# Comparative Evaluation of Mean Cumulative Regret in Multi-Armed Bandit Algorithms: ETC, UCB, Asymptotically Optimal UCB, and TS

*Yicong* Lei*

School of Science and School of Engineering, Hong Kong University of Science and Technology, 999077, Hong Kong, China

**Abstract.** This research provides insights into how to address short-term and long-term decision-making in different kinds of the Multi-Armed Bandit (MAB) problem, a classic problem in decision-making under uncertainty. In this study, four algorithms - Explore-Then-Commit (ETC), the Upper Confidence Bound (UCB), Asymptotically Optimal UCB, and Thompson Sampling Algorithms (TS) – are selected to solve the MAB problem with numerical and categorical types. Different types represent different value intervals. Each algorithm is applied to each dataset with two different horizons, which represent the number of iterations, to evaluate its short-term and long-term decision-making ability. All algorithms are then utilized in each dataset to compare which one is most suitable for solving a certain type of MAB problem. This research provides an explicit introduction to the MAB problem and the four algorithms. Furthermore, it concludes that both Asymptotically Optimal UCB and TS are suitable for decision-making in the short and long term. At the same time, Asymptotically Optimal UCB is the most appropriate for the numerical MAB problem, while TS is the most appropriate for the categorical MAB problem. Additionally, UCB only suits short-term decision-making, ETC can be efficient only in the numerical MAB problem.

## 1 Introduction

In real life, many choices arise, and the best decision must be made within limited cognition. There are many uncertain situations where decisions must be made, like medical trials, marketing, and crowdsourcing [1-3]. While these situations may not seem the same, they are similar. To better understand, the essence of all these situations is the exploration-exploitation tradeoff, which leads to the Multi-Armed Bandit Problem. The MAB problem is suitable to describe those cases where the information is incomplete, but sequential decisions must be made, which means each decision made affects what happens next, as one decision follows another in a series [4]. In medical trials, there may be several ways to treat a disease, and a decision is made when one treatment is better than the other. Similarly, in marketing, budgets are assigned in different ways to achieve the lowest cost.

* Corresponding author: yleiaq@connect.ust.hk

This study focuses on evaluating mean cumulative regret among ETC, UCB, Asymptotically Optimal UCB, and TS in solving the MAB problem. Two different types of datasets are used to evaluate the advantages and disadvantages of each algorithm. To be concrete, the same algorithm is compared using a different horizon n in each dataset. All the algorithms are then compared in the same horizon n with their different cumulative regret values in each dataset. Then, this paper presents results and evaluations through the above experiments. Finally, the study summarizes all the findings and implications.

## 2 Relevant theories

### 2.1 MAB problem

The MAB problem depicts a situation where a player chooses from multiple actions over time to maximize cumulative payoffs, where the actions are also called arms and the payoffs are also called rewards. Each arm can provide a stochastic reward, and the player must balance exploration, which means gathering the reward information about each arm, and exploitation, which means maximizing the reward by choosing the best arm each turn from the calculation. Naturally, there is a dilemma between exploration and exploitation because exploring more to improve future decisions and exploiting the optimal arm to maximize cumulative reward are both desired. In the MAB problem, 'regret' is defined as the gap between the cumulative reward from all chosen arms and the largest cumulative reward if which arm is the best is known in advance. Therefore, the goal can also be to minimize the cumulative regret.

### 2.2 ETC, UCB, Asymptotically Optimal UCB, and TS

ETC Algorithm: ETC is one of the simplest approaches to solving the MAB problem. The algorithm can be divided into two phases. In the exploration phase, each arm is explored a fixed number of times to obtain the average reward for each arm. In the commitment phase, the player chooses the arm with the highest estimated reward till the end [5].

UCB Algorithm: UCB is a powerful algorithm that balances exploration and exploitation through uncertainty. Each arm calculates a confidence bound on the reward and chooses the arm with the highest upper bound [6]. As the name says, the upper confidence bound is chosen to be the index.

