Exploration for Free: How Does Reward Heterogeneity Improve Regret in Cooperative Multi-agent Bandits?

Xuchuang Wang[1], Lin Yang[2], Yu-Zhen Janice Chen[3], Xutong Liu[1], Mohammad Hajiesmaili[3], Don Towsley[3], John C.S. Lui[1]

To Appear in UAI 2023

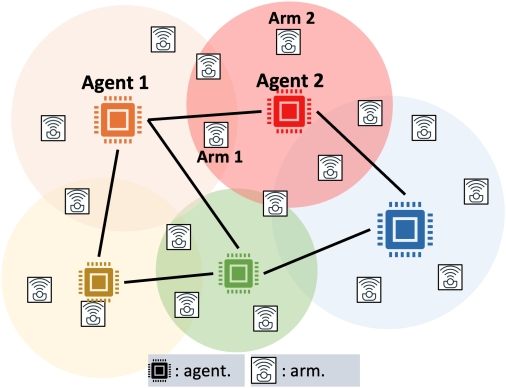
The Chinese University of Hong Kong[1], Nanjing University[2], University of Massachusetts Amherst[3]

香港中文大學 The Chinese University of Hong Kong

南京大學 NANJING UNIVERSITY

University of Massachusetts Amherst

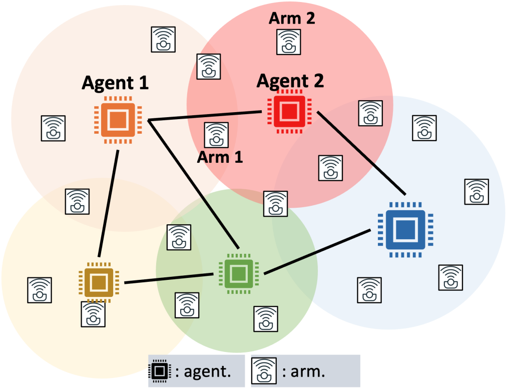
Motivation Example for Action Constrained Multi-Agent Bandits



Agents have access to its nearby arms

Motivation Example for Action Constrained Multi-Agent Bandits

1. M Agents and K Arms (set $\mathcal{K}$)
2. Each agent m has access to a subset of arms $\mathcal{K}^{(m)} \subseteq \mathcal{K}$
3. Overlap $\mathcal{K}^{(m)} \cap \mathcal{K}^{(m')}$ leads to cooperation.

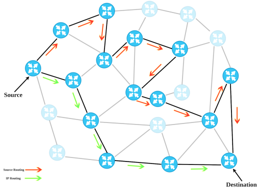


Agents have access to its nearby arms

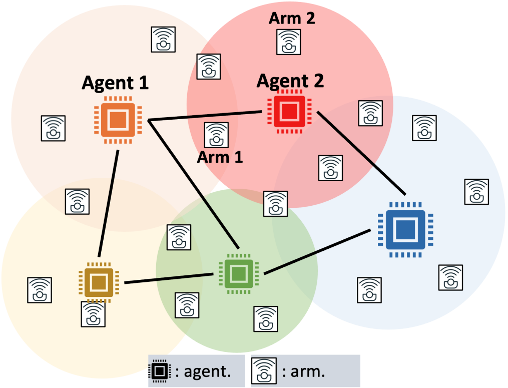
Motivation Example for Action Constrained Multi-Agent Bandits

1. M Agents and K Arms (set $\mathcal{K}$)
2. Each agent m has access to a subset of arms $\mathcal{K}^{(m)} \subseteq \mathcal{K}$
3. Overlap $\mathcal{K}^{(m)} \cap \mathcal{K}^{(m')}$ leads to cooperation.

(a) Drone swarm



(b) Path routing



Agents have access to its nearby arms

Action Constrained Multi-Agent Multi-Armed Bandits (1/2)

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$
  - Assume $\mu(1) > \cdots > \mu(K)$.

