Performance Comparison and Analysis of UCB, ETC, and Thompson Sampling Algorithms in the Multi-Armed Bandit Problem

Qijin Hu

Sejtu-Leeds Joint School, South West Jiaotong University and University of Leeds, Chengdu, 611756, China

sc21qh@leeds.ac.uk

Abstract. In this study, the nuanced interplay between problem complexity, prior knowledge, and the exploration-exploitation balance is examined, given their critical impact on algorithmic performance in diverse settings. To gain a deeper understanding of these dynamics, three algorithms – Upper Confidence Bound (UCB), Explore-Then-Commit (ETC), and Thompson Sampling – are selected for a comprehensive investigation. Each of these algorithms presents unique approaches to handling the challenges posed by varying problem contexts. The UCB algorithm, known for its robustness in balancing exploration and exploitation, is scrutinized for its performance in environments with differing levels of complexity and uncertainty. This algorithm's reliance on confidence bounds makes it particularly relevant in scenarios where accurate estimates of uncertainty can significantly enhance decision-making processes. ETC, on the other hand, is characterized by its phased approach, initially exploring options before committing to a seemingly optimal choice. This study examines how the ETC algorithm's performance is influenced by the availability of prior knowledge and the intricacy of the problem at hand. Its phased nature makes it a subject of interest in environments where initial exploration can yield substantial insights for subsequent exploitation.

Keywords: Multi-Armed Bandit; ETC algorithm; UCB algorithm; Thompson Sampling algorithm.

1. Introduction

In everyday life, we are constantly faced with the challenge of making optimal decisions under constraints of limited resources. Whether it's choosing a restaurant, strategizing an advertising campaign, or planning clinical drug trials, these situations encapsulate the broader challenge of decision-making in uncertain environments. At the heart of these decisions lies the dilemma of selecting the best course of action when outcomes are not fully predictable. To better understand and optimize decision-making in such scenarios, the concept of the Multi-Armed Bandit problem comes into play. MAB effectively simulates the process of making sequential choices when resources are limited and information is incomplete. This problem framework mirrors real-life situations where each choice, or 'arm pull', represents a decision with potentially varying rewards. The MAB model provides a structured approach to evaluating the trade-offs between exploring new options and exploiting known ones. In a restaurant selection context, for example, this might mean deciding whether to try a new eatery or return to a favored spot. In advertising, it involves allocating budget across different channels to gauge their effectiveness. Similarly, in drug trials, it entails balancing the exploration of new treatments against the known efficacy of existing ones.

2. Related Theories

2.1. Background

In the Multi-Armed Bandit problem, each "arm" represents a choice, and each selection is equivalent to pulling an arm [1]. The goal is to find the arm with the best reward in as few attempts as possible. This scenario extends beyond restaurant choices to areas like advertising placement and drug experiments. To address such problems, this study focuses on three classic MAB algorithms:

UCB, ETC, and Thompson Sampling, aiming to analyze their performance in different scenarios and provide more effective solutions for practical decision-making problems [2].

2.2. Theoretical Foundation of UCB, ETC, and Thompson Sampling Algorithms

UCB Algorithm: The UCB algorithm is based on the concept of upper confidence bounds, aiming to balance exploration and exploitation. By modeling each arm, the algorithm continuously updates the upper confidence interval for each arm to ensure a more effective balance between exploration and exploitation during the decision-making process [3]. The UCB algorithm more frequently explores unknown arms in the short term and gradually tends to choose arms with higher cumulative rewards in the long term.

ETC Algorithm: The ETC algorithm adopts the Explore-Then-Commit strategy [4]. Unlike the UCB algorithm, it actively explores in the initial phase and then continuously selects the best arm discovered during the subsequent selection phase. The performance of the ETC algorithm is significantly influenced by the length of the exploration phase, requiring a balance between the demands of exploration and exploitation [5].

