A Mechanism-Level Comparison of ETC, UCB, and Thompson Sampling Under Finite-Horizon Constraints

Yichen Fan

{yichen.fan1@student.unimelb.edu.au}

Faculty of Science, The University of Melbourne, Parkville, VIC 3010, Australia

Abstract. Multi-armed bandit (MAB) algorithms are commonly used in sequential decision tasks such as online recommendation and advertising. In real-world systems, algorithms often have limited time and data to learn from, and poor decisions can be costly. From a finite-horizon perspective, this paper presents a mechanism-level comparison of Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), and Thompson Sampling (TS). The findings suggest ETC is highly sensitive to the length of the exploration phase and does not allow parameter adjustment once it enters the commit stage. UCB tends to incur relatively high exploration costs in short horizons and is also sensitive to parameter settings. In contrast, Thompson Sampling usually exhibits smoother behavior in the early stages and more stable performance under finite horizons, with less reliance on parameter tuning, although some randomness across runs remains. Overall, algorithm selection in real-world systems should consider early-stage behavior and risk under finite-horizon constraints.

Keywords: Multi-armed bandits, Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), Thompson Sampling

1 Introduction

In many real-world decision problems, systems must make a sequence of decisions with limited information. For instance, in online advertising and personalized recommendation systems, decisions are made repeatedly, but the feedback is often partial and indirect. Each action not only brings an immediate reward but also affects what the system can learn for future decisions. As a result, a key challenge in these problems is how to design a decision strategy that balances exploring unknown options and exploiting options that currently seem better, in order to achieve good long-term performance [1, 2].

Most existing studies on the exploration-exploitation trade-off focus on theoretical settings, while real-world systems face many practical constraints. For example, decision processes may be limited by short time horizons, small user populations, or high costs when making incorrect decisions. This is common in online advertising, recommendation systems, and medical

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365150

decision-making. Thus, under these conditions, how an algorithm explores in the early stages can strongly affect overall system performance. Relying only on the classical asymptotic optimality framework and regret analysis proposed by Lai and Robbins [3] is often not sufficient to reflect how algorithms perform within practical time limits. In real deployments, useful performance is required within a limited time window, which has led to growing interest in finite-time and non-asymptotic analyses of bandit algorithms [2.4].

The multi-armed bandit (MAB) framework is a widely used model for studying the exploration–exploitation trade-off in sequential decision-making. Many algorithms have been proposed under this framework, among which Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), and Thompson Sampling (TS) are three representative approaches with fundamentally different exploration mechanisms. ETC separates exploration and exploitation into two stages, UCB performs adaptive exploration based on optimism under uncertainty, and TS introduces randomized decisions through probability matching [1, 5, 6]. These structural differences suggest that the algorithms may behave quite differently when data are limited or during cold-start periods.

The currently available research on bandit algorithms concentrates on either asymptotic optimality or empirical performance in a particular application context. In comparison, less research adopts an integrated perspective to explore the influence of various exploration systems on the behavior of the algorithm, in finite time conditions and the real world. Out of this gap, the present paper compares ETC and UCB and Thompson sampling based on the view of their exploration mechanisms, where the emphasis is on the behavior of the first stage when used with limited samples.

2.1 Background & exploration mechanisms

Multi-armed bandit problems involve having no idea about the rewards of the armed bandits, hence decisions must be made under uncertainty state. Despite the fact that this framework has been extensively theorized, in practice algorithms are limited by small user bases and limited time horizons, and it is hard to gather large volumes of data over extended time.

When this is the case, the initial behaviour of the algorithm tends to be much more influential on the final results. In finite-horizon models, when reward distributions are not known extensively, the bandit algorithms will tend to explore intensively to the start. Consequently, much of the cumulative regret has to be done before there are sufficient observations to confidently tell the difference between various options. This is why even the algorithm performance of practical time scales can vary significantly compared to what asymptotic analysis would predict, which demonstrates the significance of a finite-time view.

