Distributed Multi-agent Bandits Over Erdős-Rényi Random Networks

Jingyuan Liu   Hao Qiu   Lin Yang   Mengfan Xu

Abstract

We study the distributed multi-agent multi-armed bandit problem with heterogeneous rewards over random communication graphs. Uniquely, at each time step t agents communicate over a time-varying random graph $\mathcal{G}_t$ generated by applying the Erdős–Rényi model to a fixed connected base graph $\mathcal{G}$ (for classical Erdos-Rényi graphs, $\mathcal{G}$ is a complete graph), where each potential edge in $\mathcal{G}$ is randomly and independently present with probability p. Notably, the resulting random graph is not necessarily connected at each time step. Additionally, each agent's arm rewards follow time-invariant distributions, and the reward distribution for the same arm may differ across agents. The goal is to minimize the cumulative expected regret relative to the global mean reward of each arm, defined as the average of that arm's mean rewards across all agents. To this end, we propose a fully distributed algorithm that integrates the arm elimination strategy with the random gossip algorithm. We theoretically show that the regret upper bound is of order $\log T$ and is highly interpretable, where T is the time horizon. It includes the optimal centralized regret $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k}\right)$ and an additional term $\mathcal{O}\left(\frac{N^2 \log T}{p\lambda_{N-1}(\mathrm{Lap}(\mathcal{G}))} + \frac{KN^2 \log T}{p}\right)$ where N and K denotes the total number of agents and arms, respectively. This term reflects the impact of $\mathcal{G}$'s algebraic connectivity $\lambda_{N-1}(\mathrm{Lap}(\mathcal{G}))$ and the link probability p, and thus highlights a fundamental trade-off between communication efficiency and regret. As a by-product, we show a nearly optimal regret lower bound. Finally, our numerical experiments not only show the superiority of our algorithm over existing benchmarks, but also validate the theoretical regret scaling with problem complexity.

1 Introduction

Multi-armed Bandit (MAB) is a widely studied framework for sequential decision-making under uncertainty [Auer et al., 2002]. In this setting, an agent selects an arm from multiple options in each round, observes the reward from the chosen arm, and aims to maximize the cumulative expected reward. The emergence of large-scale cooperative systems holds true in various applications ranging from sensor networks [Ganesan et al., 2004, Zhu et al., 2016] to federated learning [Ye et al., 2023, McMahan et al., 2017] and edge computing [Wang et al., 2022a, Ghoorchian and Maghsudi, 2020]. It has naturally motivated interest in distributed multi-agent multi-armed bandit (MA-MAB) problem, where multiple agents collaboratively learn to optimize rewards. MA-MAB settings are typically categorized as homogeneous [Landgren et al., 2016a, Martínez-Rubio et al., 2019] or heterogeneous [Zhu et al., 2021, Xu and Klabjan, 2023], depending on whether the reward distributions for the same arm are identical across agents. The heterogeneous setting, in which reward distributions vary across agents, is significantly more general but more challenging, and thus has attracted growing attention. It introduces substantial difficulties, as agents must make sequential decisions under uncertainty while relying on limited information about both their own rewards and the actions or observations of other agents.

Another challenging aspect of MA-MAB lies in the underlying communication protocol, which constrains how agents share information with one another. Decentralized MA-MAB settings are more realistic than centralized ones—where all agents can communicate with any other agent—as they restrict communication to immediate neighbors defined by a graph structure. Much of the existing work has focused on time-invariant graphs [Zhu et al., 2021], where the communication graph remains fixed throughout. However, the complexity of many real-world decentralized systems, such as wireless ad-hoc networks, necessitates the use of time-varying graphs [Zhu and Liu, 2023], particularly random graphs [Xu and Klabjan, 2023]. This added complexity significantly complicates both the communication and learning dynamics. Notably, [Xu and Klabjan, 2023] is the first to consider classical Erdős–Rényi (E-R) random graphs in the MA-MAB setting, where any two agents can communicate with probability p at each time step. However, it is possible that some pairs of agents can never communicate directly due to inherent topological constraints. This scenario has been formulated as a more general version of the E-R graph, where two agents can communicate with probability p only if there is an edge between them in a base graph. Note that when the base graph is a complete graph, it is equivalent to classical (E-R) random graphs, implying consistency. To date, this setting remains unexplored, which motivates our work.

To date, a line of work has studied regret bounds under various graph structures, where connectivity or sequential connectivity is typically required. For example, in the context of time-invariant graphs, Martínez-Rubio et al. [2019] and Zhu et al. [2021] analyze distributed bandits over connected graphs and derive $\log T$ regret bounds. For time-varying graphs, Zhu et al. [2025] obtains $\log T$ regret bounds under the assumption of B-connectivity, where the union of any l consecutive graphs must be connected. In the classical Erdős–Rényi model [Erdos et al., 1960] over fully connected agent communication, Xu and Klabjan [2023] derive a regret of order $\log T$, but only under the assumption that $p \geq 1/2 + 1/2\sqrt{1 - (\epsilon/NT)^{2/N-1}}$ which is larger than $1/2$ and can even approach 1 when the number of agents N or the time horizon T is large. This is a strong assumption, as it may not hold in many real-world settings, but is required in their analysis to ensure that the graph is connected with high probability. Relaxing this connectivity requirement to allow arbitrary p presents a significant challenge. Moreover, their regret bound does not reflect how the link probability p impacts the regret. Incorporating p into the regret expression would significantly improve our understanding of how to choose p in practice—a gap that remains open. In this work, we address both of these gaps. To this end, we address the following key research question: *Can we solve MA-MAB under new Erdős–Rényi random networks and heterogeneous rewards, and derive regret bounds that captures graph complexity under much milder assumptions?*

Contribution. We provide an affirmative answer to the above question through the following contributions. Methodologically, we solve the MA-MAB problem over general Erdős–Rényi communication networks using a gossip algorithm, which is widely adopted in distributed settings. Moreover, we adopt an algorithm based on arm elimination with a minimal number of arm pulls required for each arm to guarantee sufficient information collecting for each agent, which addresses the E-R communication networks.

Analytically, we study the regret of MA-MAB under our proposed algorithm over general Erdős–Rényi communication networks for any $p \in (0, 1]$, leading to novel contributions. We are the first to 1) explore general Erdos-Rényi graphs induced by any fixed connected base graphs beyond classical setting that assumes a complete base graphs; 2) obtain the a tighter and more interpretable regret bound, which generalizes the bound from the homogeneous fixed-graph setting studied in Martínez-Rubio et al. [2019] to our more challenging heterogeneous setting, as shown in Section 5; 3) relax the assumption of a sufficiently large p used in Xu and Klabjan [2023] and reduce the order of N in the upper bound in their analysis of classical Erdős–Rényi models with complete base graphs. Precisely, we obtain a regret upper bound of order of $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k} + \frac{N^2 \log T}{p\lambda_{N-1}(\text{Lap}(\mathcal{G}))} + \frac{KN^2 \log T}{p}\right)$ where the first term accounts for an optimal centralized regret, aligning with the lower bound established in Section 5.2, and the last two terms capture the effects of both the link probability p and the algebraic connectivity $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ of the base graph. Moreover, we uniquely characterize how the regret upper bounds are influenced by p and $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$, which highlights a tradeoff between the communication cost (i.e., the number of communication rounds) and regret performance.

Numerically, we implement our proposed algorithm and conduct experiments to validate our theoretical results with multiple random graph settings and edge generation probability p. We also compare our methods with existing approaches to demonstrate their effectiveness.

2 Related Works

Distributed online algorithm. Our framework builds on the classical line of work on gossip algorithms [Xiao and Boyd, 2004, Boyd et al., 2006]. Specifically, when an agent has access only to local information and communicates solely with its immediate neighbors, it adopts a gossip algorithm to aggregate information from agents beyond its local neighborhood, based on a weight matrix. Notably, gossip algorithms are widely used in distributed optimization. For example, Duchi et al. [2011], Nedic and Ozdaglar [2009] address distributed convex optimization problems using gossip algorithms. Subsequently, Hosseini et al. [2013], Yan et al. [2013] extend the gossip approach to the online setting and achieve a regret of order $\sqrt{T}$, assuming convex loss functions. Later, Mateos-Núñez and Cortés [2014] consider distributed online optimization over B-connected and design a distributed online primal-dual algorithm coupled with a gossip protocol, also achieving a $\sqrt{T}$ regret bound. More recently, Lei et al. [2020] study the same problem over random communication networks (Erdős–Rényi networks) and obtain the same regret order. They further characterize how regret is affected by the link probability p in the Erdős–Rényi model and the algebraic connectivity of the base graph $\mathcal{G}$. We consider the same graph topology but focus on multi-agent multi-armed bandits, which differ significantly from online convex optimization and introduce the additional challenge of learning the dynamics of bandits. For a more comprehensive survey on distributed online optimization, we recommend the reader to Li et al. [2023] and Yuan et al. [2024].

Distributed multi-agent multi-armed bandit. Along the line of work on MA-MAB, several studies [Landgren et al., 2016a,b, Zhu et al., 2020, Chawla et al., 2020, Wang et al., 2022b, 2020, Zhu et al., 2025, Martínez-Rubio et al., 2019, Agarwal et al., 2022, Sankararaman et al., 2019, Zhu et al., 2021, Zhu and Liu, 2023, Xu and Klabjan, 2023, Yi and Vojnovic, 2023] have investigated both homogeneous and heterogeneous settings. In homogeneous settings, numerous work incorporate gossip algorithms to reduce regret in terms of the number of agents; information sharing among agents accelerates the concentration of reward observations. For example, Landgren et al. [2016a,b] first formulate this problem and solve it using gossip algorithms. Martínez-Rubio et al. [2019] achieve the optimal centralized regret—independent of the number of agents and matching that of the single-agent bandit—plus an additional term that depends on the spectral gap of the communication matrix. [Chawla et al., 2020] characterizes the regret-communication trade-off, considers circular ring graphs, and establishes regret improvement. Wang et al. [2022b, 2020] further focus on optimizing communication efficiency to minimize the number of communication rounds while guaranteeing regret performance. In contrast, we consider more challenging heterogeneous settings and also characterize the regret-communication trade-off. Here gossiping enables regret reduction in terms of the order of T; without information from other agents, the regret can easily be linear in T Xu and Klabjan [2025]. In this direction, Zhu et al. [2021] is the first to study heterogeneous rewards over a time-invariant connected graph and establishes regret bounds of order $\log T$. [Zhu and Liu, 2023] extends this to B-connected graphs, also achieving regret bounds of order $\log T$. Recently, Xu and Klabjan [2023] propose a gossip-based algorithm for the classical E-R model and obtain regret bounds of order $\log T$ when p is sufficiently larger than $1/2$ to ensure the graph is connected with high probability. More generally, Yi and Vojnovic [2023] consider MA-MAB with heterogeneous rewards in the adversarial environment and establishes a regret bound of order $T^{2/3}$. In contrast, we consider the stochastic setting with general Erdős–Rényi random networks, without any assumption on p.

