PAPER Special Section on Mathematical Systems Science and its Applications

Consensus-Based Distributed Exp3 Policy Over Directed Time-Varying Networks

Tomoki NAKAMURA[†], ***Nonmember***, **Naoki HAYASHI**[†a)], ***and*** **Masahiro INUIGUCHI**[†], ***Members***

SUMMARY In this paper, we consider distributed decision-making over directed time-varying multi-agent systems. We consider an adversarial bandit problem in which a group of agents chooses an option from among multiple arms to maximize the total reward. In the proposed method, each agent cooperatively searches for the optimal arm with the highest reward by a consensus-based distributed Exp3 policy. To this end, each agent exchanges the estimation of the reward of each arm and the weight for exploitation with the nearby agents on the network. To unify the explored information of arms, each agent mixes the estimation and the weight of the nearby agents with their own values by a consensus dynamics. Then, each agent updates the probability distribution of arms by combining the Hedge algorithm and the uniform search. We show that the sublinearity of a pseudo-regret can be achieved by appropriately setting the parameters of the distributed Exp3 policy.

key words: *distributed decision-making, multi-armed bandit problem, multi-agent system*

1. Introduction

The multi-armed bandit problem is a decision-making problem where a player agent repeatedly chooses an action, referred to as an arm, in order to maximize its cumulative reward over a sequence of trials. The challenge of the problem lies in balancing exploration of new arms and exploitation of known high-reward arms. Different algorithms have been developed to address this trade-off, such as the upper confidence bound algorithm [1] and Thompson sampling [2]. These algorithms aim to find the optimal balance between exploration and exploitation, and are widely applied in fields such as online advertising, recommendation systems, and clinical trials [3]–[5].

In some applications, one needs to consider the situations where an agent must deal with an opponent. This type of the bandit problem is called the adversarial multi-armed bandit problem [6], [7]. In the adversarial multi-armed bandit problem, an agent faces a more challenging scenario, where the reward distributions of the arms are not fixed but can be adversarially chosen to disrupt agent's learning process. The Exp3 (Exponential-weight algorithm for Exploration and Exploitation) policy is a widely used approach for solving the adversarial bandit problem [8]. The Exp3 algorithm balances exploration and exploitation by assigning probabilities to each arm based on their past rewards.

Manuscript received April 1, 2023.
Manuscript revised August 25, 2023.
Manuscript publicized October 16, 2023.

[†]The authors are with Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560-8531 Japan.
a) E-mail: n.hayashi@sys.es.osaka-u.ac.jp

DOI: 10.1587/transfun.2023MAP0008

Recently, the need for a distributed approach has increased in many large-scale machine learning and optimization problems, where the datasets or models are too large to be processed by a single processor [9]–[16]. In a distributed bandit algorithm, multiple agents cooperatively learn the optimal arm by communicating with each other over a network [17]–[20]. Each agent cooperatively makes a decision by combining the own estimation of the rewards of the arms with those of the nearby agents. The advantage of the distributed multi-armed bandit algorithm over the centralized one is that it can achieve a smaller upper bound on the regret. This is because in a distributed setting, agents explore the arms more effectively, which allows for faster exploration and more efficient exploitation.

For the cooperative adversarial multi-armed bandit problem, Cesa-Bianchi et al. proposed the Exp3-Coop algorithm with communication delay and analyzed the impact of delays between players' decisions and the potential benefits of cooperation [21]. Bar-On and Mansour proposed a distributed algorithm for the nonstochastic bandit problem, which allows agents to learn independently of one another while still achieving a cooperative goal [22]. Alatur et al. addressed the multi-armed bandit problem of multiple players competing for limited resources based on an adaptation of the Exp3 policy [23]. Yi and Vojnović proposed a decentralized follow-the-regularizer-leader algorithm with communication delays [24]. Although these methods with the adversarial settings assume undirected or static communication graphs, considering the case with directed and time-varying communication graphs is particularly important because agents are limited to sending messages in specific directions in many networked systems.

To relax such a limitation on the network topology, this paper focuses on a cooperative adversarial multi-armed bandit problem, in which multiple agents work together on directed and time-varying communication graphs to maximize the collective reward. We propose a distributed Exp3 policy, in which the learning process is distributed across multiple player agents. Each agent maintains a local estimate of the reward distribution of arms. These estimations are combined by the consensus algorithm [25]–[27] to update the probability distribution used for arm selection. As opposed to the existing work, such as [21]–[24], we do not make the assumption of omnidirectionality of communication. Thus, the proposed algorithm can be used in a wider range of applications.

