Near-Optimal Regret Bounds for Federated Multi-armed Bandits with Fully Distributed Communication

Haoran Zhang[1,4] **Xuchuang Wang**[2] **Haoxu Chen**[1,4] **Hao Qiu**[3] **Lin Yang***[1,4] **Yang Gao**[1,4]

[1]School of Intelligent Science and Technology, Nanjing University, Suzhou, China
[2]College of Information & Computer Science, University of Massachusetts Amherst, Massachusetts, USA
[3]Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[4]National Key Laboratory for Novel Software Technology, China

Abstract

In this paper, we focus on the research of federated multi-armed bandit (FMAB) problems where agents can only communicate with their neighbors. All agents aim to solve a common multi-armed bandit (MAB) problem to minimize individual regrets, while group regret can also be minimized. In a federated bandit problem, an agent fails to estimate the global reward means of arms by only using local observations, and hence, the bandit learning algorithm usually adopts a consensus estimation strategy to address the heterogeneity. However, up to now, the existing algorithms with fully distributed communication graphs only achieved a suboptimal result for the problem. To address that, a fully distributed online consensus estimation algorithm (`CES`) is proposed to estimate the global mean without bias. Integrating this consensus estimator into a distributed successive elimination bandit algorithm framework yields our federated bandit algorithm. Our algorithm significantly improves both individual and group regrets over previous approaches, and we provide an in-depth analysis of the lower bound for this problem.

1 INTRODUCTION

Online learning problems in federated settings, where a set of agents complete a common learning task via performing individual learning algorithms and keeping data locally used, are broadly researched due to plenty of motivating applications in the real world. For example, in the fields of finance, medicine and data processing, federated learning is a potential method for solving local training and individual privacy problems [Yang et al., 2019, Li et al., 2020, Liu et al., 2022]. In this paper, we study the FMAB problem, where multiple instances of the MAB problem are implemented on a set of agents communicating with each other. Recently, efforts have been invested in designing distributed bandit algorithms for federated learning problems [Féraud et al., 2019, Shi and Shen, 2021, Agarwal et al., 2022], where agents can only communicate with neighbors without a suitable end-to-end communication protocol due to the limitations in practical systems. FMAB with consensus communication has many real-world applications. For instance, it is common for multiple agents to collaborate on large-scale tasks in broadcasting sensor networks, which consist of several wireless sensors that communicate only with their neighbors [Li et al., 2019, Kolla et al., 2018]. For example, selecting an appropriate time to conduct an outdoor experiment requires consideration of various environmental factors such as humidity, temperature, wind speed, and others. To capture this information, a variety of sensors are deployed, viewed as agents in this context. At different time steps, these agents provide feedback based on their local observations, which serve as local samples. The ultimate objective is to integrate these local samples to identify the optimal time for the outdoor experiment. In other scenarios, data heterogeneity may arise from privacy protection policies, which require training data to be processed locally.

The major obstacles that prevent FMAB from achieving optimal learning performance are heterogeneous feedback among agents and fully distributed communication. Together, these factors make it difficult for agents to accurately track the global mean. To effectively learn the global mean, the implemented learning algorithm needs to collect the estimates or observations in each agent, as well as their number of samples. One can imagine that simply merging the global estimate of each agent without knowing the number of samples will result in unexpected bias. Furthermore, fully distributed communication indicates that each agent has a different ability to acquire information, leading to errors in tracking sample counts. These errors accumulate over time, ultimately degrading the performance of FMAB algorithms. As a result, all previous works have failed to

*Corresponding author: linyang@nju.edu.cn

Table 1: A comparison summary of prior literature and this work.

Algorithm	Individual Regret	Group Regret
`Gossip_UCB` [Zhu et al., 2021]	$O(\sum_{i:\Delta_i>0} N\Delta_i^{-1}\log T)$	$O(\sum_{i:\Delta_i>0} N^2\Delta_i^{-1}\log T)$
`Dis_UCB` [Zhu and Liu, 2023]	$O(\sum_{i:\Delta_i>0} N_{\min}^{-1}\Delta_i^{-1}\log T)$	$O(\sum_{i:\Delta_i>0} NN_{\min}^{-1}\Delta_i^{-1}\log T)$
`DRRB-bandit` (our work)	$O(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$	$O(\sum_{i:\Delta_i>0} \Delta_i^{-1}\log T)$
General regret lower bound	$\Omega(\sum_{i:\Delta_i>0} N^{-2}\Delta_i^{-1}\log T)$	$\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$
Regret lower bound for special algorithms	$\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$	$\Omega(\sum_{i:\Delta_i>0} \Delta_i^{-1}\log T)$

achieve optimal regrets.

In the presence of heterogeneous feedback, a distributed estimation approach has been proposed. This approach primarily collects information from neighboring agents and estimates the global mean of each arm to determine whether the selection is optimal. However, because of the absence of a central server in a fully distributed communication setup, real-time information remains inaccessible to individual agents. According to the above explanation, two obstacles are coupled, i.e., fully distributed communication makes heterogeneous feedback more difficult to deal with. To address the FMAB problem with only hop-by-hop communication, a kind of gossip-based communication was proposed [Kempe et al., 2003, Boyd et al., 2006], which does not need a central server or a fully connected communication graph. These methods eliminate the need for a central server or a fully connected communication graph, allowing agents to exchange information efficiently. Building on foundational research in communication methods, several scholars have developed algorithms to tackle the FMAB problem [Zhu et al., 2021, Zhu and Liu, 2023, Xu and Klabjan, 2024]. These studies employed gossip-based communication and refined the selection strategy using the Upper Confidence Bound (UCB) algorithm [Auer and Ortner, 2010]. Additionally, they introduced mechanisms to regulate agent behavior, ensuring consensus on sampling frequency across the network.

However, due to inherent limitations in the framework of UCB-based algorithms, certain challenges remain in the proposed approach, potentially leading to suboptimal results. Among those, the core challenge is to obtain unbiased global estimates from biased local observations and limited information from neighbors. Specifically, traditional UCB-based algorithms tend to select the arm with the maximum upper confidence bound, without accounting for the heterogeneity of agents. This can result in biased estimates of reward means, thereby leading to suboptimal regret performance. While previous works [Zhu et al., 2021, Zhu and Liu, 2023, Xu and Klabjan, 2024] propose estimation mechanisms to address this issue, the weight assigned to each reward in the global estimate varies. This inconsistency creates an unfair mechanism, leading to suboptimal convergence of the estimates.

Related works. The federated bandit problem can be divided into two categories from the perspective of reward, called homogeneous reward and heterogeneous reward settings. In the homogeneous reward setting [Hillel et al., 2013, Wang et al., 2019, 2020], agents pull the same arm and achieve rewards from the same distribution, which implies that their sampling directly helps them estimate the global means of the arms. In the heterogeneous reward setting [Shi and Shen, 2021, Zhu et al., 2021, Zhu and Liu, 2023, Xu and Klabjan, 2024], agents have their local reward distributions, which means that agents obtain different rewards even if they pull the same arm. The main challenge is to obtain an unbiased global estimate because local sampling is useless to learn the global mean.

From the classification of the communication network, the network can be divided into the fully distributed graph and the fully connected graph [Shamma, 2008]. A fully connected graph means that any two agents are directly connected, while a fully distributed graph, means that there is a path between two agents that may be connected through other agents. For a fully connected graph (end-to-end communication), the time delay is the 1 time slot, which can be seen as a central server because all agents have access to all agents' information. For a fully distributed graph (hop-by-hop communication), the time delay is at most D time slots, which is the diameter of the communication graph.

From the perspective of the sampling method, it can be divided into synchronous setting [Wang et al., 2019, Dubey and Pentland, 2020, Huang et al., 2021] and asynchronous setting [He et al., 2022, Wang et al., 2023b]. In the synchronous setting, each agent samples and communicates at the same frequency, which makes concentration easier than with asynchronous methods. In an asynchronous setting, agents can not implicitly coordinate their actions through time, making it difficult to cooperate. Agents pull arms and exchange information without a common rule, while each agent could act in their own interests.

Contributions. In the article, we investigate the above-mentioned federated bandit learning problem and make the following contributions.

In Section 3, we propose two algorithms called `CES` and `DRRB-bandit`. (a) `DRRB-bandit` leverages the strategy of round-robin sampling to ensure the agents' sam-

ples in a synchronized manner [Wang et al., 2023a, Perchet and Rigollet, 2013]. Specifically, the agents communicate with one another to maintain a consistent candidate set and explore the arms within that set in a round-robin manner. (b) `CES` uses a novel estimation mechanism, which is first presented in the literature, to estimate the global mean of each arm. `CES` combines the global estimates from other agents with its own latest samples in a dynamic proportion, even when agents are not directly connected. This mechanism fairly allocates the weight of each sample reward in the global estimates, effectively mitigating the effects of heterogeneity and producing a more accurate global estimate.

In Section 4, through theoretical analysis, `DRRB-bandit` is proved to achieve a near-optimal individual regret $O(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ for each agent, where N is the number of agents, Δ_i is the gap between the optimal global mean and the global mean of arm i, and T is the time horizon. As a straightforward result, the group regret is $O(\sum_{i:\Delta_i>0}\Delta_i^{-1}\log T)$. We also provide two kinds of lower individual regret bounds: the first one, which is $\Omega(\sum_{i:\Delta_i>0} N^{-2}\Delta_i^{-1}\log T)$, is general and holds for all algorithms; the second one, which is $\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$, holds for all algorithms with round-robin sampling, implying that `DRRB-bandit` is near-optimal among all round-robin-based algorithms. Additionally, the total communication cost is bounded by $O(K\Delta_{\min}^{-1}\log T)$, where $\Delta_{\min}$ is the minimal non-zero gap. The above results dramatically outperform existing results in previous works, among which the best one is $O(\sum_{i:\Delta_i>0} N_{\min}^{-1}\Delta_i^{-1}\log T)$ for the individual regret ($O(\sum_{i:\Delta_i>0} NN_{\min}^{-1}\Delta_i^{-1}\log T)$ for group regret), where $N_{\min}$ denotes the smallest number of neighbors for any agent, including the agent itself. We provide a simple account of the results in Table 1.

The improvement in this work is practically significant. Considering a practical case where $N_{\min} \ll N$, the individual and group regrets in previous works can be as large as $O(\sum_{i:\Delta_i>0}\Delta_i^{-1}\log T)$ and $O(\sum_{i:\Delta_i>0} N\Delta_i^{-1}\log T)$, respectively. Clearly, these results lead to linear regret with respect to the number of agents (or system size), whereas our approach eliminates the dependence on the number of agents, making it more practically significant for real-world applications of cooperative learning.

Finally, we will introduce the organization of the paper below. Firstly, we introduce the necessary notations and the problem formulation in Section 2. Secondly, we describe the framework of both `DRRB-bandit` and `CES` in Section 3. In Section 4, we provide theoretical results on the regret for `DRRB-bandit`, with missed details deferred to the appendix. In Section 5, we provide experimental results with varying settings.

2 PROBLEM DESCRIPTION

In the MAB problem, a player repeatedly selects an arm from a given set $\mathcal{K} = \{1, \ldots, K\}$ over time. At each time slot $t \in \{1, \ldots, T\}$, the player chooses an arm to pull and obtain a reward associated with the selected arm. The rewards for each arm are drawn from an independent and identically distributed (i.i.d.) process, with values in the interval $[0, 1]$ [1]. This reward serves as real-time feedback to the player regarding the chosen arm.

In this article, we focus on federated bandit problems. Different from general MAB problems, this setting introduces two additional elements: multiple players and heterogeneous feedback. Specifically, we consider a stochastic bandit setting containing N agents, represented by the agent set $\mathcal{N}$. At each time slot t, agent j selects an arm $A_j(t) \in \mathcal{K}$ to pull and receives a random reward $X_{A_j(t),j}(t)$. The decision-making strategy primarily depends on the agents' past actions and observed rewards.

