On Regret-optimal Cooperative Nonstochastic Multi-armed Bandits

Jialin Yi
London School of Economics and Political Science
London, United Kingdom
j.yi8@lse.ac.uk

Milan Vojnović
London School of Economics and Political Science
London, United Kingdom
m.vojnovic@lse.ac.uk

ABSTRACT

We consider the nonstochastic multi-agent multi-armed bandit problem with agents collaborating via a communication network with delays. We show a lower bound for individual regret of all agents. We show that with suitable regularizers and communication protocols, a collaborative multi-agent *follow-the-regularized-leader* (FTRL) algorithm has an individual regret upper bound that matches the lower bound up to a constant factor when the number of arms is large enough relative to degrees of agents in the communication graph. We also show that an FTRL algorithm with a suitable regularizer is regret optimal with respect to the scaling with the edge-delay parameter. We present numerical experiments validating our theoretical results and demonstrate cases when our algorithms outperform previously proposed algorithms.

KEYWORDS

Multi-agent system; Bandit problem; Regret minimization

ACM Reference Format:
Jialin Yi and Milan Vojnović. 2023. On Regret-optimal Cooperative Nonstochastic Multi-armed Bandits. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 7 pages.

1 INTRODUCTION

Coordinating multiple agents that can communicate with each other to make decisions under uncertainty is a classical problem and has many different applications in computer science [14], game theory [7] and machine learning [12]. We consider the multi-agent version of a multi-armed bandit problem which is one of the most fundamental decision making problems under uncertainty. In this problem, a learning agent needs to consider the exploration-exploitation trade-off, i.e. balancing the exploration of various actions in order to learn how much rewarding they are and selecting high-rewarding actions. In the multi-agent version of this problem, multiple agents collaborate with each other trying to maximize their individual cumulative rewards, and the challenge is to design efficient cooperative algorithms under communication constraints.

We consider the nonstochastic (adversarial) multi-armed bandit problem in a cooperative multi-agent setting, with $K \geq 2$ arms and $N \geq 1$ agents. In each time step, each agent selects an arm and then observes the incurred loss corresponding to its selected arm. The losses of arms are according to an arbitrary loss sequence, which is commonly referred to as the nonstochastic or adversarial setting. Each agent observes only the loss of the arm this agent selected in each time step. The agents are allowed to cooperate by exchanging messages, which is constrained by a communication graph G such that any two agents can exchange a message directly between themselves only if they are neighbors in graph G. Each exchange of a message over an edge has delay of d time steps. The goal of each agent is to minimize its cumulative loss over a time horizon of T time steps. We study the objective of minimizing the individual regret of agents, i.e. the difference between the expected cumulative loss incurred by an agent and the cumulative loss of the best arm in hindsight. We also study the average regret of all agents.

The multi-agent multi-armed bandit problem formulation that we study captures many systems that use a network of learning agents. For example, in peer-to-peer recommender systems, the agents are users and the arms are products that can be recommended to users [3]. The delay corresponds to the time it takes for a message to be transmitted between users. Note that in this application scenario, the number of products (i.e. arms) may be much larger than the number of users (i.e. agents).

The collaborative multi-agent multi-armed bandit problem was studied, e.g., in Cesa-Bianchi et al. [6] and Bar-On and Mansour [2], where each edge has unit delay. Our setting is more general in allowing for arbitrary delay d per edge. Cesa-Bianchi et al. [6] showed that when each agent selects arms according to a cooperative Exp3 algorithm (Exp3-Coop), the average regret is $O(\sqrt{(\alpha(G)/N + 1/K)\log(K)KT})$ for large enough T, where $\alpha(G)$ is the independence number of graph G. Bar-On and Mansour [2] have shown that individual regret of each agent v is $O(\sqrt{(1/|\mathcal{N}(v)| + 1/K)\log(K)KT})$ when $T \geq K^2\log(K)$, where $\mathcal{N}(v)$ is the set of neighbors of agent v and itself in graph G. This regret bound is shown to hold for an algorithm where some agents, referred to as center agents, select arms using the Exp3-Coop policy and other agents copy the actions of center agents. These bounds reveal the effect of collaboration on the learning and what graph properties effect the efficiency of learning. However, some fundamental questions still remain. For example, to the best of our knowledge, it is unknown from the previous literature what is the lower bound for this problem. Moreover, it is unknown whether better algorithms can be designed whose regret matches a lower bound under certain conditions.

In this work, we give a regret lower bound for any learning algorithm in which each agent can only communicate with their neighbors. We present a center-based algorithm whose regret upper bound matches the lower bound when the number of arms is large enough. We present an algorithm that has a regret upper bound with $\sqrt{d}$ dependence on the delay per edge, which is optimal. All our regret bounds are parametrized with the delay parameter d,

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

which is unlike to Cesa-Bianchi et al. [6] and Bar-On and Mansour [2] which considered only the special case when $d = 1$. In what follows we summarise our results in more details.

