Collaborative Multi-Agent Heterogeneous Multi-Armed Bandits

Ronshee Chawla [1] **Daniel Vial** [1 2] **Sanjay Shakkottai** [1] **R. Srikant** [2]

Abstract

The study of collaborative multi-agent bandits has attracted significant attention recently. In light of this, we initiate the study of a new collaborative setting, consisting of N agents such that each agent is learning one of M stochastic multi-armed bandits to minimize their group cumulative regret. We develop decentralized algorithms which facilitate collaboration between the agents under two scenarios. We characterize the performance of these algorithms by deriving the per agent cumulative regret and group regret upper bounds. We also prove lower bounds for the group regret in this setting, which demonstrates the near-optimal behavior of the proposed algorithms.

1. Introduction

The multi-armed bandit (MAB) problem is a paradigm for sequential decision-making under uncertainty, which involves allocating a resource to an action, in order to obtain a reward. MABs address the tradeoff between exploration and exploitation while making sequential decisions. Owing to their utility in large-scale distributed systems (such as information retrieval (Yue & Joachims, 2009), advertising (Chakrabarti et al., 2008), etc.), an extensive study has been conducted on multi-agent versions of the classical MAB in the last few years. In multi-agent MABs, there are multiple agents learning a bandit and communicating over a network. The goal is to design communication strategies which allow efficient exploration of arms across agents, so that they can perform better than single agent MAB algorithms.

There exist many versions of multi-agent MABs in the literature (see Section 1.2 for an overview). We propose a new collaborative setting where each of the N agents is learning one of M stochastic MABs (where each of the bandits have K arms and $M < N$) to minimize the group cumulative regret, i.e., the sum of individual cumulative regrets of all the agents. We assume that every bandit $m \in [M]$ is learnt by $N_m = \Theta(\frac{N}{M})$ agents. At time $t \in \mathbb{N}$, for all $m \in [M]$, any agent learning the m^{th} bandit plays one of the K arms and receives a stochastic reward, independent of everything else (including the other agents learning the m^{th} bandit pulling the same arm). The network among the agents is denoted by a $N \times N$ gossip matrix G (fixed and unknown to the agents), where every n^{th} row of G ($n \in [N]$) is a probability mass function over the set $[N]$. We investigate our setting under two scenarios: (a) *context unaware*, in which no agent is aware of the other agents learning the same bandit, and (b) *partially context aware*, in which every agent is aware of $r - 1$ other agents learning the same bandit. In the *context unaware* scenario, in addition to the arm pulls conducted at each time, each agent can choose to pull information from another agent who is chosen at random based on the gossip matrix G. In the *partially context aware* scenario, every agent can also choose to exchange messages with the $r - 1$ agents that they know are learning the same bandit, in addition to the arm pulls and the information pulls. Agents behave in a decentralized manner, i.e., the arm pulls, information pulls and messages exchanged are only dependent on their past actions, obtained rewards and the messages received from other agents.

Problem Motivation: Our setting finds its utility in applications involving multiple agents, with possibly differing contexts. While agents with the same context have the same objective, they may not be aware of who else shares the same context. Analyzing the proposed setting is an attempt in figuring out the best way agents should utilize recommendations from others. The quality of the recommendations received from other agents depends on whether those agents share the same context (and hence have the same objective).

For example, consider a social recommendation engine like Yelp, where the N agents correspond to the users, each choosing one from among K pizzerias (arms) in a particular location; this can be modeled as a K-armed MAB. Suppose that each pizzeria serves multiple pizza styles such as Neapolitan, Chicago deep dish, New York style, Detroit style, etc., where each pizza style corresponds to a bandit. Each user has a preference for one of the pizza styles and

[1] Chandra Family Department of Electrical and Computer Engineering, University of Texas, Austin, TX, USA [2] Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL, USA. Correspondence to: Ronshee Chawla <ronsheechawla@utexas.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

is looking for the corresponding best pizzeria. For instance, pizzeria 1 (arm 1) might excel in Chicago deep dish, while pizzeria 2 might be best for Detroit style. If a user (agent) does not know who else has similar tastes (i.e., has the same context), the user would not be able to determine whether a review recommendation is “useful”, especially if the review only states that a pizzeria is excellent, but does not describe the type of pizza that the reviewer consumed. What we have here is a setting where an arm (pizzeria) is being recommended by some user, but for a specific user who is perusing the reviews, it is not clear if that recommendation is actually useful (e.g., the recommendation is actually for a Detroit style pizza, but the user is interested in Chicago deep dish). This example corresponds to the *context unaware* scenario in our model. Moreover, if a user knows another user or a group of users who share the same pizza style preference, they can exchange pizzeria recommendations while still scrolling through the reviews of all the pizzerias by themselves, which corresponds to the *partially context aware* scenario in our model.

1.1. Key Contributions

Algorithms: For the *context unaware* scenario, we modify and subsequently analyze the GosInE algorithm (Algorithm 1), which (Chawla et al., 2020) proposed for the case of a single bandit ($M = 1$). For the *partially context aware* scenario, we utilize the insights obtained from the analysis of Algorithm 1 and propose a new algorithm (Algorithm 3). Algorithm 1 proceeds in phases, such that agents play from a subset of the K arms within a phase, and recommend the arm ID of the best arm during information pulls at the end of a phase. The received arm recommendations are used to update the active sets before the next phase begins. In the *partially context aware* scenario, the agents additionally (a) share the arm recommendation with the $r - 1$ agents known to be learning the same bandit, (b) determine the M most recent unique arm recommendations among their set of r agents, and (c) distribute these recent recommendations among themselves to update their respective active sets.

Gossip despite multiple bandits: Our main analytical contribution is to show that under a mild assumption (Assumption 3.1), agents running Algorithm 1 for the *context unaware* scenario and running Algorithm 3 for the *partially context aware* scenario are able to identify the best arm of the bandit they are learning, and are able to spread it quickly to the other agents learning the same bandit. Even though the outcome is the same as the setting in which agents are collaboratively learning a single bandit (Chawla et al., 2020; Newton et al., 2022), the spreading of the best arms for each of the M bandits is extremely complicated because the M spreading processes are intertwined and evolving simultaneously (since every agent interacts with the agents learning other bandits). Consequently, agents cannot trust the arm recommendations received from the agents learning other bandits. Thus, unlike (Chawla et al., 2020; Newton et al., 2022), the spreading process of each of the M arms isn’t bounded by a standard rumor spreading process, hence requiring an involved analysis (detailed in Section 3.5).

Upper bounds: We show that the expected cumulative regret of an individual agent running Algorithms 1 and 3 scales as $O\left(\frac{\frac{MK}{N}+M}{\Delta_m}\log T\right)$ (Theorem 3.2 and Corollary 3.3) and $O\left(\frac{\frac{MK}{N}+\frac{M}{r}}{\Delta_m}\log T\right)$ (Theorem 4.1 and Corollary 4.2) respectively, for large T, where Δ_m is the minimum arm gap of the m^{th} bandit. It is evident that when $M << \min\{K, N\}$, agents learning the m^{th} bandit for all $m \in [M]$ experience regret reduction compared to the case of no collaborations, because of the distributed exploration of the K arms among $N_m = \Theta(\frac{N}{M})$ agents. Furthermore, the expected group regret incurred by agents running Algorithms 1 and 3 scales as $O\left(\sum_{m\in[M]}\frac{K+N}{\Delta_m}\log T\right)$ (Corollary 3.4) and $O\left(\sum_{m\in[M]}\frac{K+\frac{N}{r}}{\Delta_m}\log T\right)$ (Corollary 4.3) respectively.

Lower bounds: We show in Theorem 5.1 that the expected group regret of our model scales as $\Omega\left(\sum_{m\in[M]}\sum_{k\neq k^*(m)}\frac{1}{\Delta_{m,k}}\log T\right)$ for large T, where $k^*(m)$ is the best arm and $\Delta_{m,k}$ is the arm gap of the k^{th} arm of the m^{th} bandit. This demonstrates that the first terms in our group regret upper bounds are near-optimal; we also show the second terms (which scale as $N \log T$) are unavoidable in general. See Section 5 for details.

1.2. Related Work

Our work falls broadly under the category of cooperative MABs. To the best of our knowledge, the setting considered in this work hasn’t been studied previously. Our work is closest to the line of work in (Sankararaman et al., 2019; Chawla et al., 2020; Newton et al., 2022; Vial et al., 2021; 2022), which involves multiple agents learning the same K-armed MAB and communicating sub-linear number of times (during the entire time horizon) through bit-limited pairwise gossip style communications to minimize individual cumulative regret. (Vial et al., 2021; 2022) considers a multi-agent system with honest and malicious agents, where honest agents are learning the same K-armed MAB. Their algorithms can be used in our setting, where an agent learning some bandit treats agents learning other bandits as malicious. In our setting, an agent running those algorithms incur regret scaling as $O\left(\frac{1}{\Delta_m}\left(\frac{MK}{N}+N\left(1-\frac{1}{M}\right)\right)\log T\right)$ after T time steps. This regret scaling is problematic when $\frac{K}{N} = \Theta(1)$, as there is no benefit of collaboration: the regret scales as $O\left(\frac{K}{\Delta_m}\log T\right)$, which is akin to learning a K-armed MAB without communications. In our work, we show that an agent using the GosInE Algorithm in (Chawla et al., 2020) with a slight modification results in lesser regret in the *context unaware* scenario and is further reduced in the *partially context aware* scenario. Due to space constraints, we refer the reader to Appendix A for other related work.

2. Problem Setup

We consider a multi-agent multi-armed bandit (MAB) setting consisting of N agents, where each agent attempts to learn one of M stochastic MABs (where $M < N$), each containing K arms. Formally, for each $m \in [M]$, let $\mathcal{I}_m \subset [N]$ denote the set of agents learning the m^{th} bandit. We assume that every bandit $m \in [M]$ is learnt by N_m agents, such that $\sum_{m \in [M]} N_m = N$ and $c_1 \frac{N}{M} \leq N_m \leq c_2 \frac{N}{M}$ for all $m \in [M]$, where $\frac{M}{N} < c_1 \leq c_2 \leq \frac{M}{2}$ are known absolute constants. For every bandit $m \in [M]$, the K arms have unknown mean rewards, denoted by $\{\mu_{m,k}\}_{k \in [K]}$, where $\mu_{m,k} \in [0,1]$ for all $k \in [K]$. Let $k^*(m)$ denote the best arm for the m^{th} bandit, i.e., $k^*(m) = \arg\max_{k \in [K]} \mu_{m,k}$, and assume that $\mu_{m,k^*(m)} > \mu_{m,k}$ for all $k \neq k^*(m)$. We define $\mathcal{B}$ to be the set of M best arms, i.e., $\mathcal{B} = \{k^*(m)\}_{m \in [M]}$ and $\mathcal{B}^{(-m)} = \mathcal{B} \backslash \{k^*(m)\}$. Let $\Delta_{m,k} := \mu_{m,k^*(m)} - \mu_{m,k}$ denote the arm gaps for all $k \neq k^*(m)$ and $m \in [M]$. The assumption on the arm means implies that $\Delta_{m,k} \in [0,1]$ for all $k \neq k^*(m)$. Let Δ_m denote the minimum arm gap of the m^{th} bandit, i.e., $\Delta_m = \min_{k \in [K] \backslash k^*(m)} \Delta_{m,k}$.

For all $m \in [M]$, each agent $i \in \mathcal{I}_m$ at time $t \in [T]$ pulls an arm $I_t^{(i)} \in [K]$ and receives a reward $X_t^{(i)}(I_t^{(i)})$, where $X_t^{(i)}(k) = \mu_{m,k} + \delta_t^{(i)}$ for each $k \in [K]$ and $\delta_t^{(i)}$ is 1-subgaussian noise (independent of everything else).

The network between the agents is represented by a $N \times N$ gossip matrix G (fixed and unknown to the agents), where each row of the matrix $G(n, .)$ denotes a probability distribution for all $n \in [N]$. In this work, we consider that the agents are connected by a complete graph, i.e., for all $n \in [N]$, $G(n,i) = (N-1)^{-1}$ for all $i \neq n$.

The problem setup is investigated under two scenarios:

(i) *Context Unaware* - No agent $i \in [N]$ knows which other agents are learning the same bandit.

(ii) *Partially Context Aware* - For all bandits $m \in [M]$, each agent i learning the m^{th} bandit knows $r-1$ other agents (where $1 < r \leq \min_{m' \in [M]} N_{m'}$) who are also learning the m^{th} bandit, such that N_m is an integral multiple of r for all $m \in [M]$[1].

In the *context unaware* scenario, after pulling an arm, agents can choose to receive a message from another agent through an information pull. In particular, if an agent $n \in [N]$ decides to pull information, it does so by contacting another agent $i \in [N]$ according to the probability distribution $G(n, .)$, independently of everything else. The agent i who is contacted is then allowed to communicate $\lceil \log_2 K \rceil$ number of bits during this information pull.

In the *partially context aware* scenario, in addition to the information pulls allowed in the *context unaware* scenario, each agent learning the m^{th} bandit can also exchange messages with the $r-1$ other agents who they know are also learning the m^{th} bandit. As a result, agents in this scenario are allowed to communicate $r\lceil \log_2 K \rceil$ number of bits during information pulls.

Agents operate in a decentralized fashion, i.e., all the decisions that an agent makes can solely depend on their past actions, rewards and the messages received from other agents during information pulls. Moreover, decisions made by agents during the information pulling slots (i.e., what to communicate if asked for information) are allowed to be dependent on the arm pulls in those slots.

Under both the scenarios, we would like to leverage collaboration between the agents to minimize the expected group cumulative regret, i.e., the sum of the individual cumulative regrets for all the agents. Mathematically, let $I_t^{(i)}$ denote the arm pulled by agent $i \in [N]$ at time $t \in [T]$ and $c(i)$ denote the index of the (unknown) bandit that agent i is trying to learn, i.e., if $i \in \mathcal{I}_m$, $c(i) = m$. Then, the cumulative regret of an agent $i \in [N]$ after playing for T time steps is denoted by $R_T^{(i)} := \sum_{t=1}^{T} (\mu_{c(i),k^*(c(i))} - \mu_{c(i),I_t^{(i)}})$ and the expected group cumulative regret is given by $\mathbb{E}[\text{Reg}(T)]$[2], where $\text{Reg}(T) = \sum_{i=1}^{N} R_T^{(i)}$.

3. *Context Unaware* Algorithm

For the *context unaware* scenario, we consider the GosInE Algorithm in (Chawla et al., 2020) with a slight modification (Algorithm 1). Subsequently, we demonstrate that under a mild assumption (Assumption 3.1), agents running Algorithm 1 incur less regret compared to the case when they are learning their bandit without collaboration, despite being unaware of the other agents learning the same bandit. Furthermore, we will show that this regret is near-optimal by stating a lower bound in Section 5.

3.1. Key Algorithmic Principles

The GosInE Algorithm in (Chawla et al., 2020) has the following key components:

Phases - The algorithm proceeds in phases $j \in \mathbb{N}$, such that during a phase, agents play from a set $S_j^{(i)} \subset [K]$,

[1] since $N = \sum_{m \in [M]} N_m$, N is an integral multiple of r

[2] the expectation is with respect to the rewards, communications and the algorithm

also known as active set. At the end of a phase, agents exchange arm recommendations through pairwise gossip communication and update their active sets.

Active Sets - The active set for any agent $i \in [N]$ is a combination of two sets: (i) the time-invariant sticky set, denoted by $\widehat{\mathcal{S}}^{(i)} \subset [K]$, (ii) the non-sticky set. As the name suggests, the sticky set $\widehat{\mathcal{S}}^{(i)} \subset S_j^{(i)}$ for all $j \in \mathbb{N}$, i.e., it is always present in the active set. The non-sticky set contains two arms at all times, which are updated across the phases through arm recommendations received from other agents.

Arm Recommendations - At the end of phase j, every agent contacts another agent at random based on the network's gossip matrix (which in this work is a complete graph), and the contacted agent sends the "best" arm in their active set. After receiving the recommendation, agents update their active set by adding the recommended arm to their sticky set and discarding the "worst" performing arm out of the two non-sticky arms. We provide an efficient modification to this update rule in Algorithm 1, and provide the details about what it means to be the "best" and the "worst" performing arm in the next sub-section.

3.2. Algorithm Description

We provide a detailed description of the GosInE Algorithm with an efficient modification to the active set updates at the end of a phase, with the pseudo-code in Algorithm 1.

Initialization: The algorithm is initialized with the following inputs - (i) the standard exploration parameter of the UCB Algorithm, denoted by $\alpha > 0$, (ii) the parameter $\beta > 1$ which controls the length of the phases, where a phase j runs from the (discrete) time instants $1 + A_{j-1}, \cdots, A_j$ with $A_j := \lceil j^\beta \rceil$, and (iii) a sticky set $\widehat{\mathcal{S}}^{(i)}$ of cardinality S for each agent i. Note that the phases grow longer as the algorithm progresses (since $A_j - A_{j-1} = \Theta(j^{\beta-1})$ and $\beta > 1$). Details regarding the size of the sticky set are provided in the remarks at the end of this sub-section. For the first phase ($j = 1$), we initialize the active set of each agent to be their sticky set, i.e., $S_1^{(i)} = \widehat{\mathcal{S}}^{(i)}$.

