Dynamic Games and Applications (2023) 13:307–325
https://doi.org/10.1007/s13235-022-00451-1

Asymptotic Optimality for Decentralised Bandits

Conor J. Newton[1] · **Ayalvadi Ganesh**[1] · **Henry W. J. Reeve**[1]

Accepted: 4 May 2022 / Published online: 20 June 2022

Abstract
We consider a large number of agents collaborating on a multi-armed bandit problem with a large number of arms. The goal is to minimise the regret of each agent in a communication-constrained setting. We present a decentralised algorithm which builds upon and improves the *Gossip-Insert-Eliminate* method of Chawla et al. (International conference on artificial intelligence and statistics, pp 3471–3481, 2020). We provide a theoretical analysis of the regret incurred which shows that our algorithm is asymptotically optimal. In fact, our regret guarantee matches the asymptotically optimal rate achievable in the full communication setting. Finally, we present empirical results which support our conclusions.

1 Introduction

The classical stochastic multi-armed bandit problem is specified by a collection of probability distributions $\{P_k\}_{k=1}^K$, commonly referred to as arms. Here, there is a single agent which plays an arm I_t taking values in $[K] := \{1, \ldots, K\}$ at each time step $t \in [T]$ and receives an associated reward $X_t \sim P_{I_t}$. The agent's goal is to minimise the expected regret $\mathbb{E}[R_T] = T\mu_\star - \sum_{t=1}^T \mathbb{E}[X_t]$, where μ_k is the expectation of a random variable with distribution P_k, and $\star := \text{argmax}_{k\in[k]}\, \mu_k$ is the largest mean of the arms. The agent's decisions must be made using only the knowledge acquired from previous actions and observed rewards.

We consider an extension of this problem where there are multiple agents collaborating on a multi-armed bandit problem [3, 17]. The agents may communicate with one another, and the agents decision's of which arms to play can be made using the information from both their own reward history, and from the sequence of messages received from other agents. However, communication between agents is tightly restricted as described in Sect. 2. Specifically, time is divided into growing phases and each agent may receive only one message per phase. Furthermore, a message is limited to recommending the id of a single arm; no additional information may be exchanged. We show in Theorem 3.1 that, even with these restrictions

This article is part of the topical collection "Multi-agent Dynamic Decision Making and Learning" edited by Konstantin Avrachenkov, Vivek S. Borkar and U.Jayakrishnan Nair.

✉ Conor J. Newton
conor.newton@bristol.ac.uk

[1] School of Mathematics, University of Bristol, Bristol , UK

on communication, it is possible to asymptotically match the optimal total regret achievable with unlimited communication.

The multi-agent version of the multi-armed bandit problem is motivated by multiple applications:

- *Decentralised web advertising* Consider the problem of selecting an advertisement to be displayed on a website. The website will want to optimise which advertisements it chooses to display with the goal of maximising click-through rate. To do this, the website should react to its users interactions. Additionally, the website could be hosted on multiple web servers operating in parallel to serve many users at once. Here, each web server will select an advertisement to display each time a user requests the page. The web servers can benefit from sharing information on the performance of each advertisement. The web server communication will be limited by bandwidth and potentially geographical constraints. This motivates the communication constraints imposed in our setting. A bandit-based approach to ad selection is considered in [16].
- *Decentralised network routing* In this problem, a user wants to send data over a network between two computers as fast as possible. In the network, there are many paths the data can be sent along. These paths will have different latency's and the user can measure the latency of a path after using it. A bandit algorithm can be used with this information to choose the best path to send the data along. In addition, it is likely that multiple people are using the same network at once. These users can collaborate to the find the best paths faster. In [4, 8], bandit algorithms are applied to network routing problems.
- *Multi-robot systems* Multi-agent multi-armed bandit algorithms can be used to operate multi-robot systems. In particular [18], consider the problem of foraging using a group of robots. In this problem, the robots need to search for the sites which they can forage the most from. The robots can communicate with other nearby robots over a wireless network which can be used to quickly identify the best sties to forage from. Since the communication is constrained locally, it is similar to the setting we are considering.

It should be noted that in these examples contextual information could be used to improve the decision-making and additionally the expected rewards may be non-stationary. In the network routing example, there could be a penalty if more than one user chooses a single path. However, we work in a simplified settings where we do not make these assumptions, which is currently more feasible to prove results in.

There has recently been growing interest in multi-agent multi-armed bandits. A setting in which agents communicate with a central node is considered in [9], while [2, 5, 15, 19] consider settings where agents can communicate *rewards* (not just arm ids) with their neighbours. Kola et al. [10] considered a model where agents observe the rewards of their neighbours. We follow the setting introduced in [3, 17] where agents may only communicate arm ids and do this through a gossiping PULL protocol. This ensures that in each round the number of bits communicated is bounded and relatively small. Additionally, we prefer a decentralised system over a centralised system as it does not have a single point of failure. Furthermore, a centralised system would have a high-communication overhead through its central node which may be limiting in applications. In a recent work [1], the authors introduced a method for achieving nearly minimax optimal regret in the gossiping and decentralised setting.

A central problem in the multi-armed bandit literature is the search for algorithms which perform optimally in the asymptotic regime of the time horizon T tending to infinity. Returning to the single-agent setting, Lai and Robbins [11] proved a fundamental lower bound on the regret incurred by any *consistent* algorithm. Here, we say that an algorithm is *consistent*

if it achieves subpolynomial regret for all possible values of $\{P_k\}_{k=1}^K$. (This precludes trivial algorithms like one which always selects a specific arm and has zero regret if that happens to be the best arm.) Lai and Robbins [11] showed that the regret of any consistent algorithm satisfies the following lower bound:

$$\liminf_{T\to\infty} \frac{\sum_n \mathbb{E}[\mathcal{R}_T]}{\log(T)} \geq \sum_{i\neq\star} \frac{\mu_\star - \mu_i}{\mathrm{KL}(P_i,\, P_\star)}, \tag{1}$$

where KL denotes the Kullback–Leibler divergence. A significant breakthrough was achieved by [6, 14] who demonstrated that this bound is attained by the KL-UCB algorithm in the Bernoulli reward setting.

In this work, we consider the question of asymptotic optimality in the decentralised multi-agent setting. Our contributions are as follows:

- We present a decentralised algorithm which builds upon and improves the *Gossip-Insert-Eliminate* method of Chawla et al. [3]. This algorithm leverages two innovations which reduce the amount of superfluous exploration. Firstly, we include a more efficient elimination mechanism which reduces the number of arms considered by each agent at any given time. Secondly, in the spirit of [6, 14], we use KL-type confidence intervals, rather than Hoeffding-type confidence intervals.
- We provide a theoretical analysis of the expected regret of the algorithm we propose (Theorem 3.1). We show that it is optimal in the asymptotic regime. In particular, the aggregate expected regret matches the lower bound implied by (1), showing that our algorithm performs at least as well as any multi-agent algorithm, even with access to unlimited communication resources, in the asymptotic regime.
- We find a regret bound which has a clear dependence on the graph structure (Theorem 3.15). This is done in the setting where agents pull recommendations uniformly at random from their neighbours. This will allow us to leverage an existing result of Giakkoupis [7] on the spreading time of a rumour on a network following a PULL protocol where time is discrete. We conclude this section by comparing the impact that three different graphs (complete, star and cycle) have on the scaling of regret bound.
- We present empirical results that demonstrate that our algorithm performs well in a wide variety of settings, with lower finite sample regret than the baseline of [3] (Figs. 1, 2). Interestingly, both modifications lead to a consistent improvement for a range of different values of the gap between best and second-best arm.

2 Setting and Algorithm

We now present our problem setting and algorithm. Throughout N will denote the number of agents, T the number of time steps and K the number of arms. Let $X_{k,s}^n$ taking values in $\{0, 1\}$ denote the reward that agent $n \in [N]$ receives by playing arm $k \in [K]$ for the s^{th} time. We assume that these are i.i.d. Bernoulli(μ_k) random variables. Let $\star \in \operatorname{argmax} \mu_k$ and let $\mu_\star := \max_{k\in[K]} \mu_k$. We assume throughout that there is a unique best arm, so $\star$ is uniquely defined.

Communication between agents is constrained by a strictly increasing sequence $(A_j)_{j\in\mathbb{N}}$ of communication rounds and an $N \times N$ probability matrix P as follows. The time horizon $[T]$ is partitioned into phases, with phase j consisting of time steps t for which $A_{j-1} < t \leq A_j$ where $A_0 := 0$. Communication between agents only occurs once per phase, on the time steps A_j, after each agent has played an arm. On these time steps, each agent PULLs a message

from exactly one of their neighbours chosen at random, independently of everything else. The neighbouring agent is selected randomly according to P, with $P(n,q)$ denoting the probability that agent n will receive a message from agent q at the end of each phase j. We let $Q \equiv Q_j^n \sim P(n,\cdot)$ be the random variable corresponding to the agent who sends a recommendation to agent n at the end of phase j. The message, from agent Q_j^n to n, is an arm recommendation O_n^j taking values in $[K]$.