The remainder of this paper is organized as follows.

Section 2 presents the distributed Exp3 policy for the adversarial multi-armed bandit problem. The regret analysis of the proposed policy is conducted in Sect. 3. The numerical example of the proposed method is shown in Sect. 4. Finally, concluding remarks are given in Sect. 5.

2. Distributed Exp3 Policy

In the distributed multi-armed bandit problem, a group of agents works together to learn the best arm to maximize a reward. Each agent can only observe the reward for the arm it chooses. Thus, agents share information about the rewards with other agents over a communication network. In this paper, we model the communication network as a time-varying directed graph without a self-loop $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$, where $\mathcal{V} = \{1, 2, \ldots, N\}$ and $\mathcal{E}(t) \subset \mathcal{V} \times \mathcal{V}$ are the sets of agents and communication links at time $t \in \mathcal{T} = \{1, 2, \ldots, T-1\}$. We consider an adversarial multi-armed bandit problem with K arms. Let $X_{i,k}(t) \in [0, 1]$ be the reward of agent i for arm $k \in \mathcal{K} = \{1, 2, ..., K\}$ at time $t \in \mathcal{T}$. In this paper, the unconstrained reward model is considered, that is, if two or more agents choose the same arm, they receive the same reward independently [18].

In the proposed distributed Exp3 algorithm, the probability of choosing arm k is updated by

$$p_{i,k}(t) = (1-\alpha)\frac{w_{i,k}(t)}{W_i(t)} + \frac{\alpha}{K}, \tag{1}$$

where $\alpha \in (0, 1)$ is a trade-off parameter. Equation (1) implies that the probability of choosing arm k is computed by combining the Hedge algorithm to exploit the learned information and the uniform search to explore better arms. The weights for exploitation are initialized as $w_{i,k}(1) = 1/K^{\nu}$ for all $i \in \mathcal{V}$ and $k \in \mathcal{K}$, and $W_i(1) = W(1) = K^{1-\nu}$ for all $i \in \mathcal{V}$, where $0 < \nu \leq 1$.

After updating the probabilities $p_{i,1}(t)$, $p_{i,2}(t)$, $\ldots$, $p_{i,K}(t)$, agent i chooses arm $k_i(t)$ according to these probabilities. Then, the reward $X_{i,k_i(t)}(t)$ for arm $k_i(t)$ is feedbacked to the agent. The information of the reward $\hat{X}_{i,k}(t)$ of the nearby agents is unified by the consensus dynamics as follows:

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

where $a_{ij}(t)$ is the edge weight for the directed edge $(j, i) \in \mathcal{E}$. We note that if arm k is not chosen, then $X_{i,k}(t) = 0$ for all $i \in \mathcal{V}$ and $t \in \mathcal{T}$. The edge weight is defined as

$$a_{ij}(k) \begin{cases} \geq \underline{a}, & j \in \mathcal{N}_i, \\ = 0, & j \notin \mathcal{N}_i,\ j \neq i, \end{cases} \tag{3}$$

$$a_{ii}(k) = 1 - \sum_{j\in\mathcal{N}_i} a_{ij}(k) \geq \underline{a}, \tag{4}$$

where $\mathcal{N}_i(t) = \{\ell \in \mathcal{V} \mid (\ell, i) \in \mathcal{E}\}$ is the set of the nearby agents and $\underline{a}$ is a positive constant.

Finally, the weights for exploitation are updated by

$$w_{i,k}(t+1) = w_{i,k}(t) e^{\beta \hat{X}_{i,k}(t)}, \tag{5}$$

$$W_i(t+1) = \sum_{k\in\mathcal{K}} w_{i,k}(t+1), \tag{6}$$

where $0 < \beta \leq \alpha/K$ is a learning parameter.

In this paper, we make the stochasticity for the edge weight.

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

3. Regret Analysis

To evaluate the performance of the distributed Exp3 algorithm, we consider the following pseudo-regret for the multi-agent system:

$$\overline{\text{Regret}} = R^* - \sum_{i\in\mathcal{V}} \mathbb{E}\left[\sum_{t\in\mathcal{T}} X_{i,k_i(t)}(t)\right], \tag{7}$$

where $R^* = \max_{k\in\mathcal{K}} \sum_{i\in\mathcal{V}} \mathbb{E}\left[\sum_{t\in\mathcal{T}} X_{i,k}(t)\right]$ is the maximum cumulative reward for continuing to choose the same arm.