1.1 Summary of our contributions

We show that any algorithm has individual regret for each agent v lower bounded as

$$\Omega\left(\sqrt{\left(\frac{1}{|\mathcal{N}(v)|}+\frac{1}{K}\log(K)\right)KT}\right),$$

when $T \geq K/|\mathcal{N}(v)|$. This implies the average regret lower bound $\Omega(\sqrt{(\delta(G)^2+\log(K)/K)KT})$, when $T \geq \delta(G)^2K$, where $\delta(G) = (1/N)\sum_{v\in\mathcal{V}}(1/\sqrt{|\mathcal{N}(v)|})$. Hence, there is a $\sqrt{\log(K)}$ factor gap between the previously known upper bounds and the lower bound.

We show an algorithm that guarantees individual regret for each agent v to be

$$O\left(\sqrt{\left(\frac{1}{|\mathcal{N}(v)|}+\frac{|\mathcal{N}(C(v))|}{K}\log(K)^2\right)KT}\right),$$

whenever $K \geq \max_v |\mathcal{N}(v)|$ and $T \geq \Omega(K\max_{c\in C}|\mathcal{N}(c)|)$, where C is the set of center agents and $C(v)$ is the nearest center agent to agent v. This regret bound improves the best known upper bound on individual regret when the number of arms is large enough relative to agents' degrees. Moreover, we note that the algorithm has optimal regret up to a constant factor when when $|\mathcal{N}(C(v))||\mathcal{N}(v)|\log(K)^2/K = O(1)$, i.e. when the number of arms is large enough.

Our algorithm is based on using a cooperative Follow-the-Regularized-Leader (FTRL) policy with a Tsallis entropy regularizer. This is in contrast to Bar-On and Mansour [2], Cesa-Bianchi et al. [6], which both use a cooperative Exp3 policy.

Our regret analysis relies on a key new lemma that bounds the change of the action selection strategy of an agent under Tsallis entropy regularization. This result may be of independent interest.

We also present a decentralized follow-the-regularized-leader algorithm that has regret with optimal dependency on the delay parameter d, namely scaling as $\sqrt{d}$. This algorithm uses a hybrid regularizer, which combines an Exp3 type regularizer with a Tsallis entropy regularizer. This algorithm is decentralised with all agents applying the same strategy.

1.2 Related work

The multi-armed bandit problem in a multi-agent setting, where agents collaborate with each other subject to some communication constraints, has received considerable attention in recent years. Awerbuch and Kleinberg [1] introduced the cooperative nonstochastic multi-armed bandit problem setting where communication is through a public channel (corresponding to a complete graph) and some agents may be dishonest. Kar et al. [10] considered a special collaboration network in which only one agent can observe the loss of the selected arm in each time step. Szörényi et al. [17] discussed two specific P2P networks in which at each time step, each agent can send messages to only two other agents. Cesa-Bianchi et al. [5] studied an online learning problem where only a subset of agents play in each time step. They showed that an optimal average regret bound for this problem is $\Theta(\sqrt{\alpha(G)T})$ when the set of agents that play in each time step is chosen randomly, while $\Omega(T)$ bound holds when the set of agents can be chosen arbitrarily in each time step. Kolla et al. [11], Landgren et al. [13, 13] and Martínez-Rubio et al. [15] considered a setting in which communication is constrained by a communication graph such that any two agents can communicate *instantly* if there is an edge connecting them.

The communication model considered in our paper was introduced by Cesa-Bianchi et al. [6]. Here, agents communicate via messages sent over edges of a fixed connected graph and sending a message over an edge incurs a delay of value d. Cesa-Bianchi et al. [6] considered the case when $d = 1$ whereas in this paper, we consider $d \geq 1$. They proposed an algorithm, referred to as Exp3-Coop, in which each agent constructs loss estimators for each arm using an importance-weighted estimator. The Exp3-Coop algorithm has an upper bound of $O(\sqrt{(\alpha(G)/N+1/K)\log(K)KT}+\log(T))$ on the average regret. Bar-On and Mansour [2] combines the idea of center-based communication from Kolla et al. [11] with the Exp3-Coop algorithm, showing that the center-based Exp3 algorithm has a regret upper bound of $O(\sqrt{(1/|\mathcal{N}(v)|+1/K)\log(K)KT})$ for each individual agent when $d = 1$. We show that a better regret bound can be guaranteed with respect to the scaling with the number of arms K.

Multi-armed bandits with delayed feedback have been studied extensively in the single-agent setting [8, 9, 18]. Specifically, Zimmert and Seldin [21] considered a setting in which the agent has no prior knowledge about the delays and showed an optimal regret of $O(\sqrt{KT}+\sqrt{d\log(K)T})$ where d is the average delay over T time steps. We present, in the multi-agent setting, a distributed learning algorithm whose regret upper bound can also achieve this optimal $\sqrt{d}$ dependence on the delay per edge d.

We use the *Tsallis entropy* family of regularizers proposed by Tsallis [19]. Zimmert and Seldin [22] have shown that an online mirror descent algorithm with a Tsallis entropy regularizer achieves optimal regret for the single-agent bandit problem. We show a distributed learning algorithm for the multi-agent bandit setting, which uses a Tsallis entropy regularizer.

1.3 Organization of the paper

Section 2 provides problem formulation and definitions of notation. Section 3 presents our two algorithms and their regret bounds. In Section 4, we present a lower bound on individual regret of each agent. Section 5 contains numerical results. Finally, conclusion remarks are given in Section 6. Proofs of our results are available in the supplementary material [20].

