IJCAI
International Joint Conferences on Artificial Intelligence Organization

Multi-player Multi-armed Bandits with Delayed Feedback

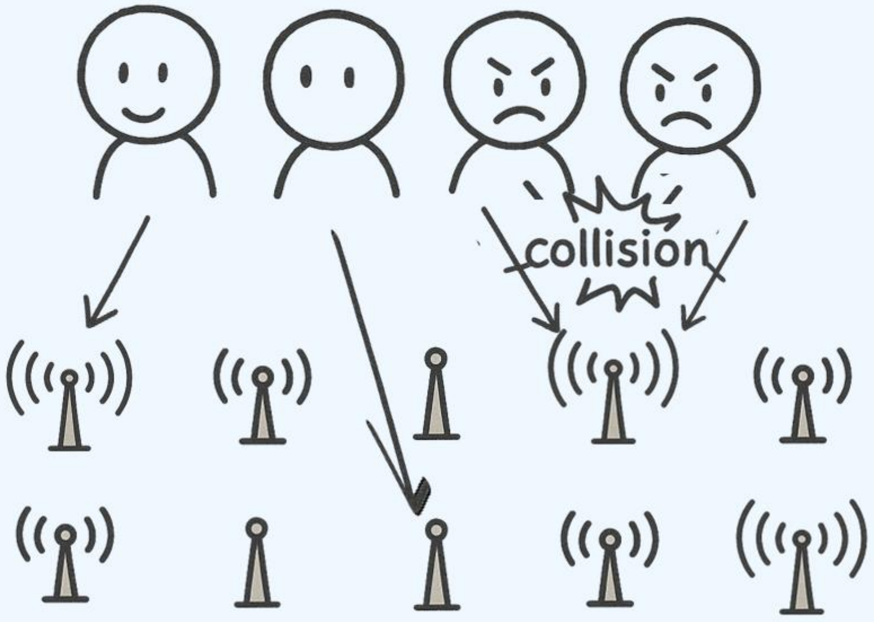
Presenter: Jingqi Fan
Paper ID: 1412

Jingqi Fan[1], Zilong Wang[2], Shuai Li[2], Linghe Kong[2]

1 Northeastern University, China   2 Shanghai Jiao Tong University

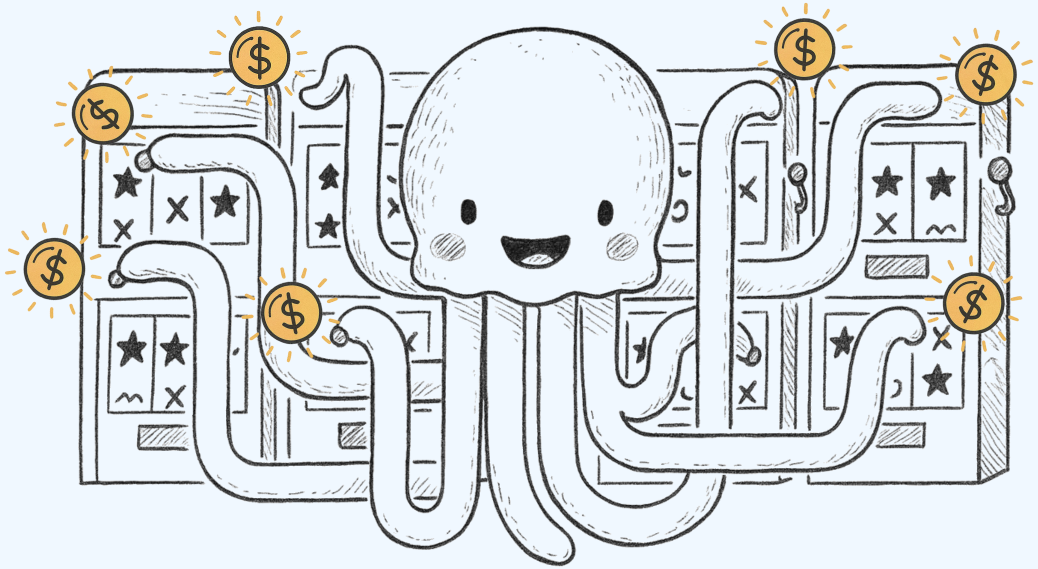
Cognitive Radio Networks

- A new type of nework.
- Users in cognitive radio networks choose different channels.
- If users choose the same channel, all of them will fail to pass information. We say that there is a collision occur.
- Users experience delay in cognitive radio networks.