Thompson Sampling Algorithm: The Thompson Sampling algorithm is based on Bayesian inference, modeling the probability distribution of each arm and sampling based on posterior probabilities to find a balance between exploration and exploitation [6]. The algorithm excels in both short-term and long-term performance, adapting flexibly to different environments through Bayesian inference.

3. Related Work

3.1. History and Development of the Multi-Armed Bandit Problem

The Multi-Armed Bandit problem first appeared in the early 20th century, simulating the decision-making process of slot machines. Over time, it has evolved into a complex probability model used in various scenarios such as optimizing online advertising and designing clinical trials [7]. The research history of this problem witnesses the integration and development of fields such as statistics, information theory, and behavioral economics.

Research Methodology:

Simulated Experiment Setup:

- Choose the MovieLens dataset, considering each movie category as an "arm" and user ratings as "rewards."

Experiment 1.

Select a round number (n = 5000), run ten experiments, record the cumulative regret for each round, and then average the cumulative regret charts. For the ETC algorithm, the exploration phase accounts for 10% of the total length. Set c = 4 in the UCB algorithm, initialize the rating as 5, and alternate the selection of all arms in the first round for the TS algorithm.

Experiment 2.

Modify the round number (n) for the three algorithms to 500, 5000, 50000, 100000, with the exploration phase of the ETC algorithm approximately 10% of each round [8].

Experiment 3.

Set the round number (n) to 50000 and study the ETC algorithm, setting the exploration rounds m*k to 50, 500, 2000, 5000, 100000. Then, conduct one hundred experiments and average the results.

Experiment 4.

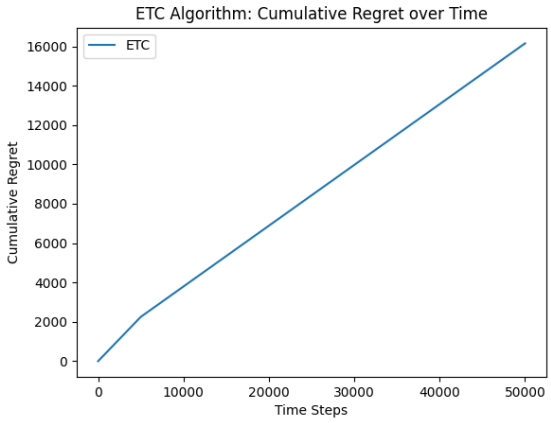
Set the round number (n) to 10000 and study the UCB algorithm, setting c to 4, double square root of 2, 2, init _ prob = 5. Conduct one hundred experiments and average the results.

Experiment 5.

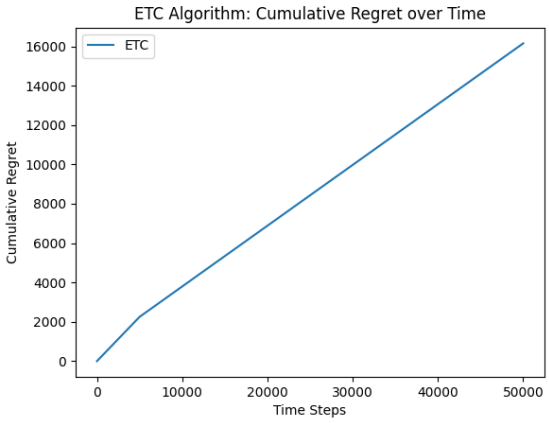
Set the round number (n) to 10000, with parameters similar to Experiment 1. Conduct one hundred experiments and average the results.