3 Setting and Notations

In this section, we formally define the problem of interest, starting by the presentation of general notations. A full notation chart can be found in Section E.

General Notations. For a matrix $M \in \mathbb{R}^{p \times q}$, let $[M]_{i,j}$ denote the entry in the i-th row and j-th column. Given a doubly stochastic matrix $P \in \mathbb{R}^{d \times d}$, we denote by $\lambda_2(P)$ its second largest eigenvalue. Let $e_i \in \mathbb{R}^d$ be the i-th standard basis vector, $\mathbf{1} \in \mathbb{R}^d$ the all-ones vector, and I_d the $d \times d$ identity matrix. We use $\mathbb{I}\{\cdot\}$ to denote the indicator function, which equals 1 if the condition inside holds, and 0 otherwise. Moreover, we use $[n] = \{1, 2, \cdots, n\}$ to denote a set of indices.

Multi-agent Multi-armed Bandit. We consider a multi-agent multi-armed bandit (MA-MAB) setting involving N agents. The bandit problem is run over a time horizon of T. In each round $t \in [T]$, every agent $i \in [N]$ selects an arm $A_i(t) \in [K]$. The local reward of arm k for agent i follows an unknown, time-invariant probability distribution $\mathbb{P}_{i,k}$ supported on $[0,1]$, with mean $\mu_{i,k} = \mathbb{E}[X \sim \mathbb{P}_{i,k}] \in [0,1]$. When agent i selects arm $A_i(t)$ at round t, it only observes a local stochastic reward $X_{i,A_i(t)}(t)$ drawn independently from the distribution $\mathbb{P}_{i,A_i(t)}$ each round. However, the agent's true objective depends on the global reward, defined as $X_{A_i(t)}(t) := \frac{1}{N} \sum_{j \in [N]} X_{j,A_i(t)}(t)$. Accordingly, we define the global mean reward for arm k as $\mu_k := \frac{1}{N} \sum_{j \in [N]} \mu_{j,k}$, where $\mu_{j,k}$ is the mean of $\mathbb{P}_{j,k}$. We denote $T_{i,k}(t)$ to be the total number of times agent i has selected arm k up to time t by $T_{i,k}(t) = \sum_{s=1}^{t} \mathbb{I}(A_i(s) = k)$.

Throughout, we focus on a decentralized setting where N agents are distributed on undirected time-varying graphs and communicate via the graphs. Specifically, we consider a communication protocol based on **Erdős–Rényi random graphs**. More precisely, agents communicate via a time-varying graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, where the vertex set $\mathcal{V} = [N]$ is fixed, but the edge set $\mathcal{E}_t$ may vary at each round t. Uniquely, each **communication graph** $\mathcal{G}_t$ is generated based on an underlying undirected, fixed, and connected (but not necessarily complete) **base graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that defines all feasible communication edges. The connectedness of the base graph is essential to avoid linear regret as proved in Theorem 4 in Xu and Klabjan [2025]. In every round t, each edge in graph $\mathcal{G}$ is independently in $\mathcal{G}_t$ with probability $p \in (0,1]$. Two agents can communicate if and only if there is an edge between them in $\mathcal{G}_t$—namely, an active edge. Formally, the random graph generation reads as follows.

Assumption 3.1 (Erdős–Rényi Random Graph). In each round t, the random communication graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, generated from the base graph $\mathcal{G}$ meets:

$$\mathbb{P}\left((i,j) \in \mathcal{E}_t\right) = \begin{cases} p, & \text{if } (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise}, \end{cases}$$

for all vertices $i, j \in \mathcal{V}$.

Thus, $\mathcal{E}_t \subseteq \mathcal{E}$ for all t. Let $\mathcal{N}_i = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$ denote the neighbors of agent i in the base graph $\mathcal{G}$, and $\mathcal{N}_i(t) = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}_t\}$ denote the active neighbors of agent i in round t, clearly satisfying $\mathcal{N}_i(t) \subseteq \mathcal{N}_i$.

The objective is to design a distributed algorithm π that minimizes the total global regret over T rounds. Let $\mu^* = \max_{k \in [K]} \mu_k$ be the optimal global average reward. The total regret for algorithm π is defined as

$$\text{Reg}_T(\pi) = NT\mu^* - \sum_{i \in [N]} \sum_{t=1}^{T} \mu_{A_i(t)} = \sum_{i \in [N]} \sum_{t=1}^{T} \Delta_{A_i(t)} \tag{1}$$

where $\Delta_k = \mu^* - \mu_k$ is the global suboptimality gap for arm k.

4 Algorithm: `Gossip Successive Elimination`

In this section, we present the proposed methodology. Unlike prior work on gossip bandits, which assumes fixed graphs or graphs that become connected within a few rounds (with high probability), our setting faces the core challenges of randomness and disconnection in the communication graph. To address this, our algorithm guarantees an upper bound on the number of rounds needed for any agent's information to reach its neighbors via gossip, enabling accurate estimate construction. It also uses a round-robin strategy combined with an arm elimination strategy [Even-Dar et al., 2006] to ensure that agents maintain consensus on arm pull counts. More specifically, we present the precise steps in Algorithm 1, namely, `Gossip Successive Elimination` (GSE), which integrates the arm elimination strategy with a gossip-based communication protocol.

The agent needs several parameters, including the link probability p and algebraic connectivity $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ as inputs. Initially, agent i's active arm set $\mathcal{S}_i$ is set to the full set of arms $[K]$, while both the local reward estimate $\widehat{\mu}_{i,k}(t)$ and global reward estimate $z_{i,k}(t+1)$ are initialized to zero ($t = 0$). Notably, Algorithm 1 requires knowledge of the link probability p and the algebraic connectivity $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$. Following the standard practice in existing work Martínez-Rubio

et al. [2019], Zhu et al. [2021], Xu and Klabjan [2023], we assume that $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ is known, while allowing the link probability p to remain unknown. In this case, the algorithm can still be implemented, and the corresponding regret upper bound preserved, by adding a short burn-in phase to estimate a value $\hat{p} \in (p/2, p]$. This estimation requires only $\mathcal{O}(\log T)$ time steps. We refer the reader to Appendix D for the details and theoretical analysis of this procedure.

At each round t, agent i observes its active neighbors $\mathcal{N}_i(t)$ in the communication graph $\mathcal{G}_t$, which is randomly generated from the base graph $\mathcal{G}$ according to Theorem 3.1. As part of our key contributions, we consider a weighting matrix W_t for gossip as

$$W_t = I_N - \frac{1}{N}\text{Lap}(\mathcal{G}_t), \tag{2}$$

where $\text{Lap}(\mathcal{G}_t)$ denotes the Laplacian matrix of the communication graph $\mathcal{G}_t$.

Algorithm 1 `Gossip Successive Elimination` for Agent $i \in [N]$

1: **Input:** Algebraic connectivity $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$, total time horizon T, set of arms $[K]$, link probability p
2: **Initialization:** Active set $\mathcal{S}_i \leftarrow [K]$, local reward estimate $\widehat{\mu}_{i,k}(0) \leftarrow 0$, global reward estimate $z_{i,k}(1) \leftarrow 0$ for all $k \in [K]$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4:     Select arm $A_i(t) \in \mathcal{S}_i$ with the minimum pull count $T_{i,k}(t)$ and update $T_{i,k}(t)$
5:     Receive feedback and update statistics for each arm using Equation (3)
6:     Remove arm $k \in \mathcal{S}_i$ if there exists an arm $k' \in \mathcal{S}_i$, $k' \neq k$, satisfying the elimination condition in Equation (6)
7:     Update the active set $\mathcal{S}_i$ according to Equation (7)
8: **end for**

The execution steps of agents run as follow. Agent i selects arm $A_i(t)$ from the active set $\mathcal{S}_i$ with the least number of pulls $T_{i,k}(t)$, observes the local reward of $A_i(t)$, and updates the reward estimates as follows:

$$\widehat{\mu}_{i,k}(t) = \frac{1}{T_{i,k}(t) \vee 1} \sum_{\tau=1}^{t} \mathbb{I}\{A_i(\tau) = k\} X_{i,k}(\tau),$$

$$z_{i,k}(t+1) = \sum_{j \in \mathcal{N}_i(t) \cup \{i\}} [W_t]_{i,j} z_{j,k}(t) + \widehat{\mu}_{i,k}(t) - \widehat{\mu}_{i,k}(t-1), \tag{3}$$

where $X_{i,k}(\tau)$ is the feedback observed by agent i from pulling arm k at round τ. The global estimate $z_{i,k}(t)$ is updated via a gossip protocol: at each round t, agent i aggregates its own and its active neighbors' estimates, weighting each neighbor $j \in \mathcal{N}_i(t) \cup \{i\}$ by the corresponding entry $[W_t]_{i,j}$ of the matrix W_t. We also define the gossip-based upper and lower confidence bounds for arm k, which remain key criteria for arm elimination and updating $\mathcal{S}_i$, as

$$\text{GUCB}_{i,k}(t) = z_{i,k}(t) + c_{i,k}(t), \ \text{GLCB}_{i,k}(t) = z_{i,k}(t) - c_{i,k}(t).$$

Here the confidence bound $c_{i,k}(t)$ reads as

$$c_{i,k}(t) = \sqrt{\frac{4\log(T)}{N \max\{T_{i,k}(t) - KL^*, 1\}}} + \frac{4(\sqrt{N} + \tau^*)}{\max\{T_{i,k}(t) - KL^*, 1\}}, \tag{4}$$

with

$$\tau^* = \left\lceil \frac{2N\log(T)}{p\lambda_{N-1}(\text{Lap}(\mathcal{G}))} \right\rceil, \ L^* = N\left\lceil -\frac{2\log(NT)}{\log(1-p)} \right\rceil. \tag{5}$$

The confidence radius in Equation (4) decomposes into two parts. The first term captures the **estimation error**, arising from the statistical variance of independently sampled rewards for each arm. The second term captures the **consensus error**, which stems from the cumulative approximation error error incurred as agents reach consensus via a gossip-based communication protocol.