UCB within a phase: We denote $T_k^{(i)}(t)$ to be the number of times agent i has played an arm k up to and including time t, and $\widehat{\mu}_k^{(i)}(t)$ to be the empirical mean reward among those plays, i.e., $\widehat{\mu}_k^{(i)}(t) = \frac{1}{T_k^{(i)}(t)} \sum_{s \le t : I_s^{(i)} = k} X_s^{(i)}(I_s^{(i)})$ [3]. In phase j, every agent i plays UCB Algorithm on their active set $S_j^{(i)}$, i.e., for $t \in \{1 + A_{j-1}, \cdots, A_j\}$, the chosen arm $I_t^{(i)} = \arg\max_{k \in S_j^{(i)}} \widehat{\mu}_k^{(i)}(t-1) + \sqrt{\frac{\alpha \log T}{T_k^{(i)}(t-1)}}$.

Arm recommendation at the end of a phase: The arm recommendation received by agent i when $t = A_j$ is denoted by $\mathcal{O}_j^{(i)} \in [K]$. During the information pull request at time $t = A_j$, every agent shares the ID of the most played arm in their active set during phase j, which is what we refer to as the "best" performing arm. Similarly, the "worst" performing arm in the active set during a phase refers to the non-sticky arm that was played the least number of times in that phase. The intuition behind sharing the most played arm during a phase is that for large time horizons, UCB chooses to play the best arm more than any other arm (Bubeck et al., 2011). Therefore, if the true best arm is present in the active set, it will be recommended at the end of a phase as the algorithm progresses and the phases grow longer.

Active set update for the next phase: We update the active set in a more efficient manner (Newton et al., 2022) compared to the GosInE Algorithm in (Chawla et al., 2020). Specifically, GosInE uses the update $S_{j+1}^{(i)} = \widehat{\mathcal{S}}^{(i)} \cup \{U_j^{(i)}\} \cup \{\mathcal{O}_j^{(i)}\}$, where $U_j^{(i)}$ is the most played non-sticky arm in phase j, i.e., $U_j^{(i)} = \arg\max_{k \in S_j^{(i)} \setminus \widehat{\mathcal{S}}^{(i)}} T_k^{(i)}(A_j) - T_k^{(i)}(A_{j-1})$. In contrast, we use the update $S_{j+1}^{(i)} = \widehat{\mathcal{S}}^{(i)} \cup \{\widehat{\mathcal{O}}_j^{(i)}\} \cup \{\mathcal{O}_j^{(i)}\}$, where $\widehat{\mathcal{O}}_j^{(i)} = \arg\max_{k \in S_j^{(i)}} T_k^{(i)}(A_j) - T_k^{(i)}(A_{j-1})$ is the most played among *all* arms (not just the non-sticky arms). As observed in (Newton et al., 2022), the latter update ensures that once the best arm spreads, the active set becomes $\widehat{\mathcal{S}}^{(i)} \cup \{k^*\}$ thereafter, where k^* is the true best arm. This reduces the cardinality of the active set by up to two arms if $k^* \in \widehat{\mathcal{S}}^{(i)}$, subsequently reducing the regret in the single bandit case. As our analysis will show, a similar regret reduction is possible for our setting.

Algorithm 1 (at agent i)

```
Input: UCB Parameter $\alpha > 0$, phase parameter $\beta > 1$,
sticky set $\widehat{\mathcal{S}}^{(i)}$ with $|\widehat{\mathcal{S}}^{(i)}| = S \le K - 2$
Initialize $A_j = \lceil j^\beta \rceil$, $j \leftarrow 1$, $S_1^{(i)} \leftarrow \widehat{\mathcal{S}}^{(i)}$
for $t \in \mathbb{N}$ do
    Play $I_t^{(i)} = \arg\max_{k \in S_j^{(i)}} \widehat{\mu}_k^{(i)}(t-1) + \sqrt{\frac{\alpha \log T}{T_k^{(i)}(t-1)}}$
    if $t == A_j$ then
        $\mathcal{O}_j^{(i)} \leftarrow$ GetRec$(i, j)$                    (Algorithm 2)
        $\widehat{\mathcal{O}}_j^{(i)} \leftarrow \arg\max_{k \in S_j^{(i)}} T_k^{(i)}(A_j) - T_k^{(i)}(A_{j-1})$
        $S_{j+1}^{(i)} \leftarrow \widehat{\mathcal{S}}^{(i)} \cup \{\widehat{\mathcal{O}}_j^{(i)}\} \cup \{\mathcal{O}_j^{(i)}\}$
        $j \leftarrow j + 1$
    end if
end for
```

3.3. Assumption on the Sticky Set

It is possible that during the initialization of Algorithm 1, none of the agents learning a particular bandit have the best

[3] $\mu_k^{(i)}(t) = 0$ if $T_k^{(i)}(t) = 0$

Algorithm 2 Arm Recommendation

```
Input: Agent $i \in [N]$, phase $j \in \mathbb{N}$
function GetRec$(i, j)$
    $n \sim G(i, .)$ (sample a neighbor)
    return $\widehat{\mathcal{O}}_j^{(n)}$ (most played arm by agent $n$ in phase $j$)
end function
```

arm $k^*(m)$ in their sticky sets, i.e., $k^*(m) \notin \widehat{S}^{(i)}$ for all $i \in \mathcal{I}_m$ for some $m \in [M]$. This will result in all agents learning that bandit to incur linear regret. In order to avoid this situation, we will follow (Chawla et al., 2020; Vial et al., 2021; 2022; Newton et al., 2022; Sankararaman et al., 2019) and make a mild assumption:

Assumption 3.1. For all $m \in [M]$, there exists an agent $i_m^* \in \mathcal{I}_m$ such that $k^*(m) \in \widehat{S}^{(i_m^*)}$.

Remarks:

1. In fact, if $S = \left\lceil \frac{MK}{N} \log \frac{M}{\gamma} \right\rceil$ for some $\gamma \in (0, 1)$ and we construct $\widehat{S}^{(i)}$ for each i by sampling S arms independently and uniformly at random from K arms, then Assumption 3.1 holds with probability at least $1 - \gamma$. We prove this claim as Proposition D.1 in Appendix D. This choice of S, scaling as $\frac{MK}{N} = \frac{K}{\left(\frac{N}{M}\right)}$, implies that for every bandit $m \in [M]$, the K arms are equally distributed across the set of $N_m = \Theta(\frac{N}{M})$ agents $\mathcal{I}_m$ with high probability, ensuring that every arm remains active for some agent learning that bandit.

2. We can alternatively define the set of arms to be those present among their sticky sets of agents to avoid Assumption 3.1 altogether. In such a scenario, agents learning a particular bandit will learn and spread the best arm among the arms in their sticky sets.

3.4. Regret Guarantee

Theorem 3.2 characterizes the performance of Algorithm 1.

Theorem 3.2. *Consider a system of $N \geq 2$ agents connected by a complete graph (for each $i \in [N]$, $G(i, n) = (N-1)^{-1} \forall n \neq i$) and learning one of the $M \geq 2$ bandits with $K \geq 2$ arms, satisfying Assumption 3.1. Let the UCB parameter $\alpha > 10$ and the phase parameter $\beta > 2$. Then, the regret incurred by an agent $i \in \mathcal{I}_m$ running Algorithm 1 for each $m \in [M]$ after T time steps is bounded by:*

$$\mathbb{E}[R_T^{(i)}] \leq \lceil (\tau^*)^\beta \rceil + (K + g)\frac{\pi^2}{3} + g_{\text{spr}} + \sum_{k \in \{\widehat{S}^{(i)} \cup \mathcal{B}^{(-m)}\} \setminus \{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}} \log T,$$

where $\tau^* = 2\max\{2, \max_{m \in [M]} \tau_m^*\}$, $\tau_m^* = \inf\left\{j \in \mathbb{N} : \frac{A_j - A_{j-1}}{S+2} \geq 1 + \frac{4\alpha}{\Delta_m^2} \log A_j\right\}$, $g = \frac{N(2^\beta+1)2^{\beta(\frac{\alpha}{2}-3)}(S+1)}{\frac{\alpha}{2}-3}\binom{K}{2}$, g_{spr} *scales as* $O\left(M^{\beta+1}\left(\left(\log \frac{N}{M}\right)^2 \log\left(\log \frac{N}{M}\right)\right)^\beta\right)$ *and* $O(.)$ *only hides the absolute constants.*

Remarks:

1. Scaling of τ^* - Proposition C.4 in Appendix C implies that $\tau^* = O\left(\frac{S}{\Delta^2}\right)^{\frac{1}{\beta-2}}$, where $\Delta = \min_{m \in [M]} \Delta_m$. This is expected, because it takes longer for the best arm to be identified and spread for bandits with the smaller gaps.

2. Regret scaling - Essentially, Theorem 3.2 says that the regret of any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ scales as $O\left(\frac{S+M}{\Delta_m} \log T\right)$ for T large.

3. Regret guarantee for arbitrary values of arm means - The result in Theorem 3.2 can be easily extended for arm means not restricted to $[0, 1]$, as assumed above. The only modification that occurs is the following: the terms $K + g$, g_{spr} and $\lceil (\tau^*)^\beta \rceil$ will be multiplied by $\widetilde{\Delta}_m := \max_{k \in [K] \setminus k^*(m)} \Delta_{m,k}$. It is noteworthy that the modification only affects the second-order term in regret.

4. Benefit of collaboration - We have the following corollary when the K arms are equally distributed across all the agents learning the same bandit, i.e., when $S = \Theta\left(\frac{MK}{N}\right)$.

Corollary 3.3. *When $S = \Theta\left(\frac{MK}{N}\right)$, the regret incurred by an agent $i \in \mathcal{I}_m$ for each $m \in [M]$ after T time steps scales as $O\left(\frac{1}{\Delta_m}\left(\frac{MK}{N} + M\right) \log T\right)$.*

Corollary 3.3 implies that when $S = \Theta\left(\frac{MK}{N}\right)$ and $M << \min\{K, N\}$, agents experience regret reduction compared to the case of no collaborations.

5. Group Regret - Corollary 3.4 quantifies the performance of Algorithm 1 in terms of the group regret.

Corollary 3.4. *For all bandits $m \in [M]$, when $\{\widehat{S}^{(i)}\}_{i \in \mathcal{I}_m}$ is a partition of the set of K arms, i.e., $\widehat{S}^{(i_1)} \cap \widehat{S}^{(i_2)} = \phi$ for $i_1 \neq i_2 \in \mathcal{I}_m$ and $\cup_{i \in \mathcal{I}_m} \widehat{S}^{(i)} = [K]$, the group regret* $\text{Reg}(T)$ *of the system playing Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{E}[\text{Reg}(T)] \leq &\sum_{m \in [M]} \sum_{k \in [K] \setminus \{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ &+ c_2 \frac{N}{M} \sum_{m \in [M]} \sum_{k \in \mathcal{B}^{(-m)}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ &+ N\lceil (\tau^*)^\beta \rceil + N(K+g)\frac{\pi^2}{3} + N g_{\text{spr}}.\end{aligned}$$

Essentially, Corollary 3.4 implies expected group regret scales as $O\left(\sum_{m \in [M]} \frac{K+N}{\Delta_m} \log T\right)$ for large T.

3.5. Proof Sketch (Theorem 3.2)

The proof of Theorem 3.2 is detailed in Appendix C and we highlight the key ideas here. Akin to the scenario of learning the single bandit in (Chawla et al., 2020), we first show the existence of a random phase τ, after which all the agents starting from phase τ contain the best arm of the bandit they are learning in their respective active sets and recommend it during information pulls (Claim 3). This allows us to decompose the expected cumulative regret incurred by an agent into two parts: the regret up to phase τ and the regret after phase τ. The regret after phase τ is the regret incurred by playing the UCB algorithm on the sticky set and the M best arms, because agents recommend their respective best arms during information pulls post phase τ.

The technical challenge lies in bounding the expected cumulative regret up to phase τ. In particular, we prove a generalization of the result in (Chawla et al., 2020) that irrespective of the bandit an agent is learning, the probability of an agent not recommending their best arm and thus dropping it from their active set at the end of a phase is small and decreases as the phases progress, such that it doesn't happen infinitely often (Lemma C.2). This is a consequence of agents playing UCB Algorithm on their active sets during a phase and the fact that UCB chooses to play the best arm more often than any other arm for large time horizons (Bubeck et al., 2011). This implies the existence of a random phase (denoted by τ_{stab}), after which agents aware of their best arm (i.e., agents with the relevant best arm in their active set) will always recommend it moving forward.

Our analysis differs significantly from the analysis in (Chawla et al., 2020) post phase τ_{stab}, where we characterize the time $\tau_{\text{spr},m}$ required for the best arm of the m^{th} bandit to spread via recommendations to the agents $\mathcal{I}_m$ learning that bandit. By definition of τ_{stab}, we know that if an agent $i_1 \in \mathcal{I}_m$ contacts another agent $i_2 \in \mathcal{I}_m$ who is aware of the true best arm $k^*(m)$ after phase τ_{stab}, then i_1 will become informed of the true best arm as well. This arm spreading process therefore resembles a rumor process, where one agent (i^*_m in Assumption 3.1) initially knows the rumor (i.e., the best arm), and any agent who contacts someone aware of the rumor becomes informed itself.

However, our rumor process is extremely complicated compared to the process for the single bandit case in (Chawla et al., 2020). This is because we have M intertwined rumor spreading processes evolving simultaneously: every agent can interact with agents learning a different bandit, and post phase τ_{stab}, agents recommend whichever of the M rumors (i.e., best arms) is relevant to them. Hence, none of these M rumor spreading processes are standard rumor spreading processes (unlike (Chawla et al., 2020)), so analyzing them directly is infeasible.

To tackle this issue, we disentangle the M intertwined processes via a coupling argument. In particular, we define M independent *noisy* rumor processes and show via coupling that the spreading times of these processes upper bound the time of the actual arm spreading processes. The m^{th} of the noisy rumor processes, denoted by $\{\bar{\mathcal{R}}_{m,j}\}_{j=0}^{\infty}$, unfolds on the subgraph of agents $\mathcal{I}_m$ learning the m^{th} bandit and tracks the agents $\bar{\mathcal{R}}_{m,j}$ aware of the m^{th} rumor at phase j. Initially, only i^*_m is aware, i.e., $\bar{\mathcal{R}}_{m,0} = \{i^*_m\}$ by Assumption 3.1. In each phase $j \in \mathbb{N}$, each agent $i \in \mathcal{I}_m$ contacts another agent $ag \in \mathcal{I}_m$ uniformly at random. If $ag \in \bar{\mathcal{R}}_{m,j-1}$ (ag is aware of the rumor), then $i \in \bar{\mathcal{R}}_{m,j}$ (i becomes aware as well) subject to Bernoulli(η) noise, where $\eta = \frac{N_m - 1}{N-1}$. Therefore, i becomes aware with probability $|\bar{\mathcal{R}}_{m,j-1} \cap \mathcal{I}_m|\eta/N_m \leq |\bar{\mathcal{R}}_{m,j-1} \cap \mathcal{I}_m|/(N-1)$. Observe that the right hand side of the inequality is equal the probability with which agent i becomes aware of the best arm in the real process (since in the real process, i contacts $ag \in [N] \setminus \{i\}$ uniformly at random), which allows us to upper bound the spreading time via coupling as discussed above. Thereafter, we further couple the noisy processes to a noiseless one as in (Vial et al., 2022), then use an existing bound for the noiseless setting (Chierichetti et al., 2010).

4. *Partially Context Aware* Algorithm

From Corollary 3.3, it can be noticed that Algorithm 1 incurs additional regret scaling of $O\left(\frac{M}{\Delta_m}\log T\right)$ after T time steps. This is because agents playing Algorithm 1 have the best arm in their active sets and recommend it during information pulls after a random phase τ. Hence, in the absence of knowledge about which other agents are learning the same bandit, any agent playing Algorithm 1 is unable to determine this information with certainty, due to the random nature of τ. One can think of ways in which agents can stop communicating with the agents not learning the same bandit, for example, the blocking approaches considered in (Vial et al., 2021; 2022). These works considered agents learning a single bandit collaboratively in the presence of adversarial agents. However, such blocking strategies incur worse regret: under the assumptions of Corollary 3.3, both (Vial et al., 2021; 2022) incur regret of $O\left(\frac{1}{\Delta_m}\left(\frac{MK}{N} + N\left(1 - \frac{1}{M}\right)\right)\log T\right)$ for large T. This regret scaling is problematic when $\frac{K}{N} = \Theta(1)$, as there is no benefit of collaboration: the regret scales as $O\left(\frac{K}{\Delta_m}\log T\right)$, which is akin to learning a single agent regret.

