Incentivized Bandit Learning with Self-Reinforcing User Preferences

Tianchen Zhou

Dept. of Electrical and Computer Engineering
The Ohio State University

Joint work with Jia Liu, Chaosheng Dong and Jingyuan Deng

June 19, 2021

Motivation: Online Recommender Systems

$299.99

Acer 27" Class Curved WQHD FreeSync Gaming Monitor

(319)

$329.99

Acer 24" Class ConceptD FHD IPS Widescreen Monitor

(6)

Motivation: Online Recommender Systems

$299.99

Acer 27" Class Curved WQHD FreeSync Gaming Monitor

(319)

$329.99 $50 cash back

Acer 24" Class ConceptD FHD IPS Widescreen Monitor

(6)

Challenges and Contributions

Challenges:

- Unknown optimal product.
- Balance between exploration and exploitation.
- Induce user preferences to one product with low incentives.

Challenges and Contributions

Challenges:

- Unknown optimal product.
- Balance between exploration and exploitation.
- Induce user preferences to one product with low incentives.

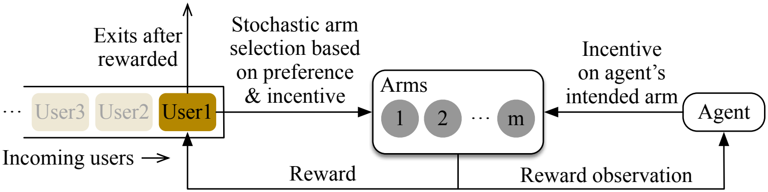
Contributions:

- A new MAB model with random arm selection that considers the relationship of self-reinforcing preferences and incentives.
- Two policies termed "At- Least-n Explore-Then-Commit" and "UCB-List", both achieve $O(\log T)$ expected regret with $O(\log T)$ expected incentive over a time horizon T.

Related Work

- Self-reinforcing preferences
  - Preferential attachment [Barabasi et al. 1999]
    - Modeling by multinomial logit model
    - Convergence to one action in social network [Acemoglu et al. 2011]
  - Positive externalities [Shah et al. 2018]
    - Incorporated in MAB framework and proposed optimal algorithms
    - Full control of arm selection
  - Balls and bins models with feedback [Drinea et al. 2002]
    - Convergence under various feedback functions
- Incentivized MAB
  - Adopted incentive schemes into Bayesian MAB [Frazier et al. 2014]
  - Non-Bayesian setting with non-discounted rewards [Wang et al. 2018]
- Bandit with budgets: the budget constraints are pre-determined
  - Approximation algorithms for a large class of budgeted learning problems [Guha et al. 2007]
  - Index-based algorithms [Goel et al. 2009]

Modeling



- Preference on arm a at time t:

$$\lambda_a(t) = \frac{F\big(S_a(t-1) + \theta_a\big)}{\sum_{i \in A} F\big(S_i(t-1) + \theta_i\big)},$$

  - $F(\cdot)$: unknown feedback function
  - θ_a: unknown initial bias

- Incentive Impact on Preference:

$$\hat{\lambda}_i(t) = \begin{cases} \dfrac{G(b,t) + \lambda_i(t)}{G(b,t) + 1}, & i = a, \\ \dfrac{\lambda_i(t)}{G(b,t) + 1}, & i \neq a. \end{cases}$$

  - $G(\cdot)$: unknown incentive impact

Policies: Basic Idea

Structure of the three-phased policies:

1. Exploration: Incentivize arm exploration until finding a best-empirical arm $\hat{a}^*$.
2. Exploitation: Incentivize pulling arm $\hat{a}^*$ until it dominates.
3. Self-Sustaining: Users pull arms based on their preferences until T.

Remark

- *After exploitation, for certain $F(\cdot)$, arm $\hat{a}^*$ is expected to dominate and proved to have exponentially increasing probability to "win" in the monopoly.*
- *The incentive stops after exploitation, which is proved $O(\log T)$, thus $\mathbb{E}[B_T] = O(\log T)$.*

Policies: Basic Idea

At-Least-n Explore-Then-Commit:

1. Exploration: Evenly incentivize arms until each arm generates at least n accumulative reward.
2. Exploitation: Incentivize pulling arm $\hat{a}^*$ until it dominates.
3. Self-Sustaining: Users pull arms based on their preferences until T.

Remark

- *After exploitation, for certain $F(\cdot)$, arm $\hat{a}^*$ is expected to dominate and proved to have exponentially increasing probability to "win" in the monopoly.*
- *The incentive stops after exploitation, which is proved $O(\log T)$, thus $\mathbb{E}[B_T] = O(\log T)$.*

Policies: Basic Idea

UCB-List:

1. Exploration: Evenly incentivize arms. Meanwhile, eliminate all arms that have bad upper confidence bound.
2. Exploitation: Incentivize pulling arm $\hat{a}^*$ until it dominates.
3. Self-Sustaining: Users pull arms based on their preferences until T.

Remark

- *After exploitation, for certain $F(\cdot)$, arm $\hat{a}^*$ is expected to dominate and proved to have exponentially increasing probability to "win" in the monopoly.*
- *The incentive stops after exploitation, which is proved $O(\log T)$, thus $\mathbb{E}[B_T] = O(\log T)$.*

Policies: Upper bounds of Regret and Incentive

- At-Least-n Explore-Then-Commit:

$$\mathbb{E}[R_T] \le \sum_{a \in A} \frac{2(G(b,t) - L_{a^*})\Delta_{max}}{\big(G(b,t)-1\big)\mu_a} \cdot q \ln T + o(\log T),$$

$$\mathbb{E}[B_T] \le \sum_{a \neq a^*} \frac{2b(G(b,t)+1)}{\mu_a(G(b,t)-1)} \cdot q \ln T.$$

- UCB-List:

$$\mathbb{E}[R_T] \le \sum_{a \neq a^*} \left[\frac{8\Delta_a\big(G(b,t)-1\big) + 8\Delta_{max}}{\big(G(b,t)-1\big)\Delta_a^2} \ln T + 4\Delta_a + \frac{4\Delta_{max}}{G(b,t)-1} \right],$$

$$\mathbb{E}[B_T] \le \frac{2G(b,t)+1}{G(b,t)-1} \left[\frac{8b\ln T}{\Delta_{min}^2} + \sum_{a \neq a^*} \left(\frac{8b \ln T}{\Delta_a^2} + 4b \right) \right].$$

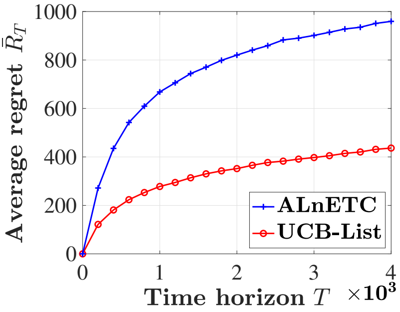
Remark

Both achieve $O(\log T)$ expected regret with $O(\log T)$ expected incentive.

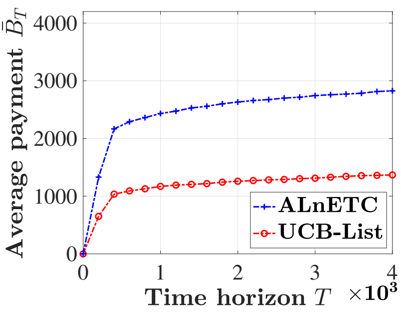
Simulations

Up to time T:

– Expected Regret $\mathbb{E}[R_T]$:



– Expected incentive $\mathbb{E}[B_T]$:



Thanks!