Revised Introduction and Related Work

Introduction

Multi-armed Bandit (MAB) is a well-known online sequential decision making paradigm where a player selects arms, receives corresponding rewards at each time step, and aims to maximize their cumulative reward over a process of length T. Regret minimization is at the heart of MAB, where regret measures the difference between the cumulative reward obtained by always selecting the best arm and the cumulative reward achieved by a player's policy. To this end, balancing exploration (gaining information) and exploitation (maximizing current reward) is key to the player's success. Several classical algorithms have been developed for different MAB settings with proven upper bounds on the regret. Furthermore, to establish optimality of these algorithms, it is essential to prove lower bounds of the same order (in terms of the time horizon T) for all algorithms in specific problem instances. If such lower bounds exist, we refer to them as tight. These worst-case scenario analyses determine the fundamental complexity of bandit problems, validate whether the algorithms are optimal or not, and motivate the development of optimal algorithms. Specifically, in the instance-dependent case, KL-divergence plays a crucial role in characterizing the hardness of distinguishing between optimal and sub-optimal arms. The seminal work by [8] establishes an asymptotic regret lower bound of order $O(\log T)$ for consistent algorithms using an elegant regret decomposition approach that incorporates KL-divergence. The key idea behind these results is to construct problem instances where the optimal arm is very close to the sub-optimal arms but not too close, making it challenging for the player to distinguish between them and resulting in a risk of getting less rewards and significant regret. The gap is precisely chosen and is the main technique.

Recently, the field of multi-agent Multi-armed Bandit (multi-agent MAB) has gained significant attention, driven by the application of cooperative learning processes in federated learning to various real-world scenarios, including e-commerce, healthcare, and autonomous driving, as well as the increasing demand for large-scale distributed decision learning processes in sensor networks and robotic systems. A specific motivating example of the MA-MAB problem is as follows. Consider a ride-sharing platform offering various product lines—premium, luxury, and regular cars—operated by operational units in different areas. Each unit (client) suggests a discount (arm) to users and obtains the revenue (reward) often observing users' behavior. Multiple units collaborate to optimize the total revenues of the platform. This represents an MA-MAB problem aiming to enhance the overall platform performance. Formally, in MA-MAB, multiple agents, also referred to as clients or players, face multiple MABs, and depending on whether the reward distribution of MAB is the same for all agents, we have homogeneous and heterogeneous MA-MAB. The objective of the clients is to optimize the overall system performance, which is quantified using regret. Regret measures the difference between the cumulative reward obtained by pulling the optimal arm, where optimality is defined based on the average rewards across all clients, and the cumulative reward obtained by all the clients. The multi-agent MAB framework presents additional challenges compared to the traditional MAB. Similar to MAB, it deals with the exploration-exploitation trade-off as a major challenge. However, in the multi-agent setting, each client faces this challenge while potentially lacking complete information about other clients. This limitation arises from the fact that optimality is defined based on average rewards across clients, requiring each client to obtain information from other clients, which, however, is constrained by the distribution of clients within the system.

Similar to the categorization in the traditional MAB framework, problem settings in multi-agent MAB are classified as either stochastic or adversarial, depending on the nature of reward distributions. In stochastic multi-agent MAB, the rewards for each client are independently and identically distributed over time, while in adversarial multi-agent MAB, the rewards are chosen by an adversary. Assuming the existence of a central server addresses the problem where the central server can communicate all clients' information. However, the assumption of centralization may not be realistic in real-world scenarios, where clients are often limited to pairwise transmissions constrained by underlying graph structures. A fully decentralized framework characterized by means of graph structures has been proposed in several studies. This decentralized approach removes the centralization assumption, making it more general while introducing non-trivial challenges. To this end, certain assumptions on the graphs are incorporated in these studies. Examples include complete graphs [16], regular graphs [7], and connected graphs under the doubly stochasticity assumption [21, 22]. In all cases, the regret upper bounds that are of order $O(\log T)$, are consistent with those in the MAB setting. Furthermore, recent research has focused on time-varying graphs, such as B-connected graphs under the doubly stochasticity assumption [20], as well as random graphs, including the Erdős-Rényi model and random connected graphs [17]. Likewise, in these cases, the regret upper bounds maintain the order $O(\log T)$. However, it is important to note that the

corresponding regret lower bounds have not yet been addressed in the existing literature, which is one of the main focuses of this study.

