Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/34/20230326

An investigation of progress related to stochastic stationary bandit algorithms

Yizhi Liu

Department of Computer Science, University of Wisconsin-Madison, Madison, Wisconsin, 53711, United States

yliu996@wisc.edu

Abstract. The Multi-armed Bandit algorithm stands as a consequential tool for informed decision-making, distinct from reliance on intuitive selections, given its systematic proclivity to meticulously assess accessible alternatives with the intent of discerning the most auspicious outcome. Amid the repertoire of algorithmic variations, the Stochastic Stationary Bandit algorithm assumes a foundational and enduring role, finding versatile application across diverse domains, including but not limited to digital advertising, price optimization, and recommendation systems. With these considerations in view, the present study embarks upon a comprehensive scrutiny of this subject matter. This paper reviews on the Explore-Then-Commit algorithm, Upper Confidence Bound algorithm, and Thompson Sampling algorithm by explaining, comparing their formulation, features, and expected results. Explore-Then-Commit algorithm has distinct phase to explore all the choices uniformly. Upper Confidence Bound algorithm make decisions by calculate an upper confidence index which is an overestimate for each choice. Thompson Sampling algorithm depends on randomness to make choices. Explore-Then-Commit algorithm faces the problem of when to explore and when to stop. Upper Confidence Bound algorithm and Thompson Sampling algorithm solve this problem by avoid certain phases. Multi-armed Bandit algorithm could deal with the process of displaying items of potential interest to users in a recommendation system, the delivery of resources in resource allocation, or the way to maximize revenue in a business.

Keywords: Muli-armed bandit, stochastic stationary bandit, explore-then-commit, upper confidence bound, Thompson sampling.

1. Introduction

Multi-armed Bandit algorithm, introduced by Robbins [1], represents a valuable algorithm for decision-making processes. Unlike relying on instinctive choices, the MAB algorithm is designed to systematically analyze available options to identify the most promising outcome. A sequential game between the learner and the uncertainty in making decisions is a multi-armed bandit problem. The rounds in this game are played repeatedly. The learner selects an action from a list of options in each round, and then they are rewarded. The objective of the learner is to maximize the cumulative reward over the entire process. In another way, the objective can also be stated as minimize the regret which is the reward lost by making sub-optimal actions. It can be computed as the difference between the

DOI: 10.54254/2755-2721/34/20230326

cumulative reward achieved by the best possible decision in the rounds played and the actual cumulative reward attained by the learner.

In the pursuit of maximizing reward, the learner endeavors to identify the optimal decision and insists on that choice. To ascertain the best optimal decision, the learner must initially explore all actions to gain information. When the learner selects a sub-optimal action, it may increase its regret. This iterative process of trial and error gives rise to a fundamental quandary known as the exploration-exploitation dilemma within the context of multi-armed bandit problems. The balance between choosing based on current results (exploitation) and choosing something uncertain and new (exploration) is essential in multi-armed bandit problems.

This research encompasses an investigation into diverse classifications of bandit problems, each characterized by distinct reward distribution properties. For instance, stochastic stationary bandit has a reward distribution which do not change over time. Non-stationary has a changing reward distribution. Structured bandit means that the rewards are distributed in a structured way. Contextual bandit would receive some contextual information about the environment before making an action. Correlated bandit means that the rewards of different actions are correlated with each other. This paper would focus on Stochastic stationary bandit since this bandit is traditional and basic.

In terms of Stochastic stationary bandit, there are different algorithms to minimize the regret, such as Explore-Then-Commit algorithm (ETC), Upper Confidence Bound algorithm (UCB), and Thompson sampling algorithm. The ETC algorithm would continuously choose the arm that performed the best during the exploration while initially exploring by choosing all arms for a predetermined amount of time. For this algorithm, one question would be how and how much to explore. Some theories consider doing the exploration in the beginning, while some theories propose to explore separately over the entire game. If one can choose the exploration phase properly, the regret might show a sublinear distribution. The UCB algorithm means that the learner uses current data to give each arm a value. The value represents an overestimate of the unknown mean and is known as the upper confidence bound. If this value is higher than optimal arm's value, the learner would explore this arm and the suboptimal arm's value would fall below the optimal arm in the end. There are also different versions of UCB algorithm. This UCB algorithm needs to know the horizon, the total number of trials, to perform. Another version of UCB algorithm, called asymptotically optimal UCB, modified the analysis and eliminate the need of horizon. The UCB algorithm has some advantages over ETC algorithm, for example, it may achieve lower regrets as well as require lesser information, such as the horizon. The Thompson sampling, also known as posterior sampling, mainly uses randomness to function. For each step, the algorithm would make a decision based on the current distribution of rewards from each action. After one action is made, its distribution of rewards is updated. Usually, the lesser one action is chosen, the greater its uncertain and dispersive it would be. In this way, the algorithm can explore unknown arms. The Thompson sampling differs with ETC and UCB algorithms by randomization. It often shows superior performance than those algorithms, while it can also show large variance between experiments. Since this algorithm can be applied to different fields, such as digital advertising, price optimization, and recommendation systems, this paper aims to conduct a review on this topic.