Considering these factors, this paper reviews the conduct of various bandit algorithms in the exploration, considering limited samples. And this paper examines three representative strategies, ETC, UCB, and TS, which embrace very dissimilar exploration mechanisms. This paper take an alternate approach to emphasize the way such mechanisms develop early-stage behavior and influence performance in finite-horizon decision-making problems [7].

2.2 Explore-Then-Commit: phased irreversible exploration strategy

The ETC algorithm refers to a kind of phased exploration approach where the decision process is segmented into two phases: exploration and exploitation. In the exploration phase, every arm is sampled by the algorithm based on a predefined rule. Once exploration is done, ETC selects the best arm and follows it throughout the horizon.

This two-stage framework makes ETC relatively simple to carry out, as well as aligns with the real-world experimental method of the test-then-deploy model of A/B testing and clinical tests [8]. But there are also limitations to this simplicity. When a decision is arrived at, the conclusion of the exploration phase, it is not possible to amend or reform it. In case the optimum arm is falsely identified because of limited samples or random noise, the algorithm will persist in utilising a suboptimal arm over the persevering horizon, which will result in a high cumulative regret.

This irreversibility is of special interest when viewed in terms of a finite horizon perspective. When the exploration phase is too brief, ETC will have committed herself too early to an erroneous choice; when it is too lengthy, the exploration itself will become expensive. The choice of the exploration length, in its turn, directly influences the performance of ETC, and inappropriate choices may result in significantly worse outcomes [6]. Most importantly, under conditions of small sample size or great expenditures of exploration, ETC may prove to be a hazardous option.

2.3 Upper Confidence Bound: Uncertainty-driven adaptive exploration

The UCB algorithm explores by considering uncertainty. At each round, it computes an upper confidence bound for each arm by combining the current reward estimate with an uncertainty term and then selects the arm with the highest score. Unlike ETC, UCB does not separate exploration and exploitation into two fixed phases, but balances the two throughout the decision process, following the principle of optimism under uncertainty proposed by Auer et al. [1].

The main idea of UCB is to allocate more trials to arms with higher uncertainty. Arms that have been selected fewer times have larger uncertainty terms and are therefore more likely to be chosen, while arms with more observations gradually receive less exploration and are selected mainly based on their observed rewards. By so doing, the algorithm has a tendency to explore more with limited information and gradually turns to exploitation as the information obtained increases to enable it to rectify initial suboptimal decisions.

But on limited horizons, UCB is over-exploratory. The first frames of the algorithm in the initial phases, with uncertainty prevailing across the board, can mean that the algorithm switches dynamically among alternative arms, resulting in high costs in exploration. Moreover, the performance of UCB in the finite time is extremely sensitive to the way in which the exploration parameter is chosen, and various decisions may bring quite different results. Empirical and theoretical results affirm that, though UCB achieves effective asymptotic guarantees, it performs poorly when using limited data due to the sensitivity of the exploration component [9].

2.4 Thompson Sampling: randomized implicit exploration

TS has a different method of exploration as compared to UCB, which involves random choices. The algorithm operates on the probability distribution of each arm's reward, and at a given step draws a random number between the distributions and chooses the arm with the largest sampled number. In contrast to UCB, TS introduces uncertainty into the action-selection procedure by random sampling, which is sometimes called probability matching.

This random approach results in an exploration behavior that is less painful. The probability of arms with less uncertainty to be selected is higher due to the greater variation in the sampled values of the arms, and probability is higher that they are large. The more information gathered by the algorithm, the less uncertainty, and decisions are gradually changed in favor of those arms that are currently more effective. By so doing, TS will not go overboard during exploration and yet consider any good arms. This moderated conduct has been witnessed several times in past empirical research [5].

In practical terms, TS can be more stable in execution with a lower cumulative regret at an early learning stage than deterministic exploration strategies. Concurrently, due to the use of random sampling, results may be different on the various runs. Despite its long history of successful practice, theoretical studies of TS have emerged comparatively later, with, in part, the difficulty of analysing its randomized exploration mechanism [6].

