Cascading Reinforcement Learning

Yihan Du
UIUC

R. Srikant
UIUC

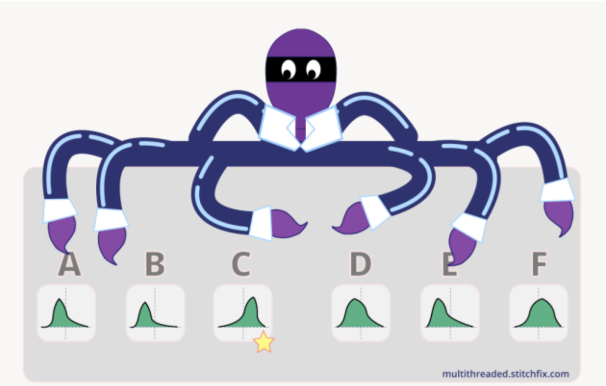
Wei Chen
Microsoft Research

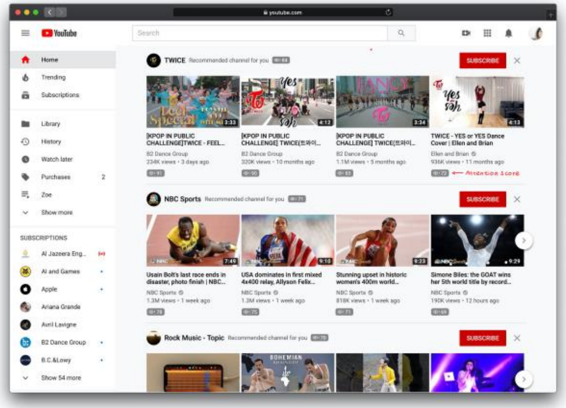
ICLR 2024 (Spotlight)

Contact: yihandu@Illinois.edu

Motivation

- Cascading bandits [Kveton et al., 2015; Combes et al., 2015]
  - Items $a_1, \ldots, a_N$, each with unknown attraction probability $q(a)$
  - Recommend an item list $A_t = (a_1^t, \ldots, a_m^t)$
  - The user clicks the first attractive item → receives a reward
  - Goal: maximize the cumulative reward
- Limitation: ignore the user states (e.g., past behavior) and state transition
- Example – video recommendation:
  - Recommend according to user profiles and viewing records
  - If a user clicks a video, his/her interest (state) may transition
  - Should recommend similar videos as the clicked one

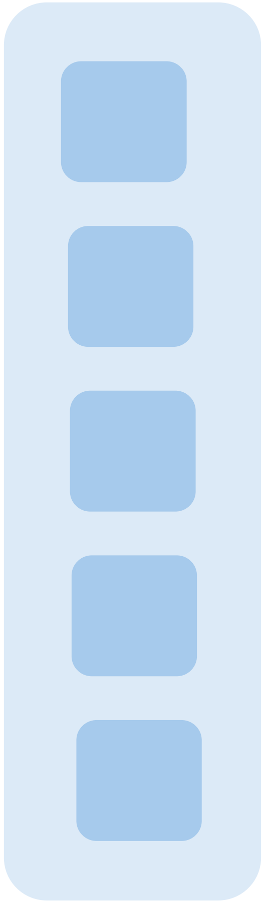




Formulation

- Cascading Markov decision process:
  - s: state
  - $A^{ground} := \{a_1, \ldots, a_N, a_\perp\}$. $a_\perp$: a virtual item denoting that no item in the list is clicked
  - A: a feasible item list, including at most m regular items and $a_\perp$ at the end
  - $\mathcal{A}$: the collection of all feasible A
  - $q(s,a)$: attraction probability. $q(s, a_\perp) = 1$
  - $p(s'|s,a)$: transition probability
  - $r(s,a)$: reward. $r(s, a_\perp) = 0$
- Policy $\pi_h(s)$: specify what item list to select

(s, A)