The remainder part of the paper would be organized as follows. In section 2, this paper will introduce and analyze these algorithms. In section 3, this paper would explore the application of multi-armed bandit in really life. Section 4 would discuss the conclusion and future improvements.

2. Method

2.1. Explore-then-commit algorithm

In situations where individuals are presented with multiple options without prior knowledge of their relative performance, the instinctive approach of A/B Testing is often employed to identify the best option. A/B Testing aims to explore all the choices uniformly. The learner would select each choice for a bunch of times to explore the performance of each choice. After the test, all the choices are

DOI: 10.54254/2755-2721/34/20230326

compared with each other to find the best performed one. It is believed that the best choice would perform the best for the rest of the trials, and it is optimal to stick to this choice. This is the formulation of ETC algorithm. This algorithm might be the most natural and widely used approach to human. However, it is shown that this algorithm which has distinct separation of exploration and exploitation would not be the optimal strategy to maximize the rewards [2]. If the expected reward for each arm differs greatly or the exploration phase is too long, this strategy may waste too much reward during exploration. Conversely, if the expected rewards are relatively similar or the exploration phase is excessively short, the ETC algorithm may lead to a loss of rewards by failing to select the optimal choice.

2.2. Upper confidence bound algorithm

In addition to ETC, UCB is another algorithm stands as another viable approach for facilitating decision-making processes. Compared to ETC which has clear phases, it can deal with exploration-exploitation dilemma more effectively. The UCB algorithm make decisions by calculate an upper confidence index which is an overestimate for each choice [3]. For each step, the algorithm would choose the one with the highest upper confidence index. The upper confidence index is affected by its expected reward given by past results and an algorithm defined bonus. The goal of this bonus is to explore the unknown or uncertain choices, which is exploration, and to avoid choose suboptimal choices, which is exploitation. Normally, the algorithm would choose the choice with the largest expected reward. After this move, the bonus of the selected choice would decrease while the unselected choices' bonus would increase. After a few rounds, the upper confidence index of the suboptimal choices would be higher than the current best choice. Then, the algorithm would explore this choice and update its expected reward and its bonus. Compared to the ETC algorithm, the UCB algorithm do not have certain exploration or exploitation phase. It continuously explores the uncertainty and avoid regrets according to the distribution of rewards from each choice. UCB algorithm is much more flexible in front of the exploration-exploitation dilemma.

2.3. Thompson sampling algorithm

Thompson Sampling (TS), also known as posterior sampling, is a random algorithm trying to minimize regrets. It is assumed that its reward distribution for each arm is prior distribution [4]. The prior distribution means that the distribution is updated after a data is obtained. For next step, updated new distribution, called posterior distribution, is used for decisions. In TS algorithm, the learner would first make a decision based on the current distribution of rewards from the choices. After it is done, the learner would update the current distribution using the new data from the step to get a new posterior distribution. For the choices unselected, theirs reward distributions are not changing.

There are many differences between UCB and TS. For example, UCB is deterministic. It calculates the value and choose the greatest one. The result of one step remains the same. Nevertheless, TS uses random values from the reward distributions. It is hard to get the same result by rerun the same step. Some research considered TS as a comparable or better algorithm in performance than UCB [5]. TS is considered more resistance to delayed feedback [5]. Since UCB is deterministic, it is required to update the algorithm to switch between explore and exploit. If UCB do not receive an update and stuck on a suboptimal choice, it would gain lots of regret. For TS, even if the value is not updated, it still has chance to switch between exploration and exploitation by randomness. It is shown that TS could gain a lower regret than a normal UCB algorithm. When However, UCB algorithm has different versions. TS does have advantages over some versions of UCB algorithms, while it could not be superior to all UCB algorithms. For example, AdaUCB could achieve lower regrets as well as lower variance than TS [6].