4. Data Source

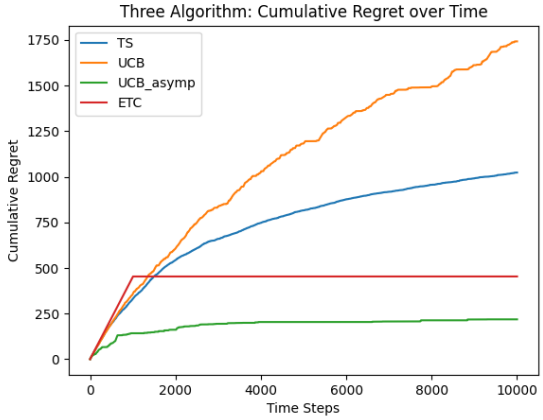
The MovieLens dataset, a rich source of movie ratings, can be accessed from the GroupLens website. This dataset forms the foundation for our experimental design, which is structured around testing three distinct algorithms: the Upper Confidence Bound (UCB), Explore-Then-Commit (ETC), and Thompson Sampling algorithms.



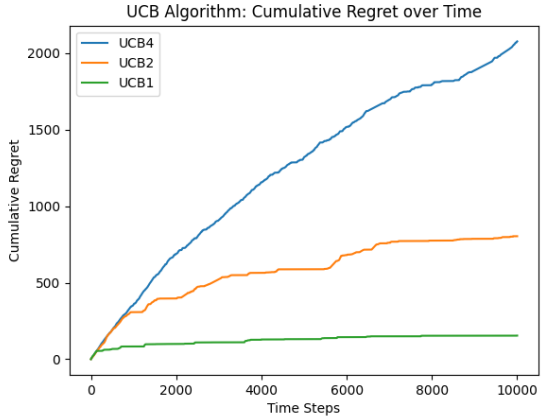
(a)



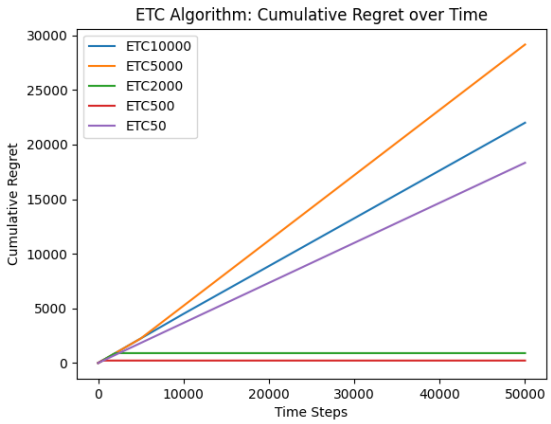
(b)



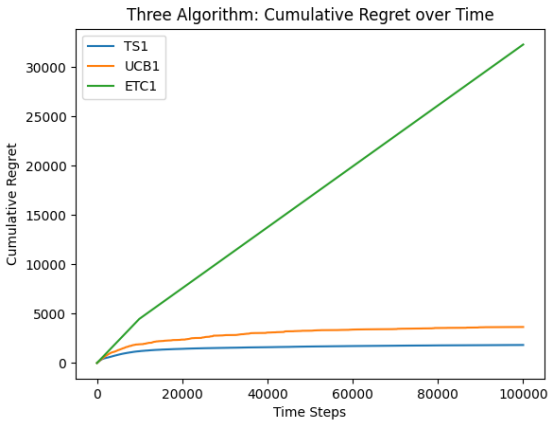
(c)



(d)



(e)



(f)

