Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Navigating Complexity in Collaborative Environments through Innovations in Multi-Agent Multi-Armed Bandit Algorithms

Siqi Zhu*

Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China

*zhusq8@mail2.sysu.edu.cn

Abstract. As Multi-Armed Bandit (MAB) applications grow increasingly complex, particularly when multiple agents collaborate or compete, traditional bandit algorithms face fresh challenges, underscoring the rising importance of research in multi-agent multi-armed bandits (MAMAB). Developments in MAMAB algorithms have spurred significant advances across a variety of fields, addressing challenges in dynamic and uncertain environments. This paper offers an exhaustive review of recent progress in MAMAB algorithms, emphasizing major strides in enhancing cooperative decision-making and operational efficiency. Our focus is particularly on the contributions of Filippo Vannella et al., who have explored sample complexity within the MAMAB framework. Their research signals a shift towards optimizing global actions by minimizing sample complexity and harnesses mean field techniques in contexts such as optimizing wireless networks. Additionally, this paper addresses communication complexity, a crucial aspect of MAMAB systems, where numerous novel algorithms have been developed. These algorithms strike a balance between performance and communication overhead, diminishing the need for frequent and costly interactions among agents. In application terms, the incorporation of MAMAB algorithms in sectors like clinical trials and wireless network spectrum management showcases their potential to revolutionize conventional approaches. Through a detailed examination of current research trends and prospective future directions, this article contributes to the broader discourse on harnessing MAMAB algorithms to navigate the complexities of collaborative environments effectively.

Keywords: Multi-Agent Multi-Armed Bandit; Communication; Reward; Fairness.

1. Introduction

1.1. Definition of Multi-Agent Multi-Armed Bandits

The multi-armed bandit model, a cornerstone of machine learning, was first conceptualized in 1952 when Herbert Robbins introduced a convergent multiple-choice strategy. MAB represents an online sequential decision-making paradigm where an individual—be it an agent or client—is presented with numerous options (arms). The participant learns about the reward associated with an arm by selecting it, using historical reward data to guide future selections with the objective of maximizing rewards or minimizing cumulative regret. A critical aspect of this process is the need to balance exploration of untested options with exploitation of known rewarding ones. Given its broad applicability, the multi-armed bandit theory has spurred extensive research.

As application contexts grow increasingly intricate, numerous variants of MAB have emerged, among which the Multi-Agent Multi-Armed Bandit model, first proposed by Gai, Krishnamachari, and Jain in 2010, is especially pertinent in large-scale scenarios. With the advent of federated learning, interest in MAMAB has surged. This model involves multiple agents concurrently making decisions within the same setting, each striving to maximize its cumulative reward through a series of arm selections [1,2]. The principal challenge for agents in a MAMAB framework is to deduce the reward distribution of each arm from limited trials, optimizing overall performance by striking an optimal balance between exploration and exploitation. The complexity of MAMAB primarily arises from the interactions among agents, which can be either cooperative or competitive depending on the agents' objectives and the specificities of the application environment.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

1.2. Literature Review Protocol

This paper is a literature review on various research directions of MAMAB algorithm, based on applications in multi-agent scenarios. It shows the current hot research directions of MAMAB algorithm and the latest research achievements and progress.

The articles for this paper are searching from the CNKI Foreign Language Database and Google Scholar, and the search condition set the string, "Multi-Armed Bandit", combined with hot research directions, and the stings chosen for hot research directions including "Communication", "Sample", "Reward", " "Fairness", "Fair". After primary reading and removing irrelevant articles, book reviews, and consultations, a total of 70 articles were obtained. Ater that. Through deeper reading, several articles about the advanced achievements mainly in four aspects (sample complexity, communication efficiency, reward method, the balance between fairness and total rewards) are selected for detailed description in the next part of this paper.

1.3. Work in this paper

Although there have been many researches on MAMAB, none of them aims to describe the research directions and advanced achievements in this field. Therefore, this paper reviews MAMAB through the above-selected articles, covering four hot research directions: sample complexity, communication complexity, reward methods, and fairness and total rewards. It aims to provide the future development direction of the field.

In particular, the research by Filippo Vannella et al. demonstrated its potential in practical applications by minimizing sample complexity in the MAMAB setting, employing mean field techniques. P. Pankayaraj and D. H. S. Maithripala proposed a decentralized communication strategy that reduces the need for frequent and expensive interactions between agents [3, 4].

