Cooperative and Stochastic Multi-Player Multi-Armed Bandit: Optimal Regret With Neither Communication Nor Collisions

Sébastien Bubeck (Microsoft Research), Thomas Budzinski (ENS Lyon), Mark Sellke (Stanford)

COLT 2021

K-Arm, m-Player Bandits

- Fix $\mathbf{p} = (p_1, p_2, \ldots, p_K) \in [0,1]^K$. Let $(r_t(i))_{1 \le i \le K, 1 \le t \le T}$ be independent variables with

$$\mathbb{P}\left(r_t(i) = 0\right) = 1 - p_i \quad \text{and} \quad \mathbb{P}\left(r_t(i) = 1\right) = p_i.$$

- At time t, each player $(P_X)_{X \in [m]}$ picks arm i_t^X *without communication*, and observes the reward:

$$r_t(X) = r_t(i_t^X) \cdot \mathbb{1}_{i_t^X \neq i_t^Y \ \forall Y \neq X}.$$

- Collisions $\to$ *no reward*.
- Regret: $R_T = \left(\sum_{t=1}^T \sum_{X=1}^m r_t(X) \right) - T\mathbf{p}^*$, where $\mathbf{p}^* = \max_{1 \le i_1 < \cdots < i_m \le K} \left(\sum_{j=1}^m p_{i_j} \right)$ is sum of the top m arms.
- Goal: find a (randomized) strategy minimizing $\max_{\mathbf{p}} \mathbb{E}[R_T]$.

Bounds on the minimax regret

- Some of the previous works:
  - Regret $\widetilde{O}(\sqrt{T})$, $p_1, p_2, p_3 \leq 1 - \varepsilon$ [Lugosi-Mehrabian 18].
  - Regret $\widetilde{O}(T^{1-\frac{1}{2m}})$, non-stochastic [Bubeck-Li-Peres-Sellke 19].
  - Regret $O\left(\sum_i \frac{\log(T)}{\Delta_i}\right)$ [Huang-Combes-Trinh 21].
- All "cheat" by using *collisions to implicitly communicate*.

Theorem (BBS 21)

There is a strategy (using public shared randomness) with

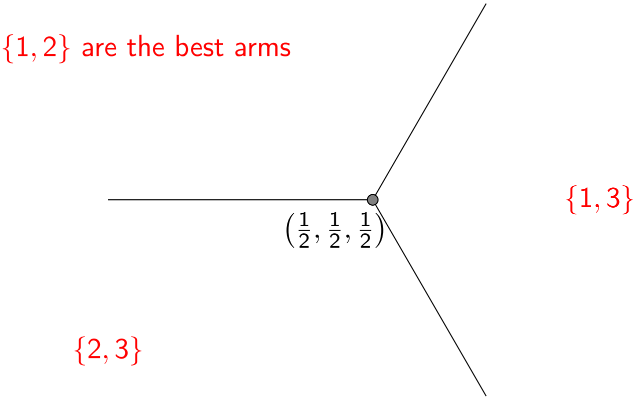
$$\max_{\mathbf{p}} \mathbb{E}[R_T] = O\left(mK^{11/2}\sqrt{T \log T}\right),$$

$$\mathbb{P}\,(\textit{there is a collision}) = O(T^{-2}).$$

- $(K, m) = (3, 2)$: $\Theta(\sqrt{T \log T})$ optimal [Bubeck-Budzinski 20].

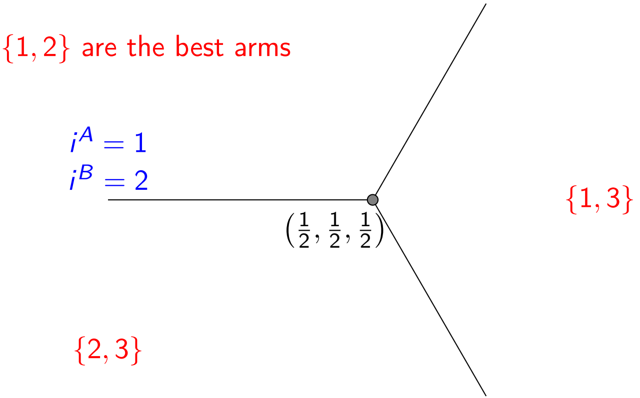
Topological Obstruction for 2 players, 3 arms

- Work in the plane $\{p_1 + p_2 + p_3 = \text{constant}\}$.
- No communication $\rightarrow$ can assume player strategies are functions of empirical average rewards.
- Topological obstruction: playing top 2 arms forces collision.



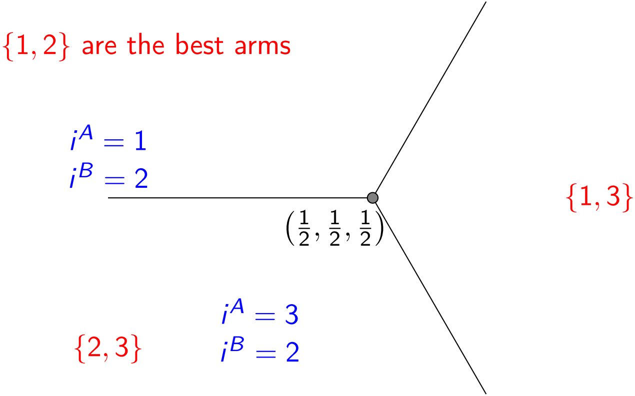
Topological Obstruction for 2 players, 3 arms

- Work in the plane $\{p_1 + p_2 + p_3 = \text{constant}\}$.
- No communication $\rightarrow$ can assume player strategies are functions of empirical average rewards.
- Topological obstruction: playing top 2 arms forces collision.