To ensure that the recommendations can spread to all agents, we assume that P is strongly connected, meaning that for any two agents $i, j \in [N]$ with $i \neq j$ there exists a sequence of agents $n_1, \ldots, n_l \in [N]$ such that $P(i,n_1), P(n_1,n_2), \ldots, P(n_{k-1},n_l), P(n_l,j) > 0$.

Let I_t^n denote the random variable, taking values in $[K]$, which specifies the index of the arm played by agent n in round t. This must be a measurable function of an agent's previous reward history and the previous messages they have received. We let $V_k^n(t) := \sum_{s=1}^t \mathbb{1}\{I_s^n = k\}$ denote the number of times agent n plays arm k in the first t rounds. Let $X^n(t) := X^n_{I_t^n, V_k^n(t)}$ denote the reward received by agent n in round t.

The goal for each agent $n \in [N]$ is to minimise their expected regret,

$$\mathbb{E}[\mathcal{R}_T^n] := T \cdot \mu_\star - \sum_{t\in[T]} \mathbb{E}[X^n(t)].$$

Our algorithm (Algorithm 1) is based on the Gossip-Insert-Eliminate algorithm of [3]. A key feature of this algorithm is that, during each phase j, each agent plays only a small subset of the K arms which we call its active set. This is made up of a sticky set of arms, which remains unchanged over time for each agent, and additional arms which evolve over time based on recommendations.

In our algorithm, we begin by partitioning $[K]$ into nearly equal-sized sets $\{S_\circ^n\}_{n\in[N]}$, so that for each agent $n \in [N]$, $S_\circ^n$ will act as the associated sticky set. The active sets are initialised to be the same as the sticky sets, but will grow over time due to recommendations and shrink due to eliminations of non-sticky arms. In each phase $j \in \mathbb{N}$, each agent $n \in [N]$ will only play arms from the active set S_j^n. For the first phase $j = 1$, we initialise each $S_1^n = S_\circ^n$. In subsequent phases $j > 1$ the active set S_{j+1}^n consists of $S_\circ^n$, along with (potentially) additional arms.

We assume that each agent n is aware of $S_\circ^n$, its own set of arms within the partition, a priori. That is, $S_\circ^n$ may be taken as an input to our algorithm. Let $\hat{\mu}_{k,s}^n := \frac{1}{s}\sum_{i=1}^s X_{k,i}^n$. Denote by $\hat{\mu}_k^n(t) := \hat{\mu}^n_{k,V_k^n(t)}$ the mean reward obtained by agent n from arm k in the first t time steps.

We let M_j^n denote the most played arm by agent n in phase j so

$$M_j^n = \text{argmax}_{k\in[K]}\{V_k^n(A_j) - V_k^n(A_{j-1})\}.$$

Following [3], when an agent $q \in [N]$ is asked for an arm recommendation at the end of phase j, its recommendation will be its most played arm for that phase. Hence, when $Q \equiv Q_j^n \sim P(n,\cdot)$ communicates with agent $n \in [N]$ at the end of phase j, the recommendation will be $O_n^j = M_j^Q$.

Our algorithm (Algorithm 1) differs from that of [3] in two important respects.

Firstly, we use a more efficient elimination scheme. More precisely, in each phase $j+1$, the new active set S_{j+1}^n will be constituted by the sticky set $S_\circ^n$, together with the agent's most played arm M_j^n during phase j, and the recommendation, O_j^n, it receives at the end of phase j. We assume that the random variable Q_j^n is independent from everything else.

The intuition is that, eventually, the best arm will become known to all agents, and M_j^n and O_j^n will both be equal to $\star$; consequently, S_j^n will be $S_\circ^n \cup \{\star\}$.

Secondly, we use tighter KL based confidence intervals, following [6]. To define our KL upper confidence bounds, we first let $\mathrm{KL} : [0,1]^2 \to \mathbb{R} \cup \{\infty\}$ be the Kullback–Leibler divergence for two Bernoulli random variables and introduce a function $f_\alpha(t) = 1 + t^\alpha \log^2(t)$ indexed by α. The upper confidence bound for arm k at agent n at time t is defined by

$$\mathrm{U}_{k,\alpha}^n(t-1) := \max\left\{u \in [0,1] : \mathrm{KL}(\hat{\mu}_k^n(t-1), u) \le \frac{\log(f_\alpha(t))}{V_k^n(t-1)}\right\} \tag{2}$$

when $V_k^n(t-1) > 0$ and $\mathrm{U}_k^n(t-1) := \infty$ otherwise. When α is clear from context we suppress it for notational convenience.

Algorithm 1 gives the steps that each agent $n \in [N]$ will perform synchronously.

Algorithm 1: Asymptotically Optimal Gossiping Bandits (AOGB) for agent n

Input: Communication rounds A_j, Sticky set $S_\circ^n$, Communication probabilities matrix P, exploration parameter α.

Init: $j \leftarrow 1$ **and** $S_1^n \leftarrow S_\circ^n$
for $t \in \mathbb{N}$ **do**
  $I_t^n \leftarrow \mathrm{argmax}_{k \in S_j^n} \mathrm{U}_{k,\alpha}^n(t-1)$; `// Select arm to play according to (2)`
  if $t == A_j$ **then**
    $Q \leftarrow P(n, \cdot)$; `// Choose neighbour to get recommendation from`
    $O_j^n \leftarrow M_j^Q$; `// Save recommendation from neighbour`
    $S_{j+1}^n \leftarrow S_\circ^n \cup \{O_j^n, M_j^n\}$; `// Update the active set of arms`
    $j \leftarrow j + 1$
  end
end

3 Theoretical Analysis and Regret Bound

We now present our asymptotically optimal regret bound for Algorithm 1.

Theorem 3.1 *Suppose there exists $C \ge 1$, $\theta > 0$ such that $C^{-1} j^\theta \le A_j - A_{j-1} \le C j^\theta$ for all $j \in \mathbb{N}$ and suppose that all agents select arms with Algorithm 1 with $\alpha = 1$. Then for each agent $n \in [N]$, we have the asymptotic bound*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathcal{R}_T^n]}{\log T} \le \sum_{k \in S_\circ^n \setminus [\star]} \frac{\mu_\star - \mu_k}{\mathrm{KL}(\mu_k, \mu_\star)}.$$

Let us consider the class of centralised algorithms $\mathcal{A}$ in which an arm I_t^n in $[K]$ is selected for each agent $n \in [N]$ and each time step $t \in [T]$ based on the combined reward history of all the agents up to time t. We let $\mathcal{A}_{\mathrm{const}} \subseteq \mathcal{A}$ denote the subset of those which are *consistent*, i.e. achieve subpolynomial total regret $\sum_n \mathbb{E}[\mathcal{R}_T^n]$ for any instance of the multi-armed bandit problem. It follows from the result of Lai and Robbins (1) that for any algorithm in the class $\mathcal{A}_{\mathrm{const}}$,

$$\liminf_{T\to\infty} \frac{\sum_n \mathbb{E}[\mathcal{R}_T^n]}{\log(T)} = \liminf_{T\to\infty} \left(\frac{\log(NT)}{\log(T)} \cdot \frac{\sum_n \mathbb{E}[\mathcal{R}_T^n]}{\log(NT)}\right) \ge \sum_{i \ne \star} \frac{\mu_\star - \mu_i}{\mathrm{KL}(\mu_i, \mu_\star)}. \tag{3}$$

Now note that we can view the class $\mathcal{A}$ as the collection of all multi-agent algorithms, with or without communication constraints. In particular, the class of decentralised multi-agent with strong communication constraints we consider in this paper correspond to a computationally attractive subset of $\mathcal{A}$. Observe that by summing over $n \in [N]$ in the regret bound given in Theorem 3.1, we see that total regret of the system for our algorithm matches the lower bound given by (3) for the full communication setting. This implies that our algorithm (with limited communication) performs just as well as any algorithm, even with access to unlimited communication constraints, in the asymptotic regime.

In addition to Theorem 3.1, which only certifies the performance in the asymptotic regime, in Sect. 3.1, we continue the analysis of Algorithm 1 to derive a finite sample bound which has a clear dependence on the graph structure. Furthermore, in Sect. 4 we shall see that our algorithm also performs well empirically on a broad range of simulated data.