In a separate line of research, [6] have introduced a regret upper bound in MAB of order $\sqrt{T}$, which is independent of the sub-optimality gap Δ_i representing the difference between the mean value of the optimal arm and the mean value of the sub-optimal arms. Their setting is standard MAB. Unlike the above regret bound of order $O(\log T) = O\left(\frac{\log T}{\Delta_i}\right)$ that tends to grow rapidly when Δ_i approaches zero, this mean-gap independent regret bound remains stable even when Δ_i is very small and thereby holding universally across different problem settings. Building upon this, [17] analyze the decentralized multi-agent MAB framework with random graphs, and establish a regret upper bound of order $O(\sqrt{T}\log T)$, which aligns with [6] up to a logarithmic factor. However, despite these advancements in the regret upper bounds, the corresponding regret lower bounds in the mean-gap independent sense have not yet been explored. Addressing this research gap is one of the primary objectives of this paper.

In addition to the classical stochastic settings, adversarial multi-agent MAB problem has been proved to have a regret upper bound of order $\sqrt{T}$ and $O(T^{\frac{2}{3}})$, in homogeneous and heterogeneous settings, respectively, demonstrating its consistency and additional challenge with the adversarial MAB problem under the EXP3 algorithm. The presence of heterogeneous adversaries poses a significant challenge. However, the authors establish a regret lower bound of order $\sqrt{T}$, which, while informative, is smaller than the proposed regret upper bound $O(T^{\frac{2}{3}})$. It remains unexplored whether this lower bound is optimal and whether it is possible to develop even larger lower bounds or smaller upper bounds in order to claim optimality. This paper improves the lower bound in this setting and highlights its fundamental challenge by incorporating mini batches and constructing a novel graph instance.

This research gap partly motivates the present study, where we aim to address this knowledge gap and provide a comprehensive analysis of the regret lower bound within the multi-agent MAB framework.

We introduce a novel contribution to the decentralized multi-agent MAB problem by investigating the regret lower bounds in various settings, accounting for different graph structures and reward assumptions. In the context of stochastic rewards and instance-dependent regret bounds, we provide the first formal analysis of the regret lower bound for the centralized setting, demonstrating its tightness. We leverage the aforementioned classical idea in MAB and incorporate it into this multi-agent MAB setting. Additionally, we conduct a comprehensive study on the regret lower bounds in decentralized settings under various graph assumptions by proposing instances that capture the problem complexities of multi-agent systems on a brand new temporal graph. We show that the regret bounds are of order $\Omega(\log T)$, aligning with the existing work's regret upper bounds and establishing their optimality and tightness.

Apart from the instance-dependent regret lower bounds of order $\Omega(\log T)$, we further extend our analysis to mean-gap independent regret lower bounds, presenting a novel contribution as well. Specifically, we establish mean-gap independent regret bounds of order $\Omega(\sqrt{T})$, which not only validate near optimality of the algorithm proposed in [17] up to a $\log T$ factor but also coincide with the existing literature on MAB. This study enhances the understanding of the decentralized problem settings and provides valuable insights for future research in terms of robust methodologies in this context.

Furthermore, our research extends to adversarial settings, where we establish regret lower bounds and demonstrate their tightness across various graph assumptions, including both centralized and decentralized scenarios. Firstly, we show that the regret lower bound is of order $\Omega(\sqrt{T})$ for complete graphs, which aligns with the results for traditional MAB problems, highlighting their inherent similarities. Particularly noteworthy is our finding that the regret lower bound for decentralized multi-agent MAB with connected graphs is of order $\Omega(T^{\frac{2}{3}})$. Notably, we construct a novel graph instance in the connected graph family and adopt a more complicated random shuffling mini batches, which increases the complexity of the problem. This result effectively bridges the gap between the regret upper and lower bounds presented in [19] and establishes that achieving a regret upper bound of $O(\sqrt{T})$ is infeasible in this adversarial setting. Our work uncovers the inherent limitations and challenges of addressing adversarial multi-agent MAB problems even with good connectivity properties compared to traditional MAB problems. Moreover, we explore the regret lower bounds in disconnected graphs with a clique connected component and demonstrate regret lower bounds of order $\Omega(T)$. These findings provide valuable insights into the performance limitations of multi-agent MAB algorithms in graph structures with limited connectivity.