In this scenario, agent j could only observe a random reward $X_{i,j}(t)$, which consists of both the mean and noise components. The reward $X_{i,j}(t)$ follows an independent and identically distributed (i.i.d.) process with a reward mean $\mu_{i,j}$, bounded within $[0, 1]$. If agent j selects arm i at time step t, i.e., $A_j(t) = i$, the global reward at time t is defined as $X_{A_j(t)}(t) = X_i(t) := \frac{1}{N}\sum_{j=1}^{N} X_{i,j}(t)$. Similarly, the global mean of $X_i(t)$ is given by $\mu_i := \frac{1}{N}\sum_{j=1}^{N}\mu_{i,j}$. Without loss of generality, denote $i^\star$ by the unique optimal arm with the highest global mean among all arms in the set $\mathcal{K}$, i.e., $i^\star = \arg\max_i \mu_i$. The reward gap between any arm i and the optimal arm $i^\star$ is then defined as $\Delta_i = \mu_{i^\star} - \mu_i$.

After sampling, agents exchange information with their neighbors. The neighborhood of agent j is defined as $\mathcal{N}_j$, which consists of all agents connected to agent j, excluding j itself. To represent the relationships among all agents, we use a communication matrix $\boldsymbol{W} = [\omega_{a,b}]_{N\times N}$ to describe the connectivity structure of the multi-agent system. Further details about this matrix are provided in Appendix B. We assume that there are no collisions; that is, when multiple agents pull the same arm, each agent independently receives a reward sample drawn from the arm's reward distribution. It is important to note that the problem is set in a heterogeneous environment, meaning that the expected reward of arm i varies across different agents. Specifically, $\mu_{i,j_1} \neq \mu_{i,j_2}$ for $j_1 \neq j_2$.

Group regret: In this paper, group regret is defined as the cumulative loss of reward incurred by selecting a suboptimal arm instead of the optimal arm. This metric serves as the primary measure for evaluating federated bandit algorithms. The optimal strategy for all agents is to consistently pull the optimal arm throughout the entire time horizon T. Therefore,

[1] Via Lemma 5, the results in this work also hold for other distributions, such as sub-Gaussian distributions, etc.

for a distributed algorithm $\mathcal{A}$, the expected group regret of the system is defined as follows:

$$\mathbb{E}[R^T(\mathcal{A})] \coloneqq NT\mu_{i^\star} - \sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{E}[X_{A_j(t),j}(t)]. \quad (1)$$

Individual regret: While group regret is a key performance metric for distributed algorithms, minimizing individual regret is also a crucial challenge in the federated bandit problem. In a heterogeneous setting, agents may pull the same arm but receive different local rewards, leading to variations in their global estimates. Moreover, an agent's ability to access information depends on the structure of its neighborhood, resulting in disparities in decision-making. Therefore, considering individual regret is essential to prevent overly aggressive behavior from any single agent. In practical applications, optimizing individual regret becomes even more critical. For instance, in cooperative systems like drone swarms, the failure of a single agent can significantly impact overall performance—a phenomenon known as the "cask effect." To quantify the impact of individual decision-making, individual regret is defined as follows:

$$\mathbb{E}[R_j^T(\mathcal{A})] \coloneqq T\mu_{i^\star} - \sum_{t=1}^{T}\mathbb{E}[X_{A_j(t)}(t)]. \quad (2)$$

Communication cost: In our setting, we do not impose any restrictions on the type or size of messages exchanged during each communication round. When one agent sends one message, the communication round incurs a unit cost. For algorithm $\mathcal{A}$, the communication cost of the global systems is defined as

$$\mathbb{E}[C^T(\mathcal{A})] = \sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{I}\{\mathcal{I}_j(t)\}, \quad (3)$$

where $\mathbb{I}\{\cdot\}$ is an indicator function and $\mathcal{I}_j(t)$ represents the event that agent j send messages to its neighbors at time slot t.

3 ALGORITHM

The first core challenge in the federated bandit problem is estimating the global mean based on biased local observations. During the game, each agent maintains its own observations or estimates of the local arms, which deviate from the global mean due to heterogeneous feedback. Consequently, to learn the global reward mean, agents have to aggregate the estimates or observations of all agents, and each agent is responsible for sampling the arms of themselves in the heterogeneous setting. During the above procedure, insufficient sampling by any agent will result in an imprecise estimate of the global mean. Upon revisiting previous works [Zhu et al., 2021, Zhu and Liu, 2023, Xu and Klabjan, 2024], we identify a key limitation that leads to suboptimal results: the UCB-based algorithm framework typically favors selecting the arm with the largest upper confidence bound, which leads to biased global estimates. Although these algorithms reduce some bias in decision-making at each round, they still produce biased decisions overall, which leads to uneven learning. As a result, they must rely on the worst-case scenario to compute concentration errors, which limits their effectiveness.

To optimize algorithms for federated bandit problems, it is crucial to fully leverage the sample information from each agent. In a homogeneous bandit setting [Hillel et al., 2013, Shahrampour et al., 2017, Zhu and Liu, 2021], the samples from one agent can directly benefit the learning process of other agents through communication, as all agents share the same learning objectives. However, in heterogeneous bandit problems, additional challenges arise in algorithm design. Specifically, in a heterogeneous setting, simply aggregating information from other agents does not ensure that it benefits an agent, as the reward distributions or environments may differ across agents. To achieve optimal performance, agents need to accurately track both the observations/estimates and the sample counts associated with the observations from all other agents. Given that the setting is fully distributed, with each agent only able to communicate with its neighbors, it becomes challenging for agents to learn the global mean. Therefore, addressing the heterogeneity in federated bandit problems is the second key challenge explored in this paper.

To address the first challenge, we adopt a round-robin-based algorithm framework, where each agent uniformly explores its local arms at each round. We apply this framework to federated bandit problems and introduce the Distributed Round-Robin-Based Bandit Algorithm (`DRRB-bandit`) in Section 3.1. In the algorithm framework, each agent can maintain a dynamic candidate arm set and sample arms in the set equally until one arm is judged as suboptimal and eliminated from the set. The agents can receive real-time implicit information, i.e., the concrete sample counts of other agents, equal to the sample counts themselves. All agents share the confidence interval for the same arm because all agents uniformly explore these arms. Hence, the worst case can be avoided and the algorithm obtains a near-optimal result.

For the second challenge, one intuition is to design a suitable online estimation algorithm based on the quality of networks. We provide an estimation policy called consensus estimation subroutine (`CES`) in Section 3.2. In `CES`, each agent combines other agents' global estimates and its latest sampling in a dynamic proportion. These global estimates contain information about other unconnected agents. Hence, the policy can counter the information congestion caused by the incomplete communication graph. Over a few rounds, the latest global estimate can gradually get rid of the effects of heterogeneity.

Based on the two ideas, the exploration efficiency will increase by N times, as each single agent can fully utilize the exploration of all agents and the influence of the heterogeneous feedback could be reduced.

3.1 DISTRIBUTED ROUND-ROBIN-BASED BANDIT ALGORITHM (DRRB-BANDIT)

We present a federated bandit learning algorithm called DRRB-bandit, which employs round-robin sampling as the underlying arm-pulling policy. A key idea behind DRRB-bandit is that agents uniformly sample arms to track the global mean of each arm. Using DRRB-bandit, agents select arms through round-robin sampling and eliminate suboptimal arms by comparing the upper confidence bounds of each suboptimal arm with the lower confidence bound of the optimal arm. By incorporating time labels on the suboptimal arms, the algorithm ensures that all agents avoid asynchronous elimination, which is typically caused by time delays in a fully distributed communication graph.

To ensure synchronous sampling, the algorithm maintains a candidate arm set, containing arms to be explored in a round-robin manner. The candidate arm set is initialized as the arm set $\mathcal{K}$. As the sample count increases, the algorithm gradually identifies suboptimal arms and removes them from the candidate arm set until only one remains. When an agent identifies a suboptimal arm, it will notify its neighbors of this information. To ensure all agents eliminate a suboptimal arm synchronously, a time label will be transmitted along with this arm. The time label indicates the time slot at which all agents have received the suboptimal arm, accounting for the time delay caused by fully distributed communication. The design of the time label ensures that all agents update the candidate arm set simultaneously. The pseudocode of DRRB-bandit is summarized in Algorithm 1.

Round-Robin Policy for Exploration. In successive elimination algorithms, each agent pulls arms from the arm set using round-robin sampling. Each agent j maintains a dynamic candidate arm set $\mathcal{S}_j(t)$, which initially includes all arms, i.e., $\mathcal{S}_j(0) = \mathcal{K}$. Over time, this set updates based on the agent's exploration and the information it shares with neighbors. The bandit algorithm operates on the candidate arm set, with agents using round-robin sampling to learn the local reward distribution and estimate the global mean for each arm. By learning the global means, agents can identify suboptimal arms. In addition, each agent shares information about suboptimal arms with its neighbors. Based on the identified suboptimal arms, agents update their candidate arm set accordingly. Arms identified as suboptimal are eliminated from the candidate arm set at a predetermined time, as specified in Lines 14 and 15 of Algorithm 1. The policy for arm elimination will be further explained below.

Arm Elimination. To manage the candidate arm set based on information from other agents, we introduce an *elimination arm set*, denoted as $\mathcal{B}_j$, which stores the indices of arms identified as suboptimal and slated for elimination. At the beginning of each round, the algorithm selects the arm with the highest global estimate, $\tilde{\mu}_{i^{\max},j}$, as the benchmark. Then it compares the global estimates of the arms in the candidate set to this maximum value. If one arm's global estimate is lower than the benchmark by a threshold related to the radius of its confidence interval, that arm is considered suboptimal and is added to the elimination arm set $\mathcal{B}_j$ (Line 8, Algorithm 1).

In a fully distributed communication graph, each agent has distinct capabilities for collecting and processing information. To manage the updates of the candidate sets across all agents, a time label t_i is assigned to each suboptimal arm i. This label incorporates both the communication delay inherent in the distributed system and the time at which the arm is identified as suboptimal. Using the predetermined time label t_i, agents can synchronize the elimination of the suboptimal arm, ensuring that all agents remove it from their candidate sets at the same time.

At each time slot t, agent j samples all arms in the candidate set $\mathcal{S}_j(t)$. Let $\tau_{i,j}(t)$ denote the number of samples of arm i by agent j up to time slot t. Since all agents update their candidate set synchronously, the number of observations of all agents on arm i is equal. Thus, the total sample count for arm i is $\tau_i(t) = N\tau_{i,j}(t)$. Let $\tilde{\mu}_{i,j}(t)$ denote the estimate of the global mean on arm i by the j-th agent (A detailed explanation is given in CES (Algorithm 2)). Based on the global estimate $\tilde{\mu}_{i,j}$ and the sample count $\tau_{i,j}$, we can construct a confidence interval for the global reward mean μ_i, which typically follows the Hoeffding's inequality [Hoeffding, 1994]. Define $U_{i,j}(t,\delta)$ as the radius of the confidence interval for the rewarding process with $\tau_{i,j}(t)$ samples and confidence level $1-2\delta$, which is written as

$$U_{i,j}(t,\delta) \coloneqq \sqrt{\frac{\log \delta^{-1}}{2N\tau_{i,j}(t)}} + \frac{Q}{(1-\lambda_2)(\tau_{i,j}(t)+1)}, \quad (4)$$

where δ specifies the violation probability that the true mean lies outside the above confidence interval (The details and analysis are introduced in Lemma 1). The global reward mean μ_i is contained within the confidence interval $(\tilde{\mu}_{i,j}(t) - U_{i,j}(t,\delta), \tilde{\mu}_{i,j}(t) + U_{i,j}(t,\delta))$ with at least $1-2\delta$ probability. For simplicity, we use $\mathtt{UCB}_{i,j}$ and $\mathtt{LCB}_{i,j}$ to represent the upper and lower confidence bounds of μ_i, respectively.