In terms of advanced application, the MAMAB model can be applied to traffic flow control and signal optimization in intelligent transportation systems, like the signal lights (intelligent agents) at crossroads learn to optimize traffic flow and reduce traffic jam. In the area of robot, the MAMAB model can be used to guide a group of robots (agents) on how to work together to complete specific tasks, such as search and rescue, environmental exploration, etc., and maximized the overall efficiency and effectiveness of the team at the same time. In the area of energy management and smart grids, the MAMAB model can help multiple distributed energy resources (intelligent agents) make independent decisions on how to regulate production or consumption to achieve stable contributions to the grid.

2. System Analysis and Application Research

2.1. Sample Complexity

The goal of studying the sample complexity of muti-agent multi-armed bandits (MAMB) is to find the minimum number of communications requires between multi-agent to achieve a better performance, where agents learn the possible rewards for each action by exploring the environment. So far, research results have shown the bounds of sample complexity in different setting, and some researches have successful declined the number of samples requires to achieve the best total result in cooperative environment by introducing efficient ways of information sharing and cooperative strategies. Besides, in terms of competitive environments, researches have already identified how competitive behavior increase the sample complexity and how to reduce the impact by algorithms designing.

Recently, the advance researches include Filippo Vannella et al. studying the challenge of identifying the best arm in Multi-Agent Multi-Armed Bandits, where rewards are defined by factor graphs to minimizing sample complexity. The author also uses the mean field (MF) technology to approximate the lower bound of sample complexity. According to the number of factors in the graph, the number

of possible actions of each agent, and the maximum degree of the factor graph, the author proposes the MF-TaS algorithm, which has a good performance on large scales.

Although significant has been made in this area, the sample complexity of multi-agent multi-armed bandits is still worth to research, and it also faces the challenges of designing algorithms that can complex interactions and minimizing the number of samples required to consider. In the future, work can be done in exploring more efficient cooperative methods, models that performance better in more complex competitive environment and some new algorithms. All of those improving directions are aiming to further reduce sample complexity in the learning process and maintain or improve performance at the same time.

2.2. Communication Complexity

The balance between communication complexity and efficiency is an important problem in the practical application of MAMAB algorithms, especially in scenarios with limited resources and frequent communication needs. In 2019, P. Pankayaraj et al. proposed a decentralized communication policy for the MAMAB problems. This policy allows agents to solve the MAMAB problem, individually and collectively address, based on their own sampling information and communication with neighbors. It provides theoretical support for agents to effectively use communication for exploration and exploitation in Decentration environments.

Recently, researches on communication complexity are mainly focused on how to reduce communication requirements and don't reduce performance of algorithm at the same time. The researches are mainly based on an Upper Confidence Based (UCB), and focused on the cooperative interactions between agents in the stochastic multi-armed bandit (SMAB) problem, and the aim of it is to minimize the accumulated regret and to reduce communication rounds and the bit size of each communication. Based on this method, Mridul Agarwal et al. introduced a variant called LCC-UCB-GRAPH [5]. This variant adjusts the communication policy for networks represented by sparse graphs, further optimizing algorithm performance and effectively addressing more complex scenarios. Udari Madhushani et al. explored the cooperative multi-armed bandit problem under practical communication limitations, like stochastic time-varying networks, stochastic delays and adversarial damage, including Byzantine Fault Tolerance (BFT) [6]. Due to the limitations of MAMAB algorithms that assume perfect communication, this research proposed Robust Communication Learning (RCL) algorithm， which can maintain competitive performance in the environment without the assumption of perfect communication. In addition, Yu-Zhen Janice Chen et al. proposed on-demand communication for multi-agent multi-armed bandits, which explored an on-demand communication (ODC) protocol that tailors the communication of each pair of agents based on their empirical pull times, and it effectively handles the asynchrony and heterogeneity of agent activities [7]. The ODC protocol significantly reduces unnecessary communication, especially when the pull rates of agents are highly inconsistent, without sacrificing the performance of the MAMAB algorithm. This research provides new perspectives on communication strategies in networks requiring Asynchronous Operations.

In general, current research directions in the field of multi-armed bandit communication focus on designing algorithms that can operate effectively in limited or constrained communication environments. These algorithms aim to achieve a balance between communication cost and performance, and the challenges, like randomness and dynamism also take into consideration.

2.3. Multi-Armed Bandit Reward Scenarios

Researches about MAMAB algorithm on the different reward methods has also been an important area. The research directions range from setting different conditions, like homogeneous rewards and heterogeneous rewards, to research under scenarios, like sub-Gaussian distribution, heavy-tailed distribution, and collision-related scenarios.