Arm elimination occurs if and only if there exists an arm $k' \neq k$ in $\mathcal{S}_i$ that meets the condition

$$\text{GLCB}_{i,k'}(t) \geq \text{GUCB}_{i,k}(t), \tag{6}$$

which means that arm k' has a higher global reward estimate than arm k with high probability. Subsequently, the active set $\mathcal{S}_i$ is updated as

$$\mathcal{S}_i \leftarrow \bigcap_{j \in \mathcal{N}_i(t) \cup \{i\}} \mathcal{S}_j. \qquad (7)$$

This procedure coupled with the arm-elimination strategy ensures that, for every agent, each arm in the active set is pulled a roughly equal number of times, ensuring consensus.

5 Regret Analyses

In this section, we analyze the regret of the proposed algorithm and establish an upper bound on the corresponding regret, demonstrating its theoretical effectiveness. Additionally, we derive a lower bound for our new problem setting, highlighting the problem's inherent complexity and showing that the algorithm is nearly optimal up to some interpretable factors.

5.1 Upper Bound

We start by presenting the regret upper bound for Algorithm 1. To that end, we first introduce several technical lemmas that play a key role in the regret analysis. The proofs can be found in Section C.

Lemma 5.1. *Let us assume that the communication protocol follows Theorem 3.1. Then we have that for Algorithm 1, for any agent $i \in [N]$ and any arm $k \in [K]$, and any $t \in [T]$, the following holds with probability at least $1 - \frac{3NK}{T}$,*

$$|z_{i,k}(t) - \mu_k| \leq c_{i,k}(t),$$

where $c_{i,k}(t)$ is the confidence bound defined in Equation (4), and τ^* and L^* are the parameters introduced in Equation (5).

Importantly, we show that the estimation error compared to the global mean value is upper bounded by the confidence bound $c_{i,k}(t)$. Notably, $c_{i,k}(t)$ is monotonically decreasing in the number of pulls of arm k by agent i, i.e., $T_{i,k}(t)$, when $T_{i,k}(t) > KL^*$. Here, KL^* is the minimal number of pulls required to ensure that agent i has collected sufficient information from all other agents. This also implies that the global estimate $z_{i,k}(t)$ becomes increasingly accurate and approaches the true mean reward μ_k as the number of pulls increases. As a result, based on Theorem 5.1, we derive the following regret bound for individual agents.

Lemma 5.2. *Let $\text{Reg}_{i,T}(\pi) = T\mu^* - \sum_{t=1}^{T} \mu_{a_i(t)}$ denote the regret incurred by agent i using policy π. Under the communication protocol assumed in Theorem 3.1, the regret of agent i under Algorithm 1 is guaranteed to meet*

$$\text{Reg}_{i,T}(\text{GSE}) \leq \sum_{k:\Delta_k > 0} \left(\frac{64 \log(T)}{N \Delta_k} + 16 \left(\sqrt{N} + \tau^* \right) \right) + KL^* + 3KN\Delta_{\max}$$

where $\Delta_{\max} = \max_{k \in [K]} \Delta_k$ denotes the largest reward gap across all arms.

In other words, Theorem 5.2 shows that the regret incurred by any individual agent can be effectively bounded. This result naturally extends to the global regret, as stated in the theorem below.

Theorem 5.3. *The global regret defined in Equation (1) for Algorithm 1 is bounded as:*

$$\text{Reg}_T(\text{GSE}) = \sum_{i \in [N]} \text{Reg}_{i,T}(\text{GSE}) \leq \mathcal{O} \left(\sum_{k:\Delta_k > 0} \frac{\log(T)}{\Delta_k} + \frac{N^2 \log(T)}{p\lambda_{N-1}(\text{Lap}(\mathcal{G}))} + \frac{KN^2 \log(NT)}{p} \right),$$

where $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ is the second smallest eigenvalue of $\text{Lap}(\mathcal{G})$.

We continue our discussion on how the base graph $\mathcal{G}$ topology affects the regret bound. In addition to the parameter p, which determines the difference between $\mathcal{G}$ and $\mathcal{G}_t$, another key factor is $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$, which reflects the topology of the base graph $\mathcal{G}$. This value is the algebraic connectivity or Fiedler value of $\mathcal{G}$, which reflects how well connected the overall graph is. To illustrate

this, we next provide more explicit regret upper bounds by specifying $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ for different base graph topologies commonly used in distributed optimization Duchi et al. [2011]. The following corollary summarises how the choice of the random gossip matrix in Equation (2) interacts with different topologies of the base graph. The proof of $\lambda_{N-1}(\text{Lap}(\mathcal{G}))$ for various base graphs $\mathcal{G}$ can be found in Corollary 1 of Duchi et al. [2011].

Corollary 5.4. *For specific choices of the base graph $\mathcal{G}$, the regret upper bound in Theorem 5.3 simplifies as follows:*

1) When $\mathcal{G}$ is a complete graph with $\lambda_{N-1}(\text{Lap}(\mathcal{G})) = N$, and refining $L^* = \left\lceil -\frac{2\log(NT)}{\log(1-p)} \right\rceil$ in Equation (5), the regret upper bound simplifies to: $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k} + \frac{KN\log T}{p}\right)$.

2) When $\mathcal{G}$ is a $\sqrt{N} \times \sqrt{N}$ 2D grid, we have $\lambda_{N-1}(\text{Lap}(\mathcal{G})) = 2\left(1 - \cos\left(\frac{\pi}{\sqrt{N}}\right)\right) = \Theta(1/N)$. The corresponding regret bound becomes: $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k} + \frac{N^2(K+N)\log T}{p}\right)$.

3) When $\mathcal{G}$ is an expander graph with a bounded ratio between minimum and maximum node degrees, we have $\lambda_{N-1}(\text{Lap}(\mathcal{G})) = \Theta(1)$. In this case, the regret bound simplifies to: $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k} + \frac{KN^2\log T}{p}\right)$.

Remark 5.5 (Comparison of Regret Bounds). In Theorem 5.3, for any fixed connected base graph $\mathcal{G}$ and $0 < p \leq 1$, we obtain the optimal centralized regret $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k}\right)$ (see the lower bound in Section 5.2) plus an additional term $\mathcal{O}\left(\frac{N^2\log T}{p\lambda_{N-1}(\text{Lap}(\mathcal{G}))} + \frac{KN\log T}{p}\right)$. We emphasize that our regret bound outperforms existing work—many of which are special (degenerated) cases of our more general framework—and is easier to interpret. Notably, Xu and Klabjan [2023] study MA-MAB under the classical E-R model (where $\mathcal{G}$ is a complete graph) with p largely over $1/2$, and derive a regret bound of $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{N\log T}{\Delta_k}\right)$. In the same setting, based on Corollary 5.4, we obtain a regret bound of $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\log(T)}{\Delta_k} + KN\log(T)\right)$, which is significantly smaller. Also, Zhu and Liu [2023] consider B-connected graphs, which do not capture E-R random graphs; the distinction between the two models is discussed in detail in Yuan et al. [2024]. Finally, when $p = 1$, the communication graph becomes time-invariant, which corresponds to most existing work where connected graphs are assumed. For example, Zhu et al. [2021] and Zhu and Liu [2023] study such settings. The former derive a regret bound of $\mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{N^2\log T}{\Delta_k}\right)$, which is worse than our results with a dependency on N. The latter obtain $\mathcal{O}\left(\max\left(\sum_{k:\Delta_k>0} \frac{N\log T}{\mathcal{N}_k\Delta_k}, K_1, K_2\right)\right)$, where K_1 and K_2 depend on T but lack explicit formulas, may grow arbitrarily large, and are difficult to interpret—at least to the best of our knowledge.

Remark 5.6 (Regret and Communication Trade-off). We emphasize that our regret upper bound exhibits a novel trade-off between regret performance and communication efficiency in the presence of random graphs. It is straightforward to observe that increasing p reduces the regret bound but increases communication overhead, thereby lowering communication efficiency. For classical E-R graphs, the expected number of agent-to-agent communications per agent over the given time horizon is pNT, while the expected regrets decrease as p increases. Therefore, for a fixed time horizon T and a given base graph $\mathcal{G}$, the parameter p can be tuned to balance communication overhead and reward maximization, informing practical decision making.

5.2 Lower Bound

In this section, we establish a lower bound on the global regret, as defined in Equation (1). Unlike the upper bound, this lower bound applies to any reasonable algorithm under a specific problem instance, highlighting the problem complexity. Detailed proofs are deferred to Appendix C. We begin by introducing several definitions related to the problem instance and the notion of a reasonable algorithm.

Definition 5.7 (Gaussian Instance). An instance ν is called a *Gaussian instance* if, for every agent $i \in [N]$ and arm $k \in [K]$, the reward distribution $\mathbb{P}_{i,k}$ is Gaussian with unit variance.

Definition 5.8 (Consistent Policy). Let $\mathcal{I}$ be a class of problem instances. Let $\mathrm{Reg}_T^\nu(\pi)$ denote the regret incurred by policy π on instance ν. A policy (algorithm) π is said to be *consistent* on $\mathcal{I}$ if there exist constants $C > 0$ and $s \in (0,1)$ such that $\mathrm{Reg}_T^\nu(\pi)$ meets $\mathrm{Reg}_T^\nu(\pi) \leq CT^s$ for all instances $\nu \in \mathcal{I}$.

