Multi-agent Multi-armed Bandit Regret Complexity and Optimality

Mengfan Xu
mengfanxu@umass.edu
Mechanical and Industrial Engineering
University of Massachusetts Amherst
Amherst, MA, USA

Diego Klabjan
d-klabjan@northwestern.edu
Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, USA

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.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

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 (12) establishes an asymptotic regret lower bound of order $O(\log T)$ for consistent algorithms using an elegant regret decomposition approach that incorporates KL-divergence. The key idea behind these results is to construct problem instances where the optimal arm is very close to the sub-optimal arms but not too close, making it challenging for the player to distinguish between them and resulting in a risk of getting less rewards and significant regret. The gap is precisely chosen and is the main technique.

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

Similar to the categorization in the traditional MAB framework, problem settings in multi-agent MAB are classified as either stochastic or adversarial, depending on the nature of reward distributions. In stochastic multi-agent MAB, the rewards for each client are independently and identically distributed over time, while in adversarial multi-agent MAB, the rewards are chosen by an adversary. Assuming the existence of a central server addresses the problem where the central server can communicate all clients' information. However, the assumption of centralization may not be realistic in real-world scenarios, where clients are often limited to pairwise transmissions constrained by underlying graph structures. A fully decentralized framework characterized by means of graph structures has been proposed in several studies. This decentralized approach removes the centralization assumption, making it more general while introducing non-trivial challenges. To this end, certain assumptions on the graphs are incorporated in these studies. Examples include complete graphs (20), regular graphs (10), and connected graphs under the doubly stochasticity assumption (26; 27). 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 (25), as well as random graphs, including the Erdős-Rényi model and random connected graphs (21). 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, (9) 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, (21) 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 (9) up to a logarithmic factor. However, despite these advancements in the regret upper bounds, the corresponding regret lower bounds in the mean-gap independent sense have not yet been explored. Addressing this research gap is one of the primary objectives of this paper.

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

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

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

Apart from the instance-dependent regret lower bounds of order $\Omega(\log T)$, we further extend our analysis to mean-gap independent regret lower bounds, presenting a novel contribution as well. Specifically, we establish mean-gap independent regret bounds of order $\Omega(\sqrt{T})$, which not only validate near optimality of the algorithm proposed in (21) 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 (24) 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 in Appendixes A and B.

2 Related Work

Classical MAB MAB has a rich history, with regret bounds extensively studied in both instance-dependent

and mean-gap independent settings, as well as in stochastic and adversarial scenarios. In stochastic settings, where the reward distribution is time-invariant, numerous studies have established instance-dependent regret upper bounds of order $\log T$. The work of (9) characterizes mean-gap independent regret upper bounds of order $\sqrt{T}$. In adversarial settings, where the reward distribution can change over time, existing work has demonstrated a regret upper bound of order $O(\sqrt{T})$. However, these algorithms cannot be directly applied to multi-agent MAB problems due to the collaborative nature required among multiple agents.

Regret Lower Bounds Regret lower bounds are critical for understanding the problem complexity of MAB and for claiming the optimality of algorithms. (12) established the first asymptotic regret lower bounds of order $O(\log T)$ using KL-divergence. Subsequent work, such as (15), relaxed these assumptions, deriving regret bounds for two-arm settings. For mean-gap independent cases, (19) introduced regret bounds of order $\sqrt{T}$, constructing problem instances where distinguishing between arms is deliberately challenging.

Multi-Agent MAB Centralized and Decentralized: The multi-agent MAB framework has gained prominence due to its relevance in distributed systems. To address this, previous work has extensively studied settings that incorporate a central server, also referred to as a controller, as discussed in (3; 28; 8; 17; 18; 23). In this setup, the central server integrates and distributes information among the clients at each time step, leading to a regret upper bound of order $O(\log T)$ in stochastic multi-agent MAB, matching the regret bounds in stochastic MAB. However, despite being mentioned in (16) regarding the instance-dependent lower bound of order $\log T$, a formal lower bound statement in this centralized structure remains unexamined. A fully decentralized framework, characterized by graph structures, has been proposed in several studies (13; 14; 27; 16; 1; 20; 10; 26; 28). This decentralized approach removes the centralization assumption, making it more general but introducing non-trivial challenges. To address these, studies incorporate specific assumptions on graph types, such as complete graphs (20), regular graphs (10), and connected graphs under the doubly stochasticity assumption (26; 27). Across all cases, regret upper bounds of order $O(\log T)$ remain consistent with those in traditional MAB settings.

