Multi-Player Multi-Armed Bandits with Finite Shareable Resources Arms: Learning Algorithms & Applications

Xuchuang Wang[1], Hong Xie[2], John C.S. Lui[1]

The Chinese University of Hong Kong[1], Chongqing University[2]

June 8, 2022

(Single-Player) Multi-Armed Bandits

- K arms: each associated with a $[0,1]$-supported reward X_k with mean μ_k.
  - Assume $\mu_1 > \mu_2 > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls an arm $k_t \in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$.
- Goal: maximize total reward; or minimize the regret

$$\mathbb{E}[\mathrm{Reg}(T)] := T\mu_1 - \sum_{t=1}^{T} \mu_{k_t}.$$

(Single-Player) Multi-Armed Bandits

- K arms: each associated with a $[0,1]$-supported reward X_k with mean μ_k.
  - Assume $\mu_1 > \mu_2 > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls an arm $k_t \in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$.
- Goal: maximize total reward; or minimize the regret

$$\mathbb{E}[\mathrm{Reg}(T)] := \boxed{\underbrace{T\mu_1}_{\text{Optimal}}} - \sum_{t=1}^{T} \mu_{k_t}.$$

(Single-Player) Multi-Armed Bandits

- K arms: each associated with a $[0,1]$-supported reward X_k with mean μ_k.
  - Assume $\mu_1 > \mu_2 > \cdots > \mu_K$.
- For $t = 1, \ldots, T$:
  - Pulls an arm $k_t \in \{1, 2, \ldots, K\}$.
  - Collects reward $X_{k,t}$.
- Goal: maximize total reward; or minimize the regret

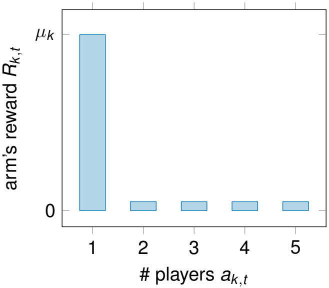
$$\mathbb{E}[\mathrm{Reg}(T)] := \underbrace{T\mu_1}_{\text{Optimal}} - \underbrace{\sum_{t=1}^{T} \mu_{k_t}}_{\text{Algorithm's}}.$$

Multi-Player Multi-Armed Bandits

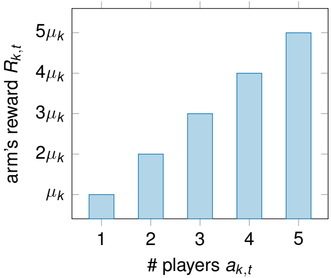
- K arms and M players.
- For $t = 1, \ldots, T$: For each player $i \in \{1, 2, \ldots, M\}$
  - Pulls an arm $k \in \{1, 2, \ldots, K\}$.
  - Collects reward $R_{k,t}$.
- Goal: minimize the **regret** of all M players

When More Than One Player Chooses The Same Arm

- Collision (e.g., [1]): if two players i, j collides, then zero reward.
- Non-Collision (e.g., [4]): each player obtains an independent reward $X_{k,t}^{(i)}$



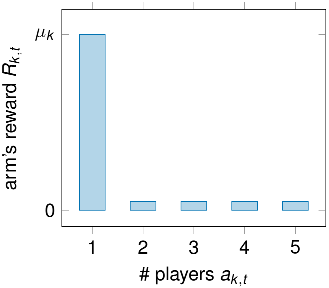
(a) Collision



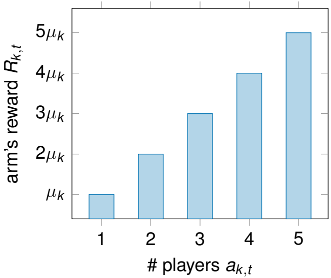
(b) Non-Collision

When More Than One Player Chooses The Same Arm

- Collision (e.g., [1]): if two players i, j collides, then zero reward.
- Non-Collision (e.g., [4]): each player obtains an independent reward $X_{k,t}^{(i)}$



(a) Collision



(b) Non-Collision

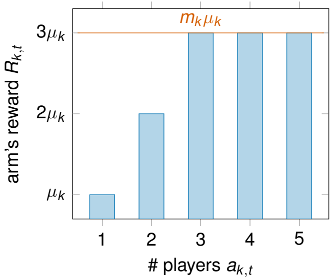
However, both can be too restrictive in practice.

Finite Shareable Resources Arm (MMAB-SA)

- Each arm has two **unknowns**:
  - “per-load” reward mean μ_k and integer resources m_k.
- If $a_{k,t}$ players share arm k with m_k resources, then

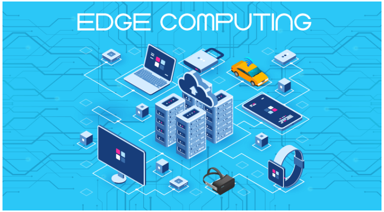
$$R_{k,t} := \min\{a_{k,t}, m_k\} X_{k,t} = \begin{cases} a_{k,t} X_{k,t}, & a_{k,t} \leqslant m_k \\ m_k X_{k,t}, & a_{k,t} > m_k \end{cases},$$

  - $X_{k,t}$ is the “per-load” reward random variable.

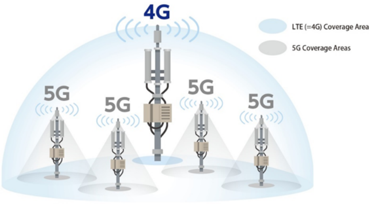


Two Types of Sharing Demand Feedback

- **Sharing Demand Information (SDI):**
  - observe the number of players $a_{k,t}$ that selects the arm k
- **Sharing Demand Awareness (SDA):**
  - know the sharing condition of the pulled arm, i.e., $\mathbb{1}\{a_{k,t} > 1\}$.



