Regret Lower Bounds in Multi-agent Multi-armed Bandit

Anonymous submission

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. While previous works have explored these settings with upper bounds, we provide a thorough study on tight lower bounds.

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, Robbins 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 healthcare and autonomous driving, as well as the increasing demand for large-scale distributed decision learning processes in sensor networks and robotic systems. 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 chal-

lenges 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, Hassani, and Pappas 2021; Réda, Vakili, and Kaufmann 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, Kanade, and Rebeschini 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, Srivastava, and Leonard 2016b,a, 2021; Zhu, Sandhu, and Liu 2020; Martínez-Rubio, Kanade, and Rebeschini 2019; Agarwal, Aggarwal, and Azizzadenesheli 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; Zhu, Sandhu, and Liu 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 (Zhu and Liu 2023). 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, Shi, and Shen 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, Shi, and Shen 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.

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 $\sqrt{T}$.

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. Finally, we summarize the paper and point out future possibilities based on the findings.

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 $r^m_i(t)$, which is independently and identically distributed across time with a mean value of μ^m_i. 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^t_m and obtains the corresponding reward $r^m_{a^t_m}(t)$ 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.

Following the common definition of the global reward, we define the global reward of arm i as $r_i(t) = \frac{1}{M}\sum_{m=1}^{M} r^m_i(t)$, and the corresponding expected global reward as $\mu_i = \frac{1}{M}\sum_{m=1}^{M}\mu^m_i$. 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) $\pi = \{a^t_m\}^{1\leq t\leq T}_{1\leq m\leq M}$. 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 $\sigma_F^{t,m} = \sigma(\{\{I_j^s\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$ where I_j^s represents the information of all arms contained at client j at time step s and, denote $\sigma_B^{t,m} = \sigma(\{\{I_j^s(a_s^j)\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$ 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.

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 $R_T^F \le R_T^B$.*

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 $\min_{\pi\in\Pi_F} R_T^\pi \le \min_{\pi\in\Pi_B} R_T^\pi$, or equivalently $R_T^F \le R_T^B$. □

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; we summarize the main idea 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. □

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. These 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. The instance-dependent regret bound presents non-trivial challenges to the analysis. We start with complete graphs. We specify $K = 2$ and assume $\mu_1 > \mu_2$ without loss of generality. Consider the centralized problem which has times when the clients pull the same arm (agreement) and times when the clients pull distinct arms (disagreement). We denote the number of time steps of agreement and disagreement as T_a and T_d, respectively. We observe that $T_a + T_d = T$. For T_d, there exist clients pulling the worse arm, which implies that for any policy $\pi \in \Pi_B$

$$\begin{aligned} R_T^\pi &= \frac{1}{M}\sum_m\sum_{t\in T_d}(\mu_1 - \mu_{a_t^m}) + \frac{1}{M}\sum_m\sum_{t\in T_a}(\mu_1 - \mu_{a_t^m}) \\ &= \sum_{t\in T_d}\Delta_2 + \frac{1}{M}\sum_m\sum_{t\in T_a}(\mu_1 - \mu_{a_t^m}) \\ &= T_d\Delta_2 + \frac{1}{M}\sum_m\sum_{t\in T_a}(\mu_1 - \mu_{a_t^m}). \end{aligned} \tag{1}$$

Note that when $T_d = \Omega(\log T)$, we immediately derive that $E[R_T^B] \ge \Omega(\log T)$, which concludes the proof.

From now on, we assume $T_d = o(\log T)$, which implies that $T_a = T - o(\log T)$ and $\frac{T_a}{T} \to 1$ as T goes to ∞. We denote the value $t_0 = \log T$ and divide the time horizon into $\bigcup_{j=0}^{t_0}[2^j, 2^{j+1}-1]$. It is clear that 1) the number of intervals is $\log T$ and 2) the length of the j^{th} interval is 2^{j-1}. Let $t_d = \max\{t | t \in T_d\} + 1$. Since $T_d = o(\log T)$, we have $|[t_d, T]| \ge 2^{\frac{1}{2}\log T}$ for all large enough T.

Meanwhile, we observe that for T_a, it is equivalent to a single-agent multi-objective bandit problem (Xu and Klabjan 2023b)

since the global reward of a single arm i is given as a reward vector $(r_i^{m,t})_{m=1}^M$ and is revealed to all the clients at each time step.

Note that $\frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1-\mu_{a_t^m}^m) = \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1-\mu_{a_t}^m) = \sum_{t\in T_a}(\mu_1-\mu_{a_t})$ where the first equality is by the definition of T_a and the second equality uses the definition of μ_1 and μ_{a_t}. We denote $T_a^d = T_a \cap [t_d, T] = [t_d, T]$.