Recent research has also explored time-varying graphs, such as B-connected graphs under doubly stochasticity (25), and random graphs, including Erdős-Rényi and random connected graphs (21). In these cases, regret upper bounds similarly maintain the order $O(\log T)$. However, corresponding regret lower bounds have not yet been addressed in existing literature, a key focus of this study.

Stochastic and Adversarial: In classical stochastic settings, (5) investigated an adversarial multi-agent MAB problem and provided a regret upper bound of order $\sqrt{T}$, demonstrating consistency with adversarial MAB problems under the EXP3 algorithm. More recently, (24) examined heterogeneous adversarial environments, where adversaries vary across clients. The heterogeneity introduces significant challenges, resulting in a regret upper bound of order $O(T^{\frac{2}{3}})$, larger than the standard MAB regret bound of $\sqrt{T}$. Furthermore, they established a regret lower bound of order $\sqrt{T}$, which, while informative, remains smaller than their proposed upper bound. By leveraging results from (19) and constructing problem instances with mini batches of adversarial rewards, they provided a foundation for further exploration of these bounds.

Instance-Free and Mean-Gap Independent Regret: The concept of mean-gap independent regret, introduced by (9), ensures robustness across varying gap sizes Δ_i. Recent work, such as (21), extended this idea to decentralized settings, achieving regret upper bounds of $O(\sqrt{T}\log T)$. However, lower bounds for mean-gap independent regret in these settings remain unexplored, another focus of this study.

In standard MAB, (9) introduced a regret upper bound of order $\sqrt{T}$, independent of the sub-optimality gap Δ_i, which represents the difference between the mean values of optimal and sub-optimal arms. Unlike regret bounds of $O(\log T) = O\left(\frac{\log T}{\Delta_i}\right)$, which grow rapidly as Δ_i approaches zero, mean-gap independent regret bounds remain stable even for small Δ_i, making them universally applicable across problem settings. Building on this, (21) analyzed decentralized multi-agent MAB with random graphs, establishing regret upper bounds of $O(\sqrt{T}\log T)$, aligning with (9) up to a logarithmic factor. Despite these advancements, corresponding regret lower bounds in the mean-gap independent sense remain unexplored, and addressing this research gap is one of the objectives of this paper.

3 Problem Formulation

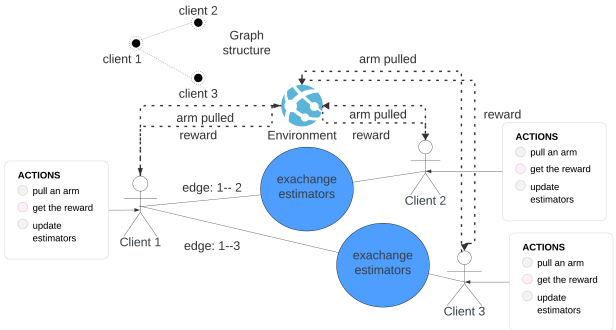
Throughout the paper, we study a decentralized system with $M \geq 3$ clients, and T represents the time horizon. More specifically, the clients are labeled as nodes $1, 2, \ldots, M$ on a network, where the underlying graph at each time step $1 \leq t \leq T$ is represented by an undirected graph G_t. 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 $(X^t_{i,j})_{1\le i,j\le M}$, 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 $1 \le i \le M$. 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 $1 \le m \le M$, there are $K \ge 2$ arms to be pulled. At each time step t, the reward of arm $1 \le i \le K$ 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. Below, we present a visualization of the proposed problem framework for illustrative purposes.

Following (24; 25), 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 $\max_i T\cdot\mu_i = \max_i E[\sum_{t=1}^T r_i(t)] \le E[\max_i \sum_{t=1}^T r_i(t)]$, 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.