Given that agents are collaborative in our setting in the sense that they divide the exploration of sub-optimal arms in addition to identifying their best arm, Algorithm 1 has a structure to it: post phase τ, there are only M rumors (corresponding to the M best arms) flowing through the network. If each agent is aware of $r-1$ other agents learning the same bandit, they can distribute the exploration of the received arm recommendations as follows: after receiving the arm recommendations, each group of r agents learning the same bandit can select the M most recent unique arm recommendations from all the recommendations they have received so far and divide them (almost) equally among themselves. The intuition is that post phase τ, given that there are only M rumors flowing through the network, we know from the coupon collector problem that after a (finite) random number of phases post phase τ, the M most recent unique arm recommendations will be the M best arms, and it will stay that way from then on. This reduces the additional regret due to arm recommendations by a factor of r.

Algorithm 3 (at agent i)

```
Input: UCB Parameter $\alpha > 0$, phase parameter $\beta > 1$, sticky set $\widehat{S}^{(i)}$ with $|\widehat{S}^{(i)}| = S \leq K - 2 - \lceil \frac{M}{r} \rceil$, the set $f(i)$ of agents known to be learning the same bandit
Initialize $A_j = \lceil j^{\beta} \rceil$, $j \leftarrow 1$, $S_1^{(i)} \leftarrow \widehat{S}^{(i)}$, $\mathrm{rec}(ag) = \{\}$ for all $ag \in i \cup f(i)$.
for $t \in \mathbb{N}$ do
  Play $I_t^{(i)} = \arg\max_{k \in S_j^{(i)}} \widehat{\mu}_k^{(i)}(t-1) + \sqrt{\frac{\alpha \log T}{T_k^{(i)}(t-1)}}$
  if $t == A_j$ then
    $\mathcal{O}_j^{(i)} \leftarrow \mathrm{GetRec}(i, j)$                (Algorithm 2)
    $\widehat{\mathcal{O}}_j^{(i)} \leftarrow \arg\max_{k \in S_j^{(i)}} T_k^{(i)}(A_j) - T_k^{(i)}(A_{j-1})$
    $\mathrm{rec}(i) \leftarrow \mathrm{rec}(i) \cup \{(j, \mathcal{O}_j^{(i)})\}$
    Obtain $\mathcal{O}_j^{(ag)}$ from all $ag \in f(i)$ to maintain the set of arm recommendations $\mathrm{rec}(ag) \leftarrow \mathrm{rec}(ag) \cup \{(j, \mathcal{O}_j^{(ag)})\}$ for all $ag \in f(i)$
    $\mathrm{uniquerec}(i, f(i), j) \leftarrow M$ most recent unique arm recommendations in $\bigcup_{ag \in \{\{i\} \cup f(i)\}} \mathrm{rec}(ag)$
    if $\mathrm{uniquerec}(i, f(i), j) \neq \mathrm{uniquerec}(i, f(i), j-1)$ then
      $\mathrm{sortrec}(i, f(i), j) \leftarrow$ elements of $\mathrm{uniquerec}(i, f(i), j)$ in the ascending order
      $\widetilde{S}_j^{(i)} \leftarrow \mathrm{DivideRec}(i, j, f(i), \mathrm{sortrec}(i, f(i), j)))$ (Algorithm 4)
      $S_{j+1}^{(i)} \leftarrow \widehat{S}^{(i)} \cup \{\widehat{\mathcal{O}}_j^{(i)}\} \cup \{\mathcal{O}_j^{(i)}\} \cup \{\widetilde{S}_j^{(i)}\}$
    else
      $S_{j+1}^{(i)} \leftarrow S_j^{(i)}$
    end if
    $j \leftarrow j + 1$
  end if
end for
```

We use the intuition in the previous paragraph and propose Algorithm 3, which builds upon Algorithm 1 and uses the following extra input: for each $m \in [M]$ and $i \in \mathcal{I}_m$, let $f(i)$ satisfy: (i) $f(i) \subset \mathcal{I}_m$, (ii) $|f(i)| = r - 1$, and (iii) if $n \in f(i)$, then $i \in f(n)$. Thus, $f(i)$ consists of $r - 1$ other agents learning the same bandit who are known to agent i. Algorithm 3 divides the M most recent unique arm recommendations among r agents in $\{i\} \cup f(i)$ using the subroutine DivideRec, described in Algorithm 4. DivideRec can be best understood for the case when $\frac{M}{r}$ is a (positive) integer. For example, if $M = 6$ and $r = 3$, each agent in the set $\{i\} \cup f(i)$ will get 2 arm IDs from $\mathrm{sortrec}(i, f(i), .)$. Suppose that i is the second smallest element in the sorted version of $\{i\} \cup f(i)$ ($pos(i) = 2$), it will get the third and the fourth entries in $\mathrm{sortrec}(i, f(i), .)$.

It is important to note that the array $\mathrm{sortrec}(i, f(i), .)$ is identical for all $ag \in \{i\} \cup f(i)$. This observation is crucial in dividing the M most recent unique recommendations among the r agents in $\{i\} \cup f(i)$, without violating the constraint on the number of communication bits per agent.

Remarks:

1. Constructing $\widetilde{S}_j^{(\cdot)}$ when $|\mathrm{uniquerec}(., f(.), j)| < M$ **-** when enough phases haven't elapsed such that there are less than M most recent unique arm recommendations, we can construct $\widetilde{S}_j^{(\cdot)}$ with $\lceil \frac{|\mathrm{uniquerec}(., f(.), j)|}{r} \rceil$ elements.

2. Satisfying the communication bit constraint of $r\lceil \log_2 K \rceil$ **bits -** Each agent $i \in \mathcal{I}_m$ uses $\lceil \log_2 K \rceil$ bits to receive an arm recommendation via an information pull, and uses $\lceil \log_2 K \rceil$ bits per agent to obtain the arm recommendations received by $r - 1$ agents in $f(i)$.

Algorithm 4 Dividing M most recent unique arm recommendations

```
Input: Agent $i \in [N]$, phase $j \in \mathbb{N}$, sets $f(i)$ and $\mathrm{sortrec}(i, f(i), j))$
function DivideRec($i, j, f(i), \mathrm{sortrec}(i, f(i), j))$)
  $ags \leftarrow$ elements of $\{i\} \cup f(i)$ in the ascending order
  $pos(i) \leftarrow$ position of $i$ in the array $ags$
  $\widetilde{S}_j^{(i)} \leftarrow$ elements of $\mathrm{sortrec}(i, f(i), j))$ in the positions $\Big\{ \big((pos(i) - 1) \lceil \frac{M}{r} \rceil \mod M\big) + 1, \cdots, \big(pos(i) \lceil \frac{M}{r} \rceil - 1 \mod M\big) + 1 \Big\}$
  return $\widetilde{S}_j^{(i)}$
end function
```

4.1. Regret Guarantee

Theorem 4.1 characterizes the performance of Algorithm 3.

Here, for any bandit $m \in [M]$, $k_{m,1}, k_{m,2}, \cdots, k_{m,K}$ denotes the order statistics of the arm means, i.e., $k_{m,1} = k^*(m)$ and $\mu_{m,k_{m,1}} > \mu_{m,k_{m,2}} \geq \cdots \geq \mu_{m,k_{m,K}}$.

Theorem 4.1. *Under the assumptions of Theorem 3.2, the regret incurred by an agent $i \in \mathcal{I}_m$ running Algorithm 3 for* *each $m \in [M]$ after T time steps is bounded by:*

$$\begin{aligned}\mathbb{E}[R_T^{(i)}] \leq & \sum_{k \in \widehat{S}^{(i)} \setminus \{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ & + \sum_{k \in \{k_{m,l}\}_{l=2}^{\lceil \frac{M}{r} \rceil + 2}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ & + \lceil (j^*)^\beta \rceil + (K + \widehat{g}) \frac{\pi^2}{3} + \widehat{g}_{\mathrm{spr}} + \widehat{g}_{\mathrm{rec}},\end{aligned}$$

where $j^* = 3 \max\{2, \max_{m \in [M]} j_m^*\}$,
$j_m^* = \inf \left\{ j \in \mathbb{N} : \frac{A_j - A_{j-1}}{S + 2 + \lceil \frac{M}{r} \rceil} \geq 1 + \frac{4\alpha}{\Delta_m^2} \log A_j \right\}$,
$\widehat{g} = \frac{N(3^\beta + 1) 3^{\beta(\frac{\alpha}{2} - 3)} (S + 1 + \lceil \frac{M}{r} \rceil)}{\frac{\alpha}{2} - 3} \binom{K}{2 + \lceil \frac{M}{r} \rceil}$,
$\widehat{g}_{\mathrm{rec}} = \frac{N}{r} \left(\lceil (3M)^\beta \rceil + 2 \left(\frac{3}{(\frac{c_1}{M} - \frac{1}{N}) r} \right)^\beta \frac{M}{(1 - \frac{c_1}{M})^r} \Gamma(\beta + 1) \right)$,
$\Gamma(z) = \int_{t=0}^\infty t^{z-1} e^{-t} \, \mathrm{d}t$ *for* $z > 0$, $\widehat{g}_{\mathrm{spr}}$ *scales as* $O\left(M^{\beta+1} \left(\left(\log \frac{N}{M} \right)^2 \log \left(\log \frac{N}{M} \right) \right)^\beta \right)$ *and* $O(.)$ *only hides the absolute constants.*

Remarks:

1. Scaling of j^* - j^* scales just like τ^* in Theorem 3.2, except with S replaced by $S + \frac{M}{r}$. Proposition E.5 in Appendix E formalizes this scaling.

2. Regret Scaling - Essentially, Theorem 4.1 says that the regret of any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ scales as $O\left(\frac{S + \frac{M}{r}}{\Delta_m} \log T \right)$ for large T.

3. Benefit of collaboration - We have the following corollary when the K arms are equally distributed across all the agents learning the same bandit, i.e., when $S = \Theta\left(\frac{MK}{N}\right)$.

Corollary 4.2. *When $S = \Theta\left(\frac{MK}{N}\right)$, the regret incurred by an agent $i \in \mathcal{I}_m$ for each $m \in [M]$ after T time steps scales as $O\left(\frac{1}{\Delta_m} \left(\frac{MK}{N} + \frac{M}{r} \right) \log T \right)$.*

It is clear from Theorem 4.1 and Corollary 4.2 that agents having knowledge of $r - 1$ other agents learning the same bandits results in lesser regret, compared to the *context unaware* scenario where agents aren't aware of the other agents are learning the same bandit.

4. Group Regret - Corollary 4.3 quantifies the performance of Algorithm 1 in terms of the group regret.

Corollary 4.3. *For all bandits $m \in [M]$, when $\{\widehat{S}^{(i)}\}_{i \in \mathcal{I}_m}$ is a partition of the set of K arms, i.e., $\widehat{S}^{(i_1)} \cap \widehat{S}^{(i_2)} = \phi$ for $i_1 \neq i_2 \in \mathcal{I}_m$ and $\cup_{i \in \mathcal{I}_m} \widehat{S}^{(i)} = [K]$, the group regret* $\mathrm{Reg}(T)$ *of the system playing Algorithm 3 satisfies*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}(T)] \leq & \sum_{m \in [M]} \sum_{k \in [K] \setminus \{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ & + c_2 \frac{N}{M} \sum_{m \in [M]} \sum_{k \in \{k_{m,l}\}_{l=2}^{\lceil \frac{M}{r} \rceil + 2}} \frac{4\alpha}{\Delta_{m,k}} \log T \\ & + N \lceil (j^*)^\beta \rceil + N(K + \widehat{g}) \frac{\pi^2}{3} + N \widehat{g}_{\mathrm{spr}} + N \widehat{g}_{\mathrm{rec}}.\end{aligned}$$

Essentially, Corollary 4.3 implies that agents running Algorithm 3 incur expected group regret scaling as $O\left(\sum_{m \in [M]} \frac{K + \frac{N}{r}}{\Delta_m} \log T \right)$ for large T.

4.2. Proof Sketch (Theorem 4.1)

Similar to the regret analysis of Algorithm 1, we first show the existence of (finite) random phases: (i) τ_{stab}, such that if agents have the best arm in their active sets, that will be their most played arm and recommended henceforth during information pulls, and (ii) additional τ_{spr} phases post the phase τ_{stab}, after which all the agents have their best arms in their active sets. The proof for showing that the random phases τ_{stab} and τ_{spr} are finite proceeds identically to the proof for Theorem 3.2.

Moreover, for Algorithm 3, we also show the existence of additional (finite) random phases post the phase $\tau_{\mathrm{stab}} + \tau_{\mathrm{spr}}$, denoted by τ_{rec}, after which the M most recent unique arm recommendations is equal to the set of M best arms from then onwards. As a consequence of the active set updates in Algorithm 3, the active sets of agents remain unchanged in all the subsequent phases and freeze like the GosInE Algorithm in (Chawla et al., 2020), which helps improve the regret by distributing the exploration of M best arms across $r - 1$ other agents learning the same bandit.

Remark: This freezing does not happen in Algorithm 1, where the active sets are still time-varying post phase τ and the randomness of τ prevents agents from gathering information about others learning the same bandit with certainty.

5. Lower Bounds

We state lower bounds for Gaussian noise with unit variance. As is standard, we restrict to *uniformly efficient* policies, i.e., those that ensure small regret on any problem instance. In our case, instances are defined by the arm means $\mu = \{\mu_{m,k}\}_{m \in [M], k \in [K]}$ of the M bandits and the partition $\mathcal{I} = \{\mathcal{I}_m\}_{m \in [M]}$ of the N agents to the M bandits. Policies are called uniformly efficient if $\mathbb{E}[\mathrm{Reg}_{\mu,\mathcal{I}}(T)] = o(T^\gamma)$ for any $\gamma \in (0, 1)$ and any instance $(\mu, \mathcal{I})$.

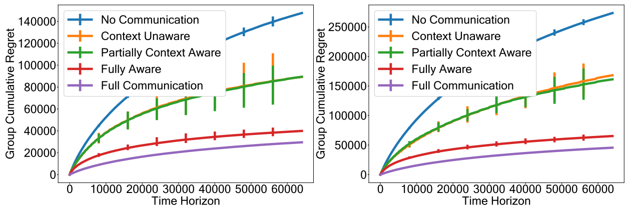

Figure 1. (K, M, N, r) are $(20, 5, 25, 5)$ and $(30, 6, 36, 6)$ respectively. Arm means are in $[0, 1)$ and the UCB parameter $\alpha = 15$.

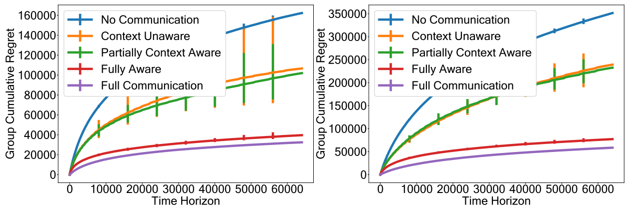

Figure 2. (K, M, N, r) are $(20, 5, 25, 5)$ and $(30, 6, 36, 6)$ respectively. Arm means are in $[2, 4)$ and the UCB parameter $\alpha = 30$.

Theorem 5.1. *For any uniformly efficient policy,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_{\mu,\mathcal{I}}(T)]}{\log T} \geq \sum_{m\in[M]} \sum_{k\in[K]\setminus\{k^*(m)\}} \frac{2}{\Delta_{m,k}}.$$

Theorem 5.1 is proved in Appendix F, by reducing our model to the setting of (Réda et al., 2022). The result shows that the first terms in Corollaries 3.4 and 4.3 are optimal up to constant factors. Hence, for large T, the suboptimality of our upper bounds is due to the second terms, which grow linearly in N and logarithmically in T. Under a further assumption on μ, we can show these dependencies are unavoidable. Again, see Appendix F for a proof.

Theorem 5.2. *Let $(\mu, \mathcal{I})$ be any instance satisfying $\mu_{m,k^*(m)} = \mu_{m+1,k^*(m)}$ and $k^*(m) \neq k^*(m+1)$ for each $m \in [M-1]$.*[4] *Then for any uniformly efficient policy in the context unaware scenario,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_{\mu,\mathcal{I}}(T)]}{\log T} \geq 2\sum_{m=1}^{M-1} |\mathcal{I}_m|\Delta_m \geq N\Delta.$$

6. Numerical Results

We evaluate Algorithm 1 in the *context unaware* setting and Algorithm 3 in the *partially context aware* setting, and verify their insights through synthetic simulations. The algorithms are compared with respect to the following benchmarks: (a) no communication, corresponding to all the agents running UCB-α algorithm on K-armed MAB from (Auer et al., 2002) without any communications, (b) fully aware, corresponding to agents playing GosInE algorithm from (Chawla et al., 2020), but agents only communicate with all the other agents learning the same bandit, and (c) full communication, where for each bandit, all agents play the UCB-α algorithm on K-armed MAB from (Auer et al., 2002), but with the entire history of all arms pulled and the corresponding rewards obtained by all the agents.

