Multi-Armed Bandit Algorithms: Innovations and Applications in Dynamic Environments

Lei An *

College of science, Da Lian Jiao Tong University, Da Lian, 116028, China

* Corresponding Author Email: anlei@djtu.edu.cn

Abstract. This paper delves into the fundamental concept of the Multi-Armed Bandit (MAB) problem, structuring its analysis around two primary phases. The initial phase, exploration, is dedicated to investigating the potential rewards of each arm. Subsequently, the exploitation phase utilizes insights from exploration to maximize returns. The discussion then progresses to elucidate the core methodologies and workflows of three principal MAB algorithms: Upper Confidence Bound (UCB), Thompson Sampling, and Epsilon-Greedy. These algorithms are meticulously analyzed for their unique approaches and efficiencies in handling the MAB problem. Expanding the scope further, the paper spotlights three practical applications of MAB algorithms. The first application involves Dynamic Resource Allocation in Multi-Unmanned Aerial Vehicle (UAV) Air-Ground Networks, leveraging the K-armed Bandit framework. This is followed by an exploration of Product Pricing Algorithms grounded in MAB principles, offering innovative solutions for dynamic pricing strategies. Lastly, the paper examines a cost-effective MAB algorithm tailored for dense wireless networks, addressing the complexities and demands of modern network infrastructures. This comprehensive study not only highlights the versatility of MAB algorithms but also underscores their growing importance in diverse real-world applications.

Keywords: Multi-Armed Bandit Algorithms, Upper Confidence Bound (UCB), Thompson Sampling, Epsilon-Greedy, K-armed Bandit framework.

1. Introduction

The Multi-Armed Bandit problem presents a scenario akin to a hypothetical gambler visiting a casino filled with numerous slot machines, each with a distinct probability of yielding a win. This problem poses the question: in the absence of prior knowledge about the winning odds of each machine and with a limited number of attempts, how should one play to maximize returns [1]. The nomenclature 'multi-armed bandit' stems from the visual and functional aspects of traditional slot machines. Historically, these machines featured levers resembling arms, and the often disappointing outcomes of gambling were likened to being robbed by a bandit. In the MAB problem, individuals confront multiple such 'slot machines', each representing a unique choice with uncertain outcomes [2]. In this complex problem, decision-making is a critical challenge, involving a strategic balance between 'exploration', where various options are tested to gather information, and 'exploitation', where decisions are based on accumulated knowledge to optimize outcomes. This dilemma is central to various fields, such as economics, computer science, and psychology, reflecting the universal nature of decision-making under uncertainty. The MAB problem has evolved into a rich field of study, with numerous variations and extensions being explored. These include contextual bandits, where additional information influences the decision process, and multi-agent bandits, involving several decision-makers with potentially competing interests. The applications of MAB algorithms are equally diverse, ranging from optimizing online advertising strategies to personalized medical treatment plans, showcasing their adaptability and relevance in addressing real-world problems. This exploration into the depths of the MAB problem not only provides insights into optimal decision-making strategies but also contributes significantly to the broader understanding of managing uncertainty in various domains.

2. Theoretical Foundations and Historical Development

2.1. Fundamental Concepts of Multi-Armed Bandit Problems

The meaning of Multi-armed Bandit problem is that a gambler goes to a casino to play a multi-armed bandit that has k arms. And how can make the largest reward? In order to solve the problem there are two stages which are exploration and exploitation [3]. In the stage of exploration the gambler try every machine record the reward and calculate the average profit and determine the arm with the highest yield, which called A. In the stage of exploitation try to only use arm A and calculate the rewards. That is the MAB problem.

2.2. Review of Existing Algorithms and Their Comparative Analysis

In the current landscape of algorithmic decision-making, three prominent algorithms stand out: Upper Confidence Bound, Thompson Sampling, and Epsilon-Greedy [4]. These algorithms are renowned for their efficacy in resolving the complexities inherent in the Multi-Armed Bandit problem. UCB excels in balancing exploration and exploitation by considering the uncertainty in the estimate of each arm's reward. Thompson Sampling, meanwhile, adopts a probabilistic approach, dynamically adjusting the selection strategy based on observed outcomes. The Epsilon-Greedy algorithm, known for its simplicity, periodically explores alternative options while generally exploiting the best-known arm. Each of these algorithms demonstrates unique strengths in various scenarios, from online content recommendation to adaptive clinical trials. For instance, UCB's precision in estimating rewards makes it suitable for environments where accurate predictions are crucial. Thompson Sampling's adaptability is particularly advantageous in dynamic environments where reward probabilities change over time. Epsilon-Greedy, with its straightforward implementation, offers a robust solution in situations where computational resources are limited.

2.2.1. UCB

UCB Score each choice by calculating the upper bound of the confidence interval:

$$\text{Sorcea} = \overline{\mu_a} + \sqrt{\frac{2\log T}{n_a}} \tag{1}$$

In this formula, $\overline{\mu_a}$ means the average reward of choice ‘a’, which shows the effect of ‘a’ has been observed so far. The $\sqrt{\frac{2\log T}{n_a}}$ is upper boundary of the confidence interval, which reflects the uncertainty of the effect of choice ‘a’. And T shows the total attempts na means selecting the number of attempts for ‘a’ [5].

