

Utilizing Reinforcement Learning Bandit Algorithms in Advertising Optimization

Shuning Zhang *

Aien Institute, Shanghai Ocean University/University of Tasmania, Shanghai, 201306, China

* Corresponding Author Email: snzhang@utas.edu.au

Abstract. This research provides a comprehensive analysis of the application of Multi-Armed Bandit (MAB) algorithms in the field of advertising, particularly highlighting the crucial balance between exploration and exploitation strategies. The implementation of MAB algorithms, especially within the framework of reinforcement learning, introduces a dynamic approach to optimizing advertisement placements and mixtures. This paper conducts a critical review of traditional advertising technologies such as rule engines and keyword targeting, drawing a comparison with more advanced techniques like the Explore-Then-Commit (ETC) algorithm and Deep Q-Networks (DQN). The study pays particular attention to the challenges inherent in integrating these algorithms. These challenges include managing the delicate exploration-exploitation equilibrium, amalgamating MAB algorithms with deep learning techniques, and addressing delays in user feedback. To address these issues, the paper proposes novel solutions like intelligent exploration strategies, the implementation of real-time updates, and the development of scalable algorithms. In conclusion, the paper asserts that the synergy of MAB algorithms with deep learning has the potential to substantially improve the efficiency and effectiveness of advertising systems. This integration facilitates more personalized and intelligent decision-making in ad delivery, representing a significant advancement over conventional advertising approaches.

Keywords: Multi-Armed Bandit Algorithms; Advertising Delivery; Reinforcement Learning; Exploration-Exploitation Balance.

1. Introduction

In the contemporary digital landscape, advertising is a pivotal strategy for businesses to market their products and services. However, escalating market competition and evolving consumer demands pose fresh challenges to traditional advertising methods. To achieve more precise and efficient ad delivery, the focus has shifted towards reinforcement learning and its integration with the multi-armed bandit model in the advertising sector. This paper delves into the application of the multi-armed bandit algorithm within the context of reinforcement learning in advertising. It concentrates on critical technologies involving the balance of exploration and exploitation, integration with deep learning, and the challenge of user feedback latency [1]. Through a comprehensive examination of this domain, this study aims to uncover how reinforcement learning can enhance the sophistication of ad delivery systems. It also seeks to provide innovative insights and methodologies for businesses to gain a competitive edge in the market.

1.1. Reinforcement learning and multi-arm bandit algorithm

Reinforcement learning is a machine learning approach that learns the best behavior strategy through the interaction of an agent with the environment to maximize cumulative rewards in an unknown environment. The agent performs an action, the environment returns an immediate reward, and the agent adjusts its strategy through trial and error, balancing exploration and exploitation to optimize decisions in a given state [2]. The multi-armed robber algorithm is a special algorithm in the field of reinforcement learning, which is used to solve resource allocation and decision problems. The algorithm takes its name from the casino slot machine (one-arm bandit logic), in which the player faces multiple slot machines, each of which is an "arm", and the algorithm's task is to choose among these arms to maximize the accumulated reward.

Unlike general reinforcement learning problems, however, the multiple-armed robber problem usually does not involve changes in the state of the environment, but instead emphasizes making informed choices without knowing the distribution of rewards for each arm [3]. This involves a tradeoff between exploration and utilization, namely, finding the best arms in the shortest possible time (exploration) and choosing those arms that have performed well in the past more frequently in subsequent choices (exploitation). The algorithm needs to dynamically adjust the selection strategy of the pair arm through continuous experiment and learning, so as to gradually converge to the best strategy. This process is closely related to the learning and decision-making process in reinforcement learning but emphasizes the ability to make optimal decisions in the face of uncertainty and lack of prior information.

1.2. Application of multi-arm bandit model in advertising

In the field of advertising, the multi-arm robber algorithm can be used to dynamically adjust the advertising strategy to maximize the advertising effect. Advertising requires making decisions with limited advertising resources to select the ads that are most suitable for displaying to users in order to achieve the maximum click-through rate or other business objectives [4]. This is similar to the multiple-armed bandit problem, where each AD can be seen as an "arm" and the AD click-through rate acts as a reward signal. The multi-armed Bandit algorithm can be adapted to a variety of practical applications in advertising, such as online advertising platforms such as Google and Facebook, or shopping sites with promotional activities to use dynamic pricing, and personalized recommendation ads like TikTok.

1.3. Process of advertising

In the application of multi-arm bandit algorithm, there are four main entities, namely users, advertising Spaces, advertisements and click records. The formal definition of personalized advertising is as follows:

K: The set of all users.

S: A collection of all possible AD placements, such as placement on a web page, in-app AD placements, etc.