2 PROBLEM FORMULATION

We consider a multi-armed bandit problem with a finite set $\mathcal{A} = \{1,\dots,K\}$ of actions (arms) played by N agents. The agents can communicate through a communication network $G = (\mathcal{V},\mathcal{E})$ where $\mathcal{V}$ is the set of N agents and $\mathcal{E}$ is the set of edges such that $(u,v) \in \mathcal{E}$ if, and only if, agent u can send/receive messages to/from agent v. We denote the neighbors of the agent v and itself by the set $\mathcal{N}(v) = \{u \in \mathcal{V} : (u,v) \in \mathcal{E}\} \cup \{v\}$. Sending a message

over edge $e \in \mathcal{E}$ incurs a delay of value $d_e \geq 0$ time steps. We consider the *homogeneous* setting under which $d_e = d \geq 1$ for every edge $e \in E$. Note that the delayed communication network model in Cesa-Bianchi et al. [6] and Bar-On and Mansour [2] is restricted to the special case $d = 1$.

At each time step $t = 1, 2, \ldots, T$, each agent $v \in \mathcal{V}$ chooses an action $I_t(v) \in \mathcal{A}$ according to distribution p_t^v over $\mathcal{A}$ and then observes the loss value, $\ell_t(I_t(v)) \in [0, 1]$. Notice that the loss does not depend on the agent, but only on the time step and the chosen action. Hence, if two agents choose the same action at the same time step, they incur the same loss. We consider the *nonstochastic setting* where the losses are determined by an oblivious adversary, meaning that the losses do not depend on the agent's realized actions.

At the end of each time step t, each agent $v \in \mathcal{V}$ sends a message $S_t(v)$ of size $b_t(v)$ information bits to all its neighbors and after this, each agent $v \in \mathcal{V}$ has messages $\cup_{u \in \mathcal{N}(v)} \{S_s(u) : s+d = t\}$. We assume that at each time step t, each agent v can send to each of its neighbors a message $S_t(v) = \left\langle v, t, I_t(v), \ell_t\left(I_t(v)\right), p_t^v \right\rangle$, i.e. the agent id, the time step, the chosen arm id, the instant loss received and the instant action distribution. We denote with $b_t(v)$ the number of information bits to encode $S_t(v)$. The total communication cost in each time step is $\sum_{v \in \mathcal{V}} \sum_{u:(u,v) \in \mathcal{E}} b_t(u)$ information bits.

The *individual regret* of each agent v is defined as the difference between its expected accumulated loss and the loss of the best action in hindsight, i.e.

$$R_T^v = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t\left(I_t(v)\right)\right] - \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(i).$$

The *average regret* of N agents is defined as

$$R_T = \frac{1}{N} \sum_{v \in \mathcal{V}} R_T^v.$$

Additional notation. We define $\mathcal{P}_{K-1}$ to be the $K-1$ simplex. Let $\alpha(G)$ be the size of a maximal independent set of graph G, where the maximal independent set is the largest subset of nodes such that no two nodes in this set are connected by an edge.

3 ALGORITHMS AND REGRET UPPER BOUNDS

In this section, we propose two collaborative multi-agent bandit algorithms, the center-based cooperative follow-the-regularized-leader (CFTRL) algorithm and the decentralized cooperative follow-the-regularized-leader (DFTRL) algorithm. The first algorithm has optimal regret up to a constant factor when the number of arms is large enough. The second algorithm has optimal dependence on the delay parameter d.

3.1 A center-based cooperative follow-the-regularized-leader algorithm

We consider an algorithm where some agents, referred to as *centers*, run a FTRL algorithm, and each other agent copies the action selection distribution from its nearest center. The strategy based on using center agents was proposed in Bar-On and Mansour [2], where agents played the Exp3 strategy instead. These centre agents collaboratively update their strategies by exchanging messages with other agents, and each non-center agent copies the strategy of its nearest center agent. The center agents are selected such that they have a sufficiently large degree, which can be shown to reduce individual regret of center agents. Moreover, the center agents are selected such that each non-center agent is within a small distance to a center agent.

Let $C \subseteq \mathcal{V}$ be the set of centers. The set of agents $\mathcal{V}$ is partitioned into disjoint components $\mathcal{V}_c$, $c \in C$. Each non-center agent v belongs to a unique component. For each agent v, let $C(v)$ denote its center agent, $c = C(v)$ if and only if $v \in \mathcal{V}_c$. Let $d(v)$ be the distance between a non-center agent v and its center $C(v)$. The set of centers C and the partitioning $\{\mathcal{V}_c : c \in C\}$ are computed according to Algorithms 3 and 4 in Bar-On and Mansour [2].