The initial version of UCB is UCB1, which is defined as Formula (1).

$$UCB_i(t) = \mu_i(t) + \sqrt{\frac{4logn}{N_i(t)}} \quad (1)$$

Where $\mu_i(t)$ refers to the average reward from arm i till round t, and $N_i(t)$ refers to the number of samples from arm i.

Then, select the arm: $A_t = arg_i max UCB_i(t-1)$.

Asymptotically Optimal UCB Algorithm: This variant of UCB is similar to the original UCB. It only adjusts the exploration term to achieve minimal regret over time. This approach is particularly designed for long-term decision-making since the selection of arms is more flexible and even those arms that have not been selected for a long time may be chosen [7]. The index is defined as Formula (2).

$$UCB_i(t) = \mu_i(t) + \sqrt{\frac{2logf(t)}{N_i(t)}} \quad (2)$$

Where $f(t) = 1 + t(logt)^2$, but in this paper, $f(t) = t$ is used for simplicity. The way to select arms is the same as UCB1.

TS Algorithm: TS is a Bayesian approach to solving the MAB problem. The reward of each arm can be assumed to be normally distributed. At each step, the algorithm samples from these distributions, and the arm with the highest reward is chosen [8]. This approach can be adapted to changes in the environment by continuously updating the posterior distribution. The selection of arms is defined as Formula (3).

$$\theta_i(t) \sim F_i(t) \tag{3}$$

Where '∼' refers to generating a random value that comes from the distribution $F_i(t)$. Then, select the arm $A_t = arg_i max \theta_i(t-1)$. $F_i(t)$ is defined as Formula (4).

$$F_i(t) \sim N(\mu_i(t) + \frac{1}{N_i(t)}) \tag{4}$$

Where $\mu_i(t), N_i(t)$ are the same as defined in UCB1. $\tau 2$ is replaced by 1 because $\tau = 1$ is chosen when the reward distributions are assumed to be 1-subGaussian.

## 3 Experiments and results

### 3.1 Data source

The experiment is based on two datasets. These two datasets have different types. The first one is numerical while the second one is categorical [9,10].

The first one is from the GroupLens website. This dataset has a rich set of movie ratings with 18 genres as the bandit arms and rewards from 1 to 5.

The second one is from the Open Bandit Dataset [11]. The dataset is constructed during A/B testing of two multi-armed bandit strategies on ZOZOTOWN, a large fashion e-commerce platform. It consists of approximately 26 million rows, with each row, representing a user impression that includes feature values, selected items as actions, true propensity scores, and click metrics as outcomes.

These datasets are the basis for experimental design, which is centred around testing four different algorithms: ETC, UCB, Asymptotically Optimal UCB, and TS.

### 3.2 Experiment: the first dataset

In the first dataset, each movie genre is an arm, and user ratings are considered as the reward. The regret per round is the difference between the actual rating of the selected genre and the best genre with the highest average rating.

#### *3.2.1 Experiment 1: comparison of the same algorithm under different horizons in the first dataset*

In this experiment, each algorithm is used for the first dataset with the horizon set to be 1000 and 10000. The performances of each algorithm are compared under these two horizons. Cumulative regret is used to evaluate the performance of each algorithm. The reason for this design is to evaluate how the performance of the algorithm changes as the decision-making horizon varies. To be concrete, it can reveal how well the algorithm adapts and improves

over time and show how well the algorithm performs in less stable or more volatile conditions in numerical data.

The results of Experiment 1 contain Figures 1-8.