Before presenting the main proof of Theorem 3.1 we shall present a brief sketch. The argument hinges upon a random time $\hat{\tau}$ which corresponds to a phase after which all of the active sets S_j^n become fixed. After this time, all of the active sets become $S_\circ^n \cup \{\star\}$, which leads to an asymptotic regret bound for agent n governed by the relationship between μ_k and $\mu_\star$ for $k \in [K]$. The crucial difficulty then is to bound $\mathbb{E}[A_{\hat{\tau}}]$, the expected time until the end of phase $\hat{\tau}$. To bound $\mathbb{E}[A_{\hat{\tau}}]$, we show that, provided the phase lengths $A_j - A_{j-1}$ are sufficiently large in relationship to the gap, the probability of a suboptimal arm being the most played and subsequently being recommended decays exponentially.

To bound the per agent expected regret of this system, we divide time into two parts; before $A_{\hat{\tau}}$ and after $A_{\hat{\tau}}$. The regret before time $A_{\hat{\tau}}$ is trivially upper bounded by $\mathbb{E}[A_{\hat{\tau}}]$, and since, after time $A_{\hat{\tau}}$, the set of active arms for each remains fixed, this reduces to bounding the expected regret of a single-agent multi-armed bandit problem. For this, we consider the approach given in [12], where we show that for a late enough time, we expect that the KL-UCB for the optimal arm does not fall far below its true mean, and additionally, the KL-UCB for all suboptimal arms does not exceed this value often.

The proof of Theorem 3.15 (finite sample bound) is similar to that of 3.1, but they differ when bounding $\mathbb{E}[A_\tau]$. The main difference is that Theorem 3.15 uses Lemma 3.13 instead of Lemma 3.8.

We now proceed with proof itself, which goes through a sequence of lemmas. Let us begin by introducing some notation used throughout. Firstly, fix the exploration function $f(t) := 1 + t\log^2(t)$ (i.e. $\alpha = 1$). Next we define the suboptimality gap for each arm $k \in [K]$ by

$$\Delta_k := \mu_\star - \mu_k,$$

and we define the smallest suboptimality gap,

$$\Delta_{\min} := \min_{k \in [K]\setminus\{\star\}} \Delta_k > 0.$$

For each $\epsilon \in (0, \Delta_{\min})$ and each agent $n \in [N]$, we define a random variable

$$\kappa_\epsilon^n := \min\left\{ t \in \mathbb{N} : \max_{s \in [T]} \left(\underline{d}\left(\hat{\mu}_{\star,s}^n, \mu_\star - \epsilon\right) - \frac{\log(f(t))}{s} \right) \leq 0 \right\},$$

where $\underline{d}(p,q) := \mathrm{KL}(p,q) \cdot \mathbb{1}\{p \leq q\}$. This random variable is the time whereafter the KL upper confidence bound of the optimal arm will not fall below $\mu_\star - \epsilon$, no matter how many times the optimal arm has been played.

Next we define for every $\epsilon \in (0, \Delta_{\min})$, for every agent $n \in [N]$ and for every suboptimal arm $k \in [K] \setminus \{\star\}$,

$$\nu^n_{\epsilon,k} := \sum_{s=1}^{T} \mathbb{1}\left\{ \mathrm{KL}(\hat{\mu}^n_{k,s}, \mu_\star - \epsilon) \leq \frac{\log(f(T))}{s} \right\}.$$

This random variable is an upper bound for the number of times the KL upper confidence bound of a suboptimal k arm exceeds $\mu_\star - \epsilon$.

Together, these random variables allow us to bound the regret. This is because after time κ^n_ϵ the number of times any suboptimal arm is played is bounded above by $\nu^n_{\epsilon,k}$. To do this, we require the following lemmas (Lemmas 3.2, 3.3), which are effectively the same as [12, Lemma 10.7 & Lemma 10.8].

Lemma 3.2 *For $\epsilon \in (0, \Delta_{\min})$, $\max_{n\in[N]} \mathbb{E}[\kappa^n_\epsilon] \leq 2/\epsilon^2$.*

Lemma 3.3 *For $\epsilon \in (0, \Delta_{\min})$ and $n \in [N]$, we have*

$$\mathbb{E}[\nu^n_{\epsilon,k}] \leq \inf_{\tilde{\epsilon}\in(0,\Delta_k-\epsilon)} \left(\frac{\log f(T)}{\mathrm{KL}(\mu_k + \tilde{\epsilon}, \mu_\star - \epsilon)} + \frac{1}{2\tilde{\epsilon}^2} \right).$$

To continue the proof, we define some further random variables that concern the optimal arm and its movement around the network.

Firstly, for each agent $n \in [N]$ and each phase j we define a random variable

$$\chi^n_j := \mathbb{1}\{\star \in S^n_j,\ M^n_j \neq \star, A_{j-1} \geq \kappa^n_\circ\},$$

where $\kappa^n_\circ := \kappa^n_{\Delta_{\min}/2}$. This variable indicates whether an agent has the best arm but has not played it most over the phase j (and therefore it will not recommend it). Additionally, the condition $A_{j-1} \geq \kappa^n_\circ$ demands that we are in a late enough phase which is necessary for Lemma 3.7.

For each agent $n \in [N]$, we define the following random variables:

$$\begin{aligned}
\hat{\tau}^n_{\mathrm{stab}} &:= \min\{j \in \mathbb{N} \ :\ A_{j-1} \geq \kappa^n_\circ, \forall j' \geq j,\ \chi^n_{j'} = 0\} \\
\hat{\tau}_{\mathrm{stab}} &:= \max_{n\in[N]} \hat{\tau}^n_{\mathrm{stab}} \\
\hat{\tau}^n_{\mathrm{spr}} &:= \min\{j \geq \hat{\tau}_{\mathrm{stab}} \ :\ \star \in S^n_j\} - \hat{\tau}_{\mathrm{stab}} \\
\hat{\tau}_{\mathrm{spr}} &:= \max_{n\in[N]} \hat{\tau}^n_{\mathrm{spr}} \\
\hat{\tau} &:= \hat{\tau}_{\mathrm{stab}} + \hat{\tau}_{\mathrm{spr}}.
\end{aligned}$$

These random variables highlight two key timings of the system (for each agent). The first being the *stabilisation* phase $\hat{\tau}^n_{\mathrm{stab}}$; this is the phase whereafter agent n will always recommend the best arm if it has the best arm. The second is the *spreading* time $\hat{\tau}_{\mathrm{spr}}$; this is the number of phases after $\hat{\tau}^n_{\mathrm{stab}}$, where agent n will have the best arm for all subsequent phases. After phase $\hat{\tau}$, each agent will have the best arm and only recommend the best arm; therefore, the set of active arms for each agent will be subsequently fixed. Lemma 3.4 proves this.

Lemma 3.4 *For all phases $j > \hat{\tau}$ and all $n \in [N]$, we have $S^n_j = S^n_\circ \cup \{\star\}$.*

Proof For each agent $n \in [N]$, we see by induction that for any phase $j \geq \hat{\tau}^n_{\mathrm{spr}} + \hat{\tau}_{\mathrm{stab}}$, we have that $M^n_j = \star \in S^n_j$.

Moreover, since $S^n_{j+1} = S^n_\circ \cup \{M^n_j, M^Q_j\}$ for some agent Q in $[N]$, it follows that $S^n_{j+1} = S^n_\circ \cup \{\star\}$, for all $j \geq \hat{\tau} = \hat{\tau}_{\mathrm{stab}} + \hat{\tau}_{\mathrm{spr}}$. □

In the following lemma, we bound the number of times a suboptimal arm is played after the phase $\hat{\tau}$.

Lemma 3.5 *For each agent $n \in [N]$ and each suboptimal arm $k \in [K]\backslash\{\star\}$, we have*

$$\sum_{t=A_{\hat{\tau}}+1}^{T} \mathbb{1}\left\{I_t^n = k\right\} \leq \begin{cases} \inf_{\epsilon\in(0,\Delta_{\min})} \left\{ v_{\epsilon,k}^n + \kappa_\epsilon^n \right\} & \text{if } k \in S_\circ^n \\ 0 & \text{if } k \notin S_\circ^n. \end{cases}$$

Proof Fix an agent $n \in [N]$. First note that by Lemma 3.4 we have $S_j^n = S_\circ^n \cup \{\star\}$ for all phases $j > \hat{\tau}$. In particular, this means that $I_t^n \notin S_\circ^n \cup \{\star\}$ cannot occur for $t \geq A_{\hat{\tau}} + 1$. Now take $\epsilon \in (0, \Delta_{\min})$ and consider a suboptimal arm $k \in S_\circ^n\backslash[\star]$. If $I_t^n = k$ for some $t \geq (A_{\hat{\tau}} + 1) \vee \kappa_\epsilon^n$, then we must have $\mathrm{U}_k^n(t-1) \geq \mathrm{U}_\star^n(t-1) \geq \mu_\star - \epsilon$, and hence,

$$\mathrm{KL}(\hat{\mu}_{k,V_k^n(t-1)}^n, \mu_\star - \epsilon) \leq \frac{\log(f(t))}{V_k^n(t-1)} \leq \frac{\log(f(T))}{V_k^n(t-1)}.$$

Consequently,

$$\sum_{t=(A_{\hat{\tau}}+1)\vee\kappa_\epsilon^n}^{T} \mathbb{1}\left\{I_t^n = k\right\} \leq \sum_{t=(A_{\hat{\tau}}+1)\vee\kappa_\epsilon^n}^{T} \mathbb{1}\left\{ I_t^n = k \text{ and } \mathrm{KL}(\hat{\mu}_{k,V_k^n(t-1)}^n, \mu_\star - \epsilon) \leq \frac{\log(f(T))}{V_k^n(t-1)} \right\} \leq v_{\epsilon,k}^n,$$

and therefore,

$$\sum_{t=A_{\hat{\tau}}+1}^{T} \mathbb{1}\left\{I_t^n = k\right\} \leq v_{\epsilon,k}^n + \kappa_\epsilon^n.$$

The result then follows by taking an infimum over $\epsilon \in (0, \Delta_{\min})$. □

This leads to the following regret bound.