Let $\mathcal{J}_t(v) = \{I_t(v') : v' \in \mathcal{N}(v)\}$ be the set of actions chosen by agent v or its neighbors at time step t. Each center agent $c \in C$ runs a FTRL algorithm with the collaborative importance-weighted loss estimators observable up to time step t,

$$\hat{L}_t^{c,obs}(i) = \sum_{s=1}^{t-1} \hat{\ell}_s^{c,obs}(i)$$

and

$$\hat{\ell}_t^{c,obs}(i) = \begin{cases} \frac{\ell_{t-d}(i)}{q_{t-d}^c(i)} \mathbb{I}\{i \in \mathcal{J}_{t-d}(c)\} & \text{if } t > d \\ 0 & \text{otherwise} \end{cases}$$

where

$$q_t^c(i) = 1 - \prod_{v \in \mathcal{N}(c)} \left(1 - p_t^v(i)\right)$$

is the *neighborhood-aggregated importance weight.*

In each time step, the center agents update their action selection distributions according to the FTRL algorithm, i.e.

$$p_t^c = \text{argmin}_{p \in \mathcal{P}_{K-1}} \left\{ \left\langle p, \hat{L}_t^{c,obs} \right\rangle + F_t(p) \right\}.$$

where $F_t(p)$ is the *Tsallis entropy regularizer* [22] with the learning rate $\eta(c)$,

$$F_t(p) = -2 \sum_{i=1}^{K} \sqrt{p_i} / \eta(c). \tag{1}$$

Each non-center agent $v \in \mathcal{V} \backslash C$ selects actions according to the uniform distribution until time step $t > d(v)d$. Then the non-center agent copies the action selection distribution from its center, i.e. $p_t^v = p_{t-d(v)d}^{C(v)}$. The details of the CFTRL algorithm are described in Algorithm 1.

3.1.1 *Individual regret upper bound.*

We show an individual regret upper bound for Algorithm 1 in the following theorem.

Theorem 3.1. *Assume that $K \geq \max_{v \in \mathcal{V}} |\mathcal{N}(v)|$ and $T \geq 36(d+1)^2 K \max_c |\mathcal{N}(c)|$, and agents follow the CFTRL algorithm with each center agent $c \in C$ using the learning rate $\eta(c) = \sqrt{|\mathcal{N}(c)|/(3T)}$. Then, the individual regret of each agent $v \in \mathcal{V}$ is bounded as*

$$R_T^v = O\left(\frac{1}{\sqrt{|\mathcal{N}(v)|}} \sqrt{KT} + d \log(K) \sqrt{|\mathcal{N}(C(v))|T}\right).$$

The proof of the theorem is provided in the supplementary material [20]. In the following, we provide a proof sketch. The proof relies on two key lemmas which are shown next.

Algorithm 1: Center-based cooperative FTRL (CFTRL)

Input : Tsallis regularizer Eq. (1), learning rate $\eta(c)$ and the delay d.

Initialization : $\hat{L}_1^{c,obs}(i) = 0$ for all $i \in \mathcal{A}$ and $c \in C$, $p_1^v(i) = 1/K$ for all $i \in \mathcal{A}$ and $v \notin C$.

for *each time step* $t = 1, 2, \ldots, T$ **do**
- Each $c \in C$ updates $p_t^c = \mathrm{argmin}_{p \in \mathcal{P}_{K-1}} \{\langle p, \hat{L}_t^{c,obs} \rangle + F_t(p)\}$;
- Each $c \in C$ chooses $I_t(c) = i$ with probability $p_t^c(i)$ and receives the loss $\ell_t(I_t(c))$;
- Each $c \in C$ sends the message $S_t(c) = \left\langle c, t, I_t(c), \ell_t(I_t(c)), p_t^c \right\rangle$ to all their neighbors;
- Each $c \in C$ receives messages $\{S_{t-d}(v) : v \in \mathcal{N}(c)\}$ and computes $\hat{L}_{t+1}^{c,obs}$;
- Each $v \in \mathcal{V} \setminus C$ updates $p_t^v = p_{t-d(v)d}^{C(v)}$ when $t > d(v)d$ and $p_t^v = p_1^v$ otherwise;
- Each $v \in V \setminus C$ chooses $I_t(v) = i$ with probability $p_t^v(i)$ and receives $\ell_t(I_t(v))$;
- Each $v \in \mathcal{V} \setminus C$ sends $S_t(v) = \left\langle v, t, I_t(v), \ell_t(I_t(v)), p_t^v \right\rangle$ to all its neighbors.

end

Lemma 3.2. *Assume that the delay of each edge is $d \geq 1$, then the individual regret of each center agent v with the regularizer $F_t(p) = \sum_{i=1}^K f_t(p_i)$ satisfies*

$$R_T^v \leq M + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{i \in \mathcal{A}} \frac{1}{q_t^v(i) f_t''(p_t^v(i))}\right] + d \cdot \mathbb{E}\left[\sum_{t=1}^T \sum_{i \in \mathcal{A}} \frac{1}{f_t''(p_t^v(i))}\right]$$

where

$$M = \max_{x \in \mathcal{P}_{K-1}} -F_1(x) + \sum_{t=2}^T \max_{x \in \mathcal{P}_{K-1}} \left(F_{t-1}(x) - F_t(x)\right).$$

Lemma 3.3. *For any $\delta > 1$, assume that agent v runs a FTRL algorithm with Tsallis entropy (1) and learning rate $\eta(v) \leq (1 - 1/\sqrt{\delta})/(\delta^{3d/2}\sqrt{K})$ and $K \geq 2$, then for all $t \geq 1$ and $i \in \mathcal{A}$*

$$(1 - (1+\delta)\eta(v)\hat{\ell}_t^{v,obs}(i))p_t^v(i) \leq p_{t+1}^v(i) \leq \delta p_t^v(i).$$

The proofs of the two lemmas are provided in the supplementary material [20]. Similar property as in Lemma 3.3 was known to hold for the Exp3 algorithm by a result in Cesa-Bianchi et al. [6]. To the best of our knowledge, this property was previously not known to hold for the FTRL algorithm with Tsallis entropy.