A: A collection of all possible ads, including the attributes of the AD such as ID, type, target audience, etc.

U: a utility function that measures the click utility of A user $k \in K$ on an AD $a \in A$ on an AD space $s \in S$, i.e., $U: K \times S \times A \rightarrow T$, where T is a perfectly ordered set.

Therefore, when the user $k \in K$, we want to choose the AD space $s' \in S$ and the AD $a' \in A$, so that the goal is to maximize the user's click utility.

That is: $(s, a) = \arg \max U(c, s, a), s \in S, a \in A$.

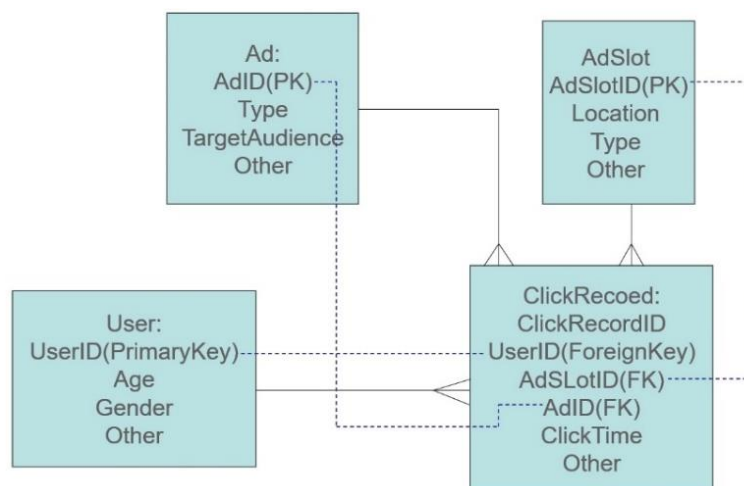


Fig 1. ER chart of advertising delivery (Photo/Picture credit: Original).

Under the multiple-arm robber algorithm, we can use the click record as the feedback of the algorithm. Click history usually indicates whether a user clicked on an AD on a certain AD space. As shown in Fig 1. This can be represented by a binary variable ($r_{\{k, s, a\}}$), where ($r_{\{k, s, a\}} = 1$) means that user k clicked on AD a on AD space s , and ($r_{\{k, s, a\}} = 0$) means that user K did not click. The goal is to maximize cumulative click effectiveness through experimental placement on different AD Spaces and ads. The algorithm strikes a balance between exploration (trying new AD Spaces and ads) and exploitation (selecting known efficient AD Spaces and ads) to maximize long-term click utility. This can be done by using algorithms such as the Bayesian multi-armed bandit algorithm to dynamically adjust strategies and maximize click-through revenue. Bayesian multi-armed bandit algorithm:

$$(s_t, a_t) = \arg \max_{s \in S, a \in A} \left(\theta_{s,a} + \sqrt{\frac{2 \ln(t)}{n_{s,a}}} \right) \quad (1)$$

Where (s_t, a_t) is the AD placement and AD combination selected at time t , $\theta_{s,a}$ is the real click-through rate of AD a on AD placement s , and $n_{s,a}$ is the number of times AD a has been selected on AD placement s . This formula expresses the tradeoff between the estimated click-through rate and the degree of exploration to find the best AD space and AD mix within a limited number of runs.

2. Technology of advertisement delivery

2.1. Traditional technology

Traditional AD delivery technology mainly includes rule engine, keyword targeting, behavior targeting, targeted AD network and A/B testing. A rule engine is a rule-based system that realizes the orderly display of advertisements by defining and executing the rules of advertising delivery, such as advertising orientation, delivery time, frequency control, etc. Keyword targeting is a common search engine advertising technique that uses keywords in a user's search query to match an advertisement. Behavioral targeting is based on a user's past behavior data, such as browsing history and click history, to more accurately target ads. Targeted AD networks allow advertisers to place ads on specific websites or apps to ensure exposure to their target audience. Finally, A/B testing is a traditional technique of experimenting with different AD ideas or delivery strategies to determine which is more effective.

2.2. Deep Technology

A The technique of the 'explore and exploit' strategy

The Explore-Then-Commit algorithm is an interesting and effective exploration-exploit strategy in deep reinforcement learning advertising. The core idea of the algorithm is that during a certain exploration period, the advertising system will randomly select the advertisement with a certain probability in order to obtain more information about the effect of the advertisement. At the end of the exploratory period, the system will select the current AD strategy it believes is best based on the collected data, and focus on this strategy for the rest of the delivery process [5].