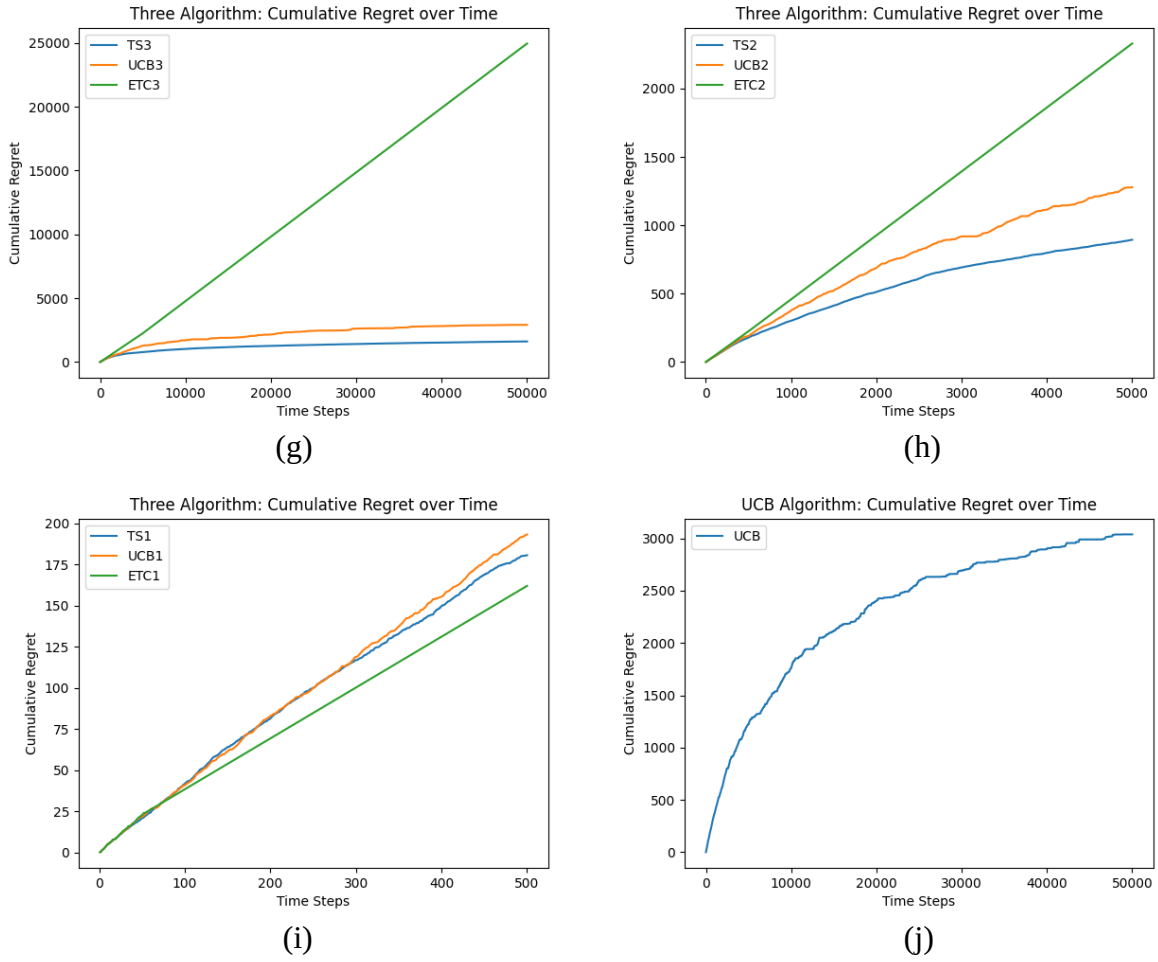


Fig 1. Experimental result

Implementation and Parameter Settings of UCB Algorithm, As shown in Fig 1:

Experiment 1:

The UCB algorithm is implemented, with rounds (n) set at 50,000, a confidence level (c) of 2, and an initial probability (init_prob) of 5 (reflecting the movie ratings range from 1 to 5).

Experiment 2:

For this experiment, rounds (n) vary (500, 5,000, 50,000, 100,000), maintaining c at 2 and init_prob at 5.

Experiment 3:

In this setup, rounds are fixed at 10,000, with varying c values (4, the square root of 2, and 2), and init_prob remains at 5.

Experiment 4:

Here, rounds (n) are set at 10,000, c at 2, and init_prob at 5.

Implementation and Parameter Settings of ETC Algorithm:

Experiment 1:

The ETC algorithm is tested with 50,000 rounds and an exploration phase length (m*k) of 5000.

Experiment 2, As shown in Fig 6:

Varying rounds (500, 5,000, 50,000, 100,000) are tested with different exploration phase lengths (50, 500, 5,000, 10,000).

