IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Cooperative Learning for Adversarial Multi-Armed Bandit on Open Multi-Agent Systems

Tomoki Nakamura, Naoki Hayashi, and Masahiro Inuiguchi

Abstract— **This paper considers a cooperative decision-making method for an adversarial bandit problem on open multi-agent systems. In an open multi-agent system, the network configuration changes dynamically as agents freely enter and leave the network. We propose a distributed Exp3 policy in which a group of agents exchanges the estimation of the expected reward of each arm with active neighboring agents. Then, each agent updates the probability distribution of choosing arms by combining the estimated rewards of neighboring agents. We derive a sufficient condition for a sublinear bound of a pseudo regret. The numerical example shows that active agents can cooperatively find the optimal arm by the proposed Exp3 policy algorithm.**

I. INTRODUCTION

The bandit problem is a decision-making problem in which a player agent iteratively learns to choose the best option from a set of candidates [1]. Many bandit algorithms that maximize the total rewards over time, or equivalently, minimize the regret, have been applied to a broad range of decision-making systems such as recommendation systems, clinical trials, and anomaly detection [2]. The bandit problem is divided into two categories: the stochastic bandit problem and the adversarial bandit problem. In the stochastic bandit problem, the distribution of rewards is given in a stochastic manner, while in the adversarial bandit problem, an adversary can set the rewards of arms arbitrarily. The UCB policy and the Thompson sampling are commonly applied for the stochastic bandit problem [3], [4]. For the adversarial bandit problem, the Exp3 policy is a typical approach to finding the best arm against a hostile environment [5].

Recently, according to the growing progress of communication and IoT technology, many learning tasks are expected to be performed in a collaborative way. Control and optimization of multi-agent systems have received significant attention due to their broad applications in such a collaborative setting [6]–[10]. The distributed approach is also applied to the algorithms of the bandit problem. For example, the distributed UCB-based algorithms for the stochastic bandit problem have been widely investigated in [11]–[17]. For the adversarial settings, Cesa-Bianchi et al. proposed the Exp3-Coop algorithm with communication delay [18]. The authors showed a sublinear bound of the average welfare regret can be achieved. Alatur et al. considered a multi-player bandit algorithm based on the Exp3 policy with a coordinator-metaplayer architecture [19]. Yi and Vojnović proposed a cooperative follow-the-regularizer-leader algorithm with delayed information exchange [20].

Most existing distributed methods for the bandit problem assume a closed multi-agent system, where the number of active agents participating in the network is fixed. On the other hand, open multi-agent systems are networked systems where the number of agents and communication links changes dynamically. Many practical networks have such an open structure, where agents freely enter and leave the network on their timing. One application of distributed decision-making is a peer-to-peer file-sharing system where agents decide which peer to connect in order to optimize download speed and resource utilization [21]. In such a peer-to-peer network, agents join and leave the network at any time based on battery life or dynamic network connectivity. The bandit problem arises when agents decide which connections are the most beneficial under the limited bandwidth of the network. Moreover, as the size of a network becomes large, the likelihood of an agent failing increases. In response, additional agents temporarily join the network to maintain its functionality until the affected agent recovers. The motivation for considering a distributed bandit problem in open networks lies in developing a robust and flexible learning architecture in such dynamic environments. In recent years, distributed control and optimization on open multi-agent networks have been investigated to address the dynamic nature of networked systems [22]–[27]. However, to the best of the authors' knowledge, the cooperative bandit problem on open multi-agent networks has not been previously investigated.

This paper focuses on a distributed approach to the adversarial multi-armed bandit problem on open multi-agent systems. We propose a distributed Exp3 policy in which a group of active agents collaboratively searches for the best arm with the highest expected value of the reward. An adversary sets the rewards of arms arbitrary before the active agents take actions. Then, each agent chooses an arm according to the estimated probability distribution. After receiving the reward of the chosen arm, the agents exchange the estimations of the expected rewards with the neighboring agents. The probability of choosing an arm is computed by combining the Hedge algorithm [28] that leverages learned information and the uniform search that explores better arms. The reward estimation is updated by integrating the information of the neighboring active agents with the consensus-based algorithm [6]. We conduct the analysis of the distributed Exp3 policy under an assumption

This work was supported in part by JSPS KAKENHI Grant Number JP21K04121.