The probability of AD selection during exploration can be expressed by the following formula:

$$P(a_i) = 1/K \quad (1)$$

Where $P(a_i)$ is the probability of selecting AD i and K is the total number of ads. This indicates that each AD has an equal probability of being selected during the exploration period. The reward collection for the exploration period is to record the number of clicks and conversions for each AD during the exploration period in order to evaluate the performance of the AD in subsequent decisions. The decision rule of the ETC algorithm is to calculate the performance metrics of each AD, such as click-through rate or conversion rate, at the end of the exploration period. Then, choose the best performing ads as your primary delivery strategy.

Here is an example:

Consider an advertising system with 3 ads ($K = 3$). During the exploration period, each ad is chosen with equal probability ($P(a_1) = P(a_2) = P(a_3) = 1/3$). After the exploration period, assuming that Ad 1 has the highest click-through rate, the system decides to focus on Ad 1 in the subsequent delivery period.

The process can be shown in the Table 1:

Table 1. Performance Metrics of Advertising Campaigns.

ADs	clicks	conversions	click-through rates
1	100	10	0.1
2	80	8	0.1
3	60	6	0.1

At the end of the exploration period, the system selects AD 1 as the primary delivery strategy. In the subsequent launch period, the probability of selecting AD 1 is 1, while the probability of selecting other ads is 0. This approach can help the system fully explore in the initial phase, understand the performance of each AD, and focus on the optimal advertising strategy in subsequent decisions, thereby improving the overall advertising effectiveness.

B Deep Q network: In advertising, the multi-armed robber model of deep learning combined with reinforcement learning involves the use of deep neural networks to estimate the value function of advertising strategies. A classic application of Deep technology is the use of a Deep Q-Network (DQN) to optimize advertising decisions.

DQN is a deep reinforcement learning algorithm, which is widely used to solve the multi-armed robber problem. In AD delivery, each AD strategy can be seen as an action, and the goal of DQN is to learn a value function that represents the long-term cumulative reward for each AD strategy. DQN uses the idea of Q-learning, and its update rules are:

$$Q(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot (r + \gamma \cdot \max_{a'} Q(s', a'))$$

Where: $Q(s, a)$ is the value function of selecting action a under states,

α is the learning rate

r is the instant reward you get when states selects Action a ,

γ is the discount factor,

s' is to move on to the next too.

Deep Q Networks (DQN) include the following key principles [6]:

State Representation: In advertising, the states is provided by a deep learning model, which usually contains user characteristics, AD placement information, and so on. This can be done through a deep neural network that maps incoming ads and user information into state representations.

Action Selection: Each advertising strategy is considered an Action a . The goal of DQN is to learn the Q value of each action, which is the potential long-term cumulative reward for choosing an advertising strategy. The selection of actions is carried out by maximizing Q value, which is, selecting the advertising strategy with the highest probability.

The Reward Signal: Instant Reward r is used to represent the advertising effect obtained after States selects action a , such as user clicks or conversions. This is the feedback signal provided by the advertising system to evaluate the effectiveness of each advertising strategy.

Learn and update: DQN updates its value function by constantly learning the Q value. The learning rate α controls the updating amplitude of the model parameters. The deep learning model is responsible for estimating the Q -value function, where the weights of the deep neural network act as the parameters of the Q network. Over time, networks continue to adapt to new information, improving accurate estimates of the value of advertising strategies.

3. Challenges

Technology is always evolving, so there are always challenges. Here we will analyze the challenges of Exploration-Exploitation Balance, Integration with Deep Learning and Integration with Deep Learning.

3.1. Challenge on Exploration-Exploitation Balance

In advertising, there is a trade-off between exploring new advertising strategies and leveraging known strengths. Too much focus on exploration can lead to performance degradation in the short term, while too much focus on exploiting strategies that can make the system miss out on better. Bandit algorithms usually make initial AD selection based on prior knowledge (such as the prior distribution of CTR), but this knowledge is limited [7]. In practice, finding a balance in a changing environment to maximize long-term rewards is a complex optimization problem. In the long term, algorithms need to explore to find potentially high-grossing ads, but in the short term, they need to make the most of what they know.

3.2. Challenge on Integration with Deep Learning

When integrating deep learning model with multi-arm bandit algorithm, it is necessary to solve the complex interaction problem between deep learning training and updating and bandit algorithm decision-making. The deep learning model is used to estimate the value function of the advertising strategy to provide prior knowledge for the bandit algorithm. However, the training process and parameter updating of deep learning may need to take into account the decisions of the bandit algorithm to ensure consistency between the two. In addition, the training cycle of the deep learning model may be long, and it needs to be effectively combined with the real-time decision-making of the bandit algorithm [8].

