Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30166

Multi-arm Bandit Machine Exploration - Investigating the Performance Differences of Classical Algorithms Through Trade-off Analysis

Haokai Tang

Faculty of Science, University of Bristol, Bristol, UK
kevin_tang_5200@icloud.com

Abstract. The exploration-exploitation dilemma in multi-arm bandit problems has long been a classic challenge and serves as the foundation of reinforcement learning. It has applications in various industries, such as advertising online, A/B testing, and clinical medicine and so on. There are many MAB algorithms and each has its own advantages and disadvantages. This paper analyzes the performance of three classic MAB algorithms: the simple and effective ε-Greedy; the Upper Confidence Bound algorithm (UCB1)—which is more optimistic when facing uncertainty; and Thompson Sampling, an approach rooted in Bayesian inference. This paper conducts simulation experiments under the Bernoulli Machine environment using three evaluation criteria: cumulative regret, convergence speed, and parameter dependence, and comprehensively analyzes the performance of the three algorithms. The results show that Thompson sampling achieved the lowest cumulative regret and the fastest convergence speed, followed by UCB1. The performance of ε-Greedy is highly sensitive to its hyperparameters. These findings may provide some practical guidance for algorithm selection in real-world scenarios with similar properties and validate the theoretical advantages of the probability matching strategy.

Keywords: Multi-Armed Bandit, Exploration-Exploitation Dilemma, Cumulative Regret, Thompson Sampling, Algorithm Comparison

1. Introduction

The multi-armed bandit problem (MAB) is a core issue in fields such as reinforcement learning, clinical trials, decision theory, and A/B testing[1]. In the context of a stochastic MAB, in each decision round, humans need to select one of the limited arms, and the rewards of each arm come from a fixed but unknown distribution; and the goal of humans is to maximize the cumulative reward (equivalent to minimizing the cumulative regret) within a limited number of rounds by balancing the costs of exploration and exploitation. In recent years, MAB research has produced many classic algorithms with theoretical regret bounds: such as UCB1 and its variants, Thompson Sampling, KL-UCB, etc. Lai & Robbins proved that under certain gap conditions, some algorithms can achieve logarithmic regret and provided lower bounds to illustrate that these results are optimal in many scenarios [2]. Subsequently, Agrawal proved that there do exist algorithms that can achieve

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30166

logarithmic regret, turning the theory into reality [3]. However, in practical applications, the performance of these algorithms is significantly influenced by the environmental parameters set in experiments—including the total number of trials, the magnitude of gaps between the expected rewards of different arms, the time horizon of decision-making, and computational complexity constraints. Therefore, the current shortcomings of MAB research lie in the fact that although there are numerous diverse theoretical regret analyses, the empirical studies that systematically compare the performance of classic algorithms under different parameter environments are very limited, especially in terms of cumulative regret, convergence speed, and parameter sensitivity comparisons. This paper compares the performance of three classic algorithms, ε-greedy, UCB1, and Thompson Sampling, in a controlled 10-arm random slot machine environment. It examines cumulative regret growth, convergence speed, and parameter settings sensitivity. Multiple experiments are repeated and analyzed for accuracy.

The significance of this study lies in: For practitioners who need to deploy MAB strategies in real-world scenarios (such as recommendation systems, advertising placement, and clinical trial design), our results provide a reference for algorithm selection and demonstrate the trade-offs in different environments. At the same time, this study supplements the theoretical regret bounds through experimental results, providing a basis for understanding the gap between theory and practice.

2. Theoretical foundation

This section will introduce the basic definitions that the experiment will involve and explain the three classic algorithms that we will use to evaluate its performance.

2.1. Cumulative regret

Cumulative regret measures the difference between the expected reward of always choosing the optimal arm and the cumulative expected reward of the arm chosen by the algorithm over time:

$$R(\mathrm{T}) = T\mu^* - \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{E}[\mu_{a_t}] \tag{1}$$

Where μ^* is the expected reward of choosing the optimal arm, and at is the arm chosen at time t. Minimizing cumulative regret is one of the main goals of all algorithms.