The above notations allow us to precisely quantify the regret associated with the action sequence (policy) $\pi = \{a^t_m\}^{1\le t\le T}_{1\le m\le M}$. 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), a client incurs certain regret using a policy π. The regret of a policy π in the bandit setting is defined as $R^\pi_T = T\mu_{i^*} - \frac{1}{M}\sum_{t=1}^T\sum_{m=1}^M \mu_{a^t_m} = \sum_{i=1}^K\sum_{m=1}^M n_{m,i}(T)\Delta_i$ which compares the cumulative expected reward of the actual pulls of arms using π with the cumulative expected reward of the ideal case where we always pull the globally-optimal arm. Let us denote the σ-algebra induced by full information I^s_j of client j at time step s as $\sigma^{t,m}_F = \sigma(\{\{I^s_j\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$ where I^s_j represents the information of all arms contained at client j at time step s and, denote the σ-algebra induced by information $I^s_j(a^j_s)$ that limited to the pulled arm as $\sigma^{t,m}_B = \sigma(\{\{I^s_j(a^j_s)\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$ where $I^s_j(a^j_s)$ represents the information of arm a^j_s contained at client j at time step s. In other words, σ^t_F captures the history of all arms up to time t, whereas $\sigma^{t,m}_B$ only contains the information of client m's time dependent actions up to time t. Henceforth, we have $\sigma^{t,m}_B \subset \sigma^{t,m}_F$. With these notations at hand, we consider two types of policies, by further defining policy sets Π_F and Π_B. For the former, $\Pi_F = \{f_t\}$, where the domain of f_t is on $\sigma^t_F = \{\sigma^{t,m}_F\}_m$, while $\Pi_B = \{g_t\}$ with the domain of g_t being $\sigma^t_B = \{\sigma^{t,m}_B\}_m$ where the domain means the input information a policy can take. The two types of policies necessitates two definitions of regret lower bounds that consider the minimal regret of all the policies, reading as $R^B_T = \min_{\pi\in\Pi_B} R^\pi_T$ and $R^F_T = \min_{\pi\in\Pi_F} R^\pi_T$. Here R^B_T refers to the bandit setting where only partial information $\sigma^{t,m}_B$ is observable, while R^F_T assumes the observations of all arms are visible to the clients (σ^t_F), which is referred to as the full-information setting.

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.

4 Stochastic Settings

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 E; 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 (7) and (19), respectively. □

4.1 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 suboptimality 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. *Let the reward distributions belong to any distribution with finite moments and univariate density. 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 E; 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 (22) 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 (22). 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 (16) 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. *Let the reward distributions belong to any distribution with a finite $1+\epsilon$ moment. 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 E; 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. A further discussion on $\Omega(T)$ is in Appendix D. □

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. *Let the reward distributions of the constructed instances follow Bernoulli distributions. 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 E; 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. □

4.2 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 and the proof is in Appendix E.

Theorem 6. *Let the reward distributions belong to distributions with support on $[0, 1]$. 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 $\Omega(\sqrt{T})$.*

Remark. *First, this regret lower bound also applies to reward distributions with a finite $1 + \epsilon$ moment. Meanwhile, this lower bound of order $\sqrt{T}$ corresponds to the mean-gap upper bounds presented in (21) and (9) for multi-agent and single-agent MAB problems, respectively. This consistency further shows the tightness of the lower bound we have derived.*

5 Adversarial Settings

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 (24).

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. It is worth noting that a single client with arm feedback distributed on graphs is considered in (4), while we consider multiple clients. It also closes the gap between the existing lower bound $\sqrt{T}$ and upper bound $T^{\frac{2}{3}}$ as in (24). 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 $\Omega(T^{\frac{2}{3}})$.*

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 (11)). 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. The proof is in Appendix E.

Theorem 9. *For decentralized multi-agent MAB problems, if the number of clients $M = T^{\frac{2}{15}}$ and the graph*

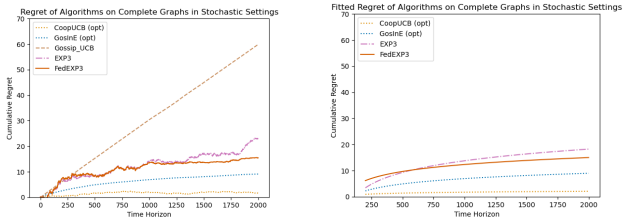
G_t is a connected graph with two expanders of size $\frac{M}{4}$ having distance $d \geq \frac{\eta M}{8}$ given constant $4 > \eta > 8 \cdot T^{-\frac{2}{15}}$, then there exists a problem instance such that the regret of any online distributed learning algorithms is at least $\Omega(T^{\frac{2}{3}})$.