Lemma 3.2 bounds the individual regret by the sum of a constant, the regret due to the *instant* bandit feedback, and the regret due to the *delayed* full-information feedback, which is $O(\sqrt{KT/|\mathcal{N}(c)|})$. Since the action selection distributions of non-center agents are copied from their centers in the past rounds, Lemma 3.3 bounds the difference between action selection distributions of non-center agent v and its center $C(v)$ in the same rounds when $T > d(v)d$. Consequently, the difference between the individual regret of a non-center agent v and its center $C(v)$ is bounded by $O(d(v)d\eta(C(v))T) = O(d\log(K)\sqrt{|\mathcal{N}(C(v))|T})$.

Algorithm 2: Decentralized cooperative FTRL (DFTRL)

Input : Hybrid regularizer Eq. (2), learning rates η_t, ζ_t, and delay d.

Initialization : $\hat{L}_1^{v,obs}(i) = 0$ for all $i \in \mathcal{A}$ and $v \in \mathcal{V}$.

for *each time step* $t = 1, 2, \ldots, T$ **do**
- Each $v \in \mathcal{V}$ updates $p_t^v = \mathrm{argmin}_{p \in \mathcal{P}_{K-1}} \{\langle p, \hat{L}_t^{v,obs} \rangle + F_t(p)\}$;
- Each $v \in \mathcal{V}$ chooses $I_t(v) = i$ with probability $p_t^v(i)$ and receives the loss $\ell_t(I_t(v))$;
- Each $v \in \mathcal{V}$ sends the message $S_t(v) = \left\langle v, t, I_t(v), \ell_t(I_t(v)), p_t^v \right\rangle$ to all their neighbors;
- Each $v \in \mathcal{V}$ receives messages $\{S_{t-d}(v') : v' \in \mathcal{N}(v)\}$ and computes $\hat{L}_{t+1}^{v,obs}$;

end

3.2 A decentralized cooperative follow-the-regularizer-leader algorithm

Theorem 3.1 provides a bound for individual regrets, which increases linearly in the edge-delay parameter d. This can be problematic when the delay in the communication network is large. We show that the effect of delays on regret can be reduced by using a decentralized follow-the-regularized-leader (DFTRL) algorithm.

In the DFTRL algorithm, each agent runs a FTRL algorithm with a *hybrid regularizer* $F_t(p)$ defined in Zimmert and Seldin [21] as follows

$$F_t(p) = \sum_{i=1}^K \left(\frac{-2\sqrt{p_i}}{\eta_t} + \frac{p_i \log(p_i)}{\zeta_t}\right) \tag{2}$$

where η_t and ζ_t are some non-increasing sequences.

As is shown in Theorem 4.1, there is a regret lower bound that consists of two parts: the first part is the regret lower bound of the multi-armed bandit problem and the second part is the regret lower bound of the bandit problem with full-information but delayed feedback [9]. The hybrid regularizer combines the Tsallis entropy regularizer with an optimal regularizer in the full-information setting, the negative entropy regularizer. The learning rates of the two regularizers can be tuned separately to minimize the regret from the two parts. The details of the DFTRL are described in Algorithm 2.

3.2.1 Average regret upper bound. We show a bound on the average regret for the DFTRL algorithm in the following theorem. The sequences η_t and ζ_t are assumed to be set as

$$\eta_t = (1/(1-1/e))(\alpha(G)/N + 1/K)^{-1/4}\sqrt{2/t}$$

and

$$\zeta_t = \sqrt{\log(K)/(dt)}.$$

Theorem 3.4. *Assume that each agent follows the DFTRL algorithm and the delay of each edge is $d \geq 1$, then the average regret*

over N agents is bounded as

$$R_T = O\left(\left(\frac{\alpha(G)}{N} + \frac{1}{K}\right)^{1/4}\sqrt{KT} + \sqrt{d\log(K)T}\right).$$

Proof of the theorem is provided in the supplementary material [20]. We note that the average regret scales as $\sqrt{d}$ which is better than linear scaling of the CFTRL algorithm.

For the special case when $d = 1$, as in Cesa-Bianchi et al. [6],Theorem 3.4 shows that when the number of arms K is large enough, then the DFTRL algorithm has an $O((\alpha(G)/N)^{1/4}\sqrt{KT})$ regret, which is better than $O((\alpha(G)/N)^{1/2}\sqrt{KT}\sqrt{\log(K)})$ of Exp3-Coop from Cesa-Bianchi et al. [6]. Specifically, this improvement holds when $K = \Omega(\exp(\sqrt{N/\alpha(G)}))$.

In what follows we provide a proof sketch of Theorem 3.4. First we present a key lemma whose proof is provided in the supplementary material [20].

Lemma 3.5. *For every agent $v \in \mathcal{A}$ and any probability distribution p^v over $\mathcal{A}$, it holds*

$$\sum_{i\in\mathcal{A}}\sum_{v\in\mathcal{V}}\frac{p^v(i)^{\frac{3}{2}}}{q^v(i)} \le N\sqrt{\frac{1}{1-1/e}\left(\frac{\alpha(G)}{N}+\frac{1}{K}\right)K}$$

where $q^v(i) = 1 - \prod_{v'\in\mathcal{N}(v)}(1-p^{v'}(i))$.