Multi-armed Bandits

- A player pulls an arm and gets a reward.
- We can design some algorithm to maximize the accumlated rewards.
- Then we want to prove this algorithm can works well theoretically.



Multi-player Multi-armed Bandits (MP-MAB) with Delayed Feedback

Problem Formulation:

- M players, K arms, T total steps.
- Let $[M] := \{1, ..., M\}$ and $[K] := \{1, ..., K\}$.
- At each step s, each player $j \in [M]$ pulls an arm $\pi^j(t) \in [K]$.
- The environment generates $X^j(s) \sim \mathrm{Bernoulli}(\mu_{\pi^j(s)})$ and $r^j(s) := X^j(s)[1 - \eta^j(s)]$.
- The environment also generates $d^j(s) \sim D_{\pi^j(s)}$, where $D_{\pi^j(s)}$ is an unknow distribution.
- Then, at step $s + d^j(s) - 1$, player j receives the feedback $[r^j(s), \eta^j(s), s]$.

Multi-player Multi-armed Bandits (MP-MAB) with Delayed Feedback

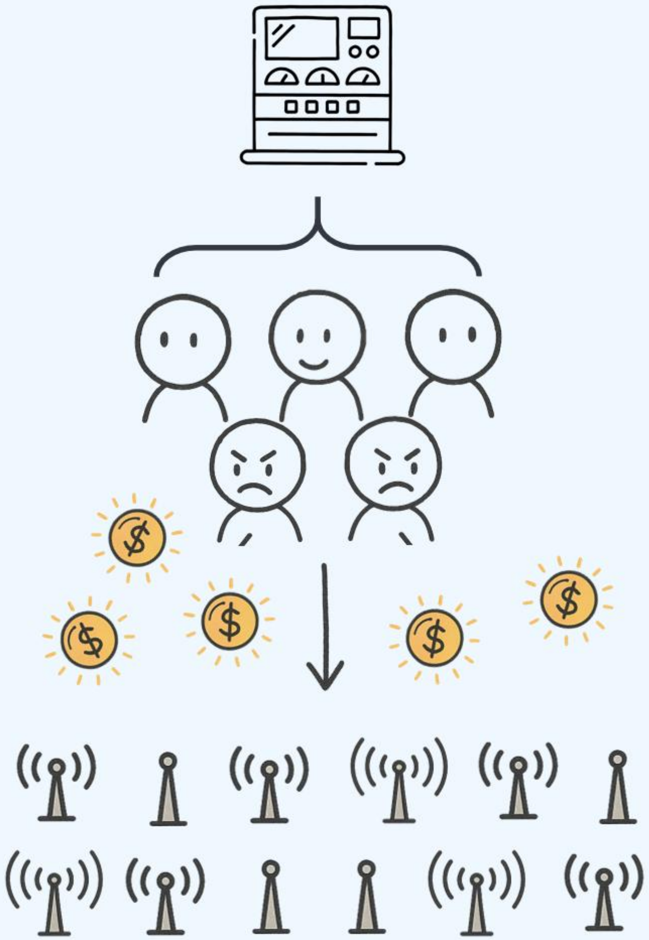
Goal: minimize the regret

$$\mathbb{E}[R(T)] := \sum_{s \le T} \sum_{k \le M} \mu_k - \mathbb{E}\left[\sum_{s \le T} \sum_{j \le M} r^j(s)\right],$$

