Coordination without communication: optimal regret in two players multi-armed bandits

Sébastien Bubeck (Microsoft) and Thomas Budzinski (UBC)

COLT 2020

Two players stochastic three-armed bandits

- Fix $\mathbf{p} = (p_1, p_2, p_3) \in [0,1]^3$. Let $(\ell_t(i))_{1 \le i \le 3, 1 \le t \le T}$ be independent variables with

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

- At time t, player A (resp. B) picks arm i_t^A (resp. i_t^B) *without to communicate*, and observes the loss:

$$\mathbb{1}_{i_t^A = i_t^B} + \mathbb{1}_{i_t^A \neq i_t^B} \, \ell_t(i_t^A) \quad (\text{resp. } \ell_t(i_t^B)).$$

- Regret: $R_T = \sum_{t=1}^T \left(2 \cdot \mathbb{1}_{i_t^A = i_t^B} + \mathbb{1}_{i_t^A \neq i_t^B} (p_{i_t^A} + p_{i_t^B}) - \mathbf{p}^* \right)$, where $\mathbf{p}^* = \min(p_1 + p_2, p_2 + p_3, p_3 + p_1)$.
- Goal: find a randomized strategy such that $\max_{\mathbf{p}} \mathbb{E}[R_T]$ is as small as possible.

Bounds on the minimax regret

- Some of the previous works:
  - Regret $\widetilde{O}(T^{3/4})$ [Bubeck–Li–Peres–Sellke 2019] (2 players, k arms, not restricted to stochastic bandits).
  - Regret $\widetilde{O}(\sqrt{T})$ for p_1, p_2, p_3 bounded away from 1 [Lugosi–Mehrabian 2018] (m players, k arms, stochastic).
- Both "cheat" by using *collisions as an implicit form of communication*.

Theorem (BB. 2020)

There is a randomized strategy (using shared randomness) such that

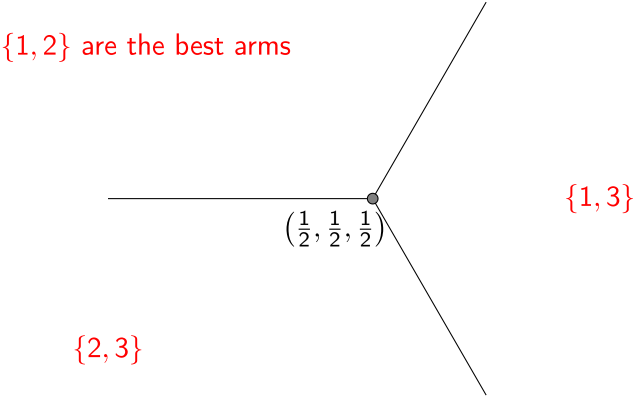
$$\max_{\mathbf{p}} \mathbb{E}[R_T] = O\left(\sqrt{T \log T}\right)$$

and

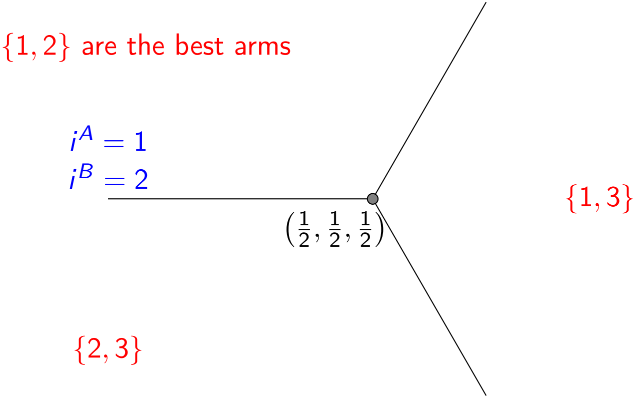
$$\mathbb{P}\left(\textit{there is at least a collision}\right) = o(1).$$

Why not $\sqrt{T}$?

- We work in the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$.

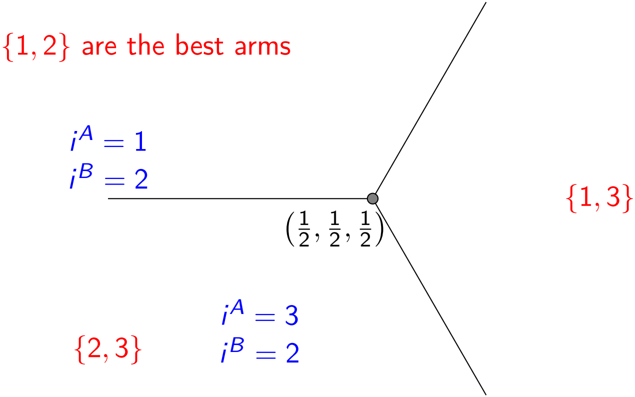


- We work in the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$.



