Regret Lower Bounds in Multi-agent Multi-armed Bandit

Anonymous submission
Anonymous affiliation

Abstract

Multi-armed Bandit motivates methods with provable upper bounds on regret and also the counterpart lower bounds have been extensively studied in this context. Recently, Multi-agent Multi-armed Bandit has gained significant traction in various domains, where individual clients face bandit problems in a distributed manner and the objective is the overall system performance, typically measured by regret. While efficient algorithms with regret upper bounds have emerged, limited attention has been given to the corresponding regret lower bounds, except for a recent lower bound for adversarial settings, which, however, has a gap with let known upper bounds. To this end, we herein provide the first comprehensive study on regret lower bounds across different settings and establish their tightness. Specifically, when the graphs exhibit good connectivity properties and the rewards are stochastically distributed, we demonstrate a lower bound of order $O(\log T)$ for instance-dependent bounds and $\sqrt{T}$ for mean-gap independent bounds which are tight. Assuming adversarial rewards, we establish a lower bound $O(T^{\frac{2}{3}})$ for connected graphs, thereby bridging the gap between the lower and upper bound in the prior work. We also show a linear regret lower bound when the graph is disconnected. These lower bounds are made possible through our newly constructed instances. In the numerical study, we assess the performance of various algorithms on these hard instances. While previous works have explored these settings with upper bounds, we provide a thorough study on tight lower bounds.

1 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 [Lai *et al.*, 1985] establishes an asymptotic regret lower bound of order $O(\log T)$ for consistent algorithms using an elegant regret decomposition approach that incorporates KL-divergence. Subsequent work relaxes the assumptions of consistency and asymptotics [Lattimore and Szepesvári, 2020] assuming 2 arms. For the mean-gap independent case, [Lattimore and Szepesvári, 2020] demonstrate a minimax regret lower bound of order $\sqrt{T}$. Furthermore, [Shamir, 2014] establishes a general regret lower bound of order $\sqrt{T}$ for MAB variants where multiple arms can be pulled at each time step. 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 multi-agent MAB, multiple agents, also referred to as clients or players, face multiple MABs. 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. 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.

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. To tackle this issue, previous work has extensively studied settings that incorporate a central server, also referred to as a controller, as discussed in [Bistritz and Leshem, 2018; Zhu *et al.*, 2021b; Huang *et al.*, 2021; Mitra *et al.*, 2021; Réda *et al.*, 2022; Yan *et al.*, 2022]. In this setup, the central server integrates and distributes information among the clients at each time step, which has led 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 [Martínez-Rubio *et al.*, 2019] regarding the instance-dependent lower bound of order $\log T$, a formal lower bound statement has yet to be thoroughly examined in this centralized structure. 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 centralized multi-agent MAB framework.

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. In response to this, a fully decentralized framework characterized by means of graph structures has been proposed in several studies [Landgren *et al.*, 2016b,a, 2021; Zhu *et al.*, 2020; Martínez-Rubio *et al.*, 2019; Agarwal *et al.*, 2022; Wang *et al.*, 2021; Jiang and Cheng, 2023; Zhu *et al.*, 2021a,b]. 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 [Wang *et al.*, 2021], regular graphs [Jiang and Cheng, 2023], and connected graphs under the doubly stochasticity assumption [Zhu *et al.*, 2021a, 2020]. 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 [Zhu and Liu, 2023], as well as random graphs, including the Erdős-Rényi model and random connected graphs [Xu and Klabjan, 2023a]. 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, [Jia *et al.*, 2021] 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, [Xu and Klabjan, 2023a] 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 [Jia *et al.*, 2021] 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, [Cesa-Bianchi *et al.*, 2016] investigate an adversarial multi-agent MAB problem and provide a regret upper bound of order $\sqrt{T}$, demonstrating its consistency with the adversarial MAB problem under the EXP3 algorithm. More recently, [Yi and Vojnović, 2023] have focused on the heterogeneous variant, where different adversaries are different across clients. The presence of heterogeneous adversaries poses a significant challenge, resulting in a regret upper bound of order $O(T^{\frac{2}{3}})$, which is larger than the regret bound for the standard MAB problem of order $\sqrt{T}$. Furthermore, in the adversarial setting, they establish a regret lower bound of order $\sqrt{T}$, which, while informative, is smaller than their proposed regret upper bound. They achieve this by leveraging the results from the MAB setting presented in [Shamir, 2014] and constructing problem instances with mini batches of adversarial rewards. Nevertheless, 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.

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 [Xu and Klabjan, 2023a] 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 [Yi and Vojnović, 2023] 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 suboptimal 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.

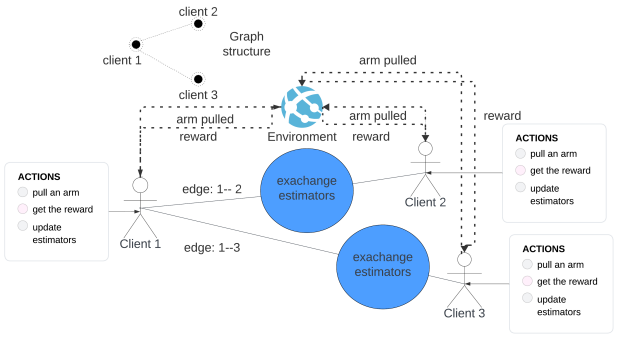
2 Problem Formulation

Throughout the paper, we study a decentralized system with $M \geq 3$ clients, and T represents the time horizon. More specifically, the clients are labeled as nodes $1, 2, \ldots, M$ on a network, where the underlying graph at each time step $1 \leq t \leq T$ is represented by an undirected graph G_t. It is worth emphasizing that the centralization structure is equivalent to communications on a complete graph since every pair of clients communicates through the central server.

Formally, $G_t = (V, E_t)$ is described by a unique vertex set $V = 1, 2, \ldots, M$ and an edge set E_t that contains pairwise nodes and conveys the neighborhood information of G_t. We use $\mathcal{N}_m(t)$ to denote the neighbor set of client m, which represents all the neighbors of client m in G_t. It is worth noting that the graph G_t can be equivalently described by its adjacency matrix, denoted as $(X^t_{i,j})_{1\leq i,j\leq M}$, where the element $X^t_{i,j}$ is equal to 1 if there is an edge between clients i and j, and 0 otherwise. For simplicity, we specify $X_{i,i} = 1$ for any client $1 \leq i \leq M$. We use $\mathcal{G}_M$ to denote the set of all connected graphs with M nodes. If $G = G_t$, we call it stationary and otherwise temporal. In the Erdős-Rényi model we use superscript c where c is the edge probability, e.g. $\mathcal{N}^c_m(t)$ is defined based on probability c. In the random connected graph model we denote by c the probability of an edge being in such a graph.

