Bandit Learning with Positive Externalities

Virag Shah, Jose Blanchet, Ramesh Johari

Management Science and Engineering Department, Stanford University

Stanford University

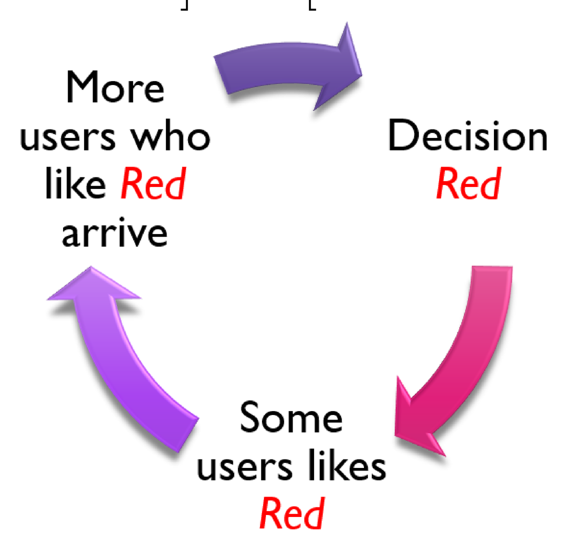
Positive externalities in online platforms

ebay Etsy upwork airbnb Thumbtack urbansitter STITCH FIX instacart

- A positive experience attracts more users of the same type.
- Aka **Positive Externalities / Self- Reinforcement / Network Effects**, etc.
- Thus, the arrival process is influenced by decisions.
- This makes simultaneous learning and decision-making more challenging

What could go wrong?

- Suppose Blue users like Blue items (but not Red items)
- Suppose Red users like Red items (but not Blue items)
- Suppose user type not known upon arrival
- $\mathbb{E}$[Blue-Blue match reward] > $\mathbb{E}$[Red-Red match reward]



- Successful Red-Red matches made early on may trigger more Red user arrivals.
- So the platform might learn to prefer Red-Red matches even if it is suboptimal!

Main insights from our results

- There is a cost to being optimistic in the face of uncertainty as initial mistakes are amplified. UCB algorithm, in particular, fails miserably.
- It is possible to reduce the impact of transients by structuring the exploration procedure well.
- Once enough evidence is gathered, one may use the externalities to shift the arrivals to the reward-maximizing population.

Model

Standard bandit setting:

- m: Number of arms (items)
- T : Time horizon; one user arrives per time step
- μ_a: Expected reward when arm a pulled (Bernoulli)
- a^*: best arm
- $T_a(t)$: number of times arm a pulled up to time t
- $S_a(t)$: total reward at arm a up to time t
- Goal: maximize expected total reward (Γ_T).
- We study performance asymptotic in T.

Positive externalities:

- Let θ_a be initial "bias" of arm a.
- We assume the user arriving at time t *likes* arm a independently with probability:

$$\lambda_a(t) = \frac{f(\theta_a + S_a(t))}{\sum_b f(\theta_b + S_b(t))}.$$

- f is the *externality function*. We consider $f(x) = x^\alpha$, $\alpha \geq 0$. Here, α determines the strength of the positive externality.
- $\mathbb{P}$(reward at t| arm a pulled) $= \mu_a$ if user t likes a, otherwise zero.

The baseline oracle

Since we study performance that is asymptotic in T, natural to consider a baseline oracle that *always chooses arm* a^*.

Proposition

The oracle earns $\mathbb{E}[\Gamma_T^*] = \begin{cases} \mu_{a^*}T - \Theta(T^{1-\alpha}), & 0 < \alpha < 1 \\ \mu_{a^*}T - \Theta(\ln T), & \alpha = 1 \\ \mu_{a^*}T - \Theta(1), & \alpha > 1 \end{cases}$

Intuition: Suppose $\alpha = 1$. We need $\Omega(\log T)$ time to remove any initial bias toward suboptimal arms, since:

$$\mathbb{P}(\text{user } t \text{ likes } a^*) \approx 1 - \frac{\sum_{a \neq a^*} \theta_a}{O(t) + \sum_{a \neq a^*} \theta_a}.$$

We measure performance of any algorithm against baseline oracle as *expected regret*: $R_T = \mathbb{E}[\Gamma_T^*] - \mathbb{E}[\Gamma_T]$.