At the same time, the Pareto pseudo regret reads $R_{T_a^d,M} = Dist(\sum_{t\in T_a^d}(\mu_{a_t}^m)_m, O)$ where $Dist(\cdot)$ is the distance measure between a reward vector and the Pareto optimal set O as introduced in (Xu and Klabjan 2023b), and satisfies that $R_{T_a^d,M} \geq \Omega(\log T_a^d)$ for any policy $\{a_t\}$ based on Theorem 6 in (Xu and Klabjan 2023b).

By specifying the rewards homogeneous, i.e. $\mu_{a_t}^1 = \mu_{a_t}^2 = \ldots = \mu_{a_t}^M$ and following a similar analysis as on Theorem 6 in (Xu and Klabjan 2023b), we obtain $R_{T_a^d,M} = Dist(\sum_{t\in T_a^d}(\mu_{a_t}^m)_m, O) = \sum_{t\in T_a^d}(\mu_1-\mu_{a_t})$ which yields

$$\begin{aligned}\sum_{t\in T_a}(\mu_1-\mu_{a_t}) &\geq \sum_{t\in T_a^d}(\mu_1-\mu_{a_t})\\ &\geq \Omega(\log T_a^d) = \Omega(\log(2^{\frac{1}{2}\log T})) = \Omega(\log T). \qquad (2)\end{aligned}$$

To put everything together, we have that for any policy $\pi \in \Pi_B$ $R_T^\pi \geq T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1-\mu_{a_t^m})) \geq \Omega(\log T)$ where the second inequality holds by (2).

Subsequently, we obtain $\min_{\pi\in\Pi_B} R_T^\pi \geq T_d\Delta_2 + \frac{1}{M}\sum_m \sum_{t\in T_a}(\mu_1-\mu_{a_t^m})) \geq \Omega(\log T)$, which concludes the analysis of complete graphs.

The remaining cases follow from the monotonicity of the regret in the graph complexity as follows. We first consider the full-information setting. For any $0 < c \leq 1$, we denote $\sigma_c^t = \sigma(\{\{I_j^s\}_{j\in\mathcal{N}_m^c(s)}\}_{s\leq t})$. We observe that $\sigma_1^t = \sigma(\{I_1^s, \ldots, I_M^s\}_{s\leq t})$. We have $\sigma_c^t \subset \sigma_1^t$. We define policy set Π_c as $\{f_t\}$ where the domain of f_t is on σ_c^{t-1}.

For any policy $\pi \in \Pi_c$, i.e. $\pi = \{h_t\}_{t=1}^T$, we have that it only leverages the neighborhood information σ_c^{t-1} to determine a decision rule at each time step. Since $\sigma_c^{t-1} \subset \sigma_1^{t-1}$, σ_1^{t-1} also has the neighborhood information that h_t requires. This leads to $\pi \in \Pi_1$, and subsequently yields $\Pi_c \subset \Pi_1$. We hence obtain that in the full-information setting $\min_{\pi\in\Pi_1} R_T^\pi \leq \min_{\pi\in\Pi_c} R_T^\pi$.

By the above discussion on c and the statement for complete graphs, or equivalently, with respect to Π_1, we obtain $\Omega(\log T) \leq \min_{\pi\in\Pi_1} R_T^\pi$, in the instance-dependent sense and subsequently $\Omega(\log T) \leq \min_{\pi\in\Pi_c} R_T^\pi$.

By Theorem 1, we have $R_T^B \geq \Omega(\log T)$. This completes the E-R case. All remaining cases follow the same logic.

□

Remark. *While (Martínez-Rubio, Kanade, and Rebeschini 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.*

Mean-gap independent

Apart from the instance-dependent regret lower bounds, we also investigate the mean-gap independent regret lower bound that is 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. *Similarly, this result aligns with the lower bound established in the single-agent MAB setting. Furthermore, this lower bound of order $\sqrt{T}$ corresponds to the mean-gap upper bounds presented in (Xu and Klabjan 2023a) and (Jia, Shi, and Shen 2021) for multi-agent and single-agent MAB problems, respectively. This consistency further shows the tightness of the lower bound we have derived.*

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 $M \bmod 4 = 0$ 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 $D_{TV}(Q_j^{-1}, Q_j^i) \leq \sqrt{\frac{D_{KL}(Q_j^{-1}, Q_j^i)}{2}} \leq \epsilon\sqrt{2d}$. (3)

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

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

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

$$E[R_T^B]$$
$$= E[E[R_T^B|X_1, \ldots, X_D]]$$
$$= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(T - n_{m,1}(T)))|X_1, \ldots, X_D]]$$
$$= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(\sum_{j=1}^{D} d - \sum_{j=1}^{D} n_{m,1}(T,j)))|X_1, \ldots, X_D]]$$
$$= E[\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D} E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_1, \ldots, X_D]]$$
$$= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D} E[E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j]]$$
$$= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}}\frac{E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j = i]]}{|I_0|+1}$$
$$\geq \frac{1}{2M^2}(\frac{1}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}} E[\epsilon\cdot(d - n_{1,1}(T,j))|X_j = i])$$
$$= \frac{1}{2M^2}(\epsilon\cdot T - \frac{\epsilon}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}} E_{Q_j^i}[(n_{1,1}(T,j))]) \quad (4)$$