We next present the problem instance constructed to establish the regret lower bound. Note that an alternative, equivalent expression of the global regret reads as

$$\mathrm{Reg}_T^\nu(\pi) = \sum_{k \in [K]} \Delta_k \sum_{i \in [N]} \mathbb{E}[T_{i,k}(T)].$$

Thus, to derive a lower bound on regret, it suffices to lower bound the total expected number of pulls $\sum_{i \in [N]} \mathbb{E}[T_{i,k}(T)]$ for suboptimal arms $k \in [K]$. To this end, we consider a Gaussian instance ν where each $\mathbb{P}_{i,k} = \mathcal{N}(\mu_{i,k}, 1)$ with the random graph communication protocal based any connected base graph $\mathcal{G}$ and connection probability p, and construct a perturbed instance ν' such that:

$$\mathbb{P}'_{i,a} = \begin{cases} \mathcal{N}(\mu_{i,a}, 1), & \text{if } a \neq k, \\ \mathcal{N}(\mu_{i,a} + (1+\varepsilon)\Delta_a, 1), & \text{if } a = k, \end{cases}$$

for a small constant $\varepsilon \in (0,1)$ representing the level of perturbation. The communication protocal for ν' is the same as that for ν. Under this perturbation, which defines a new problem instance, we derive the following information-theoretic inequality:

$$\sum_{j \in [N]} \mathbb{E}[T_{j,k}(T)] \cdot \frac{(1+\varepsilon)^2 \Delta_k^2}{2} \geq \log\left(\frac{NT\varepsilon\Delta_k}{4\left(\mathrm{Reg}_T^\nu(\pi) + \mathrm{Reg}_T^{\nu'}(\pi)\right)}\right), \tag{8}$$

which imposes a lower bound on the total number of pulls of arm k across all agents.

By applying this inequality to a consistent policy π and rearranging the terms, we obtain the following lower bound on regret. The formal statement reads as follows.

Theorem 5.9. *Let π be a consistent policy on the class $\mathcal{I}$ of Gaussian instances for some $s \in (0,1)$. Then, for all instances $\nu \in \mathcal{I}$ and any $\varepsilon \in (0,1]$, the following holds:*

$$\lim_{T \to \infty} \frac{\mathrm{Reg}_T^\nu(\pi)}{\log T} \geq \sum_{k:\Delta_k > 0} \frac{2(1-s)}{(1+\varepsilon)^2 \Delta_k}.$$

Remark 5.10 (Comparison with Upper Bounds). Recall that the regret upper bound in Theorem 5.3 consists of two components: a centralized term and additional terms that capture the influence of the communication graph. The lower bound in Theorem 5.9 shows that the centralized component, $\mathcal{O}\left(\sum_{k:\Delta_k > 0} \frac{\log T}{\Delta_k}\right)$, is tight, thereby establishing the optimality of the centralized regret achieved by Algorithm 1. This result is intuitive: the problem effectively involves N agents collaboratively solving a global multi-armed bandit task. With appropriate information sharing, the collective performance can match that of a single-agent bandit problem with full access to all rewards. Hence, achieving a global regret of the same order as the classical centralized bandit setting is both natural and optimal.

6 Experiments

In this section, we demonstrate the effectiveness of our algorithm through numerical experiments on both synthetic and real-world datasets. [1] The objective is twofold. First, we show that the cumulative regret of our algorithm grows logarithmically with respect to T and is significantly smaller than that of existing benchmarks, thereby validating our theoretical findings. We use DrFed-UCB, proposed by [Xu and Klabjan, 2023], as the baseline. Second, we conduct a simulation study to examine how the regret depends on the link probability p and the algebraic connectivity of the base graph $\mathcal{G}$, as reflected in the regret bound. We evaluate the impact of different values of p across various base graphs, including the complete graph, grid, and Petersen graph [Holton and Sheehan, 1993]. Each edge in the base graph $\mathcal{G}$ appears in the communication graph $\mathcal{G}_t$ with probability p.

---

[1]The code for the experiments is available at https://anonymous.4open.science/r/multi-agent-bandit-856B/
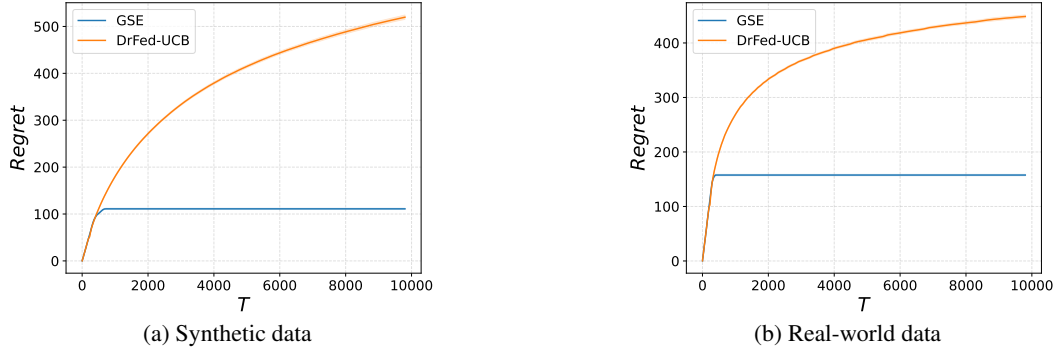
(a) Synthetic data

(b) Real-world data

Figure 1: Comparison of the empirical results of our algorithm and DrFed-UCB. The base graph is a complete graph and the link probability $p = 0.9$.

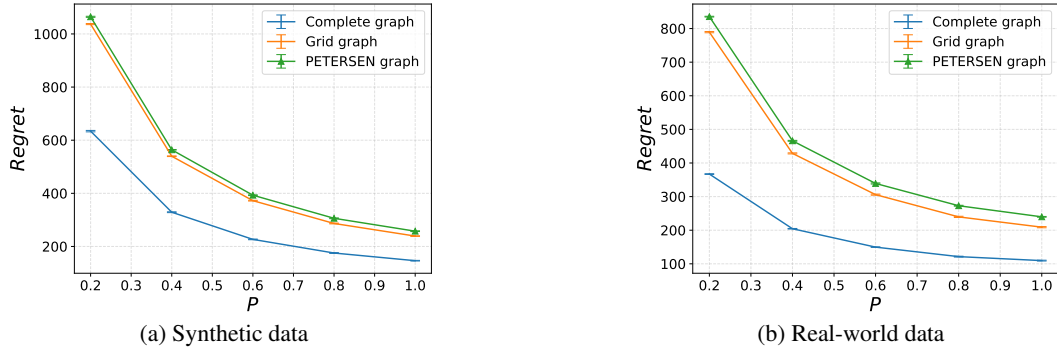(a) Synthetic data

(b) Real-world data

Figure 2: Regret of our algorithm on different base graphs with different p value.

Experimental Settings. For synthetic experiment setting, We set $T = 10000$, $N = 16$, and $K = 5$; for the Petersen graph, we use $N = 10$ by definition. For the comparison with DrFed-UCB, we consider a complete graph and a high link probability ($p = 0.9$), as required therein. Before the game starts, we sample each independently and uniformly from the interval $[0, 1]$ for each agent i. The local mean reward of arm k on agent i is given by $\mu_{i,k} = q_i \cdot \frac{k-1}{K-1}$, and the global mean reward of arm k is $\mu_k = \frac{k-1}{K-1} \cdot \frac{\sum_{i \in [N]} q_i}{N}$. At each time step t, each agent $i \in [N]$ selects an arm and observes the local reward. For real-world experiments, we use the MovieLens dataset and refer to Yi and Vojnovic [2023] for details. We set the horizon $T = 10,000$, and select 20 users as agents ($N = 20$) and 5 genres as arms ($K = 5$). At each time step t, each agent randomly selects a movie from the genres. All ratings (rewards) of movies are normalised to $[0, 1]$.

Experimental results. All experiments are performed with 20 independent replications. The shaded areas consider a range centered around the mean with half-width corresponding to the empirical standard deviation over 20 repetitions. In Figure 1, we observe that our algorithm consistently outperforms DrFed-UCB on both synthetic and real-world datasets. In all runs, after an initial exploration period, our algorithm eliminates a significant number of suboptimal actions, resulting in near-constant regret thereafter. In Figure 2, we observe that increasing the link probability p improves the algorithm's performance, clearly validating the regret–communication trade-off. Additionally, different base graphs significantly impact the regret under the same p value—with the complete graph yielding the lowest regret.

We summarize this paper and point out exciting future directions in Appendix A.

References

Juliette Achddou, Nicolò Cesa-Bianchi, and Hao Qiu. Distributed online optimization with stochastic agent availability. *arXiv preprint arXiv:2411.16477*, 2024.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

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

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

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3): 592–606, 2011.

Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Deepak Ganesan, Alberto Cerpa, Wei Ye, Yan Yu, Jerry Zhao, and Deborah Estrin. Networking issues in wireless sensor networks. *Journal of parallel and distributed computing*, 64(7):799–814, 2004.

Saeed Ghoorchian and Setareh Maghsudi. Multi-armed bandit for energy-efficient and delay-sensitive edge computing in dynamic networks with uncertainty. *IEEE Transactions on Cognitive Communications and Networking*, 7(1):279–293, 2020.

Derek Allan Holton and John Sheehan. *The petersen graph*, volume 7. Cambridge University Press, 1993.

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 1484–1489. IEEE, 2013.

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

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

Jinlong Lei, Peng Yi, Yiguang Hong, Jie Chen, and Guodong Shi. Online convex optimization over erdos-rényi random networks. *Advances in neural information processing systems*, 33: 15591–15601, 2020.

Xiuxian Li, Lihua Xie, and Na Li. A survey on distributed online optimization and online games. *Annual Reviews in Control*, 56:100904, 2023.

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

David Mateos-Núnez and Jorge Cortés. Distributed online convex optimization over jointly connected digraphs. *IEEE Transactions on Network Science and Engineering*, 1(1):23–37, 2014.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3 (3):1–35, 2019.

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.

Xiong Wang, Jiancheng Ye, and John CS Lui. Decentralized task offloading in edge computing: A multi-user multi-armed bandit approach. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1199–1208. IEEE, 2022a.