We show the group cumulative regret of Algorithms 1 and 3 over 30 random runs with 95% confidence intervals. The K arm means for each of the M bandits are generated uniformly at random from $[0, 1)$ in Figure 1 and $[2, 4)$ in Figure 2. We assume an equal number ($\frac{N}{M}$) of agents learning each bandit, set $\beta = 3$ and the size of the sticky set $S = \frac{MK}{N}$ in these simulations. The UCB parameter α is set to 15 in Figure 1 and 30 in Figure 2.

From our simulations, it is evident that Algorithms 1 and 3 incur lower regret than the case when agents don't communicate, despite limited communication among the agents and agents interacting with agents learning other bandits. Furthermore, our simulations also demonstrate that for each bandit, when an agent knows other agents learning the same bandit in the *partially context aware* scenario, it incurs lesser regret compared to the *context unaware* scenario.

Acknowledgements

This work was supported by NSF Grants 2207547, 1934986, 2106801, 2019844, 2112471, and 2107037; ONR Grant N00014-19-1-2566; the Machine Learning Lab (MLL) at UT Austin; and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Avner, O. and Mannor, S. Concurrent bandits and cognitive radio networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 66–81. Springer, 2014.

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

[4] Such instances $\mu \in [0, 1]^{M\times K}$ exist; for example, if $\mu_{m,k} = ((m-1) + \mathbf{1}(m = k))/M$, where $\mathbf{1}$ is the indicator function.

Boursier, E. and Perchet, V. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Buccapatnam, S., Tan, J., and Zhang, L. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2605–2613. IEEE, 2015.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. Mortal multi-armed bandits. *Advances in neural information processing systems*, 21, 2008.

Chakraborty, M., Chua, K. Y. P., Das, S., and Juba, B. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pp. 164–170, 2017.

Chawla, R., Sankararaman, A., Ganesh, A., and Shakkottai, S. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3471–3481. PMLR, 2020.

Chawla, R., Sankararaman, A., and Shakkottai, S. Multi-agent low-dimensional linear bandits. *IEEE Transactions on Automatic Control*, 2022.

Chierichetti, F., Lattanzi, S., and Panconesi, A. Almost tight bounds for rumour spreading with conductance. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 399–408, 2010.

Dakdouk, H., Féraud, R., Laroche, R., Varsier, N., and Maillé, P. Collaborative exploration and exploitation in massively multi-player bandits. 2021.

Dubey, A. and Pentland, A. t. S. Differentially-private federated linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6003–6014, 2020.

Dubey, A. et al. Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, pp. 2740–2750. PMLR, 2020.

Hillel, E., Karnin, Z. S., Koren, T., Lempel, R., and Somekh, O. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Kalathil, D., Nayyar, N., and Jain, R. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

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

Korda, N., Szorenyi, B., and Li, S. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pp. 1301–1309. PMLR, 2016.

Lalitha, A. and Goldsmith, A. Bayesian algorithms for decentralized stochastic bandits. *IEEE Journal on Selected Areas in Information Theory*, 2(2):564–583, 2021.

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

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

Liu, K. and Zhao, Q. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing*, 58(11):5667–5681, 2010.

Liu, L. T., Mania, H., and Jordan, M. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1618–1628. PMLR, 2020.

Madhushani, U. and Leonard, N. When to call your neighbor? strategic communication in cooperative stochastic bandits. *arXiv preprint arXiv:2110.04396*, 2021.

Mansour, Y., Slivkins, A., and Wu, Z. S. Competing bandits: Learning under competition. *arXiv preprint arXiv:1702.08533*, 2017.

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

Newton, C., Ganesh, A., and Reeve, H. Asymptotic optimality for decentralised bandits. *ACM SIGMETRICS Performance Evaluation Review*, 49(2):51–53, 2022.

Réda, C., Vakili, S., and Kaufmann, E. Near-optimal collaborative learning in bandits. In *Advances in Neural Information Processing Systems*, 2022.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

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

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

Szorenyi, B., Busa-Fekete, R., Hegedus, I., Ormándi, R., Jelasity, M., and Kégl, B. Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*, pp. 19–27. PMLR, 2013.

Tekin, C. and Van Der Schaar, M. Distributed online learning via cooperative contextual bandits. *IEEE transactions on signal processing*, 63(14):3700–3714, 2015.

Vial, D., Shakkottai, S., and Srikant, R. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 161–170, 2021.

Vial, D., Shakkottai, S., and Srikant, R. Robust multi-agent bandits over undirected graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6 (3):1–57, 2022.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208, 2009.

Zhu, Z., Zhu, J., Liu, J., and Liu, Y. Federated bandit: A gossiping approach. *Proc. ACM Meas. Anal. Comput. Syst.*, 5(1), 2021.

A. Related Work Revisited

Apart from the works mentioned in Section 1.2, there exist several other works in the space of collaborative multi-agent MABs with different models of communication among agents, some of which are shared here. Agents exchange information in (Buccapatnam et al., 2015; Chakraborty et al., 2017) via broadcasting instead of pairwise gossip style communications. There is more frequent communication between the agents in (Kolla et al., 2018; Lalitha & Goldsmith, 2021; Martínez-Rubio et al., 2019). In (Madhushani & Leonard, 2021) the number of communications is inversely proportional to the minimum arm gap so could be large. Arm mean estimates are exchanged instead of arm indices in (Landgren et al., 2016). (Wang et al., 2020) employs a leader-follower mechanism, wherein the leader explores the arms and estimates their mean rewards, while the followers play the arm with the highest estimated arm mean based on the samples collected by the leader.

Collaborative multi-agent bandits have also been studied in the contextual setting (Dubey et al., 2020; Tekin & Van Der Schaar, 2015) and the linear reward model setting (Dubey & Pentland, 2020; Chawla et al., 2022; Korda et al., 2016), which are significantly different from the setting considered in our work. Some of the works in this space also focus on minimizing the simple regret (Hillel et al., 2013; Szorenyi et al., 2013), instead of minimizing the cumulative regret. There also exist works such as (Bistritz & Leshem, 2018; Shahrampour et al., 2017; Zhu et al., 2021; Réda et al., 2022) which assume different bandits across agents. While this list of works is by no means exhaustive, it represents a broad class of settings studied in the collaborative multi-agent bandits thus far.

Another long line of work in multi-agent MABs comprises of collision-based models (Anandkumar et al., 2011; Avner & Mannor, 2014; Bistritz & Leshem, 2018; Boursier & Perchet, 2019; Dakdouk et al., 2021; Kalathil et al., 2014; Liu & Zhao, 2010; Liu et al., 2020; Mansour et al., 2017; Rosenski et al., 2016), where if multiple agents play the same arm, they receive a small or no reward. However, our work assumes that if multiple agents learning the same bandit play the same arm, they receive independent rewards and thus, is different than the collision-based models.

B. Notations Revisited

Recall that for every bandit $m \in [M]$ and arm $k \in [K]$, the (unknown) mean rewards are denoted by $\mu_{m,k}$, $k^*(m)$ denotes the best arm, $\Delta_{m,k} := \mu_{m,k^*(m)} - \mu_{m,k}$ denotes the arm gap, $\Delta_m := \min_{k \neq k^*(m)} \Delta_{m,k}$ is the minimum arm gap, $\mathcal{I}_m \subset [N]$ denotes the set of agents learning the m^{th} bandit and $c(.) : [N] \to [M]$ is a function mapping the set of agents to the set of bandits, i.e., if $i \in \mathcal{I}_m$, $c(i) = m$. Let $\mathcal{B}$ denote the set of all the M best arms $\{k^*(m)\}_{m \in [M]}$ and $\mathcal{B}^{(-m)} = \mathcal{B} \backslash \{k^*(m)\}$ denote the set of best arms excluding the best arm of the m^{th} bandit. For every agent $i \in [N]$ and phase $j \in \mathbb{N}$, $\widehat{\mathcal{O}}_j^{(i)} \in S_j^{(i)}$ denotes the most played arm by agent i in phase j (where $S_j^{(i)}$ is the set of active arms for agent i in phase j), and subsequently recommended at the end of the phase if requested by another agent through information pull. Each phase j lasts from $t \in \{1 + A_{j-1}, \cdots, A_j\}$ and

$$A^{-1}(t) = \sup\{j \in \mathbb{N} : A_j \leq t\}. \tag{1}$$

For $A_j = \lceil j^{\beta} \rceil$, $A^{-1}(t) = \lfloor t^{\frac{1}{\beta}} \rfloor$.

C. Proof of Theorem 3.2

We extend the proof ideas from (Chawla et al., 2020) and (Vial et al., 2022) to prove our results. Fix an agent $i \in [N]$ and phase $j \in \mathbb{N}$. Let $\chi_j^{(i)}$ be a boolean random variable associated with the event $\left\{k^*(c(i)) \in S_j^{(i)}, \widehat{\mathcal{O}}_j^{(i)} \neq k^*(c(i))\right\}$, i.e.,

$$\chi_j^{(i)} = \mathbf{1}\left(k^*(c(i)) \in S_j^{(i)}, \widehat{\mathcal{O}}_j^{(i)} \neq k^*(c(i))\right), \tag{2}$$

indicating whether agent i does not recommend the best arm of the bandit it plays at the end of the phase j, *if* it is present in its active set $S_j^{(i)}$. We also extend the definitions of random times defined in (Chawla et al., 2020), which will capture the key aspects in the system dynamics, as well as highlight the differences from the single bandit case, as follows:

$$\begin{aligned}
\tau_{\text{stab}}^{(i)} &= \inf\{j \geq \tau^* : \forall l \geq j, \chi_l^{(i)} = 0\}, \\
\tau_{\text{stab}} &= \max_{i \in [N]} \tau_{\text{stab}}^{(i)}, \\
\tau_{\text{spr}}^{(i)} &= \inf\{j \geq \tau_{\text{stab}} : k^*(c(i)) \in S_j^{(i)}\} - \tau_{\text{stab}}, \\
\tau_{\text{spr},m} &= \max_{i \in \mathcal{I}_m} \tau_{\text{spr}}^{(i)}, \\
\tau_{\text{spr}} &= \max_{m \in [M]} \tau_{\text{spr},m}, \\
\tau &= \tau_{\text{stab}} + \tau_{\text{spr}}.
\end{aligned}$$

$\tau_{\text{stab}}^{(i)}$ is the earliest phase, following which if agent i has their best arm in its active set, will play it most number of times and recommend it during information pulls requested by other agents. Furthermore, starting from phase τ, all the agents contain their best arms in their respective active sets, such that it will also be their most played arm and hence, will recommend it at the end of any phase $j \geq \tau$. Mathematically, it implies the following: for all $i \in [N]$,

$$k^*(c(i)) \in S_j^{(i)}, \widehat{\mathcal{O}}_j^{(i)} = k^*(c(i)) \; \forall j \geq \tau. \tag{3}$$

Claim (3) can be shown by induction as follows: it is evident from the definition of $\tau_{\text{spr}}^{(i)}$ that $k^*(c(i)) \in S^{(i)}_{\tau_{\text{stab}}+\tau_{\text{spr}}^{(i)}}$. Furthermore, $\widehat{\mathcal{O}}_j^{(i)} = k^*(c(i))$ for any phase $j \geq \tau_{\text{stab}}$ if $k^*(c(i)) \in S_j^{(i)}$. Therefore, $\widehat{\mathcal{O}}^{(i)}_{\tau_{\text{stab}}+\tau_{\text{spr}}^{(i)}} = k^*(c(i))$ the update step of the Algorithm 1 ensures that $k^*(c(i)) \in S^{(i)}_{\tau_{\text{stab}}+\tau_{\text{spr}}^{(i)}+1}$, and hence claim (3) follows.

We now highlight the similarities and differences of the random times defined in this work with respect to the random times defined in (Chawla et al., 2020). τ_{stab} is defined exactly the same as in (Chawla et al., 2020), because we will demonstrate in Lemma C.3 that the bound on the tail probability of the random variable $\tau_{\text{stab}}^{(i)}$ is independent of the bandit played by an agent. However, in contrast to (Chawla et al., 2020), given that the set of agents are learning M different bandits, the spread of M best arms is non-trivial (compared to spreading a single best arm), because agents learning different bandits are communicating with each other, hence we have M intertwined spreading processes occuring simultaneously. Consequently, we simply cannot bound the spreading time $\tau_{\text{spr},m}$ for each of the M best arms by the spreading time of a standard rumor spreading process, unlike (Chawla et al., 2020). This necessitates coupling the M intertwined rumor spreading processes happening in our model to a set of M independent fictitious noisy rumor spreading process happening on the subgraph of the agents for each bandit.

C.1. Intermediate Results

We begin by providing a roadmap for the proof of Theorem 3.2 and then proving the intermediate results needed to complete it. Claim (3) states that all the agents starting from phase τ contain the best arm of the bandit they are learning in their respective active sets and recommend it during information pulls. This allows us to decompose the expected cumulative regret incurred by an agent into two parts: the regret up to phase τ and the regret after phase τ.

We first show that the expected cumulative regret up to phase τ is bounded by a constant that depends only on the system parameters (number of agents, number of bandits and their respective arm gaps) and independent of the time horizon T (Proposition C.1). It follows from the observation that the probability of an agent not recommending their best arm and thus dropping it from their active set at the end of a phase is small and decreases as the phases progress, such that it doesn't happen infinitely often (Lemma C.2). This implies the existence of a random phase (defined by τ_{stab}), after which agents always recommend and never drop their respective best arms. Post phase τ_{stab}, we characterize the time taken by the best arms of each of the M bandits to spread across their respective agents (denoted by $\tau_{\text{spr},m}$). Unlike (Chawla et al., 2020), we cannot bound the spreading time $\{\tau_{\text{spr},m}\}_{m \in [M]}$ of each of the M best arms through a standard rumor spreading process. This is because each agent communicates with either other agents learning the same bandit or the agents learning different bandits. For each $m \in [M]$, $\tau_{\text{spr},m}$ is bounded by multiple layers of stochastic domination, described in the order in which they are applied below:

- first, reducing the system to the case when only one agent learning the m^{th} bandit is aware of their best arm $k^*(m)$

- second, lower bounding the spreading time of $k^*(m)$ through a fictitious noisy rumor spreading process unfolding on the subgraph of the agents learning the m^{th} bandit (described in the proof sketch of Theorem 3.2)
- third, coupling the fictitious noisy rumor spreading process in the subgraph of the agents learning the m^{th} bandit in the previous layer with a fictitious noiseless process (described in Appendix C.2)

The last two layers of stochastic domination are absent in (Chawla et al., 2020), as all the agents are playing the same bandit. The preceding discussion characterizes the regret up to phase τ.

We subsequently show that the expected cumulative regret incurred by an agent after phase τ is bounded by the regret due to the arms in their sticky set and the regret due to the other $M-1$ best arms (Proposition C.1). This is a consequence of all the agents recommending their respective best arms after phase τ (Claim (3)) and every agent communicating with agents learning other bandits.

Proposition C.1. *The expected cumulative regret of any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ after playing for T time steps is bounded by:*

$$\mathbb{E}[R_T^{(i)}] \leq \mathbb{E}[A_\tau] + \frac{\pi^2}{3}K + \sum_{k \in \{\widehat{\mathcal{S}}^{(i)} \cup \mathcal{B}^{(-m)}\} \setminus \{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}} \log T.$$

Proof. Fix a bandit $m \in [M]$ and an agent $i \in \mathcal{I}_m$. By regret decomposition principle, we know that

$$\begin{aligned} R_T^{(i)} &= \sum_{k \in [K]} \Delta_{m,k} \sum_{t=1}^{T} \mathbf{1}(I_t^{(i)} = k), \\ &= \sum_{t=1}^{T} \sum_{k \in [K]} \Delta_{m,k} \mathbf{1}(I_t^{(i)} = k), \\ &\leq A_\tau + \sum_{k \in [K]} \Delta_{m,k} \sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k), \end{aligned}$$

where $I_t^{(i)}$ is the arm played by agent i at time t and the last step follows from the fact that $\Delta_{m,k} \in (0,1)$. Taking expectation on both sides, we get

$$\mathbb{E}[R_T^{(i)}] \leq \mathbb{E}[A_\tau] + \sum_{k \in [K]} \Delta_{m,k} \mathbb{E}\left[\sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k)\right]. \tag{4}$$

The first term $\mathbb{E}[A_\tau]$ is bounded in Proposition C.3. We now bound the second term. Let $X_{k,t}^{(i)} = \mathbf{1}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) \leq \frac{4\alpha}{\Delta_{m,k}^2} \log T\right)$ and $Y_{k,t}^{(i)} = \mathbf{1}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right)$. Then, the inner sum in the second term can be re-written as follows:

$$\sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k) = \sum_{t=1}^{T} (X_{k,t}^{(i)} + Y_{k,t}^{(i)}). \tag{5}$$

We first bound the sum $\sum_{t=1}^{T} \mathbb{E}[Y_{k,t}^{(i)}]$. Notice that

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[Y_{k,t}^{(i)}] &= \sum_{t=1}^{T} \mathbb{P}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right), \\ &\leq \sum_{t=1}^{T} \mathbb{P}\left(I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right), \\ &\leq \sum_{t=1}^{T} 2t^{2-\frac{\alpha}{2}} \leq \frac{\pi^2}{3}, \end{aligned} \tag{6}$$

where we substitute the classical estimate of the probability that UCB plays a sub-optimal arm using Chernoff-Hoeffding bound for 1-subgaussian rewards and $\alpha > 10$ in the last line.