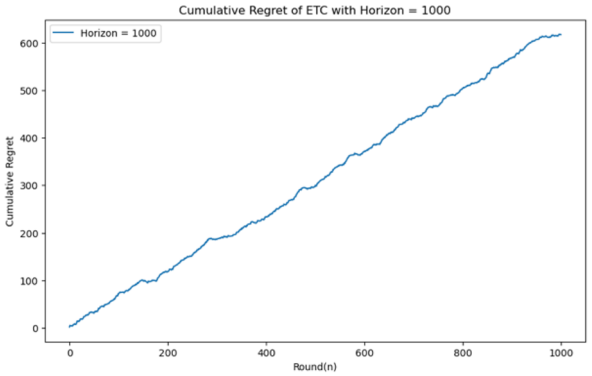


**Fig. 1.** Cumulative regret of ETC with Horizon=1000 in the first dataset

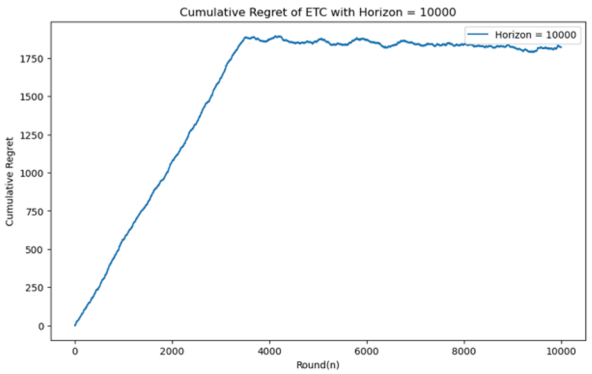


**Fig. 2.** Cumulative regret of ETC with Horizon=10000 in the first dataset

For ETC, the cumulative regret is high when the horizon is small, but it is low to some extent when the horizon is large, which implies that ETC has a good performance when the horizon is large. Particularly, for stable numerical data such as movie ratings, ETC can adapt and improve well over time.

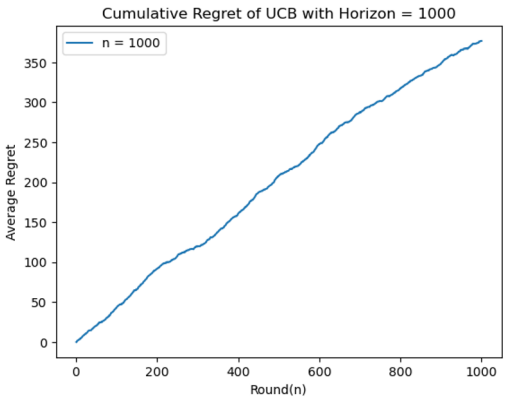


**Fig. 3.** Cumulative regret of UCB with Horizon=1000 in the first dataset

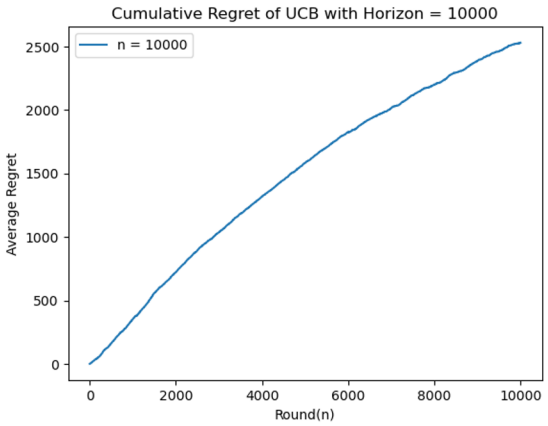


**Fig. 4.** Cumulative regret of UCB with Horizon=10000 in the first dataset

For UCB, the cumulative regret is low when the horizon is small, but it is high when the horizon is large, which implies that UCB has a good performance when the horizon is small. In other words, UCB is quite flexible and more suitable for less stable or more volatile conditions in numerical data.

