Adversarial Combinatorial Bandits with General Non-linear Reward Functions

Yanjun Han*, Yining Wang†, Xi Chen‡

*Department of Electrical Engineering, Stanford University. Email: yjhan@stanford.edu

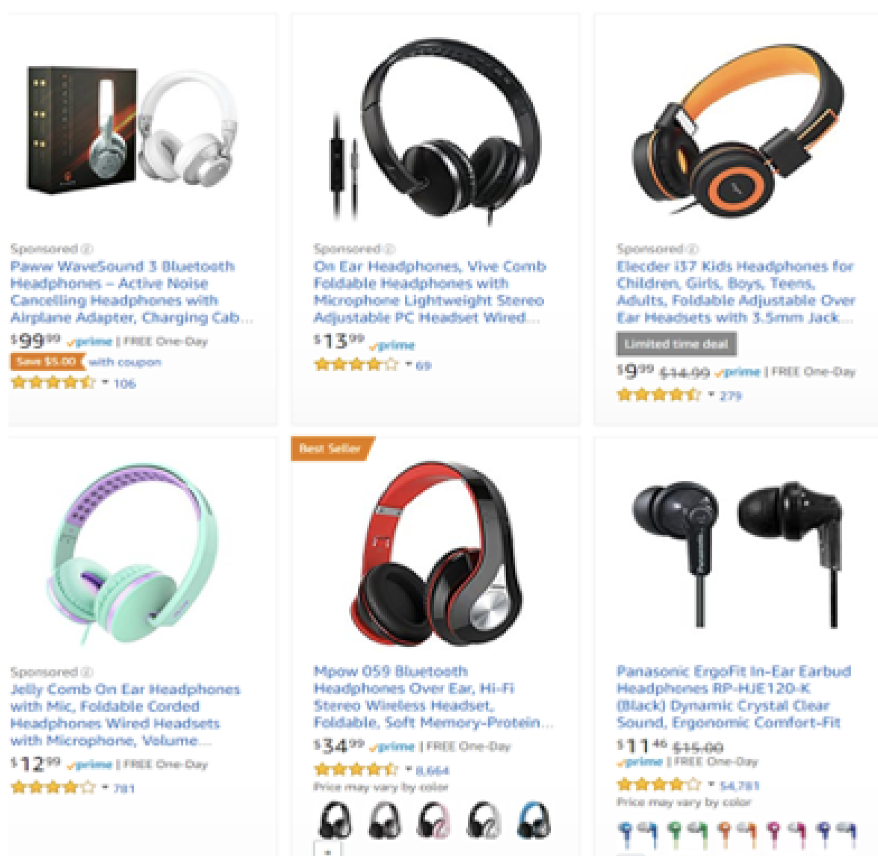
†Warrington College of Business, University of Florida. Email: yining.wang@warrington.ufl.edu

‡Stern School of Business, New York University. Email: xc13@stern.nyu.edu

Motivation: assortment optimization

Assortment optimization:

- Select a subset of substitutable items to maximize expected revenue
- Key step of recommendation in online retailing



Mathematical model

Multinomial Logit (MNL) model:

- N available items in the pool
- each item has a revenue $r_i \in [0,1]$, and a choice probability $v_i \in [0,1]$
- seller offers an assortment $S \subseteq [N]$ of size K
- customer selects item i with probability

$$p_i(S,v) = \frac{v_i}{\underbrace{1}_{\text{"no-purchase"}} + \sum_{j\in S} v_j}$$

- seller's observation: the chosen item or "no-purchase"
- seller's expected revenue when offering assortment S:

$$R(S,v) = \sum_{i\in S} p_i(S,v) r_i = \frac{\sum_{j\in S} r_j v_j}{1+\sum_{j\in S} v_j}$$

Static vs. dynamic model

Regret in repeated assortment optimization:

$$\mathbb{E}\left[\max_{S:|S|=K} \sum_{t=1}^T R(S,v_t) - \sum_{t=1}^T R(S_t,v_t)\right]$$

Static model: $v_t \equiv v$ for all $t\in[T]$

- $\widetilde{O}(\sqrt{NT})$ regret achievable [Rusmevichientong et al. 2010, Agrawal et al. 2019, ...]

Dynamic model: v_t may change across time

- open question: is $O(\sqrt{\text{poly}(N,K)T})$ regret still achievable?