3. Exploration Strategies under Finite-Horizon Constraints

3.1 Finite Horizons and Sensitivity to Early-Stage Behavior

In problems of finite-horizon decision-making, exploration behavior during the initial stages is the key to the overall performance. Algorithms will be forced to make crucial decisions with limited time and data, and in these situations, a few sufficient data points can allow the algorithms to stabilize in a steady long-term regime. The losses incurred in exploration in the initial days are usually hard to recover in the later cases and thus may create a compelling effect on the overall performance. Due to this, similar (asymptotically guaranteed) algorithms may still have significantly different behavior on finite horizons, and this behavior is often observed when studied in the context of finite-time regret [10]. This observation is supported by empirical comparisons of UCB, ETF, and TS in horizons. Hu switches the number of rounds and demonstrates that the ranking of different algorithms and regret curves varies significantly in short-run and long-run contexts, which prove to be quite sensitive to finite-horizon effects [11].

Algorithms also cannot wait until adequate learning is achieved, so they must make consequential decisions in practice in small user experiments or short product trial time. In these circumstances, it is essential to understand the behavior of exploration at an initial stage and the accompanying risks to determine the feasibility of practicality.

3.2 Costs and Risks of Different Exploration Mechanisms under Finite Horizons

The distribution of finite time performance due to differences in the exploration strategies is largely due to the difference in the treatment of the expenses involved in exploration, and the irreversibility of the decision. In this regard, ETF, UCB, and TS represent three different design options.

In ETC-type techniques, after the exploration stage is complete, subsequent decisions cannot be put back. In case the algorithm takes the wrong step in the process of exploration, future choices are hard to rectify. This irreversibility enhances the risks that were involved in early-stage exploration, and it is permanent when faced under finite horizons [11].

UCB still has the capacity to make amends on the initial mistakes; however, its exploratory growth would incur short-term expenses. This aggressive exploration can significantly impair the overall performance in an environment where exploration costs are high or have short horizons.

Conversely, Thompson Sampling obtains a more amenable equilibrium between exploration and exploitation by way of randomization in choices. This allows it get away with initial errors and incur very high costs of exploration. This is further in support of the fact that TS tends to be quite stable when it comes to both short-term and long-term performances in empirical comparison [5,11].

3.3 Parameter Sensitivity and Robustness Differences

In the context of finite time constraints, consideration of sensitivity to parameter choices would make a significant impact on the performance of bandit algorithms, as well as their practical value.

ETC needs the exploration length pre-specified, and this is a parameter that is hard to tune in practice. When the exploration phase is not very long, the algorithm may more importantly select the incorrect arm; when it is too lengthy, the expenditure on exploration is unreasonable. Consequently, ETC is very sensitive to parameter settings and not very resilient in small sample space conditions. It is also experimentally observed that the regret curves of ETC vary significantly when the exploration budget m x k is varied [11].

The UCB also relies on exploration parameters where the magnitude of the uncertainty bonus is regulated. Parameters have a variety of influences on the frequency of the initial searches of the algorithm, which may result in significant deviations in the finite-time performance. In practice, this reliance complicates tuning of parameters and may decrease stability between settings, despite UCB possessing good asymptotic properties [1].

In comparison, TS is not so sensitive to explicit parameter choices. It has a behavior of exploration that is largely guided by posterior uncertainty, as opposed to exploration intensities manually defined, which tends to make it more resilient to a variety of conditions and time scales [6]. Concurrently, such strongness is accompanied by randomness, hence variations between runs in terms of performance can occur.

4. Implications for real-world decision systems

In continuation of the mechanism-level discussion of ETC, UCB, and TS in the foregoing sections, this section can talk of how the finite-time behavior of the three algorithms can have practical consequences with regards to design of real-world decision systems.

4.1 Early Deployment and Short-Horizon Applications

In high levels of initial deployment or in short decision horizon, the cumulative or overall system performance is highly propelled by the original decisions. When an algorithm has high exploration costs when it starts then such costs are sometimes hard to recoup within a restricted time constraint. This follows the previous results: at an expression of round diversity and UCB, ETC, and TS considered critically on variable horizons, the regret plots observed vary significantly with horizon a small number of rounds turns to a large number of rounds.