For any arm i in the candidate set $\mathcal{S}_j(t)$, it will be considered suboptimal if the global estimate of arm i and $i^{\max}$ satisfies

$$\underbrace{\tilde{\mu}_{i,j}(t) + U_{i,j}(t,\delta)}_{\mathtt{UCB}_{i,j}} \geq \underbrace{\tilde{\mu}_{i^{\max},j}(t) - U_{i,j}(t,\delta)}_{\mathtt{LCB}_{i^{\max},j}}, \quad (5)$$

where $i^{\max}$ is the arm with the maximum global mean estimate among all arms in $\mathcal{S}_j(t)$ and is determined at the

Algorithm 1 Distributed Round-Robin-based Bandit Algorithm (`DRRB-bandit`) (for agent j)

Input: The time horizon T, the diameter D and the arm set $\mathcal{K}$

Initialization: $t = 0$, $\tau_{i,j} = 0$, $U_{i,j} = 1$, $\mathcal{S}_j = \mathcal{K}$, $\mathcal{B}_j = \varnothing$

1: Pull each arm one time and receive a local reward $X_{i,j}$
2: $\tilde{\mu}_{i,j} \leftarrow X_{i,j}$, $\tau_{i,j} \leftarrow \tau_{i,j} + 1$, $t \leftarrow t + K$, $i \in \mathcal{K}$
3: **while** $t \leq T$ **do**
4: $\quad i^{\max} \leftarrow \arg\max_i \{\tilde{\mu}_{i,j} : i \in \mathcal{S}_j\}$
5: $\quad$ **for** $i \in \mathcal{S}_j$ **do**
6: $\quad\quad$ Pull arm i and obtain the reward $X_{i,j}$
7: $\quad\quad$ $t \leftarrow t + 1$, $\tau_{i,j} \leftarrow \tau_{i,j} + 1$
8: $\quad\quad$ **if** $\tilde{\mu}_{i,j} < \tilde{\mu}_{i^{\max},j} - 2U_{i,j}$ **then**
9: $\quad\quad\quad$ $t_i \leftarrow t + |\mathcal{S}_j|D$, $\mathcal{B}_j \leftarrow \mathcal{B}_j \cup \{i\}$
10: $\quad\quad$ **end if**
11: $\quad\quad$ Update $U_{i,j}$ via equation (4)
12: $\quad$ **end for**
13: $\quad$ Operate Subroutine 2 for the latest global estimates
14: $\quad$ **for** each arm i in $\mathcal{B}_j$ whose $t \geq t_i$ **do**
15: $\quad\quad$ **if** $|\mathcal{S}_j| > 1$ **then** $\mathcal{S}_j \leftarrow \mathcal{S}_j \backslash \{i\}$; **else** $\mathcal{S}_j \leftarrow \mathcal{S}_j$
16: $\quad$ **end for**
17: **end while**

beginning of each round (Line 4, Algorithm 1).

In `DRRB-bandit`, once agent j identifies arm i as suboptimal, it will add the arm to the elimination arm set $\mathcal{B}_j$ and broadcast its index to all other agents. For each arm i, a predetermined time label t_i is assigned, which indicates the time at which the arm will be removed. The value of t_i is set according to the following equation:

$$t_i = t + |\mathcal{S}_j(t)|D, \tag{6}$$

where D is the diameter of the communication graph $\mathcal{G}$, and t represents the time slot when the arm is identified as suboptimal. $|\mathcal{S}_j(t)|$ represents the element number in $\mathcal{S}_j(t)$. Given the indexes of suboptimal arms, an elimination arm set $\mathcal{B}_j(t)$ of agent j at time t is constructed, which contains all arms identified as suboptimal, i.e.,

$$\begin{aligned}\mathcal{B}_j(t) = \{i, i \in \mathcal{S}_j : \exists i' \in \mathcal{S}_j \setminus \{i\} \text{ such that} \\ \tilde{\mu}_{i,j}(t) \leq \tilde{\mu}_{i',j}(t) - 2U_{i,j}(t,\delta)\}.\end{aligned} \tag{7}$$

By continuously monitoring the elimination set, the algorithm iteratively updates the candidate set until the optimal arm is identified.

3.2 CONSENSUS ESTIMATION SUBROUTINE (`CES`)

To mitigate the biased estimation arising from the heterogeneous setting, we propose a novel consensus estimation subroutine in the federated bandit setting with fully distributed communication, which can be integrated to `DRRB-bandit` (Introduced in Section 3.1).

Algorithm 2 Consensus Estimation Subroutine (`CES`) (for agent j)

Input: The local reward $X_{i,j}$, the candidate arm set $\mathcal{S}_j$, the function of weight coefficient $\sigma_i(\tau) = \frac{1}{\tau+1}$, the sample count $\tau_{i,j}$ and the weight matrix $W = [\omega_{j,j'}]_{N\times N}$

Output: The latest estimate $\tilde{\mu}_{i,j}$ and elimination arm set $\mathcal{B}_j$

1: Send $\tilde{\mu}_{i,j}$, t_i, $i \in \mathcal{S}_j$ and $\mathcal{B}_j$ to neighbors
2: Receive $\tilde{\mu}_{i,j'}$, t_i and $\mathcal{B}_{j'}$ from neighbors $j' \in \mathcal{N}_j$
3: **for** $i \in \mathcal{S}_j$ **do**
4: $\quad$ Update the weight coefficient σ_i and compute the latest global estimate as follows

$$\tilde{\mu}_{i,j} \leftarrow (1 - \sigma_i) \sum_{j' \in \mathcal{N}_j \cup \{j\}} \omega_{j,j'} \tilde{\mu}_{i,j'} + \sigma_i X_{i,j}$$

5: **end for**
6: **for** $j' \in \mathcal{N}_j$ **do**
7: $\quad$ Update the elimination arm set via $\mathcal{B}_j \leftarrow \mathcal{B}_j \cup \mathcal{B}_{j'}$
8: **end for**

The key idea of `CES` is synthesizing the information exchanged from each agent's neighborhood and estimating the global mean without bias. In this section, we propose a *fair* mechanism where the samples of all agents are equally used to estimate the global mean. By properly configuring `CES`, each agent ensures a fair global estimate, which identifies suboptimal arms more accurately and rapidly.

In `CES`, agent j combines the historical data from its neighborhood and its own real-time reward to obtain biased global estimates. As an example, we focus on demonstrating the consensus process in estimating the global mean of arm i. Up to time slot t, agent j has sampled arm i for $\tau_{i,j}(t)$ times and the reward obtained at $\tau_{i,j}(t)$-th sample is defined as $X_{i,j}^{\tau_{i,j}(t)}$. The global estimate of agent j on arm i is also defined as $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$. In the communication phase, agent j exchanges its previous estimate $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)-1}$ among its neighborhood $\mathcal{N}_j$ (Line 1-2, Algorithm 2). Based on the historical observations from the neighborhood and the real-time reward $X_{i,j}^{\tau_{i,j}(t)}$, each agent j updates its latest global estimate $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$ as follows

$$\begin{aligned}\tilde{\mu}_{i,j}^{\tau_{i,j}(t)} :=& (1 - \sigma_i(\tau_{i,j}(t))) \sum_{j' \in \mathcal{N}_j \cup \{j\}} \omega_{j,j'} \tilde{\mu}_{i,j'}^{\tau_{i,j'}(t)-1} \\ &+ \sigma_i(\tau_{i,j}(t)) X_{i,j}^{\tau_{i,j}(t)},\end{aligned} \tag{8}$$

where $\sigma_i(\tau_{i,j}(t))$ represents the weight coefficient that adjusts the contribution of each piece of information in the global estimate $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$. Additionally, the elimination arm set $\mathcal{B}_j(t)$ is also updated in `CES` (Lines 6-7, Algorithm 2).

Algorithm 2 provides the latest global estimates and the updated elimination arm set to `DRRB-bandit`.

4 REGRET ANALYSIS

In this section, we summarize the theoretical results for the FMAB problem and present the near-optimal results from the perspectives of both individual and group regret. To derive these regret bounds, we first introduce the following lemma, which characterizes a tighter confidence interval for the estimates compared to the previous work. Then, we present the results for FMAB in the form of Theorems.

The following lemma demonstrates the performance of `CES`, which achieves a bounded estimation error, with the upper bound of the error decreasing as the agent number N based on limited samples.

Lemma 1. *Assume that $X_{i,j}$ is an i.i.d. reward process with unknown mean $\mu_{i,j}$. Set $\sigma_i(\tau_{i,j}(t)) = \frac{1}{\tau_{i,j}(t)+1}$. Then, for any arm $i \in \mathcal{K}$, agent $j \in \mathcal{N}$ and time slot $t \in \{1, \dots, T\}$, with probability $1-2\delta$, $\delta \in (0, 0.5)$, we have*

$$|\tilde{\mu}_{i,j}^{\tau_{i,j}(t)} - \mu_i| \leq \sqrt{\frac{\log \delta^{-1}}{2N\tau_{i,j}(t)}} + \frac{Q}{(1+\tau_{i,j}(t))(1-\lambda_2)},$$

where Q is determined by the communication graph $\mathcal{G}$ and λ_2 is the second largest eigenvalue of matrix W. We have $Q = 1$ if the graph $\mathcal{G}$ is balanced, otherwise, $Q = \sqrt{N}$.

Proof Sketch of Lemma 1. The term $|\tilde{\mu}_{i,j}^{\tau_{i,j}(t)} - \mu_i|$ can be upper bounded by $|\hat{\mu}_{i,j}^{\tau_{i,j}(t)} - \mu_i| + |\tilde{\mu}_{i,j}^{\tau_{i,j}(t)} - \hat{\mu}_i^{\tau_{i,j}(t)}|$ via the triangle inequality. The variable $\hat{\mu}_i^{\tau_{i,j}(t)}$ represents the global estimate under the full information communication, i.e., the sample rewards of the arm i of all agents are accessible. The first term $|\hat{\mu}_{i,j}^{\tau_{i,j}(t)} - \mu_i|$ is bounded by Hoeffding's inequality (Lemma 4). For the second term, it is obtained by exploiting the properties of graph theory (Lemma 2): Iterating equation (8) yields the relation between $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$ and $X_{i,j}^{\tau_{i,j}(t)}$, where each part matches that in $\hat{\mu}_i^{\tau_{i,j}(t)}$. Detailed proofs are provided in Appendix D.1.

Lemma 1 provides a better confidence interval for the global estimates $\tilde{\mu}_{i,j}$ compared to previous works. The radius of the confidence interval is N^2 and N times smaller than that in Zhu et al. [2021] and Xu and Klabjan [2024], respectively. While Zhu and Liu [2023] achieved similar performance, their results only hold for fully connected communication graphs, i.e., each agent is directly connected to all others. The superiority of our interval is clearly reflected in equation (8), which ensures that the proportion of each reward $X_{i,j}^{\tau_{i,j}(t)}$ in the global estimate $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$ is $\frac{1}{N\tau_{i,j}(t)}$, allowing agents to estimate the global mean more accurately.

4.1 UPPER BOUNDS

Theorem 1 (Regret upper bound)**.** *Let $U_{i,j}(t,\delta)$ in equation* (4) *with $\delta = T^{-2}$ be the radius of the confidence interval of a random $[0,1]$-valued i.i.d. process. Given $\gamma > 0$, `DRRB-bandit` for FMAB problems achieves the following performance, with a probability of at least $1 - 2TKN\delta$.*

(i) Individual regret:

$$\mathbb{E}[R_j^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{N\Delta_i} + \sum_{i:\Delta_i>0} (D+1)\Delta_i + \frac{8Q(K-1)}{1-\lambda_2} + 1,$$

(ii) Group regret:

$$\mathbb{E}[R^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i} + \sum_{i:\Delta_i>0} N(D+1)\Delta_i + \frac{8NQ(K-1)}{1-\lambda_2} + 1.$$

Proof Sketch of Theorem 1. To bound the individual regret $R_j^T(\mathcal{A})$, we first need to determine an upper bound for the sample counts. Based on equation (5), we can derive an instance-dependent upper bound for the sample counts. However, in distributed networks, communication delays between agents cause the upper bound derived from equation (5) to be inaccurate. To ensure synchronization in sampling, agents will additionally sample suboptimal arms. Therefore, we bound the sample count by considering the diameter of the communication graph and the theoretical result from equation (5). Subsequently, by performing a regret decomposition, we combine the regrets for each arm sampled by agent j to obtain the individual regret $R_j^T(\mathcal{A})$. Detailed proofs are provided in Appendix D.4.

In Theorem 1, there exists an uncertain term $\frac{1}{1-\lambda_2}$ which is bounded in the following corollary.

