Coordination without communication in two players multi-armed bandits

Thomas Budzinski (CNRS and ENS Lyon)
Joint works with Sébastien Bubeck (Microsoft Research) and Mark Sellke (Stanford)

March 4th, 2021
Lyon Probability Seminar

Stochastic three-armed bandits

- Let $T \geq 1$ fixed, and let $\mathbf{p} = (p_1, p_2, p_3) \in [0, 1]^3$ be unknown from the player.
- Loss functions: let $(\ell_t(i))_{1 \leq i \leq 3,\, 1 \leq t \leq T}$ be independent variables with

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

- At each step t, the player chooses an arm i_t and receives the loss $\ell_t(i_t)$.
- Regret: $R_T = \left(\sum_{t=1}^{T} \ell_t(i_t)\right) - \mathbf{p}^* T$, where $\mathbf{p}^* = \min\left(p_1, p_2, p_3\right)$.
- Goal: find a strategy for which $\max_{\mathbf{p}} \mathbb{E}[R_T]$ is small.

Stochastic three-armed bandits

- Motivations: clinical trials, online advertising...
- Two settings:
  - Full information: at time t, the player observes $(\ell_t(1), \ell_t(2), \ell_t(3))$.
  - Bandits: at time t, the player only observes $\ell_t(i_t)$.
- In both settings, the minimax expected regret is of order $\sqrt{T}$:
  - If $|p_1 - p_2| \approx \frac{1}{\sqrt{T}}$, difficult to distinguish the best arm with T observations.
  - Full information strategy: follow the best arm.
  - Bandit strategy: explore everything at the beginning, discard an arm when it is significantly behind others.

Two players stochastic three-armed bandits

- Again: $T \geq 1$, a vector $\mathbf{p} = (p_1, p_2, p_3)$ and $\ell_t(i)$ are independent Bernoulli with parameter p_i.
- Two players A and B. At time t, player A (resp. B) picks arm i_t^A (resp. i_t^B), with *no communication between players*.
- Collisions are penalized: player A (resp. B) observes the loss:

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

- Regret:
$R_T = \sum_{t=1}^{T} \left(2 \cdot \mathbb{1}_{i_t^A = i_t^B} + \mathbb{1}_{i_t^A \neq i_t^B} \left(\ell_t(i_t^A) + \ell_t(i_t^B) \right) \right) - \mathbf{p}^* T,$
where $\mathbf{p}^* = \min \left(p_1 + p_2, p_2 + p_3, p_3 + p_1 \right)$.
- Again, we want to minimise $\max_{\mathbf{p}} \mathbb{E}[R_T]$.

Two players stochastic three-armed bandits

- Motivations:
  - Situations where gains on an arm have to be "shared" between the players who played this arm.
  - Cognitive radios (finding available channels).
- Naive algorithms:
  - *A* plays the best arms and *B* the second best? But then what if $p_1 = p_2 << p_3$?
  - *A* plays preferably arm 1 and *B* plays preferably arm 3? Then what if $p_2 << p_1 = p_3$?

Bounds on the minimax regret

- Some of the previous works:
  - Regret $\widetilde{O}(\sqrt{T})$ for p_1, p_2, p_3 bounded away from 1 [Lugosi–Mehrabian 2018] (m players, k arms, stochastic).
  - Regret $\widetilde{O}(T^{3/4})$ [Bubeck–Li–Peres–Sellke 2019] (2 players, k arms, works for adversarial bandits).
- Both "cheat" by using *collisions as an implicit form of communication*.

Theorem (Bubeck–B. 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

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

A full-information toy model

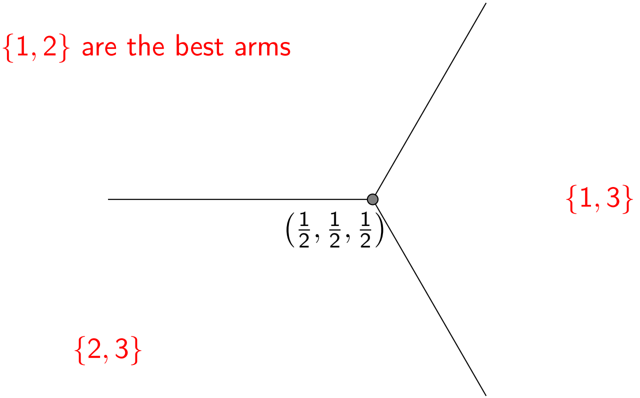
- To isolate the problem of collisions from the usual *exploration vs exploitation* trade-off, we look at a full information toy model:
  - Fix $\mathbf{p} = (p_1, p_2, p_3) \in [0,1]^3$.
  - $\big(\ell_t^A(i), \ell_t^B(i)\big)_{1 \le i \le 3,\, 1 \le t \le T}$ are independent Bernoulli with parameter p_i.
  - At time t, player A picks i_t^A and observes $\big(\ell_t^A(1), \ell_t^A(2), \ell_t^A(3)\big)$ (even if there is a collision), and similarly for B.
  - Regret computed as in the 2-player bandit model.
- No way to use collisions to communicate!