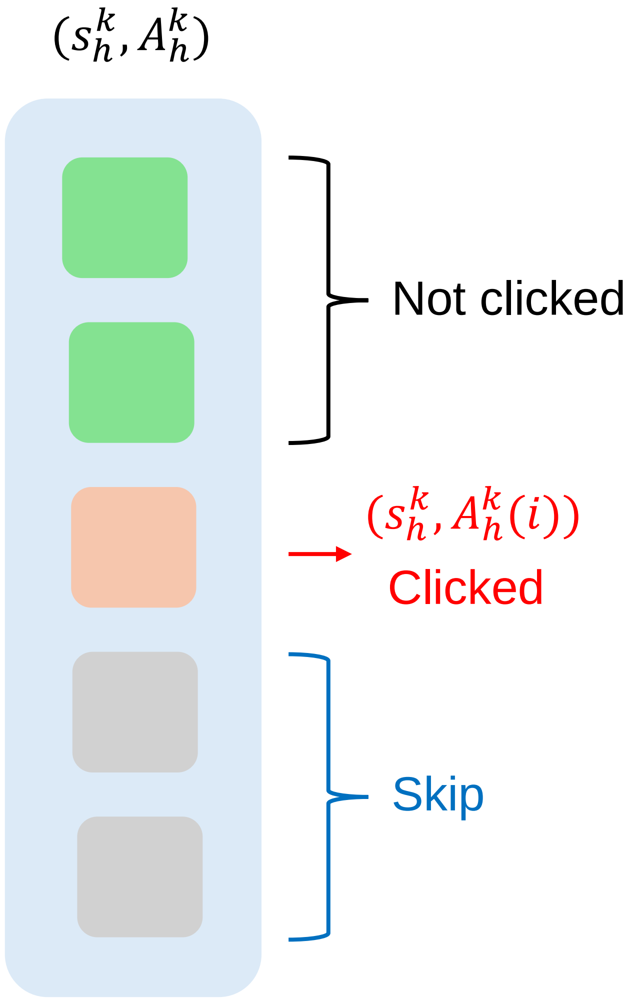


$\rightarrow (s, a)$

- Click w.p. $q(s,a)$
- Receive $r(s,a)$
- Transition according to $p(\cdot|s,a)$

- Cascading Reinforcement Learning (RL):

  In episode k:
  - Choose π^k, and start from $s_1^k \coloneqq s_1$
  - At step h:
    - Observe s_h^k, and select $A_h^k = \pi_h^k(s_h^k)$
    - The user browses the items in A_h^k one by one
    - Once an item $A_h^k(i)$ is clicked:
      - Receive $r(s_h^k, A_h^k(i))$, and transition to $s_{h+1}^k \sim p(\cdot\,|s_h^k, A_h^k(i))$. Skip the following items
    - No item in A_h^k is clicked (i.e., $a_\perp$ is clicked):
      - Receive 0 reward, and transition to $s_{h+1}^k \sim p(\cdot\,|s_h^k, a_\perp)$



- Cascading value functions:

$$\begin{cases} Q_h^\pi(s,A) = \sum_{i=1}^{|A|} \prod_{j=1}^{i-1} \Big(1 - q(s, A(j))\Big) q(s, A(i)) \Big(r(s, A(i)) + p(\cdot|s, A(i))^\top V_{h+1}^\pi\Big) \\ V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)) \\ V_{H+1}^\pi(s) = 0, \quad \forall s \in \mathcal{S}, \end{cases}$$

- Optimal policy π^*: maximize $V_h^\pi(s)$ for all $s \in \mathcal{S}, h \in [H]$
- Regret: $R(K) = \sum_{k=1}^{K} V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)$

Algorithms and Results

- Oracle **BestPerm:**
  - Utilize the property of $V_h^\pi(s)$: sorting items a by descending $r(s,a) + p(\cdot\,|s,a)^\top V_{h+1}^\pi(\cdot)$ gives the optimal item list
  - Find the optimal permutation by a dynamic programming
- Algorithm **CascadingVI:**
  - Employ oracle BestPerm to enable computation efficiency
  - Optimistic value iteration with exploration bonuses for $q(s,a)$ and $p(\cdot\,|s,a)^\top V_{h+1}^\pi(\cdot)$