Corollary 1 (An extension of Theorem 1)**.** *Under the condition of Theorem 1, the further bound of the regret is*

(i) Individual regret:

$$\mathbb{E}[R_j^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{N\Delta_i} + 8KQDN^2 + \sum_{i:\Delta_i>0} (D+1)\Delta_i + 1,$$

(ii) Group regret:

$$\mathbb{E}[R^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i} + 8KQDN^3 + \sum_{i:\Delta_i>0} N(D+1)\Delta_i + 1.$$

Corollary 2 (Instance-independent regret bound)**.** *Under the conditions of Theorem 1, the instance-independent upper bound of* `DRRB-bandit` *for FMAB problems achieves the following performance:*

(i) Individual regret:

$$\mathbb{E}[R_j^T(\mathcal{A})] \leq 8\sqrt{\frac{KT\log T}{N}} + K(D+1) + \frac{8QK}{1-\lambda_2} + 1,$$

(ii) Group regret:

$$\mathbb{E}[R^T(\mathcal{A})] \leq 8\sqrt{KNT\log T} + KN(D+1) + \frac{8NQK}{1-\lambda_2} + 1.$$

Theorem 2 (Communication cost)**.** *Under the conditions of Theorem 1,* `DRRB-bandit` *suffers the communication cost at most*

$$\mathbb{E}[C^T(\mathcal{A})] \leq \frac{16K\log T}{\Delta_{\min}^2} + \frac{8KNQ}{(1-\lambda_2)\Delta_{\min}} + KN(D+1),$$

where $\Delta_{\min} = \min_{i:\Delta_i>0}\Delta_i$.

Proof Sketch of Theorem 2. In the proof of Theorem 1, one can deduce that the suboptimal arm i is sampled by agent j at most $\frac{8\log\delta^{-1}}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + D + 1$ times. In Theorem 1, the violation probability is denoted by $\delta = \frac{1}{T^2}$, then the sample count is bounded by

$$\tau_{i,j} \leq \frac{16\log T}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + D + 1.$$

In each round, `DRRB-bandit` collects information about all arms and communicates it with other agents in a single batch. Therefore, to determine the maximum number of communications, it suffices to consider the number of samples of the arm that remain in the candidate set for the second longest period.

Remark 1. *Although the proposed algorithm requires knowledge of the time horizon T to set δ and achieve near-optimal regret performance, in practice, when T is unknown, the algorithm can still perform similarly when designing δ as a tunable function, such as $\delta = 1/t^2$.*

4.2 LOWER BOUNDS

Besides the upper bounds, we also present lower bounds for FMAB problems. We investigate the lower bounds of both individual and group regrets. For the regret lower bound, we derive two separate lower bounds corresponding to two distinct cases (Theorems 3 and 4).

Theorem 3 (General regret lower bound)**.** *For FMAB problems with any number of agents, arms, and stochastic rewards satisfying a* 1*-Gaussian distribution, if the graph $\mathcal{G}$ is connected, any federated bandit algorithm must incur regrets at least:*

(i) Individual regret:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_j^T(\mathcal{A})]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{N^2\Delta_i}.$$

(ii) Group regret:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R^T(\mathcal{A})]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{N\Delta_i}.$$

Theorem 4 (Regret lower bound for algorithms with round-robin sampling)**.** *For FMAB problems with any number of agents, arms, and stochastic rewards satisfying a* 1*-Gaussian distribution, if the graph $\mathcal{G}$ is connected, any federated bandit algorithm using round-robin sampling must incur regrets at least:*

(i) Individual regret:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_j^T(\mathcal{A})]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{N\Delta_i}.$$

(ii) Group regret:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R^T(\mathcal{A})]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}.$$

Remark 2. *In Section 4.2, we present two types of lower bounds for FMAB problems: one general bound in Theorem 3, and a specific bound for the class of round-robin-based algorithms in Theorem 4. Theorem 3 provides general lower bounds for individual regret $\Omega(\sum_{i:\Delta_i>0} N^{-2}\Delta_i^{-1}\log T)$ and group regret $\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ under the strict assumption that all other agents' reward means are equal. This implies that each agent only needs to learn its local reward means. Theorem 4 gives the individual regret bound $\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ and group regret bound $\Omega(\sum_{i:\Delta_i>0} \Delta_i^{-1}\log T)$ for all round-robin-based algorithms.*

Remark 3. *According to Theorems 1 and 3, we have shown that the lower and upper bounds match in terms of agent number N, reward gap Δ_i, and time horizon T for the class of algorithms based on round-robin sampling. However, for general algorithms, we have been unable to prove the optimality of* `DRRB-bandit` *due to the complexity of decision-making in multi-agent systems. Recalling the algorithms from previous works, we have improved both the individual and group regret bounds to $O(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ and $O(\sum_{i:\Delta_i>0} \Delta_i^{-1}\log T)$, respectively.*

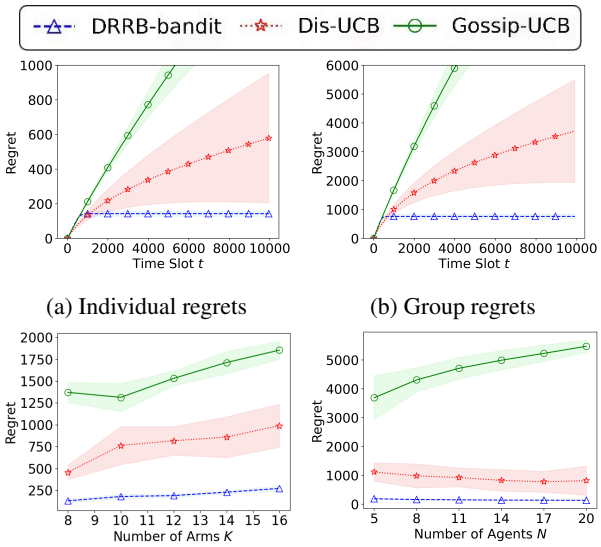
Remark 4. *In the special case of homogeneous FMAB problems ($\mu_{i,j} = \mu_i$ for all agents), the regret upper bounds of Theorem 1 match the known individual and group lower bounds, $\Omega(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ and $\Omega(\sum_{i:\Delta_i>0} \Delta_i^{-1}\log T)$ [Wang et al., 2020, Wang and Yang, 2023]. Therefore, our algorithm, reduced to the easier homogeneous setting, is near-optimal.*

5 EXPERIMENTS

In this section, we conduct a series of numerical experiments to evaluate the performance of `DRRB-bandit`. All experiments are repeated for 50 trials, with the means plotted as lines and the standard deviations represented by shaded regions.

Setups and Baselines. In `DRRB-bandit`, we set the parameters as $N = 8$, $K = 10$, $Q = \sqrt{N}$ and $\delta = 1/T^2$. To ensure a fair comparison, we use a ring graph, which is a connected graph with a second-largest eigenvalue $\lambda_2 = 0.5713$. Each agent is connected to four neighbors with whom it exchanges information. We compare both individual and group regrets of `DRRB-bandit` against two baselines, `Gossip-UCB` Zhu et al. [2021] and `Dis-UCB` Zhu and Liu [2023], as outlined in Table 1. Both algorithms tackle the federated bandit problem within the UCB framework but fail to exploit the advantages of distributed learning fully. Consequently, the individual regret of `Dis-UCB` remains independent of the number of agents N. In contrast, the group regret grows linearly with N, highlighting a fundamental limitation in their distributed learning design. This shortcoming stems from the fact that these algorithms do not effectively utilize the multi-agent system's collaborative potential, and their consensus-based decision-making overlooks the benefits that can arise from coordinated exploration.

Observations. Figure 1 reports regrets for the proposed algorithm and the baselines. Figures (1a) and (1b) show individual and group regrets for the three algorithms. These figures demonstrate that `DRRB-bandit` does not perform well during the initial phase, as it performs uniform sampling across all arms. However, after sufficient sampling, all agents successfully eliminate the suboptimal arms, and the regret stabilizes, remaining almost unchanged for the remainder of the period. In Figure (1c), it is obvious that the increasing trend aligns with the individual regret bound $O(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$, which increases with the number of arms K. This phenomenon can be easily explained: as the number of arms increases, the task of learning each arm's reward becomes more difficult, leading to higher regret. Finally, when varying the number of agents, we observe a decreasing trend that corresponds to the $O(\sum_{i:\Delta_i>0} N^{-1}\Delta_i^{-1}\log T)$ individual regret bounds, which decrease with the number of agents N. In contrast, `Gossip-UCB` shows increasing regret, consistent with its regret bound $O(\sum_{i:\Delta_i>0} N\Delta_i^{-1}\log T)$. For `Dis-UCB`, since the number of neighbors for each agent remains fixed when the number of agents changes, we also observe increasing regret, as shown in Figure (1d). We also provide additional simulations of the homogeneous setting in Appendix E.



(a) Individual regrets (b) Group regrets

(c) Regrets with different numbers of arms (d) Regrets with different numbers of agents

Figure 1: Performance comparison with different arms and agents.

6 CONCLUSION

This work focuses on the study of FMAB problems and introduces a fully distributed algorithm called `DRRB-bandit`. To address the challenge of heterogeneous feedback, we propose a consensus estimation subroutine that allows agents to estimate the global mean of each arm by only communicating with their neighbors. This approach improves convergence speed compared to previous methods by effectively balancing the contribution of each agent's data. According to the works above, the proposed algorithm reduces individual and group upper regrets. Additionally, we discuss the lower bounds for the heterogeneous federated bandit problem, proving that our algorithm achieves near-optimal performance among those using Round-Robin sampling.

7 ACKNOWLEDGMENT

We thank our colleague Mengfan Xu, as well as our anonymous reviewers, for their valuable feedback. This work was supported by NSFC (no. 62306138), JiangsuNSF (no. BK20230784), and the Innovation Program of State Key

Laboratory for Novel Software Technology at Nanjing University (no. ZZKT2025B25).

References

Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *Journal of Machine Learning Research*, 23(212):1–24, 2022.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

Raphaël Féraud, Réda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909. PMLR, 2019.

Jonathan L Gross and Jay Yellen. *Handbook of graph theory*. CRC press, 2003.

Jiafan He, Tianhao Wang, Yifei Min, and Quanquan Gu. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. *Advances in neural information processing systems*, 35:4762–4775, 2022.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in neural information processing systems*, 34:27057–27068, 2021.

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 482–491. IEEE, 2003.

Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

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

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

Xinbin Li, Yi Zhou, Lei Yan, Haihong Zhao, Xiaodong Yan, and Xi Luo. Optimal node selection for hybrid attack in underwater acoustic sensor networks: A virtual expert-guided bandit algorithm. *IEEE Sensors Journal*, 20(3): 1679–1687, 2019.

Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.

Michael Molloy and Bruce Reed. *Graph colouring and the probabilistic method*, volume 23. Springer Science & Business Media, 2002.

Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. 2013.

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

Jeff Shamma. *Cooperative control of distributed multi-agent systems*. John Wiley & Sons, 2008.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9603–9611, 2021.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

Xuchuang Wang and Lin Yang. Achieving near-optimal individual regret low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Xuchuang Wang, Lin Yang, Yu-zhen Janice Chen, Xutong Liu, Mohammad Hajiesmaili, Don Towsley, and John C.S. Lui. Achieve near-optimal individual regret & low communications in multi-agent bandits. In *International Conference on Learning Representations*, 2023a.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*, 2019.

Zichen Wang, Chuanhao Li, Chenyu Song, Lianghui Wang, Quanquan Gu, and Huazheng Wang. Pure exploration in asynchronous federated bandits. *arXiv preprint arXiv:2310.11015*, 2023b.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Feng Yan, Shreyas Sundaram, SVN Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25 (11):2483–2493, 2012.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Jingxuan Zhu and Ji Liu. Distributed multi-armed bandit over arbitrary undirected graphs. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6976–6981. IEEE, 2021.

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

Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(1):1–29, 2021.

Supplementary Material

Haoran Zhang[1,4] **Xuchuang Wang**[2] **Haoxu Chen**[1,4] **Hao Qiu**[3] **Lin Yang**[†1,4] **Yang Gao**[1,4]

[1]School of Intelligent Science and Technology, Nanjing University, Suzhou, China
[2]College of Information & Computer Science, University of Massachusetts Amherst, Massachusetts, USA
[3]Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[4]National Key Laboratory for Novel Software Technology, China

A APPENDIX / SYMBOL EXPLANATION

For all symbols in this article, we give explanations of them in Table 2.