Theorem (Bubeck–B. 2020)

In the full-information toy model, the minimax expected regret is at least $c\sqrt{T \log T}$.

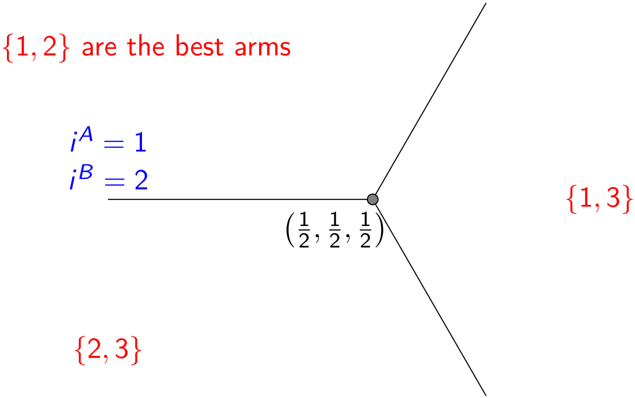
Why not $\sqrt{T}$? A topological obstruction

- We represent the set of possible **p** (restricted to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$).



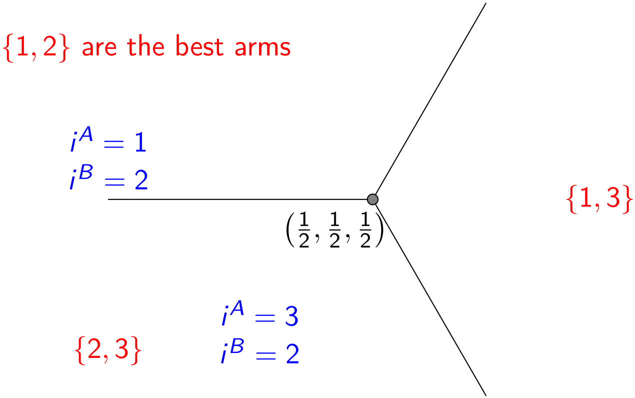
Why not $\sqrt{T}$? A topological obstruction

- We represent the set of possible $\mathbf{p}$ (restricted to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$).



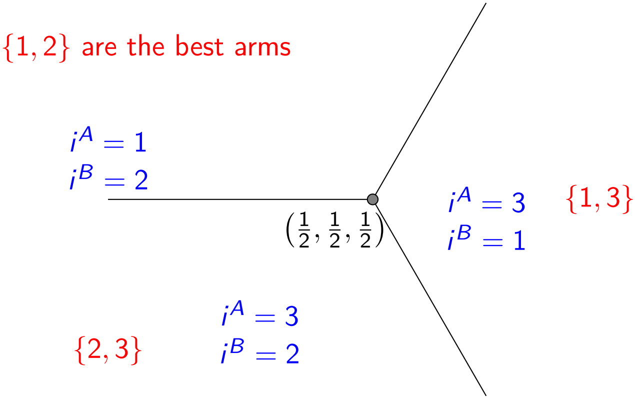
Why not $\sqrt{T}$? A topological obstruction

- We represent the set of possible $\mathbf{p}$ (restricted to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$).



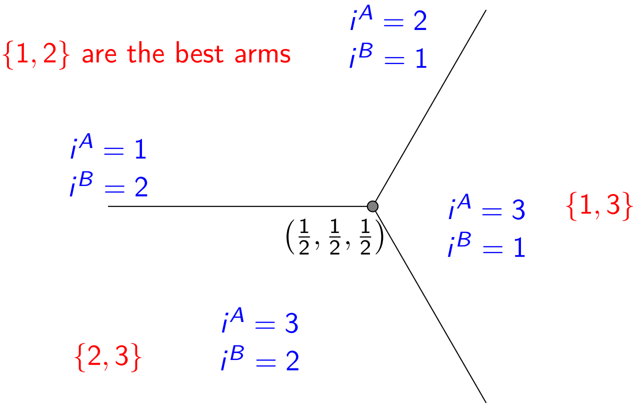
Why not $\sqrt{T}$? A topological obstruction

- We represent the set of possible $\mathbf{p}$ (restricted to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$).