Lemma 3.5 shows that the average regret from the FTRL algorithm with the hybrid regularizer with instant feedback is $O((\alpha(G)/N + 1/K)^{1/4}\sqrt{KT})$. For the regularizer in Eq. (2), the delay effect term is $O(\sqrt{dT\log(K)})$. Lemma 3.2 bounds the average regret by the sum of two terms.

4 REGRET LOWER BOUNDS

We present lower bounds on individual regret R_T^v for every agent $v \in \mathcal{V}$ and average regret R_T.

Theorem 4.1. *The worst-case individual regret of each agent $v \in \mathcal{V}$, R_T^v, is*

$$\Omega\left(\max\left\{\min\left\{T, \frac{1}{\sqrt{|\mathcal{N}(v)|}}\sqrt{KT}\right\}, \sqrt{d\log(K)T}\right\}\right)$$

and the worst-case average regret, R_T, is

$$\Omega\left(\max\left\{\min\left\{T, c_G\sqrt{KT}\right\}, \sqrt{d\log(K)T}\right\}\right)$$

where $c_G = (1/N)\sum_{v\in\mathcal{V}} 1/\sqrt{|\mathcal{N}(v)|}$.

The proof is provided in the supplementary material [20]. The lower bounds contain two parts. The first part is derived from the lower bounds in Shamir [16] for a class of online algorithms. The second part handles the effect of delays by showing that the individual regret of each agent cannot be smaller than the regret of a single agent with delayed full information.

We note that the individual regret of Algorithm 1 is optimal with respect to scaling with the number of arms K and the average regret of Algorithm 2 is optimal with respect to scaling with delay d.

5 NUMERICAL EXPERIMENTS

In this section, we present results of numerical experiments whose goal is to compare performance of CFTRL and DFTRL algorithms with some state-of-the-art algorithms and demonstrate the tightness of our theoretical bounds. We consider the classic stochastic multi-armed bandit problem with agents communicating via different networks.

The stochastic multi-armed bandit problem is defined as follows: each arm i is associated with a Bernoulli distribution with mean μ_i for $i = 1, 2, \ldots, K$. The loss $\ell_t(i)$ from choosing arm i at time step t is sampled independently from the corresponding Bernoulli distribution. In our experiments, we set $\mu_i = (1 + 8(i-1)/(K-1))/10$ so that $\mu_1, \ldots, \mu_K$ is a linearly decreasing sequence. Each problem instance is specified by a tuple (K, G, d). The two baseline algorithms we choose are the center-based Exp3 algorithm in Bar-On and Mansour [2] and the Exp3-Coop algorithm in Cesa-Bianchi et al. [6] whose regret upper bounds are suboptimal as discussed in the introduction.

The numerical results are for four experiments whose goals are as follows:

- the first experiment compares the performance of CFTRL, DFTRL and the baselines when the number of arms increases,
- the second experiment validates the effect of the graph degree in the regret upper bound on CFTRL,
- the third experiment validates the effect of the delay parameter on the regret upper bounds of CFTRL and DFTRL, and, finally,
- the fourth experiment compares CFTRL and the center-based Exp3 algorithm on some sparse random graphs.

In summary, our numerical results validate theoretical results and demonstrate that CFTRL and DFTRL can achieve significant performance gains over some previously proposed algorithms. The code for producing our experimental results is available online in the GitHub repository: [link].

5.1 The effect of the number of arms

In the first experiment, we evaluate the performance of CFTRL and DFTRL against the baselines on a r-regular graph (all nodes have the same degree of r). Note that in a regular graph, each agent has equal probability to be a center agent. For a r-regular graph, CFTRL has an individual regret upper bound of $O(\sqrt{(1/r)}\sqrt{KT})$ and DFTRL has an average regret upper bound of $O(\sqrt[4]{1-r/N}\sqrt{KT})$ when the number of arms K is large enough according to our analysis. We first demonstrate numerical results showing that CFTRL and DFTRL can achieve significant performance gains over the center-based Exp3 algorithm whose individual regret upper bound is $O(\sqrt{(1/r)\log(K)}\sqrt{KT})$ and the Exp3-Coop algorithm whose average regret upper bound is $O(\sqrt{(1-r/N)\log(K)}\sqrt{KT})$ when the number of arms K is large enough.

Figure 1 shows the regret R_T versus T for different number of arms, namely 20, 30 and 40. The results demonstrate that CFTRL and DFTRL achieve better regret than Exp3-COOP and center-based Exp3 when the number of arms is large enough.