Corollary 3.6 *For each agent $n \in [N]$, we have*

$$\mathbb{E}[\mathcal{R}_T^n] \leq \mathbb{E}[A_{\hat{\tau}}] + \sum_{k\in S_\circ^n\backslash[\star]} \Delta_k \inf_{\epsilon\in\left(0,\frac{\Delta_{\min}}{2}\right)} \left\{ \frac{\log f(T)}{\mathrm{KL}(\mu_k + \epsilon, \mu_\star - \epsilon)} + \frac{3}{\epsilon^2} \right\}.$$

For the remainder of the proof, we must show that $\mathbb{E}[A_{\hat{\tau}}]$ may be bounded independently of T.

We do this as follows: In Lemma 3.7, we show that if the length of a phase is large enough, then the expected value of χ_j^n decays exponentially with phase length; in Lemmas 3.8 and 3.10, we find high probability bounds for $\hat{\tau}_{\text{stab}}$ and $\hat{\tau}_{\text{spr}}$, respectively; and we conclude in 3.11 by showing $\mathbb{E}[A_{\hat{\tau}}]$ is finite and does not depend on the time horizon T.

Lemma 3.7 *For every phase $j \in \mathbb{N}$ such that $A_j - A_{j-1} \geq \frac{8}{\Delta^2}\left(\frac{K}{N} + 3\right)\log f(A_j)$, we have*

$$\mathbb{E}[\chi_j^n] \leq \frac{8K}{\Delta_{\min}^2} \exp\left(-\frac{\Delta_{\min}^2 (A_j - A_{j-1})}{16(K/N + 3)} \right).$$

Proof First observe that if $\chi_j^n = 1$ then $\star \in S_j^n$, $A_{j-1} \geq \kappa_\circ^n$ and $M_j^n \neq \star$. Since $M_j^n \neq \star$, we deduce that for some $k \in [K]\backslash\{\star\}$, we have

$$V_k^n(A_j) - V_k^n(A_{j-1}) \geq \frac{A_j - A_{j-1}}{|S_j^n|} \geq \frac{A_j - A_{j-1}}{K/N + 3},$$

and so, for some $A_{j-1} < t \leq A_j$ we have $s = V_k^n(t-1) \geq \frac{A_j - A_{j-1}}{K/N+3} - 1$ and $I_t^n = k$, so $\mathrm{U}_k^n(t-1) \geq \mathrm{U}_\star^n(t-1)$ as $\star \in S_j^n$. Since $t \geq A_{j-1} \geq \kappa_\circ^n$ we deduce that $\mathrm{U}_k^n(t-1) \geq \mathrm{U}_\star^n(t-1) \geq \mu_\star - \Delta_{\min}/2$. Hence, by Pinsker's inequality

$$2\left(\hat{\mu}_{k,s}^n - \mu_\star + \frac{\Delta_{\min}}{2}\right)^2 = 2\left(\hat{\mu}_k^n(t-1) - \mu_\star + \frac{\Delta_{\min}}{2}\right)^2 \leq \mathrm{KL}\left(\hat{\mu}_k^n(t-1), \mu_\star - \frac{\Delta_{\min}}{2}\right) \leq \frac{\log(f_\alpha(t))}{V_k^n(t-1)} \leq \frac{\log f(A_j)}{s}.$$

Thus, for some $k \in [K]\backslash\{\star\}$ and $s \geq \frac{A_j - A_{j-1}}{K/N+3} - 1$,

$$\hat{\mu}_{k,s}^n \geq \mu_\star - \frac{\Delta_{\min}}{2} - \sqrt{\frac{\log f(A_j)}{2s}} \geq \mu_k + \frac{\Delta_{\min}}{2} - \sqrt{\frac{\log f(A_j)}{2s}} \geq \mu_k + \frac{\Delta_{\min}}{4},$$

since $A_j - A_{j-1} \geq \frac{8}{\Delta^2}\left(\frac{K}{N} + 3\right)\log f(A_j)$. Thus, by Hoeffding's inequality we have

$$\begin{aligned}
\mathbb{E}[\chi_j^n] &\leq \sum_{k\in[K]\backslash\{\star\}} \sum_{s \geq \frac{A_j - A_{j-1}}{K/N+3} - 1} \mathbb{P}\left[\hat{\mu}_{k,s}^n \geq \mu_k + \frac{\Delta_{\min}}{4}\right] \\
&\leq (K-1) \sum_{s \geq \frac{A_j - A_{j-1}}{K/N+3} - 1} \exp\left(-\frac{s\Delta_{\min}^2}{8}\right) \\
&\leq K \int_{\frac{A_j - A_{j-1}}{K/N+3} - 2}^{\infty} \exp\left(-\frac{s\Delta_{\min}^2}{8}\right) \mathrm{d}s \\
&\leq \frac{8K}{\Delta_{\min}^2} \exp\left(-\frac{\Delta_{\min}^2 (A_j - A_{j-1})}{16(K/N+3)}\right).
\end{aligned}$$

□

In what follows, we let $p_{\min} := \min\left(\{P(i,j)\}_{(i,j)\in[N]^2}\backslash\{0\}\right)$ and let $\mathrm{diam}(P)$ denote the maximum length of a directed path between two distinct nodes corresponding to the graph induced by P.

Lemma 3.8 *For $\xi \in \mathbb{N}$, $\mathbb{P}(\hat{\tau}_{\mathrm{spr}} \geq \xi) \leq N(1 - p_{\min}^{\mathrm{diam}(P)})^{\left\lfloor \frac{\xi}{2\mathrm{diam}(P)} - 1 \right\rfloor}$.*

Proof Recall that $\hat{\tau}_{\mathrm{spr}}$ is the number of phases *since* $\hat{\tau}_{\mathrm{stab}}$, so we can assume that if an agent has the best arm it will recommend it. Therefore, to find an upper bound for this probability, we consider a single path from an agent with the optimal arm ($n_\star$) to the chosen node n and the probability that there exists a single node in this path that does not request a recommendation from the prior node. And therefore, the best arm does not spread along this path.

Fix an agent $n \in [N]$ and choose a sequence of nodes $(\ell_i)_{i\in[q]\cup\{0\}} \in [N]^q$ with $q \leq \mathrm{diam}(P)$ and such that $\ell_0 = n_\star$, $\ell_q = n$ and $P(\ell_i, \ell_{i-1}) > 0$ for each $i \in [q]$. Note that the definition of $\mathrm{diam}(P)$ entails the existence of at least one such a sequence. Recall that we let $Q_j^{\tilde{n}}$ denote the node which sends a message to agent $\tilde{n}$ and the end of phase j. Let $m = \lfloor \xi/(2q) - 1 \rfloor$ and observe that if for some $j_0 \in \{\hat{\tau}_{\mathrm{stab}}, \ldots, \hat{\tau}_{\mathrm{stab}} + 2mq\}$ we have $Q_j^{\ell_{j-j_0}} = \ell_{j-j_0-1}$ for $j \in \{j_0+1, \ldots, j_0+q\}$ then $\hat{\tau}_{\mathrm{spr}}^n + \hat{\tau}_{\mathrm{stab}} \leq j_0 + q < \xi + \hat{\tau}_{\mathrm{stab}}$. Hence,

