Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness

Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin and Mingyan Liu
xueru@umich.edu; khalili@umich.edu; cemtekin@ee.bilkent.edu.tr; mingyan@umich.edu

Paper

Objectives

Motivation: ML model may be less favorable to groups contributing less to training process –> degrade population retention –> exacerbate representation disparity

Goal:
* What happens to the group representation over time when fair ML models are used.
* How it is affected when underlying feature distributions are also reshaped by decisions.

Problem Formulation

Two demographic groups G_a, G_b with time-varying X and $Y \in \{0,1\}$

- feature distribution at t: $f_{k,t}(x) = g^0_{k,t} f^0_{k,t}(x) + g^1_{k,t} f^1_{k,t}(x)$, $x \in \mathbb{R}^d$
- representation disparity: $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ with $\overline{\alpha}_a(t) + \overline{\alpha}_b(t) = 1$.

One-shot problem:

$$\min_{\theta_a,\theta_b} \quad \boldsymbol{O}_t(\theta_a,\theta_b;\overline{\alpha}_a(t),\overline{\alpha}_b(t)) = \overline{\alpha}_a(t)O_{a,t}(\theta_a) + \overline{\alpha}_b(t)O_{b,t}(\theta_b)$$
$$\text{s.t.} \quad \Gamma_{\mathcal{C},t}(\theta_a,\theta_b) = 0$$

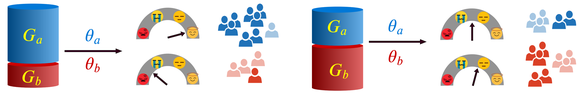
Participation dynamics:

$$N_k(t+1) = N_k(t)\pi_{k,t}(\theta_k(t)) + \beta_k$$
$$\overline{\alpha}_k(t+1) = \frac{N_k(t+1)}{N_a(t+1)+N_b(t+1)}$$

Goal: what happens to $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ when one-shot fair decisions are applied $\forall t$.

Monotonicity Condition (MC)

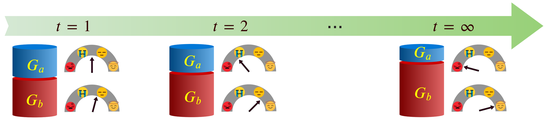
Two one-shot problems $\widehat{\boldsymbol{O}}(\theta_a,\theta_b;\widehat{\alpha}_a,\widehat{\alpha}_b)$ and $\widetilde{\boldsymbol{O}}(\theta_a,\theta_b;\widetilde{\alpha}_a,\widetilde{\alpha}_b)$ defined over $\widehat{f}_k(x)$ and $\widetilde{f}_k(x)$ satisfy MC under a dynamic model if $\forall\ \widetilde{\alpha}_a + \widetilde{\alpha}_b = 1$ and $\widehat{\alpha}_a + \widehat{\alpha}_b = 1$ such that $\frac{\widehat{\alpha}_a}{\widehat{\alpha}_b} < \frac{\widetilde{\alpha}_a}{\widetilde{\alpha}_b}$, the resulting retention rates satisfy $\widehat{\pi}_a(\widehat{\theta}_a) < \widetilde{\pi}_a(\widetilde{\theta}_a)$ and $\widehat{\pi}_b(\widehat{\theta}_b) > \widetilde{\pi}_b(\widetilde{\theta}_b)$.



Exacerbation of Representation Disparity

- **Monotonic change of $\frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ under MC:**

Consider $\left\{\boldsymbol{O}_t(\theta_a,\theta_b;\overline{\alpha}_a(t),\overline{\alpha}_b(t))\right\}_{t=1}^{\infty}$, if $\boldsymbol{O}_t$ and $\boldsymbol{O}_{t+1}$ satisfy MC $\forall t$ under a dynamic model, then we have: if $\pi_{a,1}(\theta_a(1)) < \pi_{b,1}(\theta_b(1))$(different retention), then $\frac{\overline{\alpha}_a(t+1)}{\overline{\alpha}_b(t+1)} < \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$ (exacerbation of disparity) and $\pi_{a,t+1}(\theta_a(t+1)) < \pi_{a,t}(\theta_a(t)) < \pi_{b,t}(\theta_b(t)) < \pi_{b,t+1}(\theta_b(t+1))$ (discrepancy increases).



Cases satisfying MC under the same $f_k(x)$

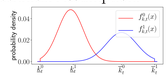
Two one-shot problems $\widehat{\boldsymbol{O}}(\theta_a,\theta_b;\widehat{\alpha}_a,\widehat{\alpha}_b) = \widehat{\alpha}_a O_a(\theta_a) + \widehat{\alpha}_b O_b(\theta_b)$ and $\widetilde{\boldsymbol{O}}(\theta_a,\theta_b;\widetilde{\alpha}_a,\widetilde{\alpha}_b) = \widetilde{\alpha}_a O_a(\theta_a) + \widetilde{\alpha}_b O_b(\theta_b)$ satisfy MC when:
* $O_k(\widehat{\theta}_k) \neq O_k(\widetilde{\theta}_k)$ for any possible $\widehat{\alpha}_k \neq \widetilde{\alpha}_k$
* they are defined over the same $f_k(x)$
* dynamics satisfy $\pi_k(\theta_k) = h_k(O_k(\theta_k))$ for some decreasing function $h_k(\cdot)$

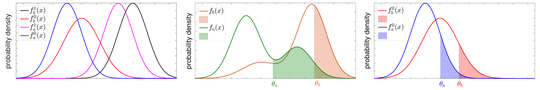
$\Longrightarrow$ hold under commonly used objectives, dynamics and fairness criteria!

