Expanding the Horizon: Diverse Applications and Insights from Multi-Armed Bandit Algorithms

Haoyang Li *

College of Arts & Science, New York University, New York, 10012, United States

* Corresponding Author Email: hl3951@nyu.edu

Abstract. In an era marked by the ever-increasing convergence of applications and algorithms, the imperative to maximize efficiency while mitigating the adverse effects of uncertainty has emerged as a critical objective for application developers. Despite this, a significant body of research on Multi-Armed Bandit (MAB) algorithms has predominantly focused on the comparative analysis of their performance, often overlooking their tangible impact on diverse applications. This study builds upon the data and findings of several preceding investigations that examine the application of MAB algorithms across various domains, including e-commerce, clinical trials, and dynamic pricing strategies. Our findings underscore the versatility and adaptability of MAB algorithms in enhancing the performance of applications across these varied fields. Notably, certain MAB algorithms demonstrate a higher suitability for specific scenarios compared to others. Consequently, this research posits that achieving an optimal balance between exploration and exploitation, and thereby maximizing rewards in uncertain environments, necessitates not just the application of MAB algorithms but also the strategic selection of the most effective algorithm tailored to each unique context. Overall, this study aims to showcase the wide-ranging applicability of MAB algorithms, offering a comprehensive exploration of their capabilities and impact across multiple sectors.

Keywords: Multi-armed bandit; explore-then-Commit (ETC); upper confidence bound (UCB); Thompson sampling (TS); exploration-exploitation, decision-making.

1. Introduction

In today's digital landscape, applications offer value and convenience in myriad ways, yet the underlying mechanisms that power these functionalities remain obscure to many users. For instance, the interactivity of websites, a common yet intricate aspect, is largely driven by complex algorithms. Among these, Multi-Armed Bandit algorithms stand out for their role in facilitating decision-making under uncertain conditions.

MAB algorithms, a subset of reinforcement learning, originate from the challenge of selecting from various options with unknown rewards. Broadly, the MAB problem revolves around the optimal allocation of limited resources among diverse choices to maximize the returns. Central to this is the exploration-exploitation dilemma in machine learning and artificial intelligence: exploration involves probing various possibilities, while exploitation leverages gathered information for optimal decision-making. MAB algorithms address this challenge by enabling effective decision-making with limited resources.

The diverse applications of MAB algorithms underscore their significance across various sectors, such as recommendation systems, online advertising, and dynamic pricing. In recommendation systems, for instance, MAB algorithms optimize content suggestions, enhancing click-through rates [1]. Airlines, similarly, utilize MAB algorithms for dynamic pricing of airfares to boost profits [2]. These instances demonstrate the versatility of MAB algorithms in tackling complex real-world challenges. However, much research, like Singh's work on "Reinforcement Learning Based Empirical Comparison of UCB, ε-greedy, and Thompson Sampling" [3], primarily focuses on comparing algorithmic performances rather than their practical applications. Consequently, there remains a gap in comprehensive research exploring the applications of MAB algorithms in various domains. This gap presents an opportunity for deeper investigation into the potential of MAB algorithms in everyday life.

This paper, therefore, shifts the focus from algorithmic performance to the application of MAB algorithms in recommendation systems, clinical trials, and dynamic pricing. It delves into the applications and underpinning strategies of MAB algorithms, moving beyond mere explanations to a detailed analysis. The paper is structured into several sections, with the second section introducing key theories of MAB problems and essential algorithms. The third section showcases the varied applications and strategies of different MAB algorithms, while the fourth discusses encountered challenges and offers recommendations for future research. In summary, this study aims to elucidate the versatility of MAB algorithms by highlighting their extensive applications and the insights they provide for optimal decision-making. By thoroughly analyzing the use of MAB algorithms, this paper seeks to deepen the understanding of their capabilities and stimulate further research in this promising field.