Action Constrained Multi-Agent Multi-Armed Bandits (1/2)

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$
  - Assume $\mu(1) > \cdots > \mu(K)$.
- M Agents: each agent m has access to a subset of local arms $\mathcal{K}^{(m)} \subseteq \mathcal{K}$
  - Local optimal arm $k_*^{(m)} := \arg\max_{k \in \mathcal{K}^{(m)}} \mu(k)$

Action Constrained Multi-Agent Multi-Armed Bandits (1/2)

- K arms: each associated with a Bernoulli variable $X_t(k)$ with mean $\mu(k)$
  - Assume $\mu(1) > \cdots > \mu(K)$.
- M Agents: each agent m has access to a subset of local arms $\mathcal{K}^{(m)} \subseteq \mathcal{K}$
  - Local optimal arm $k_*^{(m)} := \arg\max_{k \in \mathcal{K}^{(m)}} \mu(k)$
- T Rounds: in each round $t \leqslant T$
  - Each agent m pulls an arm $k_t^{(m)} \in \mathcal{K}^{(m)}$ and collects reward $X_k^{(m)}(k_t^{(m)})$.

Action Constrained Multi-Agent Multi-Armed Bandits (2/2)

Group regret with K arms M Agents T Rounds

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] := \sum_{m \in \mathcal{M}} \underbrace{\sum_{t \in \mathcal{T}} (\mu(k_*^{(m)}) - \mu(k_t^{(m)}))}_{\text{agent } m\text{'s regret}}$$

Action Constrained Multi-Agent Multi-Armed Bandits (2/2)

Group regret with K arms M Agents T Rounds

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] := \sum_{m\in\mathcal{M}} \underbrace{\sum_{t\in\mathcal{T}} (\mu(k_*^{(m)}) - \mu(k_t^{(m)}))}_{\text{agent } m\text{'s regret}}$$

$$= \sum_{m\in\mathcal{M}} \sum_{t\in\mathcal{T}} \Delta^{(m)}(k)$$

- $\Delta^{(m)}(k) := \mu(k_*^{(m)}) - \mu(k)$ is the reward gap of arm k with respect to agent m's local optimal arm $k_*^{(m)}$.

Action Constrained Multi-Agent Multi-Armed Bandits (2/2)

Group regret with K arms M Agents T Rounds

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] := \sum_{m\in\mathcal{M}} \underbrace{\sum_{t\in\mathcal{T}}(\mu(k_*^{(m)}) - \mu(k_t^{(m)}))}_{\text{agent } m\text{'s regret}}$$

$$= \sum_{m\in\mathcal{M}} \sum_{t\in\mathcal{T}} \Delta^{(m)}(k)$$

$$= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)$$

- $\Delta^{(m)}(k) := \mu(k_*^{(m)}) - \mu(k)$ is the reward gap of arm k with respect to agent m's local optimal arm $k_*^{(m)}$.
- $n_T^{(m)}(k)$ is the number of times that agent m pulls arm k till the end.

Individual *v.s.* Cooperative: Algorithm Design

1. Individual UCB arm pull policy: at time t, agent m pulls

$$k_t^{(m)} = \arg\max_{k\in\mathcal{K}^{(m)}} \mathtt{UCB}_t^{(m)}(k) = \arg\max_{k\in\mathcal{K}^{(m)}} \underbrace{\hat{\mu}_t^{(m)}(k)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\log t}{n_t^{(m)}(k)}}}_{\text{confidence radius}},$$

where $n_t^{(m)}(k)$ is the number of times that agent m pulls arm k up to time t, and $\hat{\mu}_t^{(m)}(k)$ is the average of these $n_t^{(m)}(k)$'s observations.

Individual *v.s.* Cooperative: Algorithm Design

1. Individual UCB arm pull policy: at time t, agent m pulls

$$k_t^{(m)} = \arg\max_{k\in\mathcal{K}^{(m)}} \mathrm{UCB}_t^{(m)}(k) = \arg\max_{k\in\mathcal{K}^{(m)}} \underbrace{\hat{\mu}_t^{(m)}(k)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\log t}{n_t^{(m)}(k)}}}_{\text{confidence radius}},$$

   where $n_t^{(m)}(k)$ is the number of times that agent m pulls arm k up to time t, and $\hat{\mu}_t^{(m)}(k)$ is the average of these $n_t^{(m)}(k)$'s observations.

