Optimizing Personalized Assortment Decisions in the Presence of Platform Disengagement

Mika Sumida* and Angela Zhou*
USC Marshall Data Sciences and Operations

Intro

- Assortment optimization: optimize an assortment (set) of items to optimize revenue under a choice model, e.g. Multinomial choice (MNL)
- But decisions regarding revenue and recommendations can impact customer engagement over time
- Interleaving *dynamic decision-making* over time (long-term customer dynamics) with *single-timestep* contextual learning
- **Our contributions**:
  - **Model**: episodic RL setting with disengagement based on purchase history
  - **Static** characterization: structural results when
  - **Dynamic** learning setting: episodic RL algorithm combining UCRL and ideas from generalized linear UCB

Setup

- Choose an assortment out of M items, $S \subseteq M$
- Each item has attractiveness v_i (later, contextual in item utility)

  Customer purchases item i w.p. $\phi_i(S) = \dfrac{v_i}{1 + \sum_{j \in S} v_j}$, revenue r_i

  Expected revenue is $R(S) = \sum_{i \in S} r_i \phi_i(S)$

Model

- **Dynamic model**:

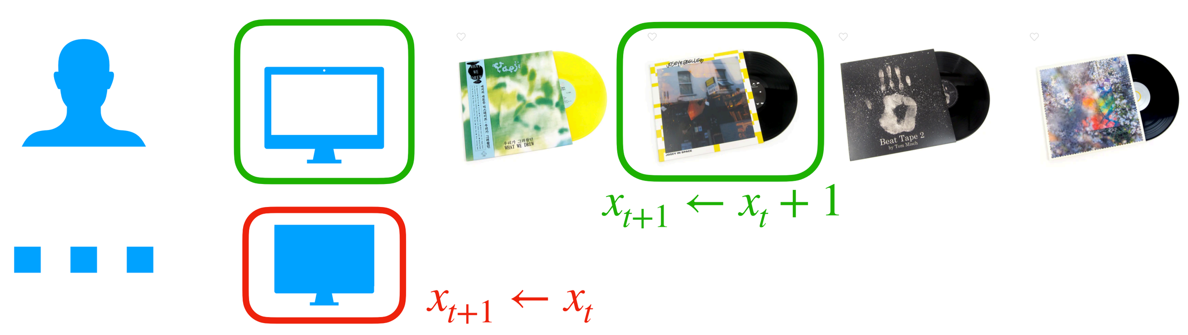
  State:

  $x \in [T] \cup 0$ cumulative number of purchases

  $Z \in \{0,1\}$ customer engagement

Assumption: Engagement dynamics.

Iff a customer engages, she is shown S and may make a purchase. If she purchases, engagement level x increases by 1.



Contextual Dynamic Decision protocol:

- For episode 0 ... K
  - New customer.
  - If contextual:

    Observe M many d-dim item contexts, $W \in \mathbb{R}^{d \times M}$
  - For timestep 0 ... T
    - Customer "engages" (logs on) w.p. $p(x_{k,t})$.

      If customer engages, you can sell them something

      Action: choose an assortment $S_{k,t} \in [M]$
    - Customer purchases with prob. $\phi(S; \beta)$, collect **reward** $R(S; \beta)$
    - *If* customer purchased (and engaged), **increment state by 1**, $x_{k,t+1} = x_{k,t} + 1$

Results

- Non-learning setting - known v_i
- - **Dynamic formulation**:

  Value function:

  $$J_t(x) = p(x) \max_{S \subseteq M} \left\{ \sum_{i \in M} \phi_i(S)(r_i + J_{t+1}(x+1)) + (1 - \sum_{i \in M} \phi_i(S)) J_{t+1}(x) \right\} + (1 - p(x)) J_{t+1}(x$$
- And $J_{T+1}(x) = 0 \quad \forall x \in \mathbb{Z}$.
- **Assumption:** $p(x)$ is monotone increasing
- **Lemma: Revenue-ordered assortments.** Denote $\Delta_t(x) = J_{t+1}(x+1) - J_{t+1}(x)$.

  The optimal solution to the problem $\max_{S \subseteq M} \left\{ R(S) + \dfrac{V(S)}{1 + V(S)} \Delta_t(x) \right\}$ is **revenue-ordered**. That is, if the items are indexed such that $r_1 \geq r_2 \geq \ldots \geq r_m$, then the optimal assortment $S^* = \{i \leq i^*\}$ for some index $i^* \in M$.

Dynamic Contextual Learning

- Contextual linear utility: $v_i = \exp(w_i^\top \beta^*)$
- **Assumption:**

  $P(x+1 \mid x, S) = \phi(S; \beta) p(x)$

  Transitions factorize into state- and time-invariant $\phi_{k,t}(S; \beta)$ *contextual* probabilities, and *dynamic/sequential* p(x)
- Episodic regret

  $$\text{Regret}(K) = \sum_{k=1}^{K} J^*_{k,0}(0) - J^{\pi_k}_{k,0}(0)$$

Algorithm

- Estimators: $\hat{\beta}$ solves regularized max-likelihood from engagement data

  $\hat{p}(x)$: empirical engagement probabilities at x, $\hat{p}_{k,t}(x) = N^k_t(1,x)/N^k_t(x)$
- Confidence intervals $b_{k,t}(x) = 2T\sqrt{\ln(Tk/\delta)/N^k_t(x)}$ (for $p(x)$),
- Optimistic estimates $\overline{p}_{k,t}(x) = \hat{p}_{k,t}(x) + b_{k,t}(x)$, $\overline{\beta}_{k,t} = \hat{\beta}_k + \zeta_k(\delta)$
- Assumptions: There exists some $\kappa > 0$ such that for all $S, i \in S$, and w, we have $\min_{\beta : \|\beta - \beta^*\| \leq 1} \phi(S, \beta)\phi(S, \beta) \geq \kappa$

Algorithm: UCRL & linUCB

- For episode 0 ... K
  - New customer. Observe item contexts and update covariance matrix.
  - Update estimates $\hat{p}(x)$, $\hat{\beta}$
  - Optimistic parameters $\overline{p}(x), \overline{\beta}$; $\overline{v}$
  - Optimistic planning: For timestep 0 ... T
    - $\overline{Q}^{\pi_k}_{k,t}(x, S) \leftarrow \overline{J}^{\pi_k}_{k,t+1}(x) + \overline{p}_{k,t}(x)\left\{R(S) + \phi_{k,t}(S; \overline{\beta})\overline{\Delta}_{k,t+1}(x)\right\}$
    - $\overline{S}_{k,t}(x) \in \arg\max_S \overline{Q}_{k,t}(x, S)$, $\pi_{k,t}(x) \leftarrow \overline{S}_{t,k}(x)$

** Use revenue-ordering Lemma for computationally easy planning

- **Theorem**: **Regret bound**

  $$\mathbb{E}\left[\text{Regret}(K)\right] = \tilde{O}\left(\max\left(\frac{\sigma}{\kappa} d\lambda, T^2\right)\sqrt{KT}\right)$$

Related work (abridged)

Abbasi-Yadkori, Y. and Neu, G. Online learning in mdps with side information. arXiv 2014
Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. ICML 2019
Modi, A. and Tewari, A. No-regret exploration in contextual reinforcement learning. UAI 2020
Oh, M.-H. and Iyengar, G. Multinomial logit contextual bandits: Provable optimality and practicality. AAAI 2021