Experiment 3:

This experiment also uses 50,000 rounds but tests varying exploration phase lengths (50, 500, 2,000, 5,000, 10,000).

Experiment 4 8:

Rounds are set at 10,000, with an exploration phase length of 1,000.

Implementation and Parameter Settings of Thompson Sampling Algorithm:

Experiment 1:

Here, the Thompson Sampling algorithm is tested with 50,000 rounds, initially selecting each "arm" once.

Experiment 2:

The setup includes varying rounds (500, 5,000, 50,000, 100,000), maintaining the initial selection of each "arm" once.

Experiment 3:

Rounds are fixed at 10,000, with the initial selection of each "arm" once.

Cumulative Regret Comparison: Each algorithm's cumulative regret was measured across different experiments. The UCB algorithm showed stable performance but varied in long-term efficacy. The ETC algorithm's performance was heavily influenced by the exploration phase length. Thompson Sampling excelled in both short-term and long-term scenarios.

Comparative Analysis: The UCB algorithm's stability is attributed to its upper confidence bounds strategy, quickly adapting to the environment for correct short-term choices. However, its adaptability issues may impact long-term performance. The ETC algorithm's performance hinges on the exploration phase length, requiring a balance to avoid under or over-exploration. Thompson Sampling's strength lies in its Bayesian inference-based decision-making, allowing flexibility and consistent performance over time..

5. Conclusion

This research thoroughly assesses the efficacy of three algorithms - Upper Confidence Bound, Explore-Then-Commit, and Thompson Sampling - in the context of the Multi-Armed Bandit problem, employing both theoretical analysis and simulated experiments. The findings reveal distinct strengths and ideal application scenarios for each algorithm. The UCB algorithm demonstrates high efficiency in stable environments, proving particularly adept at facilitating swift decision-making processes. Its robustness and simplicity make it a preferred choice in situations where environmental parameters remain consistent over time. In contrast, the ETC algorithm shows its strengths in contexts where a clear demarcation between exploration and exploitation phases is beneficial. This characteristic makes it suitable for scenarios where initial broad exploration is crucial before committing to a specific strategy.

References

[1] Kalvit, A., & Zeevi, A. (2021). A closer look at the worst-case behavior of multi-armed bandit algorithms. Advances in Neural Information Processing Systems, 34, 8807-8819.

[2] Shi, Z., Zhu, L., & Zhu, Y. Thompson Sampling: An Increasingly Significant Solution to the Multi-armed Bandit Problem.

[3] Wang, Y. Thompson Sampling for Multi-armed Bandit Problems: From Theory to Applications.

[4] Zhang, T. H. (2023). Optimistic Thompson sampling: strategic exploration in bandits and reinforcement learning (Doctoral dissertation, University of British Columbia).

[5] Lykouris, T., Tardos, E., & Wali, D. (2020, January). Feedback graph regret bounds for Thompson sampling and UCB. In Algorithmic Learning Theory (pp. 592-614). PMLR.

[6] Zhu, X., Zhao, Z., Wei, X., & others. (2021). Action recognition method based on wavelet transform and neural network in wireless network. In 2021 5th International Conference on Digital Signal Processing (pp. 60-65).

[7] Gupta, S., Chaudhari, S., Mukherjee, S., Joshi, G., & Yağan, O. (2021, June). A unified approach to translate classical bandit algorithms to structured bandits. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3360-3364). IEEE.

[8] Jin, T., Xu, P., Xiao, X., & Anandkumar, A. (2022). Finite-time regret of thompson sampling algorithms for exponential family multi-armed bandits. Advances in Neural Information Processing Systems, 35, 38475-38487.

[9] Zhong, Z., Chueng, W. C., & Tan, V. Y. (2021). Thompson sampling algorithms for cascading bandits. The Journal of Machine Learning Research, 22(1), 9915-9980.

[10] Bayati, M., Hamidi, N., Johari, R., & Khosravi, K. (2020). Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. Advances in Neural Information Processing Systems, 33, 1713-1723.