Xuchuang Wang, Lin Yang, Yu-Zhen Janice Chen, Xutong Liu, Mohammad Hajiesmaili, Don Towsley, and John CS Lui. Achieving near-optimal individual regret & low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations*, 2022b.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 36: 74799–74855, 2023.

Mengfan Xu and Diego Klabjan. Multi-agent multi-armed bandit regret complexity and optimality. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

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, 2013.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

Jialin Yi and Milan Vojnovic. Doubly adversarial federated bandits. In *International Conference on Machine Learning*, pages 39951–39967. PMLR, 2023.

Deming Yuan, Alexandre Proutiere, Guodong Shi, et al. Multi-agent online optimization. *Foundations and Trends® in Optimization*, 7(2-3):81–263, 2024.

Jiang Zhu, Yonghui Song, Dingde Jiang, and Houbing Song. Multi-armed bandit channel access scheme with cognitive radio technology in wireless sensor networks for the internet of things. *IEEE access*, 4:4609–4617, 2016.

Jingxuan Zhu and Ji Liu. Distributed multiarmed bandits. *IEEE Transactions on Automatic Control*, 68(5):3025–3040, 2023.

Jingxuan Zhu, Romeil Sandhu, and Ji Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3078–3083. IEEE, 2020.

Jingxuan Zhu, Ethan Mulle, Christopher S Smith, Alec Koppel, and Ji Liu. Decentralized upper confidence bound algorithms for homogeneous multi-agent multi-armed bandits. *IEEE Transactions on Automatic Control*, 2025.

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.

A   Conclusion and Future Work

In this paper, we study the multi-agent multi-armed bandit (MA-MAB) problem under general Erdős–Rényi random networks with heterogeneous rewards. To the best of our knowledge, we are the first to formulate MA-MAB with Erdős–Rényi random networks, where the communication graph is induced by a base graph and each edge in the base graph appears in the communication graph with probability p. This formulation generalizes the classical Erdős–Rényi model, in which the base graph is complete. We propose an algorithmic framework that incorporates the gossip communication protocol into arm elimination. Importantly, we analyze the regret bound of the algorithm and show that it improves the regret even under the classical Erdős–Rényi model. Moreover, our regret bound holds for any p, explicitly characterizes its dependency on p and the algebraic connectivity of the base graph. This naturally reveals a trade-off between regret and communication efficiency.

Moving forward, while we focus extensively on the stochastic setting, it would be valuable and exciting to explore other reward models—such as contextual bandits, where rewards depend on dynamic, non-stationary contexts. In addition, achieving the optimal trade-off among regret, communication, and privacy, as previously studied in homogeneous MA-MAB settings, points out a meaningful direction for future research.

B   Auxiliary results

In this section, we show several auxiliary lemmas that will be helpful throughout the paper.

The following lemma is the concentration bound for the estimation of the observed expected rewards,

Lemma B.1. *Let $\widehat{\mu}_{i,k}(t)$ be the observed empirical average of the expected reward up to the end of round $t-1$. Then,*

$$\mathbb{P}\left[|\widehat{\mu}_{i,k}(t) - \mu_{i,k}| > \sqrt{\frac{2\log T}{T_{i,k}(t)}}\right] \leq \frac{2}{T^2}$$

for $i \in [N]$, $k \in [K]$ and $t \in [T]$ holds.

Proof of Lemma B.1. This lemma follows immediately from Hoeffding's inequality and a union bound. □

The following lemma is the concentration bound for the estimation of the observed expected rewards.

Lemma B.2. *Let $\widehat{\mu}_{i,k}(t)$ be the observed empirical average of the expected reward for agent i pulling arm k. Then,*

$$\mathbb{P}\left[\left|\sum_{i\in[N]}\widehat{\mu}_{i,k} - \sum_{i\in[N]}\mu_{i,k}\right| > \sqrt{\frac{4N\log T}{\min_{i\in[N]}\{T_{i,k}(t)\}}}\right] \leq \frac{2}{T^2}$$

holds for any $i \in [N], k \in [K]$ and $t \in [T]$.

Proof of Lemma B.2. We sample $T_{i,k}(t)$ number of random variables iid from $\mathbb{P}_{i,k}$ for each $i \in [N]$, each taking values in $[0,1]$, and hence $X_{i,k}(\tau) - \mu_{i,k}$ is 1-subgaussian for each τ.

According to the property of subgaussian variables, we can obtain that

$$\sum_{i\in[N]}(\widehat{\mu}_{i,k}(t) - \mu_{i,k}) \text{ is } \left(\sum_{i\in[N]}\frac{1}{T_{i,k}(t)}\right)^{1/2} \text{ - subgaussian.}$$

Moreover, note that $\sum_{i\in[N]}\frac{1}{T_{i,k}(t)} \leq \frac{N}{\min_{i\in[N]}T_{i,k}(t)}$, then by applying Chernoff's bound, we have

$$\mathbb{P}\left(\left|\sum_{i\in[N]}\widehat{\mu}_{i,k} - \mu_{i,k}\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2\min_{i\in[N]}\{T_{i,k}(t)\}}{2N}\right).$$

Taking $\epsilon = \sqrt{\frac{4N \log T}{\min_{i\in[N]}\{T_{i,k}(t)\}}}$, we obtain that

$$\mathbb{P}\left(\left|\sum_{i\in[N]} \widehat{\mu}_{i,k}(t) - \mu_{i,k}\right| \geq \sqrt{\frac{4N \log T}{\min_{i\in[N]}\{T_{i,k}(t)\}}}\right) \leq \frac{2}{T^2}$$

which ends the proof of Theorem B.1 $\qquad\square$

We provide lemmas for the convergence bound of randomised gossip algorithms. Similar proof could be found Lei et al. [2020], Achddou et al. [2024].

Lemma B.3 (Random Graph). *Let the communication protocol based on random graphs defined in Theorem 3.1 hold. For $t = 1\ldots T$, W_t is doubly stochastic matrix and symmetric and i.i.d. $\forall v \in V$, $\forall s, t \in [T]$ such that $t > s$,*

$$\mathbb{P}\left(\left\|W_t \cdots W_{s+1} e_v - \frac{1}{N}\mathbf{1}\right\|_2 \geq \delta\right) \leq \frac{\lambda_2(\mathbb{E}[W^2])^{t-s}}{\delta^2}.$$

When $t - s \geq \left\lceil \frac{3\log(T)}{\log \lambda_2(\mathbb{E}[W^2])^{-1}} \right\rceil = \tau'$, we have

$$\mathbb{P}\left(\left\|W_t \cdots W_{s+1} e_v - \frac{1}{N}\mathbf{1}\right\|_2 \geq \delta\right) \leq \delta. \tag{9}$$

Furthermore, when $t - s \geq \tau^* = \left\lceil \frac{3N\log(T)}{p\lambda_{N-1}(\mathrm{Lap}(\mathcal{G}))} \right\rceil \geq \tau'$, Equation (9) still holds.

Proof. Using Markov's inequality we have

$$\mathbb{P}\left(\left\|W_t \cdots W_{s+1} e_v - \frac{1}{N}\mathbf{1}\right\|_2 \geq \delta\right) \leq \frac{\mathbb{E}\left(\left\|W_t \cdots W_{s+1} e_v - \frac{1}{N}\mathbf{1}\right\|_2^2\right)}{\delta^2}.$$

Let $\tilde{W}_k = W_k - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ and assume

$$\mathbb{E}\left[\left\|W_{k-1} \cdots W_{s+1} e_v - \frac{1}{N}\mathbf{1}\right\|_2^2\right] \leq e_v^T e_v \left\|\mathbb{E}[W_1 W_1^\top] - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}}^{k-s-1} \qquad \text{for some } k-1 > s.$$

Let $\mathcal{F}_{k-1}$ be the σ-algebra generated by all random events up to time $k-1$. We have that

$$\mathbb{E}\left[\left\|W_k^\top \cdots W_{s+1}^\top e_v - \frac{1}{N}\mathbf{1}\right\|_2^2\right] = \mathbb{E}\left[e_v^T \tilde{W}_{s+1}^\top \cdots \tilde{W}_{k-1}^\top \tilde{W}_k^\top \tilde{W}_k \tilde{W}_{k-1} \cdots \tilde{W}_{s+1} e_v\right]$$

$$= \mathbb{E}\left[e_v^T \tilde{W}_{s+1}^\top \cdots \tilde{W}_{k-1}^\top \mathbb{E}[\tilde{W}_k^\top \tilde{W}_k \mid \mathcal{F}_{k-1}] \tilde{W}_{k-1} \cdots \tilde{W}_{s+1} e_v\right]$$

$$= \mathbb{E}\left[e_v^T \tilde{W}_{s+1}^\top \cdots \tilde{W}_{k-1}^\top \mathbb{E}[\tilde{W}_1^\top \tilde{W}_1] \tilde{W}_{k-1} \cdots \tilde{W}_{s+1} e_v\right]$$
$$\text{(by independence of } W_k\text{)}$$

$$\leq \left\|\mathbb{E}[W_1 W_1^\top] - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}} e_v^T e_v \left\|\mathbb{E}[W_1 W_1^\top] - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}}^{k-s-1}$$

$$\leq \lambda_2(\mathbb{E}[W^2])^{t-s} e_v^T e_v \left\|\mathbb{E}[W_1 W_1^\top] - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}}^{k-s-1}$$

which by induction, suffices to prove the lemma. Moreover, the proof of $\tau^* \geq \tau'$ follows from the fact that

$$\frac{1}{\log \lambda_2^{-1}} \leq \frac{1}{1 - \lambda_2}, \lambda_2(\mathbb{E}[W^2]) \leq 1 - \frac{p\lambda_{N-1}(\mathrm{Lap}(\mathcal{G}))}{N} \text{ and } \frac{1}{\log(1-p)^{-1}} \leq \frac{1}{p},$$

where the second inequality is taken from Theorem 6.1 in Achddou et al. [2024]. $\qquad\square$

The following lemmas guarantee local consistency between agents.