And the algorithm flow is as follows:

Initialize: let us try each option once and get the initial score for each option.

For every time try t in [T]:

Try the highest selection for the current score. (If you have more than one, just pick one at random) Re-score each choice and update the score.

However, UCB algorithm also has some disadvantages:

The algorithm is deterministic, and the score does not change until a new attempt is fed back in.

Failure to incorporate prior knowledge, such as knowing in advance that certain choices are better.

2.2.2. Thompson Sampling

In the Thompson Sampling algorithm uses the distribution called beta, which assigns a score to each choice, using α shows the number of times each choice is reward after trying [6]. And the algorithm uses β showing the number of times each option failed to get a payoff after trying. So α+β are the total number of attempts for a selection.

For each choice ‘a’ the payoff probability can be described as (αa, βb).

And the algorithm flow is as follows:

Initialize: given a prior distribution beta (αat, βbt) for each choice A, based on prior knowledge.

For every time try t in [T]:

1 A sample θat is randomly drawn in each selected distribution beta, so getting {θ1θ2……θk} as the score of each choice.

2 Try the choice with the highest current score. (If you have more than one, just pick one at random)

3 Update the beta distribution selected by the attempt based on the attempt feedback [7]:

If get reward, beta (αat, βbt) ⟶beta (αat+1, βbt)

If not get reward, beta (αat, βbt)⟶ beta (αat, βbt+1)

First, explore each option equally, and then uniformly the highest-earning option. Epsilon greedy algorithm will use the probability ε to determine whether to explore or exploit in each round of attempts, that is, alternating exploration and exploitation in the process of trying [8]. The algorithm flow is as follows:

Initialize a probability of ε∈ [0, 1]

For every time try t in [T]:

Generate a random number i ∈ [0, 1]

If i<ε:

Exploration: try a choice at random

Else:

Exploitation: try the option with the highest average yield currently observed â (If you have more than one, just pick one at random).

If the choice has a wider distribution of returns, more exploration is required, corresponding to a larger ε, and when the choice has a more concentrated distribution, a small number of attempts can estimate returns well, only a small ε is required [9]. Typically, the ε is set to a smaller number, such as 0.1 or 0.01.

3. Problem Analysis and Application in Diverse Fields

3.1. Optimizing Dynamic Resource Allocation in UAV Air-Ground Networks

With the rapid development of drones, more and more people are using drones for entertainment and work. This makes communication between the drone and the pilot on the ground a very important problem. Most drones fly beyond the visual range, and the pilot can only observe the object in front of the drone from the first perspective through the screen. If there is a problem with the communication between the ground and the sky so far, the pilot will not avoid the obstacle in time, which will lead to the crash of the drone. Reference was studied how to use multi-arm Bandit problem to solve the problem of dynamic contact between UAV and ground [10].

In traditional K-arm bandit model, and the problem is modeled as a Markov process for solving. This model can be represented by a 4-tupleΦ= {M, S, A, R}, where M represents the number of agents, S represents the system state, A represents the joint action of multiple agents, and R={R1(s(t), a(t)), …, RM(s(t), a(t))} represents the return function selected by the m-th agent in the state.

User selection and power allocation algorithm of episodic multi-agent multi-state K-arm gambling machine is follow:

Initialization: Explore parameters, maximum number of training acts Nepi, state-action value function Qm1(s, a) = 0, ∀m ∈ M

For all drones, the initial state is given sm (0);

Nepi=Nepi-1

Loop t=1, 2… T, for each step in the screen, and each drone performs the following steps independently

Select Action am (t) according to policy πmε;

Perform action am (t), get immediate return Rm (t+1), status change to sm (t+1);

Update status - action value function Qmt+1(s, a) = Qmt (s, a) +α (Rmt - Qmt (s, a));

Repeat steps (1) - (3) until Nepi=0.

Through the above algorithm, it can be concluded that the multi-arm slot machine problem can solve the dynamic resource allocation problem of UAV air-ground network.

3.2. Adaptive Product Pricing Strategies

In recent years, the number of online retailers has increased significantly. Online retail can sell a variety of goods at the same time, and shopping can be carried out at any time and place without time constraints. In order to attract consumers to spend, it is important for retailers to price their products. But at the same time, online retailers face the phenomenon of incomplete user information, such as: consumer demand. Literature studies how to use the multi-arm slot machine to solve the pricing problem of goods in the case of incomplete user information.

The price to be selected is the interval {pt… pm} of the dispersion, and the price is regarded as the handle in the MAB problem. When $t \leq K$, the price pt is selected as the current pricing; when $t \geq K + 1$, the product is priced according to the strategy Ψ, and the quantity of products sold in the current period is known after the end of each period, that is, the reward value in the MAB model. In the MAB model, the expected cumulative regret value is used to measure the performance of the algorithm, that is, the expectation of the difference between the actual selected handle and the cumulative return obtained by consistently selecting the optimal handle, and represents the loss caused by not selecting the optimal handle. The goal of the MAB model is to minimize the cumulative regret value, which is equivalent to maximizing the cumulative profit value in the pricing problem under incomplete demand information.