In the most basic MAB model, the reward distribution of each arm is fixed, and the reward expectation of all arms maybe the same. Jingxuan Zhu et al. proposed a new distributed algorithm for the distributed multi-armed bandit problem under homogeneous reward, based on the classic UCB1 algorithm and Metropolis algorithm [8]. This algorithm introduces Metropolis weights and new information transfer methos, which makes sure that each agent can effectively find the best arm even in a distributed environment.

In more complex MAB problems, the rewards for each arm may come from different distributions with different parameters, which are called heterogeneous reward distributions. This requires the agent not only to find which arm provides the best reward during exploration, but to evaluate the variability and uncertainty of each arm's reward. In 2023, Xu et al. researched the decentralized MAMAB problem for heterogeneous rewards under random graphs [9]. This research both focused on sub-Gaussian and sub-exponential reward distribution settings, and proposed an algorithmic framework using robust simulation, averaging-based consensus, novel weight methods, and UCB strategy, aims to minimize the total regret of the whole system through cooperation. This framework provided a solution based on UCB which also took graph stochasticity into consideration. This research solves the challenges of heterogeneous rewards and dynamic graph structures in multi-agent settings, improving the efficiency of decision-making in distributed systems.

In MAMAB problem, multi agents may choose the same arms in the same time which may cause collisions, especially in application scenarios such as wireless communications and network spectrum allocation. When two or more agents select the same arm, they may have influence on each other, causing the reward gained by all agents decreasing or even becoming zero. To solve this problem, researchers often design protocols to coordinate the behavior of each arm, or explore algorithms to automatically find the best way to distribute best arm for each agent to avoid collisions. Chengshuai Shi et al. researched a new stochastic multi-agent multi-armed bandit collision problem (MP-MAB), and proposed the Error-Correction Collision Communication (EC3) algorithm [10]. This algorithm models used a random coding error index to establish an optimal regret that cannot be surpassed by any communication protocol, which solved the problem of how to learn effectively and make decisions when players are unaware of collisions.

Many primary researches on MAMAB problem often assume the reward following a sub-Gaussian distribution (such as a normal distribution), and the tail probabilities often decrease quickly. In this assumption, there are very little extreme values of reward and it's relatively easy to estimate the arm's true reward expectation. In more advanced MAB models, researcher often assume rewards following heavy-tailed distributions, in which extreme reward values are more likely to occur. In 2020, Abhimanyu Dubey et al. discussed the cooperative decision-making problem in multi-agent systems and introduced a novel decentralized algorithm MP-UCB. This algorithm is specially for heavy-tailed reward distributions environments and adopted a robust estimation method and message passing protocol [11]. This algorithm can gain best group regret bounds under heavy-tailed conditions, solving the challenges of heavy-tailed reward distributions and delayed communication between agents.

Researches on multi-armed bandit problems cover a wide range of scenarios from basic reward models to complex distribution and multi-agent environments. As more complex reward distribution and practical limitations, the strategies and algorithms for solving MAMAB problems must to be future developing to make sure that it can be used in more fields of machine learning and artificial intelligence scenarios.

2.4. Fairness and Reward Maximization

In the resource allocation problem of multi-agent systems, how to achieve fairness and to ensure efficiency at the same time is a problem that has not been well solved for a long time. By introducing the Nash Social Welfare (NSW) concept into the multi-armed bandit problem, related research not only focuses on maximizing the total reward, but also pays attention to the fair distribution of rewards.

The exploring of solution for this question, provides new perspectives and tools for solving fairness problems in multi-agent systems, and also provides more humane and social considerations for distributed resource allocation.

In 2020, Aristide et al. researched the egalitarian bargaining solution (EBS) in a stochastic MAB problem between two players, where each player faces different expected rewards [12]. This research highlighted the importance of fairness and individual rational criterion in cooperative settings, and it also showed that cooperative behavior can lead to higher rewards than competitive approaches. This work not only advances the theoretical understanding of the multi-agent MAB problem, but also provides algorithms that can be used in practical scenarios where fairness and cooperative decision-making are crucial.

Then Hossain et al., designed classic algorithm variants such as multi-agent UCB and Thompson sampling based on the fairness of Nash social welfare measurement, and both of them can achieve sublinear regret in losing Nash social welfare situation [13]. This provides a way for resource allocation and policy decision-making in cooperative decision-making. Xiong Wang et al. considered fairness and solving the capacity limitations of edge servers through the concept of Nash social welfare [14]. This algorithm achieves almost best service performance and offloading regret.