2. Cooperative UCB arm pull policy: at time t, agent m pulls

$$k_t^{(m)} = \arg\max_{k\in\mathcal{K}^{(m)}} \mathrm{UCB}_t(k) = \arg\max_{k\in\mathcal{K}^{(m)}} \underbrace{\hat{\mu}_t(k)}_{\substack{\text{empirical mean}\\ \text{global}}} + \underbrace{\sqrt{\frac{2\log t}{n_t(k)}}}_{\substack{\text{confidence radius}\\ \text{global}}},$$

   where $n_t(k)$ is the number of times that all M agents pull arm k up to time t, and $\hat{\mu}_t(k)$ is the average of these $n_t(k)$'s observations.

Individual *v.s.* Cooperative: Regret Analysis

1. Individual UCB's regret:

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\}} \frac{\log T}{\Delta^{(m)}(k)}\right),$$

where $\Delta^{(m)}(k) := \mu(k_*^{(m)}) - \mu(k)$ is the reward gap of arm k with respect to agent m's local optimal arm.

Individual *v.s.* Cooperative: Regret Analysis

1. Individual UCB's regret:

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\}}\frac{\log T}{\Delta^{(m)}(k)}\right),$$

where $\Delta^{(m)}(k) := \mu(k_*^{(m)}) - \mu(k)$ is the reward gap of arm k with respect to agent m's local optimal arm.

2. Cooperative UCB's regret:

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\cup_{m\in\mathcal{M}}\left(\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\}\right)}\frac{\log T}{\tilde{\Delta}(k)}\right),$$

where $\tilde{\Delta}(k) := \min_{m:k\in\mathcal{K}^{(m)}}\Delta^{(m)}(k)$ is the smallest reward gap of local suboptimal arm k with respect to any feasible agent m's local optimal arm.

Individual *v.s.* Cooperative: Regret Analysis

1. Individual UCB's regret:

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\}} \frac{\log T}{\Delta^{(m)}(k)}\right),$$

   where $\Delta^{(m)}(k) := \mu(k_*^{(m)}) - \mu(k)$ is the reward gap of arm k with respect to agent m's local optimal arm.

2. Cooperative UCB's regret: **How good?**

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\cup_{m\in\mathcal{M}}\left(\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\}\right)} \frac{\log T}{\tilde{\Delta}(k)}\right),$$

   where $\tilde{\Delta}(k) := \min_{m:k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k)$ is the smallest reward gap of local suboptimal arm k with respect to any feasible agent m's local optimal arm.

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathsf{R}_{\mathsf{T}}(\mathcal{A})] = \sum_{m \in \mathcal{M}} \underbrace{\sum_{k \in \mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] = \sum_{m\in\mathcal{M}} \underbrace{\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

1. In single agent m's bandit, to distinguish a suboptimal arm requires $n_T^{(m)}(k) = \Omega\left(\frac{\log T}{(\Delta^{(m)}(k))^2}\right)$ pulls.

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] = \sum_{m\in\mathcal{M}} \underbrace{\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

1. In single agent m's bandit, to distinguish a suboptimal arm requires $n_T^{(m)}(k) = \Omega\left(\frac{\log T}{(\Delta^{(m)}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) \times \frac{\log T}{(\Delta^{(m)}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \frac{\log T}{\Delta^{(m)}(k)}\right)$ regret lower bound.

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] = \sum_{m\in\mathcal{M}} \underbrace{\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

1. In single agent m's bandit, to distinguish a suboptimal arm requires $n_T^{(m)}(k) = \Omega\left(\frac{\log T}{(\Delta^{(m)}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) \times \frac{\log T}{(\Delta^{(m)}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \frac{\log T}{\Delta^{(m)}(k)}\right)$ regret lower bound.
2. In cooperative multi-agent bandits, we need $\Omega\left(\frac{\log T}{(\tilde{\Delta}(k))^2}\right)$ pulls.