B APPENDIX / SOME KNOWLEDGE OF GRAPHS

Throughout this paper, we consider FMAB problems with N agents operating in a time-invariant network. The network is represented by a communication graph $\mathcal{G}(\mathcal{N}, \mathcal{E}, \mathcal{X})$, which consists of three components:

1. $\mathcal{N} = \{1, \ldots, N\}$ is the set of agents in the network, corresponding to the number of agents in the distributed system.
2. $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ is the edge set, which determines the connectivity between agents.
3. $\mathcal{X} = [a_{j,j'}]_{N\times N}$ is the adjacency matrix of the graph $\mathcal{G}$, where $a_{j,j'}$ denotes the weight of the edge between agents j and j'.

Notably, the adjacency matrix represents the importance of one agent to its neighbors and encodes neighborhood information in $\mathcal{G}$. Specifically, $a_{j,j'}$ is the weight from agent j' to agent j. Since the graph is directed, we have $a_{j,j'} \neq a_{j',j}$.

The graph has no self-loops, meaning that $a_{j,j} = 0$ for all $j \in \mathcal{N}$. An edge between agents j and j' exists if and only if $a_{j,j'} \neq 0$, i.e., $(j, j') \in \mathcal{E}$.

For each agent j, its neighborhood is denoted as $\mathcal{N}_j = \{j' \mid j' \in \mathcal{N}, a_{j,j'} \neq 0, j' \neq j\}$. Finally, we define the diameter of the graph $\mathcal{G}$ as D, which represents the longest distance between any two agents in the network.

For graph $\mathcal{G}$, its corresponding Laplacian matrix $\mathcal{L}$ is defined as follows

$$\mathcal{L}_{j,j'} = \begin{cases} -a_{j,j'}, & j \neq j' \\ \sum_{k=1}^{N} a_{j,k}. & j = j' \end{cases}$$

The maximum degree of graph $\mathcal{G}$ is defined as $\epsilon = \max_i(\sum_{j'\neq j} a_{j,j'})$. Then, for any constant $\beta \in (0, 1/\epsilon]$, the Perron matrix $W = I - \beta\mathcal{L}$ could be obtained. The Perron matrix $\boldsymbol{W} = [\omega_{i,j}]_{N\times N}$ is a doubly random matrix and both the sum of row elements and column elements in $\boldsymbol{W}$ is 1. In the multi-agent bandit setting, it is widely used to solve the consensus problem [Olfati-Saber et al., 2007].

*Corresponding author: linyang@nju.edu.cn
†Corresponding author: linyang@nju.edu.cn

Symbol/Term	Definition
$\mathcal{G}(\mathcal{N}, \mathcal{E}, \mathcal{A})$	A graph to describe a multi-agent system
$\mathcal{N} = \{1, \ldots, N\}$	Set of agents in a multi-agent system
$\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$	The edge set in graph $\mathcal{G}$
$\mathcal{X} = [a_{i,j}]_{N \times N}$	The weight matrix to describe the relations between agents
$\mathcal{N}_j$	Neighborhood of agent j, excluding agent j
$\boldsymbol{W} = [\omega_{a,b}]_{N \times N}$	Communication matrix
D	The diameter of graph $\mathcal{G}$
λ_2	The second largest eigenvalue of $\boldsymbol{W}$
Q	A symbol to describe whether the graph is balanced
$\mathcal{K} = \{1, \ldots, K\}$	Set of arms in a multi-armed bandit (MAB) problem
T	Total number of time slots
$A_j(t)$	Arm chosen by agent j at time slot t
$X_{A_j(t),j}(t)$	Random reward received by agent j after pulling arm $A_j(t)$ at time slot t
$X_{i,j}(t)$	Random reward of arm i observed by agent j at time slot t
$\mu_{i,j}$	Mean reward of arm i observed by agent j, bounded in $[0, 1]$
$X_i(t)$	Global reward of arm i at time slot t
μ_i	Global mean reward of arm i
$\tilde{\mu}_{i,j}(t)$	The global estimate of agent j of arm i at time slot t
$i^\star$	The unique optimal arm with the largest global mean reward
$\mathcal{S}_j(t)$	The candidate arm set of agent j at time slot t
$\mathcal{B}_j(t)$	The elimination arm set of agent j at time slot t
t_i	The time label attached to arm i
$\Delta_i = \mu_{i^\star} - \mu_i$	Reward gap between the optimal arm and arm i
$\tau_{i,j}(t)$	The sample count of agent j on arm i until time slot t
$\tau_i(t)$	The global sample count on arm i
δ	The violation probability of confidence interval
$U_{i,j}(t, \delta)$	The radius of confidence interval
$\texttt{UCB}_{i,j}$	The upper confidence bound of agent j on arm i
$\texttt{LCB}_{i,j}$	The lower confidence bound of agent j on arm i
$\mathbb{E}[R^T(\mathcal{A})]$	Expected group regret for a distributed algorithm $\mathcal{A}$
$\mathbb{E}[R_j^T(\mathcal{A})]$	Expected individual regret for agent j in a distributed algorithm $\mathcal{A}$

Table 2: Summary of symbols and Definitions

C APPENDIX / PRELIMINARIES OF THE PROBLEM

Lemma 2. *[Yan et al., 2012] For a doubly random matrix W, it is an irreducible, doubly stochastic matrix with strictly positive diagonal entries. Then, there exists a positive constant Q such that*

$$\sum_{j=1}^{N} \left| \omega_{i,j}^k - \frac{1}{N} \right| < Q\lambda_2^k,$$

where $\omega_{i,j}^k$ represents the element in the i-th row and j-th column of the matrix W^k, k represents the iteration step, and λ_2 is the second largest eigenvalue of matrix W. Q is equal to 1 if the graph $\mathcal{G}$ is balanced; otherwise, $Q = \sqrt{N}$.

Lemma 3. *[Lattimore and Szepesvári, 2020] Suppose that X_i is σ_i^2 sub-Gaussian and X_i are all independent for $i \in \{1, \ldots, N\}$. Then we have $\frac{1}{N}\sum_{i=1}^{N} X_i$ is $\frac{\sum_i^N \sigma_i^2}{N^2}$ sub-Gaussian.*

Lemma 4. *[Molloy and Reed, 2002] Assume that $X(t) - \mu$ is independent, σ^2 sub-Gaussian random variable. Then for any $\epsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}),$$
$$\mathbb{P}(\hat{\mu} \leq \mu - \epsilon) \geq \exp(-\frac{n\epsilon^2}{2\sigma^2}),$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X(t)$ *and* n *is the sample count.*

Lemma 5. *[Dubhashi and Panconesi, 2009] If a random variable* X *has a finite mean and* $a \leq X \leq b$ *almost surely, then* X *is* $\frac{1}{4}(b-a)^2$ *sub-Gaussian.*

Lemma 6. *[Gross and Yellen, 2003] For a strong connected graph* $\mathcal{G}$ *with* N *nodes and diameter* D*, the second largest eigenvalue of Perron matrix* W *is bounded by*

$$\lambda_2 \leq 1 - \frac{\beta}{ND},$$

with $\beta \in (0, 1/\epsilon]$ *and* $\epsilon = \max_i(\sum_{j' \neq j} a_{j,j'})$.

Lemma 7. *Lattimore and Szepesvári [2020] In FMAB problems, let* $\mathcal{D}_j(i)$ *denote the event that agent* j *eliminates the optimal arm* $i^\star$ *in favor of some suboptimal arm* i*. Then, the probability of this event is bounded by*

$$\mathbb{P}(\mathcal{D}_j(i)) \leq \delta.$$

Assuming the violation probability is $\delta = \frac{1}{T^2}$*, the regrets incurred from the erroneous elimination of the optimal arm are of* $O(1)$ *order.*

D APPENDIX / MISSED PROOFS

D.1 PROOF OF LEMMA 1

Proof. To clearly illustrate the relationship between the sampling count and the global estimate, let $X_{i,j}^{\tau_{i,j}(t)}$ represent the reward, and $\tilde{\mu}_{i,j}^{\tau_{i,j}(t)}$ the global estimate, both for agent j when pulling arm i at the $\tau_{i,j}(t)$-th sample. Here, $\tau_{i,j}(t)$ denotes the number of times agent j has pulled arm i up to time slot t. Benefiting from the design of Algorithm 1, all agents sample arm i at the same frequency. In the proof, we focus on arm i and use τ to represent the sampling count $\tau_{i,j}$ for simplicity.

For the sake of computation, stack the value of $\tilde{\mu}_{i,j}^{\tau}$ and $X_{i,j}^{\tau}$ into vectors as follows

$$\begin{aligned}\boldsymbol{\tilde{\mu}_i^{\tau}} &:= [\tilde{\mu}_{i,1}^{\tau}, \ldots, \tilde{\mu}_{i,N}^{\tau}]^T, \\ \boldsymbol{X_i^{\tau}} &:= [X_{i,1}^{\tau}, \ldots, X_{i,N}^{\tau}]^T.\end{aligned}$$

Stacking all global estimates $\tilde{\mu}_{i,j}^{\tau}$, equation (8) can be rewritten as

$$\boldsymbol{\tilde{\mu}_i^{\tau}} = (1 - \sigma_i(\tau))\boldsymbol{W}\boldsymbol{\tilde{\mu}_i^{\tau-1}} + \sigma_i(\tau)\boldsymbol{X_i^{\tau}}. \tag{9}$$

Substituting $\sigma_i(\tau) = \frac{1}{\tau+1}$ into (9) and iterating it, we have

$$\boldsymbol{\tilde{\mu}_i^{\tau}} = \frac{1}{\tau+1}\boldsymbol{W}^{\tau}\boldsymbol{\tilde{\mu}_i^{0}} + \frac{1}{\tau+1}\sum_{k=1}^{\tau}\boldsymbol{W}^{\tau-k}\boldsymbol{X_i^{k}}.$$

In Algorithm 1, when $t = 0$, there is no communication between agents. Hence, we denote $\tilde{\mu}_{i,j}^{0} = X_{i,j}^{0}$ (Line 2, Algorithm 1). Then, the above equation could be rewritten as

$$\boldsymbol{\tilde{\mu}_i^{\tau}} = \frac{1}{\tau+1}\sum_{k=0}^{\tau}\boldsymbol{W}^{\boldsymbol{\tau-k}}\boldsymbol{X_i^{k}}.$$

Spitting the elements from $\boldsymbol{\tilde{\mu}_i^{\tau}}$, the global estimate $\tilde{\mu}_{i,j}^{\tau}$ of agent j for arm i is as follows

$$\tilde{\mu}_{i,j}^{\tau} = \frac{1}{\tau+1}\sum_{k=0}^{\tau}\sum_{j'=1}^{N}\omega_{j,j'}^{\tau-k}X_{i,j'}^{k}.$$

Meanwhile, the global estimate under the full information communication is written as

$$\hat{\mu}_i^\tau = \frac{1}{N}\sum_{j=1}^{N} \bar{X}_{i,j}^\tau, \tag{10}$$

where

$$\bar{X}_{i,j}^\tau = \frac{1}{\tau+1}\sum_{k=0}^{\tau} X_{i,j}^k.$$

Subtracting $\hat{\mu}_i^\tau$ from $\tilde{\mu}_{i,j}^\tau$, we have

$$\begin{aligned}\tilde{\mu}_{i,j}^\tau - \hat{\mu}_i^\tau &= \frac{1}{\tau+1}\sum_{k=0}^{\tau}\sum_{j'=1}^{N} \omega_{j,j'}^{\tau-k} X_{i,j'}^k - \frac{1}{\tau+1}\sum_{k=0}^{\tau}\sum_{j=1}^{N}\frac{1}{N}X_{i,j}^k \\ &= \frac{1}{\tau+1}\sum_{k=0}^{\tau}\sum_{j'=1}^{N}(\omega_{j,j'}^{\tau-k} - \frac{1}{N})X_{i,j'}^k.\end{aligned}$$

Then, substituting Lemma 2 into $\tilde{\mu}_{i,j}^\tau - \hat{\mu}_i^\tau$, we have

$$|\tilde{\mu}_{i,j}^\tau - \hat{\mu}_i^\tau| < \frac{1}{\tau+1}\sum_{k=0}^{\tau} Q\lambda_2^{\tau-k} = \frac{1-\lambda_2^{\tau+1}}{1-\lambda_2}\cdot\frac{Q}{\tau+1} \le \frac{Q}{(1-\lambda_2)(\tau+1)}, \tag{11}$$

where the constant Q depends on the matrix $\boldsymbol{W}$. When the doubly random matrix $\boldsymbol{W}$ is symmetric, we have $Q = 1$. Otherwise, we have $Q = \sqrt{N}$. The details are also shown in Lemma 8 in Zhu and Liu [2023].