2.2. ε-Greedy algorithm

The ε-greedy algorithm was named and widely promoted by Sutton & Barto: in each round, an arm is randomly selected with a probability of ε, and the arm with the highest current experience value is selected with a probability of 1-ε [4]. This is a very simple algorithm, but its performance is greatly affected by the choice of ε.

2.3. UCB1 algorithm

Auer, Cesa-Bianchi, & Fischer proposed the Upper Confidence Bound 1 (UCB1) algorithm [5]. The key innovation of the UCB1 algorithm lies in selecting the arm that maximizes the sum of its empirical average reward and the confidence bound, and this confidence bound decreases as the

DOI: 10.54254/2755-2721/2025.LD30166

number of observations increases. Therefore, the UCB1 algorithm effectively balances the cost of exploration and exploitation in the two stages by using this rule. Its excellent theoretical performance, simplicity, ease of implementation, and 'optimistic' attitude make it a classic algorithm for solving the exploration-exploitation problem.

2.4. Thompson sampling algorithm

"Thompson Sampling" is an algorithm based on Bayesian thinking. Thompson proposed its initial idea, but it was buried due to the insufficient computing power at that time [6]. Later, Chapelle & Li systematically reintroduced the Thompson Sampling algorithm [7]. In the modern era, with the advancement of science, the computing power of computers has significantly improved. People began to pay attention to the Thompson Sampling algorithm again and invested computing resources into it, achieving good results. It samples from the posterior distribution of the reward probabilities of each arm and selects the arm with the highest sampling value. Then, based on the actual return of the action taken, it updates the probability distribution corresponding to that action. This process excellently integrates exploration and exploitation, enabling the Thompson Sampling to frequently achieve outstanding performance, and it is the most promising algorithm in this experiment.

3. Methodology

This section specifically describes how the experiment was conducted, including the design of the experimental environment and the experimental procedures. The experiment involved a 10-armed bandit environment with a Bernoulli distribution for reward and success probability. A total of the above three algorithms were tested. The ε-greedy algorithm used 0.03, 0.05, and 0.07 as evaluation parameters respectively, which can better demonstrate the parameter dependence of the ε-greedy algorithm. UCB1 using standard parameters, and Thompson Sampling initializing the prior distribution as Beta(1,1). Each algorithm was interacted with the bandit environment for 20,000 rounds, recording cumulative regret and reward trajectories in real time. The parameter dependence can be derived from the above two sets of data. To ensure reliability, outcomes were averaged across 100 independent runs and the standard error calculated.

4. Results

This section presents the results and charts obtained from the experiment.

DOI: 10.54254/2755-2721/2025.LD30166

4.1. Cumulative regret analysis

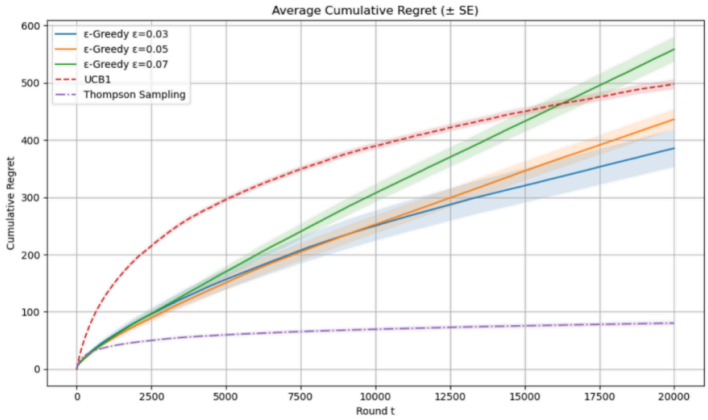

Figure 1. Average cumulative regret

Figure 1 shows the performance differences of three different strategies (ε-greedy, UCB1, and Thompson Sampling algorithm) in terms of cumulative regret over 20,000 rounds. As expected, the cumulative loss of the Thompson sampling method is the lowest, and its cumulative loss curve has gradually flattened over time, contrasting sharply with the increasing cumulative loss curves of the ε-greedy family and UCB1. This indicates that the Thompson sampling method achieves a relatively better balance between exploration and exploitation and has better adaptability and coordination. The UCB1 algorithm experiences rapid cumulative loss growth in the initial stage due to excessive exploration, but its loss curve also gradually slows down in subsequent rounds, with an overall trend very similar to the logarithmic regret characteristic. The ε-greedy variants display a linear cumulative regret growth pattern. From the three curves of the ε-greedy series, it can be seen that the exploration rate has a significant impact on the performance of the algorithm. Among these three curves, the ε=0.03 variant achieves the best balance and lower cumulative regret value, indicating poor long-term performance.

