Combinatorial Multi-armed Bandit Problem with Probabilistically Triggered Arms - A Case with Bounded Regret

A. Omer Saritac, Cem Tekin

Bilkent University
Electrical and Electronics Engineering Department

November 14, 2017
The multi-armed bandit problem

**Classical MAB** [Lai and Robbins 85]:

- System operates over epochs $t = 1, 2, \ldots$ (learning over time)
- Set of arms: $\mathcal{M} = \{1, \ldots, m\}$
- Select arm $a_t$, receive reward $X_{a_t}^{(t)}$
- Goal: Maximize $\mathbb{E} \left[ \sum_{t=1}^{T} X_{a_t}^{(t)} \right]$
- Distribution of $X_{i}^{(t)}$ is fixed but unknown
The multi-armed bandit problem

**Classical MAB** [Lai and Robbins 85]:

- System operates over epochs $t = 1, 2, \ldots$ (learning over time)
- Set of arms: $\mathcal{M} = \{1, \ldots, m\}$
- Select arm $a_t$, receive reward $X_{a_t}^{(t)}$
- Goal: Maximize $\mathbb{E}[\sum_{t=1}^{T} X_{a_t}^{(t)}]$
- Distribution of $X_{i}^{(t)}$ is fixed but unknown

**Combinatorial MAB (CMAB)** [Gai et al 12]:

- Select $S_t \subset \mathcal{M}$
- Reward is a combination of the rewards of arms in $S_t$
The multi-armed bandit problem

Classical MAB [Lai and Robbins 85]:
- System operates over epochs \( t = 1, 2, \ldots \) (learning over time)
- Set of arms: \( \mathcal{M} = \{1, \ldots, m\} \)
- Select arm \( a_t \), receive reward \( X_{a_t}^{(t)} \)
- Goal: Maximize \( \mathbb{E}[\sum_{t=1}^{T} X_{a_t}^{(t)}] \)
- Distribution of \( X_i^{(t)} \) is fixed but unknown

Combinatorial MAB (CMAB) [Gai et al 12]:
- Select \( S_t \subset \mathcal{M} \)
- Reward is a combination of the rewards of arms in \( S_t \)

CMAB w Prob. Triggered Arms (CMAB-PTA) [Chen et al 16]
- Select \( S_t \subset \mathcal{M} \)
- \( \tau_t \subset \mathcal{M} \) gets triggered
- Reward is a combination of the rewards of arms in \( S_t \cup \tau_t \)
Motivation: Viral marketing

Network: $n$ nodes, $m$ edges

Action: select $k < n$ node seed set $S$

Influence spread model: Nodes in $S$ can influence their neighbors, and so on ... (influence probabilities unknown)

$$k = 1 \quad G(V, E, p)$$
Example: Influence maximization (IM)

In epoch $t$, select $S_t$ based on $G(V, E, \hat{p})$

- $X_{(i,j)}$: state of edge $(i, j)$
- $X_{(i,j)} = 1$: influence successful (triggered)
- $X_{(i,j)} = 0$: influence unsuccessful (not triggered)

**Expected state:**

- $\mu_{i,j} := \mathbb{E}[X_{(i,j)}] = p_{i,j}$ [unknown]
Example: Influence maximization (IM)

Set of actions:

\[ S = \{ \text{All } k \text{ out of } n \text{ combinations of nodes} \} \]

Set of triggered edges:

\[ \tau \]

Reward:

\[ R(S, X, \tau) = \text{num. influenced nodes} \]
\[ = \text{influence spread} \]}
Maximize the cumulative expected reward by epoch $T$, for all $T$:

$$\maximize \mathbb{E} \left[ \sum_{t=1}^{T} R(S_t, X^{(t)}, \tau_t) \right]$$

- Need to learn $p_{i,j}$s!
Arms and actions

- $X_i^{(t)}$: state of arm $i$ at epoch $t$
- $X^{(t)} = (X_1^{(t)}, \ldots, X_m^{(t)})$: state vector [not known beforehand]
- $X^{(t)} \sim D$
- Expected state: $\mu_i = \mathbb{E}[X_i^{(t)}]$
- Expectation vector: $\mu = (\mu_1, \ldots, \mu_m)$
- Set of actions: $S$
What happens in epoch $t$?