The pseudo-regret (7) measures how much reward a group of agents loses by not selecting the arm with the highest expected reward. The upper bound on the pseudo-regret is sublinear if the total regret grows slower than the number of iterations of the algorithm. This is desirable because it means that the agents choose the optimal arm with high probability as the iteration goes on [7], [8]. Therefore, the purpose of the multi-agent system is to search the optimal arm by achieving a sublinear regret bound.

The next result evaluates the upper bound of the pseudo-regret by the distributed Exp3 algorithm.

Theorem 1: The upper bound of the pseudo-regret by the distributed Exp3 algorithm is given by

$$\overline{\text{Regret}} \leq (\alpha + \beta K) R^* + \frac{1-\alpha}{\beta} N \ln K. \tag{8}$$

Proof : From (1) and (6), for all $t \in \mathcal{T}$, we have

$$\sum_{k\in\mathcal{K}} p_{i,k}(t) = (1-\alpha)\frac{\sum_{k\in\mathcal{K}} w_{i,k}(t)}{W_i(t)} + \sum_{k\in\mathcal{K}} \frac{\alpha}{K} = 1.$$

Moreover, from (1), the probability of choosing arm k is lower bounded by

$$p_{i,k}(t) \geq \frac{\alpha}{K} > 0. \tag{9}$$

From (5) and (6), we have

$$W_i(T) = \sum_{k\in\mathcal{K}} w_{i,k}(T) = \sum_{k\in\mathcal{K}} w_{i,k}(T-1) e^{\beta \hat{X}_{i,k}(T-1)}. \tag{10}$$

From (2), we also have

$$
\begin{aligned}
0 &\le \beta \hat{X}_{i,k}(t) \\
&\le \beta \frac{\sum_{j\in\mathcal{V}} a_{ij}(t) X_{j,k}(t)}{p_{i,k}(t)} \\
&\le \beta \frac{\sum_{j\in\mathcal{V}} a_{ij}(t)}{\frac{\alpha}{K}} \le \beta \frac{K}{\alpha} \le 1, \qquad (11)
\end{aligned}
$$

where the third inequality follows from (9) and $X_{i,k}(t) \le 1$, and the forth inequality follows from Assumption 1.

We note that $e^x \le 1 + x + x^2$ holds for any $x \in [0,1]$. Then, from (10) and (11), we have

$$
\begin{aligned}
W_i(T) &= \sum_{k\in\mathcal{K}} w_{i,k}(T-1) e^{\beta \hat{X}_{i,k}(T-1)} \\
&\le \sum_{k\in\mathcal{K}} w_{i,k}(T-1)(1 + \beta \hat{X}_{i,k}(T-1) \\
&\quad + (\beta \hat{X}_{i,k}(T-1))^2) \\
&= W_i(T-1) \\
&\quad \times \left(1 + \beta \sum_{k\in\mathcal{K}} \frac{w_{i,k}(T-1)}{W_i(T-1)} \hat{X}_{i,k}(T-1) \right. \\
&\quad \left. + \beta^2 \sum_{k\in\mathcal{K}} \frac{w_{i,k}(T-1)}{W_i(T-1)} \hat{X}_{i,k}(T-1)^2 \right). \qquad (12)
\end{aligned}
$$

From (1), we have

$$
p_{i,k}(T-1) = (1-\alpha) \frac{w_{i,k}(T-1)}{W_i(T-1)} + \frac{\alpha}{K}.
$$

Thus, we have

$$
\begin{aligned}
\frac{w_{i,k}(T-1)}{W_i(T-1)} &= \frac{1}{1-\alpha} p_{i,k}(T-1) - \frac{\alpha}{(1-\alpha)K} \\
&\le \frac{1}{1-\alpha} p_{i,k}(T-1). \qquad (13)
\end{aligned}
$$

From (13), we obtain

$$
\begin{aligned}
&\beta \sum_{k\in\mathcal{K}} \frac{w_{i,k}(T-1)}{W_i(T-1)} \hat{X}_{i,k}(T-1) \\
&\le \frac{\beta}{1-\alpha} \sum_{k\in\mathcal{K}} p_{i,k}(T-1) \hat{X}_{i,k}(T-1) \\
&= \frac{\beta}{1-\alpha} p_{i,k_i(T-1)}(T-1) \hat{X}_{i,k_i(T-1)}(T-1) \\
&\le \frac{\beta}{1-\alpha} p_{i,k_i(T-1)}(T-1) \\
&\quad \times \frac{\sum_{j\in\mathcal{V}} a_{ij}(T-1) X_{j,k_i(T-1)}(T-1)}{p_{i,k_i(T-1)}(T-1)} \\
&= \frac{\beta}{1-\alpha} \sum_{j\in\mathcal{V}} a_{ij}(T-1) X_{j,k_i(T-1)}(T-1), \qquad (14)
\end{aligned}
$$

where the second inequality follows from (2).