Deployment wise, ETC is especially susceptible to a short-horizon decentre in that the initial misidentification of the optimal arm at an early stage cannot be orchestrated once commitment has started. Whereas, UCB can always make changes in decisions as time goes by, it is generally associated with high initial expenses through aggressive exploration. Conversely, Thompson sampling is commonly associated with smoother initial behaviour and is described as having done well on short and long horizons in relative experimentation [11].

4.2 High-Cost and Irreversible Decision Environments

The management of irreversible risk is an issue of focus in the context of high-cost individual decision-making (e.g., medical interventions), or when errors are hard to correct (e.g., systems facing the user).

The fact that ETC does not do post-commit correction to a great extent, combined with the fact that it is the setting of practical value limits its applicability, and the fact that ETC results can shift drastically with the length of the exploration phase compound this perceived risk and makes a miscalibration in the early calibration especially expensive in these settings [11].

At UCB, more flexibility is provided by lifelong updating, yet can be associated with unacceptable losses in cases where even the exploration can be an expensive affair. By responding sensitively to initial uncertainties with the Sampling of Thompson by maximizing exploration and minimizing exploitation in decisions taken randomly, can provide a more cautious approach to exploration.

4.3 Robustness Under Limited Parameter Tuning

On a real-world system, a lot of parameter tuning is not viable because of modified and altering environments or scarce operational resources. Consequently, the highly sensitive algorithms can fail to be reliable in the deployment.

The sensitivity of ETC to parameter settings is particularly high, when taking into account that it is positioned on a fixed exploration length. This can be supported by empirical research: with a diversified exploration budget, there are big changes in the performance of ETF [11]. UCB also needs to exercise good calibration of exploration parameters to prevent both an early exploitative and an over-exploration. In comparison, TS is less reliant on explicit exploration parameters, hence more resistant to various deployment conditions. Simultaneously, there is increased stochastic variability with such robustness.

4.4 Design Insights for Algorithm Selection

Combined, these observations indicate that the process of selecting an algorithm in finite-horizon settings ought not only to consider the asymptotic regret guarantees, but also the behavior at an early stage, the irreversibility of decisions, and resilience to parameter settings. ETC is better adapted to a low-risk environment with the exploration phases clearly defined whereas UCB applies to the issues where the exploration costs can be accepted in exchange of continued correction. Thompson Sampling offers a more balanced approach in systems where early decisions are one-way and stability as well as long term performance matters [11].

5. Limitations and Potential Improvements

Even though this paper discusses the exploration mechanisms of ETC, UCB, and TS utilizing the finite-time point of view, there are a few limitations.

To begin with, the analysis in this paper is based on the classical stationary stochastic multi-armed bandit scenario. However, in practice, user preferences or experimental conditions can vary differently with time resulting in non-stationary reward distributions. In this case, the finite-time behaviour might vary significantly as compared to what is observed in the stationary assumptions [2].

Second, this literature is concerned with mechanism level, qualitative analysis and does not perform systematic quantitative appraisal of tasks or datasets. Whereas earlier research has shown unequivocal differences in the early-stage performance of these algorithms [5, 11], these findings are not duplicated in real-world scenarios, which limits the immediate practical applicability of the conclusions.

Third, the parameter sensitivity is not that much discussed and is more structural. Practically, exploration parameters are frequently fine-tuned online or via data-driven methods, which somewhat alleviate finite-time sensitivity to the manual selection of exploration parameters. Recent research on adaptive confidence limits also indicates that these mechanisms can have major implications on the behavior at an early stage [9].

Lastly, the paper only discusses one of the basis of the decision-makers and does not cover or discuss the multi-agent or competitive dynamics. Relations among agents that engage in various learning could add more complexities and create effects at finite time that are beyond the context of this work.

6. Future Directions

Based on the analytical model of this paper, there are various directions that would be worth researching. The next obvious extension is the application of the finite-time perspective to non-stationary or contextual bandit problems. Exploration must change with these settings or adapt to any contextual information in these settings, making exploration mechanisms more complicated [2].