- Select an action: $S_t \in \mathcal{S}$
- Arms get probabilistically triggered: $\tau_t \sim D^{\text{trig}}(S_t, X^{(t)})$  \hspace{1cm} [$\tau_t \subset \mathcal{M}$]
- Receive a non-negative reward: $R(S_t, X^{(t)}, \tau_t)$
- Observe states of triggered arms: $X^{(t)}_i, i \in \tau_t$

Assumption:

$\mathbb{E}[R(S, X, \tau)] = r_{\mu}(S)$ (expected reward only depends on $\mu$ and $S$)
Problem is NP hard, but approximations exist! [Vazirani 2001]

- Optimal expected reward: $r^*_\mu = \max_{S \in S} r_\mu(S)$
- $(\alpha, \beta)$-approximation algorithm

$$\text{action } S^O : \Pr(r_{\hat{\mu}}(S^O) \geq \alpha r^*_\mu) \geq \beta$$
Regret

Regret by epoch $T$:

$$\text{Reg}_{\mu,\alpha,\beta}(T) = \underbrace{T \alpha \beta r^*_\mu}_{(\alpha, \beta) \text{ oracle}} - \mathbb{E} \left[ \sum_{t=1}^{T} r_{\mu}(S_t) \right]$$

maximize $\mathbb{E} \left[ \sum_{t=1}^{T} r_{\mu}(S_t) \right] \approx$ minimize Regret
Assumptions on the expected reward

**Assumption (Chen 2016 - bounded smoothness)**

If \( \max_{i \in \{1, \ldots, m\}} |\mu_i - \mu'_i| \leq \Delta, \forall S \in S, \) then

\[
|r_{\mu}(S) - r_{\mu'}(S)| \leq f(\Delta)
\]

- \( f \): continuous, strictly increasing *bounded smoothness function* \((f(0) = 0)\).

**Assumption (Chen 2016 - monotonicity)**

If for all arms \( i \in \{1, \ldots, m\}, \mu_i \leq \mu'_i \), then we have

\[
r_{\mu}(S) \leq r_{\mu'}(S), \forall S \in S
\]
Positive arm triggering probabilities (CMAB-PTA$^+$)

- $p_i^S$: minimum probability that action $S$ triggers arm $i$
- CMAB-PTA: $p_i^S$ can be zero
- CMAB-PTA$^+$: $p_i^S \geq p^* > 0$

Examples of CMAB-PTA$^+$:
- Influence maximization over strongly connected graphs
- Recommender systems with word of mouth effect
## Our contributions

<table>
<thead>
<tr>
<th></th>
<th>Our work</th>
<th>CMAB PTAs [Chen 16] [Wang 17]</th>
<th>CMAB [Kveton 15] [Chen 16b]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap-dependent regret</td>
<td>$O(1)$</td>
<td>$O(\log T)$</td>
<td>$O(\log T)$</td>
</tr>
<tr>
<td>Gap-independent regret</td>
<td>$O(\sqrt{T})$</td>
<td>$O(\sqrt{T \log T})$</td>
<td>$O(\sqrt{T \log T})$</td>
</tr>
<tr>
<td>Strictly positive ATPs</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
</tr>
</tbody>
</table>

- First to show bounded regret in CMAB with PTAs
Bounded regret in other bandits

A negative result:

- [Lai and Robbins 85]: regret $\Omega(\log T)$ (arms do not provide information about each other)

Positive results:

- [Mersereau 09], [Atan 15]*, [Akbarzadeh 16]**: arm rewards are related through parameter(s) that can be learned by selecting any arm.

---

*O. Atan, C. Tekin, M. van der Schaar “Global bandits”, AISTATS 2015

**N. Akbarzadeh, C. Tekin “Gambler’s ruin bandit problem”, Allerton 2016
Greedy policy for CMAB-PTA$^+$ (pure exploitation)