The authors are with the Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan (e-mail: nakamura@inulab.sys.es.osaka-u.ac.jp, {n.hayashi, inuiguti}@sys.es.osaka-u.ac.jp). (Corresponding author: Tomoki Nakamura)

of the connectivity of the open network. We show that a pseudo regret of an active agent has a sublinear upper bound. Our results can establish the asymptotic performance of the distributed Exp3 policy for the case when the number of agents and communication links changes dynamically. Thus, the proposed method greatly differs from the existing distributed methods for the adversarial bandit problem [18]–[20], which only consider the performance on a closed system with a fixed number of agents and a time-invariant network. Moreover, in [18]–[20], the communication between agents is assumed to be bidirectional. To conduct the analysis for directed networks, we consider the row stochasticity of the edge weight that is used to obtain the upper bound.

The remainder of this paper is organized as follows. In Section II, we present the problem formulation of the adversarial multi-armed bandit problem and the distributed Exp3 policy on open multi-agent systems. The regret analysis of the proposed policy is conducted in Section III. The illustration via a numerical simulation is shown in Section IV. Finally, concluding remarks are given in Section V.

II. DISTRIBUTED EXP3 POLICY ON OPEN NETWORK

We consider an open multi-agent system over a time-varying undirected graph $\mathcal{G}(t) = (\mathcal{V}(t), \mathcal{E}(t))$, where $\mathcal{V}(t)$ and $\mathcal{E}(t)$ are the sets of agents and communication links at time $t \in \mathcal{T} = \{1, 2, \dots, T\}$. The times when agent i is connected to and disconnected from the network are represented by $t_i^{\rm in}$ and $t_i^{\rm out}$, where $1 \le t_i^{\rm in} < t_i^{\rm out}$. Agent i is said to be active at t if $t \in [t_i^{\rm in}, t_i^{\rm out}]$, and inactive otherwise.

Assumption 1: Once an agent becomes inactive, it will never be active again.

A group of agents cooperatively solves an adversarial multi-armed bandit problem with K arms. Let $X_{i,k}(t)$ be the reward of agent i for arm $k \in \mathcal{K} = \{1, 2, ..., K\}$ at time $t \in \mathcal{T}$.

Let V' be the maximum number of active agents in the time horizon, that is, $V' = \max_{t\in\mathcal{T}} |\mathcal{V}(t)|$. We assume that $V' > 2$. We make the following assumption of the connectivity of the network.

Assumption 2: There exists $Q > 0$ such that $\mathcal{G}_Q(s) = (\mathcal{V}_Q(s), \mathcal{E}_Q(s))$ is strongly connected for $s \ge 1$ with $T \ge (s+1)Q - 1$, where $\mathcal{V}_Q(s) = \bigcup_{t=sQ}^{(s+1)Q-1} \mathcal{V}(t)$ and $\mathcal{E}_Q(s) \subset \mathcal{V}_Q(s) \times \mathcal{V}_Q(s)$ is the sets of vertices and edges for the union of the graphs in the interval $[sQ, (s+1)Q - 1]$.

Assumption 2 is an extension of the strong connectivity in a bounded subinterval for a closed multi-agent system to the case for an open network [24].

In the adversarial bandit problem, each agent obtains the reward of the chosen arm. To effectively estimate the optimal arm, we consider the distributed method by sharing the information of the rewards among agents. The proposed distributed Exp3 algorithm is shown in Algorithm 1. At step 4, the probability of choosing arm k is computed by the convex combination of the Hedge algorithm to exploit the learned information and the uniform search to explore better arms. The trade-off parameter γ determines the balance between the exploitation and the exploration. At step 6, each agent i randomly chooses arm $k_i(t)$ according to the probabilities $p_{i,1}(t), p_{i,2}(t), \dots, p_{i,K}(t)$. Then, agent i receives the reward $X_{i,k_i(t)}(t) \in [0, 1]$ for arm $k_i(t)$ at step 7. At step 9, the estimation of the reward $\hat{X}_{i,k}(t)$ is updated by integrating the rewards of the neighboring active agents with the consensus-based update, where $a_{ij}(t)$ is the edge weight for the communication link $\{i, j\} \in \mathcal{E}(t)$. Finally, the weight $w_{i,k}(t+1)$ for each arm is updated at step 10, where η is a learning parameter determining the extent to which learned estimation is exploited. The parameters and the weight are initialized as $0 < \gamma < 1$, $0 < \eta \le \gamma/K$, and $w_{i,k}(t_i^{\rm in}) = 1/K^\alpha$, where $0 < \alpha < 1$.