From (13), we also have

$$
\begin{aligned}
&\beta^2 \sum_{k\in\mathcal{K}} \frac{w_{i,k}(T-1)}{W_i(T-1)} \hat{X}_{i,k}(T-1)^2 \\
&\le \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} p_{i,k}(T-1) \hat{X}_{i,k}(T-1)^2 \\
&\le \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} p_{i,k}(T-1) \hat{X}_{i,k}(T-1) \hat{X}_{i,k}(T-1) \\
&\le \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} p_{i,k}(T-1) \\
&\quad \times \frac{\sum_{j\in\mathcal{V}} a_{ij}(T-1) X_{j,k}(T-1)}{p_{i,k}(T-1)} \hat{X}_{i,k}(T-1) \\
&= \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} \left(\sum_{j\in\mathcal{V}} a_{ij}(T-1) X_{j,k}(T-1) \right) \\
&\quad \times \hat{X}_{i,k}(T-1) \\
&\le \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} \left(\sum_{j\in\mathcal{V}} a_{ij}(T-1) \right) \hat{X}_{i,k}(T-1), \\
&= \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} \hat{X}_{i,k}(T-1), \qquad (15)
\end{aligned}
$$

where the last equality follows from Assumption 1.

Substituting (14) and (15) for (12) gives

$$
\begin{aligned}
&W_i(T) \\
&\le W_i(T-1) \\
&\quad \times \left(1 + \frac{\beta}{1-\alpha} \sum_{j\in\mathcal{V}} a_{ij}(T-1) X_{j,k_i(T-1)}(T-1) \right. \\
&\quad \left. + \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} \hat{X}_{i,k}(T-1) \right). \qquad (16)
\end{aligned}
$$

This yields

$$
\begin{aligned}
&W_i(T) \\
&\le W_i(1) \prod_{t=1}^{T-1} \left(1 + \frac{\beta}{1-\alpha} \sum_{j\in\mathcal{V}} a_{ij}(t) X_{j,k_i(t)}(t) \right. \\
&\quad \left. + \frac{\beta^2}{1-\alpha} \sum_{k\in\mathcal{K}} \hat{X}_{i,k}(t) \right). \qquad (17)
\end{aligned}
$$

From (5) and (6), for any arm $k \in \mathcal{K}$, we have

$$
\begin{aligned}
W_i(T) &= \sum_{\ell=1}^{K} w_{i,\ell}(T) \\
&\ge w_{i,k}(T) \\
&= w_{i,k}(T-1) e^{\beta \hat{X}_{i,k}(T-1)} \\
&= w_{i,k}(1) e^{\beta \sum_{t\in\mathcal{T}} \hat{X}_{i,k}(t)}. \qquad (18)
\end{aligned}
$$

From (17) and (18), we have

$$
w_{i,k}(1) e^{\beta \sum_{t\in\mathcal{T}} \hat{X}_{i,k}(t)}
$$

$$\leq W_i(1)\prod_{t=1}^{T-1}\left(1+\frac{\beta}{1-\alpha}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)+\frac{\beta^2}{1-\alpha}\sum_{k\in\mathcal{K}}\hat{X}_{i,k}(t)\right). \tag{19}$$

By taking the natural logarithm for (19) and using the initialization of $w_{i,k}(1) = 1/K^\nu$, we have

$$-\ln K+\beta\sum_{t\in\mathcal{T}}\hat{X}_{i,k}(t) \leq \sum_{t\in\mathcal{T}}\ln\left(1+\frac{\beta}{1-\alpha}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)+\frac{\beta^2}{1-\alpha}\sum_{k\in\mathcal{K}}\hat{X}_{i,k}(t)\right).$$

We note that $\ln(1+x)\leq x$ holds for any $x\geq 0$. Then, we have

$$-\ln K+\beta\sum_{t\in\mathcal{T}}\hat{X}_{i,k}(t) \leq \frac{\beta}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)+\frac{\beta^2}{1-\alpha}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\hat{X}_{i,k}(t).$$