Subsequently, we introduce the bandit problems associated with the clients. Consistent with the existing literature, an environment generates graphs G_t and rewards $r^m_i(t)$. For each client $1 \leq m \leq M$, there are $K \geq 2$ arms to be pulled. At each time step t, the reward of arm $1 \leq i \leq K$ is denoted

as $r_i^m(t)$, which is independently and identically distributed across time with a mean value of μ_i^m. The clients draw rewards independently of one another. The interaction between the client and the environment works as follows; Client m pulls an arm a_m^t and obtains the corresponding reward $r_{a_m^t}^m(t)$ from the environment. Additionally, clients can communicate with their neighbors in G_t as provided by the environment. This means that two clients can exchange information if and only if they are connected by an edge. Below, we present a visualization of the proposed problem framework for illustrative purposes.



Following [Yi and Vojnović, 2023; Zhu and Liu, 2023], we define the global reward of arm i as $r_i(t) = \frac{1}{M}\sum_{m=1}^M r_i^m(t)$, and the corresponding expected global reward as $\mu_i = \frac{1}{M}\sum_{m=1}^M \mu_i^m$. An arm is called globally optimal if $i^* = \arg\max_i \mu_i$, and globally sub-optimal otherwise. The parameter $\Delta_i = \mu_{i^*} - \mu_i$ represents the sub-optimality gap of arm i.

We note that $\max_i T \cdot \mu_i = \max_i E[\sum_{t=1}^T r_i(t)] \leq E[\max_i \sum_{t=1}^T r_i(t)]$, by the Jensen's inequality. If we establish a lower bound on the regret defined with respect to $\max_i T \cdot \mu_i$ (called also pseudo regret), we establish that the expected regret with respect to $E[\max_i \sum_{t=1}^T r_i(t)]$ exhibits the same lower bound. As a result, we focus on demonstrating lower bounds on the pseudo regret throughout the paper, which is called regret for convenience.

This allows us to precisely quantify the regret associated with the action sequence (policy) $\pi = \{a_m^t\}_{1\leq m\leq M}^{1\leq t\leq T}$. In an ideal scenario where complete knowledge of $\{\mu_i\}_i$ is available, clients would prefer to pull the arm i^*. However, due to partially observed rewards from the bandits (dimension i) and limited access to information from other clients (dimension m), the regret of a policy π in the bandit setting is defined as $R_T^\pi = T\mu_{i^*} - \frac{1}{M}\sum_{t=1}^T\sum_{m=1}^M \mu_{a_m^t}^m$. This regret metric quantifies the difference between the cumulative expected reward obtained by following the globally optimal arm and the actual reward accumulated by executing the action sequence. We consider two types of policies. Denote $\sigma_F^{t,m} = \sigma(\{\{I_j^s\}_{j\in\mathcal{N}_m(s)}\}_{s\leq t})$ where I_j^s represents the information of all arms contained at client j at time step s and, denote $\sigma_B^{t,m} = \sigma(\{\{I_j^s(a_s^j)\}_{j\in\mathcal{N}_m(s)}\}_{s\leq t})$ where $I_j^s(a_s^j)$ represents the information of arm a_s^j contained at client j at time step s. In other words, σ_F^t captures the history of all arms up to time t, whereas $\sigma_B^{t,m}$ only contains the information of client m's time dependent actions up to time t. Henceforth, we have $\sigma_B^{t,m} \subset \sigma_F^{t,m}$. With these notations at hand, we further define policy set Π_F and Π_B as $\Pi_F = \{f_t\}$ where the domain of f_t is on $\sigma_F^t = \{\sigma_F^{t,m}\}_m, \Pi_B = \{g_t\}$ where the domain of g_t is on $\sigma_B^t = \{\sigma_B^{t,m}\}_m$. To this end we define $R_T^B = \min_{\pi\in\Pi_B} R_T^\pi$. Likewise, assuming the observations of all arms are visible to the clients, which is referred to as the full-information setting, we denote the regret as $R_T^F = \min_{\pi\in\Pi_F} R_T^\pi$.

The primary objective of this paper is to develop theoretical lower bounds on the regret in worst-case scenarios under different assumptions on the underlying graphs, where clients operating in decentralized settings have certain regrets regardless of the policies deployed.

3 Lower Bound Analyses

Before analyzing the regret lower bounds in bandit settings, we consider its relationship with the regret in the full information setting. The full information setting provides a less black-box approach for characterizing the regret of algorithms.

Theorem 1. *For decentralized multi-agent problems on any graph G_t, for all problem instances we have $R_T^F \leq R_T^B$.*

Proof. Consider any policy $\pi \in \Pi_B$. Since it only requires the information of clients' actions σ_B^t, and $\sigma_B^t \subset \sigma_F^t$, we obtain that $\pi \in \Pi_F$. Subsequently, we arrive at $\Pi_B \subset \Pi_F$ by the arbitrary choice of π, which yields that $\min_{\pi\in\Pi_F} R_T^\pi \leq \min_{\pi\in\Pi_B} R_T^\pi$, or equivalently $R_T^F \leq R_T^B$. □

Subsequently, we establish the following regret lower bounds in the instance-dependent and mean-gap independent sense for the full information setting.

Theorem 2. *For decentralized multi-agent online problems with full information, if the graph G is a complete graph, then there exists a problem instance such that the regret of any online distributed learning algorithms is at least $\Omega(\sqrt{T})$ and $\Omega(\log T)$ in mean-gap independent and instance-dependent settings, respectively.*

Proof sketch. The complete proof is presented in Appendix; the main idea is as follows. We note that the complete graph case is approximately equivalent to a single-agent bandit problem with full information. For the single-agent case, there exists literature establishing the corresponding instance-dependent regret bound of order $\log T$ and mean-gap independent regret bound of order $\Omega(\sqrt{T})$, as introduced in [Goldenshluger and Zeevi, 2013] and Shamir [2014], respectively. □

3.1 Instance-dependent

Next, we demonstrate the instance-dependent lower bounds in stochastic bandits for different graph structures, building upon the previously established lower bound for the full information setting. More specifically, instance-dependent lower

bounds depend on the sub-optimality gap Δ_i that varies across different problem settings, which allows for a more precise characterization of regret in various instances, reflecting their complexities. The graph structures include time-invariant complete, connected, and regular graphs, as well as time-varying complete, connected, regular graphs, and time-varying Erdős-Rényi (E-R) model and random connected graphs, which encompass the graphs studied in prior works. The formal statement is as follows.