2. Theoretical Foundations and Core Strategies

2.1. Fundamentals of MAB Problems

A bandit problem is a sequential game between a learner and an environment, and the game is played over n rounds. In each round t = 1, 2… n, the learner chooses an action A_t from a set of k possible actions/arms and receives a random rewardX_t. The objective of the learner is to maximize the cumulative reward over n rounds, i.e., maximize $\sum_{t=1}^{n} X_t = X_1 + X_2 + \cdots + X_n$. Equivalently, this objective can be stated as the minimization of regret, which is defined as the following: regret = reward lost by taking sub-optimal decisions = largest possible cumulative reward in n rounds if the learner knows which arm is the "best" - $\sum_{t=1}^{n} X_t$.

However, the environment does not reveal the reward of the actions not selected by the learner. Thus, the learner must gain information by repeatedly selecting all actions, which is exploration. When the learner chooses a bad/sub-optimal action, it loses from the cumulative reward since the learner should choose the action that returns the largest reward so far, which is exploiation. These two conditions together constitute the exploration-exploitation dilemma. The MAB algorithms aim to overcome this issue.

2.2. Key Algorithms and Strategic Approaches

Explore-then-Commit: the ETC algorithm is one of the ways to balance exploration and exploitation [4]. This algorithm explores by playing each arm a fixed number of times and then exploits by committing to the arm that appears the best during exploration, which shares the same idea as A/B testing. Assume there are k arms and m is the number of times each arm will be explored. The ETC algorithm chooses each arm in a round-robin fashion until all k arms are selected m times each for the exploration phase. In the exploitation phase, the ETC algorithm will start with round t = mk + 1, the algorithm selects the arm with the largest average reward in the exploration phase for all future rounds.

Upper Confidence Bound: The UCB algorithm is based on the principle of optimism under uncertainty [5]. Being optimistic about the unknown supports exploration of different choices, particularly those that have not been selected many times. In each round, a value is assigned to each arm (the UCB index of that arm) based on the data observed so far that is an overestimate of its mean reward (with high probability). Then the arm with the largest value/index will be chosen.

Thompson Sampling: Thompson sampling is a close-to-optimal algorithm in a wide range of settings [6]. TS often exhibits superior performance in experiments and practical settings compared to UCB and its variants. Nevertheless, TS has one disadvantage: it has larger variance in its performance from one experienment to the next. TS works by choosing actions with Bayesian approach to estimate the reward probability distributions and favoring actions that return greater rewards. Over time, TS will balance exploration and exploitation, fostering better decision-making.

3. Diverse Applications and Empirical Research

3.1. Optimization of Content Recommendation in E-commerce

MAB algorithms, a major depature from the traditional A/B testing, have transformed content recommendation in e-commerce. Specifically, MAB algorithms are used in e-commerce to tailor content suggestions to maximize the possibility of exposing certain products to potential customers. The main objective is to provide the most relevant products or services to enhance user engament and increase sales. Different from A/B testing that distributes traffic evenly among alternatives for a certain amount of time, MAB algorithms keep monitoring user reactions to various suggestions offered and make real-time adjustments to their recommendation strategies.

For example, the research Adaptively Optimize Content Recommendation Using Multi Armed Bandit Algorithms in E-commerce investigates how to apply MAB algorithms in E-commerce platforms to decrease shopping frictions and improve customer satisfactions [7]. The researchers study three MAB algorithms, which are ε-greedy, UCB, and TS. First, using simulated datasets that mimic non-stationary customer preferences, the research finds that all the three algorithms can optimize recommendations in an adaptable manner, and the UCB algorithm has a better performance than TS. Then these researchers conducted more than 1000 experiments by using historical A/B testing datasets from a primary e-commerce website. The results show that larger differences between the success rates of competing recommendations lead to greater cumulative rewards for the MAB algorithms. To solve the delayed reward issue in e-commerce, the researchers develop a batch-updated MAB algorithm to optimize online content recommendation. This batch-updated MAB algorithm is shown to significantly outform other methods, being effective in handling changing user preferences and shopping frictions. The outstanding performance of MAB algorithms demonstrates their ability to increase customer satisfaction, click-through rates, and conversion rates, which are essential e-commerce metrics. These following pictures exhibit the testing results obtained by the researchers, including traffic allocation patterns and performances of different MAB algorithms in their e-commerce scenario. As shown in Fig 1 and 2.

