Bandit Learning with Positive Externalities

Virag Shah, Jose Blanchet, Ramesh Johari
Management Science and Engineering Department, Stanford University

Positive externalities in online platforms

- 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
- \( E[\text{Blue-Blue match reward}] > E[\text{Red-Red match reward}] \)

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.

Standard bandit setting:

- \( m \): Number of arms (items)
- \( T \): Time horizon; one user arrives per time step
- \( \mu_a \): 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 (\( \Gamma_T \)).
- We study performance asymptotic in \( T \).

Positive externalities:

- Let \( \theta_a \) be initial “bias” of arm \( a \).
- We assume the user arriving at time \( t \) likes arm \( a \) independently with probability:
  \[
  \lambda_a(t) = \frac{f(\theta_a + S_a(t))}{\sum_a f(\theta_a + S_a(t))}.
  \]
- \( f \) is the externality function. We consider \( f(x) = x^\alpha, \alpha \geq 0 \).
- Here, \( \alpha \) determines the strength of the positive externality.
- \( \mathbb{P}(\text{reward at } t | \text{ arm } a \text{ pulled}) = \mu_a \) 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

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

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

\[
\mathbb{P}(\text{user } t \text{ likes } a^*) \approx 1 - \frac{\sum_a \theta_a}{\Theta(T) + \sum_a \theta_a}.
\]

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

Regret Lower Bound:

<table>
<thead>
<tr>
<th>Theorem</th>
<th>( \alpha = 0 )</th>
<th>( 0 &lt; \alpha &lt; 1 )</th>
<th>( \alpha = 1 )</th>
<th>( \alpha &gt; 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Bound</td>
<td>( \Omega(\log T) )</td>
<td>( \Omega(T^{-\alpha}\log^a T) )</td>
<td>( \Omega(\log^2 T) )</td>
<td>( \Omega(\log T) )</td>
</tr>
<tr>
<td>UCB</td>
<td>( \Omega(T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
</tr>
<tr>
<td>BE</td>
<td>( \Omega(\log T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
</tr>
<tr>
<td>BE-AE</td>
<td>( \Omega(\log T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
<td>( \Theta(T) )</td>
</tr>
</tbody>
</table>

Optimal Algorithm:

Balanced-Exploration (BE):
Suppose \( w_k = \ln k \) 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 \( \tau \).

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

Full Picture

![Full Picture](image_url)

**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.