we have

$$\begin{aligned}
\mathbb{P}(\hat{\tau}^n_{\text{spr}} \geq \xi) &\leq \mathbb{P}\left(\bigcap_{j_0 - \hat{\tau}_{\text{stab}} \in \{0,1,\ldots,2mq\}} \bigcup_{j \in \{j_0+1,\ldots,j_0+q\}} \left\{ Q_j^{\ell_{j-j_0}} \neq \ell_{j-j_0-1} \right\} \right) \\
&\leq \mathbb{P}\left(\bigcap_{j_0 - \hat{\tau}_{\text{stab}} \in \{0,2q,\ldots,2mq\}} \bigcup_{j \in \{j_0+1,\ldots,j_0+q\}} \left\{ Q_j^{\ell_{j-j_0}} \neq \ell_{j-j_0-1} \right\} \right) \\
&= \prod_{j_0 - \hat{\tau}_{\text{stab}} \in \{0,2q,\ldots,2mq\}} \mathbb{P}\left(\bigcup_{j \in \{j_0+1,\ldots,j_0+q\}} \left\{ Q_j^{\ell_{j-j_0}} \neq \ell_{j-j_0-1} \right\} \right) \\
&= \prod_{j_0 - \hat{\tau}_{\text{stab}} \in \{0,2q,\ldots,2mq\}} \left\{ 1 - \mathbb{P}\left(\bigcap_{j \in \{j_0+1,\ldots,j_0+q\}} \left\{ Q_j^{\ell_{j-j_0}} = \ell_{j-j_0-1} \right\} \right) \right\} \\
&= \prod_{j_0 - \hat{\tau}_{\text{stab}} \in \{0,2q,\ldots,2mq\}} \left\{ 1 - \prod_{j \in \{j_0+1,\ldots,j_0+q\}} \mathbb{P}\left(Q_j^{\ell_{j-j_0}} = \ell_{j-j_0-1} \right) \right\} \\
&\leq (1 - p_{\min}^q)^m \leq (1 - p_{\min}^{\text{diam}(P)})^{\left\lfloor \frac{\xi}{2\text{diam}(P)} - 1 \right\rfloor}.
\end{aligned}$$

The lemma now follows by the union bound over $[N]$. □

The following lemma gives us a bound for the time at which each phase starts (and ends) using the phase lengths.

Lemma 3.9 *Suppose that there exist $C \geq 1$, $\theta > 0$ such that $C^{-1} j^\theta \leq A_j - A_{j-1} \leq C j^\theta$ for all $j \in \mathbb{N}$. Then we have $\frac{C^{-1}}{1+\theta} j^{1+\theta} \leq A_j \leq \frac{C}{1+\theta}(1+j)^{1+\theta}$ for all $j \in \mathbb{N}$.*

Proof We have that

$$A_j = \sum_{i=1}^{j} A_j - A_{j-1}$$

since $A_0 := 0$. Therefore,

$$C^{-1} \sum_{i=1}^{j} i^\theta \leq A_j \leq C \sum_{i=1}^{j} i^\theta.$$

Since j^θ is increasing, we can bound the sums as follows

$$C^{-1} \int_0^j i^\theta \mathrm{d}i \leq A_j \leq C \int_0^j (i+1)^\theta \mathrm{d}i.$$

And this gives the desired result:

$$\frac{C^{-1}}{1+\theta} j^{1+\theta} \leq A_j \leq \frac{C}{1+\theta}(1+j)^{1+\theta}.$$

□

Now define the phase $\underline{j}(\Delta_{\min}) \in \mathbb{N}$ by

$$\underline{j}(\Delta_{\min}) := 4 + \max\left(\{0\} \cup \left\{ j \in \mathbb{N} \ : \ j^\theta < \frac{8C(1+\theta)}{\Delta_{\min}^2} \left(\frac{K}{N} + 3 \right) \log f\left(\frac{C}{1+\theta}(1+j)^\theta \right) \right\} \right).$$

Note that $\underline{j}(\Delta_{\min})$ is always finite since $f(t) = O(\log t)$. This phase is conveniently defined by considering Lemma 3.9 and with the purpose of applying Lemma 3.7 in Lemma 3.10.

Lemma 3.10 *Suppose that there exist constants $C \geq 1$, $\theta > 0$ such that $C^{-1} j^\theta \leq A_j - A_{j-1} \leq C j^\theta$ for all $j \in \mathbb{N}$. Then for all $\xi \geq \underline{j}(\Delta_{\min})$ we have*

$$\mathbb{P}(\hat{\tau}_{\text{stab}} \geq \xi) \leq \sum_{n\in[N]} \mathbb{P}(\kappa_\circ^n > C^{-1}(\xi-2)^{1+\theta}) + \frac{8KN}{\Delta_{\min}^2} \sum_{j\geq\xi} \exp\left(-\frac{\Delta_{\min}^2 j^\theta}{16C(K/N+3)}\right).$$

Proof Fix an agent $n \in [N]$ and suppose that $\hat{\tau}_{\text{stab}}^n \geq \xi$. Since $\hat{\tau}_{\text{stab}}^n := \min\{j \in \mathbb{N} : A_{j-1} \geq \kappa_\circ^n, \forall j' \geq j,\ \chi_{j'}^n = 0\}$ it follows that either $A_{\xi-2} < \kappa_\circ^n$ or $\chi_j^n = 1$ for some $j \geq \xi - 1$. Note also that by the upper bound in Lemma 3.9 for $j \geq \xi \geq \underline{j}(\Delta_{\min})$ we have

$$\begin{aligned} A_j - A_{j-1} \geq C^{-1} j^\theta &\geq \frac{8}{\Delta_{\min}^2}\left(\frac{K}{N}+3\right)\log f\left(\frac{C}{1+\theta}(1+j)^\theta\right) \\ &\geq \frac{8}{\Delta_{\min}^2}\left(\frac{K}{N}+3\right)\log f(A_j). \end{aligned}$$

Hence, by Lemmas 3.7 and the lower bound in 3.9 we have

$$\begin{aligned} \mathbb{P}(\hat{\tau}_{\text{stab}}^n \geq \xi) &\leq \mathbb{P}(A_{\xi-2} < \kappa_\circ^n) + \sum_{j\geq\xi-1} \mathbb{E}[\chi_j^n] \\ &\leq \mathbb{P}(\kappa_\circ^n > \frac{C^{-1}}{1+\theta}(\xi-2)^{1+\theta}) + \frac{8K}{\Delta_{\min}^2} \sum_{j\geq\xi-1} \exp\left(-\frac{\Delta_{\min}^2 (A_j - A_{j-1})}{16(K/N+3)}\right) \\ &\leq \mathbb{P}(\kappa_\circ^n > \frac{C^{-1}}{1+\theta}(\xi-2)^{1+\theta}) + \frac{8K}{\Delta_{\min}^2} \sum_{j\geq\xi-1} \exp\left(-\frac{\Delta_{\min}^2 j^\theta}{16C(K/N+3)}\right). \end{aligned}$$

Once again, conclusion of the lemma follows by union bounding over $n \in [N]$. □

Proposition 3.11 *Suppose that there exist $C \geq 1$, $\theta > 0$ such that $C^{-1} j^\theta \leq A_j - A_{j-1} \leq C j^\theta$ for all $j \in \mathbb{N}$. Then there exists a constant $\phi \equiv \phi(\Delta_{\min}, C, \theta, N, K, p_{\min}, \text{diam}(P))$ depending on $\Delta_{\min}, C, \theta, N, K, p_{\min}, \text{diam}(P)$ but not T such that $\mathbb{E}[A_\tau] \leq \phi$.*

Proof Given $A_{\hat{\tau}} \geq \zeta \geq \frac{C}{1+\theta}(1 + 2\underline{j}(\Delta_{\min}))^{1+\theta} \vee \frac{C}{1+\theta} \cdot \{16\text{diam}(P)\}^{1+\theta}$ then $\hat{\tau} \geq (\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}} - 1$, so $\hat{\tau}_{\text{spr}} \vee \hat{\tau}_{\text{stab}} \geq \{(\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}} - 1\}/2 \geq \underline{j}(\Delta_{\min})$. Hence, for $\zeta \geq \psi \equiv \psi(\Delta_{\min}, C, \theta) := \frac{C}{1+\theta}(1 + 2\underline{j}(\Delta_{\min}))^{1+\theta} \vee \frac{C}{1+\theta}\{16\text{diam}(P)\}^{1+\theta}$,

$$\begin{aligned} \mathbb{P}(A_{\hat{\tau}} \geq \zeta) &\leq \mathbb{P}\left(\hat{\tau}_{\text{spr}} \geq \frac{1}{2}\{\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}} - 1\}\right) + \mathbb{P}\left(\hat{\tau}_{\text{stab}} \geq \frac{1}{2}\{\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}} - 1\}\right) \\ &\leq N(1 - p_{\min}^{\text{diam}(P)})^{\left\lfloor \frac{(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}}{4\text{diam}(P)} - 2 \right\rfloor} \\ &\quad + \frac{8KN}{\Delta_{\min}^2} \int_{z \geq (\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}/2-3} \exp\left(-\frac{\Delta_{\min}^2 z^\theta}{16C(K/N+3)}\right) \text{d}z \\ &\quad + \sum_{n\in[N]} \mathbb{P}(\kappa_\circ^n > \{(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}/2 - 4\}^{1+\theta}/(C(1+\theta)) \end{aligned}$$

$$\leq N(1-p_{\min}^{\mathrm{diam}(P)})^{\frac{(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}}{2^4\mathrm{diam}(P)}} + \frac{8KN}{\Delta_{\min}^2}\int_{z\geq(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}/2-3}\exp\left(-\frac{\Delta_{\min}^2 z^\theta}{16C(K/N+3)}\right)\mathrm{d}z + \sum_{n\in[N]}\mathbb{P}(\kappa_\circ^n > (2^{1+\theta}C)^{-2}\cdot\zeta).$$

