Decentralized Asynchronous Multi-player Bandits

Paper ID: 81

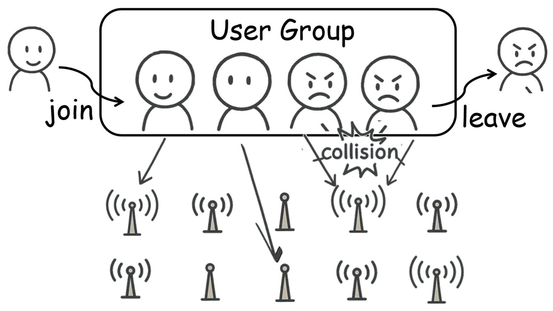
Jingqi Fan[1], Canzhe Zhao[2], Shuai Li[2], Siwei Wang[3]

1 Northeastern University, China 2 Shanghai Jiao Tong University

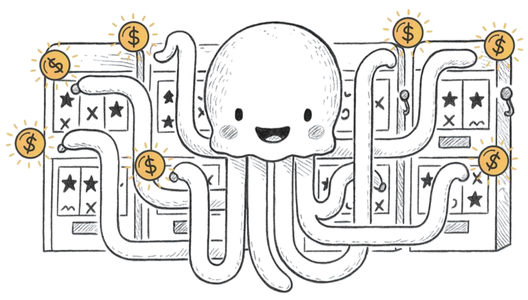
3 Microsoft Research Asia

Motivation

In cognitive radio networks, users join and leave dynamically.



Multi-armed Bandits is a classical decision-making framework.



Setting

Problem Formulation:

- M players, K arms, T total steps.
- Full decentralized environment, i.e., players can not observe others.
- Let $[M] := \{1, ..., M\}$ and $[K] := \{1, ..., K\}$.
- Let $1 \le T^j_{\text{start}} < T^j_{\text{end}} \le T$. A player is active at step t means that she needs to pull an arm at this step. Let m_t denote the number of active players at step t.
- T^j_{start} and T^j_{end} are chosen arbitrarily.
- Each player $j \in [M]$ is only active from T^j_{start} to T^j_{end}.
- Player j is only aware of T, but does not know T^j_{start} and T^j_{end}.
- At each step $t \in [T^j_{\text{start}}, T^j_{\text{end}}]$, player j pulls an arm $\pi^j(t) \in [K]$.
- She observes $< r^j(t), \eta^j(t) >$, where
  1. $r^j(t) := X^j(t)[1 - \eta^j(t)]$ is a reward, and $X^j(t) \sim \text{Bernoulli}(\mu_{\pi^j(t)})$;
  2. $\eta^j(t) := \mathbb{1}\left[\exists j' \neq j, j' \in [M] : \pi^j(t) = \pi^{j'}(t)\right]$ is a collision indicator.

Assumption:

- There exists a constant m such that for any t, $m_t \le m \le K/2$.

Regret Definition:

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

where μ_k is the k-th biggest reward expectation. $\mu_1 > \mu_2 > \cdots > \mu_K$.

Contribution

	Environment	Com	Async setting	Regret bound
Boursier and Perchet [2019]	Decentralized	No	Players arrive at different times but never leave.	$\mathcal{O}\left(\frac{KM\log T}{\Delta^2_{(1)}} + \frac{KM^2\log T}{\mu_M}\right)$
Dakdouk [2022]	Decentralized	Yes	Activation probability p	$\mathcal{O}\left(\max\left\{K^2, \frac{\log(KT)}{Mp(1-p/K)^M}\right\} T^{2/3}\right)$
Richard et al. [2024]	Centralized	Yes	Known activation probability p	$\mathcal{O}\left(\sqrt{KT\log(KT)\min\{K, Mp\}}\right)$
Richard et al. [2024]	Centralized	Yes	Known activation probability p	$\mathcal{O}\left(\frac{(K^2+(1+p)M^2)\log(KT)}{\Delta_{(2)}}\right)$
ACE	Decentralized	No	Players arrive and leave arbitrarily over time.	$\mathcal{O}\left(m^{3/2}M\sqrt{T\ln T} + \frac{mKM\log T}{\Delta^2_{(3)}}\right)$

- "Com" column indicates whether direct communication (rather than via collision) is allowed.
- $\Delta_{(1)}$ to $\Delta_{(3)}$ are different reward gaps.
- Our setting is more general and the assumption is mild.