Lemma B.4. *Let Theorem 3.1 hold. With probability $1 - N^2 T \delta$, Algorithm 1 guarantees that for fixed arm $k \in [K]$, and for every $i, j \in [N]$ and for every $t \in [T]$,*

$$|T_{i,k}(t) - T_{j,k}(t)| \leq K N L_p(\delta)$$

where $L_p(\delta) = \left\lceil \frac{\log(\delta)}{\log(1-p)} \right\rceil$ denotes the maximum number of rounds each edge within base graph $\mathcal{G}$ is connected in the communication graph $\mathcal{G}_t$ with probability $1 - \delta$.

Proof. Fix an agent $i \in [N]$ and a time step $t \in [T]$. According to Theorem 3.1 and Algorithm 1, p be the probability that agent i communicates with a fixed neighbour $j \in \mathcal{N}_i(t)$ in any given step, independently of the past. For a non-negative integer L, we have

$$\mathbb{P}\left(i \text{ does not contact j during the next } L + 1 \text{ steps} \right) = (1 - p)^{L+1}.$$

The communication gap length at time t is the number of steps starting from time t until agent i next successfully communicates with agent j. Choose a confidence parameter $\delta \in (0, 1)$, we have

$$\mathbb{P}\left(\text{time until first contact between agents } i \text{ and } j \text{ exceeds } L_p(\delta) \right) \leq \delta.$$

Applying a union bound to gives

$$\mathbb{P}\left(\exists i \in [N], j \in \mathcal{N}_i, t \in [T] : \text{time until first contact between } i \text{ and } j \text{ after time } t \text{ exceeds } L_p(\delta) \right)$$
$$\leq N^2 T \delta.$$

Note that we have $\mathcal{N}_i(t) \subseteq \mathcal{N}_i$, where $\mathcal{N}_i$ is a fixed superset of possible neighbors. Hence with probability at least $1 - N^2 T \delta$, for every agent $i \in [N]$, every time step $t \in [T]$, and every neighbor $j \in \mathcal{N}_i$, the time until the next communication between i and j is at most $L_p(\delta)$ time steps.

For any two agents i and j, let $d(i, j) \leq N$ denote the shortest path length between i and j in the base graph $\mathcal{G}$. Because with high probability, information can gossip across each edge within at most $L_p(\delta)$ steps, it follows that information originating at i at time step t reaches j at most by time step

$$t + L_p(\delta) \cdot d(i, j) \leq t + L_p(\delta) \cdot N. \tag{10}$$

In order to be cautious for notations, we use $\mathcal{S}_i(t)$ to denote the active set in round t for Algorithm 1. According to Algorithm 1, during every communication, each agent i replaces its active set by the intersection with its neighbors:

$$\mathcal{S}_i(t+1) = \bigcap_{j \in \mathcal{N}_i(t) \cup \{i\}} \mathcal{S}_j(t).$$

because for all $\mathcal{S}_i(0) = [K]$, at most K distinct arms can ever be removed. Each arm can start a new wave of disagreement. Each wave of disagreement at most last $L_p(\delta) \cdot N$. In the worst case, the waves do not overlap. Consenquently the longest possible sequence of rounds in which $\mathcal{S}_i(t) \neq \mathcal{S}_j(t)$ is $K \cdot N \cdot L_p(\delta)$. During the disagreement period, for a fixed arm k we could upper bound of maximum pull of k is $K N L_p(\delta)$ and lower bound of maximum pull of k is 0. Hence we have

$$|T_{i,k}(t) - T_{j,k}(t)| \leq K N L_p(\delta).$$

$\square$

C Omitted details in Section 5

In this section, we show the omitted details in Section 5.

Proof of Theorem 5.1. When $T_{i,k}(t) \leq KL^*$, the bound trivially holds.

Now we consider the case when $T_{i,k}(t) > KL^*$, We first define the following event:

$$E := \left\{ \bigcap_{\substack{k \in [K] \\ t \in [T]}} \left\{ \frac{\left| \sum_{j \in [N]} \left(\widehat{\mu}_{j,k}(t) - \mu_{j,k} \right) \right|}{N} \leq \sqrt{\frac{4 \log(T)}{N \min_{j \in [N]} \{T_{j,k}(t)\}}} \right\} \right\}$$

$$\cap \left\{ \bigcap_{\substack{j \in [N] \\ \tau^* \leq t-s \leq T}} \left\| W_t \cdots W_{s+1} e_j - \frac{1}{N}\mathbf{1} \right\|_2 \leq \frac{1}{T^2} \right\} \cap \left\{ \bigcap_{\substack{i,j \in [N] \\ k \in [K] \\ t \in [T]}} |T_{i,k}(t) - T_{j,k}(t)| \leq KL^* \right\}. \tag{11}$$

In Equation (11), the first event bounds the global estimation error for each arm, the second ensures near-uniform information mixing across agents after sufficient communication rounds, and the third guarantees that the number of pulls for any arm remains approximately balanced across agents at each time step. By applying Theorem B.2, B.3, B.4 and union bound, we obtain that $\mathbb{P}[E^c] \leq \frac{K}{T} + \frac{N}{T} + N^2 T\delta$ holds for Erdős–Rényi random model defined in Equation (1).

We define $\widetilde{\mu}_k(t) = \frac{1}{N}\sum_{j \in [N]} z_{j,k}(t)$ to be an intermediate variable that has access to each agent's average mean on arm k at time t.

For any agent i, we have

$$\begin{aligned}
|z_{i,k}(t) - \mu_k| &= |z_{i,k}(t) - \widetilde{\mu}_k(t) + \widetilde{\mu}_k(t) - \mu_k| \\
&\leq |\widetilde{\mu}_k(t) - z_{i,k}(t)| + |\widetilde{\mu}_k(t) - \mu_k| &\text{(Triangle inequality)} \\
&= \underbrace{|\widetilde{\mu}_k(t) - z_{i,k}(t)|}_{\text{Consensus Error}} + \underbrace{\frac{\left|\sum_{j \in [N]}(\widehat{\mu}_{j,k}(t) - \mu_{j,k})\right|}{N}}_{\text{Estimation Error}},
\end{aligned} \tag{12}$$

The last equality is due to the definition of the global mean reward and $\widetilde{\mu}_k(t)$. Let us first focus on Consensus Error. For any $i \in [N]$, According to the update in Equation (3) we have

$$\begin{aligned}
z_{i,k}(t) &= \sum_{j \in [N]}[W_{t-1}]_{i,j}z_{i,k}(t-1) + \widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2) \\
&= \sum_{j \in [N]}[W_{t-1}\cdots W_{t-s}]_{i,j}z_{i,k}(t-s) + \sum_{\tau=t-s}^{t-2}\sum_{j \in [N]}[W_{t-2}\cdots W_{\tau+1}]_{i,j}\left(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)\right) \\
&\quad + \widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2) \tag{13} \\
&= \sum_{j \in [N]}[W_{t-1}\cdots W_1]_{i,j}z_{i,k}(1) + \sum_{\tau=1}^{t-2}\sum_{j \in [N]}[W_{t-2}\cdots W_{\tau+1}]_{i,j}\left(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)\right) \\
&\quad + \widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2). &\text{(setting } s = t-1)
\end{aligned}$$

We also have

$$\begin{aligned}
\widetilde{\mu}_k(t) &= \frac{1}{N}\sum_{j \in [N]} z_{j,k}(t) \\
&= \widetilde{\mu}_k(t-s) + \frac{1}{N}\sum_{\tau=t-s}^{t-1}\sum_{j \in [N]}(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \\
&= \widetilde{\mu}_k(1) + \frac{1}{N}\sum_{\tau=1}^{t-1}\sum_{j \in [N]}(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) &\text{(setting } s = t-1) \\
&= \frac{1}{N}\sum_{j \in [N]} z_{j,k}(1) + \frac{1}{N}\sum_{\tau=1}^{t-1}\sum_{j \in [N]}(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \\
&= \frac{1}{N}\sum_{j \in [N]} z_{j,k}(1) + \frac{1}{N}\sum_{\tau=1}^{t-2}\sum_{j \in [N]}(\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \\
&\quad + \frac{1}{N}\sum_{j \in [N]}(\widehat{\mu}_{j,k}(t-1) - \widehat{\mu}_{j,k}(t-2)) \tag{14}
\end{aligned}$$

Hence, we obtain

$$\tilde{\mu}_k(t) - z_{i,k}(t) = \sum_{\tau=1}^{t-2} \left(\sum_{j\in[N]} \left(\frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right)$$
$$+ \frac{1}{N} \sum_{j\in[N]} (\widehat{\mu}_{j,k}(t-1) - \widehat{\mu}_{j,k}(t-2)) - (\widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2))$$
$$+ \frac{1}{N} \sum_{j\in[N]} z_{j,k}(1) - \sum_{j\in[N]} [W_{t-1}\cdots W_1]_{i,j} z_{j,k}(1).$$
$$= \sum_{\tau=1}^{t-2} \left(\sum_{j\in[N]} \left(\frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right)$$
$$+ \frac{1}{N} \sum_{j\in[N]} (\widehat{\mu}_{j,k}(t-1) - \widehat{\mu}_{j,k}(t-2)) - (\widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2)),$$

.Taking absolute values on both sides, we have

$$|\tilde{\mu}_k(t) - z_{i,k}(t)| \le \left| \sum_{\tau=1}^{t-2} \left(\sum_{j\in[N]} \left(\frac{1}{N}[W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right) \right|$$
$$+ \left| \frac{1}{N} \sum_{j\in[N]} (\widehat{\mu}_{j,k}(t-1) - \widehat{\mu}_{j,k}(t-2)) - (\widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2)) \right|$$
$$\le \underbrace{\left| \sum_{\tau=1}^{t-\tau^*-2} \left(\sum_{j\in[N]} \left(\frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right) \right|}_{\heartsuit}$$
$$+ \underbrace{\left| \sum_{\tau=t-\tau^*-1}^{t-2} \left(\sum_{j\in[N]} \left(\frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right) \right|}_{\spadesuit}$$
$$+ \underbrace{\left| \frac{1}{N} \sum_{j\in[N]} (\widehat{\mu}_{j,k}(t-1) - \widehat{\mu}_{j,k}(t-2)) - (\widehat{\mu}_{i,k}(t-1) - \widehat{\mu}_{i,k}(t-2)) \right|}_{\clubsuit}. \quad (15)$$

Now we analyze three terms on the right-hand side of Equation (15).