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathrm{R}_\mathrm{T}(\mathcal{A})] = \sum_{m\in\mathcal{M}} \underbrace{\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

1. In single agent m's bandit, to distinguish a suboptimal arm requires $n_T^{(m)}(k) = \Omega\left(\frac{\log T}{(\Delta^{(m)}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) \times \frac{\log T}{(\Delta^{(m)}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \frac{\log T}{\Delta^{(m)}(k)}\right)$ regret lower bound.
2. In cooperative multi-agent bandits, we need $\Omega\left(\frac{\log T}{(\tilde{\Delta}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}} \tilde{\Delta}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}} \frac{\log T}{\tilde{\Delta}(k)}\right)$ regret lower bound?

Regret Lower Bound and Free Exploration Intuition (1/2)

$$\mathbb{E}[\mathrm{R}_\mathrm{T}(\mathcal{A})] = \sum_{m\in\mathcal{M}} \underbrace{\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) n_T^{(m)}(k)}_{\text{regret of agent } m}$$

1. In single agent m's bandit, to distinguish a suboptimal arm requires $n_T^{(m)}(k) = \Omega\left(\frac{\log T}{(\Delta^{(m)}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \Delta^{(m)}(k) \times \frac{\log T}{(\Delta^{(m)}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}^{(m)}} \frac{\log T}{\Delta^{(m)}(k)}\right)$ regret lower bound.
2. In cooperative multi-agent bandits, we need $\Omega\left(\frac{\log T}{(\tilde{\Delta}(k))^2}\right)$ pulls.
   - Therefore, $\Omega\left(\sum_{k\in\mathcal{K}} \tilde{\Delta}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right) = \Omega\left(\sum_{k\in\mathcal{K}} \frac{\log T}{\tilde{\Delta}(k)}\right)$ regret lower bound?

Regret Lower Bound and Free Exploration Intuition (2/2)

1. Wait! Agents cooperate: $\Omega\left(\sum_{k\in\mathcal{K}} \Delta^{(\boldsymbol{?})}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right)$

Regret Lower Bound and Free Exploration Intuition (2/2)

1. Wait! Agents cooperate: $\Omega\left(\sum_{k\in\mathcal{K}} \Delta^{(\textbf{?})}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right)$
2. May exist $\Delta^{(m)}(k) = 0$ if the arm $k = k_*^{(m)}$ is local optimal for some agent m.

Regret Lower Bound and Free Exploration Intuition (2/2)

1. Wait! Agents cooperate: $\Omega\left(\sum_{k\in\mathcal{K}} \Delta^{(\textbf{?})}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right)$
2. May exist $\Delta^{(m)}(k) = 0$ if the arm $k = k_*^{(m)}$ is local optimal for some agent m.
3. Denote $\mathcal{F} := \{k_*^{(m)} : m \in \mathcal{M}\}$ as arms can be **freely explored**.

$$\Omega\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right)$$

Regret Lower Bound and Free Exploration Intuition (2/2)

1. Wait! Agents cooperate: $\Omega\left(\sum_{k\in\mathcal{K}} \Delta^{(\textbf{?})}(k) \times \frac{\log T}{(\tilde{\Delta}(k))^2}\right)$
2. May exist $\Delta^{(m)}(k) = 0$ if the arm $k = k_*^{(m)}$ is local optimal for some agent m.
3. Denote $\mathcal{F} \coloneqq \{k_*^{(m)} : m \in \mathcal{M}\}$ as arms can be **freely explored**.

$$\Omega\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right)$$

4. Cooperative UCB is not optimal.
   - $\mathcal{K}\setminus\mathcal{F} \subseteq \cup_{m\in\mathcal{M}}(\mathcal{K}^{(m)} \setminus \{k_*^{(m)}\})$ of Cooperative UCB's upper bound.

