Decentralized Asynchronous Multi-player Bandits

Extended Abstract

Jingqi Fan
Northeastern University, China
Shenyang, China
fanjingqi@stumail.neu.edu.cn

Canzhe Zhao
Shanghai Jiao Tong University
Shanghai, China
canzhezhao@sjtu.edu.cn

Shuai Li
Shanghai Jiao Tong University
Shanghai, China
shuaili8@sjtu.edu.cn

Siwei Wang*
Microsoft Research Asia
Beijing, China
siweiwang@microsoft.com

ABSTRACT

In recent years, multi-player multi-armed bandits (MP-MAB) have been extensively studied due to their wide applications in cognitive radio networks and Internet of Things systems. Most existing works focus on synchronized settings, whereas real-world systems are often decentralized and asynchronous, with players entering and leaving arbitrarily and no shared global clock. This introduces two major challenges: avoiding collisions without time coordination, and estimating the number of active players in every step. In this paper, we propose an algorithm to address these challenges. During exploration, players uniformly explore the arms that are not currently exploited by others, which reduces the probability of collisions and solves the first challenge. Meanwhile, players occasionally pull the arms that are currently exploited by others, enabling them to detect other players' departures and addressing the second challenge. We prove that our algorithm achieves a regret of $O(\sqrt{T \log T} + \log T/\Delta^2)$, where Δ is the minimum expected reward gap between any two arms. To the best of our knowledge, this is the first efficient algorithm in the asynchronous and decentralized environment.

KEYWORDS

Multi-armed Bandits, Multi-agent Systems, Asynchronous Coordination, Decentralized Learning

ACM Reference Format:
Jingqi Fan, Canzhe Zhao, Shuai Li, and Siwei Wang. 2026. Decentralized Asynchronous Multi-player Bandits: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/BYGO6394

1 INTRODUCTION

Multi-armed bandit (MAB) is a well-established model with broad applications in online advertising and recommendation systems [1]. However, classical bandit models consider only a single player, whereas many practical systems, such as cognitive radio networks and Internet-of-Things environments [10], involve multiple players competing for limited resources. This naturally gives rise to the multi-player multi-armed bandit (MP-MAB) problem. In this setting, M players simultaneously select arms from $[K]$, and collisions occur when multiple players choose the same arm, resulting in zero rewards. Compared with the single-player setting, MP-MAB introduces additional challenges, as players need to coordinate with others while still dealing with uncertainty in reward distributions. For example, many existing approaches in decentralized MP-MAB deliberately introduce collisions as an *implicit communication* mechanism to approximate the performance of the centralized setting [3, 4, 9]. These methods typically assume a synchronous environment, where all players enter the system simultaneously and remain active throughout. In contrast, real-world applications often involve inherently asynchronous systems [5].

In this paper, we consider a decentralized asynchronous setting in which players are unaware of the global clock and may join or leave the system at arbitrary times. The unpredictable access patterns in the decentralized environment introduce **two major challenges**: **(i)** The absence of a global clock makes implicit communication through collisions unreliable, as new players may join at arbitrary times and unintentionally collide with existing players, leading to frequent and uncontrolled collisions. **(ii)** The dynamic nature of player participation makes it important to detect the number of current active players. If the number is overestimated, then the player may exploit an arm that is not good enough, leading to unacceptable regret.

To address the above challenges, we propose a novel algorithm called **Adaptive Change between Exploration and Exploitation (ACE)**. ACE enables each player j to maintain an estimated arm set, which contains the arms believed to be currently exploited by other players. By dynamically updating the arm set based on observed collisions, players reduce competition and move to better arms when they become available. Our analysis shows that ACE achieves a regret upper bound of $O(\sqrt{T \log T} + \log T/\Delta^2)$.