A One-Dimensional Threshold Decision

- **Decision rule:**
$$h_\theta(x) = \mathbf{1}(x \geq \theta),\ X \in \mathbb{R}$$
- **Objective function:**
$O_{k,t}(\theta_k) = L_{k,t}(\theta_k)$ with 0-1 loss $\mathbf{1}(y \neq h_\theta(x))$
- **Fairness criterion** $\Gamma_{\mathcal{C},t}(\theta_a,\theta_b) \longleftrightarrow \theta_a = \phi_{\mathcal{C},t}(\theta_b)$**:**
  - Simple fair (`Simple`): $\Gamma_{\mathcal{C},t} = \theta_a - \theta_b$
  - Statistical Parity (`StatPar`): $\Gamma_{\mathcal{C},t} = \int_{\theta_a}^{\infty} f_{a,t}(x)dx - \int_{\theta_b}^{\infty} f_{b,t}(x)dx$
  - Equal Opportunity (`EqOpt`): $\Gamma_{\mathcal{C},t} = \int_{\theta_a}^{\infty} f^0_{a,t}(x)dx - \int_{\theta_b}^{\infty} f^0_{b,t}(x)dx$

Technical Assumption:

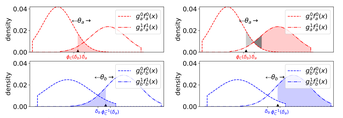




$\Longrightarrow$ **One-shot solutions under either `Simple`, `EqOpt` or `StatPar`:**
  - **Bounded by a fixed interval** $\forall \overline{\alpha}_a(t), \overline{\alpha}_b(t)$**:**
$$\theta_a(t) \in [\phi_{\mathcal{C},t}(\delta_{b,t}), \delta_{a,t}], \theta_b(t) \in [\delta_{b,t}, \phi^{-1}_{\mathcal{C},t}(\delta_{a,t})];\quad \delta_{k,t} = \text{argmin}_\theta\ L_{k,t}(\theta)$$
  - $\exists$ **a function** $\Psi_{\mathcal{C},t}(\theta_a,\theta_b)$ **increasing in** θ_a,θ_b **such that:**
$$\Psi_{\mathcal{C},t}(\theta_a(t),\theta_b(t)) = \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)}$$

Results under the fixed $f_k(x)$

- **Dynamics:**
  1. By model accuracy:
  $\pi_k(\cdot) = \nu(L_k(\cdot))$ with $\nu(\cdot)\downarrow$
  2. By intra-group disparity:
  $\pi_k(\cdot) = w(D_k(\cdot))$ with $w(\cdot)\uparrow$



$\Longrightarrow$ **Exacerbation of representation disparity:**

Consider $\left\{\boldsymbol{O}_t(\theta_a,\theta_b;\overline{\alpha}_a(t),\overline{\alpha}_b(t))\right\}_{t=1}^{\infty}$ under either `Simple`, `EqOpt` or `StatPar` criterion, when $f_{k,t}(x) = f_k(x), \forall t$:

- $\boldsymbol{O}_t$ and $\boldsymbol{O}_{t+1}$ satisfy MC $\forall t$ under above dynamics.
- $\left\{(\theta_a(t),\theta_b(t))\right\}_{t=1}^{\infty}$ converges monotonically to a constant decision.

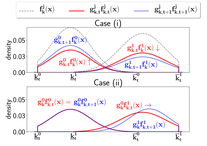
Results under time-varying $f_{k,t}(x)$

- Dynamics driven by model accuracy: $\pi_k(\cdot) = \nu(L_k(\cdot))$
- Time-varying feature distributions: $f_{k,t}(x) = g^0_{k,t} f^0_{k,t}(x) + g^1_{k,t} f^1_{k,t}(x)$

G^0_k **and** G^1_k **may react differently to the same** θ_k**:**

(i) $f^j_{k,t}(x) = f^j_k(x), \forall t$ but $g^j_{k,t}$ changes:
G^j_k's retention is determined by $L^j_{k,t}$.

(ii) $g^j_{k,t} = g^j_k, \forall t$ but $f^j_{k,t}(x)$ changes:
For G^i_k that is less favored by decisions, its members may make extra effort for lowering their losses.