Algorithm 1 Distributed Exp3 algorithm

Parameters $0 < \gamma < 1$, $0 < \eta \le \gamma/K$.
Initilization $w_{i,k}(t_i^{\rm in}) = 1/K^\alpha$ for all $k \in \mathcal{K}$.

1: **for** $t \in \mathcal{T}$ **do**
2: **for** $i \in \mathcal{V}(t)$ **do**
3: **for** $k \in \mathcal{K}$ **do**
4: Compute the probability of arm k by

$$p_{i,k}(t) = (1-\gamma)\frac{w_{i,k}(t)}{\sum_{k=1}^{K} w_{i,k}(t)} + \frac{\gamma}{K}. \tag{1}$$

5: **end for**
6: Choose arm $k_i(t)$ according to the probabilities $\{p_{i,\ell}(t)\}_{\ell\in\mathcal{K}}$.
7: Receive the reward $X_{i,k_i(t)}(t)$.
8: **for** $k \in \mathcal{K}$ **do**
9: Update the estimation of the reward by

$$\hat{X}_{i,k}(t) = \begin{cases} \frac{\sum_{j\in\mathcal{V}(t)} a_{ij}(t) X_{j,k}(t)}{p_{i,k}(t)}, & \text{if } k = k_i(t), \\ 0, & \text{if } k \ne k_i(t). \end{cases} \tag{2}$$

10: Update the weights for the exploitation by

$$w_{i,k}(t+1) = w_{i,k}(t) e^{\eta \hat{X}_{i,k}(t)}. \tag{3}$$

11: **end for**
12: **end for**
13: **end for**

The edge weight satisfying Assumption 4 can be set as $a_{ij}(t) = 1/(d_i(t) + \sigma)$ if agent i receives the estimation from j, and $a_{ii}(t) = 1 - \sum_{j\in\mathcal{V}(t)} a_{ij}(t)$, where $d_i(t)$ is the

Assumption 3: The unconstrained reward model is considered, that is, if two or more agents choose the same arm, they receive the same reward independently. The reward is given as $X_{i,k}(t) \in [0, 1]$ for all $i \in \mathcal{V}(t)$, $k \in \mathcal{K}$, and $t \in \mathcal{T}$. Moreover, we assume that $X_{i,k}(t) = 0$ and $\hat{X}_{i,k}(t) = 0$ for all $i \notin \mathcal{V}(t)$, $k \in \mathcal{K}$, and $t \in \mathcal{T}$.

In this paper, we also make the stochasticity assumption for the edge weight of an active agent.

Assumption 4: $\sum_{j\in\mathcal{V}(t)} a_{ij}(t) = 1$ for all $i \in \mathcal{V}(t)$ and $t \in \mathcal{T}$.

number of incoming edges of i and $\sigma > 0$.

III. REGRET FOR DISTRIBUTED EXP3 POLICY

In this section, we evaluate the performance of the distributed Exp3 algorithm in Algorithm 1. We consider the following team pseudo regret.

$$\overline{\mathrm{Regret}} = \bar{G}^* - \mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)} X_{i,k_i(t)}(t)\right], \tag{4}$$

where $\bar{G}^* = \max_{k\in\mathcal{K}} \mathbb{E}[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)} X_{i,k}(t)]$.

The pseudo-regret (4) evaluates the difference between the optimal expected reward and the expectation of the actual reward. Thus, the purpose of each agent is to choose arms in order to achieve smaller regret bound by sharing the information on the optimal arm among neighboring agents.

We now consider the performance of the proposed algorithm through a regret analysis. First, we derive an upper bound of the pseudo regret by the distributed Exp3 algorithm.

Theorem 1: If each agent updates the estimation of the rewards by Algorithm 1, we have $\overline{\mathrm{Regret}} \le (\gamma+\eta K)\bar{G}^* + ((1-\gamma)/\eta)V'\ln K$.

Proof: We define $W_i(t)$ as $W_i(t) = \sum_{k=1}^K w_{i,k}(t)$. From (1), for all $t \in [t_i^{\mathrm{in}}, t_i^{\mathrm{out}}]$, we have $\sum_{k=1}^K p_{i,k}(t) = (1-\gamma)(\sum_{k=1}^K w_{i,k}(t))/W_i(t) + K\cdot\gamma/K = 1$. Thus, from (1), we have

$$0 < \frac{\gamma}{K} \le p_{i,k}(t) \le 1. \tag{5}$$

From (3), we have

$$\begin{aligned} W_i(t_i^{\mathrm{out}}) &= \sum_{k=1}^K w_{i,k}(t_i^{\mathrm{out}}) \\ &= \sum_{k=1}^K w_{i,k}(t_i^{\mathrm{out}}-1)e^{\eta\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)}. \end{aligned} \tag{6}$$