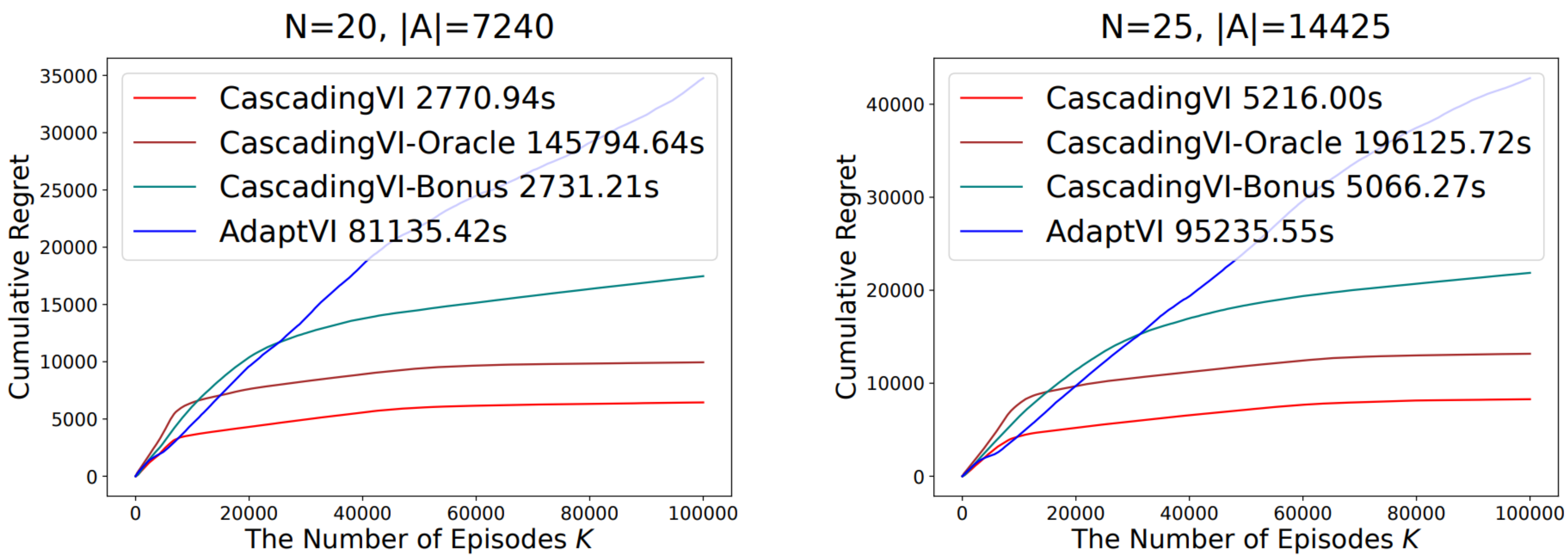
Theorem 1. With probability $1-\delta$, the regret of algorithm CascadingVI is bounded by

$$\tilde{O}(H\sqrt{HSNK})$$

Remark:

- Depend only on N, rather than $|\mathcal{A}| = O(N^m)$
- Match the optimal result in cascading bandits [Vial et al., 2022] (when $S = H = 1$)
- Match the lower bound $\Omega(H\sqrt{SNK})$ for classic RL [Osband & Van Roy, 2016] up to $\sqrt{H}$ (when $q(s,a) = 1$ for all (s,a))

Experiments



- Real-world dataset MovieLens [Harper & Konstan, 2015]

References

1. Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. ICML, 2015.
2. Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. SIGMETRICS, 2015.
3. Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, and R Srikant. Minimax regret for cascading bandits. NeurIPS, 2022.
4. Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. arXiv preprint arXiv:1608.02732, 2016.
5. F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems, 2015.

Image sources:

1. https://multithreaded.stitchfix.com/assets/posts/2020-08-05-bandits/multi_armed_bandit.png
2. https://miro.medium.com/v2/resize:fit:2000/1*u08Kuygehq0gx--2p3wy3A.png

Thank You

Yihan Du
yihandu@Illinois.edu