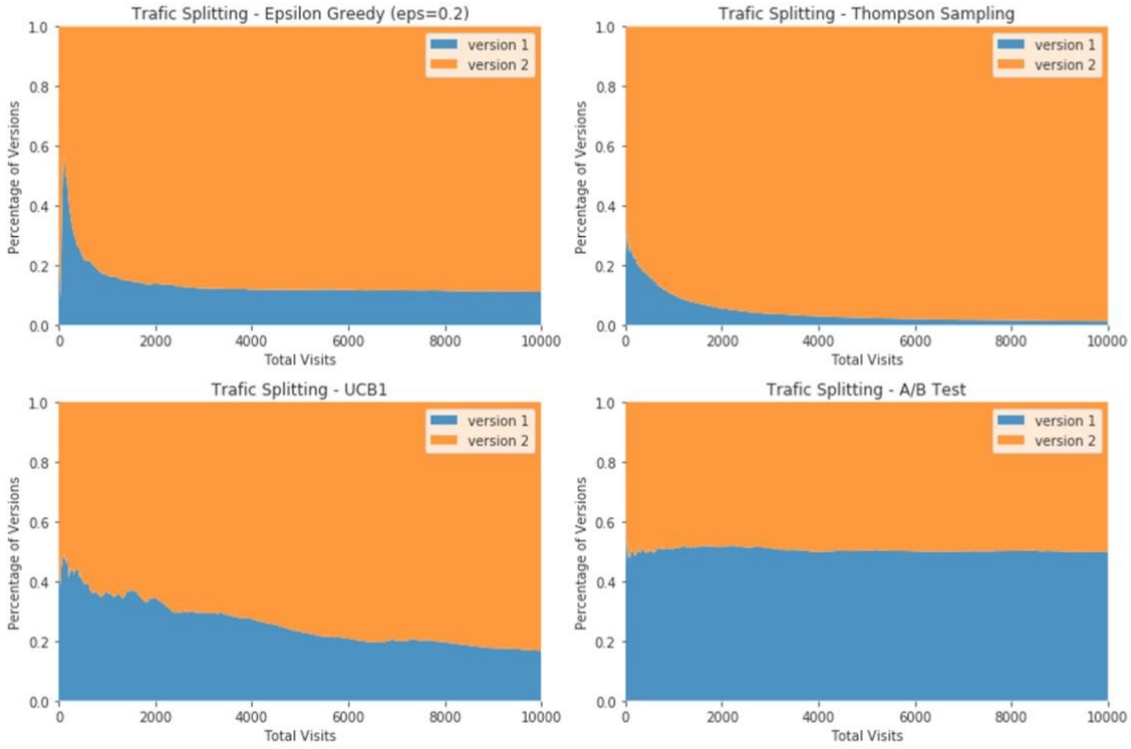


Fig 1. Traffic allocation patterns of the MAB algorithms and typical A/B testing (Photo/Picture credit: Original).