Theorem 3. *For decentralized multi-agent MAB problems with any numbers of clients and stochastic rewards, if G_t are complete, or connected or regular, and either stationary or temporal, or if G_t follow the E-R model or are random connected graph, then the instance-dependent expected regret R_T^B of any algorithm is at least $\Omega(\log T)$.*

Proof sketch. The complete proof is deferred to Appendix; the main idea is as follows. We construct an instance where the number of arms is 2 and $\Delta_2 = \mu_1 - \mu_2 > 0$. For the complete graph case, we consider the time period T_a when clients achieve an agreement and T_d when clients experience disagreement. Subsequently, we decompose the regret as $R_T^\pi = T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1 - \mu_{a_t^m})$. The case where $T_d = \Omega(\log T)$ directly leads to the conclusion. On the other hand, the scenario where $T_d = o(\log T)$ is managed by dividing the time horizon into intervals $\bigcup_{j=0}^{t_0}[2^j, 2^{j+1}-1]$, where $t_0 = \log T$. This division enables us to derive $T_a^d = T_a \cap [t_d, T] = [t_d, T] \geq 2^{\frac{1}{2}\log T}$, where $t_d = \max\{t | t \in T_d\} + 1$ and the inequality holds by $T_d = o(\log T)$.

For T_a^d, the regret $\frac{1}{M}\sum_m \sum_{t\in T_a^d}(\mu_1 - \mu_{a_t^m})$ is related to the regret in a single-agent multi-objective bandit problem [Xu and Klabjan, 2023b] and, precisely, the regret is bounded from below by the Pareto pseudo regret $R_{T_a^d,M} = Dist(\sum_{t\in T_a^d}(\mu_{a_t}^m)_m, O)$. The latter exhibits a lower bound of order $\Omega(\log T)$ as shown in Theorem 6 in [Xu and Klabjan, 2023b]. This concludes the regret lower bound in settings with complete graphs.

Using the monotonicity of regret in graph complexity, we derive the same lower bounds for scenarios with random connected graphs or the E-R model. This concludes the proof. □

Remark. *While [Martínez-Rubio* et al.*, 2019] discuss the instance-dependent regret lower bound of order $\Omega(\log T)$ in the centralized setting, we provide the first formal statement for various graphs. The result coincides with the lower bound in the single-agent MAB setting. Furthermore, the result is consistent with the established upper bounds in the multi-agent MAB settings, thereby demonstrating its tightness.*

Additionally, we also consider scenarios with disconnected graphs, which can result in linear regret due to the presence of isolated clients when the rewards are heterogeneous. The first result applies to consistent algorithms, following the classical assumption made in some existing literature. The consistency assumption states that the regret of the considered algorithms is of order $o(T^a)$ for any constant $0 < a \leq 1$. The second result applies to any algorithms, with the constraint of limiting the number of arms to 2. These results are summarized in the following statements.

Theorem 4. *For decentralized multi-agent MAB problems, if graph G is disconnected with a clique connected component, then there exists a problem instance such that the regret of any online distributed algorithms that are individually consistent at local clients is at least $\Omega(T)$.*

Proof sketch. The proof is deferred to Appendix; the main logic is as follows when the clique is an isolated vertex. We construct a problem instance as follows. For clients $1, \ldots, M-1$, their reward distributions are the same, reading as $(\Delta, 0, \ldots, 0) \in R^K$, while for client M, the reward distribution reads as $(0, 2\Delta, 0, \ldots, \ldots, 0) \in R^K$ for any $\Delta > 0$. We assume node M is isolated. Using any consistent algorithms at client M leads to $E[n_{M,2}(T)] = \Omega(T)$ and subsequently results in a linear regret. Here $n_{M,2}$ is the number of pulls of arm 2 at client M. □

As mentioned earlier, we remove the consistency assumption by assuming the number of clients is 2, which essentially deals with the trade-off between the problem setting and the considered algorithms.

Theorem 5. *For decentralized multi-agent MAB problems, if graph G is disconnected with a clique connected component, then there exists a problem instance with $K = 2$ such that the regret of any online distributed algorithms is at least $\Omega(T)$.*

Proof sketch. The proof is given in Appendix; the proof logic is as follows when the clique component is an isolated vertex. We again let client M be an isolated node. For two arms labeled as arm 1 and 2, we construct the instance at clients as follows. Let random variable x follow a uniform distribution in $\{0, 1\}$ and be fixed once determined, and for any time step t, the reward $r_k^j(t)$ is generated as $r_k^1(t) = \begin{cases} x & \text{arm 1} \\ \frac{1}{2} & \text{arm 2} \end{cases}$ and for $j > 1$ we have $r_k^j(t) = \begin{cases} \frac{1}{2} & \text{arm 1} \\ \frac{1}{2} & \text{arm 2.} \end{cases}$ The randomness of x changes the optimality of arms, and makes client M even harder to identify the global optimal arm and impossible to achieve sublinear regret even though inconsistent algorithms are deployed. □

Remark. *To the best of our knowledge, this is the first result on the regret lower bound for settings with disconnected graphs. This linear regret essentially highlights the inherent complexity of multi-agent MAB problems compared to their single-agent counterparts.*

3.2 Mean-gap independent

We note that the instance-dependent regret bounds have dependencies on Δ_i, particularly on $\frac{1}{\Delta_i}$, which can lead to large regret bounds when $\Delta_i \approx 0$ and thus necessitate more accurate regret bounds. In addition, Δ_i's are not known to the clients in

advance. Therefore, apart from the instance-dependent regret lower bounds, we also investigate the mean-gap independent regret lower bound that is independent of Δ_i and applicable to both stochastic and adversarial settings. The regret order in this case is $\sqrt{T}$, which differs from the $\log T$ bound. The following theorem summarizes these results, considering all the previously mentioned graph structures.

Theorem 6. *For decentralized multi-agent MAB problems with any numbers of clients and stochastic rewards, if G_t are complete, connected or regular, and stationary or temporal, or the E-R model or random connected graphs, then the mean-gap independent regret of any algorithm is at least $\Omega(\sqrt{T})$.*

Proof sketch. The formal proof is in Appendix; the main logic is as follows. The proof is similar to that of Theorem 6, except that we consider mean-gap independent bounds using Theorem 4 in [Shamir, 2014]. We first analyze settings with complete graphs and establish $R_T^B \geq \sqrt{\frac{KT}{1+M}} = \Omega(\sqrt{T})$. Likewise, the monotonicty in graphs of the regret bounds allow us to determine the same result for other graphs, which concludes the proof. □

Remark. *This lower bound of order $\sqrt{T}$ corresponds to the mean-gap upper bounds presented in [Xu and Klabjan, 2023a] and [Jia* et al.*, 2021] for multi-agent and single-agent MAB problems, respectively. This consistency further shows the tightness of the lower bound we have derived.*

3.3 Adversarial

Since the mean-gap independent regret bounds hold for the stochastic problem setting, they also hold for the adversarial problem setting. This is due to the fact that the set of stochastic settings is essentially a subset of the set of adversarial settings. Therefore, our result remains consistent with the result in [Yi and Vojnović, 2023].