Moreover, as part of our contributions, we implement existing popular algorithms on our proposed instances that are used to prove the regret lower bounds, report crucial findings, and provide insights into next steps. Surprisingly, the performances of theoretically optimal algorithms can sometimes be inferior compared to sub-optimal ones on such hard instances, suggesting room for improvement in the existing regret upper bounds and motivating the development of one-size-fits-all optimal algorithms. Furthermore, we examine the coefficients of the empirical regret curves among these algorithms and point out future directions for theoretical improvements. As a by-product, the computational study also validates the newly established regret lower bounds presented herein.

Our main contributions are as follows. We are the first

- to formally establish the tight instance-dependent regret lower bounds of order $\log T$ in stochastic multi-agent MAB in both centralized and decentralized settings,
- to study the mean-gap independent regret lower bounds of order $\sqrt{T}$ in multi-agent MAB,
- to prove that for adversarial settings, the regret lower bound is of order $T^{\frac{2}{3}}$ and T for connected and disconnected graphs, the first of which bridges the existing gap; a coherent analysis also extends to complete graphs, where the result is of order $\sqrt{T}$.
- to construct technically worst-case scenarios and examine the exact regret of state-of-the-art methods on them, which raises important research questions, and motivates exciting future work.

The structure of the paper is as follows. First, we formally introduce the problem settings along with the notations that are utilized throughout the paper. In the subsequent section, we provide the statements on the regret lower bounds in a wide variety of settings. Last but not least, we present a comprehensive numerical study on the newly proposed instances. Finally, we summarize the paper and point out future possibilities based on the findings in Appendixes A and B.

Related Work

Classical MAB

MAB has a rich history, with regret bounds extensively studied in both instance-dependent and mean-gap independent settings, as well as in stochastic and adversarial scenarios. In stochastic settings, where the reward distribution is time-invariant, numerous studies have established instance-dependent regret upper bounds of order $\log T$. The work of [6] characterizes mean-gap independent regret upper bounds of order $\sqrt{T}$.

In adversarial settings, where the reward distribution can change over time, existing work has demonstrated a regret upper bound of order $O(\sqrt{T})$. However, these algorithms cannot be directly applied to multi-agent MAB problems due to the collaborative nature required among multiple agents.

Regret Lower Bounds

Regret lower bounds are critical for understanding the problem complexity of MAB and for claiming the optimality of algorithms. [8] established the first asymptotic regret lower bounds of order $O(\log T)$ using KL-divergence. Subsequent work, such as [11], relaxed these assumptions, deriving regret bounds for two-arm settings. For mean-gap independent cases, [15] introduced regret bounds of order $\sqrt{T}$, constructing problem instances where distinguishing between arms is deliberately challenging.

Multi-Agent MAB

Centralized and Decentralized

The multi-agent MAB framework has gained prominence due to its relevance in distributed systems. To address this, previous work has extensively studied settings that incorporate a central server, also referred to as a controller, as discussed in [2, 23, 5, 13, 14, 18]. In this setup, the central server integrates and distributes information among the clients at each time step, leading to a regret upper bound of order $O(\log T)$ in stochastic multi-agent MAB, matching the regret bounds in stochastic MAB. However, despite being mentioned in [12] regarding the instance-dependent lower bound of order $\log T$, a formal lower bound statement in this centralized structure remains unexamined.

A fully decentralized framework, characterized by graph structures, has been proposed in several studies [9, 10, 22, 12, 1, 16, 7, 21, 23]. This decentralized approach removes the centralization assumption, making it more general but introducing non-trivial challenges. To address these, studies incorporate specific assumptions on graph types, such as complete graphs [16], regular graphs [7], and connected graphs under the doubly stochasticity assumption [21, 22]. Across all cases, regret upper bounds of order $O(\log T)$ remain consistent with those in traditional MAB settings.

Recent research has also explored time-varying graphs, such as B-connected graphs under doubly stochasticity [20], and random graphs, including Erdős-Rényi and random connected graphs [17]. In these cases, regret upper bounds similarly maintain the order $O(\log T)$. However, corresponding regret lower bounds have not yet been addressed in existing literature, a key focus of this study.