In 2023, Matthew Jones et al. introduced an efficient algorithm. Previous algorithms either lacked efficiency or failed to achieve best regret [15]. The algorithms in this paper outperforms these algorithms in terms of low regret and also used an inefficient method of the low regret bound of a single agent. This algorithms performance better than the preview algorithms, showing the practicality in achieving fairness and low regret under limited communication.

3. Conclusion

This paper offers an in-depth review of the innovative applications of the multi-agent multi-armed bandit algorithm in addressing complex decision-making challenges. Initially, it defines the MAMAB problem and describes its relevance across various domains, highlighting the imperative for and obstacles to multi-agent collaboration and competition in intricate settings. A thorough analysis of 70 pivotal publications provides a structured overview of the advances and outcomes in MAMAB research across four critical aspects: sample complexity, communication complexity, reward mechanisms, fairness, and incentives. Despite considerable progress in MAMAB research, significant challenges persist. These include accurately modeling real-world complexities, designing algorithms that can adapt to dynamic conditions and uncertainties, and enhancing communication and cooperation within large-scale agent networks. The exploration of a "Semi-Stationary" model introduces a novel approach, suggesting a potential strategy for addressing MAMAB issues in dynamic environments by incorporating periodic changes into the design of value functions. In conclusion, the study of the multi-agent multi-armed bandit algorithm has evolved from foundational theoretical work to advanced application-focused investigations, establishing it as a powerful tool for resolving real-world complex decision-making issues. Future research should focus more on the practicality and scalability of these algorithms, and on ensuring fairness and stability in the system while promoting efficient decision-making.

Reference

[1] Robbins, H. (1952). Some aspects of the sequential design of experiments.

[2] Gai, Y., Krishnamachari, B., & Jain, R. (2010, April). Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN) (pp. 1-9). IEEE.

[3] Vannella, F., Proutiere, A., & Jeong, J. (2023, July). Best arm identification in multi-agent multi-armed bandits. In International Conference on Machine Learning (pp. 34875-34907). PMLR.

[4] Pankayaraj, P., & Maithripala, D. H. S. (2020, May). A Decentralized Communication Policy for Multi Agent Multi Armed Bandit Problems. In 2020 European Control Conference (ECC) (pp. 356-361). IEEE.

[5] Agarwal, M., Aggarwal, V., & Azizzadenesheli, K. (2022). Multi-agent multi-armed bandits with limited communication. Journal of Machine Learning Research, 23(212), 1-24.

[6] Madhushani, U., Dubey, A., Leonard, N., & Pentland, A. (2021). One more step towards reality: Cooperative bandits with imperfect communication. Advances in Neural Information Processing Systems, 34, 7813-7824.

[7] Chen, Y. Z. J., Yang, L., Wang, X., Liu, X., Hajiesmaili, M., Lui, J. C., & Towsley, D. (2023, April). On-demand communication for asynchronous multi-agent bandits. In International Conference on Artificial Intelligence and Statistics (pp. 3903-3930). PMLR.

[8] Zhu, J., Sandhu, R., & Liu, J. (2020, December). A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In 2020 59th IEEE Conference on Decision and Control (CDC) (pp. 3078-3083). IEEE.

[9] Xu, M., & Klabjan, D. (2024). Decentralized Randomly Distributed Multi-agent Multi-armed Bandit with Heterogeneous Rewards. Advances in Neural Information Processing Systems, 36.

[10] Zhu, X., Huang, Y., Wang, X., & Wang, R. (2023). Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 1-19.

[11] Dubey, A. (2020, November). Cooperative multi-agent bandits with heavy tails. In International conference on machine learning (pp. 2730-2739). PMLR.

[12] Tossou, A. C., Dimitrakakis, C., Rzepecki, J., & Hofmann, K. (2020, May). A novel individually rational objective in multi-agent multi-armed bandits: Algorithms and regret bounds. In Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (pp. 1395-1403).

[13] Hossain, S., Micha, E., & Shah, N. (2021). Fair algorithms for multi-agent multi-armed bandits. Advances in Neural Information Processing Systems, 34, 24005-24017.

[14] Wang, X., Ye, J., & Lui, J. C. (2022, May). Decentralized task offloading in edge computing: A multi-user multi-armed bandit approach. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications (pp. 1199-1208). IEEE.

[15] Jones, M., Nguyen, H., & Nguyen, T. (2023, June). An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 7, pp. 8159-8167).