4.2. Convergence speed

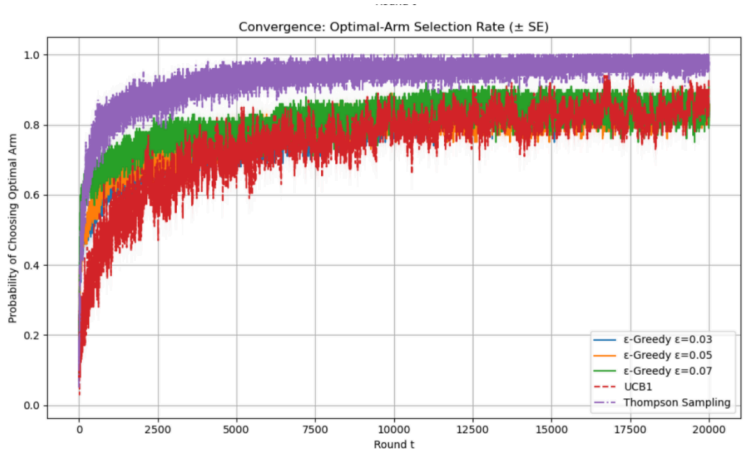

Figure 2. Convergence speed

DOI: 10.54254/2755-2721/2025.LD30166

Figure 2 shows the optimal strategy selection rate, which measures the degree to which the system converges towards always choosing the best action. The Thompson sampling algorithm can quickly achieve a nearly perfect selection rate, very close to 1, and maintains this performance with a very small variance, demonstrating its excellent stability. This fully proves its powerful performance in practical applications. For the ε-greedy algorithm, the early performance with ε = 0.07 is superior to ε = 0.03 and ε = 0.05, but the convergence speeds of the three eventually tend to be similar. A larger ε will have a higher exploration frequency, thus enabling faster sampling of the true optimal arm. The ε-Greedy variants accumulate enough samples for convergent success probability estimations. UCB1 shows stable improvement but slower convergence speed than Thompson Sampling. Thompson Sampling achieves lowest cumulative loss and fastest convergence speed, while UCB1 is reliable but has poorer stability in practical applications. ε-greedy strategy depends on parameter adjustment.

4.3. Parameter sensitivity

Figures 1 and 2 both illustrate the dramatic performance changes of ε-Greedy due to parameter variations, confirming its dependence on parameter adjustments. A smaller ε (0.03) improves the long-term performance, but there is a risk of suboptimal exploration. UCB1 and Thompson Sampling are steadier and can still exhibit stable performance without explicit parameter adjustments.

5. Discussion

The experimental results show that there are significant performance differences in terms of cumulative regret, convergence speed and parameter sensitivity among different algorithms (ε-greedy variants, UCB1 and Thompson Sampling).

Firstly, in terms of cumulative regret, Thompson Sampling consistently outperforms other algorithms. The cumulative regret curve of the algorithm gradually flattens over time, indicating that the Thompson sampling algorithm excellently balances exploration and exploitation, demonstrating its outstanding adaptability to the environment. Agrawal and Goyal rigorously proved that Thompson sampling can achieve logarithmic expected regret in the random multi-armed bandit problem, providing a theoretical basis for the excellent performance of the Thompson sampling algorithm in this experiment and proving the optimality of its regret upper bound [8]. The ε-greedy algorithm shows linear regret growth, highlighting its deficiency in long-term performance. The UCB1 algorithm exhibits logarithmic regret growth, but due to excessive exploration in the initial stage, the regret value rises sharply in the early stage [5]. Therefore, it can be concluded that this Bayesian method - the Thompson sampling method - has superior long-term performance. Secondly, in terms of convergence speed, the Thompson sampling method performs well, demonstrating its efficiency and stability. The convergence speed of the UCB1 algorithm is slow and fluctuates greatly, performing the worst in the initial stage, but similar to the ε-greedy series in the middle and later stages. The ε-greedy series algorithms increase the initial convergence speed when the exploration rate increases, but their convergence speed is slow in the later stage, highlighting their instability in practical applications.