Bounding term $\heartsuit$. Conditioning on event E, we obtain

$$\heartsuit = \left| \sum_{\tau=1}^{t-\tau^*-2} \left(\sum_{j\in[N]} \left(\frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau-1)) \right) \right|$$
$$\le \sum_{\tau=1}^{t-\tau^*-2} \sum_{j\in[N]} \left| \frac{1}{N} - [W_{t-2}\cdots W_{\tau+1}]_{i,j} \right| \qquad \text{(rewards are bounded in the interval } [0,1])$$
$$= \sum_{\tau=1}^{t-\tau^*-2} \left\| W_{t-2}\cdots W_{\tau+1} e_i - \frac{\mathbf{1}}{N} \right\|_1$$
$$\le \sqrt{N} \cdot \sum_{\tau=1}^{t-\tau^*-2} \left\| W_{t-2}\cdots W_{\tau+1} e_i - \frac{\mathbf{1}}{N} \right\|_2$$

16

$$\leq \frac{\sqrt{N}(t - \tau^*)}{T^2}$$

$$\leq \frac{\sqrt{N}}{T}$$

$$\leq \frac{\sqrt{N}}{T_{i,k}(t)}, \tag{16}$$

where the third inequality comes from the condition that $t - \tau - 1 \in [\tau^*, t - 3]$ and the event E.

Bounding term ♠. We have

$$\spadesuit = \left| \sum_{\tau = t - \tau^* - 1}^{t-2} \left(\sum_{j \in [N]} \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) (\widehat{\mu}_{j,k}(\tau) - \widehat{\mu}_{j,k}(\tau - 1)) \right) \right|$$

$$\leq \sum_{\tau = t - \tau^* - 1}^{t-2} \left(\sum_{j \in [N]} \left| \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) \right| \left| \left(\frac{\sum_{s=1}^{\tau} \mathbb{I}\{A_j(s) = k\} X_{j,k}(s)}{T_{j,k}(\tau)} - \frac{\sum_{s=1}^{\tau-1} \mathbb{I}\{A_j(s) = k\} X_{j,k}(s)}{T_{j,k}(\tau - 1)} \right) \right| \right)$$

$$(\text{ definition of } \widehat{\mu}_{j,k}(t))$$

$$\leq \sum_{\tau = t - \tau^* - 1}^{t-2} \left(\sum_{j \in [N]} \left| \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) \right| \cdot \right.$$

$$\left. \left| \left(\frac{\sum_{s=1}^{\tau-1} \mathbb{I}\{A_j(s) = k\} X_{j,k}(s) + X_{j,k}(\tau)}{T_{j,k}(\tau)} - \frac{\sum_{s=1}^{\tau-1} \mathbb{I}\{A_j(s) = k\} X_{j,k}(s)}{T_{j,k}(\tau) - 1} \right) \right| \right)$$

$$\leq \sum_{\tau = t - \tau^* - 1}^{t-2} \left(\sum_{j \in [N]} \left| \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) \right| \left| \left(\frac{-\sum_{s=1}^{\tau-1} \mathbb{I}\{A_j(s) = k\} X_{j,k}(s) + (T_{j,k}(\tau) - 1) X_{j,k}(\tau)}{T_{j,k}(\tau) \, (T_{j,k}(\tau) - 1)} \right) \right| \right)$$

$$\leq \sum_{\tau = t - \tau^* - 1}^{t-2} \sum_{j \in [N]} \left| \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) \right| \frac{1}{T_{j,k}(\tau)}$$

$$(\text{rewards are bounded in the interval } [0, 1])$$

$$\leq \sum_{\tau = t - \tau^* - 1}^{t-2} \sum_{j \in [N]} \left| \left(\frac{1}{N} - [W_{t-1} \cdots W_{\tau+1}]_{i,j} \right) \right| \frac{1}{\max\{T_{i,k}(\tau) - KL^*, 1\}} \qquad (\text{event } E)$$

$$\leq \frac{4\tau^*}{\max\{T_{i,k}(t) - KL^*, 1\}},$$

where the second inequality follows from the fact that there are only two possible cases for $T_{j,k}(\tau)$ and $T_{j,k}(\tau - 1)$. When $T_{j,k}(\tau) = T_{j,k}(\tau - 1)$ the absolute difference of means is trivially 0, so the only non-trivial case to consider is when $T_{j,k}(\tau) = T_{j,k}(\tau - 1) + 1$. The last inequality is due to that the upper bound of the total-variation distance from any distribution to the uniform distribution is 1.

Bounding term ♣. Conditioning on E, we obtain

$$\clubsuit = \left| \frac{1}{N} \sum_{j \in [N]} (\widehat{\mu}_{j,k}(t - 1) - \widehat{\mu}_{j,k}(t - 2)) - (\widehat{\mu}_{i,k}(t - 1) - \widehat{\mu}_{i,k}(t - 2)) \right|$$

$$\leq \sum_{j \in [N]} \frac{2}{N T_{j,k}(t - 1)} + \frac{2}{T_{i,k}(t - 1)}$$

$$\leq \frac{4}{\max\{T_{i,k}(t - 1) - KL^*, 1\}},$$

Next, let us analyse Estimation Error. Conditioned on E, for all $k \in [K]$ and $t \in [T]$ we obtain

$$\frac{\left| \sum_{j \in [N]} (\widehat{\mu}_{j,k}(t) - \mu_{j,k}) \right|}{N} \leq \sqrt{\frac{4 \log(T)}{N \min_{j \in [N]} \{T_{j,k}(t)\}}}$$

17

$$\leq \sqrt{\frac{4\log(T)}{N\max\{T_{i,k}(t) - KL^*, 1\}}}, \qquad (17)$$

where the last inequality is due to the event E.

Combining all the results collected so far, we can finally derive the concentration bound conditioned on the event E. For any $i \in [N]$ and $k \in [K]$, we obtain

$$|z_{i,k}(t) - \mu_k| \leq \sqrt{\frac{4\log(T)}{N\max\{T_{i,k}(t) - KL^*, 1\}}} + \frac{4\left(\sqrt{N} + \tau^*\right)}{\max\{T_{i,k}(t) - KL^*, 1\}}. \qquad (18)$$

$\square$

Proof of Theorem 5.2. First, for all agents we consider the cases under the event E. According to Algorithm 1, if arm k is eliminated, there are only two possible cases: 1) there exists some k' such that $z_{i,k}(t) + c_{i,k}(t) \leq z_{i,k'}(t) - c_{i,k'}(t)$; 2) When Algorithm 1 updates the active set $\mathcal{S}_i(t+1) \leftarrow \bigcap_{j \in \mathcal{N}_i(t) \cup \{i\}} \mathcal{S}_j(t)$, $k \notin \mathcal{S}_j$ for any $j \in \mathcal{N}_i(t)$.

For Case 1, Since we have $z_{i,k}(t) + c_{i,k}(t) \leq \mu_k + 2c_{i,k}(t)$ as well as $z_{i,k'}(t) - c_{i,k'}(t) \geq \mu_{k'} - 2c_{i,k'}(t)$, then when

$$2(c_{i,k}(t) + c_{i,k'}(t)) \leq \mu_{k'} - \mu_k \leq \Delta_k$$

arm k will be essentially eliminated. Due to the pulling rule of Algorithm 1, we have $|T_{i,k}(t) - T_{i,k'}(t)| \leq 1$. Thus when

$$\Delta_k \geq 2\left(\sqrt{\frac{4\log(T)}{N\max\{T_{i,k}(t) - KL^*, 1\}}} + \frac{4\left(\sqrt{N} + \tau^*\right)}{\max\{T_{i,k}(t) - KL^*, 1\}}\right).$$

Hence, here we know that when

$$T_{i,k}(t) \geq \frac{64\log(T)}{N\Delta_k^2} + \frac{16\left(\sqrt{N} + \tau^*\right)}{\Delta_k} + KL^*,$$

arm k will be essentially eliminated.

For Case 2,the optimal arm k^* it cannot be eliminated, because for agents i we consider the cases under the event E.

For all agents cases under event E^c, we have

$$T \cdot \mathbb{P}(E^c)\Delta_{\max} \leq 3KN\Delta_{\max}.$$

Therefore, combining all inequalities above, we have

$$\text{Reg}_{i,T}(\pi) \leq \sum_{k:\Delta_k > 0} \left(\frac{64\log(T)}{N\Delta_k} + 16\left(\sqrt{N} + \tau^*\right)\right) + KL^* + 3KN\Delta_{\max}$$

which ends the proof. $\square$

Proof of Theorem 5.3. By adding up Theorem 5.2 for all agents $i \in \mathcal{N}$, Theorem 5.3 can be proved through the facts that

$$\frac{1}{log\lambda_2^{-1}} \leq \frac{1}{1 - \lambda_2}, \lambda_2(\mathbb{E}[W^2]) \leq 1 - \frac{p\lambda_{N-1}(\text{Lap}(\mathcal{G}))}{N} \text{ and } \frac{1}{\log(1-p)^{-1}} \leq \frac{1}{p},$$

where the second inequality is taken from Theorem 6.1 in Achddou et al. [2024]. $\square$

Proof of Corollary 5.4. If $\mathcal{G}$ is complete, then $d(i,j) = 1$ for all $i,j \in \mathcal{N}$, and the gossip time simplifies to $t + L$. Recall Equation (10), with the same steps we can prove that

$$|T_{i,k}(t) - T_{j,k}(t)| \leq KL_p(\delta) = K\left\lceil \frac{\log(\delta)}{\log(1-p)} \right\rceil$$

for any arm $k \in [K]$ and agents $i, j \in [N]$ for any $t \in [T]$.

Hence by the fact that $\lambda_{N-1}(\mathrm{Lap}(\mathcal{G})) = N$ and let L^* as defined in Corollary 5.4 we can prove the case for complete graph.