1: Maintain $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_m)$ (sample mean estimate of $\mu$)
2: \textbf{while} $t \geq 1$ \textbf{do}
3: \hspace{1em} Call the $(\alpha, \beta)$-approximation algorithm with $\hat{\mu}$ as input to get $S_t$
4: \hspace{1em} Select action $S_t$, observe $X_i^{(t)}$'s for $i \in \tau_t$ and collect the reward $R$
5: \hspace{1em} \textbf{for} $i \in \tau_t$ \textbf{do}
6: \hspace{2em} $T_i = T_i + 1$
7: \hspace{2em} $\hat{\mu}_i = \hat{\mu}_i + \frac{X_i^{(t)} - \hat{\mu}_i}{T_i}$
8: \hspace{1em} \textbf{end for}
9: \hspace{1em} $t = t + 1$
10: \textbf{end while}
### Lemma (Sufficient arm observations)

For any learning algorithm, \( \eta \in (0, 1) \) and for all \( t \geq t' := 4c^2/e^2 \), where \( c := 1/(p^*(1 - \eta))^2 \), we have

\[
\Pr \left( \bigcup_{i \in \{1, \ldots, m\}} \left\{ T_i^{t+1} \leq \eta p^* t \right\} \right) \leq \frac{m}{t^2}.
\]

- \( t' \): turning point
- Num. observations of each arm is linear in \( t \) after the turning point
Gap-dependent regret

**Theorem**

\[ \text{Reg}^{\text{greedy}}(T) = O(1) \]

**Finite time version:** \( \forall T \geq 1 \)

\[
\text{Reg}^{\text{greedy}}_{\mu,\alpha,\beta}(T) \leq \nabla_{\max} \inf_{\eta \in (0,1)} \left( \lceil t' \rceil + \frac{m \pi^2}{3} \left( 1 + \frac{1}{2\delta^2} \right) + 2m \left( 1 + \frac{1}{2\delta^2 \eta p^*} \right) \right)
\]

- \( \delta := f^{-1}(\nabla_{\min}/2) \)
- \( t' := 4c^2/e^2 \) and \( c := 1/(p^*(1 - \eta))^2 \)
- \( \nabla_{\min} = \min_{S: \nabla S > 0} \nabla S \) where

\[
\nabla S = \alpha r^*_\mu - r_\mu(S) \text{ [suboptimality gap]}
\]
Gap-independent regret

Theorem

\[ \text{Reg}_{\mu, \alpha, \beta}^{\text{greedy}}(T) = O(\sqrt{T}) \]

Finite time version: \( \forall T \geq 1 \)

\[ \text{Reg}_{\mu, \alpha, \beta}^{\text{greedy}}(T) \leq \inf_{\eta \in (0,1)} \left( \lceil t' \rceil \nabla_{\max} + 4\gamma m \left[ 2\left( \frac{\pi}{2\eta \pi^*} \right)^{1/2} + 3 \right] T^{1/2} \right) \]

where \( t' := 4c^2/e^2 \) and \( c := 1/(p^*(1 - \eta))^2 \).

- Holds when the bounded-smoothness function is \( f(x) = \gamma x \) where \( \gamma > 0 \) and \( \omega \in (0, 1] \)
- Matches with the lower bound in [Wang 17] (tight). Upper bound in [Wang 17] is \( \tilde{O}(\sqrt{T}) \)
Movie recommendation example

- Movielens dataset
- Weighted bipartite graph $G = (L, R, E, p)$
- $L$: 50 movies, $R$: 881 users, $E$: movie-user pairs
- Action: select $k$ movies
- $p_{S}^{(i,j)}$: probability that action $S$ triggers edge $(i,j)$
  
  $$ p_{S}^{(i,j)} = 1 \text{ for outgoing edges of nodes in } S $$
  
  $$ > p^* > 0 \text{ otherwise [word of mouth] } $$

- $p_{i,j}$: probability user $j$ watches movie $i$ (after he/she learns about the movie)
Reported regrets are normalized, i.e., divided by the $\alpha \beta$ fraction of the optimal reward.

- Learning is faster when $p^*$ or $k$ is large.
Conclusion

- Considered a special case of CMAB with PTAs.
  - Proved that the gap-dependent regret is $O(1)$
  - Proved that worst-case regret is $O(\sqrt{T})$

Recent extensions

- $O(1)$ gap-dependent and $O(\sqrt{T})$ gap-independent regrets for Combinatorial Upper Confidence Bound (CUCB) and Combinatorial Thompson Sampling (CTS) [they both explore and exploit]
probabilistically triggered arms." The Journal of Machine Learning Research 17.1
Media, 2001