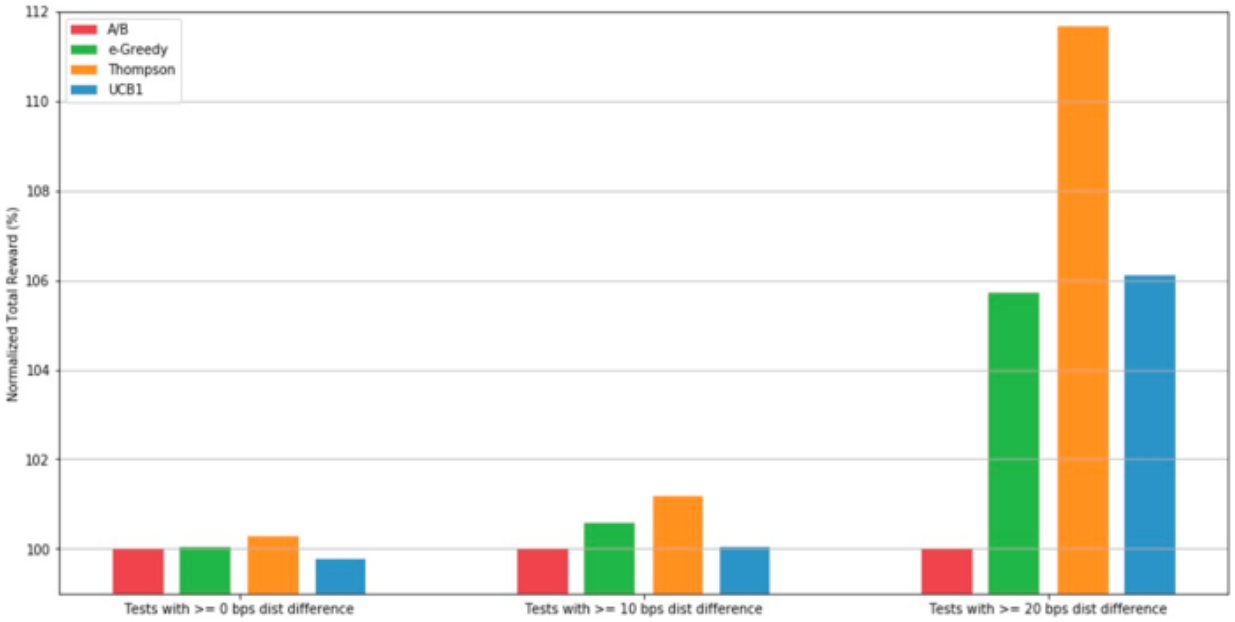
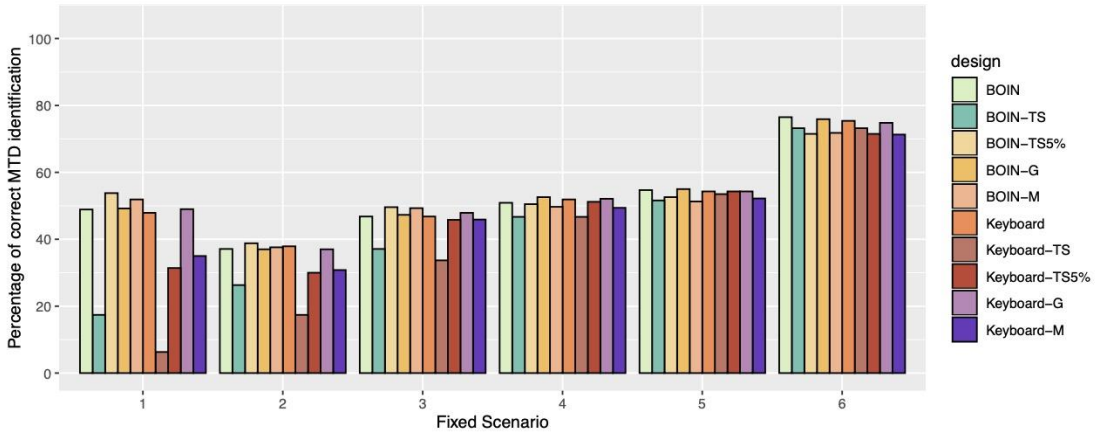


Fig 2. Comparison of normalized performance of different MAB algorithms under different scenarios (Photo/Picture credit: Original).

3.2. MAB in Model-Assisted Design for Dose-Finding Clinical Trials

Multi-Armed Bandit algorithms offer a dynamic approach to patient allocation in clinical trials, adapting to ongoing results, a stark contrast to the static assignment in traditional fixed-design trials which often leads to resource wastage and delays in identifying optimal therapies. The ability to adapt in real-time is particularly advantageous in clinical settings. MAB algorithms enable the assignment of new patients to the most promising treatments based on the observed responses to therapies. This approach not only accelerates the discovery of effective treatments but also ethically enhances patient outcomes by reducing the administration of ineffective medicines.