From (2), we also have

$$0 \le \eta\hat{X}_{i,k}(t) \le \eta\frac{\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k}(t)}{p_{i,k}(t)} \le \eta\frac{K}{\gamma} \le 1, \tag{7}$$

where the third inequality follows from (5), Assumption 4, and the fact that $0 \le X_{i,k}(t) \le 1$. We note that $e^x \le 1 + x + x^2$ holds for any $x \in [0,1]$. Thus, from (6) and (7), we have

$$\begin{aligned} W_i(t_i^{\mathrm{out}}) &= \sum_{k=1}^K w_{i,k}(t_i^{\mathrm{out}}-1)e^{\eta\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)} \\ &\le \sum_{k=1}^K w_{i,k}(t_i^{\mathrm{out}}-1)(1+\eta\hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &\quad + (\eta\hat{X}_{i,k}(t_i^{\mathrm{out}}-1))^2) \\ &= W_i(t_i^{\mathrm{out}}-1) \\ &\quad \left(1+\eta\sum_{k=1}^K \frac{w_{i,k}(t_i^{\mathrm{out}}-1)}{W_i(t_i^{\mathrm{out}}-1)}\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)\right. \\ &\quad \left. +\eta^2\sum_{k=1}^K \frac{w_{i,k}(t_i^{\mathrm{out}}-1)}{W_i(t_i^{\mathrm{out}}-1)}\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)^2\right). \end{aligned} \tag{8}$$

From (1), we have $p_{i,k}(t_i^{\mathrm{out}}-1) = (1-\gamma)w_{i,k}(t_i^{\mathrm{out}}-1)/W_i(t_i^{\mathrm{out}}-1) + \gamma/K$. Thus, we obtain

$$\begin{aligned} \frac{w_{i,k}(t_i^{\mathrm{out}}-1)}{W_i(t_i^{\mathrm{out}}-1)} &= \frac{1}{1-\gamma}p_{i,k}(t_i^{\mathrm{out}}-1) - \frac{\gamma}{(1-\gamma)K} \\ &\le \frac{1}{1-\gamma}p_{i,k}(t_i^{\mathrm{out}}-1). \end{aligned} \tag{9}$$

From (9), we have

$$\begin{aligned} &\eta\sum_{k=1}^K \frac{w_{i,k}(t_i^{\mathrm{out}}-1)}{W_i(t_i^{\mathrm{out}}-1)}\hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &\le \frac{\eta}{1-\gamma}\sum_{k=1}^K p_{i,k}(t_i^{\mathrm{out}}-1)\hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &= \frac{\eta}{1-\gamma}p_{i,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1)\hat{X}_{i,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1) \\ &\le \frac{\eta}{1-\gamma}p_{i,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1) \\ &\quad \times \frac{\sum_{j\in\mathcal{V}(t_i^{\mathrm{out}}-1)} a_{ij}(t_i^{\mathrm{out}}-1)X_{j,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1)}{p_{i,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1)} \\ &= \frac{\eta}{1-\gamma}\sum_{j\in\mathcal{V}(t_i^{\mathrm{out}}-1)} a_{ij}(t_i^{\mathrm{out}}-1)X_{j,k_i(t_i^{\mathrm{out}}-1)}(t_i^{\mathrm{out}}-1), \end{aligned} \tag{10}$$

where the first equality follows from the fact that the estimated reward has a positive value only for the chosen arm and the second inequality follows from (2).

