UMassAmherst

Manning College of Information & Computer Sciences

MULTI-AGENT MULTI-ARMED BANDIT

Tristan Carel, Ha Pham, Shiyu Zhang (Ordered alphabetically)
PhD Mentor: Fatemeh Ghaffari

INTRODUCTION

The multi-armed bandit problem is a core challenge in decision-making, where an agent must repeatedly choose between various actions, known as "arms," each with its unique reward distribution. The objective is to maximize cumulative rewards over time despite uncertainty about the rewards of each action. Hence, minimizing regrets, calculated as the difference between actual and ideal rewards, is critical to maximizing total rewards in the long run. In the context of multi-agent systems, each agent faces its own bandit problem, compounded by the need to coordinate actions with other agents. Effective communication among agents can potentially enhance performance by sharing information about arm selections and rewards. Our aim is to investigate how different strategies for communicating arms impact overall regret in multi-agent multi-armed bandit scenarios. By exploring the influence of communication strategies on regret, we seek to provide insights into optimizing cooperative decision-making in complex systems.

BACKGROUND

Multi-armed bandit (MAB) [1] olves resource allocation issues across various fields, such as sensor management, manufacturing systems, economics, queueing and communication networks, clinical trials, control theory, and search theory.

In the multi-agent, multi-armed bandit problem, multiple agents with their own sets of arms interact in a shared environment [2, 3]. Each agent operates independently to optimize rewards while also communicating effectively with others to minimize regrets collectively and maximize rewards over time. However, little work has been done to analyze the effect of different communication strategies while making the communication cost the same.

UCB ALGORITHM

The Upper Confidence Bound (UCB) algorithm is a technique used in the context of multi-armed bandit problems, which include:

1. Initialize a class (UCB) representing each arm of the bandit, assigning initial values and counts for each arm.
2. At each step, select the arm with the highest Upper Confidence Bound (UCB) value, balancing between exploiting the arm with the highest estimated reward and exploring other arms.

$$i_t \leftarrow arg\,max_i(Value_i + \sqrt{2 * (\frac{\ln totalCount}{N_i})})$$

3. Update the value and count of the chosen arm based on the observe reward.

$$Value_i \leftarrow \frac{Count_i - 1}{Count_i} + \frac{1}{Count_i} * reward_t$$

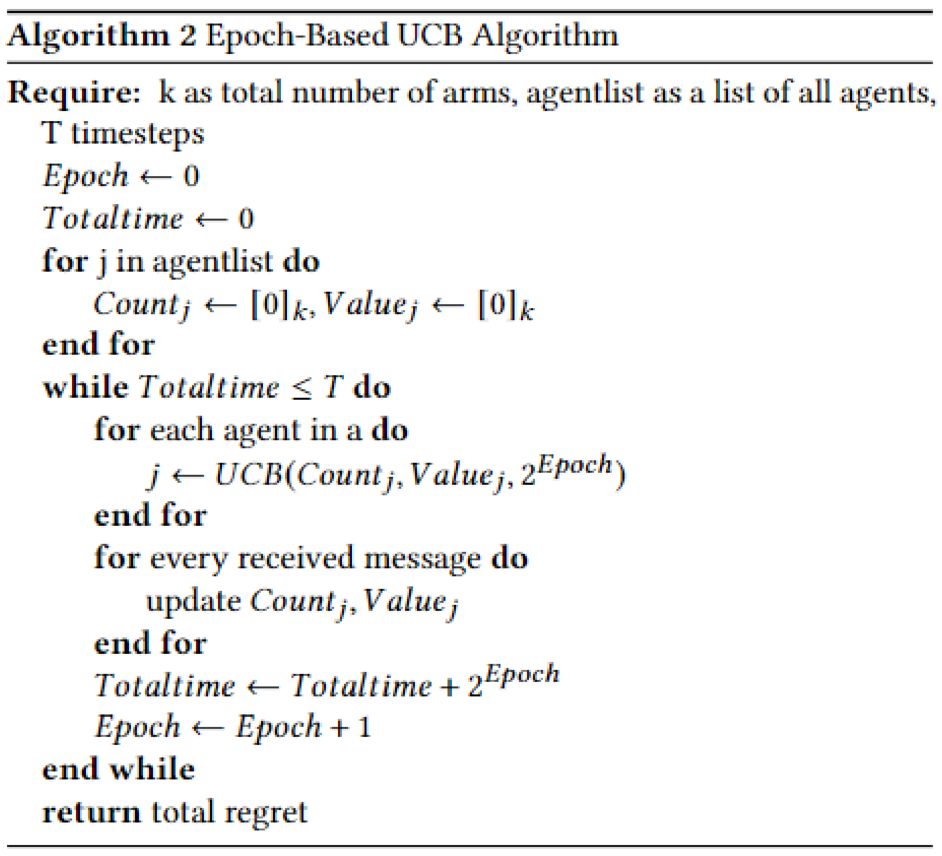
ALGORITHM

This algorithm iteratively solves a multi-armed bandit problem using an epoch-based approach with Upper Confidence Bound (UCB) arm selection:

1. Each epoch involves exploration, exploitation, and communication phases, and UCB is employed to balance exploration and exploitation.
2. During communication, every agent shares the data for only one arm and then averages them together to set the counts and values to update arm values and counts before starting the next epoch.
3. The process continues until the total allotted time is reached.