Main Results

Regret Lower Bound:

Theorem

Any feasible policy must have expected regret

$$R_T = \begin{cases} \Omega(T^{1-\alpha} \ln^\alpha T\ ), & 0 < \alpha < 1 \\ \Omega(\log^2 T\ ), & \alpha = 1 \\ \Omega(\log^\alpha T\ ), & \alpha > 1 \end{cases}.$$

Optimal Algorithm:

Balanced-Exploration (BE):

Suppose $w_k = \ln \ln k$ for each $k \geq 1$. Fix $\tau = w_T \ln T$.

- For $t \leq \tau$, pull the arm with lowest cumulative reward $S_a(t-1)$ (ties broken at random).
- For $t > \tau$, pull the arm with highest mean reward $S_a(\tau)/T_a(\tau)$ at time τ.

Balanced Exploration with Arm Elimination (BE-AE): Dynamically *eliminate poorly performing arms* while balancing the exploration of the rest. (Needs knowledge of α & ($\theta_a : 1 \leq a \leq m$).)

Full Picture

	$\alpha = 0$	$0 < \alpha < 1$	$\alpha = 1$	$\alpha > 1$
Lower bound	$\Omega(\log T\ )$	$\Omega(T^{1-\alpha} \log^\alpha T\ )$	$\Omega(\log^2 T\ )$	$\Omega(\log^\alpha T\ )$
UCB	$O(\log T\ )$	$\Omega(T\ )$	$\Omega(T\ )$	$\Omega(T\ )$
BE	$\tilde{O}(\log T\ )$	$\tilde{O}(T^{1-\alpha} \log^\alpha T\ )$	$\tilde{O}(\log^2 T\ )$	$\tilde{O}(\log^\alpha T\ )$
BE-AE	$O(\log T\ )$	$O(T^{1-\alpha} \log^\alpha T\ )$	$O(\log^2 T\ )$	$O(\log^\alpha T\ )$

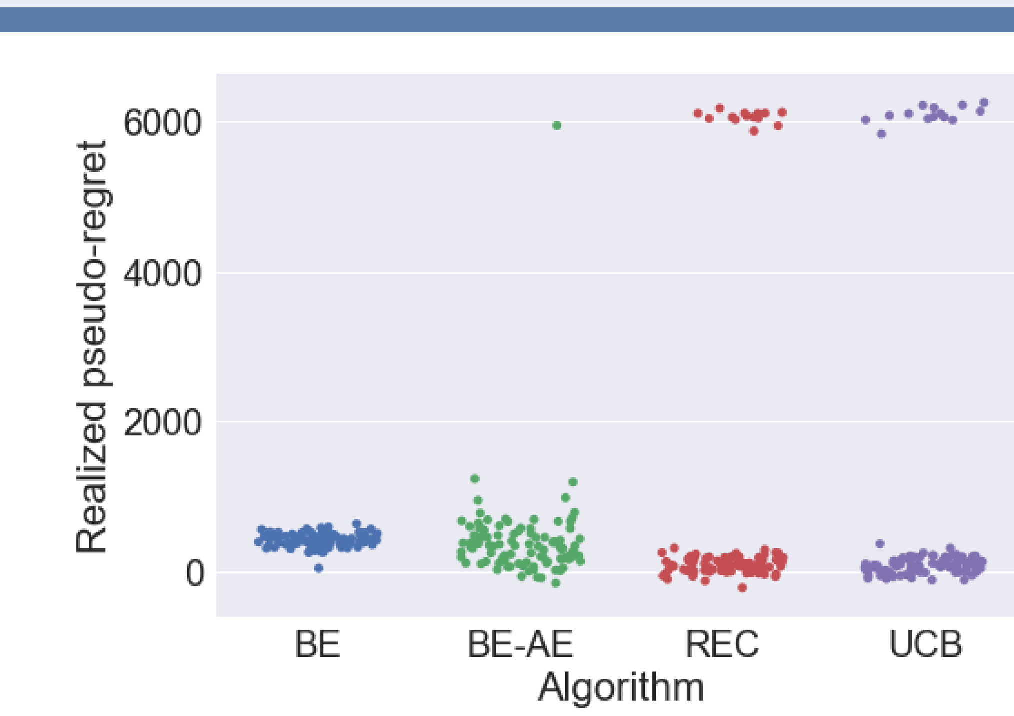


Figure 1: $T = 30,000$, $\alpha = 1$, $m = 2$, $\mu_1 = 0.5$, $\mu_2 = 0.3$, $\theta_1 = \theta_2 = 1$.

REC: Random-explore-then-commit.