From (9), we also have

$$\begin{aligned} &\eta^2\sum_{k=1}^K \frac{w_{i,k}(t_i^{\mathrm{out}}-1)}{W_i(t_i^{\mathrm{out}}-1)}\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)^2 \\ &\le \frac{\eta^2}{1-\gamma}\sum_{k=1}^K p_{i,k}(t_i^{\mathrm{out}}-1)\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)^2 \\ &\le \frac{\eta^2}{1-\gamma}\sum_{k=1}^K p_{i,k}(t_i^{\mathrm{out}}-1)\hat{X}_{i,k}(t_i^{\mathrm{out}}-1)\hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &\le \frac{\eta^2}{1-\gamma}\sum_{k=1}^K p_{i,k}(t_i^{\mathrm{out}}-1) \\ &\quad \times \frac{\sum_{j\in\mathcal{V}(t_i^{\mathrm{out}}-1)} a_{ij}(t_i^{\mathrm{out}}-1)X_{j,k}(t_i^{\mathrm{out}}-1)}{p_{i,k}(t_i^{\mathrm{out}}-1)}\hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &= \frac{\eta^2}{1-\gamma}\sum_{k=1}^K \left(\sum_{j\in\mathcal{V}(t_i^{\mathrm{out}}-1)} a_{ij}(t_i^{\mathrm{out}}-1)X_{j,k}(t_i^{\mathrm{out}}-1)\right) \\ &\quad \times \hat{X}_{i,k}(t_i^{\mathrm{out}}-1) \\ &\le \frac{\eta^2}{1-\gamma}\sum_{k=1}^K \left(\sum_{j\in\mathcal{V}(t_i^{\mathrm{out}}-1)} a_{ij}(t_i^{\mathrm{out}}-1)\right)\hat{X}_{i,k}(t_i^{\mathrm{out}}-1), \\ &= \frac{\eta^2}{1-\gamma}\sum_{k=1}^K \hat{X}_{i,k}(t_i^{\mathrm{out}}-1), \end{aligned} \tag{11}$$

where the last equality follows from Assumption 4.

From (8), (10), and (11), we obtain

$$\begin{aligned}
&W_i(t_i^{\text{out}}) \\
&\le W_i(t_i^{\text{out}}-1) \\
&\quad\Bigg(1+\frac{\eta}{1-\gamma} \\
&\qquad\times \sum_{j\in\mathcal{V}(t_i^{\text{out}}-1)} a_{ij}(t_i^{\text{out}}-1)X_{j,k_i(t_i^{\text{out}}-1)}(t_i^{\text{out}}-1) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_{i,k}(t_i^{\text{out}}-1)\Bigg). \qquad (12)
\end{aligned}$$

By iteratively solving (12), we obtain

$$\begin{aligned}
&W_i(t_i^{\text{out}}) \\
&\le W_i(t_i^{\text{in}})\prod_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\Bigg(1+\frac{\eta}{1-\gamma}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k_i(t)}(t) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_{i,k}(t)\Bigg). \qquad (13)
\end{aligned}$$

From (3), for any arm $k\in\mathcal{K}$, we have

$$\begin{aligned}
W_i(t_i^{\text{out}}) &= \sum_{\ell=1}^{K} w_{i,\ell}(t_i^{\text{out}}) \\
&\ge w_{i,k}(t_i^{\text{out}}) \\
&= w_{i,k}(t_i^{\text{out}}-1)e^{\eta\hat{X}_{i,k}(t_i^{\text{out}}-1)} \\
&= w_{i,k}(t_i^{\text{in}})e^{\eta\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\hat{X}_{i,k}(t)}. \qquad (14)
\end{aligned}$$

From (13) and (14), we obtain

$$\begin{aligned}
&w_{i,k}(t_i^{\text{in}})e^{\eta\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\hat{X}_{i,k}(t)} \\
&\le W_i(t_i^{\text{in}})\prod_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\Bigg(1+\frac{\eta}{1-\gamma}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k_i(t)}(t) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_{i,k}(t)\Bigg). \qquad (15)
\end{aligned}$$

By taking the natural logarithm for (15) and using the initialization of $w_{i,k}(t_i^{\text{in}})=K^{-\alpha}$ and $W_i(t_i^{\text{in}})=K^{1-\alpha}$, we have

$$\begin{aligned}
&-\ln K+\eta\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\hat{X}_{i,k}(t) \\
&\le \sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\ln\Bigg(1+\frac{\eta}{1-\gamma}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k_i(t)}(t) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_{i,k}(t)\Bigg).
\end{aligned}$$

Since $\ln(1+x)\le x$ holds for any $x\ge 0$, we further have

$$\begin{aligned}
&-\ln K+\eta\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\hat{X}_{i,k}(t) \\
&\le \frac{\eta}{1-\gamma}\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k_i(t)}(t) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\sum_{t=t_i^{\text{in}}}^{t_i^{\text{out}}-1}\hat{X}_{i,k}(t).
\end{aligned}$$

We note that $\hat{X}_{i,k}(t)$ is the unbiased estimator of $X_{i,k}(t)$, and $X_{i,k}(t)=0$ and $\hat{X}_{i,k}(t)=0$ for $t\notin[t_i^{\text{in}},t_i^{\text{out}}]$. Thus, by taking the expectation with respect to the estimated distribution of the rewards obtained by Algorithm 1, we have

$$\begin{aligned}
&-\ln K+\eta\sum_{t=1}^{T-1}X_{i,k}(t) \\
&\le \frac{\eta}{1-\gamma}\sum_{t=1}^{T-1}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{j,k_i(t)}(t) \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\sum_{t=1}^{T-1}X_{i,k}(t). \qquad (16)
\end{aligned}$$

