Multi-Armed Bandits and Their Classical Algorithms: Strengths and Limitations

*Sichen Wu**

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, 215000, Suzhou, China

Abstract. Multi-armed Bandits (MAB), which is the short form for multi-armed bandit, is playing a significant role in varied spaces including online advertising, recommendation system, clinical trials and medical decision-making and even in 5G. The model aims to balance "choosing the option with the highest reward known so far" and "exploring the options unknown". At last, it will reflect the results by a coefficient called regret. It is sometimes not suitable to some special situations. This article shows three of its classical algorithms: Explore Then Commit (ETC), Upper Confidence Bound (UCB) and Thompson Sampling (TS). ETC separates exploitation and commitments into two single parts but it will produce more regrets comparing to the other two algorithms. UCB selects arms based on the "average reward + upper confidence bound" metrics. It is one of the most popular algorithms. TS models uncertainty by leveraging Bayesian updates and random sampling. However, it is much harder than the others. They have different advantages and disadvantages. In summary, MAB has good application scenarios if people make good use of its different types of algorithms.

1 Introduction

Making decisions is an inevitable action in varied kinds of events. The correct decision will make the whole process smoother. At the same time, a wrong decision may ruin the plan. Therefore, making a correct decision before taking actions is necessary. Multi-armed bandit model can be an easy and powerful model. The model includes k actions to pull m bandit arms and the event will take t turns. In each turn the researcher pulls all the bandits to get a reward from them. Every bandit has different range of reward which is unknown and independent from the others. The algorithm aims to find out the different value between the best arm and the true arm(regret)in order to maximize the reward in the t turns. Aleksandrs Slivkins listed three running examples which can be unified by multi-armed bandit—news website, dynamic pricing and investment.

News website: While a new person starts, the website chooses an article header to demonstrate, observes if the one clicks on the header. The site's aim is to make the number of all clicks biggest.

* Corresponding author: Sichen.Wu24@student.xjtlu.edu.cn

Dynamic pricing A store is selling a digital good, e.g., an app or a song. When a new customer arrives, the store chooses a price offered to this customer. The customer buys (or not) and leaves forever. The store's goal is to maximize the total profit.

Investment Everyday you wake up, you choose one stock to invest into, then invest just one dollar. When the day is over, you observe the transformation in value for every stock. The aim is to make the total wealth bigger [1].

However, MAB model exists some drawbacks. While classical MAB formulations primarily focus on risk-neutral strategies aimed at maximizing expected rewards, real-world scenarios often demand the incorporation of risk preferences. This need arises because maximizing average performance may not adequately address concerns about reward variability, safety, or worst-case outcomes [2].

2 Algorithms of multi-armed bandit

2.1 Explore Then Commit Algorithm

Explring then commit algorithm, whose short form is ETC, is divided into two phases: explore phase and commit phase. Presumed that there are k bandits. In the explore phase, researchers will pull all the k arms n times and record all the n rewards written as r_{i1} , $r_{i2}, \dots, r_{in}$,analysing the average reward of each arm: $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} r_{it}$ .After the exploration,compare all the results to pick out the one with the highest average reward. Turning to the commit phase, only choose the best arm until the decision process finishes. The regret of the process is$\sum_{i=1}^{k} m\Delta i + (n - mk)\Delta i$. The multi-armed bandit problem, a framework with numerous practical applications, has undergone nearly a century of research since its early exploration. While our understanding of this problem is now relatively comprehensive, strategies that separate exploration and exploitation into distinct phases are inherently suboptimal. This conclusion carries significance: the "exploration-then-exploitation" approach is not only intuitive but also widely adopted in real-world scenarios—including the website optimization case mentioned earlier. Additionally, similar phase-separated strategies have been proposed in academic literature for more complex scenarios.

Recent advancements in optimal exploration policies, such as the work by Garivier and Kaufmann, might have led to the assumption that well-adjusted two-phase strategies could achieve near-optimal performance. However, we demonstrate the opposite: optimal strategies for multi-armed bandit problems must be fully sequential, with exploration and exploitation integrated throughout the process.