2 METHOD

2.1 Preliminaries

We consider a T-step decentralized asynchronous multi-player multi-armed bandit problem with K arms and M players. Each player $j \in [M]$ joins the system at time step T^j_{start} and leaves at

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/BYGO6394

time step T^j_{end}. Players can not observe T^j_{start}, T^j_{end} and the actual time step t. At each discrete time step $T^j_{\text{start}} \le t \le T^j_{\text{end}}$, player j selects an arm $\pi^j(t) \in [K]$ to pull (for $t < T^j_{\text{start}}$ or $t > T^j_{\text{end}}$, we let $\pi^j(t) = 0$). If more than one players choose arm k at t, then there is a collision, and $\eta_k(t) := \mathbb{1}[\sum_{j \le M} \mathbb{1}[\pi^j(t) = k] > 1]$ denotes the collision indicator. For player j, her observation at step t contains two values, $\eta^j(t) = \eta_{\pi^j(t)}(t)$ tells her whether there is a collision, and $r^j(t) := (1-\eta_{\pi^j(t)}(t))X_{\pi^j(t)}(t)$ is her reward in this step. Here $X_{\pi^j(t)}(t)$ is assumed to be a 1-subgaussian random variable with expectation $\mu_{\pi^j(t)} \in [0,1]$. Without loss of generality, we assume that $\mu_1 > \mu_2 > \cdots > \mu_K$ [6, 7, 9].

The goal of the players is to minimize the regret defined as

$$R(T) := \sum_{t=1}^{T} \sum_{k \le m_t} \mu_k - \mathbb{E}\left[\sum_{t=1}^{T} \sum_{j: T^j_{\text{start}} \le t \le T^j_{\text{end}}} r^j(t)\right],$$

where m_t denotes the number of active players at time t. We assume that there exists a constant m such that $m_t \le m \le K/2, \forall t \le T$.

2.2 Algorithm

In this section, we propose our **Adaptive Change between Exploration and Exploitation (ACE)** algorithm. Let $\mathcal{A}^j$ denote the set of arms believed by player j to be currently exploited by other players, and let ε denote the probability of maintaining exploration within $\mathcal{A}^j$. We maintain two queues $\mathcal{P}^j_k$ and $\mathcal{Q}^j_k$ of fixed lengths $L_p = \lceil 866 \ln T \rceil$ and $L_q = \lceil 570 \ln T \rceil$ for each player j and arm k. Also define $\hat{\mu}^j_k(t) := \sum_{t'=1}^{t} r^j_k(t') \mathbb{1}\{\eta_k(t') = 0\}/N^j_k(t)$ and $N^j_k(t) := \sum_{t'=1}^{t} \mathbb{1}\{\pi^j(t') = k,\ \eta_k(t') = 0\}$.

In the exploration phase, player j randomly explores arms in $[K] \setminus \mathcal{A}^j$ with probability $1-\varepsilon$. At the same time, she adds the collision indicators into $\mathcal{P}^j_k$. If there exists an arm $k \in [K] \setminus \mathcal{A}^j$ such that $\sum_{i \in \mathcal{P}^j_k} i \ge \lceil 0.85 L_p \rceil$, player j adds k into $\mathcal{A}^j$ and thus will not explore it in future exploration. Players update the upper and lower confidence bounds as

$$\text{UCB}^j_k(t) := \hat{\mu}^j_k(t) + \sqrt{\frac{6 \log T}{N^j_k(t)}}, \quad \text{LCB}^j_k(t) := \hat{\mu}^j_k(t) - \sqrt{\frac{6 \log T}{N^j_k(t)}}.$$

When detecting an arm $k [\in K] \setminus \mathcal{A}^j$ such that $\text{LCB}^j_k(t) \ge \text{UCB}^j_\ell(t)$, player j starts an exploitation phase, where she pulls the selected arm with probability $1-\varepsilon$. On the other hand, player j constantly pulls arms in $\mathcal{A}^j$ with probability ε, and adds collision indicators into $\mathcal{Q}^j_k$. When there is an arm k such that $\sum_{i \in \mathcal{Q}^j_k} i \ge \lceil 0.142 L_q \rceil$, she removes the arm from $\mathcal{A}^j$. Note that if player j finds that an arm is removed during the exploitation phase, she will switch back to the exploration phase. In this way, player j can detect other players' departures and update their exploitation choices accordingly with limited cost.

3 THEORETICAL ANALYSIS

Lemma 3.1 guarantees the correctness of adding and removing arms from $\mathcal{A}^j$, while providing an expected upper bound on a single add (remove) operation.

Lemma 3.1. *With probability at least $1 - 4MK/T$, for any player j and arm k,*

(i) if arm k is occupied and remains occupied thereafter, player j will add k to $\mathcal{A}^j(t)$ within $O(K \ln T)$ steps in expectation; If arm k is not occupied and remains not occupied thereafter, player j will not add k to $\mathcal{A}^j(t)$;

(ii) if arm k is released and never occupied again, then player j will remove k from $\mathcal{A}^j(t)$ within $O(m \ln T/\varepsilon)$ steps; If arm k is not released and remains not released thereafter, player j will not remove k from $\mathcal{A}^j(t)$.