Since (16) holds for each active agent, by taking the expectation with respect to the true distribution of the rewards, we have

$$\begin{aligned}
&-V'\ln K+\eta\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}X_{i,k}(t)\right] \\
&\le \frac{\eta}{1-\gamma}\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}\sum_{j\in\mathcal{V}(t)} a_{ij}(t)X_{i,k_i(t)}(t)\right] \\
&\quad+\frac{\eta^2}{1-\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}X_{i,k}(t)\right] \\
&\le \frac{\eta}{1-\gamma}\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}X_{i,k_i(t)}(t)\right]+\frac{\eta^2K}{1-\gamma}\bar{G}^*, \qquad (17)
\end{aligned}$$

where the first inequality follows from the unconstrained reward model in Assumption 3, and the last inequality follows from $\bar{G}^*\ge(1/K)\sum_{k=1}^{K}\mathbb{E}[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}X_{i,k}(t)]$ and Assumption 4. Since (17) holds for any $k\in\mathcal{K}$, we obtain

$$\begin{aligned}
&-V'\ln K+\eta\bar{G}^* \\
&\le \frac{\eta}{1-\gamma}\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in\mathcal{V}(t)}X_{i,k_i(t)}(t)\right]+\frac{\eta^2K}{1-\gamma}\bar{G}^*.
\end{aligned}$$

This concludes the proof. ■

Theorem 1 shows that the regret bound by the distributed Exp3 algorithm depends on the learning parameter η and the trade-off parameter γ. The next proposition addresses that a sublinear regret can be achieved by appropriately setting these parameters.

Proposition 1: Suppose that the learning parameter and the trade-off parameter are given as $\eta = \min\{c/K, \sqrt{V' \ln K/(2gK)}\}$ and $\gamma = \eta K$, where $0 < c < 1$ and $R^* \leq R$. If each agent updates the estimation of the rewards by Algorithm 1, we have $\overline{\text{Regret}} \leq (2/c)\sqrt{2V'gK \ln K}$.

Proof: If $\sqrt{V' \ln K/(2gK)} \geq c/K$, $2g \leq (V'K \ln K)/c^2$ holds. Thus, we have

$$\overline{\text{Regret}} \leq \bar{G}^* \leq 4g = 2\sqrt{2g}\sqrt{2g} \leq \frac{2}{c}\sqrt{2V'gK \ln K}. \tag{18}$$

If $\sqrt{V' \ln K/(2gK)} < c/K$, $\eta = \sqrt{V' \ln K/(2gK)}$ holds. Moreover, we have $\eta < c/K$, and hence, $1 - \gamma = 1 - \eta K > 0$ holds. Thus, from Theorem 1, we obtain

$$\begin{aligned}\overline{\text{Regret}} &\leq (\gamma + \eta K)\bar{G}^* + \frac{V'}{\eta} \ln K \\ &\leq 2\eta K \bar{G}^* + V' \sqrt{\frac{2gK}{V' \ln K}} \ln K \\ &\leq 2\sqrt{2V'gK \ln K}.\end{aligned}$$

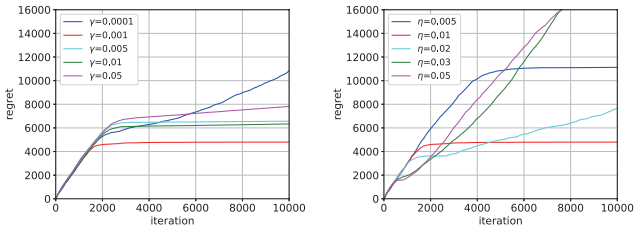
■

From the definition of the pseudo regret, $g \leq T$ holds. Thus, from Proposition 1, we have $\overline{\text{Regret}} \leq (2/c)\sqrt{2TV'K \ln K}$, which shows a sublinear regret bound of $O(\sqrt{TV'})$. If each agent estimates the optimal arm without information exchange over a network, the regret bound becomes $V' \times O(\sqrt{T})$. Hence, Proposition 1 shows that the proposed algorithm is effective for large-scale open networks.

