Preference Dynamics Under Personalized Recommendations

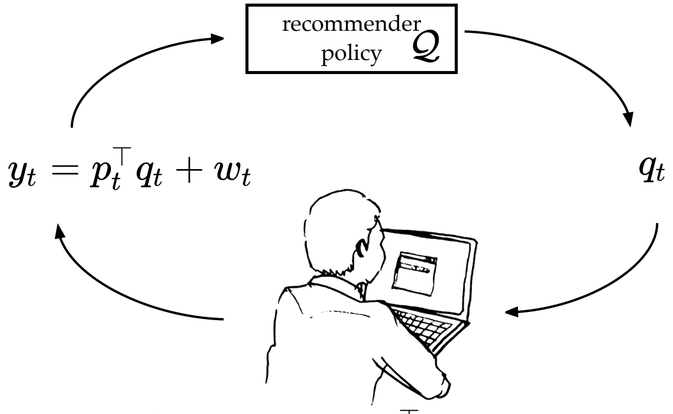
Sarah Dean *(Cornell University)* and Jamie Morgenstern *(University of Washington)*

Introduction

Individuals may be impacted by recommended content. We study a dynamical model of *biased assimilation* proposed by HJMR [2019]: preferences become more aligned with content that is enjoyed, and anti-aligned with content that is disliked.

Preference Dynamics Model

An individual's preference $p_t \in \mathcal{S}^{d-1}$ determines their response to content q_t chosen from a set $\mathcal{Q} \subset \mathcal{S}^{d-1}$ via the affinity $p_t^\top q_t$. The affinity affects both the rating y_t and the evolution of preferences along with step-size η_t.



$$\tilde{p}_{t+1} = p_t + \eta_t p_t^\top q_t \cdot q_t$$
$$p_t = \tilde{p}_t / \|\tilde{p}_t\|$$

Rating Maximization with Fixed Recommendations

The dynamics make it *easy* to achieve high ratings; as long as $\mathcal{Q}$ contains opposites very little needs to be known about users or items.

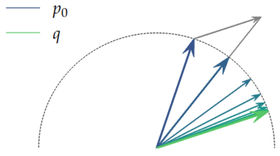
Alg: Explore-then-Commit

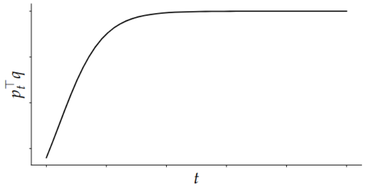
For $t = 0, \dots, \sigma^2 \log T/c^2$:
  recommend $q_t = q$ and observe y_t
For $t = \sigma^2 \log T/c^2, \dots, T$:
  if $\sum_t y_t < 0$: choose $q_t = -q$
  else: recommend $q_t = q$

Informal Result: As long as $|p_0^\top q| > c$ and noise is σ^2 sub-Gaussian,

$$R(T) = \sum_{t=0}^{T-1} 1 - p_t^\top q_t \le C_\eta(1/c^2 - 1) + \sigma^2 \log T/c^2$$

(See paper for more general setting)



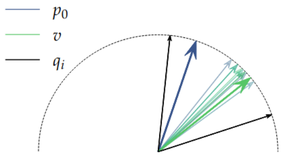


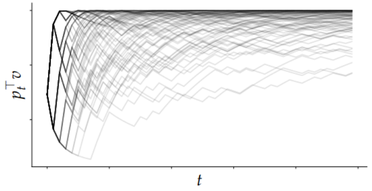
Stationary Preferences with Randomized Recommendations

Non-manipulation [KML20] is an alternative goal. Since it may be $p_0 \notin \mathcal{Q}$, a randomized strategy selects q_t i.i.d.

Informal Result: Suppose p_0 is the dominant eigenvector of $\mathbb{E}[qq^\top]$, the randomization is *aligned*, and η_t decays like $\frac{1}{1+t}$. Then

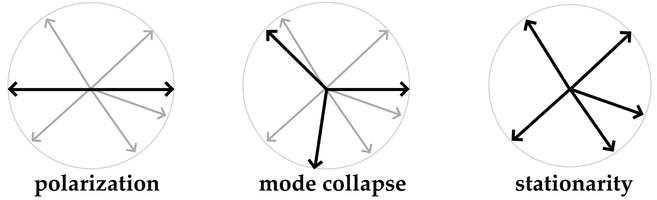
$$R(T) = \sum_{t=0}^{T-1} 1 - p_t^\top p_0 \lesssim \log T$$





Implications: Mode Collapse

While non-personalized consumption leads to polarization [HJMR19, GKT21], personalized recommendations may lead initial preference (grey) to collapse to a subset of $\mathcal{Q}$. However, randomization can keep preferences stationary.



Richness of $\mathcal{Q}$

The item set $\mathcal{Q}$ must be sufficiently rich for

1. **Estimating** initial p_0 from observations $y_{0:t}$ requires that $\text{span}(q_{0:t}) = \mathbb{R}^d$
2. **Designing** randomization for stationarity requires that $p_0 \in \text{span}(\{q \mid q \in \mathcal{Q},\ p_0^\top q > 0\})$

References

- Gaitonde, Kleinberg, Tardos, 2021. Polarization in geometric opinion dynamics. *EC*.
- Hązła, Jin, Mossel, Ramnarayan, 2019. A geometric model of opinion polarization. *arXiv:1910.05274*.
- Krueger, Maharaj, Leike, 2020. Hidden incentives for auto-induced distributional shift. *arXiv:2009.09153*.