For example, the study "Application of Multi-Armed Bandits to Model-assisted designs for Dose-Finding Clinical Trials" [8] investigates the integration of MAB algorithms into assisted clinical trial model designs, aiming to determine optimal medicine dosages. The primary objective in this context is to maximize rewards within a limited number of trials. The research focuses on improving Keyboard designs and Bayesian optimal intervals (BOIN), two model-assisted designs for dose discovery in clinical trials. These designs adjust dosages based on reported toxicity rates using straightforward criteria. By implementing Thompson sampling, greedy algorithms, and posterior median-based approaches, the researchers propose new MAB variants of these models. These methods aim to identify the most effective dose within the constraints of small sample sizes typical in dose-finding studies, relying on various statistical metrics. The performance of these designs is then evaluated through simulation experiments under diverse scenarios. The findings indicate that in certain conditions, the MAB versions of the BOIN and Keyboard designs can surpass traditional methods, especially when the accurate maximum tolerated dosage (MTD) is above the moderate dose threshold. In conclusion, the study reveals that these innovative designs could lead to more precise identification of MTDs in dose-finding clinical trials, potentially resulting in a safer and more efficient drug development process. The accompanying figures illustrate the simulation results for accurate MTD identification obtained by the researchers, as depicted in Fig 3.



TS: Thompson sampling. TS5%: Thompson sampling-0.05. G: Greedy algorithm. M: Median algorithm

Fig 3. Simulation results for correct MTD identification (Photo/Picture credit: Original).

3.3. Strategic Optimization of Pricing Models

MAB algorithms are a critical tool for strategic optimization in pricing models. Traditional pricing strategies often falter in dynamic markets due to fluctuating customer preferences and competitive landscapes. In contrast, MAB algorithms offer a robust solution by evaluating consumer responses to diverse pricing strategies in real time. This capability allows for iterative testing of various price points, garnering insights from consumer behavior, and refining pricing strategies accordingly. The learning ability inherent in MAB algorithms is essential for understanding price elasticity and customer behavior patterns. Furthermore, their adeptness in balancing exploration and exploitation ensures opportunities to discover more effective solutions while maximizing immediate revenue through the most efficient pricing models.

In competitive markets, businesses can leverage MAB algorithms to dynamically adjust prices in response to external events, thereby gaining a deeper understanding of competitors' pricing strategies and market trends. This approach allows for pricing models that are attuned to consumer preferences and market conditions, enhancing volume, revenue, and market share. An exemplary case of MAB algorithms' application in strategic pricing is evident in the research titled "Multi-Armed Bandit to Optimize the Pricing Strategy for Consumer Loans" [9]. This study establishes a baseline pricing plan using risk-based and tiered approaches. It delves into the theoretical aspects of MAB algorithms, including both stationary and non-stationary variants like UCB-based, ε-greedy, and Thompson sampling methods. The loan pricing issue is framed as an MAB problem, with synthetic data tailored to represent stationary and non-stationary settings. The study particularly emphasizes the efficacy of the Sliding-Window Thompson Sampling algorithm in this context, as demonstrated by the performance outcomes of the algorithms. Additionally, the researchers conduct a sensitivity analysis and discuss the trade-offs between regret and discrete errors. The study concludes by affirming the effectiveness of the Sliding-Window Thompson Sampling algorithm in non-stationary environments and its robustness against assumptions in non-stationary scenario modeling. It also acknowledges potential limitations and advocates for further research to enhance the model's practical applicability, particularly in the context of dynamic pricing in consumer loans within highly competitive markets. The research provides a comprehensive understanding of dynamic pricing strategies and their optimization using MAB algorithms, underpinned by the visualized results of the experiments, as depicted in Fig 4-9.