IV. NUMERICAL EXAMPLE

We consider the open multi-agent system in which the number of agents varies at each iteration. The reward of arm 1 at iteration $t \in \mathcal{T} = \{1, 2, \ldots, 10000\}$ is randomly set from the interval $[0.8, 1.0]$. The reward of arm $k \in \mathcal{K} = \{2, 3, \ldots, K\}$ is randomly set from the interval $[0.0, 0.6]$ if the indices i and k are both even or both odd, and $[0.4, 0.8]$ otherwise. Thus, arm 1 is the best arm and should be chosen as many as possible. The connected time t_i^{in} and the disconnected time t_i^{out} are randomly given in the horizon period $[1, 10000]$. The edge weight is set as $a_{ij}(t) = 1/(d_i(t) + 1)$ if agent i receives the estimation from j, and $a_{ii}(t) = 1 - \sum_{j \in \mathcal{V}(t)} a_{ij}(t)$.

We compare the performance of the proposed algorithm for different values of the trade-off parameter γ. Fig. 1(a) shows the team pseudo regret (4). The learning parameter, the number of arms, and the number of maximum agents are set to $\eta = 0.01$, $K = 20$, and $V' = 10$, respectively. Although the sublinear regret can be achieved for all cases, the evolution of the regret is different depending on the value of the trade-off parameter. As observed in (1), an agent loses the opportunity to find better arms if the value of γ is too small; conversely, if γ is too large, the agent cannot continue to select better arms. In this example, the value of 0.001 for γ can achieve a good balance between the exploitation and the exploration.



(a) Different trade-off parameter γ. (b) Different learning parameter η.

Fig. 1. Pseudo regret.

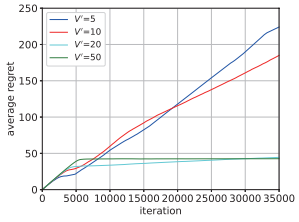


Fig. 2. Average pseudo regret per active agent with different size of networks.

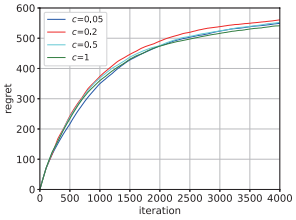


Fig. 3. Pseudo regret for different parameter c.

We also compare the performance of the proposed algorithm for different values of the learning parameter η. Fig. 1(b) shows the team pseudo regret. The trade-off parameter, the number of arms, and the number of maximum agents are set to $\gamma = 0.1$, $K = 20$, and $V' = 10$, respectively. As in the case of the trade-off parameter γ, the evolution of the regret depends on the value of the learning parameter; too large or too small parameter values result in a larger regret.

Next, the performance is compared by varying the size of the network. In this example, the horizon period and the number of arms are set to $T = 35000$ and $K = 50$. The trade-off parameter and the learning parameter are set to $\gamma = 0.001$ and $\eta = 0.01$. Fig. 2 shows the average pseudo regret with different values for the maximum number of the active agents V'. The regret curve in Fig. 2 shows the average regret per active agent. In this example, the average regret tends to be smaller for the case when the number of active agents is large. This is because the information of the best arm can be effectively searched through the information exchange over the network and the consensus-based update of the estimation of the rewards.

Finally, we compare the performance with different values of the parameter c in Proposition 1. The horizon period, the number of arms, and the number of maximum agents are set to $T = 4000$, $K = 5$, and $V' = 6$, respectively. The learning parameter and the trade-off parameter are set to $\eta = \min\{c/K, \sqrt{\ln K/(2TK)}\}$ and $\gamma = \eta T$. In this example, the reward is given in an adversarial way such that $X_{i,k}(t) \in [0.1, 0.5]$ if $k = k_i(t-1)$ and $X_{i,k}(t) \in [0.8, 1.0]$ otherwise. Thus, the adversary gives the lower reward for the arm chosen at the previous iteration to disturb the agent's estimation. The regret curves in Fig. 3 depend on the values of the parameter c. The optimal parameter value is problem-dependent and requires tuning by trial and error. Deriving optimal parameter settings is a part of the future work.

V. CONCLUSION

In this paper, we proposed a distributed Exp3 algorithm for the adversarial bandit problem in an open multi-agent system, in which each agent freely enters and leaves the communication network. We showed that an upper bound of the pseudo regret that evaluates the error between the optimal expected reward and the expectation of the actual reward. Furthermore, we derived a sufficient condition with respect to the trade-off and learning parameters to achieve a sublinear regret bound. Future work includes investigating the adversarial bandit problem with communication delays. The linear bandit problem in non-stationary and non-Markovian settings has been explored in [29], [30]. Analyzing distributed decision-making in non-stationary and non-Markovian environments is also a future direction of this research.