For an arm $k \in [K]$, we bound the sum $\sum_{t=1}^{T} \mathbb{E}[X_{k,t}^{(i)}]$ and complete the proof. By taking the expectation over $X_{k,t}^{(i)}$, we get

$$\sum_{t=1}^{T} \mathbb{E}[X_{k,t}^{(i)}] = \sum_{t=1}^{T} \mathbb{P}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) \le \frac{4\alpha}{\Delta_{m,k}^2} \log T \right). \tag{7}$$

For any arm $k \in [K]$, three cases are possible:

- if $k \in \widehat{\mathcal{S}}^{(i)}$, i.e., if arm k is one of the sticky sub-optimal arms, the sum in equation (7) is bounded above by $\frac{4\alpha}{\Delta_{m,k}^2} \log T$, This is because the event $I_t^{(i)} = k$ cannot occur more than $\frac{4\alpha}{\Delta_{m,k}^2} \log T$ times, otherwise it will contradict the event $T_k^{(i)}(t-1) \le \frac{4\alpha}{\Delta_{m,k}^2} \log T$.
- if arm k is a non-sticky arm in the active set, it has to be either the best arm $k^*(m)$ or one of the other best arms from the set $\mathcal{B}^{(-m)}$. This follows from Claim (3), as starting from phase τ, agents will recommend their respective best arms during information pulls. Given that every agent communicates with agents learning other bandits, each of the arms in $\mathcal{B}^{(-m)}$ could be present in agent i's active set after phase τ. Therefore, for all $k \in \mathcal{B}^{(-m)}$, the sum in equation (7) is bounded by $\frac{4\alpha}{\Delta_{m,k}^2} \log T$, which follows from the same argument used for a sticky sub-optimal arm.
- if arm k is neither a sticky sub-optimal arm nor one of the other best arms from the set $\mathcal{B}^{(-m)}$, the event $I_t^{(i)} = k$ cannot happen, because arm k cannot be present in the active set after phase τ from Claim (3). Thus, the sum in equation (7) is equal to zero.

From the above observations, we can conclude that for all $k \in \{\widehat{\mathcal{S}}^{(i)} \cup \mathcal{B}^{(-m)}\} \backslash \{k^*(m)\}$,

$$\sum_{t=1}^{T} \mathbb{E}[X_{k,t}^{(i)}] \le \frac{4\alpha}{\Delta_{m,k}^2} \log T. \tag{8}$$

The proof of Proposition C.1 is completed by substituting equations (6) and (8) into the expectation of the equation (5), followed by substituting the bound on the expectation of the equation (5) into equation (4). □

In order to obtain an upper bound on $\mathbb{E}[A_\tau]$, we first show that the probability of the error event that the best arm is not recommended during information pull at the end of the phase j (indicated by $\chi_j^{(i)}$ in equation (2)) decreases as the phases progress. This result is stated as a lemma below:

Lemma C.2. *For any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ and phase $j \in \mathbb{N}$ such that $\frac{A_j - A_{j-1}}{S+2} \ge 1 + \frac{4\alpha}{\Delta_m^2} \log A_j$, we have*

$$\mathbb{E}[\chi_j^{(i)}] \le \frac{2}{\frac{\alpha}{2} - 3} \binom{K}{2} (S+1) \frac{1}{A_{j-1}^{\frac{\alpha}{2}-3}}.$$

Proof. The proof of this lemma is identical to the proof of Lemma 6 in (Chawla et al., 2020), except that it is re-written in a general form here (in (Chawla et al., 2020), $S := |\widehat{\mathcal{S}}^{(i)}| = \lceil \frac{K}{N} \rceil$) and uses 1-subgaussian rewards instead of Bernoulli rewards. □

Proposition C.3. *The regret up to phase τ is bounded by*

$$\mathbb{E}[A_\tau] \le \lceil (\tau^*)^\beta \rceil + \frac{N(2^\beta + 1) 2^{\beta(\frac{\alpha}{2}-3)}}{\frac{\alpha}{2} - 3} \binom{K}{2} (S+1) \frac{\pi^2}{3} + \sum_{m \in [M]} \mathbb{E}[A_{2\tau_{\mathrm{spr},m}}],$$

where τ^* is defined in Theorem 3.2.

Proof. The random variable τ has support in $\mathbb{N}$. For $\mathbb{N}$-valued random variables, we know that

$$\begin{aligned}
\mathbb{E}[A_\tau] &= \sum_{t\geq 1}\mathbb{P}(A_\tau \geq t),\\
&\overset{(a)}{\leq} \sum_{t\geq 1}\mathbb{P}(\tau \geq A^{-1}(t)),\\
&\leq \sum_{t\geq 1}\mathbb{P}(\tau_{\text{stab}}+\tau_{\text{spr}} \geq A^{-1}(t)),\\
&\leq \sum_{t\geq 1}\mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right) + \sum_{t\geq 1}\mathbb{P}\left(\tau_{\text{spr}} \geq \frac{A^{-1}(t)}{2}\right),\\
&= \sum_{t\geq 1}\mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right) + \sum_{t\geq 1}\mathbb{P}\left(\bigcup_{m=1}^{M}\left(\tau_{\text{spr},m} \geq \frac{A^{-1}(t)}{2}\right)\right),\\
&\leq \sum_{t\geq 1}\mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right) + \sum_{m\in[M]}\sum_{t\geq 1}\mathbb{P}\left(\tau_{\text{spr},m} \geq \frac{A^{-1}(t)}{2}\right),\\
&\leq A_{\tau^*} + \sum_{t\geq A_{\tau^*}+1}\mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right) + \sum_{m\in[M]}\mathbb{E}[A_{2\tau_{\text{spr},m}}],
\end{aligned} \tag{9}$$

where we use the definition of $A^{-1}(.)$ defined in equation (1) in step (a). Unlike (Chawla et al., 2020), we cannot bound the spreading time $\{\tau_{\text{spr},m}\}_{m\in[M]}$ of each of the M best arms through a standard rumor spreading process. Instead, we bound $\mathbb{E}[A_{2\tau_{\text{spr},m}}]$ for all $m \in [M]$ in Appendix C.3 by stochastically dominating the random variable $\tau_{\text{spr},m}$ through a fictitious noisy rumor spreading process, which is described in Section 3.5 and proved in Proposition C.5.

Here, we focus on bounding the sum $\sum_{t\geq A_{\tau^*}+1}\mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right)$. We will calculate $\mathbb{P}\left(\tau_{\text{stab}} \geq x\right)$ for some fixed $x \geq \frac{\tau^*}{2}$ using Lemma C.2 and then bound the sum in the previous sentence.

$$\begin{aligned}
\mathbb{P}\left(\tau_{\text{stab}} \geq x\right) &\overset{(a)}{=} \mathbb{P}\left(\bigcup_{i\in[N]}\left(\tau_{\text{stab}}^{(i)} \geq x\right)\right),\\
&\leq \sum_{i=1}^{N}\mathbb{P}\left(\tau_{\text{stab}}^{(i)} \geq x\right),\\
&\overset{(b)}{=} \sum_{i=1}^{N}\mathbb{P}\left(\bigcup_{l\geq x}\left(\chi_l^{(i)}=1\right)\right),\\
&\leq \sum_{i=1}^{N}\sum_{l\geq x}\mathbb{P}\left(\chi_l^{(i)}=1\right),\\
&\overset{(c)}{\leq} \sum_{i=1}^{N}\sum_{l\geq x}\frac{2}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\frac{1}{A_{l-1}^{\frac{\alpha}{2}-3}},\\
&\leq \sum_{l\geq x}\frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\frac{1}{A_{l-1}^{\frac{\alpha}{2}-3}},
\end{aligned}$$

where we used the definitions of τ_{stab} and $\tau_{\text{stab}}^{(i)}$ in the steps (a) and (b) respectively, and Lemma C.2 in step (c) because it holds for any phase $j \geq \frac{\tau^*}{2}$ by definition of τ^*. Therefore,

$$
\begin{aligned}
\sum_{t \geq A_{\tau^*}+1} \mathbb{P}\left(\tau_{\text{stab}} \geq \frac{A^{-1}(t)}{2}\right) &\leq \sum_{t \geq A_{\tau^*}+1} \sum_{l \geq \frac{A^{-1}(t)}{2}} \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\frac{1}{A_{l-1}^{\frac{\alpha}{2}-3}}, \\
&\overset{(a)}{\leq} \sum_{l \geq \frac{\tau^*}{2}} \sum_{t=A_{\tau^*}+1}^{A_{2l}} \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\frac{1}{A_{l-1}^{\frac{\alpha}{2}-3}}, \\
&\leq \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\sum_{l \geq \frac{\tau^*}{2}} \frac{A_{2l}}{A_{l-1}^{\frac{\alpha}{2}-3}}, \\
&\overset{(b)}{=} \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\sum_{l \geq \frac{\tau^*}{2}} \frac{\lceil (2l)^\beta \rceil}{\lceil (l-1)^\beta \rceil^{\frac{\alpha}{2}-3}}, \\
&\overset{(c)}{\leq} \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\sum_{l \geq \frac{\tau^*}{2}} \frac{(2l)^\beta + 1}{(l-1)^{\beta(\frac{\alpha}{2}-3)}}, \\
&\overset{(d)}{\leq} \frac{2N}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)(2^\beta+1)2^{\beta(\frac{\alpha}{2}-3)}\sum_{l \geq \frac{\tau^*}{2}} \frac{l^\beta}{l^{\beta(\frac{\alpha}{2}-3)}}, \\
&\overset{(e)}{\leq} \frac{2N(2^\beta+1)2^{\beta(\frac{\alpha}{2}-3)}}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\sum_{l \geq 2} \frac{1}{l^{\beta(\frac{\alpha}{2}-4)}}, \\
&\leq \frac{N(2^\beta+1)2^{\beta(\frac{\alpha}{2}-3)}}{\frac{\alpha}{2}-3}\binom{K}{2}(S+1)\frac{\pi^2}{3}.
\end{aligned}
$$

Step (a) follows by re-writing the range of summations. In step (b), we use $A_j = \lceil j^\beta \rceil$. Step (c) uses the property of ceiling function: $x \leq \lceil x \rceil < x+1$. Steps (d) and (e) use the fact that $l-1 \geq \frac{l}{2}$ for all $l \geq 2$ and $\tau^* \geq 4$. The last step follows under the assumption that $\alpha > 10$ and $\beta > 2$.

Substituting the above bound in equation (9) completes the proof of Proposition C.3. □

Proposition C.4. *τ_m^* defined in Theorem 3.2 is bounded by*

$$\tau_m^* \leq 2 + \left(\frac{1}{\beta} + \left(\frac{1}{\beta} + \frac{8\alpha}{\Delta_m^2}\right)(S+2)\right)^{\frac{1}{\beta-2}}.$$

Proof. From Theorem 3.2,

$$\tau_m^* = \inf\left\{ j \in \mathbb{N} : \frac{A_j - A_{j-1}}{S+2} \geq 1 + \frac{4\alpha}{\Delta_m^2}\log A_j \right\}.$$

For $A_j = \lceil j^\beta \rceil$ and $j \geq 2 + \left(\frac{1}{\beta} + \left(\frac{1}{\beta} + \frac{8\alpha}{\Delta_m^2}\right)(S+2)\right)^{\frac{1}{\beta-2}}$,

$$
\begin{aligned}
1 + \frac{4\alpha}{\Delta_m^2}\log A_j &= 1 + \frac{4\alpha}{\Delta_m^2}\log\lceil j^\beta \rceil, \\
&\overset{(a)}{\leq} 1 + \frac{4\alpha}{\Delta_m^2}\log(j^\beta + 1), \\
&\leq 1 + \frac{4\alpha}{\Delta_m^2}\log(2j^\beta), \\
&= 1 + \frac{4\alpha}{\Delta_m^2}\log 2 + \frac{4\alpha\beta}{\Delta_m^2}\log j \leq 1 + \frac{8\alpha\beta}{\Delta_m^2}\log j,
\end{aligned}
$$

where we use $\lceil x \rceil < x + 1$ for any $x \in \mathbb{R}$ in step (a) and $j \geq 2 + \left(\frac{1}{\beta} + \left(\frac{1}{\beta} + \frac{8\alpha}{\Delta_m^2}\right)(S+2)\right)^{\frac{1}{\beta-2}} > 2$ in the last step. Therefore,

$$\begin{aligned} 1 + (S+2)\left(1 + \frac{4\alpha}{\Delta_m^2}\log A_j\right) &\leq 1 + \left(1 + \frac{8\alpha\beta}{\Delta_m^2}\log j\right)(S+2)\log j, \\ &\leq \beta\left(\frac{j-1}{\beta} + \left(\frac{j-1}{\beta} + \frac{8\alpha}{\Delta_m^2}(j-1)\right)(S+2)\right), \\ &\leq \beta(j-1)(j-2)^{\beta-2} \leq \beta(j-1)^{\beta-1}, \end{aligned}$$

by assumption on j and $\log j \leq j - 1$. Furthermore,

$$\begin{aligned} A_j - A_{j-1} &= \lceil j^\beta \rceil - \lceil (j-1)^\beta \rceil, \\ &\geq j^\beta - (j-1)^\beta - 1, \\ &= \beta \widehat{j}^{\beta-1} - 1 \text{ for some } \widehat{j} \in (j-1, j), \\ &\geq \beta(j-1)^{\beta-1} - 1, \end{aligned}$$

by Lagrange's Mean Value Theorem. Thus, we have shown $1 + \frac{4\alpha}{\Delta_m^2}\log A_j \leq \frac{\beta(j-1)^{\beta-1}-1}{S+2} \leq \frac{A_j - A_{j-1}}{S+2}$. This completes the proof of Proposition C.4. □

C.2. Coupling the Noisy Rumor Spreading Process with the Noiseless Process

Proposition C.5. *The random variable $\tau_{\mathrm{spr},m}$ is stochastically dominated by $\bar{\tau}_{\mathrm{spr},m}$ for all $m \in [M]$.*

Proof. Follows from the construction of the fictitious noisy rumor spreading process in Section 3.5. □

Let G_m denote the gossip matrix of the subgraph of the agents learning the m^{th} bandit. In our work, G_m is a complete graph of size N_m, i.e., $G_m(i,n) = (N_m - 1)^{-1}$ for all $i \in \mathcal{I}_m$ and $n \in \mathcal{I}_m \backslash \{i\}$. Similar to (Vial et al., 2022), we begin by defining the variables pertinent to the fictitious noisy rumor spreading process, followed by coupling the fictitious noisy rumor spreading process with a fictitious noiseless process unfolding on G_m. The fictitious noiseless process is defined in the same way as the fictitious noisy process in Section 3.5, but with $\eta = 1$.

Definition C.6. For each $i \in \mathcal{I}_m$, let $\{Y_j^{(i)}\}_{j=1}^\infty$ be i.i.d. Bernoulli (η) random variables (with $\eta = \frac{N_m - 1}{N-1}$) and $\{\bar{H}_j^{(i)}\}_{j=1}^\infty$ be i.i.d. random variables chosen uniformly at random from $\mathcal{I}_m \backslash \{i\}$. We define $\bar{\mathcal{R}}_{m,j}$ as follows: $\bar{\mathcal{R}}_{m,0} = \{i_m^*\}$ (from Assumption 3.1), and

$$\bar{\mathcal{R}}_{m,j} = \bar{\mathcal{R}}_{m,j-1} \cup \{i \in \mathcal{I}_m \backslash \bar{\mathcal{R}}_{m,j-1} : \bar{Y}_j^{(i)} = 1, \bar{H}_j^{(i)} \in \bar{\mathcal{R}}_{m,j-1}\} \ \forall\, j \in \mathbb{N}.$$

Then, we can define $\tau_{\mathrm{spr},m} = \inf\{j \in \mathbb{N} : \bar{\mathcal{R}}_{m,j} = \mathcal{I}_m\}$.

We now couple the fictitious noisy rumor spreading process introduced in the proof sketch with the fictitious noiseless process to obtain a bound on $\mathbb{E}[A_{2\bar{\tau}_{\mathrm{spr},m}}]$, which will also bound $\mathbb{E}[A_{2\tau_{\mathrm{spr},m}}]$ following Proposition C.5. We describe the fictitious noiseless process, which is restated from (Vial et al., 2022) for the sake of completeness. Fix a bandit $m \in [M]$. Let $\{\underline{H}_j^{(i)}\}_{j=1}^\infty$ be i.i.d. Uniform $(N_{\mathrm{hon}}(i))$ random variables for each $i \in \mathcal{I}_m$. Note that in our setting, $N_{\mathrm{hon}}(i) = \mathcal{I}_m \backslash \{i\}$. Let

$$\underline{\mathcal{R}}_{m,0} = \{i_m^*\}, \underline{\mathcal{R}}_{m,j} = \underline{\mathcal{R}}_{m,j-1} \cup \{i \in \mathcal{I}_m \backslash \underline{\mathcal{R}}_{m,j-1} : \underline{H}_j^{(i)} \in \underline{\mathcal{R}}_{m,j-1}\} \ \forall\, j \in \mathbb{N},$$

and let $\underline{\tau}_{\mathrm{spr},m} = \inf\{j \in \mathbb{N} : \underline{\mathcal{R}}_{m,j} = \mathcal{I}_m\}$.

