Reviewer 1:

1. What ensures the absolute sizes of both groups are not very small. The absolute size of a group is a function of both arrival and retention rates. If one assumes non-zero arrival at each time as the paper does, then size will not diminish regardless of retention. In particular, if group representation is maintained over time, then $\theta_k(t) = \theta_k, \forall t$ and the group size converges to $\frac{\beta_k}{1-\pi_k(\theta_k)}$. The only way for this to be small is if both arrival β_k and retention $\pi_k(\theta_k)$ are near-0. If we allow arrival to be a function of, say model accuracy (Sec 3.4), then arrival indeed may diminish; in this case ensuring representation (as shown in Sec 3.4) can simultaneously help prevent zero arrival.

2. Is the case of `EqLos` special regarding the experimental results? No, it is not. It works because the user retention is assumed to be driven by model accuracy in our experiments. As illustrated in Fig. 8(a) in Appendix K.3, if user retention is driven by TPR/FNR (e.g., loan application), `EqOpt` would be the proper fairness notion.

3. Pros & cons, feasibility & applicability of our framework. Since human decision making is inherently a sequential (and non-memoryless) process, we feel our framework of examining fairness in such a sequential framework is appropriate for real-world settings. The main limitation of such an approach is that it requires sufficiently accurate models capturing the underlying dynamics (what drives the adoption/abandonment of ML algorithms, etc), which is not always available. We believe there is value in performing long-term experiments to better understand such dynamics.

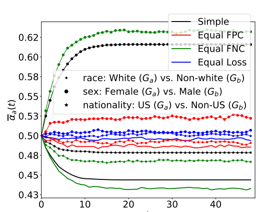
4. Can this framework illustrate when positive scenarios can be achieved. In a sense the current model captures positive feedback, for the majority group: better model performance leads to population growth. Case 2 in Sec 3.3 may be viewed as another positive instance: a group can work to change their distribution in light of perceived bias in the algorithm (and if they manage to break the condition stated therein then they may retain representation).

5. Improving readability: We will adjust figures, add forward references, fix typos, and discuss intuition/comparisons.

Reviewer 2:

1. Applicability to more general settings. Our results indeed apply more generally to non-classification problems and/or multi-dimensional features. Thm 1 states that the representation disparity worsens as long as the monotonicity condition (MC) holds; no requirement is imposed on dimensionality or objective function or dynamics. The 1D classification problem is one such case satisfying MC (Thm 3). However, it can be shown rigorously that under certain conditions for $\pi_k(\theta_k) = \nu_k(O_k(\theta_k))$ for some decreasing $\nu_k(\cdot)$, Thm 3 holds when feature vector $X \in \mathbb{R}^d$ and the underlying problem can be other supervised (e.g., regression) and unsupervised learning. We will be happy to add this result.

2. Experiments with non-synthetic data. We trained binary classifiers over *Adult* dataset by minimizing empirical loss where features are individual info (sex, race, nationality, etc.) and labels their annual income ($\geq$ \$50k or $<$ \$50k). Since the dataset does not reflect dynamics, we assume it follows (2) with $\pi_k(\theta_k) = \nu(L_k(\theta_k))$. We examine the monotonic convergence of representation disparity under `Simple`, `EqOpt` (equalized false positive/negative cost(FPC/FNC)) and `EqLos`, and consider cases where G_a, G_b are distinguished by sex, race and nationality. These results (shown on the right) are consistent with the paper.

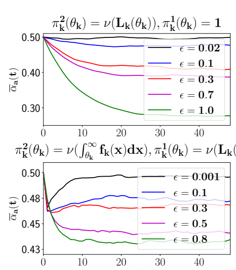


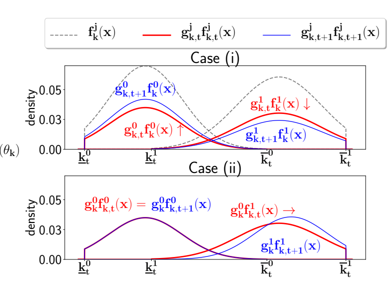
3. Clarifications (i) Goal of Thm 2 is not to find population ratios but to find one-shot solutions given population ratios; their relation is in Eqn. (3). (ii) $Y \in \{0, 1\}$ is label with distribution $Pr(Y = j|K = k) = g^j_{k,t}$ and y is its realization.

4. Distinction from (Hashimoto et al., 2018) [6]. Worsening of representation disparity is observed via simulation in [6] without using fairness ($\theta_a = \theta_b$), and a min-max fair is used to address this. We show the introduction of (any type of) fairness does not necessarily solve this problem and do so using formal analysis. Other differences include the fact we consider the case when feature distributions are reshaped by the decisions (Sec 3.3) and [6] does not.

Reviewer 3:

1. Experiment with proposed fairness constraint selection. $\Delta = \epsilon\frac{\beta_a}{\beta_b}$-fair set found with method in Sec 3.4 (left plot): each curve represents a sample path under different ϵ where $(\theta_a(t), \theta_b(t))$ is from a small randomly selected subset of Δ-fair set $\forall t$ (to model the situation where perfect fairness is not feasible) and $\frac{\beta_a}{\beta_b} = 1$. We observe that fairness is always violated at beginning in lower plot. This is because the fairness set is found based on stable fixed points, which only concerns fairness in the long run.





2. Visualization of decisions shaping feature distribution in Sec 3.3. The right plot above illustrates how distributions would change from t to $t+1$, when G^1_k (resp. G^0_k) experiences the higher (resp. lower) loss at t than $t-1$.