REFERENCES

[1] T. Lattimore and C. Szepesvári., *Bandit Algorithms*. Cambridge University Press,, 2019.

[2] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," *Proceedings of the 2020 IEEE Congress on Evolutionary Computation*, pp. 1–8, 2020.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[4] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on Thompson sampling," *Foundations and Trends in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[6] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[7] N. Hayashi, T. Sugiura, Y. Kajiyama, and S. Takai, "Event-triggered consensus-based optimization algorithm for smooth and strongly convex cost functions," *Proceedings of the 57th IEEE Conference on Decision and Control*, pp. 2120–2125, 2018.

[8] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[9] A. M. C. So, P. Jain, W. K. Ma, and G. Scutari, "Nonconvex optimization for signal processing and machine learning," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 15–17, 2020.

[10] T. Adachi, N. Hayashi, and S. Takai, "Distributed gradient descent method with edge-based event-driven communication for non-convex optimization," *IET Control Theory & Applications*, vol. 15, no. 12, pp. 1588–1598, 2021.

[11] P. Landgren, V. Srivastava, and N. E. Leonard, "Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms," *Proceeding of the 55th IEEE Conference on Decision and Control*, pp. 167–172, 2016.

[12] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2786–2790, 2017.

[13] A. Sankararaman, A. Ganesh, and S. Shakkottai, "Social learning in multi agent multi armed bandits," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 3, pp. 1–35, 2019.

[14] D. Martínez-Rubio, V. Kanade, and P. Rebeschini, "Decentralized cooperative stochastic bandits," *Proceedings of the Advances in Neural Information Processing Systems 32*, pp. 4531–4542, 2019.

[15] P. Landgren, V. Srivastava, and N. E. Leonard, "Distributed cooperative decision making in multi-agent multi-armed bandits," *Automatica*, vol. 125, p. 109445, 2021.

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

[17] Z. Yan, Q. Xiao, T. Chen, and A. Tajer, "Federated multi-armed bandit via uncoordinated exploration," *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5248–5252, 2022.

[18] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," *Proceedings of the 29th Annual Conference on Learning Theory*, vol. 49, pp. 605–622, 2016.

[19] P. Alatur, K. Y. Levy, and A. Krause, "Multi-player bandits: The adversarial case," *Journal of Machine Learning Research*, vol. 21, pp. 77:1–77:23, 2020.

[20] J. Yi and M. Vojnović, "On regret-optimal cooperative nonstochastic multi-armed bandits," *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1329–1335, 2023.

[21] N. Korda, B. Szorenyi, and S. Li, "Distributed clustering of linear bandits in peer to peer networks," *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, pp. 1301–1309, 2016.

[22] J. M. Hendrickx and S. Martin, "Open multi-agent systems: Gossiping with random arrivals and departures," *Proceedings of the 56th IEEE Conference on Decision and Control*, pp. 763–768, 2017.

[23] M. Franceschelli and P. Frasca, "Stability of open multiagent systems and applications to dynamic consensus," *IEEE Transactions on Automatic Control*, vol. 66, no. 5, pp. 2326–2331, 2021.

[24] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, "Optimization in open networks via dual averaging," 2021. [Online]. Available: arXiv:2105.13348

[25] N. Hayashi, "Distributed subgradient method in open multiagent systems," *IEEE Transactions on Automatic Control*, 2022. [Online]. Available: 10.1109/TAC.2022.3230771

[26] R. Vizuete, C. M. de Galland, J. M. Hendrickx, P. Frasca, and E. Panteley, "Resource allocation in open multi-agent systems: An online optimization analysis," *Proceedings of the 2022 IEEE 61st Conference on Decision and Control*, pp. 5185–5191, 2022.

[27] Z. A. Z. S. Dashti, G. Oliva, C. Seatzu, A. Gasparri, and M. Franceschelli, "Distributed mode computation in open multi-agent systems," *IEEE Control Systems Letters*, vol. 6, pp. 3481–3486, 2022.

[28] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[29] Y. Qin, T. Menara, S. Oymak, S. Ching, and F. Pasqualetti, "Non-stationary representation learning in sequential linear bandits," *IEEE Open Journal of Control Systems*, vol. 1, pp. 41–56, 2022.

[30] Y. Qin, Y. Li, F. Pasqualetti, M. Fazel, and S. Oymak, "Stochastic contextual bandits with long horizon rewards," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, pp. 9525–9533, 2023.