6 Numerical Experiments

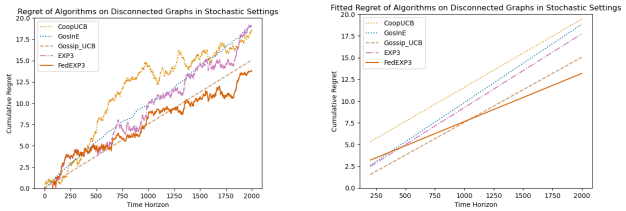
We have demonstrated regret lower bounds that apply to all algorithms in different settings by constructing novel problem instances. In this section, we conduct a comprehensive numerical study to understand how the newly constructed challenging instances affect the performance of existing algorithms. 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. The experiment details are in Appendix C. The code is available at this link. We provide the error bars of results presented in the plots in Appendix.

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 (16), Gossip_UCB in (28), and GosInE in (6), 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 (24) results in an upper bound of order $T^{\frac{2}{3}}$.

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 Appendix C.



(a) Figure 1 (b)

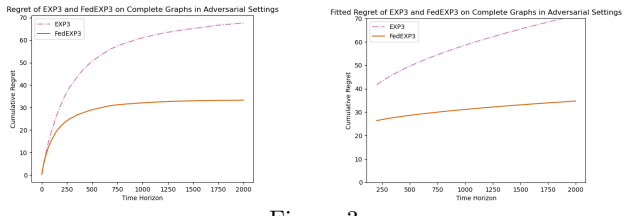


(a) Figure 2 (b)

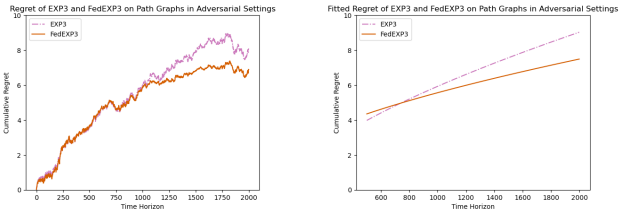
6.1 Comparison Results

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.

Fig. 1 (a) 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 $\log t$,



(a) Figure 3 (b)



(a) Figure 4 (b)

for CoopUCB, GosInE, EXP3, and FedEXP3 are 0.51, 2.94, 6.46, and 3.84, respectively. The fitted regret curves are presented in Fig. 1 (b). 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 Fig. 2 (a). 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, 0.009, and 0.006, respectively, which aligns with the above comparisons among the algorithms. The fitted regret curves are presented in Fig. 2 (b).

We next focus on Instance 3, where the graph is complete and rewards are adversarial. We show the regret of FedEXP3 and EXP3 in Fig. 3 (a). 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

Table 1: Research Gaps

Settings	Upper	Lower	Opt.
Sto. & Conn.	$\log T$	$\log T$	Opt
Adv. & Comp.	$T^{\frac{1}{2}}$	$T^{\frac{1}{2}}$	Opt
Adv. & Conn.	$T^{\frac{2}{3}}$	$T^{\frac{2}{3}}$ ~~$\sqrt{T}$~~	Opt ~~Non-Opt~~
Disconnected	T	T	Opt

FedEXP3 are 0.967 and 0.274, respectively. The fitted regret curves are presented in Fig. 3 (b). 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 Fig. 4 (a). 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 Fig. 4 (b). 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.

7 Conclusion

The detailed conclusion including a comprehensive summary table of the results (both existing work and ours) is in Appendix. A concise comparison between the existing work and our contributions is presented in Table 1. The highlighted parts represent the gaps we address in this paper, while the strikethrough indicates results from existing work that are not tight or optimal.

Acknowledgments

We sincerely appreciate Professor Barry L. Nelson's valuable comments and advice, which have greatly helped improve the paper–the final version of the second-to-last chapter of Xu's dissertation.