Theorem 7. *For decentralized multi-agent MAB problems, if the graph G_t is a complete graph, then there exists a problem instance such that the regret of any online distributed learning algorithms is at least $\Omega(\sqrt{T})$.*

Furthermore, we construct special connected graphs, in adversarial settings and demonstrate that they lead to a regret lower bound of order $\Omega(T^{\frac{2}{3}})$. This bound is larger than the commonly observed $O(T^{\frac{1}{2}})$ in single-agent adversarial settings and decentralized multi-agent adversarial settings with complete graphs. We summarize these results in the following two theorems, one for a large number of clients and the other one for a small number of clients.

Theorem 8. *For decentralized multi-agent MAB problems, if the number of clients $M \geq \Omega(T^{\frac{1}{3}})$ and the graph G_t is a connected graph with two expanders of size $\frac{M}{4}$ having distance $d \geq \frac{\eta M}{8}$ given constant $4 > \eta > 0$, then there exists a problem instance such that the regret of any online distributed learning algorithm is at least $\Omega(T^{\frac{2}{3}})$.*

Proof sketch. The proof is deferred to Appendix; the idea is summarized as follows. We consider clients are distributed on a special connected graph, e.g. a path graph and focus on two subsets of node, denoted as I_0 and I_1, respectively, that satisfy $|I_0| = |I_1| = \frac{M}{4}$, and the shortest path d_p from I_0 to I_1 meets the condition $d_p \geq \frac{\eta M}{8}$. Then the choice of M gives $d_p \geq \Omega(T^{\frac{1}{3}})$ and we import the result in [Yi and Vojnović, 2023] and obtain $R_T^B \geq \Omega(\sqrt{d_p \cdot T}) = \Omega(T^{\frac{2}{3}})$ for full-information settings. □

Remark. *Note that the existence of such graphs is guaranteed by the property of expanders of size $\frac{M}{4}$. An expander of size $\frac{M}{4}$ has a diameter of order $\log M$ (Proposition 3.1.5 in [Kowalski, 2019]). Indeed, for $\eta = 4$, a path is such an expander.*

For small values of M, achieving the same regret lower bound requires additional effort since the setting allows for more communication between clients. In this case, we present the following result that establishes the same lower bound on regret by importing techniques from information theory.

Theorem 9. *For decentralized multi-agent MAB problems, if the number of clients $M = T^{\frac{2}{15}}$ and the graph G_t is a connected graph with two expanders of size $\frac{M}{4}$ having distance $d \geq \frac{\eta M}{8}$ given constant $4 > \eta > 8 \cdot 8^{-\frac{2}{15}}$, then there exists a problem instance such that the regret of any online distributed learning algorithms is at least $\Omega(T^{\frac{2}{3}})$.*

Proof. Let $M \, mod \, 4 = 0$ and $T > 8$. Denote expanders of size $\frac{M}{4}$ as two disjoint subsets of nodes $I_0 = \{1, 2, \ldots, \frac{M}{4}\}$ and $I_1 = \{\frac{3}{4}M, \frac{3}{4}M+1, \ldots, M\}$. Note that $|I_0| = |I_1| = \frac{M}{4}$. By the definition of G_t, the shortest path distance between I_0 and I_1 is $d \geq \frac{\eta M}{8}$. We set $\epsilon = \sqrt{\frac{4}{\eta}} \frac{M^2}{2} T^{-\frac{1}{3}}$. It follows $8\epsilon^2 d \leq 1$.

Let B_1 be Bernoulli with probability $\frac{1}{2} + \epsilon$ and B_2 Bernoulli with probability $\frac{1}{2}$. Consider the bandit problem as follows. Let X be a random variable following a uniform distribution on $\{0, 1, \ldots, \frac{M}{4}\}$. For client $X \geq 1$, arm 1 follows B_1 and arm 2 follows B_2. For $i \in I_0 \backslash \{X\}$, let the arms follow B_2. All clients not in I_0 have all rewards 0.

Additionally, we re-sample random variable X every d steps, i.e. we re-specify the client X if $X \geq 1$. If $X = 0$, all clients have reward based on B_2. We denote the number of such re-sampling steps as D, $D = \lfloor \frac{T}{d} \rfloor$, which leads to a sequence $\{X_1, X_2, \ldots, X_D\}$. The following holds for $i \in I_0$. Subsequently, let us define distribution $Q_j^i(arm) = P(arm|X_j = i)$ and $Q_j^{-1}(arm) = P(arm|X_j = 0)$. Note that Q_j^{-1} represents that all clients in I_0 share the same reward distribution. Let $Q_{j,t}^i(arm) = P(arm|\sigma_t, X_j = i)$ and $Q_{j,t}^{-1}(arm) = P(arm|\sigma_t, X_j = 0)$. It is easy to verify that $D_{KL}(Q_{j,t}^{-1}, Q_{j,t}^i) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}-\epsilon} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}+\epsilon} = \frac{1}{2} \log(1 + \frac{4\epsilon^2}{1-4\epsilon^2}) \leq \frac{1}{2} \cdot \frac{4\epsilon^2}{1-4\epsilon^2} \leq 4\epsilon^2$, where the first inequality uses the fact that $\log(1+x) \leq x$ and the second inequality

holds by the choice of $\epsilon = \frac{M^2}{2}T^{-\frac{1}{3}} \leq \frac{1}{4}$ since $T > 8$.

Therefore, by the chain rule for relative entropy, we obtain $D_{KL}(Q_j^{-1}, Q_j^i) = \sum_{t=jd}^{(j+1)d} D_{KL}(Q_{j,t}^{-1}, Q_{j,t}^i) \leq \sum_{t=jd}^{(j+1)d} 4\epsilon^2 \leq 4\epsilon^2 d$. By the Pinsker's inequality we have that $D_{TV}(Q_j^{-1}, Q_j^i) \leq \sqrt{\frac{D_{KL}(Q_j^{-1}, Q_j^i)}{2}} \leq \epsilon\sqrt{2d}$. (1)

The expected reward of arm 1 is $\frac{1}{8} + \frac{1}{M}\frac{|I_0|}{|I_0|+1}\epsilon$ from

$$\begin{aligned}
\mu_1 &= \frac{1}{M}\sum_{m=1}^{M}\mu_1^m = \frac{1}{M}\sum_{m\in I_0}\mu_1^m + \frac{1}{M}\sum_{m\notin I_0}\mu_1^m \\
&= \frac{1}{M}\sum_{m\in I_0}\Big[E[\mu_1^m|X_1 \in I_0]P(X_1 \in I_0) + \\
&\qquad \sum_{m\in I_0} E[\mu_1^m|X_1 \notin I_0]P(X_1 \notin I_0)\Big] + \frac{1}{M}\sum_{m\notin I_0} 0 \\
&= \frac{1}{M}\Big(\frac{|I_0|}{|I_0|+1}\Big(\frac{1}{2} + \epsilon + \frac{1}{2}(|I_0|-1)\Big) + \\
&\qquad \frac{1}{|I_0|+1}\Big(\frac{1}{2} + \frac{1}{2}(|I_0|-1)\Big) = \frac{1}{8} + \frac{1}{M}\frac{|I_0|}{|I_0|+1}\epsilon
\end{aligned}$$