Stochastic and Adversarial

In classical stochastic settings, [3] investigated an adversarial multi-agent MAB problem and provided a regret upper bound of order $\sqrt{T}$, demonstrating consistency with adversarial MAB problems under the EXP3 algorithm. More recently, [19] examined heterogeneous adversarial environments, where adversaries vary across clients. The heterogeneity introduces significant challenges, resulting in a regret upper bound of order $O(T^{\frac{2}{3}})$, larger than the standard MAB regret bound of $\sqrt{T}$. Furthermore, they established a regret lower bound of order $\sqrt{T}$, which, while informative, remains smaller than their proposed upper bound. By leveraging results from [15] and constructing problem instances with mini batches of adversarial rewards, they provided a foundation for further exploration of these bounds.

Instance-Free and Mean-Gap Independent Regret

The concept of mean-gap independent regret, introduced by [6], ensures robustness across varying gap sizes Δ_i. Recent work, such as [17], extended this idea to decentralized settings, achieving regret upper bounds of $O(\sqrt{T}\log T)$. However, lower bounds for mean-gap independent regret in these settings remain unexplored, another focus of this study.

In standard MAB, [6] introduced a regret upper bound of order $\sqrt{T}$, independent of the sub-optimality gap Δ_i, which represents the difference between the mean values of optimal and sub-optimal arms. Unlike regret bounds of $O(\log T) = O\left(\frac{\log T}{\Delta_i}\right)$, which grow rapidly as Δ_i approaches zero, mean-gap independent regret bounds remain stable even for small Δ_i, making them universally applicable across problem settings. Building on this, [17] analyzed decentralized multi-agent MAB with random graphs, establishing regret upper bounds of $O(\sqrt{T}\log T)$, aligning with [6] up to a logarithmic factor. Despite these advancements, corresponding regret lower bounds in the mean-gap independent sense remain unexplored, and addressing this research gap is one of the objectives of this paper.

multi-agent systems.

References

[1] M. Agarwal, V. Aggarwal, and K. Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *The Journal of Machine Learning Research*, 23(1):9529–9552, 2022.

[2] I. Bistritz and A. Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

[4] R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International conference on artificial intelligence and statistics*, pages 3471–3481. PMLR, 2020.

[5] R. Huang, W. Wu, J. Yang, and C. Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34:27057–27068, 2021.

[6] H. Jia, C. Shi, and S. Shen. Multi-armed bandit with sub-exponential rewards. *Operations Research Letters*, 49(5):728–733, 2021.

[7] F. Jiang and H. Cheng. Multi-agent bandit with agent-dependent expected rewards. *Swarm Intelligence*, pages 1–33, 2023.

[8] T. L. Lai, H. Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[9] P. Landgren, V. Srivastava, and N. E. Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference*, pages 243–248. IEEE, 2016.

[10] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

[11] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[12] D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] A. Mitra, H. Hassani, and G. Pappas. Exploiting heterogeneity in robust federated best-arm identification. *arXiv preprint arXiv:2109.05700*, 2021.

[14] C. Réda, S. Vakili, and E. Kaufmann. Near-optimal collaborative learning in bandits. *Advances in Neural Information Processing Systems*, 35:14183–14195, 2022.

[15] O. Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

[16] Z. Wang, C. Zhang, M. K. Singh, L. Riek, and K. Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 1531–1539. PMLR, 2021.

[17] M. Xu and D. Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 2023.

[18] Z. Yan, Q. Xiao, T. Chen, and A. Tajer. Federated multi-armed bandit via uncoordinated exploration. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5248–5252. IEEE, 2022.

[19] J. Yi and M. Vojnovic. Doubly adversarial federated bandits. In *International Conference on Machine Learning*, pages 39951–39967. PMLR, 2023.

[20] J. Zhu and J. Liu. Distributed multi-armed bandits. *IEEE Transactions on Automatic Control*, 2023.

[21] J. Zhu, E. Mulle, C. S. Smith, and J. Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *arXiv preprint arXiv:2111.10933*, 2021.

[22] J. Zhu, R. Sandhu, and J. Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *IEEE Conference on Decision and Control*, pages 3078–3083. IEEE, 2020.

[23] Z. Zhu, J. Zhu, J. Liu, and Y. Liu. Federated bandit: A gossiping approach. In *ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.