References

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

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

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

[4] N. Cesa-Bianchi, T. R. Cesari, and R. Della Vecchia. Cooperative online learning with feedback graphs. *arXiv preprint arXiv:2106.04982*, 2021.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] M. Xu and D. Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pages 38499–38517. PMLR, 2023.

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

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

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

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

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

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Everything except code is provided]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable; a single laptop can easily run the experiment]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Table 2: Summary of Results

	Upper (existing)	Lower (new)	Graphs	Rewards	Algo.
Instance	$\log T$	$\log T$	Connected; time-invariant or time-varying	Stochastic with finite $1+\epsilon$ moment	DDUCB, FedDr-UCB, GosInE
Mean-gap	$\sqrt{T}$	$\sqrt{T}$	Connected; time-invariant or time-varying	Stochastic with finite $1+\epsilon$ moment	LCC-UCB, FedDr-UCB UCB-warm
Adversarial	$T^{\frac{1}{2}}$	$T^{\frac{1}{2},*}$	Complete graphs; time-invariant or time-varying	Adversarial (Oblivious)	EXP3
Adversarial	$T^{\frac{2}{3}}$	$T^{\frac{2}{3}}$	Expander graph family; time-invariant or time-varying	Adversarial (Oblivious)	FEDEXP3
General	T	T	Disconnected; time-invariant or time-varying	Stochastic or Adversarial	Any

A 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 (16), FedDr-UCB in (21), GosInE in (6), LCC-UCB in (1), UCB-warm in (9), and FEDEXP3 in (24). 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.

B 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.

C Additional Experiment Details

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.

C.1 Statistical Significance

We provide error bars of the regret in the following tables.

The following table, Table 3 demonstrates the error bar corresponding to $T^{\frac{1}{2}}$.

The following table, Table 4, demonstrates the error bar corresponding to $T^{\frac{2}{3}}$.

The following table, Table 5, demonstrates the error bar corresponding to $\log T$.

The following table, Table 6, demonstrates the error bar corresponding to T.

C.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 (21). 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.

D Discussions on the linear regret lower bound

As part of the contributions of this paper, it is shown that the regret lower bound is of order $\Omega(T)$ when the graph is disconnected. However, we would like to include that it is possible to get sublinear regret when the

Table 3: Numerical Experimental Results

Algorithms	Regret Order	Coefficient	Confidence Interval
EXP3	$\log T$	[0.967]	[0.967,0.967]
FedEXP3	$\log T$	[0.274]	[0.274, 0.274]

Table 4: Numerical Experimental Results

Algorithms	Regret Order	Coefficient	Confidence Interval
EXP3	$\log T$	[0.053]	[0.053,0.053]
FedEXP3	$\log T$	[0.033]	[0.033, 0.033]

Table 5: Numerical Experimental Results

Algorithms	Regret Order	Coefficient	Confidence Interval
EXP3	$\log T$	[6.46]	[6.46, 6.46]
FedEXP3	$\log T$	[3.84]	[3.84, 3.84]
CoopUCB	$\log T$	0.51	[0, 2.75]
GosInE	$\log T$	2.94	[1.54, 12.17]

Table 6: Numerical Experimental Results

Algorithms	Regret Order	Coefficient	Confidence Interval
EXP3	$\log T$	[0.009]	[0.009, 0.009]
FedEXP3	$\log T$	[0.006]	[0.006, 0.006]
CoopUCB	$\log T$	0.008	[0.001, 0.018]
GosInE	$\log T$	0.009	[0.009, 0.009]

graph is disconnected by adding some assumptions on the problem settings as follows. The assumptions could be regarding both graphs and rewards which determine the problem complexity. Uniformly strongly connected graphs, give $O(\log T)$ (25). Also, with random graphs, e.g. the E-R model where each graph observation can be disconnected, the regret is $O(\log T)$ and $O(\sqrt{T})$ (21). For other disconnected graphs, sublinear regret is ensured if each client shares the same mean reward values (homogeneous) and plays with their own MAB optimally. More broadly, having heterogeneity within each connected component while ensuring homogeneity across these components is adequate. This paper shows that linear regret results from the difference in optimal arms. If monotonicity of rewards over arms or the choice of the optimal arm is the same across the connected components, then designing optimal methods within the connected components ensures sublinear regret.