and of arm 2 is $\frac{1}{8}$ from $\mu_2 = \frac{1}{M}\sum_{m=1}^{M}\mu_2^m = \frac{1}{M}\sum_{m\in I_0}\mu_2^m + \frac{1}{M}\sum_{m\notin I_0}\mu_2^m = \frac{1}{M}\sum_{m\in I_0}\frac{1}{2} + \frac{1}{M}\sum_{m\notin I_0} 0 = \frac{1}{8}$. As a result $\Delta_1 = \frac{\epsilon}{M}\frac{|I_0|}{|I_0|+1} \geq \frac{\epsilon}{2M}$ since $|I_0| \geq 1$. Let us denote by $n_{m,1}(T, j)$ the number of pulls of arm 1 by client m during the j^{th} epoch which is the optimal arm. Therefore, we obtain

$$\begin{aligned}
&E[R_T^B] = E[E[R_T^B|X_1, \ldots, X_D]] \qquad (2)\\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(T - n_{m,1}(T)))|X_1, \ldots, X_D]] \\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(\sum_{j=1}^{D} d - \sum_{j=1}^{D} n_{m,1}(T,j)))|X_1, \ldots, X_D]] \\
&= E[\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_1, \ldots, X_D]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}E[E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}\frac{E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j = i]]}{|I_0|+1} \\
&\geq \frac{1}{2M^2}(\frac{1}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}E[\epsilon\cdot(d - n_{1,1}(T,j))|X_j = i]) \\
&= \frac{1}{2M^2}(\epsilon\cdot T - \frac{\epsilon}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}E_{Q_j^i}[(n_{1,1}(T,j))])
\end{aligned}$$

where the the first and fifth equality use the law of total expectation, the third equality is by the fact that $T = \sum_{j=1}^{D} d$ and $\sum_{j=1}^{D} n_{m,1}(T, j) = n_{m,1}(T)$, and the sixth equality uses the distribution of X_j defined by $P(X_j = i) = \frac{1}{|I_0|+1}$ for $i \in I_0 \cup \{0\}$.

Note that $E_{Q_j^i}[(n_{1,1}(T,j))] - E_{Q_j^{-1}}[(n_{1,1}(T,j))] = \sum_{t=jd}^{(j+1)d}(Q_j^i(a_t^1 = 1) - Q_j^{-1}(a_t^1 = 1)) \leq d\cdot D_{TV}(Q_j^{-1}, Q_j^i)$ where the last inequality is by the definition of the total variation D_{TV}.

This immediately gives us that

$$\begin{aligned}
&\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}E_{Q_j^i}[(n_{1,1}(T,j))] \\
&\leq \sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}\sum_{t=jd}^{(j+1)d}(Q_j^{-1}(a_t^1 = 1) + d\cdot D_{TV}(Q_j^i, Q_j^{-1})) \\
&\leq T + d\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}D_{TV}(Q_j^i, Q_j^{-1})) \\
&\leq T + d\sum_{i\in I_0\cup\{0\}}\sum_{j=1}^{D}(\epsilon\sqrt{2d}) \\
&= T + dD\epsilon\sqrt{2d}(|I_0|+1) = T + T\cdot\frac{|I_0|+1}{4}
\end{aligned}$$

where the second inequality uses $\sum_i Q_j^{-1}(a_t^1 = 1) = 1$ and $dD = T$, and the third inequality uses (1), and the last equality holds by the choices of d and ϵ that satisfy $\epsilon\sqrt{2d}(|I_0|+1) \leq \frac{|I_0|+1}{4}$. Here we also use the lower bound on η.

Consequently, we arrive at $E[R_T^B] \geq \frac{1}{2M^2}(\epsilon T - \frac{\epsilon(T + T\cdot\frac{|I_0|+1}{4})}{|I_0|+1}) \geq \frac{1}{2M^2}\frac{1}{4}\epsilon T = \Omega(T^{\frac{2}{3}})$ where the last inequality uses $|I_0| = \frac{M}{4} \geq 2$ and the equality holds by the choice of ϵ and M. □

Remark. *It is worth noting that this lower bound is consistent with the regret upper bound in [Yi and Vojnović, 2023], bridging the gap between the regret upper bound $O(T^{\frac{2}{3}})$ and the lower bound $\Omega(\sqrt{T})$ in [Yi and Vojnović, 2023]. Surprisingly, it also coincides with the regret lower bound for online learning with feedback graphs in [Alon* et al.*, 2015], where the feedback received by the client is limited to a graph structure. This connection highlights the relationship between the decentralized multi-agent MAB system and MAB with side information on graphs. Lastly, we observe that this bound is larger than $\sqrt{T}$ in the single-agent MAB, manifesting the fundamental difference between multi-agent and single-agent MAB in the presence of connected graphs, in addition to the settings with disconnected graphs.*

4 Numerical Experiments

We have demonstrated regret lower bounds that apply to all algorithms in different settings by constructing novel problem instances. In this section, we conduct a comprehensive numerical study to understand how the newly constructed challenging

instances affect the performance of existing algorithms. The results, consistent with our established regret lower bounds, also highlight opportunities for methodological and analytical improvements aimed at achieving optimal regret upper bounds. Furthermore, they provide insights into the future direction of such improvements. We select these algorithms based on the following criteria: i) corresponding regret guarantee, ii) low computational complexity, and iii) wide usage in the existing literature.

Specifically, we examine the exact regret of different algorithms on our newly proposed instances across various multi-agent MAB settings, representing worst-case scenarios. Recall that the problem complexity is determined by how the rewards and graphs are generated. In this paper, we consider both stochastic and adversarial rewards, along with graphs exhibiting different levels of connectivity, ranging from complete to random and disconnected graphs. We have constructed four hard instances (scenarios), categorized as follows - Instance 1 with stochastic rewards and complete graphs based on Theorem 3, Instance 2 with stochastic and adversarial rewards and disconnected graphs (see Theorem 4), Instance 3 with adversarial rewards and complete graphs exhibited in Theorem 7, and Instance 4 with adversarial rewards and connected graphs based on Theorem 8. Within them, we include the optimal algorithms in terms of regret's upper bounds in T. For Instance 1, the optimal algorithms include CoopUCB in [Martínez-Rubio *et al.*, 2019], Gossip_UCB in [Zhu *et al.*, 2021b], and GosInE in [Chawla *et al.*, 2020], all of which lead to a regret upper bound of order $\log T$. In Instance 2, every algorithm is optimal since they all have linear regret of order T if the reward is bounded. In Instance 3, the optimal algorithm is known to be EXP3, with regret of order $\sqrt{T}$. In Instance 4, the recently developed FEDEXP3 in [Yi and Vojnović, 2023] results in an upper bound of order $T^{\frac{2}{3}}$.