And by Lemma 3.2, we have that

$$\sum_{n\in N}\sum_{\zeta\in\mathbb{N}}\mathbb{P}(\kappa_\circ^n > (2^{1+\theta}C)^{-2}\cdot\zeta) = \sum_{n\in N}\sum_{\zeta\in\mathbb{N}}\mathbb{P}((2^{1+\theta}C)^2\kappa_\circ^n > \zeta) = (2^{1+\theta}C)^2\sum_{n\in N}\mathbb{E}[k_\circ^n] \leq (2^{1+\theta}C)^2\frac{8N}{\Delta_{\min}^2}.$$

Therefore,

$$\mathbb{E}[A_{\hat{\tau}}] \leq \psi + \sum_{\zeta>\psi}\left\{N(1-p_{\min}^{\mathrm{diam}(P)})^{\frac{(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}}{2^4\mathrm{diam}(P)}}\right\} + \sum_{\zeta>\psi}\left\{\frac{8K}{\Delta_{\min}^2}\int_{z\geq(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}/2-3}\exp\left(-\frac{\Delta_{\min}^2 z^\theta}{16C(K/N+3)}\right)\mathrm{d}z\right\} + (2^{1+\theta}C)^2\frac{8N}{\Delta_{\min}^2} \equiv \phi < \infty.$$

where ϕ is a constant not depending on the time horizon T. □

Theorem 3.1 follows from Corollary 3.6 combined with Proposition 3.11 and by taking $\epsilon \to 0$.

3.1 Finite Sample Bound

To derive a finite sample bound, we will make two additional assumptions. Firstly, we will assume that the neighbours that a recommendation is pulled from are chosen uniformly at random (Definition 3.12). This will allow for a tighter bound on $\hat{\tau}_{\mathrm{spr}}$ which will depend on the conductance and degree of the nodes. We will also assume that the phase lengths grow such that $C^{-1}j^\theta \leq A_j - A_{j-1} \leq Cj^\theta$ where $C \geq 1, \theta > 1$. We now define the required graph properties.

The degree of a node $n \in [N]$ is defined by

$$d_n := \sum_{j\in[N]}\mathbb{1}\{P(i,j)\neq 0\}.$$

The *conductance* ϕ of P is defined as

$$\phi(P) := \min_{S\subset[N],S\neq\emptyset}\frac{\sum_{i\in S,j\in S^c}P(i,j)}{\frac{1}{N}|S|\cdot|S^c|}.$$

Definition 3.12 We say that P satisfies the *uniform-pull* condition if for each $i \in [N]$, with $d_i = \sum_{j\in[N]}\mathbb{1}\{P(i,j)\neq 0\}$ we have $P(i,j) \in \{0, 1/d_i\}$ for all $j \in [N]$.

We will bound $\hat{\tau}_{\text{spr}}$ by using that it is stochastically dominated by a random variable τ_{spr} representing the time it takes for a rumour to spread on a graph according to a pull model where neighbours are chosen uniformly at random. The following result from [7, Lemma 4] gives us a bound on this rumour spreading time τ_{spr}.

Lemma 3.13 (Giakkoupis, 2011) *For all $\beta > 0$, we have that*

$$\mathbb{P}\left(\tau_{\text{spr}} > 50(\beta+2)\log N\left(\phi^{-1} + \frac{d_{\max}}{\lceil \phi \cdot d_{n_\star}\rceil}\right)\right) \leq 3N^{-\beta},$$

where $d_{\max} := \max_{n\in[N]} d_n$ is the maximal degree of the graph and $n_\star$ the agent where $\star \in S_\circ^{n_\star}$.

The following lemma, similar to Lemma 3.11, bounds $\mathbb{E}[A_\tau]$ from above with a time horizon-independent expression. Owing to the additional assumptions we have made, we are able to get a more explicit bound.

Lemma 3.14 *Suppose that P satisfies the uniform-pull condition (3.12). Suppose further that there exist $C \geq 1$, $\theta > 1$ such that $C^{-1} j^\theta \leq A_j - A_{j-1} \leq C j^\theta$ for all $j \in \mathbb{N}$. Suppose that each neighbour is equally likely to be chosen. Then we have that*

$$\mathbb{E}[A_{\hat{\tau}}] \leq \frac{C}{1+\theta}(1+2\underline{j}(\Delta_{\min}))^{1+\theta} + (2^{1+\theta}C)^2\frac{8N}{\Delta_{\min}^2} + \frac{\lceil\theta\rceil!128C^2K64^\theta(K/N+3)^{\theta+1}}{\Delta_{\min}^{4+2\theta}} + 3C\lceil\theta\rceil!\left(400\left(\phi^{-1}+\frac{d_{\max}}{\lceil\phi\cdot d_{n_\star}\rceil}\right)\right)^\theta.$$

Proof Given $A_{\hat{\tau}} \geq \zeta \geq \frac{C}{1+\theta}(1+2\underline{j}(\Delta_{\min}))^{1+\theta}$, then $\hat{\tau} \geq (\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}} - 1$, so $\hat{\tau}_{\text{spr}} \vee \hat{\tau}_{\text{stab}} \geq \{(\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}} - 1\}/2 \geq \underline{j}(\Delta_{\min})$. Hence, for $\zeta \geq \psi \equiv \psi(\Delta_{\min}, C, \theta) := \frac{C}{1+\theta}(1+2\underline{j}(\Delta_{\min}))^{1+\theta}$, and by the same approach as Proposition 3.11, we arrive at

$$\begin{aligned}\mathbb{E}[A_{\hat{\tau}}] \leq{}& \frac{C}{1+\theta}(1+2\underline{j}(\Delta_{\min}))^{1+\theta} \\ &+ (2^{1+\theta}C)^2\frac{8N}{\Delta_{\min}^2} \\ &+ \sum_{\zeta>\psi}\left\{\mathbb{P}\left(\hat{\tau}_{\text{spr}} \geq \frac{1}{2}\left\{\left(\frac{\zeta\,(1+\theta)}{C}\right)^{\frac{1}{1+\theta}} - 1\right\}\right)\right\} \\ &+ \sum_{\zeta>\psi}\left\{\frac{8K}{\Delta_{\min}^2}\int_{z\geq(\zeta(1+\theta)/C)^{\frac{1}{1+\theta}}/2-3}\exp\left(-\frac{\Delta_{\min}^2 z^\theta}{16C(K/N+3)}\right)\mathrm{d}z\right\}.\end{aligned}$$

It suffices to bound the third and fourth terms. Firstly, we will bound the third term using 3.13. For notational convenience, define

$$\Omega := 100\log N\left(\phi^{-1} + \frac{d_{\max}}{\lceil\phi d_{n_\star}\rceil}\right).$$

We have that

$$\frac{1}{3}\cdot\sum_{\zeta>\psi}\mathbb{P}\left(\hat{\tau}_{\text{spr}} \geq \frac{1}{4}\left\{\left(\frac{\zeta\,(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right\}\right) \leq \frac{1}{3}\cdot\sum_{\zeta>\psi}\mathbb{P}\left(\tau_{\text{spr}} \geq \frac{1}{4}\left\{\left(\frac{\zeta\,(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right\}\right)$$

$$
\begin{aligned}
&\leq \sum_{\zeta > \psi} N^{-\frac{1}{4\Omega}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}} \\
&\leq \int_{\psi} N^{-\frac{1}{4\Omega}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}} \mathrm{d}\zeta \\
&\leq \int_{\psi} \exp\left(-\frac{\log N}{4\Omega}\left(\frac{\zeta\,(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right) \mathrm{d}\zeta \\
&\leq C\left(\frac{4\Omega}{\log N}\right)^{\theta} \int_0^{\infty} x^{\theta} \exp(-x) \mathrm{d}x \\
&\leq C\left(\frac{4\Omega}{\log N}\right)^{\theta} \Gamma(\theta) \\
&\leq C\Gamma(\theta)\left(400\left(\phi^{-1} + \frac{d_{\max}}{\lceil \phi \cdot d_{n_\star} \rceil}\right)\right)^{\theta} \\
&\leq C\lceil\theta\rceil!\left(400\left(\phi^{-1} + \frac{d_{\max}}{\lceil \phi \cdot d_{n_\star} \rceil}\right)\right)^{\theta},
\end{aligned}
$$

where the first inequality holds since $\hat{\tau}_{\mathrm{spr}}$ is stochastically dominated by τ_{spr}, the second inequality holds from Lemma 3.13 and the fifth inequality holds from a change of variables.

