Leveraging Side Observations in Stochastic Bandits

Stéphane Caron, Technicolor Labs Palo Alto
Branislav Kveton, Technicolor Labs Palo Alto
Marc Lelarge, INRIA-ENS
Smriti Bhagat, Technicolor Labs Palo Alto

Motivation and Goals

Motivation:

- Multi-Armed Bandits (MABs) in Social Networks
- Social interactions bring *Side Observations* to the MAB
- Leverage this additional information to learn faster

Contributions:

- Novel UCB strategies
- Better theoretical guarantees depending on *graph structure*
- Experiments in Movie Rec.: 2x to 14x faster learning rates

v. # of arms in traditional bandits

Stochastic Multi-Armed Bandits

- Set of K "arms" associated with probability distributions $P_1,\dots,P_k$
- At step t, the decision maker pulls an arm i and gets reward $X_{it} \sim P_i$
- **Goal:** minimize the cumulative regret:

$$R(n) := \sum_{t=1}^{n} X_{i^*,t} - \sum_{t=1}^{n} X_{I_t,t}$$

time horizon; best arm; arm pulled at step t

- **Theorem (Lai and Robbins, 1975)**
  No strategy can achieve better asymptotic regret than

$$\mathbb{E}[R(n)] = \Omega(\ln n)$$

Upper Confidence Bound Policies

At each step t, pull the arm i maximizing

$$\mathrm{UCB}_i(t) := \overline{X}_{i,T_i(t-1)} + \sqrt{\frac{2 \ln t}{T_i(t-1)}}$$

empirical reward estimate at step t; # pulls to arm i up to step t - 1

Features:

- Well-grounded: interpretation with confidence regions
- Simplicity: easy to implement, fit for distributed computing
- Optimality: achieve the optimal $O(\log t)$ bound uniformly over time: (Auer et al. 2002)

$$\mathbb{E}[R(n)] \leq 8 \left(\sum_{i=1}^{K} \frac{1}{\Delta_i} \right) \ln n + \left(1 + \frac{\pi^2}{3} \right) \sum_{i=1}^{K} \Delta_i$$

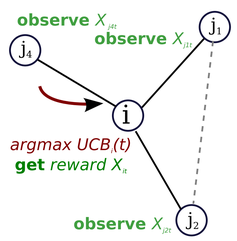
number of arms; individual expected regret of arm i

UCB with Side Observations: **UCB-N**

- Extend UCB-1 to the Side Observations setting

Algorithm 1 UCB-N

$\overline{\mathbf{X}}, \mathbf{O} \leftarrow \mathbf{0}, \mathbf{0}$
for $t \geq 1$ **do**
  $i \leftarrow \arg\max_i \left\{ \overline{X}_i + \sqrt{\frac{2 \ln t}{O_i}} \right\}$
  pull arm i — neighborhood of arm i
  for $k \in N(i)$ **do** — # observations of arm k
    $O_k \leftarrow O_k + 1$
    $\overline{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\overline{X}_k$
  end for
end for



- **Theorem**
  UCB-N achieves the following uniform guarantee: for *any* clique coverage C,

$$\mathbb{E}[R(n)] \leq 8 \left(\inf_{\mathcal{C}} \sum_{C \in \mathcal{C}} \frac{\max_{i \in C} \Delta_i}{\min_{i \in C} \Delta_i^2} \right) \ln n + O(K)$$

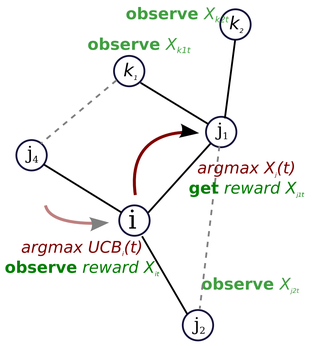
  i.e., the first term is now $O(\#\text{ cliques})$ instead of $O(K)$.

A Better Strategy: **UCB-MaxN**

- Exploration: use UCB to decide the arm we want to observe
- Exploitation bias: pull its best empirical neighbor

Algorithm 2 UCB-MaxN

$\overline{\mathbf{X}}, \mathbf{n} \leftarrow \mathbf{0}, \mathbf{0}$
for $t \geq 1$ **do**
  $i \leftarrow \arg\max_i \left\{ \overline{X}_i + \sqrt{\frac{2 \ln t}{O_i}} \right\}$
  $j \leftarrow \arg\max_{j \in N(i)} \overline{X}_j$
  pull arm j
  for $k \in N(j)$ **do**
    $O_k \leftarrow O_k + 1$
    $\overline{X}_k \leftarrow X_{k,t}/O_k + (1 - 1/O_k)\overline{X}_k$
  end for
end for



- **Theorem**
  With UCB-MaxN, the $O(K)$ upper-bound term is improved to: for *any* clique coverage C,

$$\left(1 + \frac{\pi^2}{3} \right) \sum_{C \in \mathcal{C}} \Delta_C + o_{n \to \infty}(1)$$

  i.e., asymptotically, all terms are now $O(\#\text{ cliques})$.

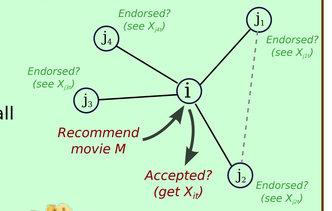
Lower Bound

Theorem
For any uniformly good allocation rule in the multi-armed bandit with side observations setting, $\mathbb{E}[R(n)] = \Omega(\ln n)$.

Experimental Setup

Movie Recommendation in Social Networks

— Target a user *i* in the network
— Send *i* a movie recommendation
  along with a promotional offer, e.g., a discount
  if he/she reacts about his experience on his Wall
— Observe *i*'s feedback
  and his/her friends' reactions



Flixster dataset
— Subset of active users
— 5K users
— 5K movies
— 1.7M ratings

Facebook dataset
— Strongly connected subgraph
— 14K users, 50K edges
— Ratings extrapolated from Flixster

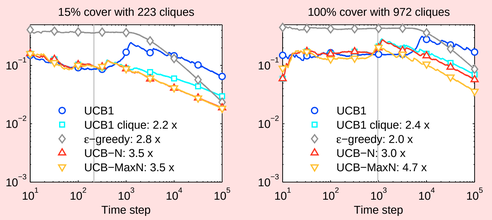
Results

Metric: per-step regret $r(n) := \frac{1}{n}\mathbb{E}[R(n)]$

Test Beds:
— *Friends*: initial social network
— *Core*: 15% cover by biggest cliques
— *Friends-of-friends* (FoF): 2-hop graph

Takeaway:
UCB-MaxN ≫ UCB-N
UCB-N ≫ UCB-1

Flixster:
— Friends: 2.2x
— Core: 3.5x
— FoF: 4.7x



Facebook:
— Friends: 2.9x
— Core: 20.6x
— FoF: 14.1x