Our experimental details are as follows. The time horizon is fixed at $T = 2000$. For Instance 1 and 3, we consider $M = 5$ clients distributed on a complete graph, with $K = 2$ arms and a heterogeneity level of $h = 0.1$. For each arm k, the associated mean reward values μ_k^m are M equal length intervals of $[0.1, 0.1 + (k+1)/K \cdot h]$ in Instance 1. In Instance 3, we use the same mean reward values but they are shuffled randomly every 6 periods. For Instance 2, we consider $M = 5$ clients distributed on a disconnected graph where clients $0, 1, 2, 3$ form a complete graph and client 4 is an isolated point, with $K = 2$ arms and a heterogeneity level of $h = 0.1$. For each arm k, the associated mean reward values follow the same partitioning as in Instance 1 and 3. Instance 4 requires further explanation. We consider $M = 10$ clients distributed along a path graph, with $K = 2$ arms. The mean reward values for arms 1 and 2 of clients 1 and 2 are either $0.5, 0.5 + \epsilon$ or $0.5, 0.5$, randomly chosen. The rewards for clients $2, 3, \ldots, 8$ are 0 at all times. The mean reward values for arms 1 and 2 of clients 9 and 10 are 0.5. We have specified these parameters to ensure relatively low computational complexity while meeting the requirements for constructing the hard instances. In Instance 1 and 2, we compare all algorithms and report their exact regret to examine whether the possibly non-optimal algorithms, which are optimal in more general settings like Instance 3 and 4, have the potential to outperform the so-called optimal algorithms in worst-case scenarios. Likewise, in Instance 3 and 4, we compare FEDEXP3 and EXP3 and derive their regret values to draw similar conclusions.

The evaluation metric is the empirical regret, calculated by averaging R_T^π as defined in Section 2 over 50 runs. In contrast, the communication cost can be computed explicitly, which is discussed in Section 4.2.

4.1 Comparison Results

Next, we present the regret performances of the aforementioned algorithms on instances 1, 2, 3, and 4, as shown in Figures 1, 2, 3, and 4, respectively. In these figures, the x-axis represents the time steps, and the y-axis represents the corresponding cumulative regret up to the corresponding time step. Furthermore, we fit the regret curves based on the theoretical regret order and report the coefficients of these curves, which ultimately demonstrate the performance comparison among the algorithms.

Figure 1 shows the regrets of CoopUCB, GosInE, Gossip_UCB, EXP3, and FedEXP3 on Instance 1. We note that, except for Gossip_UCB, all other algorithms exhibit sublinear regret, which is due to the fact that Gossip_UCB relies on the assumption that the spectral gap of the graph is strictly positive. Our constructed instance violates this assumption and thus serves as a worst-case scenario. Among the remaining algorithms, CoopUCB and GosInE have much smaller regrets compared to EXP3 and FedEXP3, which coincides with their optimal regret bounds. More precisely, the coefficients of the regret curves, with respect to $\log t$, for CoopUCB, GosInE, EXP3, and FedEXP3 are 0.51, 2.94, 6.46, and 3.84, respectively. The fitted regret curves are presented in Figure 4. This indicates room for improvement for EXP3 and FedEXP3, since they exhibit $\log t$ regret on hardest instances. Perhaps their regret in the stochastic setting is $O(\log T)$ despite being designed for adversarial reward. Such reasoning has a gap since there might be other hard instances.

For all the aforementioned algorithms, their regrets on Instance 2, where the graph is disconnected, are shown in Figure 2. Not surprisingly, all the algorithms exhibit linear-style regrets, consistent with our established regret lower bounds. Among them, FedEXP3 and EXP3 perform much better, despite their theoretical regret bounds of order $T^{\frac{2}{3}}$ and $\sqrt{T}$, respectively, in the adversarial setting on complete graphs, which are less favorable compared to the $\log T$ obtained by CoopUCB, Gossip_UCB, and GosInE with stochastic rewards on connected graphs. More specifically, FedEXP3 has the smallest regret, even though its regret bound is the largest. In the meantime, CoopUCB exhibits the largest regret, as it assumes homogeneous rewards in its original analysis and thus is more sensitive to this worst-case scenario. This suggests the robustness of FedEXP3 and EXP3 in worst-case scenarios. More surprisingly, it demonstrates and motivates studying the trade-off between regret bounds and robustness. Regarding the linear relationship, we also examine the coefficients of the regret curves. Specifically, the coefficients for CoopUCB, GosInE,

Gossip_UCB, EXP3, and FedEXP3, in terms of t, are 0.008, 0.009, 0.007, 0.008, and 0.006, respectively, which aligns with the above comparisons among the algorithms. The fitted regret curves are presented in Figure 5.

We next focus on Instance 3, where the graph is complete and rewards are adversarial. We show the regret of FedEXP3 and EXP3 in Figure 3. Consistently, both exhibit sublinear regret based on their theoretical bounds. Remarkably, unlike the theoretical regret bounds, which are of order $T^{\frac{2}{3}}$ and T respectively and suggest that FedEXP3 should have a larger regret, FedEXP3 actually leads to smaller regret in this worst-case scenario. More precisely, with respect to the function $t^{\frac{1}{2}}$, the coefficients of the regret curves for EXP3 and FedEXP3 are 0.967 and 0.273, respectively. The fitted regret curves are presented in Figure 6. This demonstrates the superior performance of FedEXP3 over the theoretically optimal one, namely EXP3. It implies it might be possible to derive a smaller theoretical bound for FedEXP3 when constrained to complete graphs. Moreover, this conclusion highlights the differences between reward dynamics and graph dynamics, uncovering the complexity of multi-agent systems and necessitating improvements in the dependency of regret on more precise graph complexities, especially since an unexpected conclusion is drawn from Instance 1.

Lastly, on Instance 4, where the rewards are adversarial and the graph is connected (path), we present the regret performances of FedEXP3 and EXP3 in Figure 7. We observe that both algorithms result in much larger regret compared to the complete graph case. FedEXP3 performs slightly better than EXP3, which is again validated by the coefficients of the regret curves with respect to $t^{\frac{2}{3}}$; the coefficients for EXP3 and FedEXP3 on this instance are 0.053 and 0.033, respectively. The fitted regret curves are presented in Figure 8. EXP3 exhibits limited learning process, coinciding with its less refined linear regret bound (which any algorithm meets). In contrast, FedEXP3 grows sublinearly, considering its regret bound of order $T^{\frac{2}{3}}$, which demonstrates its robustness in this extremely hard case. It is worth noting that EXP3 and FedEXP3 are still closely matched, which suggests the potential for establishing a smaller regret bound for EXP3, beyond the naive linear regret bound.