3.3. Challenge on User Feedback Latency

There is a lag in user feedback, such as clicks or conversions, which can cause the system to not immediately get the most up-to-date information when adjusting advertising strategies. Advertising systems often rely on user behavior data to update advertising strategies. Due to the delay in user feedback, the system may not be able to immediately understand the effectiveness of the AD in some cases. This can result in algorithms being slower to respond to rapidly changing market trends or unexpected events. The delay of user feedback is a special problem that needs to be dealt with for the AD delivery system with high real-time requirement.

3.4. Insufficient

The challenge is the problem that will be faced in the future but compared with the current use of the multi-armed bandit model in the field of advertising, there are also some subtle shortcomings. The cold start problem and the lack of scalability are the shortcomings of multi-armed bandit algorithm in the field of advertising. The cold start problem is mainly manifested in the absence of historical data for new advertising strategies, which makes it impossible for algorithms to accurately evaluate their potential value in the early stages. Since traditional multi-armed bandit algorithms usually rely on historical observations to form estimates of strategy returns, for new strategies, failure to utilize prior information may lead to excessive exploration behavior and reduce performance in the short term. Solving the cold start problem requires innovative exploration strategies and a more effective use of prior knowledge.

Insufficient scalability, on the other hand, mainly involves a linear rise in algorithmic computational complexity as the number of advertising strategies increases. The challenges include increasing computing costs, increasing storage requirements, and increasing algorithmic complexity. As the number of ads increases, the multi-armed bandit algorithm needs to make more decisions and update parameters, which may lead to a decline in real-time performance, especially in large-scale

AD libraries [9, 10]. Solving the scalability problem requires algorithms to be optimized to reduce the compute and storage burden and ensure that the system can efficiently cope with the growing number of advertising strategies.

4. Conclusion

Addressing the aforementioned challenges, the multi-armed bandit algorithm in the realm of advertising can be enhanced through several innovative strategies. Intelligent exploration tactics, the judicious application of prior knowledge, real-time updates to deep learning models, and prompt management of user feedback latency are crucial for harmonizing the algorithm's short-term and long-term performance. Moreover, leveraging multi-source information fusion, alongside active and semi-supervised learning approaches, could provide more robust solutions to the cold start issue and fortify the algorithm's resilience when integrating new advertising campaigns.

To tackle scalability concerns, optimizing the computational complexity of the algorithm is essential. Employing distributed computing and approximation techniques ensures that the system maintains efficiency even when dealing with extensive advertising inventories. These cutting-edge solutions will enhance the adaptability and practicality of the multi-armed bandit algorithm in advertising, enabling systems to make more personalized and intelligent decisions, thus elevating advertising efficacy and user experience. In summary, the application of the multi-armed bandit algorithm within reinforcement learning has brought significant innovations to advertising delivery. By striking a fine balance between exploration and exploitation, incorporating deep learning, and adeptly managing user feedback delays, these algorithms equip advertising systems with an intelligent decision-making capability. This optimizes the impact of advertisements, ensuring resourceful and effective ad placements.

Reference

- [1] Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.
- [2] Zhang, W., Zhao, X., Zhao, L., Yin, D., Yang, G. H., & Beutel, A. (2020, July). Deep reinforcement learning for information retrieval: Fundamentals and advances. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2468-2471).
- [3] Khan, M., Jan, B., Farman, H., Ahmad, J., Farman, H., & Jan, Z. (2019). Deep learning methods and applications. *Deep learning: convergence to big data analytics*, 31-42.
- [4] Ferdowsi, A., Ali, S., Saad, W., & Mandayam, N. B. (2019). Cyber-physical security and safety of autonomous connected vehicles: Optimal control meets multi-armed bandit learning. *IEEE Transactions on Communications*, 67 (10), 7228-7244.
- [5] Aramayo, N., Schiappacasse, M., & Goic, M. (2023). A Multiarmed Bandit Approach for House Ads Recommendations. *Marketing Science*, 42(2), 271-292.
- [6] Wang, J., Gou, L., Shen, H. W., & Yang, H. (2018). Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE transactions on visualization and computer graphics*, 25(1), 288-298.
- [7] Jain, R., Nagrath, P., Raina, S. T., Prakash, P., & Thareja, A. (2021). ADS Optimization Using Reinforcement Learning. In *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 53-63). Springer Singapore.
- [8] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), 452-459.
- [9] Esfahaani, M. V., Xue, Y., & Setoodeh, P. (2021). Deep reinforcement learning-based product recommender for online advertising.
- [10] Zhao, X., L., Tang, J., & Yin, D. (2019). "Deep reinforcement learning for search, recommendation, and online advertising: a survey". *ACM sigweb newsletter*, 2019(Spring), 1-15.