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

`UCB` [Auer, 2002]	
`CO-UCB` [Yang et al., 2022]	
`FreeExp` (our work)	

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

UCB [Auer, 2002]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(1,3)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
CO-UCB [Yang et al., 2022]	
FreeExp (our work)	

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

`UCB` [Auer, 2002]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(1,3)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
`CO-UCB` [Yang et al., 2022]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
`FreeExp` (our work)	

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

UCB [Auer, 2002]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(1,3)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
CO-UCB [Yang et al., 2022]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
FreeExp (our work)	$O(1)$

Key Contribution Illustrated: three arms $\mu(1) > \mu(2) > \mu(3)$

	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 1	✓	✓	✓
Agent 2	✗	✓	✓
Agent 3	✗	✗	✓

`UCB` [Auer, 2002]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(1,3)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
`CO-UCB` [Yang et al., 2022]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(2,3)}\right) \log T\right)$
`FreeExp` (our work)	$O(1)$

1. $\mathcal{K} = \mathcal{F}$, all arms can be freely explored: `FreeExp` achieves constant regret.

Free Exploration: Algorithm Design

Algorithm 1 The `FreeExp` Algorithm (for Agent m)

1: **for** each time slot t **do**
2: $\quad I_t^{(m)} \leftarrow \arg\max_{k \in \mathcal{K}^{(m)}} \hat{\mu}_t(k)$ $\triangleright$ `identify empirical optimal arm`
3: $\quad$ Send $I_t^{(m)}$ to other agents and collect their $I_t^{(m')}$

Free Exploration: Algorithm Design

Algorithm 1 The `FreeExp` Algorithm (for Agent m)

1: **for** each time slot t **do**
2: $\quad I_t^{(m)} \leftarrow \arg\max_{k \in \mathcal{K}^{(m)}} \hat{\mu}_t(k)$ $\triangleright$ `identify empirical optimal arm`
3: $\quad$ Send $I_t^{(m)}$ to other agents and collect their $I_t^{(m')}$
4: $\quad \mathcal{D}_t^{(m)} \leftarrow \{k \in \mathcal{K}^{(m)} : \texttt{UCB}_t(k) > \hat{\mu}_t(I_t^{(m)})\}$ $\triangleright$ `choose high KL-UCB arms`
5: $\quad \mathcal{D}_t^{(m)} \leftarrow \mathcal{D}_t^{(m)} \setminus \{I_t^{(m')} : \forall m' \in \mathcal{M}\}$ $\triangleright$ `free exploration`

Free Exploration: Algorithm Design

Algorithm 1 The `FreeExp` Algorithm (for Agent m)

for each time slot t **do**
  $I_t^{(m)} \leftarrow \arg\max_{k \in \mathcal{K}^{(m)}} \hat{\mu}_t(k)$ ▷ **identify empirical optimal arm**
  Send $I_t^{(m)}$ to other agents and collect their $I_t^{(m')}$
  $\mathcal{D}_t^{(m)} \leftarrow \{k \in \mathcal{K}^{(m)} : \mathtt{UCB}_t(k) > \hat{\mu}_t(I_t^{(m)})\}$ ▷ **choose high KL-UCB arms**
  $\mathcal{D}_t^{(m)} \leftarrow \mathcal{D}_t^{(m)} \setminus \{I_t^{(m')} : \forall m' \in \mathcal{M}\}$ ▷ **free exploration**
  if $\mathcal{D}_t^{(m)} = \emptyset$ **then**
    $J_t^{(m)} \leftarrow I_t^{(m)}$ ▷ **exploit, if correct only agent m pulls $k_*^{(m)}$**
  else
    $J_t^{(m)} \leftarrow \begin{cases} I_t^{(m)} & \text{w.p. } \frac{1}{2} \\ \text{uniformly pick an arm from } \mathcal{D}_t^{(m)} & \text{w.p. } \frac{1}{2} \end{cases}$ ▷ **explore**

