Reinforcement Learning and Dynamic Optimization

Assignment 1 - Recommending News Articles to Unknown Users

Manoudaki Chressida 2019030201

April 17, 2024

1 Sublinear Regret in Multi-Armed Bandits with User Diversity

1.1 Algorithm Description

In this assignment, a simple modification of the Upper Confidence Bound (UCB) algorithm is proposed to address the challenge of user diversity and achieve sublinear regret. The modification involves maintaining separate exploration counts and estimates of arm rewards for each user segment. By adapting the UCB algorithm to consider user diversity, an attempt is made to improve its performance in scenarios with heterogeneous user preferences.

More specifically, the UCB algorithm was modified to incorporate user diversity by introducing a 2D array to store exploration counts and estimate rewards for each arm and user segment. During arm selection, the algorithm selects the arm with the highest UCB value calculated independently for each user segment. This allows the algorithm to adapt its exploration-exploitation trade-off to the preferences of different user segments.

1.2 Experimental setup

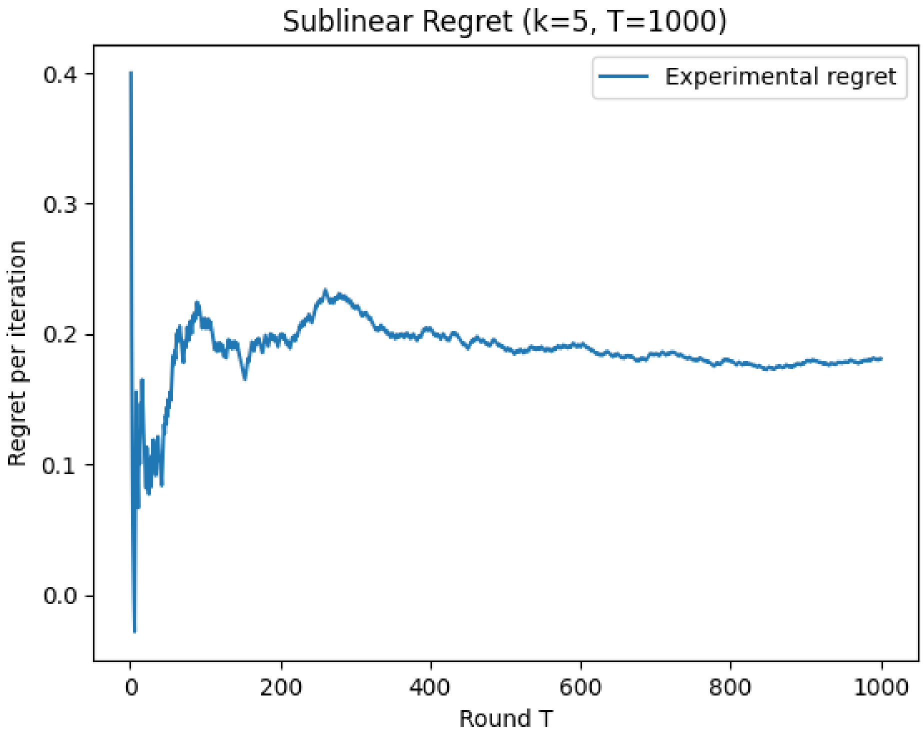
The experiments were conducted using a synthetic MAB environment with 5 arms and 4 user segments. The click probabilities for each arm were randomly generated to simulate user engagement. The proposed algorithm's performance was evaluated for two different horizons (T=1000 and T=10000) to assess its scalability.