Next, we define the variables pertinent tor the coupling between the fictitious noisy and the fictitious noiseless processes. Let

$$\sigma_0 = 0, \sigma_l = \inf\left\{ j > \sigma_{l-1} : \min_{i \in \mathcal{I}_m} \sum_{s=\sigma_{l-1}+1}^{j} \bar{Y}_s^{(i)} \geq 1 \right\} \ \forall\, l \in \mathbb{N}.$$

Furthermore, for each $i \in \mathcal{I}_m$ and $l \in \mathbb{N}$, let $Z_l^{(i)} = \min\{j \in \{1 + \sigma_{l-1}, \cdots, \sigma_l\} : Y_j^{(i)} = 1\}$. Note that $\{Z_l^{(i)}\}_{i \in \mathcal{I}_m, l \in \mathbb{N}}$ is non-empty, and since $Z_l^{(i)}$ is a deterministic function of $\{\bar{Y}_j^{(i)}\}_{j=1}^{\infty}$ (which is independent of $\{\bar{H}_j^{(i)}\}_{j=1}^{\infty}$), $\{\bar{H}_{Z_l^{(i)}}^{(i)}\}_{j=1}^{\infty}$ is also Uniform$(\mathcal{I}_m \backslash \{i\})$ for each $l \in \mathbb{N}$. Thus, we can set

$$\underline{H}_j^{(i)} = \begin{cases} \bar{H}_{Z_l^{(i)}}^{(i)} & \text{if } j = Z_l^{(i)} \text{ for some } l \in \mathbb{N} \\ \text{Uniform}(\mathcal{I}_m \backslash \{i\}) & \text{otherwise} \end{cases}$$

without changing the distribution of $\{\underline{\mathcal{R}}_{m,j}\}_{j=0}^{\infty}$. This results in a coupling where the fictitious noiseless process dominates the fictitious noisy process, as follows:

Proposition C.7. *(Claim 5 in (Vial et al., 2022)) For the coupling described above, $\underline{\mathcal{R}}_{m,j} \subset \bar{\mathcal{R}}_{m,\sigma_j}$ for all $j \in \mathbb{N}$.*

Proposition C.7 allows us to relate the rumor spreading time of the fictitious noisy and the fictitious noiseless processes, denoted by $\bar{\tau}_{\text{spr},m}$ and $\underline{\tau}_{\text{spr},m}$. In order to do so, we restate the following result from (Vial et al., 2022) in the context of our setting.

Proposition C.8. *(Claim 6 in (Vial et al., 2022)) For any $j \geq 3$ and $\iota > 1$, we have $\mathbb{P}\left(\bar{\tau}_{\text{spr},m} > \frac{\iota j \log j}{\eta}\right) \leq \mathbb{P}(\underline{\tau}_{\text{spr},m}) + 27 c_2 \frac{N}{M} j^{1-\iota}$.*

We now state the result bounding $\mathbb{E}[A_{2\bar{\tau}_{\text{spr},m}}]$ with $A_j = \lceil j^{\beta} \rceil$.

Proposition C.9. *Under the conditions of Theorem 3.2, $\mathbb{E}[A_{2\bar{\tau}_{\text{spr},m}}]$ scales as $O\left(\left(M \left(\log \frac{N}{M}\right)^2 \log\left(\log \frac{N}{M}\right)\right)^{\beta}\right)$, where $O(.)$ only hides the absolute constants.*

We refer the interested reader to (Vial et al., 2022) (Appendix D.2) for the details about the proofs of Propositions C.8 and C.9. The results in the Appendix D.2 of (Vial et al., 2022) (Claims 7 and 8 in particular) are for d-regular graphs with conductance ϕ. We would like to point out that in our setting, for each bandit $m \in [M]$, G_m is the gossip matrix of a complete graph of size N_m. Given that a complete graph of size N_m is a d-regular graph with $d = N_m - 1$, hence the conductance $\phi = \frac{N_m}{2(N_m - 1)}$. Proposition C.9 follows by substituting the aforementioned values of d and ϕ, along with η and using the assumption that $N_m = \Theta(\frac{N}{M})$ in the results in the Appendix D.2 of (Vial et al., 2022).

C.3. Completing the proof of Theorem 3.2

Propositions C.5 and C.9 imply that $\mathbb{E}[A_{2\tau_{\text{spr},m}}]$ also scales as $O\left(\left(M \left(\log \frac{N}{M}\right)^2 \log\left(\log \frac{N}{M}\right)\right)^{\beta}\right)$. Combining the above observation with Propositions C.3 and C.1 completes the proof of Theorem 3.2.

D. Random Initialization of Sticky Sets

Proposition D.1. *If $S = \left\lceil \frac{MK}{N} \log \frac{M}{\gamma} \right\rceil$ for some $\gamma \in (0, 1)$ and we construct $\widehat{S}^{(i)}$ for each agent $i \in [N]$ by sampling S arms independently and uniformly at random from K arms, then Assumption 3.1 holds with probability at least $1 - \gamma$.*

Proof. Fix a bandit $m \in [M]$ and let $i \in \mathcal{I}_m$. Then,

$$\mathbb{P}(k^*(m) \notin \widehat{S}^{(i)}) = \frac{\binom{K-1}{S}}{\binom{K}{S}} = \frac{K - S}{K} = 1 - \frac{S}{K}.$$

Let $\widehat{E}_m$ denote the event that $k^*(m) \notin \widehat{S}^{(i)}$ for all $i \in \mathcal{I}_m$.Then,

$$\mathbb{P}(\widehat{E}_m) = \mathbb{P}\left(\bigcap_{i \in \mathcal{I}_m} (k^*(m) \notin \widehat{S}^{(i)})\right) = \prod_{i \in \mathcal{I}_m} \mathbb{P}(k^*(m) \notin \widehat{S}^{(i)}),$$

$$= \left(1 - \frac{S}{K}\right)^{\frac{N}{M}} \leq e^{-\frac{NS}{MK}}.$$

In order for Assumption 3.1 to fail, there must be at least one bandit for which no agent learning that bandit has the best arm. Therefore, Assumption 3.1 fails with probability

$$\mathbb{P}\left(\bigcup_{m\in[M]}\widehat{E}_m\right) \leq \sum_{m\in[M]}\mathbb{P}(\widehat{E}_m),$$
$$\leq Me^{-\frac{NS}{MK}}.$$

Setting $S = \left\lceil \frac{MK}{N}\log\frac{M}{\gamma}\right\rceil$ completes the proof. □

E. Proof of Theorem 4.1

We need additional notation for proving Theorem 3, particularly due to the complicated active set updates at the end of a phase. Let $\{same(z)\}_{z\in[\frac{N}{r}]}$ be a partition of the set of all the agents $[N]$ consisting of $\frac{N}{r}$ sets, such that: (i) $|same(z)| = r$ for all $z \in [\frac{N}{r}]$, and (ii) for any $i, i^{'} \in same(z)$ such that $i \neq i^{'}$, $c(i) = c(i^{'})$ and if $i \in f(i^{'})$, then $i^{'} \in f(i)$. In words, for each z, $same(z)$ consists of agents learning the same bandit, which each of the agents in $same(z)$ are aware of, i.e., for some $z \in [\frac{N}{r}]$, $same(z) = \{i, f(i)\} = \{i \cup f(i)\}$ for some $i \in [N]$. Similar to the proof of Theorem 3.2, we define some random times, which will help us prove Theorem 4.1.

$$\begin{aligned}
\tau_{\text{stab}}^{(i)} &= \inf\{j \geq j^* : \forall l \geq j, \chi_l^{(i)} = 0\},\\
\tau_{\text{stab}} &= \max_{i\in[N]} \tau_{\text{stab}}^{(i)},\\
\tau_{\text{spr}}^{(i)} &= \inf\{j \geq \tau_{\text{stab}} : k^*(c(i)) \in S_j^{(i)}\} - \tau_{\text{stab}},\\
\tau_{\text{spr},m} &= \max_{i\in\mathcal{I}_m} \tau_{\text{spr}}^{(i)},\\
\tau_{\text{spr}} &= \max_{m\in[M]} \tau_{\text{spr},m},\\
\tau_{\text{rec},z} &= \inf\{j \geq \tau_{\text{stab}} + \tau_{\text{spr}} : \text{uniquerec}(same(z), j) = \mathcal{B}\} - (\tau_{\text{stab}} + \tau_{\text{spr}}),\\
\tau_{\text{rec}} &= \max_{z\in[\frac{N}{r}]} \tau_{\text{rec},z},\\
\tau &= \tau_{\text{stab}} + \tau_{\text{spr}} + \tau_{\text{rec}}.
\end{aligned}$$

For each group of agents $z \in [\frac{N}{r}]$ learning the same bandit and are aware of that, $\tau_{\text{rec},z}$ denotes the minimum number of phases it takes after the phase $\tau_{\text{stab}} + \tau_{\text{spr}}$ such that the M most recent unique arm recommendations among all the agents in $same(z)$ is the set of M best arms. The active set updates in Algorithm 3 ensure that the active sets of agents freeze after phase τ (like (Chawla et al., 2020)), unlike Algorithm 1, where the active sets are time-varying despite agents eventually identifying their respective best arms and recommending it in information pulls. We now prove this claim.

Proposition E.1. *For any agent $i \in \mathcal{I}_m$ playing Algorithm 3, the following statements hold for all $j > \tau$:*

(i) $k^*(m) \in S_j^{(i)}$,

(ii) $S_j^{(i)} = S_\tau^{(i)}$.

(iii) $\left\{\widetilde{S}_\tau^{(i)} \cup \{\mathcal{O}_\tau^{(i)}\}\right\} \subset \mathcal{B}$.

Proof. (i) follows exactly from Claim (3).

For proving (ii), notice that for any group of agents $same(z)$ and any phase $j \geq \tau_{\text{stab}} + \tau_{\text{spr}} + \tau_{\text{rec},z}$, $\text{uniquerec}(same(z), j) = \mathcal{B}$. This follows from Claim (3) because after the phase $\tau_{\text{stab}} + \tau_{\text{spr}}$, agents recommend their respective best arms during information pulls and eventually, the M most recent unique arm recommendations for all the agents in $same(z)$ (denoted by $\text{uniquerec}(same(z), j-1)$) will be the set of M best arms. Furthermore, the active set updates in Algorithm 3 ensure that if $\text{uniquerec}(same(z), j) = \text{uniquerec}(same(z), j-1)$, the active sets for all the agents in $same(z)$ remain unchanged. Since $\tau \geq \tau_{\text{stab}} + \tau_{\text{spr}} + \tau_{\text{rec},z}$, claim (ii) holds.

Claim (iii) follows from the observation that for any group of agents $same(z)$ and any phase $j \geq \tau_{\text{stab}} + \tau_{\text{spr}} + \tau_{\text{rec},z}$, $\text{uniquerec}(same(z), j) = \mathcal{B}$, $\widetilde{S}_j^{(i)} \subset \text{uniquerec}(same(z), j)$ by construction and $\mathcal{O}_j^{(i)} \in \mathcal{B}$ after any phase $j \geq \tau_{\text{stab}} + \tau_{\text{spr}}$. □

E.1. Intermediate Results

Most of the intermediate results for Algorithm 3 are similar to the results for Algorithm 1, and are proved in a similar way. The technical novelty in proving Theorem 4.1 is to prove that $\mathbb{E}[\tau_{\text{rec}}]$ is finite.

Before decomposing the regret up to phase τ and after phase τ, we define some notation. For any bandit $m \in [M]$, Let $k_{m,1}, k_{m,2}, \cdots, k_{m,K}$ denote the IDs of the arms with their arm means sorted in decreasing order, i.e., $k_{m,1} = k^*(m)$ and $\mu_{m,k_{m,1}} > \mu_{m,k_{m,2}} \geq \cdots \geq \mu_{m,k_{m,K}}$.

Proposition E.2. *The expected cumulative regret of any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ after playing for T time steps is bounded by:*

$$\mathbb{E}[R_T^{(i)}] \leq \mathbb{E}[A_\tau] + \frac{\pi^2}{3}K + \sum_{k \in \{k_{m,l}\}_{l=2}^{S+\lceil \frac{M}{r} \rceil + 2}} \frac{4\alpha}{\Delta_{m,k}} \log T.$$

Proof. Fix a bandit $m \in [M]$ and an agent $i \in \mathcal{I}_m$. By regret decomposition principle, we know that

$$\begin{aligned}
R_T^{(i)} &= \sum_{k \in [K]} \Delta_{m,k} \sum_{t=1}^{T} \mathbf{1}(I_t^{(i)} = k), \\
&= \sum_{t=1}^{T} \sum_{k \in [K]} \Delta_{m,k} \mathbf{1}(I_t^{(i)} = k), \\
&\leq A_\tau + \sum_{k \in [K]} \Delta_{m,k} \sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k),
\end{aligned}$$

where $I_t^{(i)}$ is the arm played by agent i at time t and the last step follows from the fact that $\Delta_{m,k} \in (0,1)$. Taking expectation on both sides, we get

$$\mathbb{E}[R_T^{(i)}] \leq \mathbb{E}[A_\tau] + \sum_{k \in [K]} \Delta_{m,k} \mathbb{E}\left[\sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k)\right]. \tag{10}$$

The first term $\mathbb{E}[A_\tau]$ is bounded in Proposition E.4. We now bound the second term. Let $X_{k,t}^{(i)} = \mathbf{1}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) \leq \frac{4\alpha}{\Delta_{m,k}^2} \log T\right)$ and $Y_{k,t}^{(i)} = \mathbf{1}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right)$. Then, the inner sum in the second term can be re-written as follows:

$$\sum_{t=A_\tau+1}^{T} \mathbf{1}(I_t^{(i)} = k) = \sum_{t=1}^{T} (X_{k,t}^{(i)} + Y_{k,t}^{(i)}). \tag{11}$$

We first bound the sum $\sum_{t=1}^{T} \mathbb{E}[Y_{k,t}^{(i)}]$. Notice that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[Y_{k,t}^{(i)}] &= \sum_{t=1}^{T} \mathbb{P}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right), \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(I_t^{(i)} = k, T_k^{(i)}(t-1) > \frac{4\alpha}{\Delta_{m,k}^2} \log T\right), \\
&\leq \sum_{t=1}^{T} 2t^{2-\frac{\alpha}{2}} \leq \frac{\pi^2}{3},
\end{aligned} \tag{12}$$

where we substitute the classical estimate of the probability that UCB plays a sub-optimal arm using Chernoff-Hoeffding bound for 1-subgaussian rewards and $\alpha > 10$ in the last line.

For an arm $k \in [K]$, we bound the sum $\sum_{t=1}^{T} \mathbb{E}[X_{k,t}^{(i)}]$ and complete the proof.

$$\sum_{t=1}^{T} X_{k,t}^{(i)} = \sum_{t=1}^{T} \mathbf{1}\left(t > A_\tau, I_t^{(i)} = k, T_k^{(i)}(t-1) \le \frac{4\alpha}{\Delta_{m,k}^2}\log T\right) \tag{13}$$

For any arm $k \in [K]$, two cases are possible:

- if $k \in S_\tau^{(i)}$ and arm k is one of the sub-optimal arms, the sum in equation (13) is bounded above by $\frac{4\alpha}{\Delta_{m,k}^2}\log T$, This is because the event $I_t^{(i)} = k$ cannot occur more than $\frac{4\alpha}{\Delta_{m,k}^2}\log T$ times, otherwise it will contradict the event $T_k^{(i)}(t-1) \le \frac{4\alpha}{\Delta_{m,k}^2}\log T$.
- if arm $k \notin S_\tau^{(i)}$, the event $I_t^{(i)} = k$ cannot happen. Thus, the sum in equation (13) is equal to zero.