Since Wald' s research on simple hypothesis testing, it has been recognized that sequential methods can yield substantial advantages. In the context of bandit problems, the superiority of fully sequential procedures aligns with common sense: suppose one arm initially seems more promising, but later observations reveal underperformance—fixing a "commitment point" (as in two-phase strategies) would impose unnecessary constraints. In this paper, Aurélien Garivier clearly and precisely quantifies this constraint: on average, it results in an asymptotic regret that is twice as large. The proof of this finding connects classical technology from the analysis and the existing body of multi-armed bandit research [3].

2.2 Upper Confidence Bound Algorithm

The style for the names is First Names, typed in italic 10-point Times, then Last Name, typed in 10-point Times, with a comma after all except the last author, which is separated by comma

+ “and”. Do not use academic titles. Different from ETC algorithm, the main feature of upper confidence bound algorithm (UCB)is the mix of exploitation and exploration. In the UCB framework, all the arms have different average rewards and upper bound. To make sure each arm can be explored at least one time in the whole progress, all the k arms will be chosen and then record the correspondent rewards in the first k turns, From the k+1 turn, every turn will operate as follow.

1. Calculate the UCB of each arm. The equation is UCB_i(t)=$\hat{\mu}_i$+c* $\sqrt{\frac{lnt}{n_i}}$.Among this, $\hat{\mu}_i$ means the average reward of the ith arm in the $(t-1)$th turn.The equation is $\hat{\mu}_i=\frac{1}{n_i}\sum_{s=1}^{n_i} r_{is}$.c is the confidence coefficient and normally equally to 2.
2. Choose the arm with the largest UCB_i(t).
3. Renew the information.
4. Repeat the three steps above till the test is over.

This section identifies a new stability characteristic of the Upper Confidence Bound (UCB) algorithm within the framework of multi-armed bandit problems. This characteristic simplifies subsequent statistical inference processes; for example, traditional statistical estimators still maintain asymptotic normality, even when the collected data does not follow an independent and identically distributed (iid) pattern. Additionally, we demonstrate that this stability holds when the number of arms (K) is large—and even when K is permitted to increase alongside the total number of arm selections. Finally, our results indicate that the UCB algorithm exhibits fairness: specifically, if the mean rewards of two arms are similar, the UCB algorithm will choose each of these two arms an equal number of times over the long term (asymptotically). While these findings mark an important advancement, there are still several unresolved questions. Future studies could extend these results to apply to heavy-tailed distributions, explore the stability characteristics of other widely used bandit algorithms, and analyse potential trade-offs between stability and regret reduction. In other words, it would be valuable to investigate whether we can confirm the stability of other popular bandit algorithms or develop stable alternatives without causing a significant rise in regret—an exploration that could be further examined in research on "Inference with UCB." We argue that establishing the stability of reinforcement learning algorithms will enhance the reliability and reproducibility of reinforcement learning systems in real-world applications [4].

2.3 Thompson Sampling algorithm

The Thompson Sampling (TS) algorithm is a decision-making method rooted in Bayesian principles. Its core logic lies in randomly selecting each "arm" based on the probability that the arm performs optimally. Unlike complete Bayesian approaches such as the Gittins index, Thompson Sampling can usually be implemented with high efficiency. A possible reason for its relatively low popularity in the past is the insufficient theoretical analysis supporting it [5]. First proposed in 1933 (Thompson, 1933; Thompson, 1935), Thompson Sampling was initially designed to allocate experimental resources in two-armed bandit scenarios, particularly those encountered in clinical trials. For a long time after its proposal, this algorithm received little attention in academic circles—though during this period, it was independently reinvented on multiple occasions (Wyatt, 1997; Strens, 2000) and recognized as an effective heuristic method. Today, more than 80 years since its introduction, Thompson Sampling has gained a sudden and significant increase in interest from both industry practitioners and academic researchers [6]. The main execution step of TS algorithm is shown below:

For each arm $i\ (i = 1,2,...,K)$, set the initial prior distribution of its "success probability θ i". Since the Bernoulli distribution and the Beta distribution are conjugate distributions (the posterior distribution remains a Beta distribution, which is convenient for calculation), the Beta distribution is usually chosen as the prior.