The goal is to obtain an unbiased estimation $\tilde{\mu}_{i,j}^\tau$ on the global mean μ_i. To achieve the goal, we could divide the problem into two parts: $\tilde{\mu}_{i,j}^\tau - \hat{\mu}_i^\tau$ and $\hat{\mu}_i^\tau - \mu_i$. According to the triangle inequality, we have

$$|\tilde{\mu}_{i,j}^\tau - \mu_i| \le |\tilde{\mu}_{i,j}^\tau - \hat{\mu}_i^\tau| + |\hat{\mu}_i^\tau - \mu_i|. \tag{12}$$

From equation (12) we have completed the bound between $\tilde{\mu}_{i,j}^\tau$ and $\hat{\mu}_i^\tau$. How to prove the bound of $\hat{\mu}_i^\tau - \mu_i$ is what we require to consider in the following proof. According to the definition of $\hat{\mu}_i^\tau$ in (10), $\hat{\mu}_i^\tau$ could be rewritten as

$$\hat{\mu}_i^\tau = \frac{1}{N}\sum_{j=1}^{N}\bar{X}_{i,j}^\tau = \frac{1}{N(\tau+1)}\sum_{j=1}^{N}\sum_{k=0}^{\tau} X_{i,j}^k,$$

For arm i, the global reward $X_i^\tau = \frac{1}{N}\sum_{j=1}^{N} X_{i,j}$ is considered as a linear combination reward of $X_{i,j}^\tau, j \in \mathcal{N}$. According to Lemma 5, the $[0,1]$-valued variable $X_{i,j}$ could be considered as a $\frac{1}{4}$ sub-Gaussian variable. Since that $X_{i,j}^\tau, i \in \mathcal{K}$ are all independent sub-Gaussian variables, we could deduce that X_i^τ is a $\frac{1}{4N}$ sub-Gaussian variable from Lemma 3.

Assume that arm i is pulled by N agents for τ times, then it follows from Lemma 4 that

$$\begin{aligned}\mathbb{P}(\hat{\mu}_i^\tau \ge \mu_i + \varepsilon) &\le \exp\left(\frac{-\tau\varepsilon^2}{2\sigma^2}\right), \\ \mathbb{P}(\hat{\mu}_i^\tau \le \mu_i - \varepsilon) &\le \exp\left(\frac{-\tau\varepsilon^2}{2\sigma^2}\right).\end{aligned} \tag{13}$$

where $\varepsilon = \sqrt{\frac{\log\delta^{-1}}{2N\tau}}$ and $\sigma^2 = \frac{\sum_{i=1}^{N}\sigma_i^2}{4N^2} = \frac{1}{4N}$. Make a transformation on (13), we have

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_i(t) \ge \mu_i + \sqrt{\frac{\log\delta^{-1}}{2Nt}}\right) &\le \delta, \\ \mathbb{P}\left(\hat{\mu}_i(t) \le \mu_i - \sqrt{\frac{\log\delta^{-1}}{2Nt}}\right) &\le \delta.\end{aligned} \tag{14}$$

Furthermore, combining the two inequalities in (14) yields

$$\mathbb{P}\left(\mu_i - \sqrt{\frac{\log \delta^{-1}}{2Nt}} \leq \hat{\mu}_i^{\tau} \leq \mu_i + \sqrt{\frac{\log \delta^{-1}}{2Nt}}\right) \geq 1 - 2\delta.$$

With probability at least $1 - 2\delta$, equation (12) can be written as

$$|\tilde{\mu}_{i,j}^{\tau} - \mu_i| \leq \sqrt{\frac{\log \delta^{-1}}{2N\tau}} + \frac{Q}{(1-\lambda_2)(\tau+1)}. \tag{15}$$

□

D.2 PROOF OF LEMMA 6

Proof. Due to that the relationship between the Laplacian matrix and the Perron matrix is $\boldsymbol{W} = \boldsymbol{I} - \beta\boldsymbol{\mathcal{L}}$, the second largest eigenvalue of $\boldsymbol{W}$ is equal to the smallest nonzero eigenvalue of $\boldsymbol{\mathcal{L}}$, which is the algebraic connectivity of graph $\mathcal{G}$. From reference [Gross and Yellen, 2003], the smallest nonzero eigenvalue $\hat{\lambda}_2$ of $\boldsymbol{\mathcal{L}}$ is bounded by $\hat{\lambda}_2 \geq \frac{1}{ND}$. Then, one can deduce that the second largest eigenvalue λ_2 of the Perron matrix $\boldsymbol{W}$ is

$$\lambda_2 \leq 1 - \beta\hat{\lambda}_2 \leq 1 - \frac{\beta}{ND}.$$

□

D.3 PROOF OF LEMMA 7

Proof. According to the definition of arm elimination, the emergence of event $\mathcal{D}_j(i)$ implies equation (5). Based on Lemma 4, the probability of event $\mathcal{D}_j(i)$ is as follows

$$\mathbb{P}(\mathcal{D}_j(i)) \leq \exp\left(\frac{-4\tau_{i,j}U_{i,j}^2(t,\delta)}{2 \times \frac{1}{4N}}\right) \leq \exp\left(-2N\tau_{i,j}\frac{\log \delta^{-4}}{2N\tau_{i,j}}\right) \leq \delta^4 \leq \delta.$$

In Theorem 1, the violation probability of the confidence interval is $\delta = \frac{1}{T^2}$. The expected regret caused by erroneous elimination is bounded by

$$\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{i:\Delta_i>0}\Delta_i\mathbb{P}(\mathcal{D}_j(i)) \leq NTK\delta \leq \frac{NK}{T},$$

which is in order $O(1)$. □

D.4 PROOF OF THEOREM 1

Proof. **Step 1: Bound the sample count of each arm i for agent j**

Recall that for all arms $i \in \mathcal{K} \setminus \{i^\star\}$, $\Delta_i > 0$. For the two regrets defined in equations (1) and (2), the practical sample count $\tau_{i,j}(T)$ is of primary interest. However, in the case of `DRRB-bandit`, $\tau_{i,j}(t)$ is difficult to determine directly because additional sample counts may exist, even if the arm does not satisfy the condition in equation (8). This is because agents with strong learning capabilities need to maintain the same update of the candidate arm set as other agents, which could lead to an increase in regrets.

To address this, we introduce an auxiliary variable, $\hat{\tau}_{i,j}(t)$, representing the theoretical sample count for agent j when sampling arm i. This can be computed using equation (8). The variable $\hat{\tau}_{i,j}(t)$ corresponds to the sample count when arm i has been included in the set $\mathcal{B}_j$, while $\tau_{i,j}(t)$ represents the time when arm i is excluded from the set $\mathcal{S}_j$. Therefore, when computing the upper bound of the sample count, we can replace $\tau_{i,j}(t)$ with $\hat{\tau}_{i,j}(t)$ as described in equation (8). For the two sample counts mentioned above, the relationship between them is

$$\hat{\tau}_{i,j}(t) \leq \tau_{i,j}(t) \leq \hat{\tau}_{i,j}(t) + D. \tag{16}$$

According to the explanation above, the sample counts of all arms in candidate set $\mathcal{S}_j$ are the same for all agents in agent set $\mathcal{N}$. Since that $\tilde{\mu}_{i,j}$ estimates μ_i, the reward gap Δ_i for each agent $j \in \mathcal{N}$ is related to $U_{i,j}(t,\delta)$. Equation (5) is equal to

$$2U_{i,j}(t,\delta) \geq \tilde{\mu}_{i^{\max},j}(t) - \tilde{\mu}_{i,j}(t) \overset{(a)}{\geq} \Delta_i - 2U_{i,j}(t,\delta), \tag{17}$$

where inequality (a) is from $\tilde{\mu}_{i^{\max},j}(t) \geq \tilde{\mu}_{i^\star,j}(t) \geq \mu_{i^\star} - U_{i,j}(t)$ and $\tilde{\mu}_{i,j}(t) \leq \mu_i + U_{i,j}(t)$.

Let $A_{i,j,t}$ denote the event in which agent j pulls arm i at time slot t, then we have

$$\mathbb{P}\left(\bigcap_{i,j,t} A_{i,j,t}\right) = 1 - \mathbb{P}\left(\bigcup_{i,j,t} \neg A_{i,j,t}\right) \geq 1 - \sum_{i,j,t} \mathbb{P}\left(\neg A_{i,j,t}\right) \geq 1 - 2tNK\delta.$$

Replacing $\tau_{i,j}(t)$ with $\hat{\tau}_{i,j}(t)$ in functions $U_{i,j}(t,\delta)$ and $\sigma_i(t)$, equation (17) can be written as

$$4\left(\sqrt{\frac{\log \delta^{-1}}{2N\hat{\tau}_{i,j}(t)}} + \frac{Q}{(1-\lambda_2)(1+\hat{\tau}_{i,j}(t))}\right) \geq \Delta_i, \tag{18}$$

i.e.,

$$\begin{aligned}
&\sqrt{\frac{\log \delta^{-1}}{2N\hat{\tau}_{i,j}(t)}} + \frac{Q}{(1-\lambda_2)(1+\hat{\tau}_{i,j}(t))} \geq \frac{\Delta_i}{4},\\
&\frac{\log \delta^{-1}}{2N\hat{\tau}_{i,j}(t)} \geq (\frac{\Delta_i}{4} - \frac{Q}{(1-\lambda_2)(1+\hat{\tau}_{i,j}(t))})^2,\\
&\frac{\log \delta^{-1}}{2N\hat{\tau}_{i,j}(t)} \geq \frac{\Delta_i^2}{16} - \frac{Q\Delta_i}{2(1-\lambda_2)\hat{\tau}_{i,j}(t)},\\
&\hat{\tau}_{i,j}(t) \leq \frac{8\log \delta^{-1}}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i}.
\end{aligned}$$

Define $\tau^\star$ as the maximum satisfying equation (18), then the maximum theoretic sample count $\hat{\tau}_{i,j}$ of agent j pulling arm i is

$$\hat{\tau}_{i,j} \leq \lceil \tau^\star \rceil \leq \tau^\star + 1 \leq \frac{8\log \delta^{-1}}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + 1, \tag{19}$$

with probability at least $1 - 2tNK\delta$.