Through the above modeling process, it can be seen that the MAB problem is of great help to the commodity pricing problem, so that the commodity pricing problem can be solved by algorithm.

3.3. Minimizing Loss in Dense Wireless Networks

With the development of The Times, the means of communication have also made great progress. Wireless networks have become an essential part of our daily lives. In daily life, we need the Internet anytime and anywhere, we use 5G technology to surf the Internet outdoors, and we use WIFI to surf the Internet indoors. The network signal that our mobile phone can receive is sent by the base station, when the distance between the terminal device and the base station is close, the network signal is better, but when the distance is far, the signal will become worse. This has led to intensive construction of base stations to ensure the quality of network signals. This will cause frequent network switching of the terminal, and energy consumption will be generated in the process of network switching. Literature studies how to use the multi-arm slot machine problem to improve the efficiency of user access to the network [10].

This paper presents a low complexity wireless network user access algorithm based on the multi-arm slot machine model. The effectiveness and robustness of the proposed algorithm are verified by 3 groups of comparative experiments from different angles, which provides a solution for the design of user access system in the next generation wireless communication network. This algorithm not only reduces the number of times that users trigger network switching, but also ensures that the performance of regret value will not be affected. This paper mainly considers the solution under stable environment, that is, the probability distribution of the base station generating the reward value is constant. However, as noted in the previous discussion, when users are faced with a dynamic network environment, they need to frequently restart the learning process. The algorithm proposed in this paper provides a good basis for deployment and use in dynamic environment.

Through the above introduction, it can be understood that the multi-arm slot machine problem can be a great help to solve the problem of low energy consumption of infinite dense networks.

4. Conclusion

This paper presents a thorough introduction to the fundamental mechanics and algorithms of the multi-armed bandit (MAB) problem, designed to enhance readers' comprehension of this complex

issue. It delves into the intricacies of the MAB process, elucidating how choices are made when faced with multiple options, each with uncertain rewards. To bridge the gap between theory and practice, the paper includes a range of real-world examples, demonstrating the MAB's applications in everyday scenarios. These examples serve to demystify the mathematical complexities inherent in the MAB problem, making it more accessible and relatable. By showcasing how MAB algorithms can be applied to solve practical problems in various fields - from optimizing online advertising strategies to enhancing decision-making in clinical trials - this paper aims to illustrate the real-life relevance and utility of these seemingly abstract mathematical concepts. This comprehensive approach not only fosters a deeper understanding of the MAB problem but also highlights its significance and versatility in addressing real-world challenges.

References

[1] Takeuchi, S., Hasegawa, M., Kanno, K., Uchida, A., Chauvet, N., & Naruse, M. (2020). Dynamic channel selection in wireless communications via a multi-armed bandit algorithm using laser chaos time series. Scientific reports, 10 (1), 1574.

[2] Chen, S., Tao, Y., Yu, D., Li, F., & Gong, B. (2021). Distributed learning dynamics of multi-armed bandits for edge intelligence. Journal of Systems Architecture, 114, 101919.

[3] De Curtò, J., de Zarzà, I., Roig, G., Cano, J. C., Manzoni, P., & Calafate, C. T. (2023). LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. Electronics, 12(13), 2814.

[4] Zhu, X., Zhao, Z., Wei, X., & others. (2021). Action recognition method based on wavelet transform and neural network in wireless network. In 2021 5th International Conference on Digital Signal Processing (pp. 60-65).

[5] Gonçalves, R. A., de Almeida, C. P., Venske, S. M. G. S., da Silva, M. R. D. B., & Pozo, A. T. R. (2017, October). A new hyper-heuristic based on a restless multi-armed bandit for multi-objective optimization. In 2017 Brazilian Conference on Intelligent Systems (BRACIS) (pp. 390-395). IEEE.

[6] Shi, Y., Cui, Q., Ni, W., & Fei, Z. (2020). Proactive dynamic channel selection based on multi-armed bandit learning for 5G NR-U. IEEE Access, 8, 196363-196374.

[7] Yan, C., Han, H., Zhang, Y., Zhu, D., & Wan, Y. (2022). Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. Knowledge-Based Systems, 257, 109927.

[8] GAO, S., Yang, T., Ni, H., & Zhang, G. (2020, August). Multi-armed bandits scheme for tasks offloading in MEC-enabled maritime communication networks. In 2020 IEEE/CIC International Conference on Communications in China (ICCC) (pp. 232-237). IEEE.

[9] Kamikokuryo, K., Haga, T., Venture, G., & Hernandez, V. (2022). Adversarial Autoencoder and Multi-Armed Bandit for Dynamic Difficulty Adjustment in Immersive Virtual Reality for Rehabilitation: Application to Hand Movement. Sensors, 22(12), 4499.

[10] Zhu, J., Song, Y., Jiang, D., & Song, H. (2016). Multi-armed bandit channel access scheme with cognitive radio technology in wireless sensor networks for the internet of things. IEEE access, 4, 4609-4617.