The i-th arm's initial prior distribution isθ_i~Beta ($\partial_{i,0}, \beta_{i,0}$). Among this,$\partial_{i,0}$ (the count of successful first attempts) and $\beta_{i,0}$ (the count of failure first attempts) are hyper-parameters, and are normally equals to 1. From the first turn on:

1. Sample reward expectation estimation value.
2. Choose the best arm and then obverse the rewards.
3. Test the distribution after renewing the data.
4. Repeat the three steps above.

To sum up, The TS algorithm models uncertainty as a probability distribution through the "Bayesian update + random sampling" approach, implicitly balancing exploration and exploitation.

3 Application of multi-armed bandit

One practical variant of the Multi-Armed Bandit (MAB) framework is the contextual multi-armed bandit (CMAB), also known as the contextual bandit problem. In this setup, during each decision cycle, the agent first observes an N-dimensional context—often referred to as a feature vector—before selecting an arm. To determine which arm to activate in the current iteration, the agent leverages this observed context along with the reward data collected from previously chosen arms. Over the course of multiple iterations, the agent' s key objective is to gather sufficient information about the connection between context vectors and corresponding rewards. This accumulated knowledge enables the agent to make accurate predictions about which arm will yield the optimal outcome in subsequent decision-making steps. [7].

The MAB model has many applications in daily life:

1. Online advertising and marketing: Online advertisement need to maximize the revenue from advertising placement with the limited sources. For instance, a platform needs to choose A, B, C three advertisements for a same position. It is hard to make sure which one is the best. MAB model is a good choice to solve the problem.
2. Recommendation system: Recommendation system has to match users with the content they may be interested in with the highest possibility. MAB model can solve the problems of "cold start for new content" and "dynamic changes in user interests".
3. Clinical trials and medical decision-making: In this area, MAB model can be used for optimizing the design of clinical trial minimizing ineffective treatment for patients.
4. According to S Maghhsudi, MAB can also be used in 5G, MAB modeling brings many advantages to the research in the field of wireless networking and communication; some of them are shown in the follow.

The Multi-Armed Bandit (MAB) model inherently contains uncertainty, which stems from the lack of prior knowledge and severely limited feedback. This characteristic aligns with an inherent, fundamental feature of future distributed wireless networks—providing nodes with information in such networks would lead to excessive feedback and high overhead costs. In recent years, the MAB framework has attracted considerable attention, and this growing interest has made it feasible to introduce the concept of "conflict" into the traditional single-agent MAB model. The multi-agent MAB models derived from this can be effectively applied to address wireless networking issues, such as implementing distributed resource management or solving interference coordination problems in a distributed way. Additionally, the MAB model has various variants, and this flexibility allows it to adapt to diverse conditions in wireless networks. For example, it can accommodate scenarios with different levels of information availability and different types of randomness [8].

4 Conclusion

This article mainly talks about MAB model and some of its classical algorithms. ETC is a simpler algorithm than the others. However, it always takes more resources and produces more regrets which means it cannot be the first choice in most situations. UCB is the most popular algorithm nowadays as its precise logic. It is now applicated in many fields today. TS is a complicated algorithm. Few theoretical analyses mean less application in life. With more study on it, TS is getting more and more widespread these years. MAB model is suitable in all kinds of situations as the article talked about above. Taking good use of them will surely benefits the social in varied situations from daily life to scientific research or even in marketing

References

1. A. Slivkins, Introduction to multi-armed bandits. Found. Trends Mach. Learn. 12, 1–286 (2019).
2. A. Vahidi, Risk-seeking multi-armed bandits in an explore-then-commit setting (2025).
3. A. Garivier, T. Lattimore, E. Kaufmann, On explore-then-commit strategies, in Adv. Neural Inf. Process. Syst. 29 (2016).
4. K. Khamaru, C.-H. Zhang, Inference with the upper confidence bound algorithm. arXiv:2408.04595 (2024).
5. O. Chapelle, L. Li, An empirical evaluation of Thompson sampling, in Adv. Neural Inf. Process. Syst. 24 (2011).
6. D.J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, A tutorial on Thompson sampling (Now Publishers, 2018). https://www.nowpublishers.com
7. 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).
8. S. Maghsudi, E. Hossain, Multi-armed bandits with application to 5G small cells. IEEE Wirel. Commun. 23, 64–73 (2016).