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

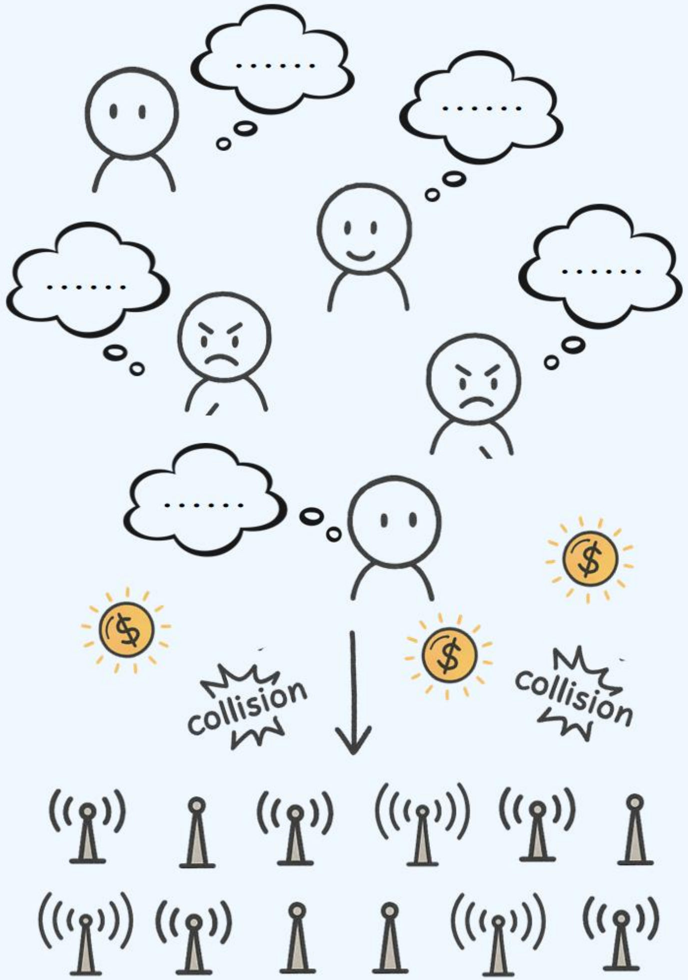
Assumption:

1. $D_k = D_{k'} = D, \forall k \in [K]$. D is sub-Gaussian.
   - σ_d^2 denotes the sub-Gaussian parameter and $\mathbb{E}[d]$ denotes the expectation.
   - Note that σ_d^2 and $\mathbb{E}[d]$ are unknown.
2. Each player is aware of her own rank j.