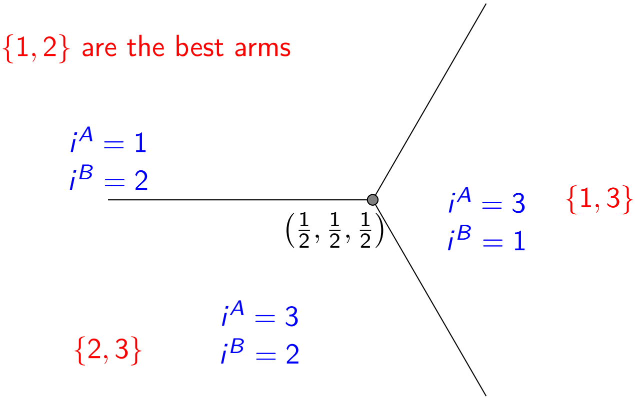
Topological Obstruction for 2 players, 3 arms

- Work in the plane $\{p_1 + p_2 + p_3 = \text{constant}\}$.
- No communication $\rightarrow$ can assume player strategies are functions of empirical average rewards.
- Topological obstruction: playing top 2 arms forces collision.



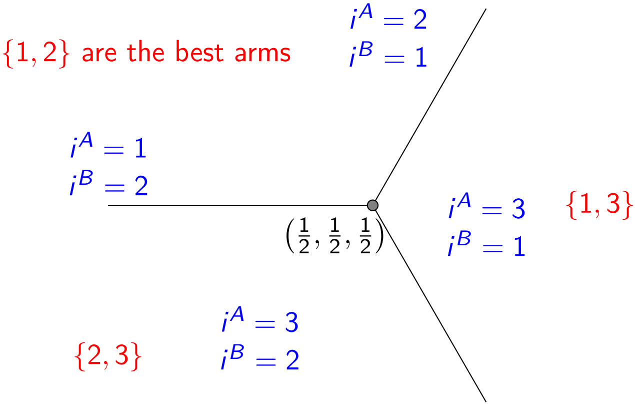
Topological Obstruction for 2 players, 3 arms

- Work in the plane $\{p_1 + p_2 + p_3 = \text{constant}\}$.
- No communication $\rightarrow$ can assume player strategies are functions of empirical average rewards.
- Topological obstruction: playing top 2 arms forces collision.



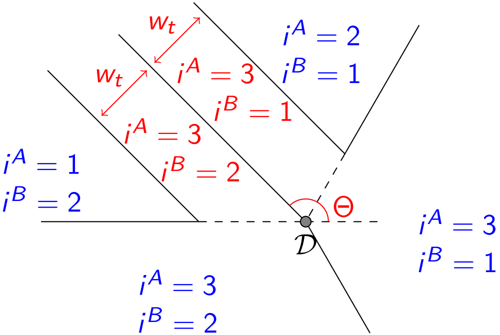
Topological Obstruction for 2 players, 3 arms

- Work in the plane $\{p_1 + p_2 + p_3 = \text{constant}\}$.
- No communication $\rightarrow$ can assume player strategies are functions of empirical average rewards.
- Topological obstruction: playing top 2 arms forces collision.



Collision-Free Solution for 2 players, 3 arms

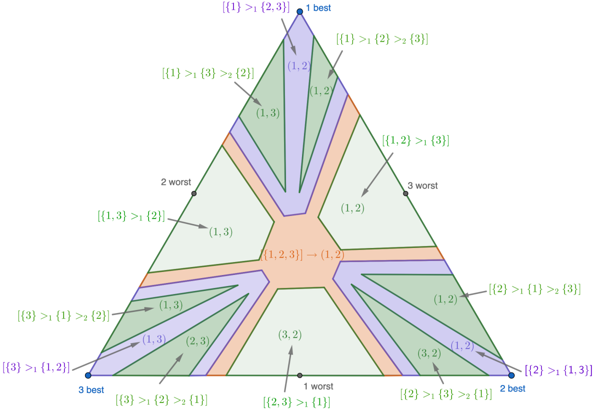
- Idea ([BB 20]): create interface between regions $\{i^A = 1, i^B = 2\}$, $\{i^A = 2, i^B = 1\}$.



- Label interface to avoid adjacent collisions with $w_t \gtrsim \sqrt{\frac{\log T}{t}}$.
- Thin interface, random $\Theta \to$ regret $O(\sqrt{T \log T})$.
- General (K, m) needs a high-dimensional analog.

General Strategy

- New partition in the case $(K, m) = (3, 2)$:



- Regions form a tree, defined by arm inequalities added *in order*.
- Example region: $\{1, 3, 5\} >_2 \{4, 8\} >_3 \{2, 6\} >_1 \{7, 9, 10\}$.
- Via shared randomness, map regions → arms w/o collision.

Some Details of the Partition

- Inequalities always separate arms that *might* be in top m. Once top m vs bottom $K - m$ is determined, stop.
- Example for $(K, m) = (10, 5)$:

$$\begin{aligned} &\{1,2,3,4,5,6,7,8,9,10\} \\ \to&\{1,2,3,4,5,6,8\} >_1 \{7,9,10\} \\ \to&\{1,3,5\} >_2 \{2,4,6,8\} >_1 \{7,9,10\} \\ \to&\{1,3,5\} >_2 \{4,8\} >_3 \{2,6\} >_1 \{7,9,10\}. \end{aligned}$$

- Generalization of random interface: stop early if gap size for new inequality lies in a small random interval.
- Each player needs to input estimate of $\mathbf{p}$, output region.
- Main step: pick 1 of $\leq K$ cuts. Efficient despite $\approx K!$ regions.

THANK YOU !