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

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

where the second inequality uses $\sum_i Q_j^{-1}(a_t^1 = 1) = 1$ and $dD = T$, and the third inequality uses (3), 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

$$E[R_T^B] \geq \frac{1}{2M^2}(\epsilon \cdot T - \frac{\epsilon}{|I_0|+1}(T + T \cdot \frac{|I_0|+1}{4})) \geq \frac{1}{2M^2}\frac{1}{4}\epsilon \cdot T = \Omega(T^{\frac{2}{3}}) \quad (5)$$

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.*

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. 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. As a next step, we suggest exploring novel algorithms with smaller coefficients that are close to the lower bounds established herein. As a concluding remark, how to show high probability lower bounds remain an important yet unexplored area of research.

References

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

Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, 23–35. PMLR.

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

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

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

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

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

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

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

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

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2016a. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control*, 167–172. IEEE.

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

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

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

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

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

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

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

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

Xu, M.; and Klabjan, D. 2023a. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *arXiv preprint arXiv:2306.05579*.

Xu, M.; and Klabjan, D. 2023b. Pareto Regret Analyses in Multi-objective Multi-armed Bandit. In *International Conference on Machine Learning*, 38499–38517. PMLR.

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

Yi, J.; and Vojnović, M. 2023. Doubly adversarial federated bandits. *arXiv preprint arXiv:2301.09223*.

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

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

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

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

Proof of Results in Section 4

Proof of Theorem 2

Proof. On a complete graph, each client can observe the rewards of all arms at M clients, where the number of observations is thereby upper bounded by KM. Henceforth, we consider Theorem 4 in (Shamir 2014) to obtain

$$R_T^F \geq \sqrt{\frac{KT}{1+KM}} = \Omega(\sqrt{T}).$$

This completes the first part of the statement.

For the instance-dependent regret lower bounds, we assume that the number of arms is 2 and the rewards of arms satisfies the assumptions in (Goldenshluger and Zeevi 2013). Then based on the result established by specifying a contextual linear bandit with $\alpha = 1$ as in (Goldenshluger and Zeevi 2013), which reads as Theorem 2, we obtain

$$R_T^F \geq \Omega(\log T).$$

We add that the lower bound result for the bandit setting holds for the full-information setting by noting the analysis essentially uses the observations that are given by the full information setting.

This concludes the instance-dependent lower bound in the full information setting and thereby completes the proof.

□

Proof of Theorem 4

Proof. Consider a disconnected graph G with a clique connected component C_G including clients $c_1, \ldots, c_Q$ without loss of generality. Since G is disconnected, for any other node $m \notin V(C_G)$, there is no path between m and any node in C_G.

Let $\Delta > 0$. For client $m \notin C_G$, the reward distributions read as $(\frac{M-1}{M-Q}\Delta, 0, \ldots, 0)$, which indicates that the optimal arm is arm 1. For client $m \in C_G$, however, the reward distribution reads as $(0, \frac{2}{Q}\Delta, 0, \ldots, \ldots, 0)$, implying that arm 2 is the optimal arm. It is straight-forward that the global mean reward value of arm 1 is $\frac{(M-1)}{M}\Delta$ that is larger than that of arm 2 which is $\frac{2\Delta}{M}$. The subsequent sub-optimality gap is $\Delta_2 = \frac{M-3}{M}\Delta$. Any no-regret (consistent as proposed in (Lattimore and Szepesvári 2020)) algorithms π at client $j \in C_G$, where the regret with respect to the available information is defined on the rewards of client $j \in C_G$, leads to $E[n_{j,2}(T)] = O(T)$. However, in this situation, the global regret satisfies

$$\begin{aligned} E[R_T^\pi] &= \frac{1}{M}\sum_m \sum_{t=1}^T (E[\mu_1 - \mu_{a_t^m}]) \\ &\geq \frac{1}{M}\sum_{t=1}^T (E[\mu_1 - \mu_{a_t^j}]) \\ &\geq \frac{1}{M} E[n_{j,2}(T)] \cdot \Delta_1 \\ &= \frac{1}{M} \cdot \frac{M-3}{M}\Delta \cdot \Omega(T) = \Omega(T) \end{aligned}$$

where the first inequality is by only considering client j and the second inequality uses the fact that arm 2 is not a global optimal arm.

This completes the proof of the linear regret in the case when clients perform local consistent learning on disconnected graphs. □

Proof of Theorem 5

Proof. Again, we consider a disconnected graph G with a clique C_G including clients $c_1, \ldots, c_Q$ without loss of generality.

