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 $\mathbb{E}[\sum_{t=1}^{T} X_{a_t}^{(t)}]$
- 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

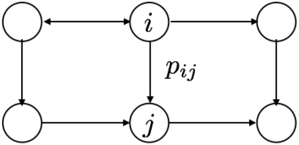
CMAB w Prob. Triggered Arms (CMAB-PTA) [Chen et al 16]

- Select $S_t \subset \mathcal{M}$
- $\tau_t \subset \mathcal{M}$ gets triggered
- Reward is a combination of the rewards of arms in $S_t \cup \tau_t$

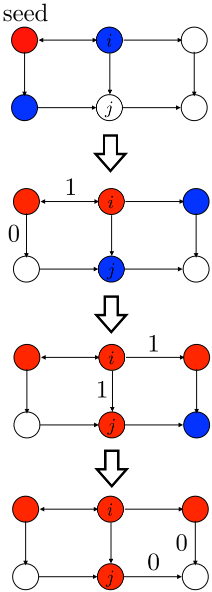
Example: Influence maximization (IM)

- 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)

$$k = 1 \qquad G(V, E, p)$$



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:

$$\mathcal{S} = \{\text{All } k \text{ out of } n \text{ combinations of nodes}\}$$

Set of triggered edges:

$$\tau$$

Reward:

$$\begin{aligned} R(S, \mathbf{X}, \tau) &= \text{ num. influenced nodes} \\ &= \text{ influence spread} \end{aligned}$$

Goal

Maximize the cumulative expected reward by epoch T, for all T:

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

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

CMAB-PTA

Arms and actions

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

CMAB with PTAs

What happens in epoch t?

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

Assumption:
$\mathbb{E}[R(S, \mathbf{X}, \tau)] = r_{\boldsymbol{\mu}}(S)$ (expected reward only depends on $\boldsymbol{\mu}$ and S)

Approximation algorithms

Problem is NP hard, but approximations exist! [Vazirani 2001]

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

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

Regret

Regret by epoch T:

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

$$\text{maximize } \mathbb{E}\left[\sum_{t=1}^{T} r_{\boldsymbol{\mu}}(S_t)\right] \approxeq \text{minimize Regret}$$

Assumptions on the expected reward

Assumption (Chen 2016 - bounded smoothness)

If $\max_{i \in \{1,\dots,m\}} |\mu_i - \mu'_i| \leq \Delta$, $\forall S \in \mathcal{S}$, *then*

$$|r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu}'}(S)| \leq f(\Delta)$$

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

Assumption (Chen 2016 - monotonicity)

If for all arms $i \in \{1, \dots, m\}$, $\mu_i \leq \mu'_i$, *then we have*

$$r_{\boldsymbol{\mu}}(S) \leq r_{\boldsymbol{\mu}'}(S), \ \forall S \in \mathcal{S}$$

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

	Our work	CMAB PTAs [Chen 16] [Wang 17]	CMAB [Kveton 15] [Chen 16b]
Gap-dependent regret	$O(1)$	$O(\log T)$	$O(\log T)$
Gap-independent regret	$O(\sqrt{T})$	$O(\sqrt{T \log T})$	$O(\sqrt{T \log T})$
Strictly positive ATPs	Yes	No	-

- 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)

Maintain $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_m)$ (sample mean estimate of $\boldsymbol{\mu}$)
while $t \geq 1$ **do**
    Call the (α, β)-approximation algorithm with $\hat{\boldsymbol{\mu}}$ as input to get S_t
    Select action S_t, observe $X_i^{(t)}$'s for $i \in \tau_t$ and collect the reward R
    for $i \in \tau_t$ **do**
        $T_i = T_i + 1$
        $\hat{\mu}_i = \hat{\mu}_i + \frac{X_i^{(t)} - \hat{\mu}_i}{T_i}$
    end for
    $t = t + 1$
end while

Key lemma

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

$$Reg^{greedy}(T) = O(1)$$

Finite time version: $\forall T \geq 1$

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

- $\delta := f^{-1}(\nabla_{\min}/2)$, $t' := 4c^2/e^2$ 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

$$Reg^{greedy}_{\boldsymbol{\mu},\alpha,\beta}(T) = O(\sqrt{T})$$

Finite time version: $\forall T \geq 1$

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

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

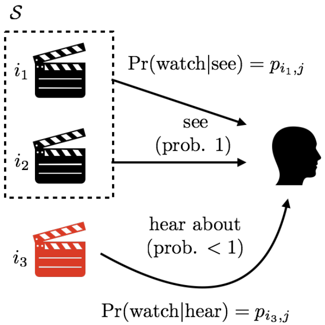
- Holds when the bounded-smoothness function is $f(x) = \gamma x$ where $\gamma > 0$ and $\omega \in (0, 1]$
- Matches with the lower bound in [Wang 17] (tight). Upper bound in [Wang 17] is $\tilde{O}(\sqrt{T})$

Movie recommendation example

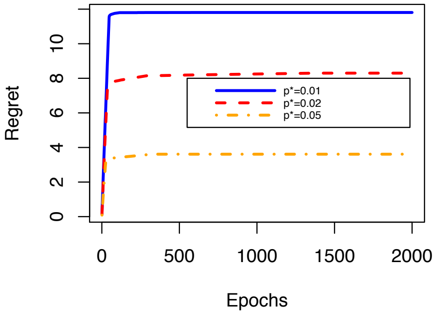
- Movielens dataset
- Weighted bipartite graph $G = (L, R, E, p)$
- L: 50 movies, R: 881 users, E: movie-user pairs
- Action: select k movies
- $p_S^{(i,j)}$: probability that action S triggers edge (i, j)

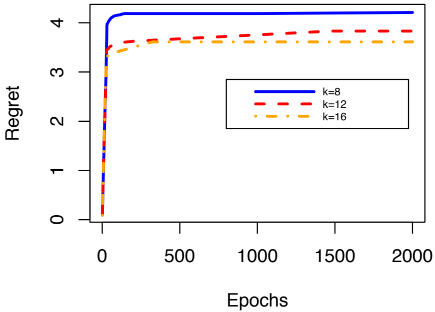
$$p_S^{(i,j)} = 1 \text{ for outgoing edges of nodes in } S$$
$$> p^* > 0 \text{ otherwise [word of mouth]}$$

- $p_{i,j}$: probability user j watches movie i (after he/she learns about the movie)



Movie recommendation example





- 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]

References

[Chen 16] Chen, Wei, et al. "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms." The Journal of Machine Learning Research 17.1 (2016): 1746-1778.

[Chen 16b] Chen, Wei, et al. "Combinatorial multi-armed bandit with general reward functions." Advances in Neural Information Processing Systems. 2016.

[Kveton 15] Kveton, Branislav, et al. "Combinatorial cascading bandits." Advances in Neural Information Processing Systems. 2015.

[Vazirani 01] Vazirani, Vijay V. Approximation algorithms. Springer Science & Business Media, 2001