E Proof of Results in Section 4

E.1 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 (19) 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 (7). Then based on the result established by specifying a contextual linear bandit with $\alpha = 1$ as in (7), 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.

□

E.2 Proof of Theorem 3

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 $E[R_T^B] \geq \Omega(\log T)$, 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 $|[t_d, T]| \geq 2^{\frac{1}{2}\log T}$ for all large enough T.

Meanwhile, we observe that for T_a, it is equivalent to a single-agent multi-objective bandit problem (22) 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 (22), and satisfies that $R_{T_a^d,M} \geq \Omega(\log T_a^d)$ for any policy $\{a_t\}$ based on Theorem 6 in (22).

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 (22), 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

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

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.

□

E.3 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 (15)) 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. □

E.4 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 (2). 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. □

E.5 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 (19) 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.

□

E.6 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

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

We star with a full-information setting. Following a similar argument and constructing the same instance as in Lemma A.4 in (24), 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 (3). Equivalently, we write it as

$$R_T^F \geq \Omega(T^{\frac{2}{3}}). \tag{4}$$

Meanwhile, by Theorem 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 (4) we obtain

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

This completes the proof of Theorem 8.

□

E.7 Proof of Theorem 9

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

$$\begin{aligned} D_{KL}(Q_{j,t}^{-1}, Q_{j,t}^i) &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2} - \epsilon} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2} + \epsilon} \\ &= \frac{1}{2}\log(1 + \frac{4\epsilon^2}{1 - 4\epsilon^2}) \leq \frac{1}{2} \cdot \frac{4\epsilon^2}{1 - 4\epsilon^2} \leq 4\epsilon^2, \end{aligned}$$

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 $D_{TV}(Q_j^{-1}, Q_j^i) \leq \sqrt{\frac{D_{KL}(Q_j^{-1}, Q_j^i)}{2}} \leq \epsilon\sqrt{2d}$. (5)

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

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

and of arm 2 is $\frac{1}{8}$ from

$$\begin{aligned} \mu_2 &= \frac{1}{M}\sum_{m=1}^{M}\mu_2^m \\ &= \frac{1}{M}\sum_{m\in I_0}\mu_2^m + \frac{1}{M}\sum_{m\notin I_0}\mu_2^m \\ &= \frac{1}{M}\sum_{m\in I_0}\frac{1}{2} + \frac{1}{M}\sum_{m\notin I_0} 0 = \frac{1}{8}. \end{aligned}$$

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

$$
\begin{aligned}
&E[R_T^B] \\
&= E[E[R_T^B|X_1,\ldots,X_D]] \\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(T - n_{m,1}(T)))|X_1,\ldots,X_D]] \\
&= E[E[\frac{1}{M}\sum_{m=1}^{M}(\frac{\epsilon}{2M}(\sum_{j=1}^{D} d - \sum_{j=1}^{D} n_{m,1}(T,j)))|X_1,\ldots,X_D]] \\
&= E[\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D} E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_1,\ldots,X_D]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D} E[E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j]] \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}} \frac{E[(\frac{\epsilon}{2M}(d - n_{m,1}(T,j)))|X_j = i]]}{|I_0|+1} \\
&\geq \frac{1}{2M^2}(\frac{1}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}} E[\epsilon\cdot(d - n_{1,1}(T,j))|X_j = i]) \\
&= \frac{1}{2M^2}(\epsilon\cdot T - \frac{\epsilon}{|I_0|+1}\sum_{j=1}^{D}\sum_{i\in I_0\cup\{0\}} E_{Q_j^i}[(n_{1,1}(T,j))]) \qquad (6)
\end{aligned}
$$

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 (5), 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

$$\begin{aligned} E[R_T^B] &\geq \frac{1}{2M^2}(\epsilon \cdot T - \frac{\epsilon}{|I_0|+1}(T + T \cdot \frac{|I_0|+1}{4})) \\ &\geq \frac{1}{2M^2}\frac{1}{4}\epsilon \cdot T = \Omega(T^{\frac{2}{3}}) \end{aligned} \tag{7}$$

where the last inequality uses $|I_0| = \frac{M}{4} \geq 2$ and the equality holds by the choice of ϵ and M. □