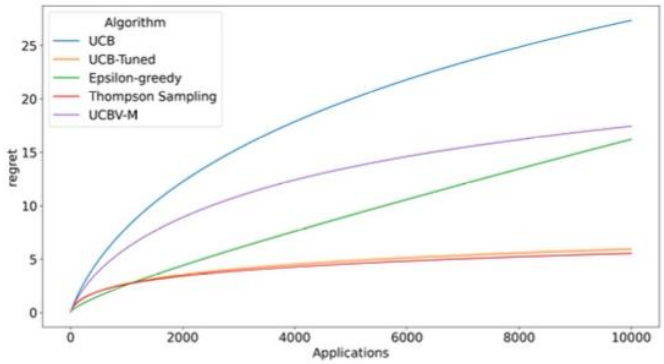


Fig 4. Average cumulative regret over 10000 applications for the statonary algorithms (Photo/Picture credit: Original).

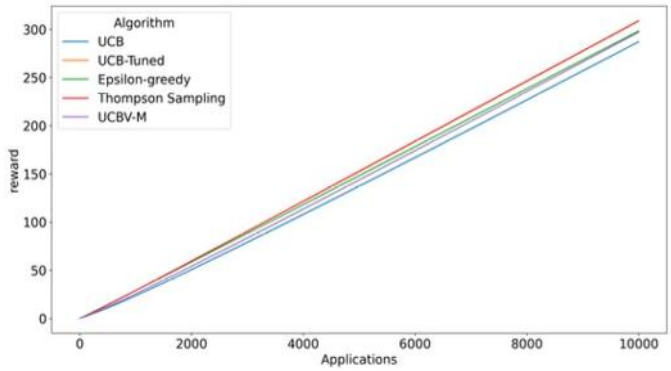


Fig 5. Average cumulative reward over 10000 applications for the statonary algorithms (Photo/Picture credit: Original).

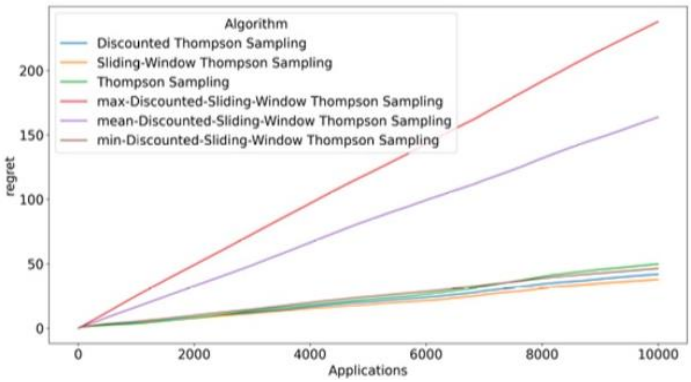


Fig 6. Cumulative regret over 10000 applications for the Thompson Sampling and the Non-Stationary algorithms using their best parameter values (Photo/Picture credit: Original).

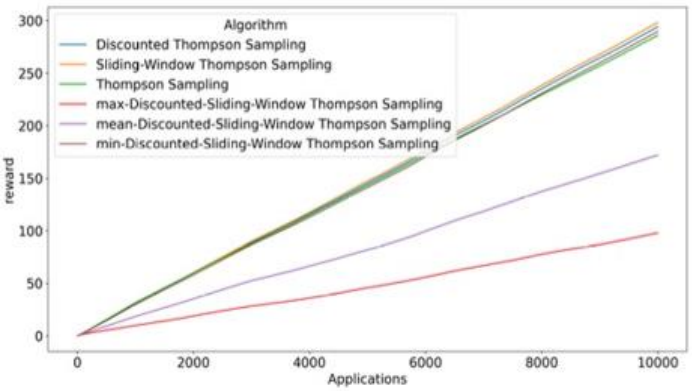


Fig 7. Cumulative reward over 10000 applications for the Thompson Sampling and the Non-Stationary algorithms using their best parameter values (Photo/Picture credit: Original).