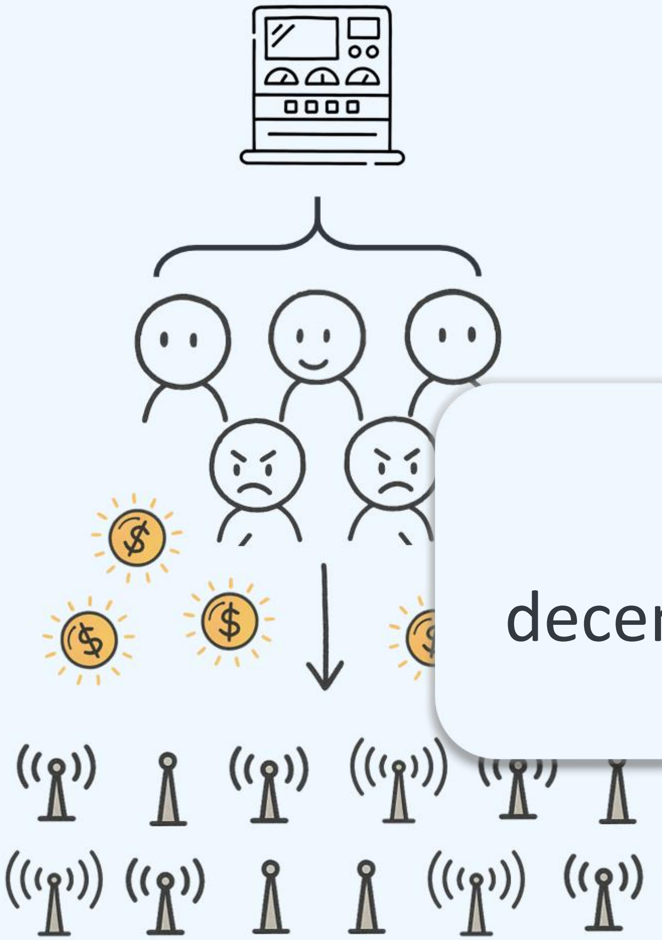
Centralized MP-MAB

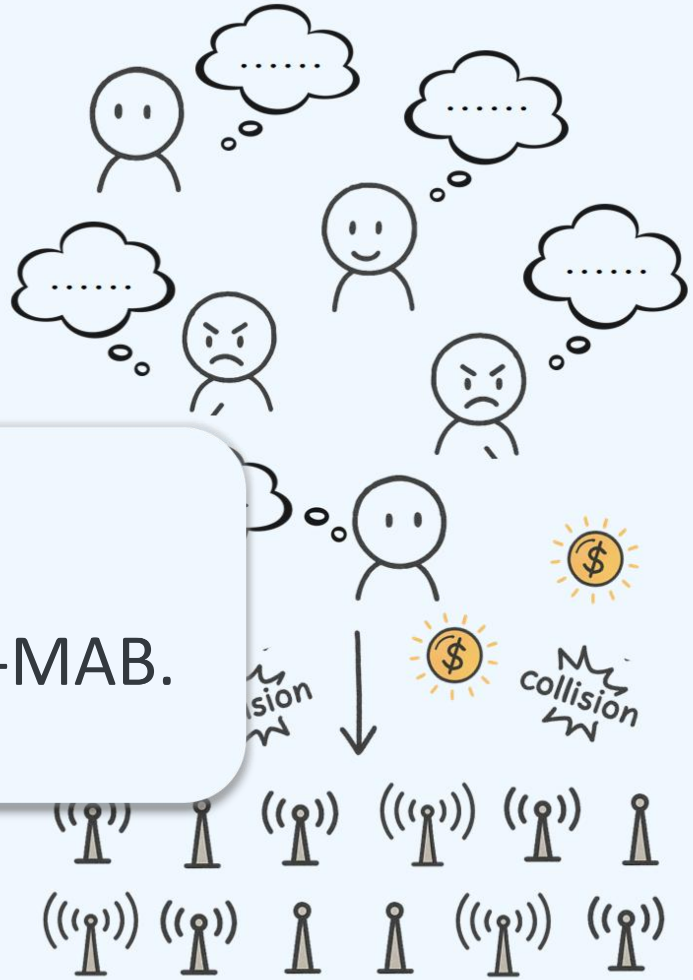
There exists a centra coordinator.
No collision occur.



Decentralized MP-MAB

Payers cannot see others' choices,
rewards, collisions.





We focus on decentralized MP-MAB.

Centralized MP-MAB

There exists a centra coordinator.
No collision occur.

Decentralized MP-MAB

Payers cannot see others' choices, rewards, collisions.

Our Algorithm: DDSE

- The algorithm is divided into many exploration-communication phases.
- Let $\mathcal{M}^j(p)$ denote the set of empirical optimal arms during the p-th phase. $|\mathcal{M}^j(p)| = M$. Players initialize $\mathcal{M}^j(1)$ which is a list with $\mathcal{M}^j(1) = \mathcal{M}^{j'}(1)$ for any $j, j' \in [M]$.
- Players are divided into a leader and many followers.
- They pull arms in a round-robin way to avoid collisions while the leader is in charge of exploring arms. **[Exploration Phase]**
- Sometimes they collide on purpose to pass messages. **[Communication Phase]**

Our Algorithm: DDSE

- When a player j receives a feedback at time t, she updates $n_k^j(t), \hat{\mu}_k^j(t), \mathrm{UCB}_k^j(t), \mathrm{LCB}_k^j(t)$ and the estimation of $\mathbb{E}[d]$ and σ_d^2 with

$$\hat{\mu}_d^j(t) := \frac{\sum_{s<t} \left(d^j(s)\mathbb{1}\{s + d^j(s) < t\}\right)}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s<t} \left([d^j(s) - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d^j(s) < t\}\right)^2}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}}.$$

the number of times to pull an arm

Our Algorithm: DDSE

- When a player j receives a feedback at time t, she updates $n_k^j(t), \hat{\mu}_k^j(t), \mathrm{UCB}_k^j(t), \mathrm{LCB}_k^j(t)$ and the estimation of $\mathbb{E}[d]$ and σ_d^2 with

$$\hat{\mu}_d^j(t) := \frac{\sum_{s<t} \left(d^j(s)\mathbb{1}\{s + d^j(s) < t\}\right)}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s<t} \left([d^j(s) - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d^j(s) < t\}\right)^2}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}}.$$

