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 \( p = (p_1, p_2, p_3) \in [0, 1]^3 \). Let \((\ell_t(i))_{1 \leq i \leq 3, 1 \leq t \leq T}\) 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 \( i^A_t \) (resp. \( i^B_t \)) \textit{without to communicate}, and observes the loss:
  \[
  1_{i^A_t = i^B_t} + 1_{i^A_t \neq i^B_t} \ell_t(i^A_t) \quad \text{(resp.} \ell_t(i^B_t)).
  \]

- Regret: \( R_T = \sum_{t=1}^{T} \left( 2 \cdot 1_{i^A_t = i^B_t} + 1_{i^A_t \neq i^B_t}(p_{i^A_t} + p_{i^B_t}) - p^* \right), \)
  where \( p^* = \min(p_1 + p_2, p_2 + p_3, p_3 + p_1) \).

- Goal: find a randomized strategy such that \( \max_{p} \mathbb{E}[R_T] \) is as small as possible.
Some of the previous works:

- Regret $\tilde{O}(T^{3/4})$ [Bubeck–Li–Peres–Sellke 2019] (2 players, $k$ arms, not restricted to stochastic bandits).
- Regret $\tilde{O}(\sqrt{T})$ 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

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

and

$$\mathbb{P}(\text{there is at least a collision}) = o(1).$$
Why not $\sqrt{T}$?

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

\{1, 2\} are the best arms

$\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$

\{2, 3\}

\{1, 3\}
Why not $\sqrt{T}$?

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

\{1, 2\} are the best arms

$i^A = 1$
$i^B = 2$

$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$

\{1, 3\}

\{2, 3\}
Why not $\sqrt{T}$?

We work in the plane \( \{ p_1 + p_2 + p_3 = \frac{3}{2} \} \).

\( \{1, 2\} \) are the best arms

\[ i^A = 1 \]
\[ i^B = 2 \]

\( \{2, 3\} \)

\[ i^A = 3 \]
\[ i^B = 2 \]
Why not $\sqrt{T}$?

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

$\{1, 2\}$ are the best arms

- $i^A = 1$
- $i^B = 2$

$\{1, 3\}$

- $i^A = 3$
- $i^B = 1$

$\{2, 3\}$

- $i^A = 3$
- $i^B = 2$
Why not $\sqrt{T}$?

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

- $\{1, 2\}$ are the best arms

- $\{2, 3\}$

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

$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$

\[\begin{align*}
i^A &= 2 \\
i^B &= 1 \\
i^A &= 3 \\
i^B &= 1
\end{align*}\]
To isolate the problem of collisions from the usual exploration vs exploitation trade-off, we look at a full information toy model:

- Fix $p = (p_1, p_2, p_3) \in [0, 1]^3$.
- $(\ell^A_t(i), \ell^B_t(i))_{1 \leq i \leq 3, 1 \leq t \leq T}$ are independent with $\mathbb{P}(\ell^X_t(i) = 0) = 1 - p_i$ and $\mathbb{P}(\ell^X_t(i) = 1) = p_i$.
- At time $t$, player $A$ picks $i_t^A$ and observes $(\ell^A_t(1), \ell^A_t(2), \ell^A_t(3))$ (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 $\Omega\left(\sqrt{T \log T}\right)$. 

S. Bubeck, T. Budzinski
• Idea: introduce a random "interface" between the regions \( \{ i^A = 1, i^B = 2 \} \) and \( \{ i^A = 2, i^B = 1 \} \).

\[
\begin{align*}
    i^A &= 1 \\
    i^B &= 2 \\
    i^A &= 2 \\
    i^B &= 1 \\
    i^A &= 3 \\
    i^B &= 1 \\
    i^A &= 3 \\
    i^B &= 2 \\

\end{align*}
\]

• Here \( w_t = 100 \sqrt{\frac{\log T}{t}} \) and \( \Theta \sim \text{Unif} \left( \left[ \frac{\pi}{3}, \pi \right] \right) \).
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 necessary? 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!