Finally, in terms of parameter sensitivity, the ε-greedy algorithm has the strongest dependence on the exploration rate (ε) among the three algorithms. Even a small change in ε can lead to significant performance differences - a larger ε value accelerates initial learning, but at the cost of a rapid increase in long-term regret; while a smaller ε value, although the cumulative regret increases more

DOI: 10.54254/2755-2721/2025.LD30166

gradually, due to insufficient exploration, may not be able to accurately find the optimal arm. This observation result emphasizes the significant value of parameter-light or parameter-free Markov decision process algorithm in practical applications, especially in actual scenarios where it is unknown or difficult to predefine "optimal parameter settings" [9].

In summary, among the three classic MAB algorithms evaluated, Thompson Sampling emerges as the most stable and practical solution, outperforming the others in both efficiency (i.e., lowest cumulative regret) and stability (i.e., consistent performance across parameter environments). UCB1, although highly theoretically appealing, may face challenges in environments that penalize excessive exploration (such as in clinical medicine). The ε-greedy strategy, although simple and easy to implement, has low stability due to its parameter sensitivity and linear regret growth. In the future, further exploration of variants of UCB algorithms based on information theory concepts, such as the KL-UCB algorithm , can be conducted to enhance the understanding of UCB algorithms [10].

6. Conclusion

In conclusion, this experiment conducted a comprehensive evaluation of the performance of three classic multi-armed bandit algorithms. The results clearly showed that the Thompson sampling algorithm had the strongest overall capability, followed by UCB1, and the ε-greedy series of algorithms performed the weakest. Thompson Sampling, rooted in Bayesian thinking, delivers outstanding comprehensive performance. It not only has the lowest long-term cumulative regret, but also can converge rapidly and stably, and has the lowest dependence on parameter adjustments. These advantages make it have both theoretical basis and practical stability and is particularly suitable for real application scenarios that require stable long-term performance. The UCB1 algorithm is more appropriate for scenarios where one is less concerned about the initial regret and where the environment is stable. Although it has provable theoretical guarantees, its "overly optimistic" principle can lead to high exploration costs in the initial stages of practice. The ε-greedy series of algorithms demonstrate the limitations of heuristic rule-based methods, as their performance is highly dependent on the choice of ε. Therefore, they face a difficult trade-off between insufficient exploration and excessive exploration, and require experience accumulation to select the appropriate ε. ε-Greedy algorithms, despite their simplicity and efficiency, are not reliable in high-stakes applications due to their linearly increasing cumulative regret and severe parameter sensitivity. Ultimately, the choice of algorithm should depend on the specific application scenario. This experiment strongly demonstrates that, when selecting an algorithm to address the exploration-exploitation dilemma, Thompson Sampling is the preferred algorithm in most cases.

References

[1] Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and no stochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 5(1).
[2] Lai, Tze Leung, & Robbins, Herbert. (1985). Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1), 4–22
[3] Agrawal, Rajeev. (1995). Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4), 1054–1078.
[4] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd Edition). MIT Press.
[5] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2), 235–256.
[6] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4), 285–294.

DOI: 10.54254/2755-2721/2025.LD30166

[7] Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In Advances in Neural Information Processing Systems(pp. 2249-2257).
[8] Agrawal, S., & Goyal, N. (2012). Analysis of Thompson Sampling for the Multi-armed Bandit Problem. Proceedings of the 25th Annual Conference on Learning Theory (COLT).
[9] Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
[10] Garivier, A., & Cappé, O. (2011). The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. Proceedings of the 24th Annual Conference on Learning Theory (COLT).