Step 2: Bound the individual regret of agent j

In the design of Algorithm 1, agents refrain from using the most recent information to ensure consensus on the candidate set. There is a time delay in updating the candidate set because agent j requires a few rounds to ensure that all other agents receive the information. The diameter of the communication graph $\mathcal{G}$ is denoted as D, which means that any agent can obtain the information initially learned by agent j after at most D communication rounds. Therefore, the practical sample count $\tau_{i,j}(t)$ is bounded by:

$$\hat{\tau}_{i,j} \leq \tau_{i,j}(T) \leq \hat{\tau}_{i,j} + D. \tag{20}$$

Therefore, the cumulative regret for agent j could be decomposed as follows

$$\begin{aligned}
\mathbb{E}[R_j^T(\mathcal{A})] &= T\mu_{i^\star} - \sum_{t=1}^{T} \mathbb{E}[X_{A_j(t)}(t)] = T\mu_{i^\star} - \sum_{i:\Delta_i>0} \mu_i \tau_{i,j}(T)\\
&= \sum_{i:\Delta_i>0} \Delta_i \tau_i(t) \leq \sum_{i:\Delta_i>0} \Delta_i(\hat{\tau}_{i,j} + D).
\end{aligned} \tag{21}$$

The regret consists of two parts: large probability events (The optimal arm persists in the candidate arm set) and small probability events (The optimal arm is eliminated). The regret caused by large probability events is bounded by the sample

counts, while the regret caused by small probability events is bounded by Lemma 7. According to the equation (21), the total regret is bounded by

$$\begin{aligned}\mathbb{E}[R_j^T(\mathcal{A})] &= \mathbb{E}[R_j^T(\text{large probability events})] + \mathbb{E}[R_j^T(\text{small probability events})] \\ &\leq \sum_{i:\Delta_i>0} \Delta_i(\hat{\tau}_{i,j} + D) + 1 \leq \sum_{i:\Delta_i>0} \frac{16\log T}{N\Delta_i} + K(D+1)\Delta_i + \frac{8KQ}{1-\lambda_2} + 1.\end{aligned}$$

According to (1), the group regret $R^T(\mathcal{A})$ could be rewritten as follows

$$\begin{aligned}\mathbb{E}[R^T(\mathcal{A})] &= \mathbb{E}[R^T(\text{large probability events})] + \mathbb{E}[R^T(\text{small probability events})] \\ &= NT\mu_{i^\star} - \sum_{t=1}^{T}\sum_{j=1}^{N} \mathbb{E}[X_{A_j(t),j}(t)] + 1 \\ &= NT\mu_{i^\star} - \sum_{i:\Delta_i>0}\sum_{j=1}^{N} \mu_i \tau_{i,j}(T) + 1 \\ &\overset{(a)}{\leq} \sum_{i:\Delta_i>0}\sum_{j=1}^{N} (\mu_{i^\star,j} - \mu_{i,j})(\hat{\tau}_{i,j} + D) + 1 \\ &\overset{(b)}{=} \sum_{i:\Delta_i>0} N(\mu_{i^\star} - \mu_i)(\hat{\tau}_{i,j} + D) + 1 \\ &= \sum_{i:\Delta_i>0} N\Delta_i(\hat{\tau}_{i,j} + D) + 1 \\ &\overset{(c)}{\leq} \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i} + \sum_{i:\Delta_i>0} NKD\Delta_i + \frac{8KNQ}{1-\lambda_2} + 1,\end{aligned}$$

where inequality (a) arises due to the time delay discussed earlier, while equality (b) holds because each agent pulls each arm at the same time, i.e., $\hat{\tau}_{i,1} = \hat{\tau}_{i,2} = \cdots = \hat{\tau}_{i,N}$. Inequality (c) follows primarily from equation (19). □

D.5 PROOF OF COROLLARY 1

Proof. According to Lemma 6, we have

$$\lambda_2 \leq 1 - \frac{\beta}{ND},$$

where $\beta \in (0, 1/\epsilon]$ is a given parameter corresponding to graph $\mathcal{G}$ and ϵ represents the largest neighbor number of any agents, which is bounded by $\epsilon = \max_i(\sum_{j'\neq j} a_{j,j'}) \leq N$.

In this paper, define β as $\beta = 1/\epsilon \geq \frac{1}{N}$. Then, one can deduce that λ_2 is bounded by $\lambda_2 \leq 1 - \frac{1}{N^2D}$ and the upper bound of $\frac{1}{1-\lambda_2}$ is

$$\frac{1}{1-\lambda_2} \leq N^2 D.$$

Then, the individual regret is bounded by

$$\mathbb{E}[R_j^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{N\Delta_i} + \sum_{i:\Delta_i>0} (D+1)\Delta_i + 8KQDN^2 + 1.$$

The group regret is bounded by

$$\mathbb{E}[R^T(\mathcal{A})] \leq \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i} + \sum_{i:\Delta_i>0} N(D+1)\Delta_i + 8KQDN^3 + 1.$$

□

D.6 PROOF OF COROLLARY 2

The individual regret R_j^T could be decomposed into

$$\begin{aligned}
\mathbb{E}[R_j^T(\mathcal{A})] &= \mathbb{E}[R_j^T(\text{large probability events})] + \mathbb{E}[R_j^T(\text{small probability events})] \\
&\leq \sum_{i:\Delta_i>0} \Delta_i \tau_{i,j}(T) + 1 \\
&= \sum_{i:\Delta_i\geq\Delta} \Delta_i \tau_{i,j}(T) + \sum_{i:\Delta_i<\Delta} \Delta_i \tau_{i,j}(T) + 1 \\
&\leq \sum_{i:\Delta_i\geq\Delta} \Delta_i \tau_{i,j}(T) + \Delta T + 1 \\
&\leq \sum_{i:\Delta_i\geq\Delta} \frac{16\log T}{N\Delta_i} + \sum_{i:\Delta_i\geq\Delta} (D+1)\Delta_i + \frac{8QK}{1-\lambda_2} + \Delta T + 1 \\
&\leq \frac{16K\log T}{N\Delta} + \Delta T + K(D+1) + \frac{8QK}{1-\lambda_2} + 1 \\
&\leq 2\sqrt{\frac{16KT\log T}{N}} + K(D+1) + \frac{8QK}{1-\lambda_2} + 1,
\end{aligned}$$

where $\Delta = \sqrt{\frac{16K\log T}{NT}}$. The group regret could also be transformed into

$$\begin{aligned}
\mathbb{E}[R^T(\mathcal{A})] &= \mathbb{E}[R^T(\text{large probability events})] + \mathbb{E}[R^T(\text{small probability events})] \\
&\leq \sum_{i=1}^{K}\sum_{j=1}^{N} (\mu_{i^\star,j} - \mu_{i,j})\tau_{i,j}(T) + 1 \\
&\overset{(a)}{=} \sum_{i:\Delta_i>0} N\Delta_i \tau_{i,j}(T) + 1 \\
&\leq \sum_{i:\Delta_i\geq\Delta} N\Delta_i \tau_{i,j}(T) + \sum_{i:\Delta_i<\Delta} N\Delta_i \tau_{i,j}(T) + 1 \\
&\leq \sum_{i:\Delta_i\geq\Delta} N\Delta_i \tau_{i,j}(T) + NT\Delta + 1 \\
&\leq \sum_{i:\Delta_i\geq\Delta} N\Delta_i (\frac{16\log T}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + 1 + D) + NT\Delta + 1 \\
&\leq \frac{16K\log T}{\Delta} + NT\Delta + \frac{8KNQ}{1-\lambda_2} + KN(D+1) + 1 \\
&\leq 8\sqrt{KNT\log T} + NT\Delta + \frac{8KNQ}{1-\lambda_2} + KN(D+1) + 1,
\end{aligned}$$

where equation (a) holds because all agents sample arms synchronously.

D.7 PROOF OF COROLLARY 2

In the proof of Theorem 1, one can deduce that the suboptimal arm i is sampled by agent j at most $\frac{8\log\delta^{-1}}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + D + 1$ times. In Theorem 1, the violation probability is denoted by $\delta = \frac{1}{T^2}$, then the sample count is bounded by

$$\tau_{i,j} \leq \frac{16\log T}{N\Delta_i^2} + \frac{8Q}{(1-\lambda_2)\Delta_i} + D + 1.$$

In each round, `DRRB-bandit` collects information about all arms and communicates it with other agents in a single batch. Therefore, to determine the maximum number of communications, it suffices to consider the number of samples of the arm

that remain in the candidate set for the second longest period.

$$C^T(\mathcal{A}) \leq \frac{16K\log T}{\Delta_{\min}^2} + \frac{8KNQ}{(1-\lambda_2)\Delta_{\min}} + KN(D+1).$$

D.8 PROOF OF THEOREM 3

Proof. Assume that $\sum_{j'\neq j}\mu_{i,j'}$ are the same for all arm $i \in \mathcal{K}$; under this assumption, the problem reduces to a single-agent regret minimization problem, where only agent j's observations matter. This assumption can be generalized to various cases, which could result in the same lower bound of the regret. That is, agent j only needs to perform regret minimization according to its own local observation. Therefore, the problem inherits the regret of classical multi-armed bandits.

According to the assumption above, define two reward distributions on arm i as follows

$$\nu_j = (P_{1,j}, \ldots, P_{i,j}, \ldots, P_{K,j}),$$
$$\nu_j' = (P_{1,j}', \ldots, P_{i,j}', \ldots, P_{N,j}'),$$

where $P_{k,j} = P_{k,j}'$ for all $k \neq i$.

Let $\mathcal{M}$ be a set of distributions with finite means, and let $\mu : \mathcal{M} \to \mathbb{R}$ be the function that maps $P_{i,j} \in \mathcal{M}$ to its mean. Let $\mu_{i^\star,j} \in \mathbb{R}$ and $P_{i,j} \in \mathcal{M}$ have $\mu(P_{i,j}) < \mu_{i^\star,j}$ and define

$$d_{i,j} = d_{\inf}(P_{i,j}, \mu_{i^\star,j}, \mathcal{M}) = \inf_{P_{i,j}'\in\mathcal{M}}\{D(P_{i,j}, P_{i,j}') : \mu(P_{i,j}') > \mu_{i^\star,j}\},$$

where $D(P_{i,j}, P_{i,j}')$ is the relative entropy between $P_{i,j}$ and $P_{i,j}'$. For arm i and $\mu(P_{i,j}') > \mu_{i^\star,j}$, there exists arbitrary $\epsilon > 0$ such that $D(P_{i,j}, P_{i,j}') \leq d_{i,j} + \epsilon$.

According to Lemma 15.1 in reference Lattimore and Szepesvári [2020], the divergence between ν_j and ν_j' is decomposed into

$$D(\mathbb{P}_{\nu_j}, \mathbb{P}_{\nu_j'}) = \sum_{k=1}^{K}\mathbb{E}[\tau_{k,j}(T)]D(P_{k,j}, P_{k,j}') \overset{(a)}{=} \mathbb{E}[\tau_{i,j}(T)]D(P_{i,j}, P_{i,j}') \leq \mathbb{E}[\tau_{i,j}(T)](d_{i,j}+\epsilon),$$

where equation (a) is obtained based on $D(P_{k,j}, P_{k,j}') = 0$ if $k \neq i$.

According to Bretagnolle–Huber inequality (Theorem 14.2 in Lattimore and Szepesvári [2020]), for any event $A_{i,j}$ (agent j pulls arm i), we have

$$\begin{aligned}\mathbb{P}_{\nu_j}(A_{i,j}) + \mathbb{P}_{\nu_j'}(A_{i,j}^c) &\geq \frac{1}{2}\exp(-D(\mathbb{P}_{\nu_j}, \mathbb{P}_{\nu_j'})) \\ &\geq \frac{1}{2}\exp(-\mathbb{E}_{\nu_j}[\tau_{i,j}(T)](d_{i,j}+\epsilon))\end{aligned}$$

Choose $A_{i,j} = \{\tau_{i,j}(T) > T/2\}$, and let $R_j^T = R_j^T(\mathcal{A}, \nu_j)$ and $R_{j'}^T = R_{j'}^T(\mathcal{A}, \nu_j')$. Then

$$\begin{aligned}R_j^T + R_{j'}^T &\geq \frac{T}{2}(\mathbb{P}_{\nu_j}(A_{i,j})\Delta_i + \mathbb{P}_{\nu_j'}(A_{i,j}^c)(\mu_i' - \mu_{i^\star})) \\ &\geq \frac{T}{2}\min\{\Delta_i, \mu_i' - \mu_{i^\star}\}(\mathbb{P}_{\nu_j}(A_{i,j}) + \mathbb{P}_{\nu_j'}(A_{i,j}^{\mathbf{c}})) \\ &\geq \frac{T}{2}\min\{\Delta_i, \mu_i' - \mu_{i^\star}\}\exp(-\mathbb{E}_{\nu_j}[\tau_{i,j}(T)](d_{i,j}+\epsilon)).\end{aligned}$$

Rearranging and taking the limit inferior leads to

$$\begin{aligned}\liminf_{T\to\infty}\frac{\mathbb{E}[\tau_{i,j}(T)]}{\log T} &\geq \frac{1}{d_{i,j}+\epsilon}\liminf_{T\to\infty}\frac{\log\frac{T\min\{\Delta_i, \mu_i'-\mu_{i^\star}\}}{4(R_j^T+R_{j'}^T)}}{\log T} \\ &\geq \frac{1}{d_{i,j}+\epsilon}(1 - \liminf_{T\to\infty}\frac{\log(R_j^T + R_{j'}^T)}{\log T}) \\ &= \frac{1}{d_{i,j}+\epsilon},\end{aligned}$$

where the last equality follows from the definition of consistency, which says that for any $p > 0$, there exists a constant C_p such that for sufficiently large T, $R_j^T + R_{j'}^T \leq C_p T^p$, which implies that

$$\liminf_{T\to\infty} \frac{\log(R_j^T + R_{j'}^T)}{\log T} \leq p.$$

Considering $p > 0$ was arbitrary and $\epsilon > 0$ is limited to zero, we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\tau_{i,j}(T)]}{\log T} \geq \frac{1}{d_{i,j}} \overset{(a)}{=} \frac{2}{N^2\Delta_i^2}, \tag{22}$$

where equation (a) is obtained from Table 16.1 given in Lattimore and Szepesvári [2020]. We have $d_{i,j} = \frac{(\mu_{i,j} - \mu_{i^\star,j})^2}{2}$. Considering that $\sum_{j'\neq j} \mu_{i,j'}$ and $\mu_i = \frac{1}{N}\sum_{j=1}^N \mu_{i,j}$, we have $\mu_{i^\star,j} - \mu_{i,j} = N(\mu_{i^\star} - \mu_i) = N\Delta_i$.

