Balanced and Incentivized Learning with Limited Shared Information in Multi-agent Multi-armed Bandit

Extended Abstract

Junning Shao
Tsinghua University
Shanghai Qi Zhi Institute
China
sjn21@mails.tsinghua.edu.cn

Siwei Wang
Microsoft Research
China
siweiwang@microsoft.com

Zhixuan Fang
Tsinghua University
Shanghai Qi Zhi Institute
China
zfang@mail.tsinghua.edu.cn

ABSTRACT

Multi-agent multi-armed bandit (MAMAB) is a classic collaborative learning model and has gained much attention in recent years. However, existing studies do not consider the case where an agent may refuse to share all her information with others, e.g., when some of the data contains personal privacy. In this paper, we propose a novel limited shared information multi-agent multi-armed bandit (LSI-MAMAB) model in which each agent only shares the information that she is willing to share, and propose the Balanced-ETC algorithm to help multiple agents collaborate efficiently with limited shared information. Our analysis shows that Balanced-ETC is asymptotically optimal, and its average regret (on each agent) approaches a constant when there are sufficient agents involved. Moreover, to encourage agents to participate in this collaborative learning, an incentive mechanism is proposed to make sure each agent can benefit from the collaboration system. Finally, we present experimental results to validate our theoretical results.

KEYWORDS

Multi-armed bandit; Multi-agent collaboration; Incentive mechanism

ACM Reference Format:
Junning Shao, Siwei Wang, and Zhixuan Fang. 2024. Balanced and Incentivized Learning with Limited Shared Information in Multi-agent Multi-armed Bandit: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-armed bandit (MAB) problem is a fundamental theoretical model and has been studied for decades. However, standard MAB setting only focuses on the game with a *single* agent, while many real-world applications face the challenge of *multiple* agents making decisions. As a concrete example, in an online shopping platform, when a buyer chooses to purchase a certain product, she will receive a corresponding reward. Hence, the buyer can be regarded as an agent and the product can be regarded as an arm. Different from the standard MAB setting with a single player, there are always a large number of buyers who are choosing products to purchase in the online shopping platform, and collaboration between them could help them learn much faster. Therefore, in this paper, we consider the multi-agent multi-armed bandit (MAMAB) model, which features multiple players playing (and collaborating in) the same instance of an MAB problem together.

Several variants of the MAMAB model have been studied in the existing literature, e.g., letting the agents collaborate to speed up the learning procedure with limited communication (e.g., [1, 5, 6]), considering decentralized matching market with multi-armed bandit (e.g., [2, 4, 9]). However, these studies do not consider the case where an agent may refuse to share her private information with others, and may even decide to withdraw from learning if she is forced to share some data that she is not willing to share. For example, in an *online shopping platform*, the shared information would be the users' comments about the products, which may help other users make their decisions. However, a user may not comment on every product she purchases in reality, and she can be reluctant to make comments for many reasons, e.g., because the user regards the comments as her privacy, or because commenting on these products does not directly improve her experiences. A survey presented in [7] indicates that customers are likely to refrain from purchasing certain types of items on online shopping platforms due to privacy concerns.

2 MODEL

We introduce a novel multi-agent multi-armed bandit model named limited shared information multi-agent multi-armed bandit (LSI-MAMAB) to characterize the structure of collaborative learning in reality, in which each agent only shares the information that she is willing to share (e.g., only her received rewards on *some* arms) with the other agents, and only decides to participate in this collaboration when she can benefit from it.

In this paper, we design an algorithm *Balanced-ETC* for the case that each agent only shares the information of some specific arms with the other agents.

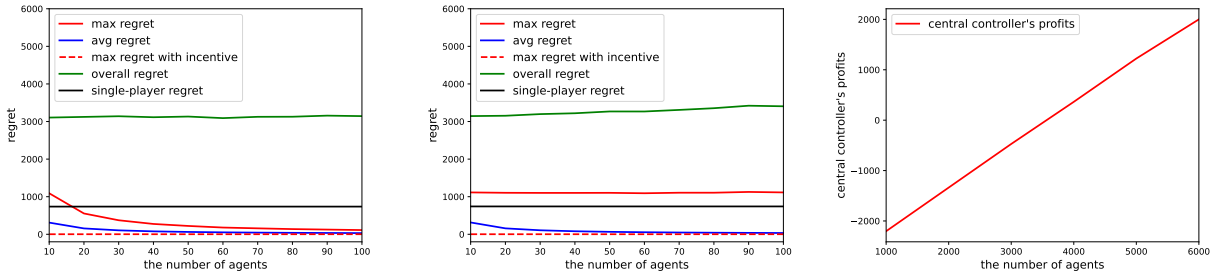
Theorem 2.1 (informal)**.** *Our Balanced-ETC algorithm achieves $O(N \log T)$ total regret, while all the users only need to share the information they are willing to share.*

This is indeed the best one can do, since the overall regret lower bound is $\Omega(N \log T)$ even if everyone shares all her information with each other [3].

This work is supported by Tsinghua University Dushi Program. Corresponding author: Zhixuan Fang at Tsinghua University, Beijing, China (zfang@mail.tsinghua.edu.cn).

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).