Since $\hat{X}_{i,k}(t)$ is the unbiased estimator of $X_{i,k}(t)$, by taking the expectation with respect to the estimated distribution of the rewards obtained by the distributed Exp3 algorithm, we have

$$-\ln K+\beta\sum_{t\in\mathcal{T}}X_{i,k}(t) \leq \frac{\beta}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)+\frac{\beta^2}{1-\alpha}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}X_{i,k}(t).$$

Furthermore, by taking the expectation with respect to the true distribution of the rewards, we have

$$-\ln K+\beta\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right] \leq \frac{\beta}{1-\alpha}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)\right]+\frac{\beta^2}{1-\alpha}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right].$$

Then, we have

$$-N\ln K+\beta\sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right] \leq \frac{\beta}{1-\alpha}\sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)\right]+\frac{\beta^2}{1-\alpha}\sum_{i\in\mathcal{V}}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right].$$

Therefore, we have

$$-N\ln K+\beta\sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right] \leq \frac{\beta}{1-\alpha}\sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)\right]+\frac{\beta^2K}{1-\alpha}R^*, \tag{20}$$

where the last inequality follows from

$$R^*\geq\frac{1}{K}\sum_{i\in\mathcal{V}}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k}(t)\right].$$

We note that (20) holds for any $k\in\mathcal{K}$. Thus, we have

$$-N\ln K+\beta R^* \leq \frac{\beta}{1-\alpha}\sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)\right]+\frac{\beta^2K}{1-\alpha}R^*.$$

It follows that

$$-\frac{1-\alpha}{\beta}N\ln K+(1-\alpha)R^* \leq \sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{V}}a_{ij}(t)X_{j,k_i(t)}(t)\right]+\beta KR^* = \sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}}X_{i,k_i(t)}(t)\right]+\beta KR^*,$$

where the last equality follows from the unconstrained reward model, and (3) and (4). This concludes the proof. □

Theorem 1 holds even for the case when the connectivity of the communication network is not guaranteed. However, to achieve a better regret bound, sharing the estimated information between agents is crucial; hence, uniform connectedness plays an important role. Investigating the relation between the connectedness of the communication graph and the regret bound is future research of this paper.

The next proposition shows that a sublinear regret can be obtained if the information on the upper bound of the accumulated reward is obtained in advance.

Proposition 1: Suppose that the trade-off parameter and the learning parameter are given as $\alpha = \beta K$ and $\beta = \min\{c/K, \sqrt{N\ln K/(2RK)}\}$, where $0 < c < 1$ and $R^*\leq R$. If each agent updates the estimation of the rewards by the distributed Exp3 algorithm, we have

$$\overline{\text{Regret}} \le \frac{2}{c}\sqrt{2NRK\ln K}. \tag{21}$$

Proof : We consider the case for $\sqrt{N\ln K/(2RK)} \ge c/K$. Then, we have $2R \le N(K\ln K)/c^2$ holds. This yields

$$\begin{aligned}\overline{\text{Regret}} &= R^* - \sum_{i\in\mathcal{V}}\mathbb{E}\left[\sum_{t\in\mathcal{T}} X_{i,k_i(t)}(t)\right] \\ &\le R^* \le 4R = 2\sqrt{2R}\sqrt{2R} \le \frac{2}{c}\sqrt{2NRK\ln K}. \end{aligned}\tag{22}$$

Next, we consider the case for $\sqrt{N\ln K/(2RK)} < c/K$. In this case, $\beta = \sqrt{N\ln K/(2RK)}$ holds. Moreover, we have $\beta < c/K$. It follows that $1-\alpha = 1-\beta K > 0$. Thus, from (8), we have

$$\begin{aligned}\overline{\text{Regret}} &\le (\alpha+\beta K)R^* + \frac{N}{\beta}\ln K \\ &\le 2\beta K R^* + N\frac{1}{\sqrt{N}}\sqrt{\frac{2RK}{\ln K}}\ln K \\ &\le 2\sqrt{2NRK\ln K}.\end{aligned}$$

□

For a single agent system, the regret bound is given by $(e-1)\alpha R^* + (1/\alpha)K\ln K$ for the case with $\alpha = \beta$ [6]. Thus, by the analysis of [6], the regret is upper-bounded by $N((e-1)\alpha R^* + (1/\alpha)K\ln K)$ for the mutliagent system with N agents. Proposition 1 implies that the tighter regret bound can be obtained if the trade-off and learning parameters are properly set.