We will now bound the fourth term. We start by bounding the integral

$$
\begin{aligned}
\int_{\frac{1}{2}(\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}}-3}^{\infty} \exp\left(-\frac{\Delta_{\min}^2 z^{\theta}}{16C(K/N+3)}\right) \mathrm{d}z &\leq \int_{\frac{1}{2}(\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}}-3}^{\infty} \exp\left(-\frac{\Delta_{\min}^2 z}{16C(K/N+3)}\right) \mathrm{d}z \\
&\leq \frac{16C(K/N+3)}{\Delta_{\min}^2} \exp\left(-\frac{1}{2}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}} - 3\right) \\
&\leq \frac{16C(K/N+3)}{\Delta_{\min}^2} \exp\left(-\frac{1}{4}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right).
\end{aligned}
$$

And therefore, we have

$$
\begin{aligned}
&\sum_{\zeta > \psi}\left\{\frac{8K}{\Delta_{\min}^2} \int_{\frac{1}{2}(\frac{\zeta(1+\theta)}{C})^{\frac{1}{1+\theta}}-3}^{\infty} \exp\left(-\frac{\Delta_{\min}^2 z^{\theta}}{16C(K/N+3)}\right) \mathrm{d}z\right\} \\
&\quad\leq \frac{128CK(K/N+3)}{\Delta_{\min}^4} \sum_{\zeta > \psi} \exp\left(-\frac{1}{4}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right) \\
&\quad\leq \frac{128CK(K/N+3)}{\Delta_{\min}^4} \int_{\psi}^{\infty} \exp\left(-\frac{1}{4}\left(\frac{\zeta(1+\theta)}{C}\right)^{\frac{1}{1+\theta}}\right) \mathrm{d}\zeta \\
&\quad\leq \frac{128C^2K64^{\theta}(K/N+3)^{\theta+1}}{\Delta_{\min}^{4+2\theta}} \int_0^{\infty} x^{\theta} \exp(-x) \mathrm{d}x \\
&\quad\leq \frac{\lceil\theta\rceil!128C^2K64^{\theta}(K/N+3)^{\theta+1}}{\Delta_{\min}^{4+2\theta}}.
\end{aligned}
$$

□

The following result is a direct consequence of Corollary 3.6 and Lemma 3.14.

Theorem 3.15 *Suppose that there exist $C \geq 1$, $\theta > 1$ such that $C^{-1} j^{\theta} \leq A_j - A_{j-1} \leq C j^{\theta}$ for all $j \in \mathbb{N}$. Then for each agent $n \in [N]$, we have the following regret bound*

$$\mathbb{E}[\mathcal{R}_T^n] \leq \frac{C}{1+\theta}(1 + 2\underline{j}(\Delta_{\min}))^{1+\theta} + (2^{1+\theta}C)^2 \frac{8N}{\Delta_{\min}^2} + \frac{\lceil\theta\rceil! 128 C^2 K 64^{\theta} (K/N+3)^{\theta+1}}{\Delta_{\min}^{4+2\theta}}$$
$$+ \underbrace{3C\lceil\theta\rceil! \left(400\left(\phi^{-1} + \frac{d_{\max}}{\lceil\phi \cdot d_{n_\star})\rceil}\right)\right)^{\theta}}_{\textit{Impact from graph}}$$
$$+ \sum_{k \in S_\circ^n \setminus [\star]} \Delta_k \inf_{\epsilon \in \left(0, \frac{\Delta_{\min}}{2}\right)} \left\{ \frac{\log f(T)}{\mathrm{KL}(\mu_k + \epsilon, \mu_\star - \epsilon)} + \frac{3}{\epsilon^2} \right\}.$$

This bound provides an insight into effect the initial phases (before τ) might have on the regret. We observe this bound is large when either $\Delta_{\min}$ or the conductance ϕ are small or when either θ or the ratio $\frac{d_{\max}}{d_{n_\star}}$ is large. Additionally, θ amplifies the affect of these parameters on the regret bound.

Figure 3 shows the results of using different networks with Algorithm 1 from a series of simulations. These simulations score the regret for different graphs, in order from lowest to highest, as *complete*, *cycle* and *star*. The *Impact from graph* term from Corollary 3.14 can help explain these results.

Firstly, the complete graph has conductance $\phi = \frac{N}{2(N-1)}$ so the impact of the complete graph on the regret bound is

$$3C\lceil\theta\rceil! \left(400\,(4)\right)^{\theta}.$$

For the cycle graph, the conductance is $\phi = 2/N$ so the graph impact is

$$3C\lceil\theta\rceil! \left(400\,(N/4)\right)^{\theta}.$$

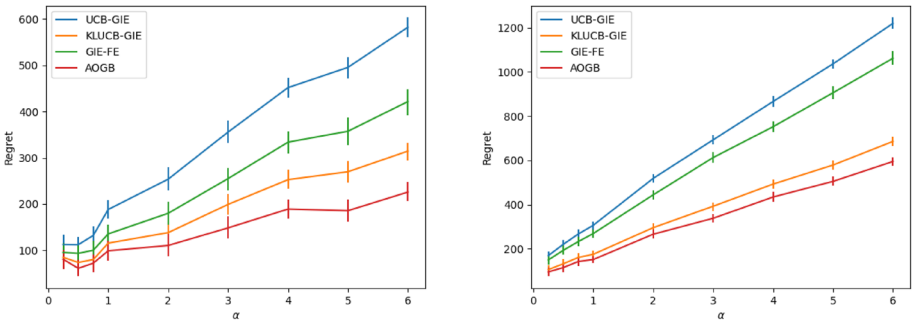
Finally, the impact of star graph depends on whether the best arm begins on the central node or a leaf node. Since the conductance of the star graph is $\phi = \frac{N-1}{N}$, we have the following scaling in each case

$$\underbrace{3C\lceil\theta\rceil! \left(400\,(2)\right)^{\theta}}_{\text{Central Node}} \qquad \underbrace{3C\lceil\theta\rceil! \left(400(N-1)\right)^{\theta}}_{\text{Leaf Node}}.$$

4 Numerical Results

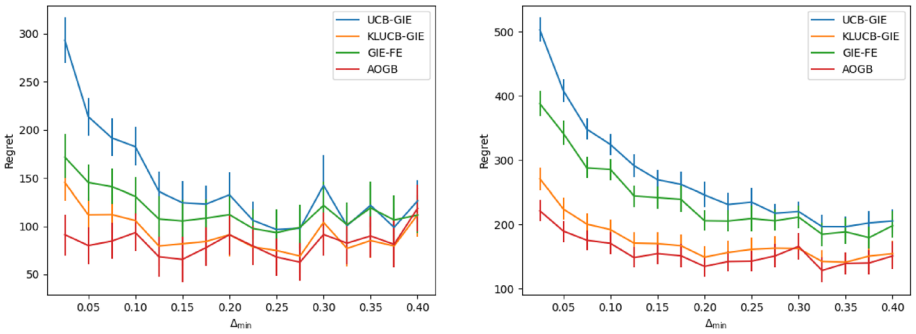
Here we will compare Algorithm 1 and the GosInE algorithm on a range of synthetic data. We compare variants of both of these algorithms using Hoeffding and KL upper confidence bounds. For GosInE, the Hoeffding and KL variants are, respectively, labelled UCB-GIE and KLUCB-GIE, and for Algorithm 1, they are labelled GIE-FE (Gossip-Insert-Eliminate with Fast Elimination) and AOGB.