Free Exploration: Algorithm Design

Algorithm 1 The `FreeExp` Algorithm (for Agent m)

for each time slot t **do**

  $I_t^{(m)} \leftarrow \arg\max_{k \in \mathcal{K}^{(m)}} \hat{\mu}_t(k)$ ▷ **identify empirical optimal arm**

  Send $I_t^{(m)}$ to other agents and collect their $I_t^{(m')}$

  $\mathcal{D}_t^{(m)} \leftarrow \{k \in \mathcal{K}^{(m)} : \mathtt{UCB}_t(k) > \hat{\mu}_t(I_t^{(m)})\}$ ▷ **choose high KL-UCB arms**

  $\mathcal{D}_t^{(m)} \leftarrow \mathcal{D}_t^{(m)} \setminus \{I_t^{(m')} : \forall m' \in \mathcal{M}\}$ ▷ **free exploration**

  if $\mathcal{D}_t^{(m)} = \emptyset$ **then**

    $J_t^{(m)} \leftarrow I_t^{(m)}$ ▷ **exploit, if correct only agent m pulls $k_*^{(m)}$**

  else

    $J_t^{(m)} \leftarrow \begin{cases} I_t^{(m)} & \text{w.p. } \frac{1}{2} \\ \text{uniformly pick an arm from } \mathcal{D}_t^{(m)} & \text{w.p. } \frac{1}{2} \end{cases}$ ▷ **explore**

  Pull arm $J_t^{(m)}$ and receive observations

  Synchronize observations with other agents and Update $\hat{\mu}_t(k)$ and $\mathtt{UCB}_t(k)$

Free Exploration: Analysis (1/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

Free Exploration: Analysis (1/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Theoretical improvement: summation range

$$\underbrace{(m\times k)\in(\mathcal{M}\times\mathcal{K})}_{\text{UCB [Auer, 2002]}} \Longrightarrow \underbrace{k\in\bigcup\nolimits_{m\in\mathcal{M}}(\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\})}_{\text{CO-UCB [Yang et al., 2022]}} \Longrightarrow \underbrace{k\in\mathcal{K}\setminus\mathcal{F}}_{\texttt{FreeExp}\text{ (ours)}}$$

Free Exploration: Analysis (1/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Theoretical improvement: summation range

$$\underbrace{(m\times k)\in(\mathcal{M}\times\mathcal{K})}_{\text{UCB [Auer, 2002]}} \Longrightarrow \underbrace{k\in\bigcup\nolimits_{m\in\mathcal{M}}(\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\})}_{\text{CO-UCB [Yang et al., 2022]}} \Longrightarrow \underbrace{k\in\mathcal{K}\setminus\mathcal{F}}_{\texttt{FreeExp}\text{ (ours)}}$$

2. Regret optimality: Match regret lower bound up to constant coefficients.

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \geqslant \Omega\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) \qquad \text{“informal”}$$

Free Exploration: Analysis (1/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[R_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Theoretical improvement: summation range

$$\underbrace{(m\times k)\in(\mathcal{M}\times\mathcal{K})}_{\text{UCB [Auer, 2002]}} \Longrightarrow \underbrace{k\in\bigcup\nolimits_{m\in\mathcal{M}}(\mathcal{K}^{(m)}\setminus\{k_*^{(m)}\})}_{\text{CO-UCB [Yang et al., 2022]}} \Longrightarrow \underbrace{k\in\mathcal{K}\setminus\mathcal{F}}_{\texttt{FreeExp}\text{ (ours)}}$$

2. Regret optimality: Match regret lower bound up to constant coefficients.

$$\mathbb{E}[R_T(\mathcal{A})] \geqslant \Omega\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) \qquad \text{“informal”}$$

3. Finite regret in special case:
When $\mathcal{K}=\mathcal{F}$ (all arms are free), the regret reduces $O(1)$.

Free Exploration: Analysis (2/2)

Theorem (FreeExp's Regret Upper Bound)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