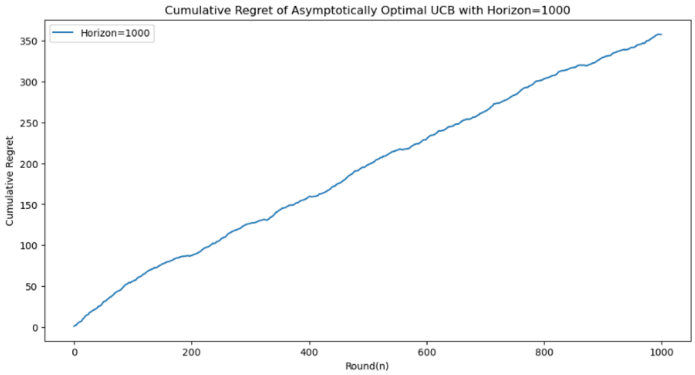


**Fig. 5.** Cumulative regret of Asymptotically Optimal UCB with Horizon=1000 in the first dataset

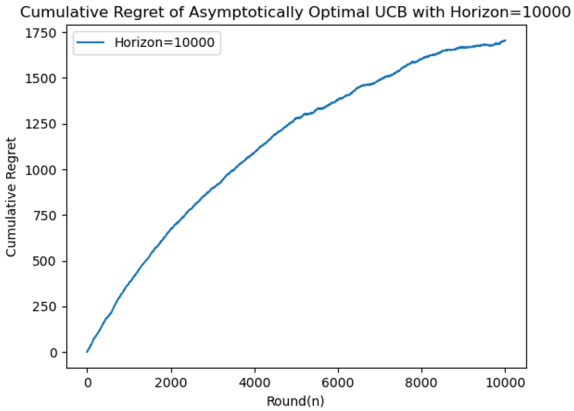


**Fig. 6.** Cumulative regret of Asymptotically Optimal UCB with Horizon=10000 in the first dataset

As for Asymptotically Optimal UCB, the cumulative regret is low when the horizon is small, and it is also low to some extent when the horizon is large. The outcomes show that Asymptotically Optimal UCB is not only flexible but also performs well in long-term performance.

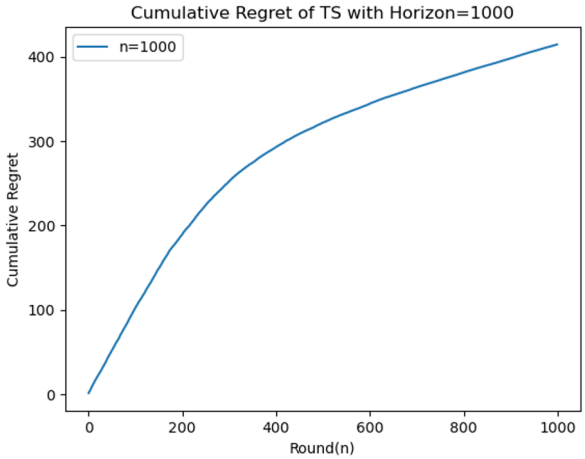


**Fig. 7.** Cumulative regret of TS with Horizon=1000 in the first dataset

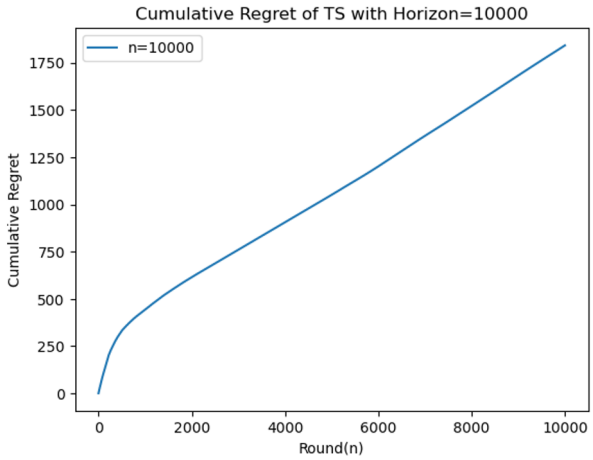


**Fig. 8.** Cumulative regret of TS with Horizon=10000 in the first dataset

In the case of TS, the cumulative regret is relatively low when the horizon is small, and it is also low when the horizon is large. The results indicate that TS is also comprehensive. It has a good performance both in stable and dynamic conditions of numerical data.

### *3.2.2 Experiment 2: comparison of all algorithms under the same horizon in the first dataset*

In this experiment, all algorithms are compared in the first dataset with the horizon set to be 1000 and 10000 respectively. Cumulative regret is used to evaluate the performance of all algorithms in one picture. The reason for this design is to evaluate the performance of different algorithms when they are exposed to the same number of iterations. More specifically, it can highlight which algorithm is more stable or reliable and help to identify the most suitable approach to numerical data.

The results of Experiment 2 are as follows.

Comparing Figures 1, 3, 5, and 7, the results show for less stable or more volatile conditions of numerical data, UCB and Asymptotically Optimal UCB perform best.