DOI: 10.54254/2755-2721/34/20230326

3. Applications and discussion

3.1. Recommendation system

In a recommendation system, the process of displaying items of potential interest to users can be effectively addressed by applying the multi-armed bandit approach [7]. For instance, there are several advertisements for a website. The company can only show one of the advertisements. Therefore, they can find out which one can be best used to attract users using multi-armed bandit. In addition to basic multi-armed bandits, several improvements were made on this recommendation system. If some of the company get access to the users' relevant information, they could make choices according to the information they got using contextual bandit. Since a user might not have the same interest forever, some researchers considered the dynamics of users and made new algorithms, which aware of user's changes [8]. The multi-armed bandit for recommendation system could be further improved by considering the rewards of each successful recommendation as well as better sensation or prediction of users' changed content.

3.2. Resource allocation

In a power system, the electricity might be unstable due to various situation. In order to keep stable power output, some of the less important electrical products could be turned off. With multi-armed bandit, the controller could deal with this situation, even when the number of electric products is changing [9]. In addition to the electricity output problem, a variety of wireless network problems can also be solved by multi-armed bandit, such as security, routing, and scheduling [10]. One of the possible limitations in this application is that multi-armed bandits consider only rewards. To enhance the efficiency of the multi-armed bandit approach, further improvements could be achieved by incorporating considerations of each arm's cost, availability, and other relevant constraints.

3.3. Business

In a biding, a buyer needs to find out a good amount to bid so that the price is higher than others and as low as possible. Multi-armed bandit could be used in operating the biding process. According to research, the usage of multi-armed bandit could significantly reduce the bidding cost [11]. In the future, more research could be conducted on which algorithm performs the best in the competition with other algorithms.

4. Conclusion

This review paper discussed and compared the algorithms for stochastic stationary bandit and their application. The explore-then-exploit, upper confidence bound, and Thompson sampling approaches are investigated in the context of multi-armed bandit problems, with a thorough examination of their respective advantages and limitations. This paper discussed the application of multi-armed bandit in real life and raised possible improvements for further studies. More research is needed in the future to systematically classify the different kinds of data, situation, and environment. Ultimately, identifying the most suitable multi-armed bandit approach for specific scenarios constitutes a key area warranting further exploration.

References

[1] Robbins H 1952 Some aspects of the sequential design of experiments Bulletin of the American Mathematical Society 58 (5): 527–535 doi:10.1090/S0002-9904-1952-09620-8

[2] Garivier Au Tor L and Emilie K 2016 On explore-then-commit strategies Advances in Neural Information Processing Systems 29

[3] Auer P et al 2002 Finite-time analysis of the multiarmed bandit problem Machine learning 47 235-256

[4] Agrawal S and Navin G 2012 Analysis of thompson sampling for the multi-armed bandit problem Conference on learning theory JMLR Workshop and Conference Proceedings

DOI: 10.54254/2755-2721/34/20230326

[5] Chapelle O et al 2011 An empirical evaluation of thompson sampling Advances in neural information processing systems 24
[6] Lattimore T and Csaba S 2020 Bandit algorithms Cambridge University Press
[7] Zhou Q et al 2017 Large-scale bandit approaches for recommender systems Neural Information Processing: 24th International Conference ICONIP 2017 Guangzhou China November 14-18 2017 Proceedings Part I 24 Springer International Publishing
[8] Bouneffouf D and Irina R 2019 A survey on practical applications of multi-armed and contextual bandits arXiv preprint arXiv:1904.10040
[9] Lesage L A et al 2017 The multi-armed bandit with stochastic plays IEEE Transactions on Automatic Control 63.7 2280-2286
[10] Maghsudi S and Ekram H 2016 Multi-armed bandits with application to 5G small cells IEEE Wireless Communications 23.3 64-73
[11] Tilli T and Leonardo E L 2021 Multi-armed bandits for bid shading in first-price real-time bidding auctions Journal of Intelligent & Fuzzy Systems 41.6 6111-6125