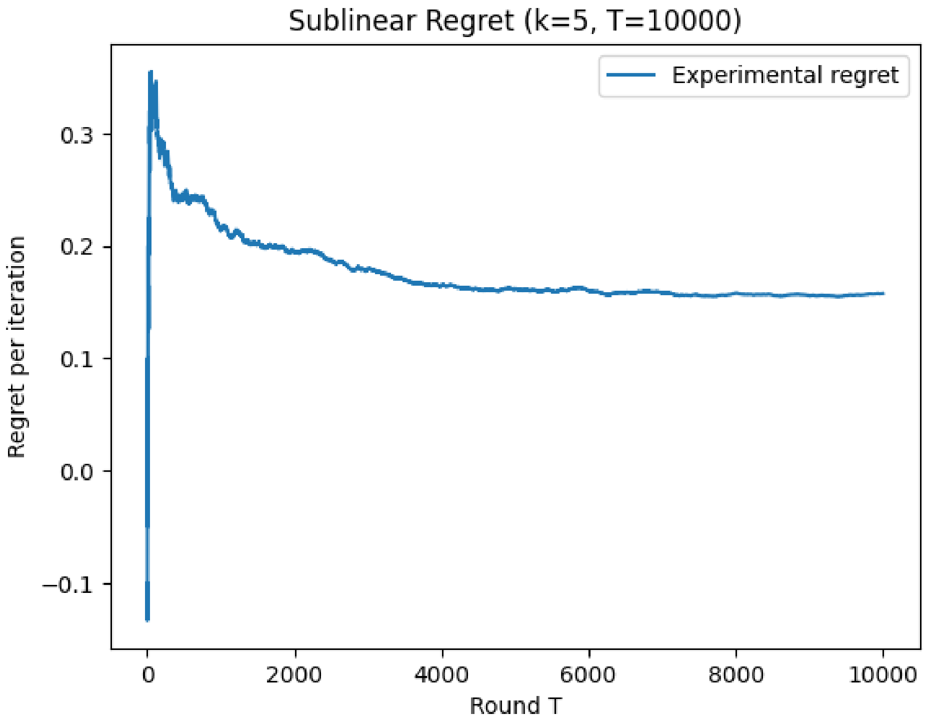
1.3 Plots of the sublinear regret

In the plot for T=1000, the regret increases at a decreasing rate as the number of rounds progresses. Initially, there might be a rapid increase in regret as the algorithm explores different arms, but this growth slows down over time as the algorithm makes better-informed decisions.

In contrast, the plot for T=10000 reveals a more pronounced sublinear trend. While there is still a rapid increase in regret at the beginning, the rate of increase diminishes gradually over time, eventually leading to a sublinear decrease. This suggests that the algorithm's performance benefits significantly from a larger horizon, allowing it more opportunities to balance exploration and exploitation effectively and achieve sublinear regret.

Therefore, while both plots show sublinear regret, the plot for T=10000 suggests that the algorithm achieves a more efficient balance between exploration and exploitation, leading to improved performance over a longer time horizon.





2 Regret derivation

From the lectures we know that these three equations hold true:

- $\mathrm{E}[R(T)] = \sum_{i=1}^{K} N_i(T)\Delta_i$ (1)
- $\Delta_i \leq 2\sqrt{\frac{2log(T)}{N_i(t)}}$ (2), on good event

Where:

- T: the number of total rounds (horizon).
- E$[R(T)]$: the theoretical expected regret of the algorithm.
- Δ_i: the estimated mean value of the arm. It is equal to $\mu^* - \mu_i$.
- $N_i(t)$: the number of times the arm "i" was recommended by the algorithm.

With the combination of equation (1) and (2) we derive the following:

$$\mathrm{E}[R(T)] \leq \sum_{i=1}^{K} N_i \cdot 2\sqrt{\frac{2log(T)}{N_i(t)}} => \mathrm{E}[R(T)] \leq \sum_{i=1}^{K} 2\sqrt{\frac{2log(T)N_i(t)^2}{N_i(t)}}$$

But we know at any round t that $N_i(t) \leq t$, so:

$$\mathrm{E}[R(T)] \leq \sum_{i=1}^{K} 2\sqrt{2log(T)N_i(t)} => \mathrm{E}[R(T)] \leq \sum_{i=1}^{K} 2\sqrt{2 \cdot t \cdot log(T)}$$

Which is the same as:

$$\mathrm{E}[R(T)] \leq K \cdot 2 \cdot \sqrt{2 \cdot t \cdot log(T)}$$

In this modification of the UCB algorithm, the above inequality is the upper bound for the theoretical expected regret of one user type. Thus, it would be more correct to symbolize it like this:

$$\mathrm{E}[R(T)|user] \leq K \cdot 2 \cdot \sqrt{2 \cdot t \cdot log(T)}$$

So, for every user type we get:

$$\mathrm{E}[R(T)] = \mathrm{E}[R(T)|user = 1] + \mathrm{E}[R(T)|user = 2] + \mathrm{E}[R(T)|user = 3] + \mathrm{E}[R(T)|user = 4] \leq \sum_{i=1}^{4} K \cdot 2 \cdot \sqrt{2 \cdot t \cdot log(T)}$$

$$\mathrm{E}[R(T)] \leq 4 \cdot K \cdot 2 \cdot \sqrt{2 \cdot t \cdot log(T)}$$

Generally, assuming U is the number of user types:

$$\mathrm{E}[R(T)] \leq U \cdot K \cdot 2 \cdot \sqrt{2 \cdot t \cdot log(T)}$$

To conclude, for K=5 and U=4, we finally derive the upper bound for the theoretical expected regret to be:

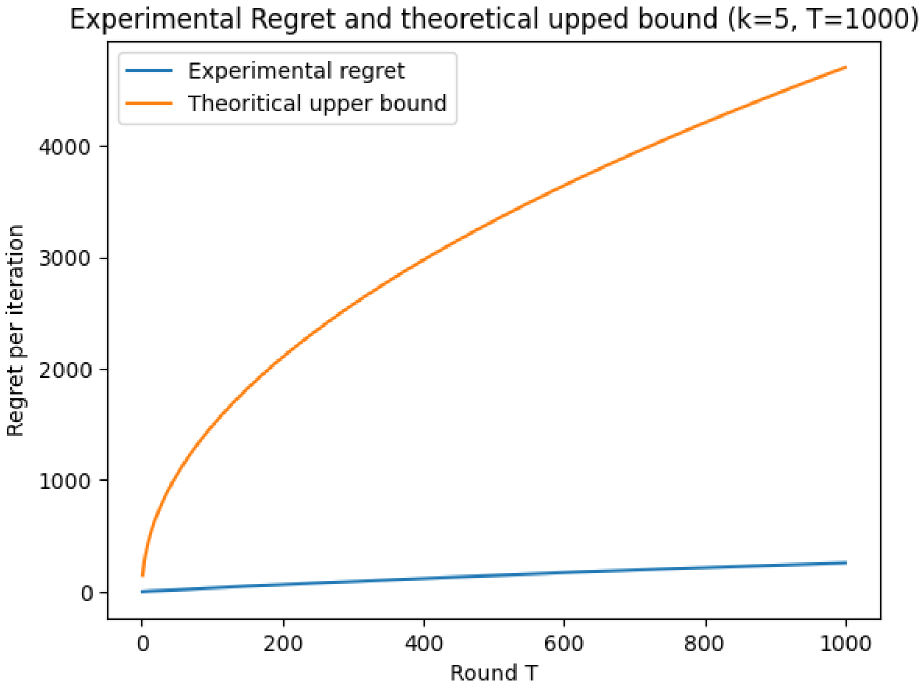
$$\mathrm{E}[R(T)] = 40\sqrt{2 \cdot t \cdot log(T)} => \mathrm{E}[R(T)] = O(\sqrt{T \cdot log(T)})$$

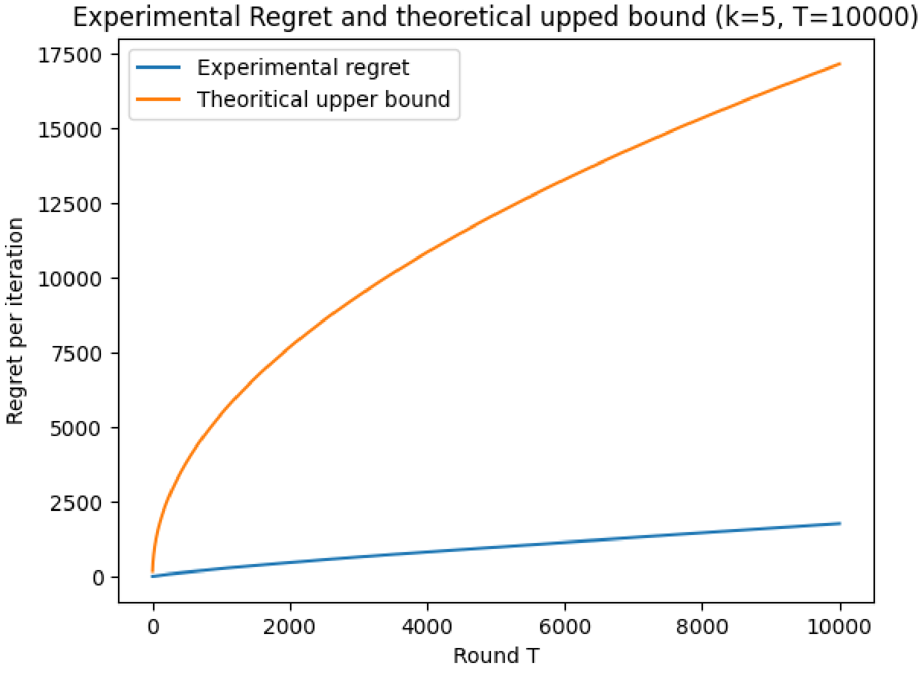
which is sublinear for large T.

3 Comparison of experimental regret with the theoretical upper bound

In the comparison between the experimental regret and the theoretical upper bound, for both T=1000 and T=10000, it was observed that the theoretical upper bound was significantly larger than the experimental regret. This indicates that the algorithm's performance, as measured by the experimental regret, was better than what was theoretically expected.

Furthermore, it's important to note that the experimental regret was plotted using the cumulative regret at each time step. This cumulative regret provides a comprehensive view of the algorithm's performance over time, capturing the accumulated regret up to each round of the experiment. By considering the cumulative regret, we gain insights into the algorithm's behavior and its ability to minimize regret over the entire horizon.


Experimental Regret and theoretical upped bound (k=5, T=1000)
Experimental regret
Theoritical upper bound
Regret per iteration
0
1000
2000
3000
4000
0
200
400
600
800
1000
Round T


Experimental Regret and theoretical upped bound (k=5, T=10000)
Experimental regret
Theoritical upper bound
Regret per iteration
0
2500
5000
7500
10000
12500
15000
17500
0
2000
4000
6000
8000
10000
Round T