$\Longrightarrow$ **Exacerbation in representation disparity can accelerate:**

(P1): $\left\{\boldsymbol{O}^f_t(\theta_a,\theta_b;\overline{\alpha}^f_a(t),\overline{\alpha}^f_b(t))\right\}_{t=1}^{\infty}$ over $f_k(x)$ with $\pi^f_a(\theta^f_a(1)) < \pi^f_b(\theta^f_b(1))$

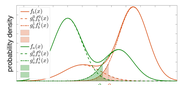
(P2): $\left\{\boldsymbol{O}^v_t(\theta_a,\theta_b;\overline{\alpha}^v_a(t),\overline{\alpha}^v_b(t))\right\}_{t=1}^{\infty}$ over $\left\{f_{k,t}(x)\right\}_{t=1}^{\infty}$ with $f_{k,1}(x) = f_k(x)$

Under either `Simple`, `EqOpt` or `StatPar`, if $f_{k,t}(x)$ changes slowly w.r.t. the change in one-shot decisions:

- $\boldsymbol{O}^v_t$ and $\boldsymbol{O}^v_{t+1}$ satisfy MC $\forall t$ under this dynamics.
- $\frac{\overline{\alpha}^v_a(t)}{\overline{\alpha}^v_b(t)} < \frac{\overline{\alpha}^f_a(t)}{\overline{\alpha}^f_b(t)}, \forall t$

Mitigation when $\pi_{k,t}(\cdot) = \nu(L_{k,t}(\cdot))$

Equalized Loss (`EqLos`) fairness:
$$\Gamma_{\mathcal{C},t} = L_{a,t}(\theta_a) - L_{b,t}(\theta_b)$$
⋆ It maintains group representation:
$$\lim_{t\to\infty} \frac{\overline{\alpha}_a(t)}{\overline{\alpha}_b(t)} = \frac{\beta_a}{\beta_b}$$



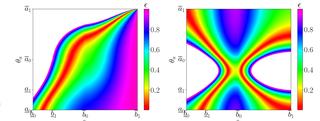
$\Longrightarrow$ Fairness has to be defined with a good understanding of underlying participation dynsmics!

Mitigation for General Dynamics

Find the proper criterion for some general dynamics when $f_{k,t}(x) = f_k(x), \forall t$ **:**
$$N_k(t+1) = \Lambda(N_k(t), \{\pi^m_k(\theta_k(t))\}_{m=1}^{M}, \beta_k)$$

- Assume $\exists(\theta_a,\theta_b)$ s.t. corresponding dynamics have stable fixed points.

$$\text{argmin}_{\theta_a,\theta_b} \left|\frac{\widetilde{N}_a}{\widetilde{N}_b} - \frac{\beta_a}{\beta_b}\right|$$
$$\text{s.t.}\quad \widetilde{N}_k = \Lambda(\widetilde{N}_k, \{\pi^m_k(\theta_k)\}_{m=1}^{M}, \beta_k)$$
$$\widetilde{N}_k \in \mathbb{R}_+, \theta_k \in \Theta, \forall k \in \{a,b\}$$

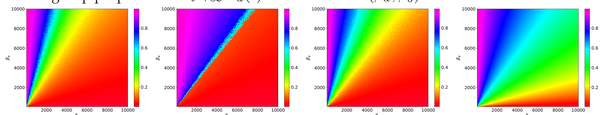


$\begin{cases}\pi^2_{\theta_a}(\theta) = \nu(\int_\theta^\infty f_k(x)dx)\\ \pi^1_k(\theta) = \nu(L_k(\theta))\end{cases}$ $\begin{cases}\pi^2_{\theta_a}(\theta) = \nu(L_k(\theta))\\ \pi^1_k(\theta) = 1\end{cases}$

- **Δ-fair set:** all (θ_a,θ_b) s.t. $\left|\frac{\widetilde{N}_a}{\widetilde{N}_b} - \frac{\beta_a}{\beta_b}\right| \leq \min\left\{\left|\frac{\widetilde{N}_a}{\widetilde{N}_b} - \frac{\beta_a}{\beta_b}\right|\right\} + \Delta$
- **Examples:** $N_k(t+1) = N_k(t)\pi^2_k(\theta_k(t)) + \beta_k\pi^1_k(\theta_k(t))$ with $\Delta = \epsilon\frac{\beta_a}{\beta_b}$

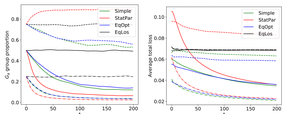
Numerical Results

- Final group proportion $\lim_{t\to\infty} \overline{\alpha}_a(t)$ under different (β_a,β_b)

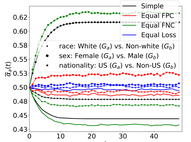


(a): `Simple` **(b):** `StatPar` **(c):** `EqOpt` **(d):** `EqLos`

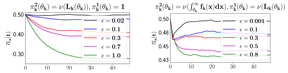
- Monotonic change of group representation



- Monotonic change of group representation when applying binary classifiers on *Adult* dataset $(\beta_a = \beta_b)$



- Effect of Δ-fair decisions $(\beta_a = \beta_b)$



Conclusions

- Group representation disparity can get exacerbated over time very easily under seemingly fair decisions.
- Exacerbation in representation disparity can accelerate when feature distributions are also reshaped by decisions.
- Develop a method of selecting a proper fairness criterion based on prior knowledge of participation dynamics.