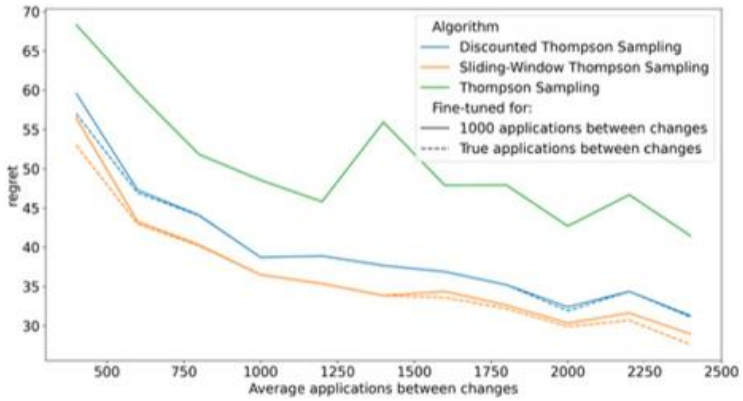


Fig 8. Cumulative regret after 10000 applications when the parameter has been tuned for the correct average number of applications between changes respectively on average 1000 applications between changes (Photo/Picture credit: Original).

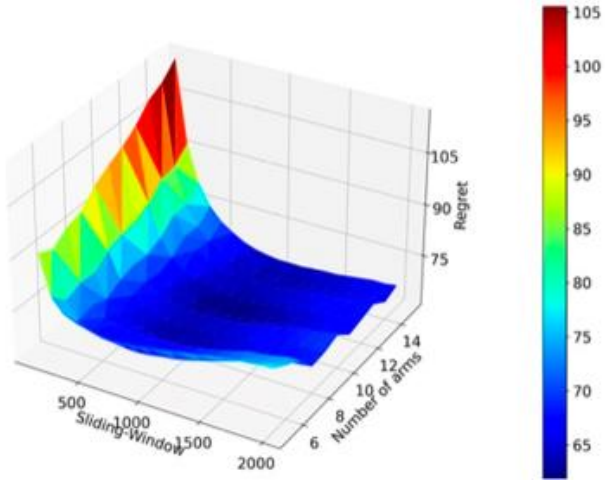


Fig 9. Cumulative regret + discretization error after 10000 applications for Sliding-Window Thompson Sampling as a function of the Sliding Window size and the number of atms (Photo/Picture credit: Original).

4. Challenges and Future Perspectives

4.1. Addressing the Challenges in Implementation and Scalability

Despite the high adaptability of MAB algorithms, there are several challenges in their implementation and scability.

Exploration and exploiation trade-off: Balancing the exploration of new options and the demand for high returns is crucial. However, to achieve this goal, the MAB algorithms need to be applied to specific large datasets to get actual performances of these algorithms in a certain context. Therefore, large-scale programming and detailed data analysis are essential parts in analyzing MAB algorithms' application. In this way, researchers can have an accurate visualization of how MAB algorithms deal with the exploration-exploitation dilemma.

Non-stationary environment: Real-world environments are always changing. Therefore, MAB algorithms must be capable of adjusting themselves to adapt to these changes. To achieve this purpose, MAB algorithms need to rely on past observations and reset their learning mechanisms in a regular fashion.

Performance monitoring: Optimal operations and adaptablity of MAB algorithms require constant monitoring of the algorithms' performances. Therefore, the actual execution of MAB algorithms on large datasets is usually a time-consuming process and requires careful attention from researchers [10].

Objective decision-making: When using MAB algorithms to make decisions, researchers have to guarantee that the design of these algorithms does not involve subjective factors that may biase the results. Instead, these algorithms should be based on objective data and working mechanisms to produce accurate results for users.

Scability: When the number of arms increases to a certain degree, the computational complexity can be a problem. Therefore, choosing appropriate MAB algorithms are crucial for ensuring the efficiency when working with large datasets. In particular, algorithms such as TS and UCB have good performances in this scenario.

4.2. Potential Directions for Future Research

There are still many promising areas for future research about MAB algorithms. The following options are some of the potential directions.

Deep learning integration: One potential area is combining deep learning with MAB algorithms to handle high-dimensional and sophisticated data spaces. Researches in this aspect can assist in handling complicated and non-linear patterns in data that conventional MAB algorithms cannot capture.