(a) Balanced Setting (b) Imbalanced Setting (c) Central Controller's Profits

Figure 1: Experimental results of Balanced-ETC

3 INCENTIVE MECHANISM

Although the overall regret of Balanced-ETC is asymptotically optimal, the algorithm cannot ensure everyone benefits, i.e., some agents may suffer from a higher individual regret (compared with not attending the collaboration system and running a 2-UCB policy herself). For example, if only one agent is willing to share the information of arm 2, then in our LSI-MAMAB model, she needs to pull (and share) arm 2 for more times than running the 2-UCB algorithm alone, since the number of pulls on sub-optimal arms in elimination-based algorithms is always larger than in UCB-based algorithms. Hence, it is necessary to apply some incentive mechanisms in Balanced-ETC to achieve IR.

In our incentive mechanism, there is a center controller which is responsible for compensating the agents (for sharing their data), and collecting the costs from them (for reading the shared data from other agents). The specific amount of cost and compensation are given in the paper. The proposed incentive mechanisms in this paper is objectively present in the real world. For example, in the current existing federal framework, it is common practice for the federated learning platform to provide compensation to the data providers and charge fees to the data users [8].

Theorem 3.1 (informal). *Our incentive mechanism makes sure that both the users and the platform can benefit by participating the collaboration when there are sufficient users.*

Our algorithms and experiments demonstrate that both agents and the platform can benefit from collaboration, indicating that the platform is viable and agents are willing to participate in the cooperation.

4 EXPERIMENTS

In this section, we present experimental results for our Balanced-ETC algorithm and its incentive mechanism. Specifically, in our experiments, there are 10 arms with an expected reward vector $\mu = [0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$. As for the information sharing structure, we assume that every agent is only willing to share the information of *one* arm, and consider two settings: the balanced setting and the imbalanced setting. In the balanced setting (Figure 1(a) and 1(c)), $|S_i|$ is the same for all arm i, i.e., the number of agents who are willing to share the information of any arm is the same. In the imbalanced setting (Figure 1(b)), $|S_1| = |S_2| = 1$, while the other arms' $|S_i|$ is the same, i.e., only few agents are willing to share the good arms. All these results take an average over 100 independent runs.

In Figure 1(a) and Figure 1(b), we compare the performance of Balanced-ETC under different number of agents. We can see that no matter the balanced setting or the imbalanced setting, the overall regret (green line) does not increase a lot when the number of agents grows up. Hence, the *average* individual regret (blue line) keeps decreasing and tends to be zero when there are more agents involved. This accords with our analysis, and demonstrates the effectiveness of our collaboration system. However, as we can see, if we do not apply any incentive mechanism, then the max original individual regret (red line) can be larger than the regret of running the 2-UCB policy alone (black line), especially when there are few agents involved or when the sharing structure is imbalanced. After adding our incentive mechanism (by receiving compensation Com_m and paying cost Cost_m), the maximum incentive individual regret (red imaginary line) is always lower than the regret of running the 2-UCB policy alone (black line), and becomes almost 0. This also accords with our analysis, and demonstrates that our incentive mechanism can achieve IR, i.e., it makes sure that every agent who joins this collaboration system can benefit. In Figure 1(c), we compare the profits of the central controller (i.e., $\sum_m \text{Cost}_m - \sum_m \text{Com}_m$) under different number of agents. We can see that the profit increases linearly in the number of agents, and becomes higher than 0 when there are about 4k agents. This also accords with our analysis, and empirically shows that the central controller can also benefit from making the platform practical in reality, as long as there are sufficient agents participating.

5 CONCLUSION

In this paper, we propose the LSI-MAMAB model, and design the Balanced-ETC algorithm and its corresponding incentive mechanism. This is the first work on multi-agent multi-armed bandit problem with limited information sharing, which sheds light on many collaborative learning scenarios with data sharing constraints. We show that our algorithm's overall regret is asymptotically optimal, and our incentive mechanism achieves individual rationality. Finally, we validate our analysis with experiments.

REFERENCES

[1] Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. 2021. Multi-agent multi-armed bandits with limited communication. *arXiv preprint arXiv:2102.08462* (2021).

[2] Fang Kong and Shuai Li. 2023. Player-optimal Stable Regret for Bandit Learning in Matching Markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 1512–1522.

[3] Tze Leung Lai, Herbert Robbins, et al. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[4] Lydia T Liu, Horia Mania, and Michael Jordan. 2020. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1618–1628.

[5] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. 2019. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 3 (2019), 1–35.

[6] Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. 2017. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2786–2790.

[7] Janice Tsai, Lorrie Faith Cranor, Alessandro Acquisti, and Christina M Fong. 2006. What's it to you? a survey of online privacy concerns and risks. *A survey of online privacy concerns and risks (October 2006) .NET Institute Working Paper* 06-29 (2006).

[8] Rongfei Zeng, Chao Zeng, Xingwei Wang, Bo Li, and Xiaowen Chu. 2021. A comprehensive survey of incentive mechanism for federated learning. *arXiv preprint arXiv:2106.15406* (2021).

[9] Yirui Zhang, Siwei Wang, and Zhixuan Fang. 2022. Matching in Multi-arm Bandit with Collision. *Advances in Neural Information Processing Systems* (2022).