Compared with the existing cooperative methods [21]–[24], in the proposed method, the condition of the communication topology is relaxed to time-varying directed networks. However, the regret bound of the proposed algorithm is inferior to those of other methods at the cost of extending the applicable class. For example, in the Exp3-Coop algorithm [21], the number of agents N affects the regret bound on the order of the square root of its reciprocal when the communication graph is fixed and undirected. This regret bound is more preferable for multiagent systems with the larger number of agents. Further theoretical analysis of the proposed algorithm for large-scale networks is a future direction of this study.

4. Numerical Experiments

We consider a cooperative adversarial multi-armed bandit problem. Arm 1 is the best arm whose reward 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.

We evaluate the effectiveness of the proposed algorithm across different values of the trade-off parameter α. The communication networks at $t = 0$, 1000, 2000, and 3000 are shown in Fig. 1. Figure 2 illustrates the pseudo-regret (7)

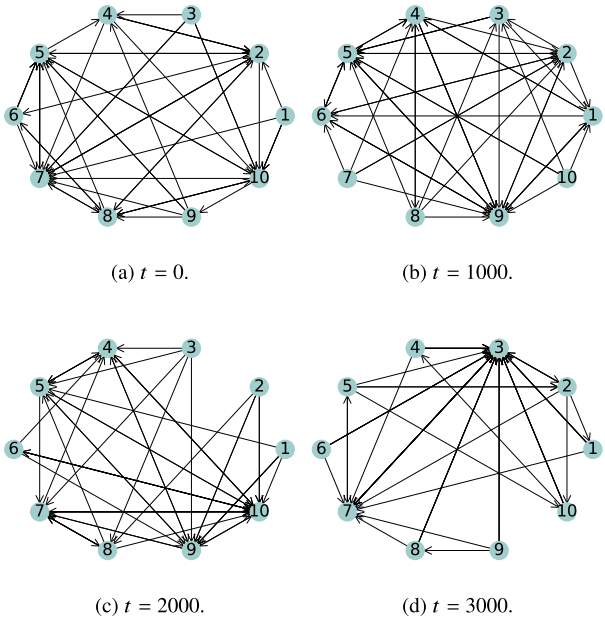

Fig. 1 Communication networks.

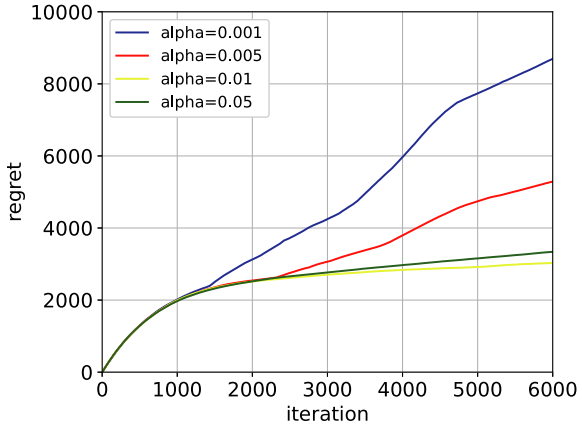

Fig. 2 Pseudo-regret with different values of the trade-off parameter α.

for agent 1. The learning parameter, number of arms, and number of agents are set to $\beta = 0.01$, $K = 10$, and $N = 10$, respectively. We see that the evolution of the regret varies depending on the value of α. When $\alpha = 0.001$ and 0.005, the regret at the initial stage of the iteration remains small but gradually increases over time. The probability of choosing an arm in (1) implies that a small value of α hinders the exploration for better arms, whereas a large value restricts the exploitation of the learned information of the rewards. In this example, a value of $\alpha = 0.01$ achieves a suitable balance between exploitation with the Hedge algorithm and exploration with uniform search.

Next, we examine the impact of different values of the learning parameter β on the convergence performance. We evaluate the pseudo-regret of agent 1 using the proposed algorithm for $\alpha = 0.01$, $K = 10$, and $N = 10$, with varying values of β. Figure 3 shows that the choice of β influences the evolution of the regret. In this example, the case with $\beta =$

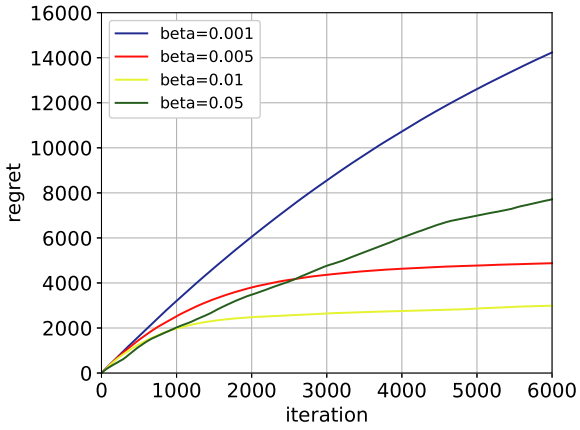

Fig. 3 Pseudo-regret with different values of the learning parameter β.