Multi-agent bandit problem: This area can explore how multiple learners deal with an identical environment. In particular, researches in this direction will provide insights for competitive market scenarios or collaborative filtering systems. The study Multi-agent multi-armed bandits with limited communication is a good example for potential relevant future researches in this field [10].

Cross-domain adaptability: Future researches can also focus on developing MAB algorithms to improve the algorithms' adaptability in diverse domains. In this way, the necessity of domain-specific tuning can be reduced.

Contextual bandits: Creating sophisticated contextual bandit algorithms with fast feedback adaptation is crucial in dynamic settings where situations can change quickly. Researches in this field may bring an improvement of contextual MAB algorithms' utility in fields like stock markets or real-time bidding systems. For instance, the research Risk-averse Contextual Multi-armed Bandit Problem with Linear Payoffs have already exhibited some potential of this type of algorithms.

Hybrid learning models: Fuure researches can also use MAB algorithms in parallel with other machine learning techniques like supervised learning to build hybrid models that can benefit from advantages of each other.

5. Conclusion

In summary, Multi-Armed Bandit algorithms have a broad range of real-world applications, with their effectiveness contingent upon both the inherent properties of the algorithms and the specific goals of the applications they are employed in. Consistent with the initial hypothesis, MAB algorithms demonstrate a robust capacity to navigate the trade-offs between exploration and exploitation. This research further reveals that different algorithms, such as ETC, UCB, and Thompson sampling, may be uniquely suited to varied applications. Consequently, selecting the most apt MAB algorithm tailored to the specific context is pivotal for optimizing algorithmic efficiency and achieving the most beneficial outcomes for users.

This study methodically explores the versatility of MAB algorithms across several applications, aiming to enhance readers' comprehension of the capabilities of MAB algorithms and the advantages of conducting further research in this area. However, it is acknowledged that this research cannot encompass all potential applications of MAB algorithms. As suggested, numerous other facets of MAB algorithms merit deeper investigation. Future researchers are encouraged to delve into these less-explored domains, offering fresh insights and innovative applications of MAB algorithms. In the long term, the integration of MAB algorithms into a wider array of applications could yield significant benefits, enhancing various aspects of everyday life.

References

[1] Louëdec, J., Chevalier, M., Mothe, J., Garivier, A., & Gerchinovitz, S. (2015). A Multiple-Play Bandit Algorithm Applied to Recommender Systems. The Florida AI Research Society.

[2] Singh, A. (2021). Reinforcement Learning Based Empirical Comparison of UCB, Epsilon-Greedy, and Thompson Sampling. Int. J. of Aquatic Science, 12(2), 2961-2969.

[3] Nie, G., Agarwal, M., Umrawal, A. K., Aggarwal, V., & Quinn, C. J. (2022, August). An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In Uncertainty in Artificial Intelligence (pp. 1541-1551). PMLR.

[4] Zhang, W., Hu, Z., & Li, G. (2023). Upper confident bound advantage function proximal policy optimization. Cluster Computing, 26(3), 2001-2010.

[5] Ding, Q., Hsieh, C. J., & Sharpnack, J. (2021, March). An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In International Conference on Artificial Intelligence and Statistics (pp. 1585-1593). PMLR.

[6] West, B., Wang, J., Cui, X., & Huang, J. (2021). Adaptively Optimize Content Recommendation Using Multi Armed Bandit Algorithms in E-commerce. ar**v preprint ar**v: 2108.01440.

[7] Kojima, M. (2022). Application of multi-armed bandits to model-assisted designs for dose-finding clinical trials.

[8] Agarwal, M., Aggarwal, V., & Azizzadenesheli, K. (2022). Multi-agent multi-armed bandits with limited communication. The Journal of Machine Learning Research, 23(1), 9529-9552.

[9] Gan, M., & Kwon, O. C. (2022). A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domains. Knowledge-Based Systems, 251, 109158.

[10] Lin, Y., Wang, Y., & Zhou, E. (2023). Risk-averse contextual multi-armed bandit problem with linear payoffs. Journal of Systems Science and Systems Engineering, 32(3), 267-288.