Comparing Figures 2, 4, 6, and 8, the results show for stable and persistent conditions of numerical data, ETC, Asymptotically Optimal UCB and TS perform best.

Thus, the results show that Asymptotically Optimal UCB is more suitable for numerical data compared to the other three algorithms.

## 3.3 Experiment: the second dataset

In the second dataset, each item_id is an arm and the click is selected as a binary reward. The regret per round is the difference between the click probability of the chosen arm and the click probability of the theoretically best arm.

### *3.3.1 Experiment 3: comparison of the same algorithm under different horizons in the second dataset*

Similar to 3.2.1. (Experiment 1), each algorithm is compared in the second dataset with the horizon set to be 1000 and 10000. Cumulative regret is used to evaluate the performance of each algorithm. The purpose of this experiment is to evaluate how effectively the algorithm adjusts and evolves and demonstrate its performance in more dynamic or unpredictable conditions with categorical data.

The results of Experiment 3 are in Figures 9-16.

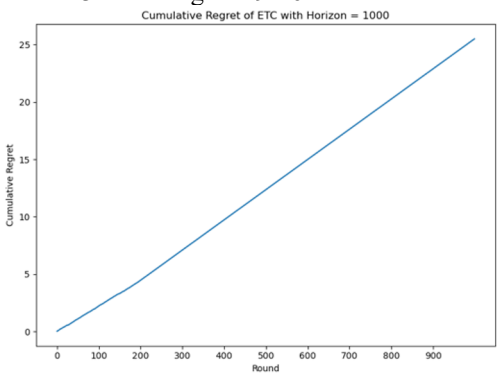


**Fig. 9.** Cumulative regret of ETC with Horizon=1000 in the second dataset

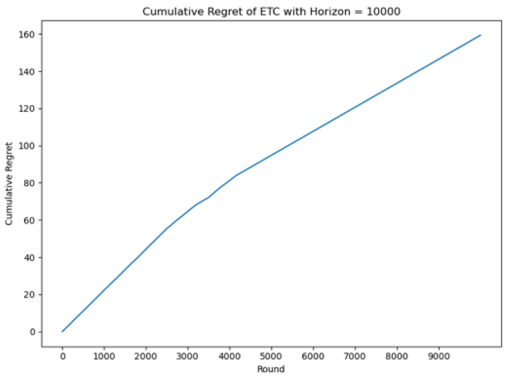


**Fig. 10.** Cumulative regret of ETC with Horizon=10000 in the second dataset

For ETC, the cumulative regret is high when the horizon is small, and it is also high when the horizon is large, which means that ETC does not perform well regardless of whether the horizon is small or large. Therefore, for categorical data, ETC can not adapt and evolve well over time.

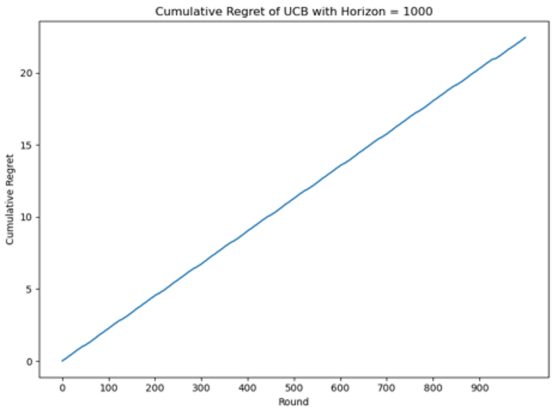


**Fig. 11.** Cumulative regret of UCB with Horizon=1000 in the second dataset

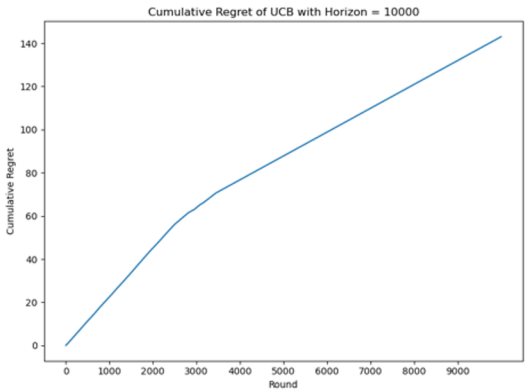


**Fig. 12.** Cumulative regret of UCB with Horizon=10000 in the second dataset

For UCB, the cumulative regret value is relatively low when the horizon is small, but relatively high when the horizon is large, which means that UCB performs better only when the horizon is small. Specifically, UCB is adaptable and more appropriate for the unpredictable or changing conditions of categorical data.