Our Algorithm: DDSE

the number of times to pull an arm

- When a player j receives a feedback at time t, she updates $n_k^j(t), \hat{\mu}_k^j(t), \mathrm{UCB}_k^j(t), \mathrm{LCB}_k^j(t)$ and the estimation of $\mathbb{E}[d]$ and σ_d^2 with

empirical reward expectation

$$\hat{\mu}_d^j(t) := \frac{\sum_{s<t} \left(d^j(s)\mathbb{1}\{s + d^j(s) < t\}\right)}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s<t} \left([d^j(s) - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d^j(s) < t\}\right)^2}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}}.$$

Our Algorithm: DDSE

the number of times to pull an arm

upper/lower confidence bound

- When a player j receives a feedback at time t, she updates $n_k^j(t), \hat{\mu}_k^j(t), \mathrm{UCB}_k^j(t), \mathrm{LCB}_k^j(t)$ and the estimation of $\mathbb{E}[d]$ and σ_d^2 with

empirical reward expectation

$$\hat{\mu}_d^j(t) := \frac{\sum_{s<t} \left(d^j(s)\mathbb{1}\{s + d^j(s) < t\}\right)}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s<t} \left([d^j(s) - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d^j(s) < t\}\right)^2}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}}.$$

Our Algorithm: DDSE

the number of times to pull an arm

upper/lower confidence bound

- When a player j receives a feedback at time t, she updates $n_k^j(t), \hat{\mu}_k^j(t), \mathrm{UCB}_k^j(t), \mathrm{LCB}_k^j(t)$ and the estimation of $\mathbb{E}[d]$ and σ_d^2 with

empirical reward expectation

$$\hat{\mu}_d^j(t) := \frac{\sum_{s<t} \left(d^j(s)\mathbb{1}\{s + d^j(s) < t\}\right)}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s<t} \left([d^j(s) - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d^j(s) < t\}\right)^2}{\sum_{s<t} \mathbb{1}\{s + d^j(s) < t\}}.$$

- Each player j aims to find q^j such that

$$q^j = \arg\min_q \left\{ q \in \mathbb{N} \mid t > \hat{\mu}_d^j(t) + (p - q)KM\log(T)\sqrt{2(\hat{\sigma}_d^2)^j(t)\log\left((M-1)(K+2M)(T)\right)} \right\}.$$

- Starting from Phase 2, players always use some old exploration results, i.e., $\mathcal{M}^j(p - q^j)$, to mitigate the influence of delay.

Upper Bound

Let $\Delta_{k,\ell} := \mu_k - \mu_\ell$ and $\Delta := \min_{k \le M} \mu_k - \mu_{k+1}$. In decentralized setting, given any K, M and a quantile $\theta \in (0,1)$, the regret of the algorithm satisfies

$$\mathbb{E}[R(T)] \le \sum_{k>M} \frac{323 \log(T)}{\theta \Delta_{M,k}} + \frac{M}{K-M} \sum_{k>M} \Delta_{1,k} d_1 + \frac{15}{\theta} d_2 + d_3 + C \,,$$

Lower Bound

For any sub-optimal gap set $S_\Delta = \{\Delta_{M,k} \mid \Delta_{M,k} = \mu_{(M)} - \mu_{(k)} \in [0,1]\}$ of cardinality $K - M$ and a quantile $\theta \in (0,1)$, there exists an instance with an order on S_Δ and a sub-Gaussian delay distribution such that

$$\mathbb{E}[R(T)] \ge \sum_{k>M} \frac{(1 - o(1)) \log(T)}{2\theta \Delta_{M,k}} + \frac{M}{2K} \sum_{k>M} \Delta_{M,k} d_4 - \frac{2}{\theta} \,,$$

where

$$d_1 = 2\mathbb{E}[d] + \sigma_d \sqrt{3 \log(K)} \,, \quad d_2 = \mathbb{E}[d] + \sigma_d \sqrt{2 \log(\frac{1}{1-\theta})} \,,$$

$$d_3 = \frac{656\sqrt{2}\sigma_d^2}{\theta K^2 M^2} + 3\sqrt{6}\sigma_d \,, \quad d_4 = \mathbb{E}[d] - \sigma_d \sqrt{\frac{\theta}{1-\theta}} \,, C = \sum_{k>M} \frac{195}{\theta \Delta_{M,k}} + \frac{4M}{\Delta^2} \,.$$