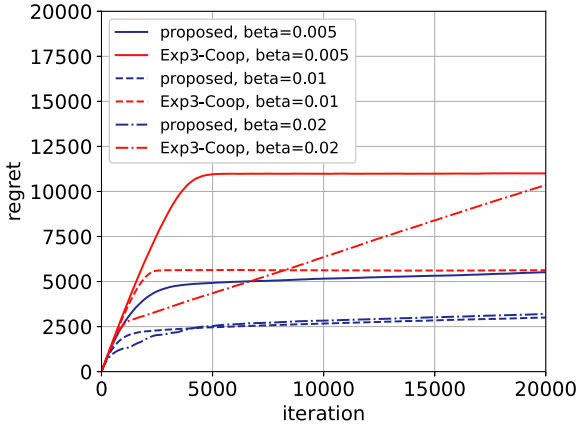

Fig. 4 Performance comparison with the Exp3-Coop Algorithm [21].

0.01 yields better performance. This result shows that the selection of an appropriate value for the learning parameter β is also crucial.

Finally, we consider a comparative performance evaluation between the proposed algorithm and the Exp3-Coop Algorithm [21]. The Exp3-Coop is a distributed variant of the Exp3 algorithm for adversarial multi-armed bandit problems. The number of arms and the number of agents are set as $K = 10$ and $N = 10$. Within the framework of the proposed algorithm, the trade-off parameter is set as $\alpha = 0.01$. It is worth noting that the theoretical analysis of the Exp3-Coop Algorithm in [21] is conducted only for fixed undirected graphs, which is a primal difference from the analysis of this paper. However, to investigate the performance comparison in more general situations, we extended its application to a time-varying directed network in this example. Figure 4 illustrates the pseudo-regret of agent 1 with varying values of β. We observe that the sublinear regret trajectories are achieved for both algorithms with suitable learning parameters. Moreover, in this specific example, we see that the proposed consensus-based Exp3 policy outperformed the Exp3-Coop algorithm regarding regret minimization. For future research, it remains to be clarified in what problem settings, such as the topology of the communication graph and the number of agents, the proposed algorithm performs better.

5. Conclusion

In this paper, we presented a distributed Exp3 algorithm for the adversarial bandit problem on directed and time-varying networks. We demonstrated that the proposed algorithm cooperatively estimates the reward distribution for each arm with nearby agents. We provided an upper bound of the pseudo-regret, which quantifies the difference between the optimal reward and the expected reward. Additionally, we derived a sufficient condition for achieving a sublinear regret bound. The numerical results illustrated that the sublinear regret can be achieved by appropriately tuning the trade-off and learning parameters. As future work, we plan to determine optimal parameter settings and to investigate the impact of communication delays between agents on the adversarial bandit problem.

Acknowledgments

This work was supported in part by Japan Society for the Promotion of Science KAKENHI Grant Number JP21K04121.

References

[1] 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.
[2] 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.
[3] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," Proc. 2020 IEEE Congress on Evolutionary Computation, pp.1–8, 2020.
[4] W. Xia, T.Q.S. Quek, K. Guo, W. Wen, H.H. Yang, and H. Zhu, "Multi-armed bandit-based client scheduling for federated learning," IEEE Trans. Wireless Commun., vol.19, no.11, pp.7108–7123, 2020.
[5] S.U. Minhaj, A. Mahmood, S.F. Abedin, S.A. Hassan, M.T. Bhatti, S.H. Ali, and M. Gidlund, "Intelligent resource allocation in LoRaWAN using machine learning techniques," IEEE Access, vol.11, pp.10092–10106, 2023.
[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire, "The nonstochastic multiarmed bandit problem," SIAM J. Comput., vol.32, no.1, pp.48–77, 2002.
[7] T. Lattimore and C. Szepesvári, Bandit Algorithms, Cambridge University Press, 2019.
[8] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," Foundations and Trends in Machine Learning, vol.5, no.1, pp.1–122, 2012.
[9] A. Nedić and J. Liu, "Distributed optimization for control," Annual Review of Control, Robotics, and Autonomous Systems, vol.1, no.1, pp.77–103, 2018.
[10] K. Ishikawa, N. Hayashi, and S. Takai, "Consensus-based distributed particle swarm optimization with event-triggered communication," IEICE Trans. Fundamentals, vol.E101-A, no.2, pp.338–344, Feb. 2018.
[11] S. Izumi, Y. Shiomoto, and X. Xin, "Mass game simulator: An entertainment application of multiagent control," IEEE Access, vol.9, pp.4129–4140, 2020.
[12] R. Adachi, Y. Yamashita, and K. Kobayashi, "Distributed optimal