Algorithm

Difficulty

Players do not know when others join the system. → **1. Unavoidable collisions:** Previous communication phase does not work. A player can join at any time and break the communication, leading to frequent collisions.

Players do not know when others leave the system. → **2. Change of optimal arms:** When a player who is exploiting her optimal arm leaves the system, the left arms that are still exploited by players may become sub-optimal.

Solve Difficulty 1

- There is no `Communication` phase; each player independently executes her own policy.
- Player j maintains a set $\mathcal{A}^j$, representing the arms believed to be occupied by other players.
- Player j explores arms in $[K] \setminus \mathcal{A}^j$ uniformly at random.
- If arms in $[K] \setminus \mathcal{A}^j$ frequently result in collisions, player j infers that those arms are likely being occupied (exploited) by others and adds them to $\mathcal{A}^j$.

Solve Difficulty 2

- Player j always pulls arms in $\mathcal{A}^j$ with a small probability ε.
- If arms in $\mathcal{A}^j$ frequently result in non-collisions, player j infers that those arms are likely being released by others and removes them from $\mathcal{A}^j$.

Proposed Algorithm: ACE

Player j Adaptively Changes between an Exploration phase and an Exploitation phase:

- **Exploration phase:** If there exists an arm k such that $\text{LCB}^j_k \ge \text{UCB}^j_\ell$ for all $\ell \neq k$, $\ell \in [K] \setminus \mathcal{A}^j$, then player j transitions to the exploitation phase and pulls arm k with probability $1 - \varepsilon$.
- **Exploitation phase:** If player j detects that an arm in $\mathcal{A}^j$ has been released, she switches back to the exploration phase.

Analysis

Theorem 1. Given K arms and M players, and let $\varepsilon = \min\{\sqrt{\frac{1141m^3\ln(T)}{2T}}, \frac{1}{K}, \frac{1}{10}\}$, the regret of Algorithm 1 is bounded by

$$\mathbb{E}[R(T)] \le \frac{576emKM\log(T)}{\Delta^2} + 96m^{3/2}M\sqrt{T\ln(T)} + 7704m^2KM\ln(T) + (4emKM)^2,$$

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

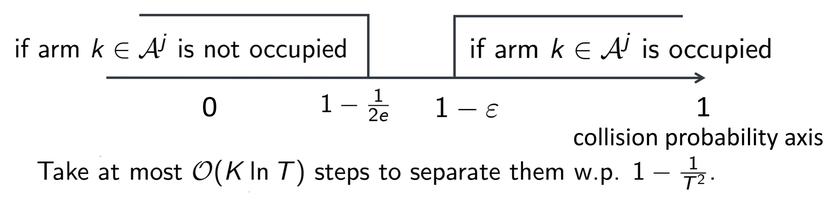
$\mathcal{O}(\log T/\Delta^2)$ arises from Difficulty 1:
Players cannot completely avoid collisions, leading to a regret of $\mathcal{O}(\log T/\Delta^2)$ instead of the standard $\mathcal{O}(\log T/\Delta)$.

$\mathcal{O}(\sqrt{T\log T})$ incurs from Difficulty 2:
The set of optimal arms may change over time, so players must pull occupied arms with a small probability. This persistent exploration contributes a regret of $\mathcal{O}(\sqrt{T\log T})$.

Proof Sketch: Distingush Events via Collison Probablity

Let $k \in \mathcal{A}^j$. player j pulls arm k. Then she receives a collision or non-collision.