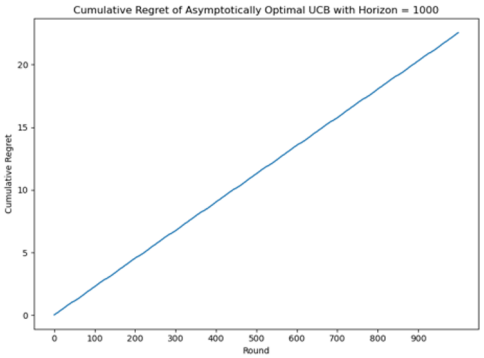


**Fig. 13.** Cumulative regret of Asymptotically Optimal UCB with Horizon=1000 in the second dataset

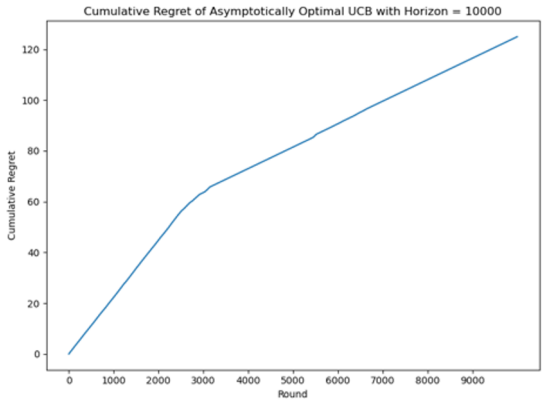


**Fig. 14.** Cumulative regret of Asymptotically Optimal UCB with Horizon=10000 in the second dataset

As for Asymptotically Optimal UCB, the cumulative regret is low when the horizon is small and when the horizon is large. The outcomes show that Asymptotically Optimal UCB is suitable for short-term decisions and also performs well in the long term for categorical data.

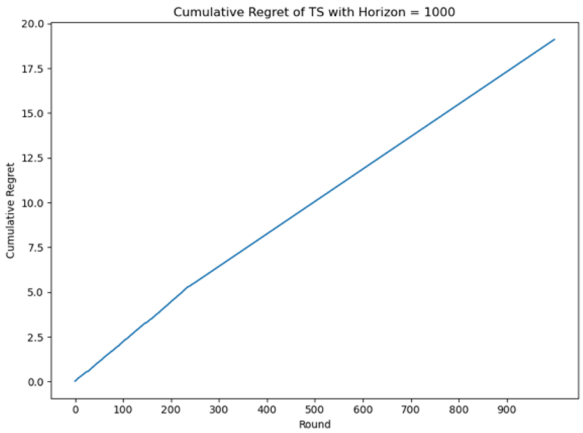


**Fig. 15.** Cumulative regret of TS with Horizon=1000 in the second dataset

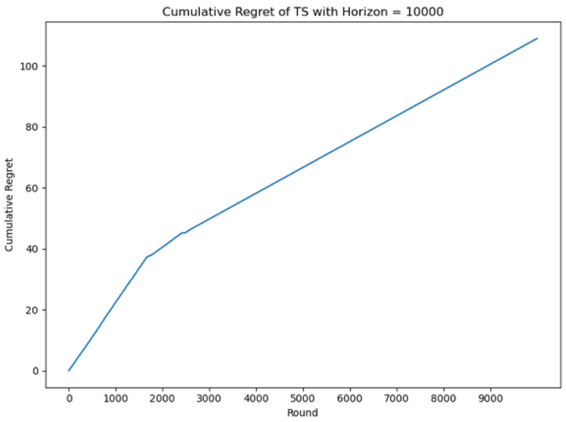


**Fig. 16.** Cumulative regret of TS with Horizon=10000 in the second dataset

In the case of TS, cumulative regret is quite low, regardless of the size of the horizon. The results show that the TS is very comprehensive. It performs very well in any condition of categorical data.