Free Exploration: Analysis (2/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Regret due to free arms in $\mathcal{F}$

2. Regret due to non-free arms in $\mathcal{K}\setminus\mathcal{F}$

Free Exploration: Analysis (2/2)

Theorem (FreeExp's Regret Upper Bound)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Regret due to free arms in $\mathcal{F}$
   - **Finite # time slots** $\underbrace{\{I_t^{(m)} : m \in \mathcal{M}\}}_{\text{estimated free arm set}} \neq \underbrace{\mathcal{F} = \{k_*^{(m)} : m \in \mathcal{M}\}}_{\text{true free arm set}} \implies$ **Finite regret**
2. Regret due to non-free arms in $\mathcal{K}\setminus\mathcal{F}$

Free Exploration: Analysis (2/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Regret due to free arms in $\mathcal{F}$
   - **Finite # time slots** $\underbrace{\{I_t^{(m)} : m \in \mathcal{M}\}}_{\text{estimated free arm set}} \neq \underbrace{\mathcal{F} = \{k_*^{(m)} : m \in \mathcal{M}\}}_{\text{true free arm set}} \Longrightarrow$ **Finite regret**
2. Regret due to non-free arms in $\mathcal{K}\setminus\mathcal{F}$
   - $\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \sum_{m\in\mathcal{M}} \Delta^{(m)}(k) n_T^{(m)}(k)$ and $\sum_{m\in\mathcal{M}} n_T^{(m)}(k) \leqslant \frac{\log T}{(\tilde{\Delta}(k))^2} \not\Longrightarrow \frac{\log T}{\tilde{\Delta}(k)}$ regret

Free Exploration: Analysis (2/2)

Theorem (FreeExp's Regret Upper Bound)

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Regret due to free arms in $\mathcal{F}$
   - **Finite # time slots** $\underbrace{\{I_t^{(m)} : m \in \mathcal{M}\}}_{\text{estimated free arm set}} \neq \underbrace{\mathcal{F} = \{k_*^{(m)} : m \in \mathcal{M}\}}_{\text{true free arm set}} \implies$ **Finite regret**
2. Regret due to non-free arms in $\mathcal{K}\setminus\mathcal{F}$
   - $\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \sum_{m\in\mathcal{M}} \Delta^{(m)}(k) n_T^{(m)}(k)$ and $\sum_{m\in\mathcal{M}} n_T^{(m)}(k) \leqslant \frac{\log T}{(\tilde{\Delta}(k))^2} \not\Longrightarrow \frac{\log T}{\tilde{\Delta}(k)}$ regret
   - $\sigma(\ell)$: **sort** agents in a **descending order** based on the magnitude of $\Delta^{(m)}(k)$

Free Exploration: Analysis (2/2)

Theorem (`FreeExp`'s Regret Upper Bound)

$$\mathbb{E}[\mathrm{R}_T(\mathcal{A})] \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \frac{\log T}{\tilde{\Delta}(k)}\right) + O\left(\sum_{k\in\mathcal{F}} 1\right)$$

1. Regret due to free arms in $\mathcal{F}$
   - **Finite # time slots** $\underbrace{\{I_t^{(m)} : m \in \mathcal{M}\}}_{\text{estimated free arm set}} \neq \underbrace{\mathcal{F} = \{k_*^{(m)} : m \in \mathcal{M}\}}_{\text{true free arm set}} \implies$ **Finite regret**
2. Regret due to non-free arms in $\mathcal{K}\setminus\mathcal{F}$
   - $\sum_{k\in\mathcal{K}\setminus\mathcal{F}} \sum_{m\in\mathcal{M}} \Delta^{(m)}(k) n_T^{(m)}(k)$ and $\sum_{m\in\mathcal{M}} n_T^{(m)}(k) \leqslant \frac{\log T}{(\tilde{\Delta}(k))^2} \not\Longrightarrow \frac{\log T}{\tilde{\Delta}(k)}$ regret
   - $\sigma(\ell)$: **sort** agents in a **descending order** based on the magnitude of $\Delta^{(m)}(k)$
   - $\sum_{\ell=1}^{L} n_T^{(\sigma(\ell))}(k) \leqslant \frac{\log T}{(\Delta^{(\sigma(L))}(k))^2} \xRightarrow{\text{Abel's summation}} \sum_{\ell=1}^{L} \Delta^{(\sigma(\ell))}(k) n_T^{(\sigma(\ell))}(k) \leqslant \frac{C \log T}{\Delta^{(\sigma(L))}(k)}$