estimation with scalable communication cost," IEICE Trans. Fundamentals, vol.E104-A, no.11, pp.1470–1476, Nov. 2021.
[13] M. Yamashita, N. Hayashi, T. Hatanaka, and S. Takai, "Logarithmic regret for distributed online subgradient method over unbalanced directed networks," IEICE Trans. Fundamentals, vol.E104-A, no.8, pp.1019–1026, Aug. 2021.
[14] K. Sakurama and T. Sugie, "Generalized coordination of multi-robot systems," Foundations and Trends in Systems and Control, vol.9, no.1, pp.1–170, 2021.
[15] R. Adachi and Y. Wakasa, "Distributed filter using ADMM for optimal estimation over wireless sensor network," IEICE Trans. Fundamentals, vol.E105-A, no.11, pp.1458–1465, Nov. 2022.
[16] K. Toda, N. Kuze, and T. Ushio, "Stability analysis and control of decision-making of miners in blockchain," IEICE Trans. Fundamentals, vol.E105-A, no.4, pp.682–688, April 2022.
[17] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.2786–2790, 2017.
[18] P. Landgren, V. Srivastava, and N.E. Leonard, "Distributed cooperative decision making in multi-agent multi-armed bandits," Automatica, vol.125, p.109445, 2021.
[19] A. Moradipari, M. Ghavamzadeh, and M. Alizadeh, "Collaborative multi-agent stochastic linear bandits," Proc. 2022 American Control Conference, pp.2761–2766, 2022.
[20] J. Zhu and J. Liu, "Distributed multi-armed bandits," IEEE Trans. Autom. Control, vol.68, no.5, pp.3025–3040, 2023.
[21] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," Proc. 29th Annual Conference on Learning Theory, vol.49, pp.605–622, 2016.
[22] Y. Bar-On and Y. Mansour, "Individual regret in cooperative nonstochastic multi-armed bandits," Advances in Neural Information Processing Systems, vol.32, 2019.
[23] 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.
[24] J. Yi and M. Vojnović, "On regret-optimal cooperative nonstochastic multi-armed bandits," CoRR, vol.abs/2211.17154, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.17154
[25] R. Olfati-Saber, J.A. Fax, and R.M. Murray, "Consensus and cooperation in networked multi-agent systems," Proc. IEEE, vol.95, no.1, pp.215–233, 2007.
[26] N. Hayashi and S. Takai, "A GTS scheduling for consensus problems over IEEE 802.15.4 wireless networks," Proc. 2013 European Control Conference, pp.1764–1769, 2013.
[27] K. Kobayashi, "Predictive pinning control with communication delays for consensus of multi-agent systems," IEICE Trans. Fundamentals, vol.E102-A, no.2, pp.359–364, Feb. 2019.
[28] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," IEEE Trans. Autom. Control, vol.60, no.3, pp.601–615, 2015.

Tomoki Nakamura received the B.E. and M.E. from Osaka University in 2022 and 2024, respectively. His research interest includes distributed learning and optimization.

Naoki Hayashi received the B.E., M.E., and Ph.D. degrees from Osaka University in 2006, 2008, and 2011, respectively. He was a Research Assistant at Kyoto University in 2011. From 2012 to 2020, he was an Assistant Professor at Osaka University. He is currently an Associate Professor at Osaka University. His research interests include cooperative control and distributed optimization. He is a member of ISCIE, SICE, and IEEE.

Masahiro Inuiguchi received B.E., M.E. and D.E. degrees at Osaka Prefecture University, in 1985, 1987 and 1991, respectively. He worked as a Research Associate at Osaka Prefecture University (1987–1992), Associate Professor at Hiroshima University (1992–1997), Associate Professor at Osaka University (1997–2003). At present, he is a Full Professor at Osaka University. His research interests include possibility theory, fuzzy mathematical programming, rough sets and multiple criteria decision analysis with interval models. He is a member of SOFT, ISCIE, SICE, JORS, IEICE, IEEE, IRSS, and INFORMS. He works as an area/regional editor of Fuzzy Sets and Systems and Fuzzy Optimization and Decision Making and members of editorial boards of many other journals including European Journal of Operational Research.