In summary, through these experiments on our novel hard instances, we not only validate the performance of existing algorithms and observe the proven regret lower bounds, but also propose the following crucial research questions. First, it is important to demonstrate the trade-off between regret bounds and robustness, especially when the regret bounds themselves are optimal in T. This allows us to push the Pareto front by developing new algorithms. Meanwhile, there is certainly room for refining the analytical bounds of existing algorithms, beyond the specific settings where their regret bounds are proven. Moreover, it would be informative to characterize how the regret bound of an algorithm continuously varies with problem settings, especially in edge cases (boundaries), instead of considering an algorithm only in a specific setting. A worst-case scenario in setting A may be a relatively good case in setting B when A subsumes B, which implies a possibly smooth transition between them. By doing so, we can more precisely select the best algorithm based on different problem settings.

4.2 Discussion of Computational Complexity

While the main focus of this paper is on regret bounds, consistent with most of the existing literature, computational complexity has been gaining recent attention. It is worth noting that for stochastic settings, the aforementioned algorithms—Gossip_UCB, CoopUCB, GosInE—have a time complexity of order $O(M^2 + M \cdot K)$, as in [Xu and Klabjan, 2023a]. In the meantime, the time complexity of FedEXP3 and EXP3 is also $O(M^2 + M \cdot K)$, resulting from message exchanges and arm updates. We would like to highlight that this complexity can be reduced by incorporating a parallel mechanism where all agents perform executions in parallel. This raises further considerations of synchronous vs asynchronous, which goes beyond the scope of this paper and is thus left for future research.

5 Conclusion

In this paper, we conduct a comprehensive study on the regret lower bounds in a decentralized multi-agent MAB framework across various settings, which provides an understanding of the fundamental challenges posed by different problem settings and insights into the development of optimal algorithms. Specifically, we establish instance-dependent and mean-gap independent lower bounds for stochastic settings, which are of order $\log T$ and $\sqrt{T}$, respectively, for all existing graphs. These results are consistent with the existing upper and lower bounds, showing their tightness and consistency, respectively. Additionally, we introduce a novel problem instance in adversarial settings that leads to a regret lower bound of order $\Omega(T^{\frac{2}{3}})$. This finding bridges the gap between the existing lower and upper bounds and highlights the distinction between the multi-agent and single-agent counterparts. In the following table, we reaffirm the tightness of the proved lower bounds by comparing them with existing regret upper bounds of algorithms for instance-dependent, mean-gap independent, and adversarial scenarios. More specifically, we consider DDUCB in [Martínez-Rubio *et al.*, 2019], FedDr-UCB in [Xu and Klabjan, 2023a], GosInE in [Chawla *et al.*, 2020], LCC-UCB in [Agarwal *et al.*, 2022], UCB-warm in [Jia *et al.*, 2021], and FEDEXP3 in [Yi and Vojnović, 2023]. The table indicates that the orders of the lower bounds and upper bounds match.

	Upper & Lower	Algo.
Instance	$\log T$	DDUCB, FedDr-UCB GosInE
Mean-gap	$\sqrt{T}$	LCC-UCB, FedDr-UCB UCB-warm
Adversarial	$T^{\frac{2}{3}}$	FEDEXP3

Furthermore, we uncover worst-case scenarios in multi-agent MAB settings by demonstrating a linear regret when the graphs are disconnected, which adds to the difference between multi-agent and single-agent MAB.