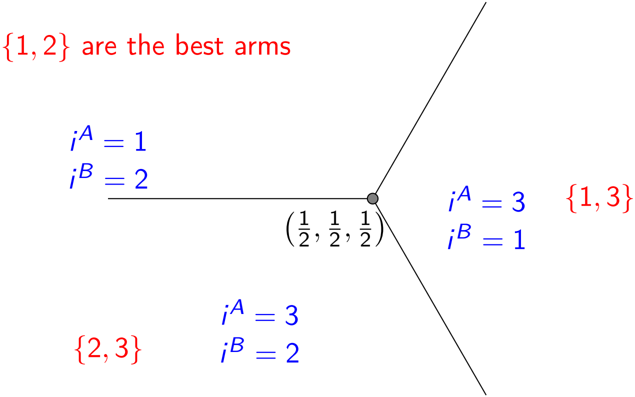
Why not $\sqrt{T}$?

- We work in the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$.



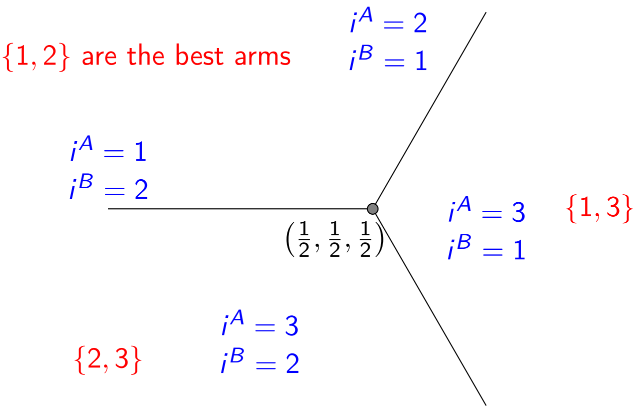
Why not $\sqrt{T}$?

- We work in the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$.



Why not $\sqrt{T}$?

- We work in the plane $\left\{p_1 + p_2 + p_3 = \frac{3}{2}\right\}$.



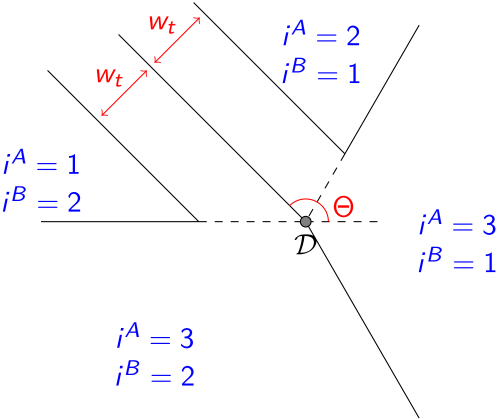
- Topological obstruction: it is not possible to always play what seems best.

A full-information toy model

- To isolate the problem of collisions from the usual *exploration vs exploitation* trade-off, we look at a full information toy model:
  - Fix $\mathbf{p} = (p_1, p_2, p_3) \in [0,1]^3$.
  - $\left(\ell_t^A(i), \ell_t^B(i)\right)_{1 \leq i \leq 3, 1 \leq t \leq T}$ are independent with $\mathbb{P}\left(\ell_t^X(i) = 0\right) = 1 - p_i$ and $\mathbb{P}\left(\ell_t^X(i) = 1\right) = p_i$.
  - At time t, player A picks i_t^A and observes $\left(\ell_t^A(1), \ell_t^A(2), \ell_t^A(3)\right)$ (even if there is a collision), and similarly for B.
  - Regret computed as in the bandit model.
- No way to use collisions to communicate!
- Using the "topological obstruction", we prove that the minimax regret for the toy model is $\Omega\left(\sqrt{T \log T}\right)$.

Strategy for the toy model

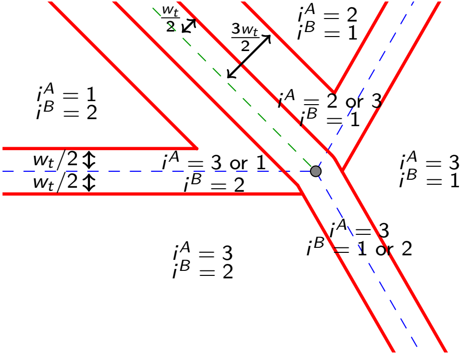
- Idea: introduce a random "interface" between the regions $\{i^A = 1, i^B = 2\}$ and $\{i^A = 2, i^B = 1\}$.



- Here $w_t = 100\sqrt{\frac{\log T}{t}}$ and $\Theta \sim \mathrm{Unif}\left(\left[\frac{\pi}{3}, \pi\right]\right)$.

The bandit strategy

- Similar to the one for the toy model, but each player needs to have some information about every arm.
  - Close to a boundary, explore both possibilities. E.g. near the boundary between $\{i^A = 2, i^B = 1\}$ and $\{i^A = 3, i^B = 1\}$, player A alternates between arms 2 and 3).
  - Players alternate roles regularly so each has a reasonable estimate of each arm.



Which assumptions are necessary?

- Is shared randomness necesssary? No by a different strategy, but then we lose the non-collision property.
- Are we limited to 2 players and 3 arms? Work in progress. The geometric picture becomes more complicated.

THANK YOU !