We assume there are two arms labeled as arm 1 and 2 and consider the instance at clients as follows by referencing (Alon et al. 2015). 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 for any $j \notin C_G$, $r_k^j(t) = \begin{cases} X & \text{arm 1} \\ \frac{1}{2} & \text{arm 2} \end{cases}$ and for any $j \in C_G$, we have $r_k^j(t) = \begin{cases} \frac{1}{2} & \text{arm 1} \\ \frac{1}{2} & \text{arm 2} \end{cases}$ where the random variable X is independent of everything at client $j \in C_G$ as client $j \in C_G$ only has the information of their own arms. We have $\Delta_2 = \frac{1}{2(M-Q)}$, no matter what value X takes since it

only changes the choice of optimal arms. Specifically, when $X = 1$, the global optimal arm is arm 1 and the suboptimality gap is $\Delta_2 = \mu_1 - \mu_2 = (1 - \frac{1}{2})/(M - Q)$. When $X = 0$, the global optimal arm is arm 2 and the suboptimality gap is $\Delta_2 = \mu_2 - \mu_0 = (\frac{1}{2} - 0)/(M - Q)$, the other way around.

Subsequently, we consider the regret at client $j \in C_G$ to obtain

$$
\begin{aligned}
&E[R_T^\pi] = \frac{1}{M}\sum_m \sum_{t=1}^{T}(E[\mu_* - \mu_{a_t^m}]) \\
&\geq \frac{1}{M}\sum_{t=1}^{T}(E[\mu_* - \mu_{a_t^M}]) \\
&= \frac{1}{M}(\frac{1}{2}E[\Delta n_{j,1}(T)|X = 0] + \frac{1}{2}E[\Delta(T - n_{j,1}(T))|X = 1]) \\
&= \frac{1}{M}(\frac{1}{2}E[\Delta n_{j,1}(T)] + \frac{1}{2}E[\Delta(T - n_{j,1}(T))]) \\
&= \frac{\Delta}{4M(M-Q)}T = \Omega(T)
\end{aligned}
$$

where the first inequality uses the non-negativity of value $\mu_* - \mu_{a_t^m}$ and the third equality leverages the independence between X and client j. □

Proof of Theorem 6

Proof. We show the mean-gap free regret lower bound starting with complete graphs. Note that a complete graph is equivalent to a centralized problem with M agents. This implies that each client can observe the reward of multiple arms by communicating with $M - 1$ neighbors, where the number of observations is thereby upper bounded by M. Henceforth, we consider Theorem 4 in (Shamir 2014) and obtain

$$R_T^B \geq \sqrt{\frac{KT}{1+M}} = \Omega(\sqrt{T}).$$

This completes the proof of the complete graphs.

Regarding the monotonicity of the regret in the graph complexity, the proof follows the proof of Theorem 3.

□

Proof of Theorem 8

Proof. Note that the graph structure determines the communication efficiency of the clients. To consider the lower bound, we leverage sparse graphs in the connected graph family to perform the worst-case scenario analysis.

Specifically, we consider the designed graph consisting of clients $1, \ldots, M$ in this order. It takes exactly $O(M)$ time steps for client 1 to obtain the information of client M, which results in a deterministic delay.

If $I_0 = \{1, \ldots, \frac{M}{4}\}$ and $I_1 = \{\frac{3M}{4}, \ldots, M\}$, then the shortest path d_p from I_0 to I_1 meets the condition

$$d_p \geq \Omega(\frac{M+1}{3}).$$

By the choice of M such that $M > \Omega(T^{\frac{1}{3}})$, we obtain

$$d_p \geq \Omega(T^{\frac{1}{3}}). \tag{6}$$

We star with a full-information setting. Following a similar argument and constructing the same instance as in Lemma A.4 in (Yi and Vojnović 2023), we arrive that in the full-information setting

$$R_T \geq \Omega(\sqrt{d_p \cdot T}).$$

Subsequently, we obtain that

$$
\begin{aligned}
R_T &\geq \Omega(\sqrt{d_p \cdot T}) \\
&= \Omega(\sqrt{T} \cdot \sqrt{d_p}) \\
&\geq \Omega(\sqrt{T} \cdot T^{\frac{1}{6}}) = \Omega(T^{\frac{2}{3}})
\end{aligned}
$$

where the last inequality is by (6). Equivalently, we write it as

$$R_T^F \geq \Omega(T^{\frac{2}{3}}). \tag{7}$$

Meanwhile, by Lemma 1, we have that the regret lower bound in the bandit setting is larger than the regret in the full information setting and thus by (7) we obtain

$$R_T^B \geq \Omega(T^{\frac{2}{3}}).$$

This completes the proof of Theorem 8.

□