RESULTS

In our implementation, each arm initially has a mean chosen uniformly between 0 and 1. In every round, a random reward is sampled for each arm from a Bernoulli distribution with the predetermined mean. The system consists of 50 arms and 10 agents, all with access to the entire arm set. Agents communicate at the end of doubling epochs, updating their counts and estimated arm values based on messages received from other agents. In Figure 1, we compare the total regret across three scenarios: no communication, communication of the best epoch arms, communication of the worst epoch arm, and communication of a random arm pulled during the epoch. Our results show that communicating the worst arm has the most beneficial effect on total regret, followed by random selection. Surprisingly, communicating the best arm can worsen regret. This observation suggests that new information may be more helpful for less observed arms. Further theoretical research could provide insights into these experimental findings.



```
Algorithm 2 Epoch-Based UCB Algorithm
Require: k as total number of arms, agentlist as a list of all agents,
  T timesteps
  $Epoch \leftarrow 0$
  $Totaltime \leftarrow 0$
  for j in agentlist do
      $Count_j \leftarrow [0]_k, Value_j \leftarrow [0]_k$
  end for
  while $Totaltime \leq T$ do
      for each agent in a do
          $j \leftarrow UCB(Count_j, Value_j, 2^{Epoch})$
      end for
      for every received message do
          update $Count_j, Value_j$
      end for
      $Totaltime \leftarrow Totaltime + 2^{Epoch}$
      $Epoch \leftarrow Epoch + 1$
  end while
  return total regret
```

OUR PROPOSED ALGORITHM

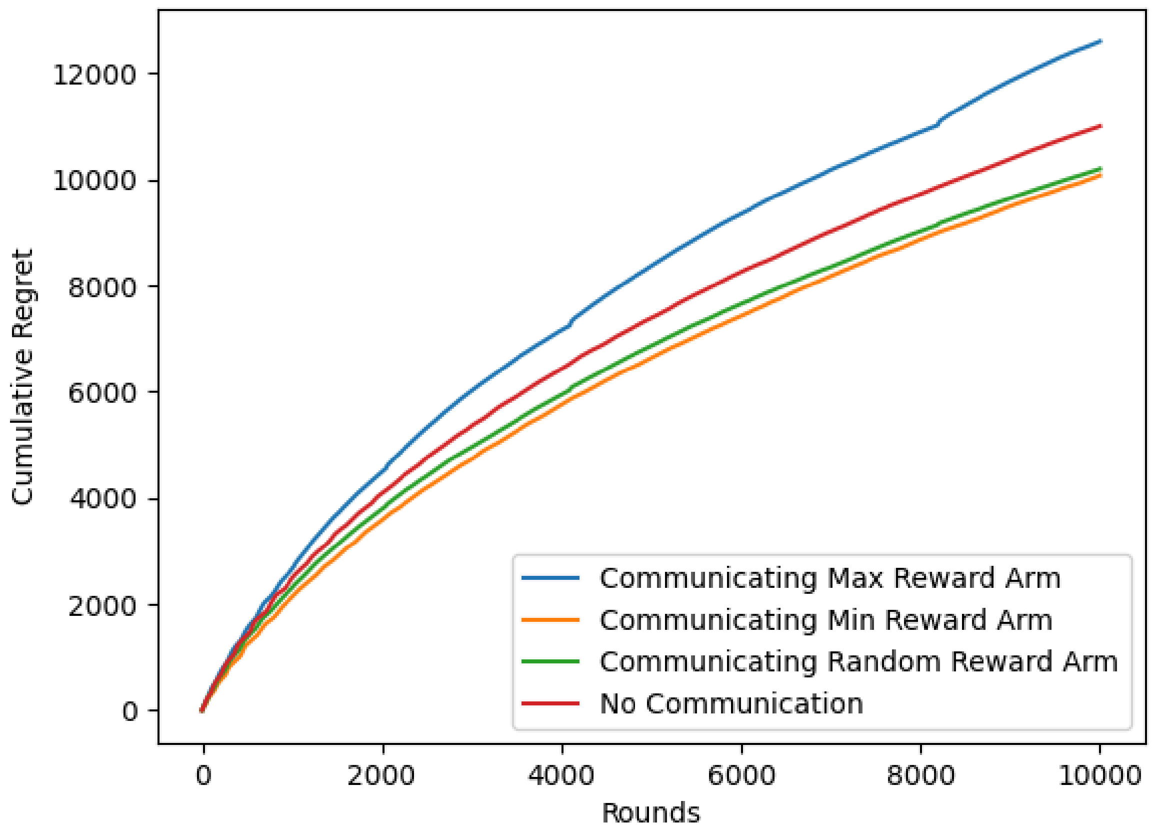


FIGURE 1: COMPARISON OF DIFFERENT COMMUNICATION METHODS

CONCLUSION

The aim of our study was to examine the impact of communicating different arms in multi-agent multi-armed bandits. Our experimental findings reveal that the choice of communicated arm significantly influences total regret, with communication of the worst arm surprisingly yielding better results than communication of the best arm. Additionally, selecting a random arm shows a more favorable effect on total regret compared to communicating the best arm. The lack of theoretical analysis on the effect of arm selection for communication highlights an area for future investigation.

REFERENCE

1. Bubeck, S. and Cesa-Bianchi, N., 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning, 5(1), pp.1-122.
2. Agarwal, M., Aggarwal, V. and Azizzadenesheli, K., 2022. Multi-agent multi-armed bandits with limited communication. The Journal of Machine Learning Research, 23(1), pp.9529-9552.
3. Sankararaman, A., Ganesh, A. and Shakkottai, S., 2019. Social learning in multi agent multi armed bandits. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 3(3), pp.1-35.