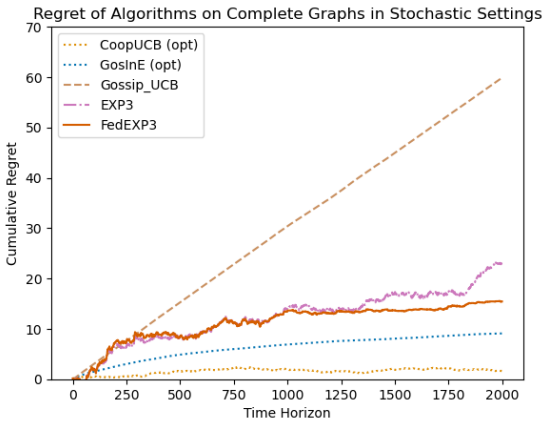


Figure 1

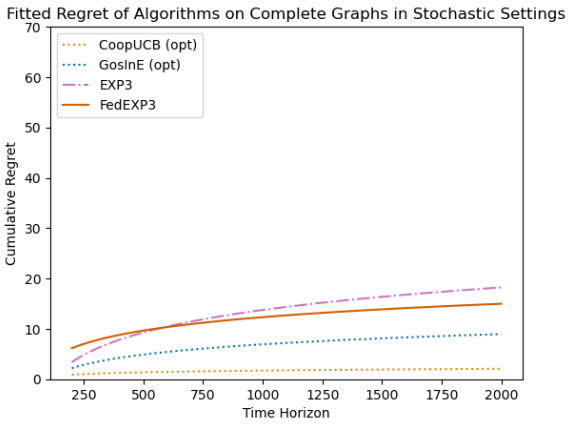


Figure 4

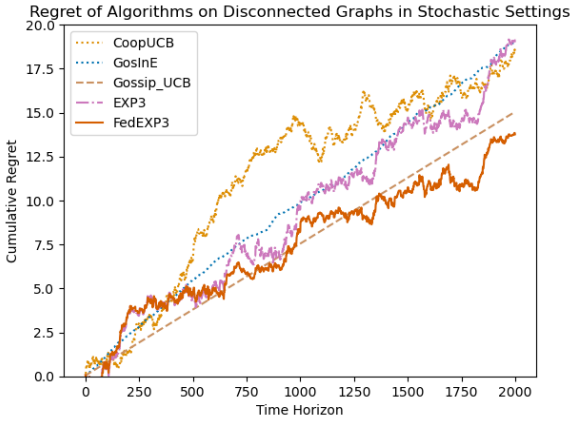


Figure 2

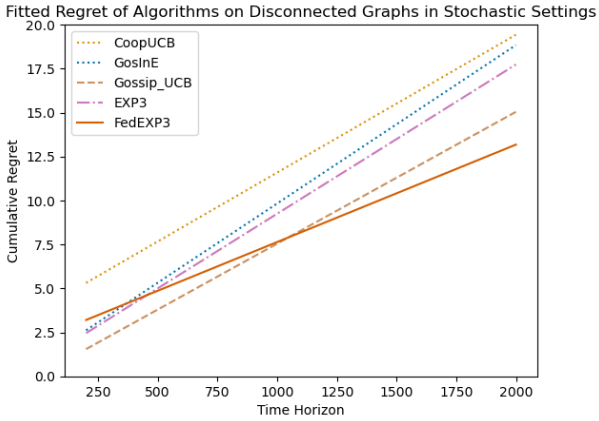


Figure 5

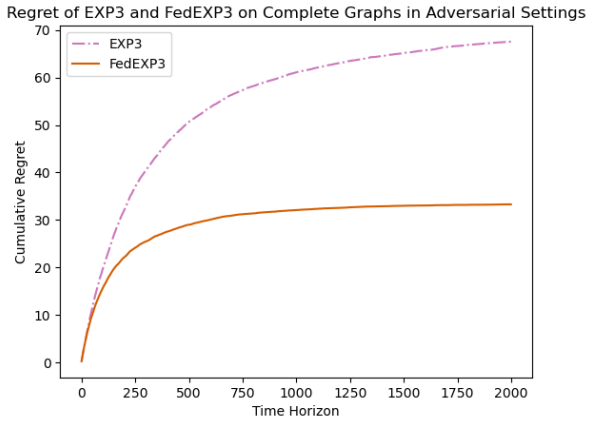


Figure 3

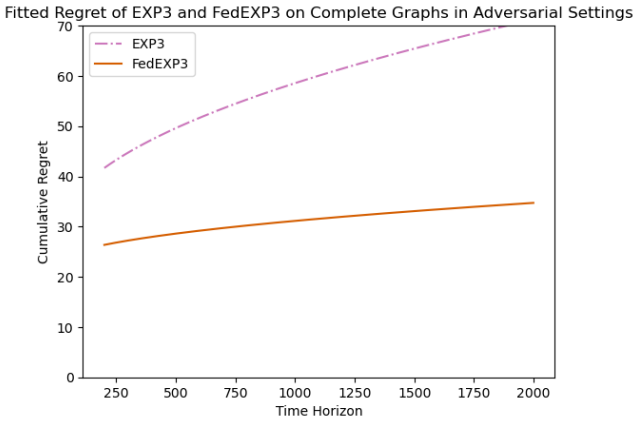


Figure 6

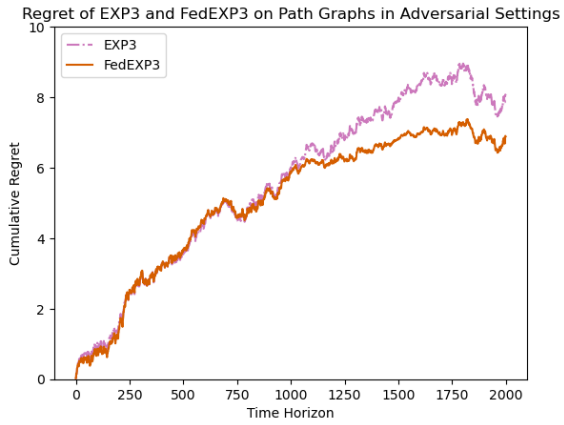


Figure 7

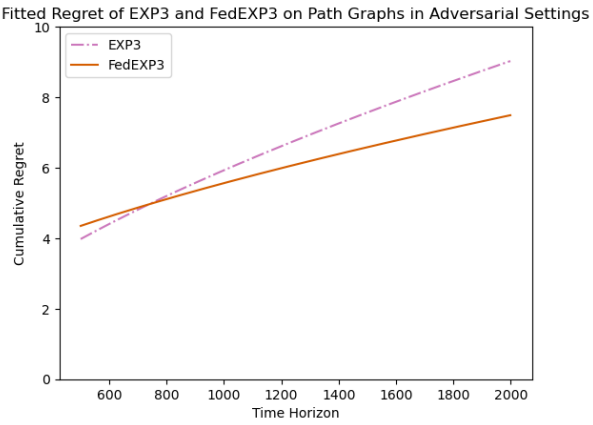


Figure 8

6 Future Work and Implications

Although the regret lower bounds match the upper bounds of some existing algorithms, implying their optimality with respect to T, this indicates that there is no possibility of achieving a smaller order in T by developing new algorithms, thereby diverting such efforts. Supported by the differences between the actual regret observed in numerical experiments and the theoretical regret bounds, we point out that the constants in terms of M, K, and graph complexity leave room for improvement through better algorithms or analyses and for measuring and improving the robustness of algorithms. One potential approach could be to measure how regret varies across problem settings, e.g., through derivatives. Ideally, these potential algorithms could push the Pareto front regarding the regret bounds and robustness. Such development may start with our constructed instances (worst-case scenarios) in the theorems. Therefore, as a next step, we suggest exploring novel algorithms with smaller coefficients that are closer to the established lower bounds and robust enough to adapt to changes in problem settings. More generally, with the recent growth of large-scale systems, it is promising to explore the dependency on M and K for both regret upper and lower bounds, considering the optimal order of T. Moreover, the problem complexity of multi-agent systems, as indicated in the numerical experiments, necessitates exploring the dependency on graph complexity induced by M clients, such as spectrum and degree. These characterizations would greatly facilitate a framework that precisely shows the effect of M, K, and graph complexity, rather than focusing solely on T. From a practical perspective, although regret bounds have been the main focus in most existing literature, computational efforts are no longer ignorable in these large-scale systems. Consequently, moving forward, it is crucial to examine the trade-off between regret and computational complexity, potentially using Pareto optimization.

References

Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *The Journal of Machine Learning Research*, 23(1):9529–9552, 2022.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International conference on artificial intelligence and statistics*, pages 3471–3481. PMLR, 2020.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34:27057–27068, 2021.

Huiwen Jia, Cong Shi, and Siqian Shen. Multi-armed bandit with sub-exponential rewards. *Operations Research Letters*, 49(5):728–733, 2021.

Fan Jiang and Hui Cheng. Multi-agent bandit with agent-dependent expected rewards. *Swarm Intelligence*, pages 1–33, 2023.

Emmanuel Kowalski. *An introduction to expander graphs*. Société mathématique de France Paris, 2019.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control*, pages 167–172. IEEE, 2016.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference*, pages 243–248. IEEE, 2016.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multiagent multi-armed bandits. *Automatica*, 125:109445, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Aritra Mitra, Hamed Hassani, and George Pappas. Exploiting heterogeneity in robust federated best-arm identification. *arXiv preprint arXiv:2109.05700*, 2021.

Clémence Réda, Sattar Vakili, and Emilie Kaufmann. Near-optimal collaborative learning in bandits. *Advances in Neural Information Processing Systems*, 35:14183–14195, 2022.

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

Zhi Wang, Chicheng Zhang, Manish Kumar Singh, Laurel Riek, and Kamalika Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 1531–1539. PMLR, 2021.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 2023.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pages 38499–38517. PMLR, 2023.

Zirui Yan, Quan Xiao, Tianyi Chen, and Ali Tajer. Federated multi-armed bandit via uncoordinated exploration. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5248–5252. IEEE, 2022.

Jialin Yi and Milan Vojnović. Doubly adversarial federated bandits. *arXiv preprint arXiv:2301.09223*, 2023.

Jingxuan Zhu and Ji Liu. Distributed multi-armed bandits. *IEEE Transactions on Automatic Control*, 2023.

Jingxuan Zhu, Romeil Sandhu, and Ji 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.

Jingxuan Zhu, Ethan Mulle, Christopher Salomon Smith, and Ji Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *arXiv preprint arXiv:2111.10933*, 2021.

Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. In *ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.