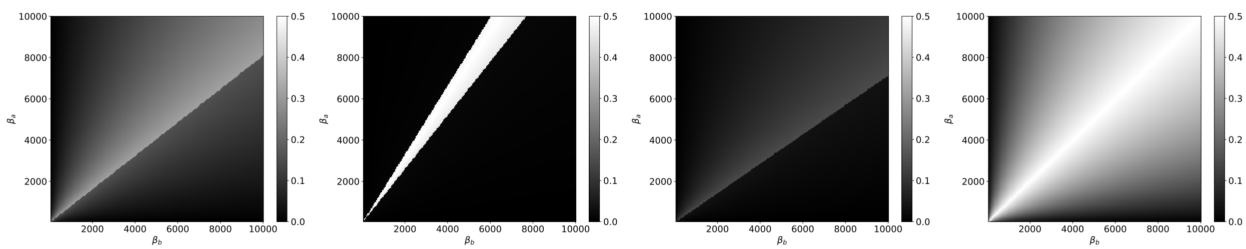
K.3 Feature distributions affect the preference



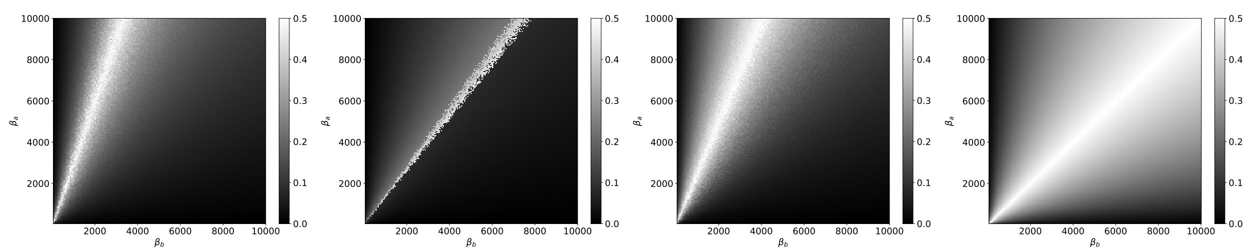
(a) Feature distributions illustration (b) Group proportion $\beta_a = \beta_b$

Figure 11: Change $f_b^0(x)$ by varying $\sigma_b^0 \in \{1, 2, 3, 4, 5, 6, 7\}$. As σ_b^0 increases, the overlap area with $f_b^1(x)$ also increases as shown in Figure. 11(a). Figure. 11(b) shows the result under `StatPar` fairness. Given $\theta_a(t)$, the larger σ_b^0 results in the larger $L_b(\theta_b(t))$ and further the larger G_a's proportion.

K.4 Representation disparity under β_a, β_b



(a) `Simple` fair (b) `StatPar` fair (c) `EqOpt` fair (d) `EqLos` fair



(e) `Simple` fair (f) `StatPar` fair (g) `EqOpt` fair (h) `EqLos` fair

Figure 12: Each dot in Figure 12(e)-12(h) represents the final group representation disparity $\min\{\lim_{t\to\infty} \overline{\alpha}_a(t), \lim_{t\to\infty} \overline{\alpha}_b(t)\}$ of one sample path under a pair of arriving rates (β_a, β_b). The darker color illustrates higher representation disparity.