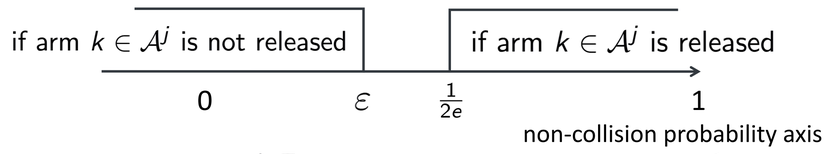
For the adding process:



Take at most $\mathcal{O}(K\ln T)$ steps to separate them w.p. $1 - \frac{1}{T^2}$.

Intuition:
Arm k is occupied.
A player is exploiting it.
The collision probability ↑.

For the Removing process:



Take at most $\mathcal{O}(\frac{m\ln T}{\varepsilon})$ steps to separate them w.p. $1 - \frac{1}{T^2}$.

Arm k is released.
No player is exploiting it.
The collision probability ↓.
The non-collision probability ↑.

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

the first mt optimal arms' expectation — active players' rewards (definition)

$$\le \sum_{t=1}^{T}\left(m_t - \mathbb{E}\left[\sum_{j: T^j_{\text{start}} \le t \le T^j_{\text{end}}} \mathbb{1}[\pi^j(t) \le m_t, \eta^j(t) = 0]\right]\right)$$

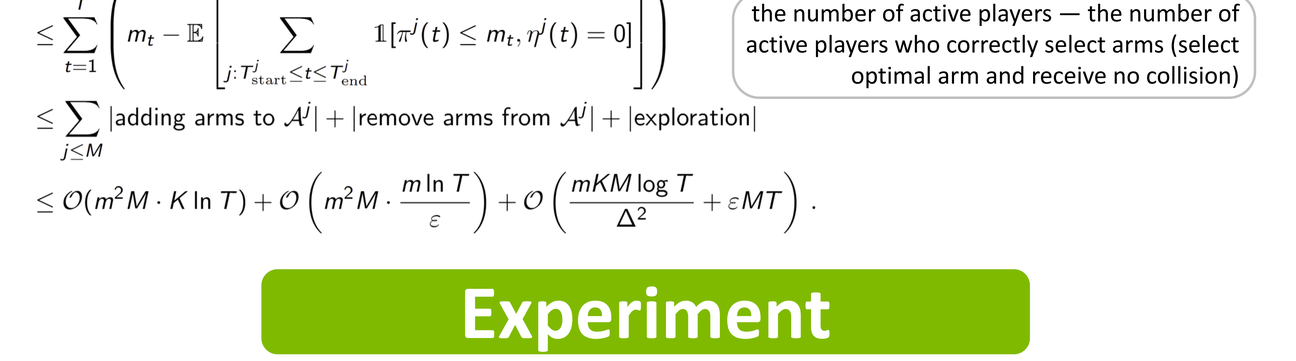
the number of active players — the number of active players who correctly select arms (select optimal arm and receive no collision)

$$\le \sum_{j \le M} |\text{adding arms to } \mathcal{A}^j| + |\text{remove arms from } \mathcal{A}^j| + |\text{exploration}|$$

$$\le \mathcal{O}(m^2M \cdot K\ln T) + \mathcal{O}\left(m^2M \cdot \frac{m\ln T}{\varepsilon}\right) + \mathcal{O}\left(\frac{mKM\log T}{\Delta^2} + \varepsilon MT\right).$$

Experiment

1. K = 20. 2. K = 50. 3. K = 100.



1. K = 20, with UCBs. 2. K = 50, with UCBs. 3. K = 100, with UCBs.

Reference:

Boursier, E. and Perchet, V., 2019. SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits.

Dakdouk, H., 2022. Massive multi-player multi-armed bandits for internet of things networks.

Richard, H., Boursier, E. and Perchet, V., 2024, April. Constant or logarithmic regret in asynchronous multiplayer bandits with limited communication.