(a) Edge Computing [3]



(b) Wireless Network [2]

Two Algorithms for Two Types of Feedback

Algorithm 1 DPE-SDI for player i

▷ **Initialization phase:** assign each player a rank $i \in \{1, \ldots, M\}$; rank 1 player becomes the leader.

while $t \leqslant T$ **do**

▷ **Exploration-exploitation phase:** estimate means μ_k and resources m_k.

▷ **Communication phase:** leader updates and sends info. to followers.

Two Algorithms for Two Types of Feedback

Algorithm 1 DPE-SDI for player i

▷ **Initialization phase:** assign each player a rank $i \in \{1, \ldots, M\}$; rank 1 player becomes the leader.

while $t \leqslant T$ **do**

▷ **Exploration-exploitation phase:** estimate means μ_k and resources m_k.

▷ **Communication phase:** leader updates and sends info. to followers.

Algorithm 2 SIC-SDA for player i

▷ **Initialization phase**

while player i does not find an optimal arm **do**

▷ **Exploration phase:** estimate reward means μ_k and resources m_k.

▷ **Communication phase:** leader receives follower's statistics and send out its updated info. to followers.

▷ **Exploitation phase**: play the identified optimal arm till the end.

Theoretical Results

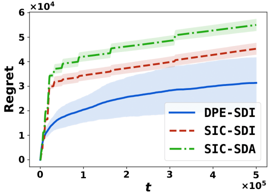
- For DPE-SDI,

$$\mathbb{E}[\mathrm{Reg}(T)] \leqslant O\left(\sum_{k=L+1}^{K} \frac{\log T}{\mu_L - \mu_k} + \sum_{k=1}^{M} \frac{m_k^2}{\mu_k^2} \log T\right).$$

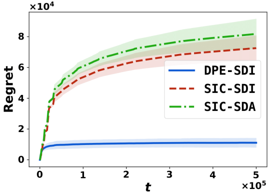
- For SIC-SDA,

$$\mathbb{E}[\mathrm{Reg}(T)] \leqslant O\left(\sum_{k=L+1}^{K} \frac{\boxed{M} \log T}{\mu_L - \mu_k} + \sum_{k=1}^{M} \frac{m_k^2}{\mu_k^2} \log T\right).$$

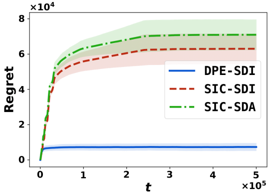
Simulations



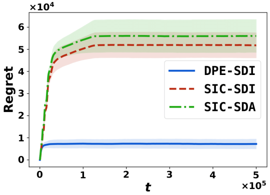
(a) $\Delta = 0.001$



(b) $\Delta = 0.012$



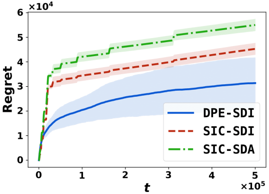
(c) $\Delta = 0.025$



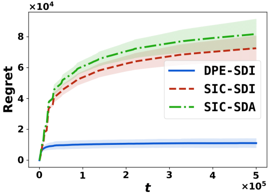
(d) $\Delta = 0.037$

Figure: Synthetic data simulations (SDI > SDA)

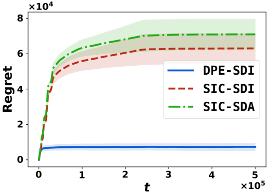
Simulations



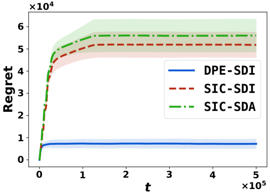
(a) $\Delta = 0.001$



(b) $\Delta = 0.012$

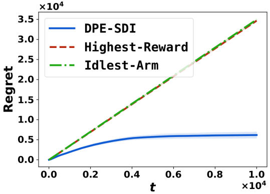


(c) $\Delta = 0.025$

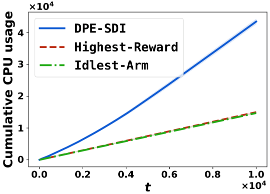


(d) $\Delta = 0.037$

Figure: Synthetic data simulations (SDI > SDA)

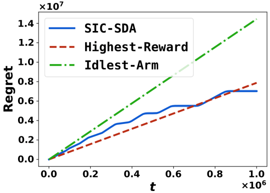


(a) Regret

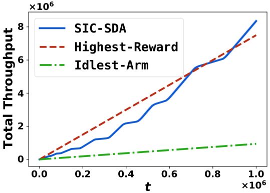


(b) Cumulative CPU usage

Figure: Edge computing (SDI)



(a) Regret



(b) Total throughput

Figure: 5G/4G network (SDA)

Thank you!

Full paper at arXiv:2204.13502

References I

[1] Etienne Boursier and Vianney Perchet. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 32, pages 12071–12080, 2019.

[2] Tokyu Corporation and Sumitomo Corporation. Launch of pilot experiment on 5g base-station-sharing business in shibuya, 2019. URL `https://www.sumitomocorp.com/en/africa/news/release/2019/group/12330`.

[3] SPEC INDIA. What is edge computing? the quick overview explained with examples, 2019. URL `https://www.spec-india.com/blog/what-is-edge-computing-the-quick-overview-explained-with-examp`

[4] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.