Combinatorial adversarial bandit

A more general bandit problem:

- time horizon T, number of arms N
- at each time $t\in[T]$, a reward vector $v_t \in [0,1]^N$ is chosen
- the learner chooses $S_t \subseteq [N]$ of size K, and observes bandit feedback

$$r_t \sim \text{Bernoulli}\left(R(S_t,v_t)\right), \quad \text{where } R(S_t,v_t) = g\left(\sum_{j\in S_t} v_{t,j}\right)$$

- $g:\mathbb{R}_+ \to [0,1]$ is a known link function
- learner's regret:

$$\mathbb{E}\left[\max_{S:|S|=K} \sum_{t=1}^T R(S,v_t) - \sum_{t=1}^T R(S_t,v_t)\right]$$

Multinomial Logit model: a special case with $g(x) = x/(1+x)$

Main results

Theorem: For general adversarial combinatorial bandits, the optimal regrets are:

- $\widetilde{\Theta}_{g,K}(\sqrt{TN^d})$ if g is a polynomial of degree $d\le K$;
- $\widetilde{\Theta}_{g,K}(\sqrt{TN^K})$ if g is not a polynomial of degree $\le K$.

Corollary: $O(\sqrt{\text{poly}(N,K)T})$ regret is impossible in dynamic assortment selection

Proof technique

High-level idea: find a distribution μ on v such that

$$\mathbb{E}_{v\sim\mu}[\mathbb{P}(\cdot\mid S,v)] = \begin{cases}\mathbb{P}_0, & \text{if } S = S^\star.\\ \mathbb{P}_1, & \text{if } S\neq S^\star.\end{cases}$$

Intuition: no information is leaked unless the learner guesses the optimal assortment $S^\star$ exactly, even if S and $S^\star$ have a lot in common

Past settings where a small regret could be obtained:

- Static model [Agrawal et al. 2019]: v is deterministic and fixed, so $\mathbb{P}(\cdot\mid S,v)$ and $\mathbb{P}(\cdot\mid S',v)$ must be correlated as long as $S\cap S'\neq\emptyset$
- Combinatorial linear bandit [Bubeck et al. 2012]: when $g(x)\propto x$, the mean of $\mathbb{E}_{v\sim\mu}[\mathbb{P}(\cdot\mid S,v)]$ is

$$\mathbb{E}_{v\sim\mu}[g(\langle 1_S, v\rangle)] = g(\langle 1_S, \mathbb{E}_{v\sim\mu}[v]\rangle),$$

  depending linearly on 1_S, so no such construction
- Combinatorial bandit with stochastic dominance [Agarwal and Aggarwal, 2018]: when an element of $[N]\backslash S^\star$ is replaced by an element of $S^\star$, the mean of $\mathbb{E}_{v\sim\mu}[\mathbb{P}(\cdot\mid S,v)]$ must increase

An example construction

Assortment optimization with $K=2$ and $g(x) = x/(1+x)$:

- choose $S^\star = (i^\star,j^\star) \in \binom{[N]}{2}$ uniformly at random
- construction of $v\sim\mu$:

$$v_k \equiv \frac{1}{2}, \quad k\notin\{i^\star,j^\star\}, \qquad (v_{i^\star},v_{j^\star}) = \begin{cases}(1,1) & \text{w.p. } 1/4,\\ (0,1) & \text{w.p. } 3/8,\\ (1,0) & \text{w.p. } 3/8.\end{cases}$$

- key property: the multinomial distribution

$$\mathbb{E}\left(\frac{1}{1+v_i+v_j}, \frac{v_i}{1+v_i+v_j}, \frac{v_j}{1+v_i+v_j}\right)$$

  is always $(1/2,1/4,1/4)$ unless the precise pair $(i^\star,j^\star)$ is chosen

General construction

Key technical lemma: Let $g\in C^m([0,b])$ be a real-valued and m-times continuously differentiable function on $[0,b]$, with $b\ge m$. Then the following two statements are equivalent:

- g is not a polynomial of degree at most $m-1$;
- there exists a random vector $(X_1,\cdots,X_m)$ supported on $[0,1]^m$, which follows an exchangeable joint distribution μ, and a scalar $x_0\in[0,1]$, such that

$$\mathbb{E}_\mu[g(X_1+\cdots+X_{\ell-1}+(b-\ell+1)x_0)] = \mathbb{E}_\mu[g(X_1+\cdots+X_\ell+(b-\ell)x_0)]$$

  for all $\ell = 1,2,\cdots,m-1$, and

$$\mathbb{E}_\mu[g(X_1+\cdots+X_{m-1}+(b-m+1)x_0)] < \mathbb{E}_\mu[g(X_1+\cdots+X_m+(b-m)x_0)].$$

Proof technique: duality existential arguments, which in turn also applies several technical tools from real analysis and functional analysis

Future direction

Interpolation between static and regret models: v_t is only allowed to change M times. How does the regret depend on M?

References

- Agarwal, M. and Aggarwal, V. *Regret bounds for stochastic combinatorial multi-armed bandits with linear space complexity.* arXiv preprint arXiv:1811.11925, 2018.
- Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnlbandit. *A dynamic learning approach to assortment selection.* Operations Research, 67(5):1453–1485, 2019.
- Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. *Towards minimax policies for online linear optimization with bandit feedback.* Conference on Learning Theory, 2012.
- Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. *Dynamic assortment optimization with a multinomial logit choice model and capacity constraint.* Operations Research, 58(6):1666–1680, 2010.