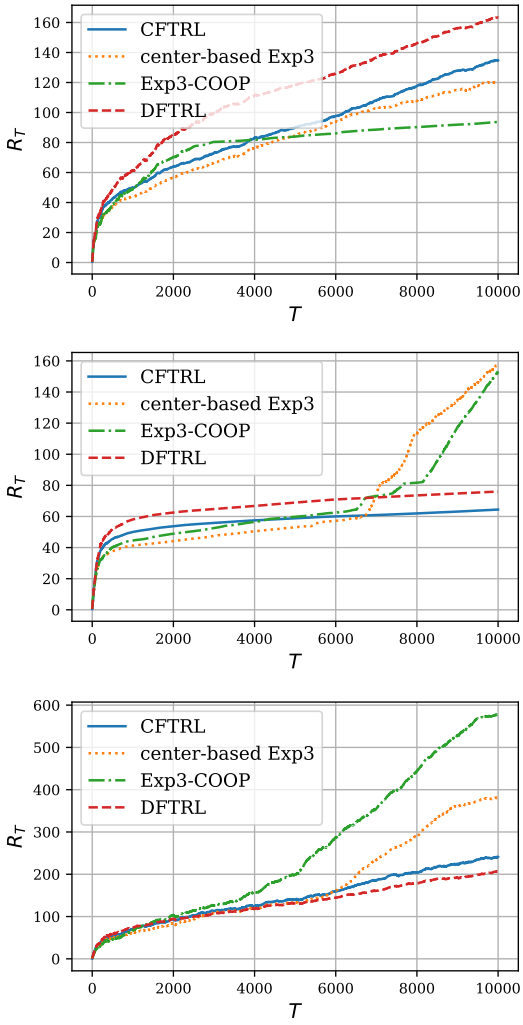

Figure 1: Average regret R_T versus T for different algorithms on a 2-regular graph with $N = 3$ agents and edge-delay $d = 1$, and varied number of arms: (top) $K = 20$ (middle) $K = 30$, and (top) $K = 40$. We used 10 independent simulation runs.

5.2 The effect of graph degree on CFTRL

In the second experiment, we validate the scaling of the graph degree in the regret upper bound of CFTRL. On the r-regular graph, CFTRL has a regret that scales as $O(1/\sqrt{r})$ according to Theorem 3.1. We run CFTRL on the problem instances with fixed number of arms K, delay d and increasing node degree r. The results in Figure 2 shows that the averaged regret decreases as the graph degree increases and the rate of decrease is approximately $O(1/\sqrt{r})$.

5.3 The effect of delay on CFTRL and DFTRL

In the third experiment, we run CFTRL and DFTRL algorithms on a fixed star graph G with a fixed number of arms K and varied edge-delay d . Figure 3 shows that the normalized regret of CFTRL is $R_T/\sqrt{d} = O(\sqrt{d})$ while the normalized regret of DFTRL

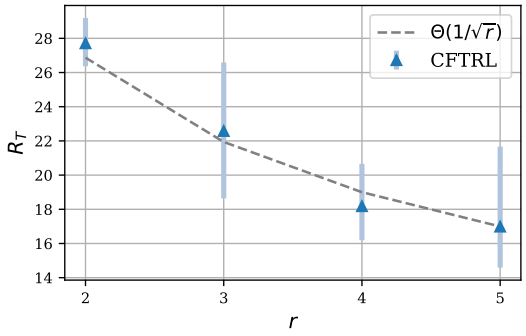

Figure 2: Average regret of CFTRL versus graph degree r, for a r-regular graph with $N = 6$ nodes, $K = 10$ and $d = 1$.

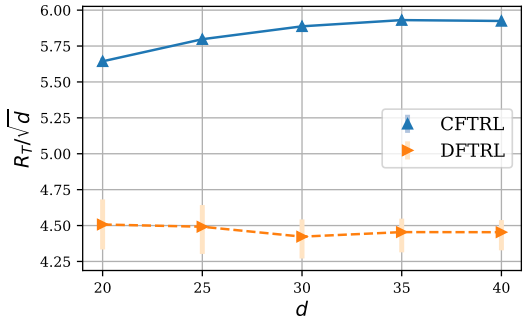

Figure 3: Average regret of CFTRL and DFTRL versus the edge-delay d on a star regular graph with $N = 20$ and $K = 3$.

is $R_T/\sqrt{d} = O(1)$. Hence, when the delay d is large enough, CFTRL has a linearly increasing regret with respect to d which is in contrast to the sub-linear increasing regret of DFTRL. This is consistent with our theoretical analysis, which states that CFTRL has a regret upper bound of $O(d)$ and DFTRL has a regret upper bound of $O(\sqrt{d})$.

5.4 The effect of graph sparsity

In the fourth experiment, we validate that our CFTRL algorithm can outperform the center-based Exp3 algorithm on some random graphs. We consider Erdős–Rényi random graphs of N nodes with probability of an edge equal to $2\log(N)/N$. This condition ensures that the graph is connected and $|\mathcal{N}(v)| = O(\log(N))$ for all $v \in \mathcal{V}$, almost surely [4, Corollary 8.2]. This random graph allows us to evaluate performance of algorithms for a large sparse random graph. We fix K and d and vary the number of nodes N and compare the performance of CFTRL and the center-based Exp3 algorithm on these graphs. We fix T to 1000 time steps. According to our analysis, CFTRL has a lower individual regret bound than the center-based Exp3 algorithm when the number of arms is large enough relative to the number of agents. The results in Figure 4 indicate that CFTRL has at least as good performance as the center-based Exp3 algorithm when K varies, and can have significantly

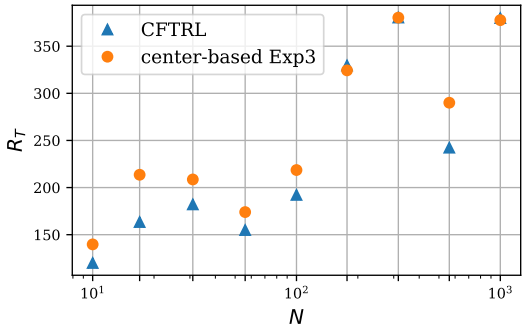

Figure 4: Average regret R_T versus the number of nodes N for sparse Erdős–Rényi random graphs, for CFTRL and center-based Exp3 algorithms.

better performance when the number of arms is large relative to the number of agents.