The individual regret of the problem is lower bounded by

$$\liminf_{T\to\infty} \frac{R_j^T(\mathcal{A})}{\log T} \geq \liminf_{T\to\infty} \sum_{i:\Delta_i>0} \frac{\Delta_i \mathbb{E}[\tau_{i,j}(T)]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_{i,j}} \geq \sum_{i:\Delta_i>0} \frac{2}{N^2\Delta_i}.$$

For the group regret, we consider it as the sum of all agents' individual regret, and the lower bound of group regret could be written as

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R^T(\mathcal{A})]}{\log T} = \liminf_{T\to\infty} \frac{\mathbb{E}[\sum_{j=1}^N R_j^T(\mathcal{A})]}{\log T} \geq \sum_{i:\Delta_i>0} \sum_{j=1}^N \frac{2}{N^2\Delta_i} \geq \sum_{i:\Delta_i>0} \frac{2}{N\Delta_i}.$$

□

D.9 PROOF OF THEOREM 4

Proof. In round-robin-based bandit algorithms, assume all agents sample the same arm at each time slot. Then, the global reward of each arm is associated with a $\frac{1}{\sqrt{N}}$-Gaussian distribution, which follows from Lemma 3.

Let $\mathcal{M}$ be a set of distributions with finite means, and let $\mu : \mathcal{M} \to \mathbb{R}$ be the function that maps $P \in \mathcal{M}$ to its mean. Let $\mu_{i^\star} \in \mathbb{R}$ and $P \in \mathcal{M}$ have $\mu(P) < \mu_{i^\star}$ and define

$$d_i = d_{\inf}(P, \mu_{i^\star}, \mathcal{M}) = \inf_{P'\in\mathcal{M}} \{D(P, P') : \mu(P') > \mu_{i^\star}\},$$

where $D(P, P')$ is the relative entropy between P and P'.

Define two reward distributions as follows

$$\nu = (P_1, \ldots, P_i, \ldots, P_K),$$
$$\nu' = (P_1, \ldots, P_i', \ldots, P_K).$$

Let all arms except arm i be the same in the two distributions. For arm i, let $\epsilon > 0$ be arbitrary such that $D(P_i, P_i') \leq d_i + \epsilon$ and $\mu(P_i') > \mu_{i^\star}$.

According to Lemma 15.1 in reference Lattimore and Szepesvári [2020], the divergence between ν and ν' is decomposed into

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{k=1}^K \mathbb{E}[\tau_{k,j}(T)] D(P_i, P_i') \overset{(a)}{=} \mathbb{E}[\tau_{i,j}(T)](d_i + \epsilon),$$

where equation (a) is obtained based on $D(P_j, P_j') = 0$ if $j \neq i$.

According to Bretagnolle–Huber inequality (Theorem 14.2 in Lattimore and Szepesvári [2020]), for any event A, we have

$$\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(A^c) \geq \frac{1}{2}\exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'})) \geq \frac{1}{2}\exp(-\mathbb{E}[\tau_{i,j}(T)](d_i + \epsilon))$$

Choose $A = \{\tau_{i,j}(T) > T/2\}$, and let $R_T = R_T(\mathcal{A}, \nu)$ and $R'_T = R'_T(\mathcal{A}, \nu')$. Then

$$\begin{aligned} R_T + R'_T &\geq \frac{T}{2}(\mathbb{P}_\mu(A)\Delta_i + \mathbb{P}_{\mu'}(A^c)(\mu'_i - \mu_{i^\star})) \\ &\geq \frac{T}{2}\min\{\Delta_i, \mu'_i - \mu_{i^\star}\}(\mathbb{P}_\mu(A)\Delta_i + \mathbb{P}_{\mu'}(A^c)) \\ &\geq \frac{T}{2}\min\{\Delta_i, \mu'_i - \mu_{i^\star}\}\exp(-\mathbb{E}[\tau_{i,j}(T)](d_i + \epsilon)). \end{aligned}$$

Rearranging and taking the limit inferior leads to

$$\begin{aligned} \liminf_{T\to\infty} \frac{\mathbb{E}[\tau_{i,j}(T)]}{\log T} &\geq \frac{1}{d_i + \epsilon}\liminf_{T\to\infty} \frac{\log \frac{T\min\{\Delta_i, \mu'_i - \mu_{i^\star}\}}{4(R_T + R'_T)}}{\log T} \\ &\geq \frac{1}{d_i + \epsilon}(1 - \liminf_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T}) \\ &= \frac{1}{d_i + \epsilon}, \end{aligned}$$

where the last equality follows from the definition of consistency, which says that for any $p > 0$, there exists a constant C_p such that for sufficiently large T, $R_T + R'_T \leq C_p T^p$, which implies that

$$\liminf_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} \leq p.$$

Considering $p > 0$ was arbitrary and $\epsilon > 0$ is limited to zero, we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\tau_{i,j}(T)]}{\log T} \geq \frac{1}{d_i}.$$

According to Table 16.1 given in Lattimore and Szepesvári [2020], we have $d_i = \frac{N\Delta_i^2}{2}$. The individual regret of the problem is lower bounded by

$$\liminf_{T\to\infty} \frac{R_j^T(\mathcal{A})}{\log T} \geq \liminf_{T\to\infty} \sum_{i:\Delta_i>0} \frac{\Delta_i \mathbb{E}[\tau_{i,j}(T)]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_i} \geq \sum_{i:\Delta_i>0} \frac{2}{N\Delta_i}.$$

According to the definition of group regret $R^T(\mathcal{A})$ in equation (1), the lower bound of group regret could be written as

$$\begin{aligned} R^T(\mathcal{A}) &= NT\mu_{i^\star} - \sum_{i:\Delta_i>0}\sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{I}\{A_j(t) = i\}\mathbb{E}[X_{i,j}(t)] \\ &= NT\mu_{i^\star} - \sum_{i:\Delta_i>0}\sum_{t=1}^{T}\sum_{j=1}^{N}\mathbb{I}\{A_j(t) = i\}\mu_{i,j} \\ &= NT\mu_{i^\star} - \sum_{i:\Delta_i>0}\sum_{j=1}^{N}\mu_{i,j}\tau_{i,j}(T), \\ &= \sum_{i:\Delta_i>0}\sum_{j=1}^{N}(\mu_{i^\star,j} - \mu_{i,j})\tau_{i,j}(T) \\ &= \sum_{i:\Delta_i>0} N\Delta_i\tau_{i,j}(T). \end{aligned}$$

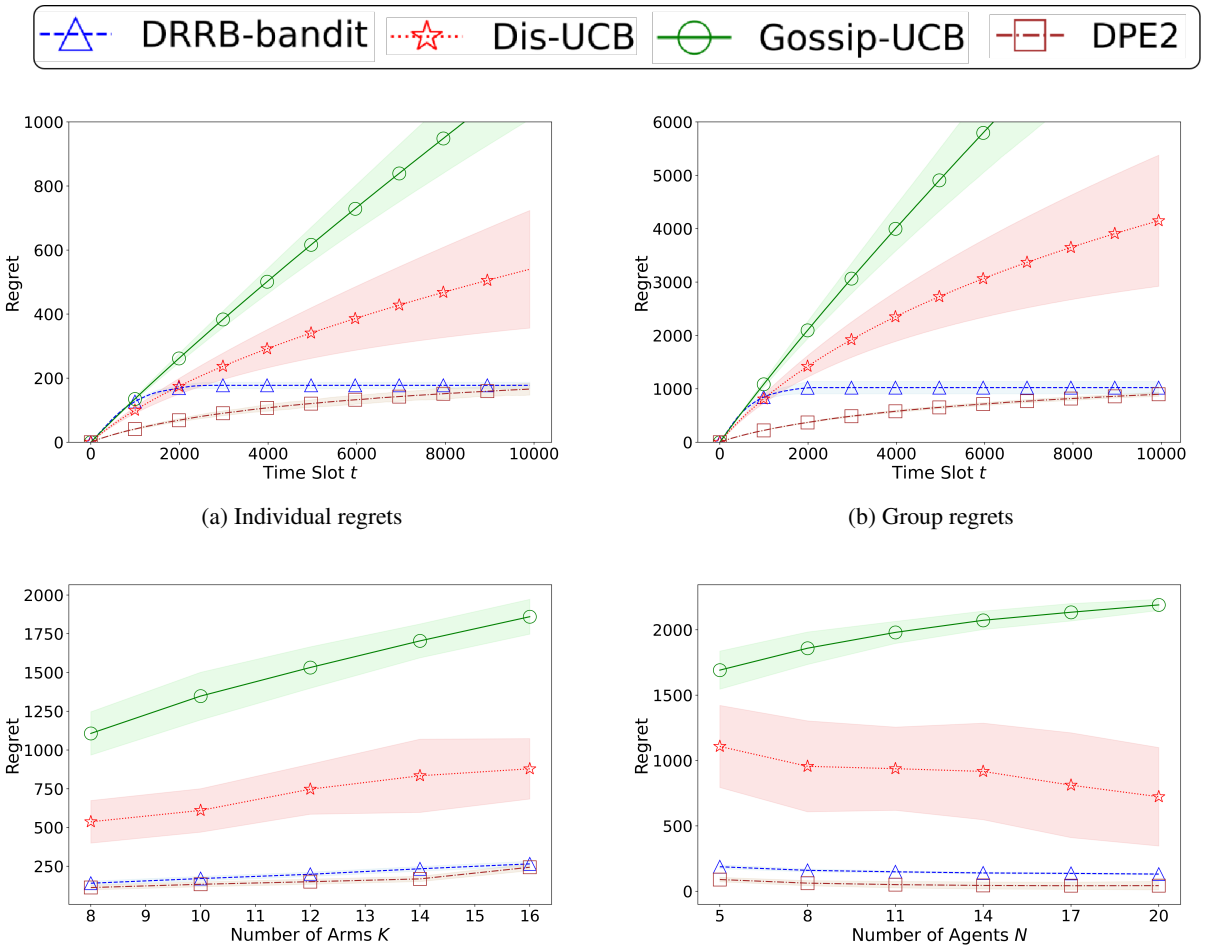
Considering equation (22), we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R^T(\mathcal{A})]}{\log T} \geq \liminf_{T\to\infty} \sum_{i:\Delta_i>0}\sum_{j=1}^{N} \frac{\Delta_{i,j}\mathbb{E}[\tau_{i,j}(T)]}{\log T} \geq \sum_{i:\Delta_i>0}\sum_{j=1}^{N} \frac{2N\Delta_i}{N\Delta_i^2} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}.$$

□

E ADDITIONAL EXPERIMENT

Our algorithm (`DRRB-bandit`) can also obtain a similar result compared to the optimal homogeneous bandit algorithm (`DPE2` [Wang et al., 2020]) in Figure 2. `DPE2` contains a leader that could uniformly allocate resources and tasks. `DRRB-bandit` relies entirely on fully distributed communication. Hence, `DRRB-bandit` has more regret compared with `DPE2` but is also better than other heterogeneous bandit algorithms.



(a) Individual regrets

(b) Group regrets

(c) Regrets with different numbers of arms

(d) Regrets with different numbers of agents

Figure 2: Performance comparison in the homogeneous setting.