For the rest two cases, we can just prove which by substituting the exact values of $\lambda_{N-1}(\mathrm{Lap}(\mathcal{G}))$ in Theorem 5.3. $\qquad\square$

Proof of Theorem 5.9. First, note that

$$\mathrm{Reg}_T^\nu(\pi) = \sum_{i \in [N]} \mathrm{Reg}_{i,T}^\nu(\pi) = \sum_{i \in [N]} \sum_{k \in [K]} \mathbb{E}[T_{i,k}(T)]\Delta_k = \sum_{k \in [K]} \Delta_k \sum_{i \in [N]} \mathbb{E}[T_{i,k}(T)].$$

In order to show Theorem 5.9, we only need to prove that

$$\lim_{T \to \infty} \frac{\sum_{i \in [N]} \mathbb{E}[T_{i,k}(T)]}{\log T} \geq \frac{2(1-s)}{(1+\varepsilon)^2 \Delta_k^2} \tag{19}$$

for p and ε and some sub-optimal arm $k \neq k^*$.

Suppose the distribution over instance ν is given by $\mathcal{P} = (\mathbb{P}_{i,a})_{i \in [N], a \in [K]}$. Consider another instance ν' with $\mathcal{P}' = (\mathbb{P}'_{i,a})_{i \in [N], a \in [K]}$ such that

$$\mathbb{P}'_{i,a} = \begin{cases} \mathcal{N}(\mu_{i,a}, 1), \ a \neq k \\ \mathcal{N}(\mu_{i,a} + (1+\varepsilon)\Delta_a, 1), \ a = k \end{cases}$$

for $i \in [N]$ where $\mathbb{P}_{i,a} = \mathcal{N}(\mu_{i,a}, 1)$ and $\varepsilon \in (0, 1]$.

According to the consistency of policy π, it holds that

$$\mathrm{Reg}_T^\nu(\pi) + \mathrm{Reg}_T^{\nu'}(\pi) \leq 2CT^s$$

for some constant C and $p \in (0, 1)$.

Simultaneously, let event $A_i = \left\{ T_{i,k}(T) \geq \frac{T}{2} \right\}$, then

$$\begin{aligned}
\mathrm{Reg}_{i,T}^\nu(\pi) + \mathrm{Reg}_{i,T}^{\nu'}(\pi) &\geq \frac{T}{2} \cdot \Delta_k \cdot \mathbb{P}_{\nu,\pi}[A_i] + \frac{T}{2} \cdot \varepsilon \cdot \Delta_k \cdot \mathbb{P}_{\nu',\pi}[A_i^c] \\
&\geq \frac{T}{2} \varepsilon \Delta_k \left(\mathbb{P}_{\nu,\pi}[A_i] + \mathbb{P}_{\nu',\pi}[A_i^c] \right) \\
&\geq \frac{T}{4} \varepsilon \Delta_k \exp\left(-\mathrm{KL}(\nu, \nu') \right) \\
&= \frac{T}{4} \varepsilon \Delta_k \exp\left(\sum_{j \in [N]} \mathbb{E}[T_{j,k}(T)] \cdot \mathrm{KL}(\mathbb{P}_{j,a}, \mathbb{P}'_{j,a}) \right) \\
&= \frac{T}{4} \varepsilon \Delta_k \exp\left(-\sum_{j \in [N]} \mathbb{E}[T_{j,k}(T)] \cdot \frac{(1+\varepsilon)^2 \Delta_k^2}{2} \right).
\end{aligned}$$

Hence, summing the above inequality for all agents $i \in [N]$, we obtain

$$\mathrm{Reg}_T^\nu(\pi) + \mathrm{Reg}_T^{\nu'}(\pi) \geq \frac{NT}{4} \varepsilon \Delta_k \exp\left(-\sum_{j \in [N]} \mathbb{E}[T_{j,k}(T)] \cdot \frac{(1+\varepsilon)^2 \Delta_k^2}{2} \right). \tag{20}$$

Rearranging the terms in Equation (20), it holds that

$$\begin{aligned}
\sum_{j \in [N]} \mathbb{E}[T_{j,k}(T)] \cdot \frac{(1+\varepsilon)^2 \Delta_k^2}{2} &\geq \log\left(\frac{NT\varepsilon\Delta_k/4}{\mathrm{Reg}_T^\nu(\pi) + \mathrm{Reg}_T^{\nu'}(\pi)} \right) \\
&\geq \log\left(\frac{NT\varepsilon\Delta_k}{8CT^s} \right) = (1-s)\log(T) + \log\left(\frac{N\varepsilon\Delta_k}{8C} \right).
\end{aligned}$$

19

Therefore, we can show

$$\lim_{T \to \infty} \frac{\sum_{i \in [N]} \mathbb{E}[T_{j,k}(T)]}{\log T} \geq \frac{2(1-s)}{(1+\varepsilon)^2 \Delta_k^2}$$

which is the goal in Equation (19). $\qquad\square$

D Estimation of unknown p

Since $\mathcal{G}$ is connected, each agent has at least one neighbor. This allows every agent to estimate the edge activation probability p by observing its connectivity status over multiple rounds. We design the following procedure for each agent $i \in [N]$ to compute an estimate $\hat{p}$ of p.

Algorithm 2 Burn-in Phase for Estimating p by Agent $i \in [N]$

1: **Input:** Confidence level $\delta \in (0,1)$, any fixed neighbor $n_i \in [N]_i$ of agent i in the base graph $\mathcal{G}$
2: **Initialization:** Set $\tilde{\tau}_i \leftarrow 0$, $t \leftarrow 0$, $\hat{p}_i(0) \leftarrow 0$ and $\mathrm{CI}_i(0) \leftarrow \infty$
3: **while** $\hat{p}_i(t) - 3\mathrm{CI}_i(t) \leq 0$ **do**
4:     Increment time $t \leftarrow t + 1$ and observe $\mathcal{N}_i(t)$
5:     **if** $n_i \in [N]_i(t)$ **then**
6:         $\tilde{\tau}_i \leftarrow \tilde{\tau}_i + 1$
7:     **end if**
8:     Select an arm uniformly at random
9:     Update $\hat{p}_i(t) = \frac{\tilde{\tau}_i}{t}$ and $\mathrm{CI}_i(t) = \sqrt{\frac{\log(2/\delta)}{2t}}$
10: **end while**
11: **Output:** $\hat{p}_i = \hat{p}_i(t) - \mathrm{CI}_i(t)$

Lemma D.1. *For the agent i and each t, it holds that*

$$\mathbb{P}\left[|\hat{p}_i(t) - p| \leq \mathrm{CI}_i(t)\right] \geq 1 - \delta.$$

Proof of Theorem D.1. Theorem D.1 comes directly from Hoeffding's inequality. $\qquad\square$

Theorem D.2. *Let $\delta = \frac{2}{T^2}$. Then, with probability at least $1 - \frac{2N}{T}$, the estimate $\hat{p}_i$ returned by Algorithm 2 satisfies $\hat{p}_i \in \left(\frac{p}{2}, p\right]$ for every agent $i \in [N]$. Furthermore, the cumulative regret incurred during the burn-in phase across all agents is bounded by $\mathcal{O}\left(\frac{N \log T}{p^2}\right)$.*

Proof of Theorem D.2. Suppose the end round for Algorithm 2 is t^*. As $\hat{p}_i = \hat{p}_i(t) - \mathrm{CI}_i(t) \leq p$ according to Theorem D.1, we only need to show $\hat{p}_i > \frac{p}{2}$. This can be verified by

$$\hat{p}_i = \hat{p}_i(t) - \mathrm{CI}_i(t) > \frac{\hat{p}_i(t) + \mathrm{CI}_i(t)}{2} \geq \frac{p}{2}.$$

By combining each $t \in [T]$ and $i \in [N]$ and union bound, we can obtain the first part of Theorem D.2. Moreover, for $t^* > \frac{16 \log(T)}{p^2}$, we have

$$\hat{p}_i(t) - 3\mathrm{CI}_i(t) \geq p - 4\mathrm{CI}_i(t) = p - 4\sqrt{\frac{\log T}{t^*}} > 0,$$

hence the stopping condition in Algorithm 2 holds. By the fact that $\Delta_{\max} \leq 1$ for all arms, we can obtain the second part of Theorem D.2 through adding the regret for all agents $i \in [N]$. $\qquad\square$

E Notation

The following notation chart provide notation used throughout the paper.

N	Number of agents.
T	Number of time steps.
K	Number of actions.
W_t	Gossip matrix.
$\lambda_1, \lambda_2, \ldots, \lambda_N$	Eigenvalues of P sorted by norm, i.e. $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_N|$. It is always $\lambda_1 = 1 > |\lambda_2|$.
$\mu_{i,k}$	global mean reward of arm k on agent i.
μ_k	global mean reward of arm k.
Δ_i	Reward gaps, i.e. $\mu_1 - \mu_i$.
p	the link probability p of the Random graph
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	base graph.
$\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$	Communication network at time t.
$\mathcal{N}_i$	the neighbors of agent i in the base graph $\mathcal{G}$
$\mathcal{N}_i(t)$	the neighbors of agent i in the communication network at time t
$T_{i,k}(t)$	Number of times arm k is pulled by node i up to time t.
$X_{i,k}(t)$	Local reward of arm k on agent i at time step t.
$\widehat{\mu}_{i,k}(t)$	Local estimate of $\mu_{i,k}$ on agent i at time step t .
$z_{i,k}(t)$	Global estimate of μ_k on agent i at time step t .
$A_i(t)$	Action played by agent i at time t.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Abstract we have stated our main contributions and scope. Moreover, in Section 1 we provide further details of our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss several limitations of the work in Section A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have stated every necessary assumptions before each theorem and lemma in the work. Detail proofs can be found in Appendix B and C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the information for reproducing the results of the paper in Section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code we used in this paper in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details for our experiments are given in both Section 6 and our supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Although the error bars in our paper are very small, the figures within our paper still include which.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Our experiments can be run just on laptop, and thus our paper do not include any further information on computer resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We believe that our paper correspond to the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Since this is a work focusing on machine learning theory, we do not have any societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: As this is a theory work, we do not have data or models that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: As this is a machine learning theory work, we do not use any existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Since our work focuses on machine learning theory, we do not provide any new asset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: As this is a theory work, we do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: As this is a machine learning theory work, we do not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any crucial components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.