### *3.3.2 Experiment 4: comparison of all algorithms under the same horizon in the second dataset*

Similar to 3.2.2. (Experiment 2), all algorithms are compared in the second dataset with the horizon set to be 1000 and 10000 respectively. Cumulative regret is used to evaluate the performance of all algorithms in one picture. The purpose of this experiment is to identify which algorithm is more suitable and optimal and assist in selecting the most appropriate algorithm for handling categorical data.

The results of Experiment 4 are as follows.

Comparing Figures 9, 11, 13, and 15, the results show for unstable or fluctuating conditions of categorical data, UCB, Asymptotically Optimal UCB and TS perform best.

Comparing Figures 10, 12, 14, and 16, the results show for continuous and steady conditions of categorical data, Asymptotically Optimal UCB and TS perform best.

Thus, the results show that TS is more appropriate for categorical data compared to the other three algorithms.

## 4 Conclusion

In conclusion, this research utterly evaluates the efficiency of four different algorithms: ETC, UCB, the Asymptotically Optimal UCB, and TS when they are used in the Multi-Armed Bandit problem. The formulas of these four algorithms are defined and all algorithms are applied to two different types of datasets. In each dataset, the same algorithm is compared with different horizons, and different algorithms are compared on the same horizon. The findings reveal that for numerical data, Asymptotically Optimal UCB performs best regardless of the size of the horizon and for categorical data, TS performs most efficiently irrespective of the horizon size. In the experimental numerical dataset, the values between each data point are relatively large. However, in the experimental categorical dataset, the values between each data point are relatively small. Therefore, the results can also demonstrate that Asymptotically Optimal UCB is more suitable for processing data with large value intervals, while TS is more appropriate for processing data with small value intervals. Nevertheless, both algorithms suit short-term and long-term decision-making. Besides, ETC is relatively good at handling numerical data with large value intervals and UCB is only good at handling short-term decisions regardless of the type of data. However, since only two datasets are used in this research, the conclusions may have some limitations. With the use of more different datasets to evaluate the efficiency of these four algorithms, more comprehensive conclusions can be reached.

## References

1. R. Degenne, V. Perchet, Anytime optimal algorithms in stochastic multi-armed bandits. In Int. Conf. Mach. Learn., pp. 1587-1595, PMLR (2016).
2. S. Vaswani, B. Kveton, Z. Wen, M. Ghavamzadeh, L.V.S. Lakshmanan, M. Schmidt, Model-independent online learning for influence maximization. In Int. Conf. Mach. Learn., pp. 3530-3539, PMLR (2017).
3. Y. Zhou, X. Chen, J. Li, Optimal PAC multiple arm identification with applications to crowdsourcing. In Int. Conf. Mach. Learn., pp. 217-225, PMLR (2014).

4. T. Lattimore, C. Szepesvári, Bandit algorithms, Cambridge University Press (2020).
5. Q. Hu, Performance comparison and analysis of UCB, ETC, and Thompson sampling algorithms in the multi-armed bandit problem. Highlights Sci. Eng. Technol. 94, 273-278 (2024).
6. A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, P. Auer, Upper-confidence-bound algorithms for active learning in multi-armed bandits. In Int. Conf. Algorithmic Learn. Theory, pp. 189-203, Springer, Berlin, Heidelberg (2011).
7. C. Qu, Enhancing UCB-tuned and asymptotically optimal UCB algorithms through weighted average techniques in multi-armed bandit scenarios. Highlights Sci. Eng. Technol. 94, 187-194 (2024).
8. D.J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, A tutorial on Thompson sampling. Found. Trends Mach. Learn. 11, 1-96 (2018).
9. F.B. Hildebrand, Introduction to numerical analysis, Courier Corporation (1987).
10. A. Agresti, Categorical data analysis, John Wiley & Sons (2012).
11. Y. Saito, S. Aihara, M. Matsutani, Y. Narita, Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. arXiv preprint arXiv:2008.07146 (2020).