Simulations (1/2): `FreeExp` vs. Baselines

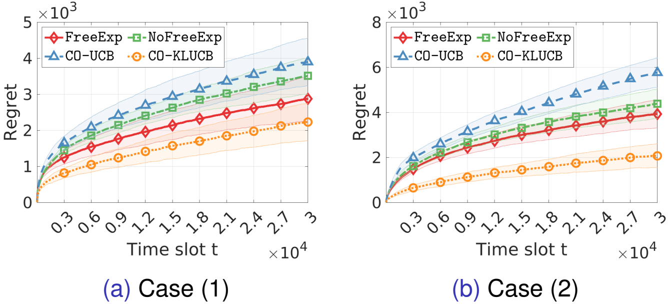
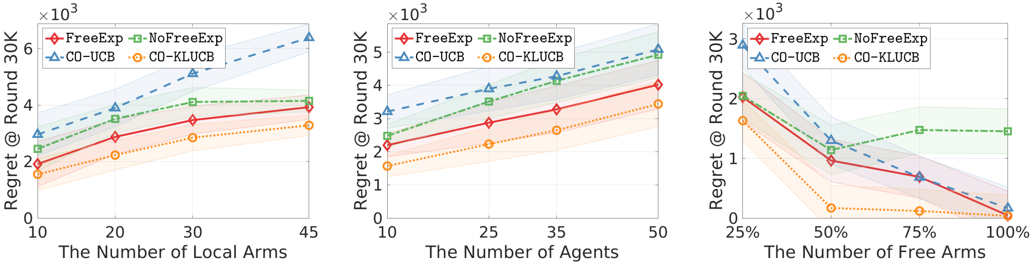


Figure 2: `FreeExp` *vs.* baselines

- Although with tighter theoretical performance, the empirical performance of `FreeExp` is not as good as `CO-KL-UCB`.

Simulations (2/2): Vary parameters of MA2B−HR



(a) Vary # local arms (b) Vary # agents (c) Vary % of free arms

Figure 3: Vary parameters of MA2B−HR

- In Figure (c), the more free arms, the better regret of `FreeExp`.

Conclusion

1. Discover the **free exploration mechanism** in multi-agent bandits with action constraints model.
2. Propose a new **regret lower bound**, echoing the free exploration mechanism.
3. Devise the **`FreeExp` algorithm** utilizing the free exploration mechanism.
4. Prove that `FreeExp`'s **regret upper bound** tightly matches the lower bound.
5. Conduct **simulations** to validate `FreeExp`'s empirical performance.

Conclusion

1. Discover the **free exploration mechanism** in multi-agent bandits with action constraints model.
2. Propose a new **regret lower bound**, echoing the free exploration mechanism.
3. Devise the **`FreeExp` algorithm** utilizing the free exploration mechanism.
4. Prove that `FreeExp`'s **regret upper bound** tightly matches the lower bound.
5. Conduct **simulations** to validate `FreeExp`'s empirical performance.

Future works:

- Fairness among heterogeneous agents?
- Reduce communications from $O(T)$ to $O(\log T)$?

Thank you!

Full paper at openreview.net/pdf?id=8kKEz1bnIEp

References I

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Lin Yang, Yu-Zhen Janice Chen, Mohammad Hajiesmaili, John C.S. Lui, and Don Towsley. Distributed bandits with heterogeneous agents. In *In Proceedings of The IEEE International Conference on Computer Communications 2022*, 2022.