Why not $\sqrt{T}$? A topological obstruction

- We represent the set of possible **p** (restricted to the plane $\{p_1 + p_2 + p_3 = \frac{3}{2}\}$).

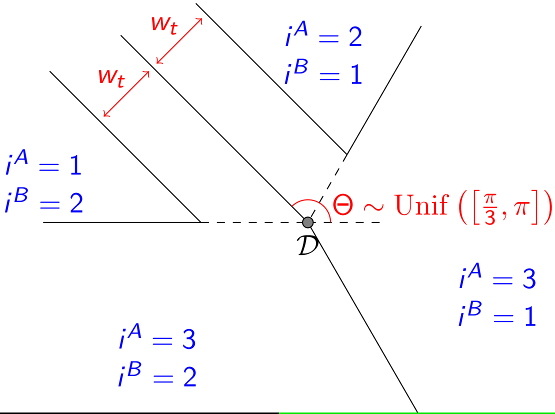


Why not $\sqrt{T}$? A topological obstruction

- Topological obstruction: it is not possible to always play what seems best.
- To fix this:
  - either take the risk of a collision in the $\{1, 2\}$ region (very costly),
  - or do a suboptimal play to pass "smoothly" from $\{i^A = 1, i^B = 2\}$ to $\{i^A = 2, i^B = 1\}$.
- The second option is less costly, provided the location of the suboptimal play is *randomized*.

Strategy for the toy model

- Let $\mathbf{q}_t^A = \left(\frac{1}{t-1}\sum_{s=1}^{t-1} \ell_s^A(i)\right)_{1\leq i\leq 3}$ be the estimate of $\mathbf{p}$ at time t according to A (and similarly define $\mathbf{q}_t^B$).
- Then A (resp. B) plays according to the position of $\mathbf{q}_t^A$ (resp. $\mathbf{q}_t^B$) in the following diagram (where $w_t = 100\sqrt{\frac{\log T}{t}}$):



Sketch of proof for the toy model

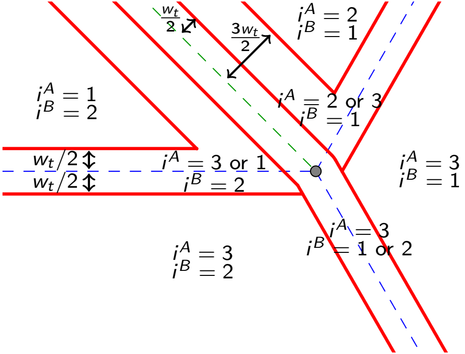
- Collisions are not possible between neighbour regions, so to have a collision, a player must make an error of more than $\frac{w_t}{2}$.
- So by Hoeffding:

$$\mathbb{P}(\text{collision}) \leq \mathbb{P}\left(\text{error} \geq \frac{w_t}{2}\right) \leq \exp\left(-\frac{t}{2}\left(\frac{w_t}{2}\right)^2\right) \leq T^{-50}.$$

- The loss caused by a suboptimal play in the interface is $O\left(d(\mathbf{p}, \mathcal{D})\right)$.
- The interface is at a random angle, so the probability to be in the interface is $O\left(\frac{w_t}{d(\mathbf{p},\mathcal{D})}\right)$.
- So the total expected loss is $O\left(\sum_{t=1}^{T} w_t\right) = O(\sqrt{T \log T})$.

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.



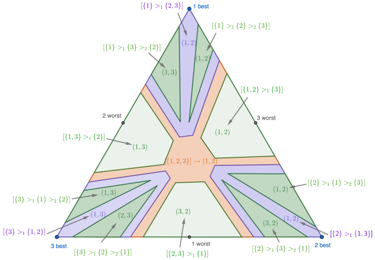
More arms, more players

Theorem (Bubeck–B–Sellke. 2020)

For multiplayers multi-armed bandits with m players and $K \geq m$ arms, there is a randomized strategy with no collision at all with high probability and

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

- Similar ideas, but the geometric picture is much more complicated:



THANK YOU !