The other administration opportunity is to come up with hybrid exploration plans. As an illustration, the structural simpleness of ETC coupled with randomized exploration of TS, or data-driven, adaptive adjustment mechanisms, to the UCB paradigm can be combined in ways that result in improved trade-offs within finite-horizon requirements [8].

Lastly, it would be interesting to apply the mechanisms level of findings of this paper to real world case studies, say a recommendation system or an internet medical decision making to determine the applicability and accuracy of these findings in real life application [5, 12].

7.Conclusion

This paper adopts a finite-horizon perspective to conduct a mechanism-level comparison of three classical multi-armed bandit algorithms: ETC, UCB, and TS. Unlike studies that focus mainly on asymptotic optimality and long-run regret guarantees, this study place greater emphasis on how these algorithms explore in the early stages under practical constraints, such as limited data and the high impact of early decisions.

The analysis shows that the finite-time performance differences among the three algorithms mainly stem from three factors related to their exploration mechanisms: early exploration cost, decision reversibility, and sensitivity to parameter choices. ETC separates exploration and exploitation into two phases, but its main risk lies in the irreversibility of the commit decision, and its performance depends strongly on the choice of exploration length. UCB performs adaptive exploration based on optimistic estimates and can correct early mistakes, but under short horizons it often explores aggressively, leading to high early costs and strong sensitivity to exploration parameters. In comparison, TS uses randomized choices based on the posterior uncertainty, which generally results in further early exploration and a less volatile finite-time performance that does not require hand-tuning. Meanwhile, it is also variant among runs due to its randomness.

On the whole, asymptotic regret guarantees should not be used in the process of selecting algorithms to use in decision systems in the real world. A combination of practical considerations must be involved, including time horizon, exploration cost, and presence of tuning resources, and the early-stage behavior and risk issues must be considered carefully. Overall, ETC is better adapted to low-risk environments where exploration budgets are defined. UCB is applicable when the applications can sustain the increased exploration expenditure in the name of continued correction. TS is more appropriate in deployment environments where satisfaction depends on a long-term performance and early stability as well. This work offers a more concrete account both nature of the behavior of classical bandit algorithms, and is an indication of upcoming investigations in less simplistic contexts, including non-stationary, contextual, and safety-critical ones.

References

[1] Auer, P.: Using upper confidence bounds for online learning. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science. pp. 270-279 (2000)

[2] Slivkins, A.: Introduction to multi-armed bandits. Foundations and Trends in Machine Learning. 12(1-2), 1-286 (2019)

[3] Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics. 6(1), 4-22 (1985)

[4] Garivier, A., Ménard, P., Stoltz, G.: Explore first, exploit next: The true shape of regret in bandit problems. Mathematics of Operations Research. 44(2), 377-399 (2019)

[5] Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. In: Advances in Neural Information Processing Systems. Vol. 24 (2011)

[6] Russo, D., Van Roy, B.: An information-theoretic analysis of Thompson sampling. Journal of Machine Learning Research. 17(68), 1-30 (2016)

[7] Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning. 5(1), 1-122 (2012)
[8] Garivier, A., Lattimore, T., Kaufmann, E.: On explore-then-commit strategies. In: Advances in Neural Information Processing Systems. Vol. 29 (2016)
[9] Hao, B., Abbasi-Yadkori, Y., Wen, Z., et al.: Bootstrapping upper confidence bound. In: Advances in Neural Information Processing Systems. Vol. 32 (2019)
[10] Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine Learning. 47(2), 235-256 (2002)
[11] Hu, Q.: Performance comparison and analysis of UCB, ETC, and Thompson sampling algorithms in the multi-armed bandit problem. Technical Report, Sejtu-Leeds Joint School, South West Jiaotong University and University of Leeds, Chengdu (2024)
[12] Silva, N., Werneck, H., Silva, T., et al.: Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. Expert Systems with Applications. 197, 116669 (2022)