From the above observations, we can conclude that for all $k \in \{S_\tau^{(i)}\}\backslash\{k^*(m)\}$,

$$\sum_{t=1}^{T} X_{k,t}^{(i)} \le \frac{4\alpha}{\Delta_{m,k}^2}\log T. \tag{14}$$

Therefore,

$$\begin{aligned}\sum_{k=1}^{K}\Delta_{m,k}\sum_{t=1}^{T} X_{k,t}^{(i)} &\le \sum_{k \in S_\tau^{(i)}} \frac{4\alpha}{\Delta_{m,k}}\log T,\\ &\le \sum_{k \in \widehat{S}^{(i)}\backslash\{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}}\log T + \sum_{k \in \{k_{m,l}\}_{l=2}^{\lceil \frac{M}{r}\rceil+2}} \frac{4\alpha}{\Delta_{m,k}}\log T,\end{aligned}$$

because $S_\tau^{(i)} = \widehat{S}^{(i)} \cup \{\widehat{\mathcal{O}}_\tau^{(i)}\} \cup \{\mathcal{O}_\tau^{(i)}\} \cup \{\widetilde{S}_\tau^{(i)}\}$, $\widehat{\mathcal{O}}_\tau^{(i)} = k^*(m)$ and we don't have any control over which arms from the set $\mathcal{B}$ are present in the set $\widetilde{S}_\tau^{(i)} \cup \mathcal{O}_\tau^{(i)}$ (Proposition E.1), so we assume the worst case scenario and bound the regret of the arms in that set by the regret of the first $\lceil \frac{M}{r}\rceil + 2$ arms in the increasing order of their arm gaps (or equivalently, decreasing order of the arm means). By taking the expectation on both sides in the above inequality, we get

$$\sum_{t=1}^{T} \mathbb{E}[X_{k,t}^{(i)}] \le \sum_{k \in \widehat{S}^{(i)}\backslash\{k^*(m)\}} \frac{4\alpha}{\Delta_{m,k}}\log T + \sum_{k \in \{k_{m,l}\}_{l=2}^{\lceil \frac{M}{r}\rceil+2}} \frac{4\alpha}{\Delta_{m,k}}\log T. \tag{15}$$

The proof of Proposition C.1 is completed by substituting equations (12) and (15) into the expectation of the equation (11), followed by substituting the bound on the expectation of the equation (11) into equation (10). □

In order to obtain an upper bound on $\mathbb{E}[A_\tau]$, we first show that the probability of the error event that the best arm is not recommended during information pull at the end of the phase j (indicated by $\chi_j^{(i)}$ in equation (2)) decreases as the phases progress. This result is stated as a lemma below:

Lemma E.3. *For any agent $i \in \mathcal{I}_m$ for all $m \in [M]$ and phase $j \in \mathbb{N}$ such that $\frac{A_j - A_{j-1}}{S+\lceil \frac{M}{r}\rceil+2} \ge 1 + \frac{4\alpha}{\Delta_m^2}\log A_j$, we have*

$$\mathbb{E}[\chi_j^{(i)}] \le \frac{2}{\frac{\alpha}{2}-3}\binom{K}{2+\lceil\frac{M}{r}\rceil}\left(S + \left\lceil\frac{M}{r}\right\rceil + 1\right)\frac{1}{A_{j-1}^{\frac{\alpha}{2}-3}}.$$

Proof. The proof of this lemma is identical to the proof of Lemma 6 in (Chawla et al., 2020) and Lemma C.2, except that $|S_j^{(i)}| \le S + \lceil\frac{M}{r}\rceil + 2$. □

Proposition E.4. *The regret up to phase τ is bounded by*

$$\mathbb{E}[A_\tau] \leq \lceil (j^*)^\beta \rceil + \frac{N(3^\beta+1)3^{\beta(\frac{\alpha}{2}-3)}}{\frac{\alpha}{2}-3}\binom{K}{2+\lceil\frac{M}{r}\rceil}\left(S+\left\lceil\frac{M}{r}\right\rceil+1\right)\frac{\pi^2}{3} + \sum_{m\in[M]}\mathbb{E}[A_{3\tau_{\mathrm{spr},m}}] + \sum_{z\in[\frac{N}{r}]}\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}],$$

where j^* is defined in Theorem 4.1.

Proof. The random variable τ has support in $\mathbb{N}$. For $\mathbb{N}$-valued random variables, we know that

$$\begin{aligned}
\mathbb{E}[\mathbb{A}_\tau] &= \sum_{t\geq 1}\mathbb{P}(A_\tau \geq t),\\
&\overset{(a)}{\leq} \sum_{t\geq 1}\mathbb{P}(\tau \geq A^{-1}(t)),\\
&\leq \sum_{t\geq 1}\mathbb{P}(\tau_{\mathrm{stab}}+\tau_{\mathrm{spr}}+\tau_{\mathrm{rec}} \geq A^{-1}(t)),\\
&\leq \sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{spr}} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{rec}} \geq \frac{A^{-1}(t)}{3}\right),\\
&= \sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{t\geq 1}\mathbb{P}\left(\bigcup_{m=1}^{M}\left(\tau_{\mathrm{spr},m} \geq \frac{A^{-1}(t)}{3}\right)\right) + \sum_{t\geq 1}\mathbb{P}\left(\bigcup_{z\in[\frac{N}{r}]}\left(\tau_{\mathrm{rec},z} \geq \frac{A^{-1}(t)}{3}\right)\right),\\
&\leq \sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{m\in[M]}\sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{spr},m} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{z\in[\frac{N}{r}]}\sum_{t\geq 1}\mathbb{P}\left(\tau_{\mathrm{rec},m} \geq \frac{A^{-1}(t)}{3}\right),\\
&\leq A_{j^*} + \sum_{t\geq A_{j^*}+1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{3}\right) + \sum_{m\in[M]}\mathbb{E}[A_{3\tau_{\mathrm{spr},m}}] + \sum_{z\in[\frac{N}{r}]}\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}], \qquad (16)
\end{aligned}$$

where we use the definition of $A^{-1}(.)$ defined in equation (1) in step (a). $\mathbb{E}[A_{3\tau_{\mathrm{spr},m}}]$ is bounded for all $m \in [M]$ in the same way as for Algorithm 1 and is stated in Proposition E.6. We bound $\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}]$ for each $z \in [\frac{N}{r}]$ in Proposition E.7.

The process for bounding the sum $\sum_{t\geq A_{j^*}+1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{3}\right)$ is identical to bounding the sum $\sum_{t\geq A_{\tau^*}+1}\mathbb{P}\left(\tau_{\mathrm{stab}} \geq \frac{A^{-1}(t)}{2}\right)$ for Algorithm 1 in Proposition C.3. □

Proposition E.5. *j_m^* defined in Theorem 4.1 is bounded by*

$$j_m^* \leq 2 + \left(\frac{1}{\beta} + \left(\frac{1}{\beta} + \frac{8\alpha}{\Delta_m^2}\right)\left(S+2+\left\lceil\frac{M}{r}\right\rceil\right)\right)^{\frac{1}{\beta-2}}.$$

Proof. The proof of Proposition E.5 is identical to the proof of Proposition C.4, except that $S+2$ is replaced by $S+2+\lceil\frac{M}{r}\rceil$. □

Proposition E.6. *Under the conditions of Theorem 4.1, $\mathbb{E}[A_{3\tau_{\mathrm{spr},m}}]$ scales as $O\left(\left(M\left(\log\frac{N}{M}\right)^2\log\left(\log\frac{N}{M}\right)\right)^\beta\right)$, where $O(.)$ only hides the absolute constants.*

Proof. The proof is identical to the proof of $\mathbb{E}[A_{2\tau_{\mathrm{spr},m}}]$ for Algorithm 1. □

Proposition E.7. *For each $z \in [\frac{N}{r}]$, $\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}]$ is bounded by*

$$\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}] \leq \lceil (3M)^\beta \rceil + 2\left(\frac{3}{\left(\frac{c_1}{M}-\frac{1}{N}\right)r}\right)^\beta \frac{M}{\left(1-\frac{c_1}{M}\right)^r}\Gamma(\beta+1),$$

where $\Gamma(\alpha) = \int_{t=0}^{\infty} t^{\alpha-1}e^{-t}\,\mathrm{d}t$ for any $\alpha > 0$ denotes the Gamma function.

Proof. Fix a $z \in [\frac{N}{r}]$. Then, we know from the definition of $same(z)$ that there exists an $i \in \mathcal{I}_m$ such that $same(z) = \{i, f(i)\} = \{\{i\} \cup f(i)\}$ for some $m \in [M]$.

Let E_m denote the event that an agent $i \in [N]$ doesn't contact an agent from $\mathcal{I}_m$ during an information pull at the end of a phase. Then, from the choice of gossip matrix G in Theorem 4.1,

$$\mathbb{P}(E_m) = \begin{cases} 1 - \frac{N_m - 1}{N-1} & \text{if } c(i) = m, \\ 1 - \frac{N_m}{N-1} & \text{otherwise.} \end{cases}$$

Furthermore,

$$\mathbb{P}(E_m) \leq \begin{cases} 1 - \frac{c_1}{M} + \frac{1}{N} & \text{if } c(i) = m, \\ 1 - \frac{c_1}{M} & \text{otherwise,} \end{cases} \tag{17}$$

by assumption on N_m for all $m \in [M]$.

For some $m \in [M]$ and $i \in \mathcal{I}_m$, $same(z) = \{i, f(i)\} = \{\{i\} \cup f(i)\}$ for some $z \in [\frac{N}{r}]$. We also know that in $\tau_{\mathrm{rec},z}$ steps, agents in $same(z)$ will receive a total of $r\tau_{\mathrm{rec},z}$ arm recommendations, which for each agent $i \in same(z)$ are i.i.d. uniform$\{[N]\backslash\{i\}\}$ and independent across all $i \in same(z)$. By the definition of $\tau_{\mathrm{rec},z}$, $(\tau_{\mathrm{rec},z} > x)$ denotes the event that agents in the set $same(z)$ don't have the M most recent unique arm recommendations equal to the set of M best arms in x steps (denoted by $\mathcal{B}$). Therefore,

$$\begin{aligned} \mathbb{P}(\tau_{\mathrm{rec},z} > x) &= \mathbb{P}\left(\cup_{m' \in [M]} (E_{m'} \text{ occurs } rx \text{ times})\right), \\ &\leq \sum_{m' \in [M]} \mathbb{P}\left(E_{m'} \text{ occurs } rx \text{ times}\right), \\ &\leq \left(\sum_{m' \neq m} \left(1 - \frac{c_1}{M}\right)^{rx}\right) + \left(1 - \frac{c_1}{M} + \frac{1}{N}\right)^{rx}, \\ &\leq (M-1)\left(1 - \frac{c_1}{M}\right)^{rx} + \left(1 - \frac{c_1}{M} + \frac{1}{N}\right)^{rx}, \end{aligned} \tag{18}$$

where the last two steps follow from equation (17) and the fact that the information pulls are independent across agents and phases.

We are now ready to bound $\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}]$. Given that $A_{3\tau_{\mathrm{rec},z}}$ is a $\mathbb{N}$-valued random variable, we have

$$\begin{aligned} \mathbb{E}[A_{3\tau_{\mathrm{rec},z}}] &= \sum_{t \geq 1} \mathbb{P}(A_{3\tau_{\mathrm{rec},z}} \geq t), \\ &\leq A_{3M} + \sum_{t \geq A_{3M}+1} \mathbb{P}(A_{3\tau_{\mathrm{rec},z}} \geq t), \\ &\overset{(a)}{\leq} A_{3M} + \sum_{j \geq 1} \mathbb{P}(A_{3\tau_{\mathrm{rec},z}} \geq A_{3(M+j-1)})(A_{3(M+j)} - A_{3(M+j-1)}), \\ &\overset{(b)}{\leq} A_{3M} + \sum_{j \geq 1} \mathbb{P}(\tau_{\mathrm{rec},z} \geq M + j - 1) A_{3(M+j)}, \\ &\overset{(c)}{\leq} A_{3M} + (M-1)\sum_{j \geq 1}\left(1 - \frac{c_1}{M}\right)^{r(M+j-1)} A_{3(M+j)} + \sum_{j \geq 1}\left(1 - \frac{c_1}{M} + \frac{1}{N}\right)^{r(M+j-1)} A_{3(M+j)}, \end{aligned} \tag{19}$$

where step (a) follows from the fact that for a random variable X, $\mathbb{P}(X \geq x)$ is non-increasing in x, step (b) follows from the definition of $A^{-1}(.)$ in equation (1) and we subsitute equation (18) in step (c). We bound each of the two sums in equation (19) to complete the proof. For any $\epsilon \in (0,1)$ and $A_j = \lceil j^\beta \rceil$,

$$\begin{aligned}
\sum_{j\geq 1}(1-\epsilon)^{r(M+j-1)}A_{3(M+j)} &= \sum_{j\geq 1}(1-\epsilon)^{r(M+j-1)}\lceil (3(M+j))^\beta \rceil, \\
&\overset{(a)}{\leq} 2(3^\beta)\sum_{j\geq 1}(1-\epsilon)^{r(M+j-1)}(M+j)^\beta, \\
&= 2(3^\beta)\sum_{l\geq M+1}(1-\epsilon)^{r(l-1)}l^\beta, \\
&\leq \frac{2(3^\beta)}{(1-\epsilon)^r}\sum_{l\geq M+1}e^{-\epsilon r l}l^\beta, \\
&\leq \frac{2(3^\beta)}{(1-\epsilon)^r}\int_0^\infty e^{-\epsilon r y}y^\beta \,\mathrm{d}y, \\
&\overset{(b)}{=} \frac{2}{(1-\epsilon)^r}\left(\frac{3}{\epsilon r}\right)^\beta \int_0^\infty u^\beta e^{-\epsilon u}\,\mathrm{d}u, = \frac{2}{(1-\epsilon)^r}\left(\frac{3}{\epsilon r}\right)^\beta \Gamma(\beta+1),
\end{aligned}$$

where we use the property that $\lceil x \rceil \leq 2x$ for all $x \geq 1$ in step (a) and we perform a change of variables $u = \epsilon r y$ in step (b). Substituting the above bound in equation (19) with $\epsilon = \frac{c_1}{M}$ and $\epsilon = \frac{c_1}{M} - \frac{1}{N}$ in each of the two sums respectively, we get

$$\begin{aligned}
\mathbb{E}[A_{3\tau_{\mathrm{rec},z}}] &\leq A_{3M} + \frac{2(M-1)}{\left(1-\frac{c_1}{M}\right)^r}\left(\frac{3}{\frac{c_1}{M}r}\right)^\beta \Gamma(\beta+1) + \frac{2}{\left(1-\frac{c_1}{M}+\frac{1}{N}\right)^r}\left(\frac{3}{\left(\frac{c_1}{M}-\frac{1}{N}\right)r}\right)^\beta \Gamma(\beta+1), \\
&\leq A_{3M} + \frac{2(M-1)}{\left(1-\frac{c_1}{M}\right)^r}\left(\frac{3}{\left(\frac{c_1}{M}-\frac{1}{N}\right)r}\right)^\beta \Gamma(\beta+1) + \frac{2}{\left(1-\frac{c_1}{M}\right)^r}\left(\frac{3}{\left(\frac{c_1}{M}-\frac{1}{N}\right)r}\right)^\beta \Gamma(\beta+1), \\
&= \lceil (3M)^\beta \rceil + 2\left(\frac{3}{\left(\frac{c_1}{M}-\frac{1}{N}\right)r}\right)^\beta \frac{M}{\left(1-\frac{c_1}{M}\right)^r}\Gamma(\beta+1),
\end{aligned}$$

which completes the Proof of Proposition E.7. □

E.2. Completing the proof of Theorem 4.1

The proof of Theorem 3 follows from the Propositions E.2, E.4, E.6 and E.7.

F. Proofs of Lower Bounds

F.1. Proof of Theorem 5.1

We derive the lower bound for our model by adapting the lower bound for the setting considered in (Réda et al., 2022). (Réda et al., 2022) considers the following problem: there are N agents and K arms. When agent $n \in [N]$ pulls arm $k \in [K]$, it observes a noisy reward with mean $\bar{\mu}_{k,n}$. However, regret is measured with respect to a "mixed reward" with mean $\mu'_{k,n} = \sum_{i=1}^N w_{i,n}\bar{\mu}_{k,i}$, where $\{w_{i,n}\}_{i=1}^N$ are (known) nonnegative weights with $\sum_{i=1}^N w_{i,n} = 1$. So the optimal arm for agent n is $k_n^\star = \arg\max_{k\in[K]} \mu'_{k,n}$ and the relevant arm gaps are $\Delta'_{k,n} = \mu'_{k_n^\star,n} - \mu'_{k,n}$. Unlike our work, (Réda et al., 2022) doesn't have any constraints on the lengths of the messages exchanged in the communication rounds.

Lower bound: Theorem 3 of (Réda et al., 2022) bounds the group regret $\mathrm{Reg}(T)$ (i.e., regret summed across agents) as follows: for uniformly efficient algorithms, assuming Gaussian rewards with unit variance,

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}(T)}{\log T} \geq \min_{x\in\mathcal{X}} f(x), \quad \text{where} \tag{20}$$

$$\mathcal{X} = \left\{ x \in \mathbb{R}_+^{K\times N} : \sum_{i:k_i^\star \neq k}\frac{w_{i,n}^2}{x_{k,i}} \leq \frac{(\Delta'_{k,n})^2}{2}\ \forall\, n \in [N], k \in [K] \right\}, \quad f(x) = \sum_{k=1}^K \sum_{n:k_n^\star \neq k} x_{k,n}\Delta'_{k,n}\ \forall\, x \in \mathcal{X}. \tag{21}$$

Applying to our setting and proving Theorem 5.1

Notation: In our setting, $c(n) \in [M]$ denotes the bandit that agent $n \in [N]$ is learning and let $c^{-1}(m) = \{n \in [N] : c(n) = m\}$ denote the set of agents learning the bandit $m \in [M]$. Notice that for any agent $n \in [N]$, $c^{-1}(c(n))$ is the set of agents learning the same bandit as agent n (also, $n \in c^{-1}(c(n))$).