The following theorem presents the regret upper bound of ACE.

Theorem 3.2. *Let $\varepsilon = \min\{\sqrt{\frac{1141 m^3 \ln(T)}{2T}}, \frac{1}{K}, \frac{1}{10}\}$. Then, given K arms and M players, the regret of ACE is upper bounded by*

$$R(T) \le O\left(\frac{emKM \log(T)}{\Delta^2} + m^{3/2} M \sqrt{T \ln(T)} + m^2 KM \ln(T)\right),$$

where $\Delta := \min_{k \le m}(\mu_k - \mu_{k+1})$.

The first $O(\log T/\Delta^2)$ term arises from Challenge (i) presented in Section 1, due to the unavoidable collisions caused by uniform exploration, resulting in a dependence on $1/\Delta^2$ rather than the standard $1/\Delta$. The $O(\sqrt{T \log T})$ term corresponds to Challenge (ii), as players must maintain exploration in $\mathcal{A}^j$ with probability ε to detect changes in availability.

4 EXPERIMENTS

We conduct experiments with 20 arms and 10 players in a random asynchronous setting. The reward of each arm k follows a Gaussian distribution $\mathcal{N}(\mu_k, 0.5^2)$, where the smallest mean μ_K is 0.1 and the gap between adjacent arms is fixed at 0.05. UCB(c) denotes the UCB algorithm with different parameters in the logarithmic term [2]. RD-UCB(c) adds random noise to the index [8]. In Figure 1, the regret of UCB and RD-UCB begins to increase rapidly once a certain point is reached. This behavior arises because UCB indices become relatively stable after players do many explorations. When some players depart and the optimal arms change, these indices adapt slowly, causing players to continue selecting previously favored arms while neglecting newly released optimal ones. In contrast, ACE allows players to update $\mathcal{A}^j$ dynamically and detect released arms quickly, leading to convergence after a brief growth phase.

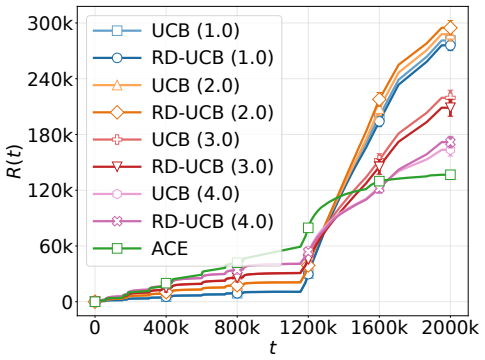

Figure 1: Comparison of cumulative regret.

REFERENCES

[1] P Auer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem.

[2] Lilian Besson and Emilie Kaufmann. 2018. Multi-player bandits revisited. In *Algorithmic Learning Theory*. PMLR, 56–92.

[3] Etienne Boursier and Vianney Perchet. 2019. SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems* 32 (2019).

[4] Wei Huang, Richard Combes, and Cindy Trinh. 2022. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Conference on Learning Theory*. PMLR, 1990–2012.

[5] Ying-Chang Liang, Kwang-Cheng Chen, Geoffrey Ye Li, and Petri Mahonen. 2011. Cognitive radio networking and communications: An overview. *IEEE transactions on vehicular technology* 60, 7 (2011), 3386–3407.

[6] Shivakumar Mahesh, Anshuka Rangi, Haifeng Xu, and Long Tran-Thanh. 2022. Multi-Player Bandits Robust to Adversarial Collisions. *arXiv e-prints* (2022), arXiv–2211.

[7] Shivakumar Mahesh, Anshuka Rangi, Haifeng Xu, and Long Tran-Thanh. 2024. Attacking multi-player bandits and how to robustify them. In *Proceedings of 23rd Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*. ACM; International Foundation for Autonomous Agents and Multiagent Systems ....

[8] Cindy Trinh and Richard Combes. 2021. A High Performance, Low Complexity Algorithm for Multi-Player Bandits Without Collision Sensing Information. *arXiv preprint arXiv:2102.10200* (2021).

[9] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. 2020. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4120–4129.

[10] Alexander M Wyglinski, Maziar Nekovee, and Thomas Hou. 2009. *Cognitive radio communications and networks: principles and practice*. Academic Press.