The experiments are conducted in two settings: $N, K = (20, 50)$ and $N, K = (10, 100)$. Each experiment consists of 100 independent runs, and in each run, the regret is averaged over the nodes. In each experiment, the algorithms encounter the same reward sequence. The first two experiments assume the agents are connected via a complete graph, while the third experiment compares different graphs. We compute the regret over a time horizon of



$N = 20, K = 50$ $N = 10, K = 100$

Fig. 1 Regret for different choices of α with $\mu_\star = 0.9$ and the rest of the arms divide the interval $[0.2, 0.8]$ uniformly



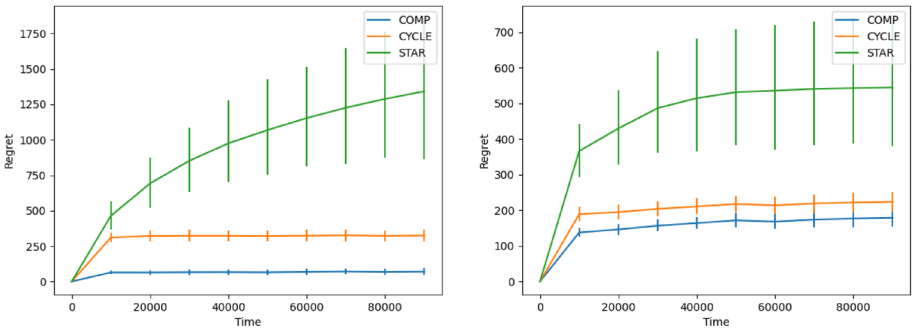
$N = 20, K = 50$ $N = 10, K = 100$

Fig. 2 Regret for different choices of $\Delta_{\min}$ with $\alpha = 1$. The best arm has mean $\mu_\star = 0.9$ and the rest of the arms divide the interval $[0.9 - \Delta_{\min}, 0.2]$ uniformly

$T = 100,000$ and plot the sample mean along with 95% confidence intervals. Other than in Fig. 4, we take the phase lengths to grow cubically, i.e. $A_j = j^3$, and other than in Fig. 3, we assume that agents are connected via a complete graph.

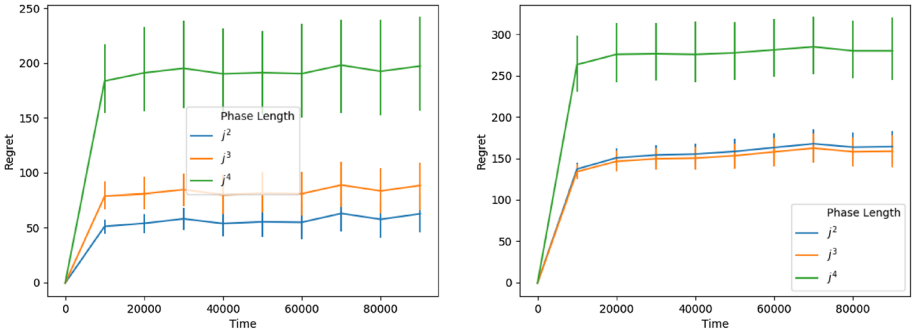
Choice of α: We begin by comparing Algorithm 1 and GosInE for the two different types of upper confidence bounds by varying the exploration function $f(t) = 1 + t^\alpha \log^2(t)$ by choosing different values for α. From Fig. 1, we identify that Algorithm 1 and GosInE perform better when equipped with KL upper confidence bound. Additionally, Algorithm 1 outperforms GosInE when they are both equipped with the same upper confidence bounds. Overall, performance is better for the smaller values of α and regret is minimised somewhere in the region $\alpha \leq 1$. This implies that there may be more practical choices for $f_\alpha(t)$ than the asymptotically optimal choice at $\alpha = 1$.

$\Delta_{\min}$ **vs Regret:** Now we consider the effect of changing the suboptimality gap $\Delta_{\min}$. This is the difference between the means of the best and the second-best arms. Figure 2 compares Algorithm 1 and the GosInE algorithm for both types of confidence intervals. Similarly to the previous experiment, we observe that both algorithms perform better when equipped with the KL upper confidence bounds and that Algorithm 1 typically outperforms GosInE on average when they are equipped with the same upper confidence bounds.



$N = 20, K = 50$ $N = 10, K = 100$

Fig. 3 Regret over time for three different networks. Each in case we consider $\alpha = 1$, $\Delta_{\min} = 0.1$ and the means of the remaining arms divide the interval [0.8, 0.2] uniformly



$N = 20, K = 50$ $N = 10, K = 100$

Fig. 4 Regret over time for different choices of A_j. Each in case we consider $\alpha = 1$, $\Delta_{\min} = 0.1$ and the means of the remaining arms divide the interval [0.8, 0.2] uniformly

Network Configurations Here we compare three different network configurations for agents implementing Algorithm 1: a complete graph, a cycle graph and a star graph.

The results in Fig. 3 show that the cycle graph performs slightly worse than the complete graph but the star graph struggles significantly along with a larger variance. In essence, this is because the best arm needs to spread to centre of the star before it can spread to all of the other nodes.

Phase Lengths

In Fig. 4, we see the effect of changing the communication rounds A_j on the regret. We consider three polynomial different functions, j^2, j^3, j^4, as these satisfy the assumptions in our theoretical analysis. We observe that increasing the phase lengths (and thus decreasing the number of communication rounds) incurs more regret in the initial time steps in both cases, which is as expected.

5 Discussion

In this paper, we presented an algorithm (Algorithm 1) for multi-agent bandits in a decentralised setting. Our algorithm builds upon the Gossip-Insert-Eliminate algorithm of [3] by making two modifications. First, we use tighter confidence intervals inspired by [6]. Second, we use a faster elimination scheme for reducing the number of arms that must be explored by an agent. Both modifications yield significant empirical improvement on simulated data (Fig. 2). Finally, we prove a regret bound (Theorem 3.1) which demonstrates asymptotically optimal performance of our algorithm, matching the asymptotic performance of a collection of agents with unlimited communication.

There is substantial scope for future work in this direction. One challenge of great practical importance is the development of distributed algorithms which are robust to both malicious agents and faulty communication [13]. An interesting theoretical challenge is to develop a multi-agent bandit algorithm which is both asymptotically optimal and nearly minimax optimal with limited communication. In very recent work of [1], an algorithm has been proposed which is minimax optimal in the distributed setting, and it would be interesting to synthesise this with the insights provided in the current paper.

References

1. Agarwal M, Aggarwal V, Azizzadenesheli K (2021) Multi-agent multi-armed bandits with limited communication. arXiv preprint arXiv:2102.08462
2. Cesa-Bianchi N, Cesari T, Monteleoni C (2020) Cooperative online learning: keeping your neighbors updated. In: Algorithmic learning theory, pp 234–250
3. Chawla R, Sankararaman A, Ganesh A, Shakkottai S (2020) The gossiping insert-eliminate algorithm for multi-agent bandits. In: International conference on artificial intelligence and statistics, pp 3471–3481
4. Dong M, Meng T, Zarchy D, Arslan E, Gilad Y, Godfrey B, Schapira M (2018) PCC vivace: online-learning congestion control. In: 15th USENIX symposium on networked systems design and implementation (NSDI 18), pp 343–356, Renton, WA, USENIX Association
5. Dubey A (2020) Cooperative multi-agent bandits with heavy tails. In: International conference on machine learning, pp 2730–2739. PMLR
6. Garivier A, Cappe O (2011) The kl-ucb algorithm for bounded stochastic bandits and beyond. In: Proceedings of the 24th annual conference on learning theory, JMLR workshop and conference proceedings, pp 359–376
7. Giakkoupis G (2011) Tight bounds for rumor spreading in graphs of a given conductance. In: Symposium on theoretical aspects of computer science (STACS2011), vol 9, pp 57–68
8. Jiang J, Sun S, Sekar V, Zhang H (2017) Pytheas: enabling data-driven quality of experience optimization using group-based exploration-exploitation. In: 14th USENIX symposium on networked systems design and implementation (NSDI 17), pp 393–406, Boston, MA, USENIX Association
9. Kanade V, Liu Z, Radunovic B (2012) Distributed non-stochastic experts. Neurips
10. Kolla RK, Jagannathan K, Gopalan A (2018) Collaborative learning of stochastic bandits over a social network. IEEE/ACM Trans Netw 26(4):1782–1795
11. Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. Adv Appl Math 6(1):4–22
12. Lattimore T, Szepesvári C (2020) Bandit algorithms. Cambridge University Press, Cambridge

13. Lynch NA (1996) Distributed algorithms. Elsevier, Amsterdam
14. Maillard OA, Munos R, Stoltz G (2011) A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In: Proceedings of the 24th annual conference on learning theory, pp 497–514. JMLR Workshop and Conference Proceedings
15. Martinez-Rubio D, Kanade V, Rebeschini P (2018) Decentralized cooperative stochastic bandits. Neurips
16. Pandey S, Agarwal D, Chakrabarti D, Josifovski V (2007) Bandits for taxonomies: a model-based approach. In: Proceedings of the 2007 SIAM international conference on data mining, pp 216–227. SIAM
17. Sankararaman A, Ganesh A, Shakkottai S (2019) Social learning in multi agent multi armed bandits. Proc ACM Meas Anal Comput Syst 3(3):1–35
18. Sugawara K, Kazama T, Watanabe T (2004) Foraging behavior of interacting robots with virtual pheromone. In: 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS) (IEEE Cat. No.04CH37566), vol 3, pp 3074–3079
19. Szorenyi B, Busa-Fekete R, Hegedus I, Ormandi R, Jelasity M, Kegl B (2013) Gossip-based distributed stochastic bandit algorithms. In: International conference on machine learning, pp 19–27. PMLR