Reduction to the setting of (Réda et al., 2022): For each agent $n \in [N]$, we can choose $\bar{\mu}_{k,n} = \mu_{c(n),k}$ and $w_{i,n} = \mathbf{1}(i \in c^{-1}(c(n)))/|c^{-1}(c(n))|$ for the arm means and weights. Then defining $\mu'_{k,n}$ as above, a simple calculation shows $\mu'_{k,n} = \mu_{c(n),k} = \bar{\mu}_{k,n}$, so $\Delta'_{k,n} = \Delta_{c(n),k}$. Note in particular that $\mu'_{k,n} = \bar{\mu}_{k,n}$ means the observed and mixed rewards are the same, as in our model, and thus, $k^\star_n = k^*(c(n))$. Given that agents are aware of the weights in the setting of (Réda et al., 2022), in our model this corresponds to the case when every agent knows all the other agents learning the same bandit. Additionally, in (Réda et al., 2022), there are no constraints on the length of the messages exchanged in the communication rounds. Due to the absence of constraints on the lengths of the messages exchanged during communications, our model is a special case of the model in (Réda et al., 2022), and thus, their lower bound is applicable to our setting.

Applying their lower bound: In our special case, for any agent n and arm $k \neq k^\star_n$, we have

$$\sum_{i:k^\star_i \neq k} \frac{w_{i,n}^2}{x_{k,i}} = \frac{1}{|c^{-1}(c(n))|^2} \sum_{i \in c^{-1}(c(n)):k^\star_i \neq k} \frac{1}{x_{k,i}} = \frac{1}{|c^{-1}(c(n))|^2} \sum_{i \in c^{-1}(c(n))} \frac{1}{x_{k,i}}, \tag{22}$$

where the first equality uses our choice of $w_{i,n}$ and the second holds since $k^\star_i = k^\star_n \neq k$ for each $i \in c^{-1}(c(n))$. Therefore, the constraint for such n, k pairs simplifies to

$$\frac{1}{|c^{-1}(c(n))|^2} \sum_{i \in c^{-1}(c(n))} \frac{1}{x_{k,i}} \leq \frac{(\Delta'_{k,n})^2}{2} = \frac{\Delta^2_{c(n),k}}{2},$$

which can be rearranged to obtain

$$\frac{2}{|c^{-1}(c(n))|\Delta^2_{c(n),k}} \leq \frac{|c^{-1}(c(n))|}{\sum_{i \in c^{-1}(c(n))} \frac{1}{x_{k,i}}} = \mathrm{HM}(\{x_{k,i}\}_{i \in c^{-1}(c(n))}),$$

where $\mathrm{HM}(\{y_j\}_j)$ is the harmonic mean of $\{y_j\}_j \subset \mathbb{R}_+$. On the other hand, when $k = k^\star_n$, we have

$$w_{i,n} = 0 \; \forall \; i \notin c^{-1}(c(n)), \quad k^\star_i = k^\star_n = k \; \forall \; i \in c^{-1}(c(n)),$$

so the summation in (22) is zero and the corresponding constraint is satisfied for any $x \in \mathbb{R}_+^{K \times N}$. Combining these observations, the constraint set in our special case simplifies to

$$\mathcal{X} = \left\{ x \in \mathbb{R}_+^{K \times N} : \frac{2}{|c^{-1}(c(n))|\bar{\Delta}^2_{k,c(n)}} \leq \mathrm{HM}(\{x_{k,i}\}_{i \in c^{-1}(c(n))}) \; \forall \; n \in [N], k \neq k^*(c(n)) \right\}.$$

Now notice that these inequality constraints only depend on the bandit $c(n)$, not the individual agent n. Therefore, we further simplify the constraint set as follows:

$$\mathcal{X} = \left\{ x \in \mathbb{R}_+^{K \times N} : \frac{2}{|c^{-1}(m)|\Delta^2_{m,k}} \leq \mathrm{HM}(\{x_{k,n}\}_{n \in c^{-1}(m)}) \; \forall \; m \in [M], k \neq k^*(m) \right\}.$$

Next, consider the objective function in our special case. We rewrite it as

$$f(x) = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{n=1}^{N} x_{k,n} \Delta'_{k,n} \mathbf{1}(k^\star_n \neq k, c(n) = m) = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{n=1}^{N} x_{k,n} \Delta_{m,k} \mathbf{1}(k^*(m) \neq k, c(n) = m) \tag{23}$$

$$= \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \Delta_{m,k} \sum_{n \in c^{-1}(m)} x_{k,n} = \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \Delta_{m,k} |c^{-1}(m)| \mathrm{AM}(\{x_{k,n}\}_{n \in c^{-1}(m)}), \tag{24}$$

where $\mathrm{AM}(\{y_j\}_j)$ denotes the arithmetic mean. Then for any $x \in \mathcal{X}$, we have

$$f(x) \geq \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \Delta_{m,k} |c^{-1}(m)| \mathrm{HM}(\{x_{k,n}\}_{n \in c^{-1}(m)}) \geq \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \frac{2}{\Delta_{m,k}},$$

where the first inequality is $\mathrm{AM}(\{y_j\}_j) \geq \mathrm{HM}(\{y_j\}_j)$ for any $\{y_j\}_j \subset \mathbb{R}_+$ and the second holds since $x \in \mathcal{X}$. On the other hand, if we set $x^\star_{k,n} = 2/(|c^{-1}(c(n))|\Delta^2_{c(n),k})$, we clearly have

$$\mathrm{AM}(\{x^\star_{k,n}\}_{n \in c^{-1}(m)}) = \mathrm{HM}(\{x^\star_{k,n}\}_{n \in c^{-1}(m)}) = 2/(|c^{-1}(m)|\Delta^2_{m,k}) \; \forall \, m \in [M].$$

Combined with the above, this shows that

$$x^\star \in \mathcal{X}, \quad f(x^\star) = \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \frac{2}{\Delta_{m,k}} \leq f(x) \; \forall \, x \in \mathcal{X},$$

so $x^\star$ is optimal. Hence, using the lower bound from (Réda et al., 2022), we get

$$\liminf_{T \to \infty} \frac{\mathrm{Reg}(T)}{\log T} \geq \sum_{m=1}^{M} \sum_{k \neq k^*(m)} \frac{2}{\Delta_{m,k}}.$$

F.2. Proof of Theorem 5.2

The bulk of the proof involves showing that for any uniformly efficient policy,

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T^{(i)}]}{\log T} \geq 2\Delta_m \; \forall \, i \in \cup_{m=1}^{M-1} \mathcal{I}_m. \tag{25}$$

Assuming we can prove (25), the first lower bound in the theorem statement follows easily, since

$$\mathbb{E}[\mathrm{Reg}(T)] = \sum_{i=1}^{N} \mathbb{E}[R_T^{(i)}] \geq \sum_{m=1}^{M-1} \sum_{i \in \mathcal{I}_m} \mathbb{E}[R_T^{(i)}],$$

and the second follows from the first and the fact that, by assumption in Section 2,

$$2 \sum_{m=1}^{M-1} |\mathcal{I}_m| \Delta_m \geq 2\Delta \sum_{m=1}^{M-1} |\mathcal{I}_m| = 2\Delta(N - |\mathcal{I}_M|) \geq 2\Delta N(1 - c_2/M) \geq \Delta N.$$

Hence, for the remainder of the proof, we fix $m \in [M-1]$ and $i \in \mathcal{I}_m$ and prove (25).

Toward this end, we first reduce the real system to a fictitious system with a larger set of uniformly efficient policies. In the fictitious system, agent i initially (i.e., at time $t = 0$) observes the full sequence of rewards for all other agents, i.e., $\mathcal{X} := (X_t^{(n)}(k) : t \in [T], k \in [K], n \in [N] \setminus \{i\})$, and thereafter does not communicate with other agents. Note that any policy in the real system can also be implemented in the fictitious system, since any information communicated by other agents is a function of $\mathcal{X}$. Hence, it suffices to prove (25) for any uniformly efficient policy in the fictitious system.

To do so, we modify a standard lower bound approach that comprises Section 4.6, Lemma 15.1, and Theorem 16.2 of (Lattimore & Szepesvári, 2020). For brevity, we focus on the modifications needed in our setting and refer the reader to the associated references for details.

To begin, we modify the probability measure construction from Section 4.6. First, we define a measurable space $(\Omega_T, \mathcal{F}_T)$, where $\Omega_T = ([K] \times \mathbb{R})^T \times \mathbb{R}^{T \times K \times (N-1)}$, $\mathcal{F}_T = \mathcal{B}(\Omega_T)$, and $\mathcal{B}$ is the Borel σ-algebra. Next, let $\pi = (\pi_t)_{t=1}^T$ be a uniformly efficient policy in the fictitious system. More precisely, each π_t is a mapping from the history of arm pulls and rewards for agent i (i.e., $I_1^{(i)}, X_1^{(i)}(I_1^{(i)}), \ldots I_{t-1}^{(i)}, X_{t-1}^{(i)}(I_{t-1}^{(i)})$), along with the rewards of other agents (i.e., $\mathcal{X}$), to the distribution of the action $I_t^{(i)}$. Further, let p_ν be the density for a Gaussian random variable with mean ν and unit variance. Finally, let ρ denote the counting measure and λ any measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for which all of the reward distributions $\{p_{\mu_{m,k}}\}_{(m,k)\in[M]\times[K]}$ are absolutely continuous with respect to λ. Then we define the probability measure

$$\mathbb{P}_{\mu,\mathcal{I},\pi}(B) = \int_B p_{\mu,\mathcal{I},\pi}(\omega) \left((\rho \times \lambda)^T \times \lambda^{T\times K\times(N-1)} \right) d\omega \; \forall \, B \in \mathcal{F}_T, \tag{26}$$

where, for any $((j_t, x_t)_{t\in[T]}, (x_{t,k,n})_{t\in[T],k\in[K],n\in[N]\setminus\{i\}}) \in \Omega_T$, $p_{\mu,\mathcal{I},\pi}$ is the density given by

$$p_{\mu,\mathcal{I},\pi}\left((j_t, x_t)_{t\in[T]}, (x_{t,k,n})_{t\in[T],k\in[K],n\in[N]\setminus\{i\}}\right) \tag{27}$$

$$= \prod_{t\in[T]} \pi_t\left(j_t \middle| (j_s, x_s)_{s\in[t-1]}, (x_{t,k,n})_{t\in[T],k\in[K],n\in[N]\setminus\{i\}}\right) p_{\mu_{m,j_t}}(x_t) \prod_{t\in[T],k\in[K],n\in[N]\setminus\{i\}} p_{\mu_{c(n),k}}(x_{t,k,n}). \tag{28}$$

Next, we adapt the divergence decomposition from Lemma 15.1 to our setting. More precisely, we show it holds for *certain* pairs of instances $(\mu, \mathcal{I})$ and $(\mu', \mathcal{I}')$. Specifically, let $\mu' = \mu$ and $\mathcal{I}' = \{\mathcal{I}'_j\}_{j\in[M]}$, where

$$\mathcal{I}'_j = \begin{cases} \mathcal{I}_j \setminus \{i\}, & j = m \\ \mathcal{I}_j \cup \{i\}, & j = m+1 \\ \mathcal{I}_j, & \text{otherwise} \end{cases} .$$

In words, the instance $(\mu', \mathcal{I}')$ is identical to $(\mu, \mathcal{I})$, except i plays the $(m+1)^{\text{th}}$ bandit in the former and the m^{th} in the latter. For simplicity, we thus write $\mathbb{P}_{\mathcal{I}} = \mathbb{P}_{\mu,\mathcal{I},\pi}$ and $\mathbb{P}_{\mathcal{I}'} = \mathbb{P}_{\mu',\mathcal{I}',\pi}$ for the probability measures (26), and $\mathbb{E}_{\mathcal{I}}$ and $\mathbb{E}_{\mathcal{I}'}$ for the associated expectations. We claim

$$D(\mathbb{P}_{\mathcal{I}}, \mathbb{P}_{\mathcal{I}'}) = \sum_{k=1}^K \mathbb{E}_{\mathcal{I}}[T_k^{(i)}(T)] D(p_{\mu_{m,k}}, p_{\mu_{m+1,k}}), \tag{29}$$

where D denotes the KL divergence. The proof of (29) follows the essentially the same logic as that of Lemma 15.1. The key difference is that our density (27) includes the product term $\prod_{t\in[T],k\in[K],n\in[N]\setminus\{i\}} p_{\mu_{c(n),k}}(x_{t,k,n})$, which arises from the reward sequences $\mathcal{X}$. However, since agents $n \neq i$ have the same reward distributions $p_{\mu_{c(n)},k}$ under both instances $(\mu, \mathcal{I})$ and $(\mu', \mathcal{I}')$, these product terms cancel in the proof, which yields (29).

Finally, we prove (25) using an approach similar to Theorem 16.2. We begin by upper bounding the summands on the right side of (29). For $k = k^*(m)$ (the optimal arm for i in the instance $\mathcal{I}$), we have $\mu_{m,k^*(m)} = \mu_{m+1,k^*(m)}$ by assumption, which implies $D(p_{\mu_{m,k^*(m)}}, p_{\mu_{m+1,k^*(m)}}) = 0$. For $k \neq k^*(m)$, we have

$$D(p_{\mu_{m,k}}, p_{\mu_{m+1,k}}) = \frac{(\mu_{m,k} - \mu_{m+1,k})^2}{2} \leq \frac{1}{2} \leq \frac{\Delta_{m,k}}{2\Delta_m},$$

where we used the fact that p_ν is Gaussian with mean ν and unit variance, the assumption $\mu_{m,k} \in [0,1]$, and the definition of Δ_m, respectively. Combining arguments and using (29), we thus obtain

$$D(\mathbb{P}_{\mathcal{I}}, \mathbb{P}_{\mathcal{I}'}) \leq \frac{1}{2\Delta_m} \sum_{k\in[K]\setminus\{k^*(m)\}} \mathbb{E}_{\mathcal{I}}[T_k^{(i)}(T)]\Delta_{m,k} = \frac{\mathbb{E}_{\mathcal{I}}[R_T^{(i)}]}{2\Delta_m}. \tag{30}$$

Next, we lower bound $D(\mathbb{P}_{\mathcal{I}}, \mathbb{P}_{\mathcal{I}'})$. Let $A = \{T^{(i)}_{k^*(m+1)}(T) > T/2\}$ be the event that i plays the best arm for the $(m+1)^{\text{th}}$ bandit at least $T/2$ times. Then by assumption $k^*(m) \neq k^*(m+1)$, we know that

$$\mathbb{E}_{\mathcal{I}}[R_T^{(i)}] \geq \frac{T\Delta_{m,k^*(m+1)}}{2}\mathbb{P}_{\mathcal{I}}(A) \geq \frac{T\Delta_m}{2}\mathbb{P}_{\mathcal{I}}(A) \geq \frac{T\Delta}{2}\mathbb{P}_{\mathcal{I}}(A),$$

and by definition of $\mathcal{I}'$ (where i plays the $(m+1)^{\text{th}}$ bandit), we similarly have $\mathbb{E}_{\mathcal{I}'}[R_T^{(i)}] \geq T\Delta\mathbb{P}_{\mathcal{I}'}(A^C)/2$. Combining these inequalities and using Theorem 14.2 of (Lattimore & Szepesvári, 2020), we get

$$\mathbb{E}_{\mathcal{I}}[R_T^{(i)}] + \mathbb{E}_{\mathcal{I}'}[R_T^{(i)}] \geq \frac{T\Delta}{2}\left(\mathbb{P}_{\mathcal{I}}(A) + \mathbb{P}_{\mathcal{I}'}(A^C)\right) \geq \frac{T\Delta}{4}\exp\left(-D(\mathbb{P}_{\mathcal{I}}, \mathbb{P}_{\mathcal{I}'})\right). \tag{31}$$

Therefore, combining (30) and (31) and letting $T \to \infty$, we obtain

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\mathcal{I}}[R_T^{(i)}]}{\log T} \geq 2\Delta_m \liminf_{T\to\infty} \frac{D(\mathbb{P}_{\mathcal{I}}, \mathbb{P}_{\mathcal{I}'})}{\log T} \geq 2\Delta_m \liminf_{T\to\infty} \left(1 - \frac{\log(4(\mathbb{E}_{\mathcal{I}}[R_T^{(i)}] + \mathbb{E}_{\mathcal{I}'}[R_T^{(i)}])/\Delta)}{\log T}\right) = 2\Delta_m,$$

where the equality holds by the uniformly efficient assumption (which states that the group regret on any problem instance, and hence the individual regrets $\mathbb{E}_{\mathcal{I}}[R_T^{(i)}]$ and $\mathbb{E}_{\mathcal{I}'}[R_T^{(i)}]$, are $o(T^\gamma)$ for arbitrarily small $\gamma > 0$).