6 CONCLUSION

We presented new results for the collaborative multi-agent non-stochastic multi-armed bandit with communication delays. We showed a lower bound on the regret of each individual agent and proposed two algorithms (CFTRL and DFTRL) together with their regret upper bounds. CFTRL provides an optimal regret of each individual agent with respect to the scaling with the number of arms. DFTRL has an optimal average regret with respect to the scaling with the edge-delay. Our numerical results validate our theoretical bounds and demonstrate that significant performance gains can be achieved by our two algorithms compared to state-of-the-art algorithms.

There are several open research questions for future research. The first question is to consider the existence of a decentralized algorithm which can provide $O((1/\sqrt{|\mathcal{N}(v)|})\sqrt{KT})$ individual regret for each agent v. It is unclear whether a center-based communication protocol is necessary to achieve this regret. The second question is to consider whether an algorithm exists with an optimal scaling with the number of arms and the edge-delay parameter. The third question is to understand the effect of edge-delay heterogeneity on individual regrets of agents.

REFERENCES

[1] Baruch Awerbuch and Robert Kleinberg. 2008. Competitive collaborative learning. *J. Comput. Syst. Sci.* 74, 8 (2008), 1271–1288.

[2] Yogev Bar-On and Yishay Mansour. 2019. Individual Regret in Cooperative Nonstochastic Multi-Armed Bandits. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 3110–3120.

[3] Ranieri Baraglia, Patrizio Dazzi, Matteo Mordacchini, and Laura Ricci. 2013. A peer-to-peer recommender system for self-emerging user communities based on gossip overlays. *J. Comput. Syst. Sci.* 79, 2 (2013), 291–308.

[4] Avrim Blum, John Hopcroft, and Ravindran Kannan. 2020. *Foundations of Data Science.* Cambridge University Press.

[5] Nicolò Cesa-Bianchi, Tommaso Cesari, and Claire Monteleoni. 2020. Cooperative Online Learning: Keeping your Neighbors Updated. In *Algorithmic Learning Theory, ALT 2020, 8-11 February 2020, San Diego, CA, USA (Proceedings of Machine Learning Research, Vol. 117)*, Aryeh Kontorovich and Gergely Neu (Eds.). PMLR, 234–250.

[6] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2019. Delay and Cooperation in Nonstochastic Bandits. *J. Mach. Learn. Res.* 20 (2019), 17:1–17:38.

[7] Satya R. Chakravarty, Manipushpak Mitra, and Palash Sarkar. 2014. *A Course on Cooperative Game Theory.* Cambridge University Press.

[8] Genevieve Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. 2021. Online Learning with Optimism and Delay. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 3363–3373.

[9] Pooria Joulani, András György, and Csaba Szepesvári. 2013. Online Learning under Delayed Feedback. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013 (JMLR Workshop and Conference Proceedings, Vol. 28)*. JMLR.org, 1453–1461.

[10] Soummya Kar, H. Vincent Poor, and Shuguang Cui. 2011. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *50th IEEE Conference on Decision and Control and European Control Conference, 11th European Control Conference, CDC/ECC 2011, Orlando, FL, USA, December 12-15, 2011*. IEEE, 1771–1778.

[11] Ravi Kumar Kolla, Krishna P. Jagannathan, and Aditya Gopalan. 2018. Collaborative Learning of Stochastic Bandits Over a Social Network. *IEEE/ACM Trans. Netw.* 26, 4 (2018), 1782–1795.

[12] Marc Lanctot, Vinícius Flores Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4190–4203.

[13] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. 2016. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *55th IEEE Conference on Decision and Control, CDC 2016, Las Vegas, NV, USA, December 12-14, 2016*. IEEE, 167–172.

[14] Nancy A. Lynch. 1996. *Distributed Algorithms.* Morgan Kaufmann.

[15] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. 2019. Decentralized Cooperative Stochastic Bandits. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 4531–4542.

[16] Ohad Shamir. 2014. Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 163–171.

[17] Balázs Szörényi, Róbert Busa-Fekete, István Hegedüs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. 2013. Gossip-based distributed stochastic bandit algorithms. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013 (JMLR Workshop and Conference Proceedings, Vol. 28)*. JMLR.org, 19–27.

[18] Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. 2019. Nonstochastic Multiarmed Bandits with Unrestricted Delays. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 6538–6547.

[19] Constantino Tsallis. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics* 52, 1 (1988), 479–487.

[20] Jialin Yi and Milan Vojnovic. 2022. On Regret-optimal Cooperative Nonstochastic Multi-armed Bandits. *arXiv preprint arXiv:2211.17154* (2022).

[21] Julian Zimmert and Yevgeny Seldin. 2020. An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 3285–3294.

[22] Julian Zimmert and Yevgeny Seldin. 2021. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits. *J. Mach. Learn. Res.* 22 (2021), 28:1–28:49.