SIC - MMAB: Synchronisation Involves Communication in Multiplayer Multi-Armed Bandits

Etienne Boursier

ENS Paris-Saclay

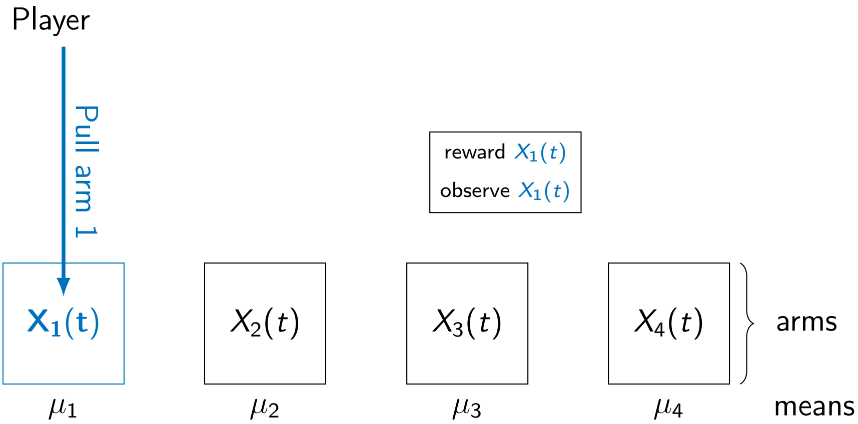
Vianney Perchet

ENSAE Paris
Criteo AI Lab

NeurIPS 2019, Vancouver

stochastic bandit game at round $t \in \{1, \dots, T\}$

K arms



Player

Pull arm 1

reward $X_1(t)$

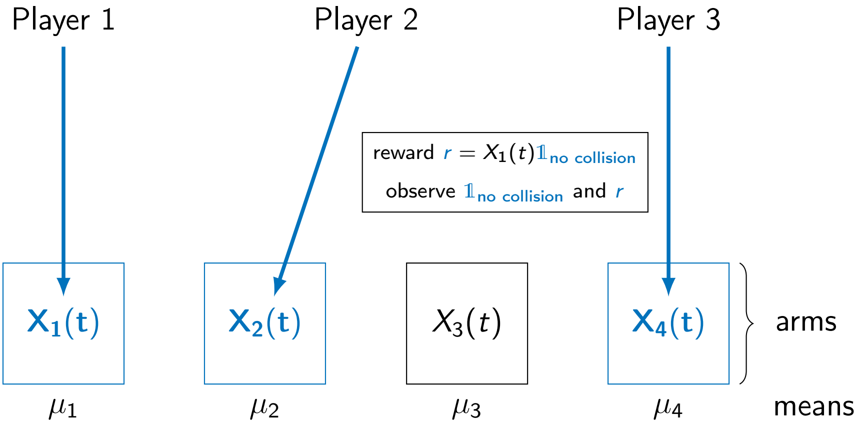
observe $X_1(t)$

$\mathbf{X_1(t)}$ $X_2(t)$ $X_3(t)$ $X_4(t)$ arms

μ_1 μ_2 μ_3 μ_4 means

Multiplayer stochastic bandit game at round $t \in \{1, \dots, T\}$

K arms, M players



Player 1

Player 2

Player 3

reward $r = X_1(t)\mathbb{1}_{\text{no collision}}$

observe $\mathbb{1}_{\text{no collision}}$ and r

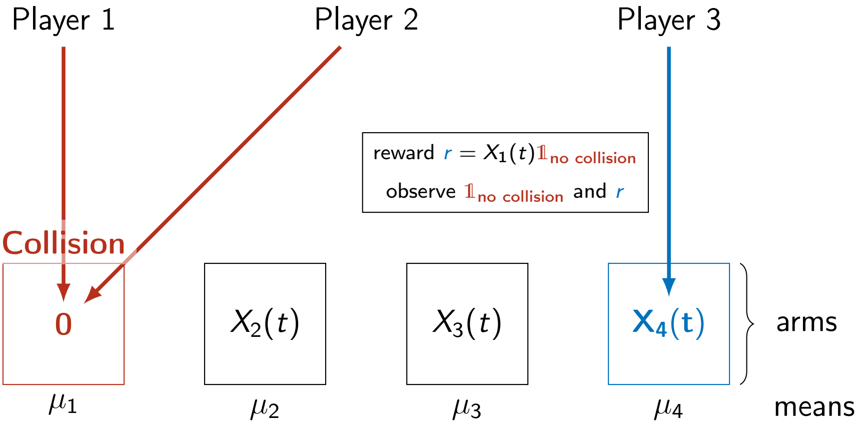
$\mathbf{X_1(t)}$ $\mathbf{X_2(t)}$ $X_3(t)$ $\mathbf{X_4(t)}$ arms

μ_1 μ_2 μ_3 μ_4 means

Motivated by cognitive radio networks (5G)

Multiplayer stochastic bandit game at round $t \in \{1, \dots, T\}$

K arms, M players



Motivated by cognitive radio networks (5G)

What is the best possible algorithm?

Performance measured in regret.
w.l.o.g. $\mu_1 > \mu_2 > \ldots > \mu_K$

Centralized model: a meta-agent controls all the players
→ Regret must scale as

$$\sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

Decentralized model: no communication between players
→ Regret must scale as [Liu and Zhao, 2010]

$$M \sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

What is the best possible algorithm?

Performance measured in regret.
w.l.o.g. $\mu_1 > \mu_2 > \ldots > \mu_K$

Centralized model: a meta-agent controls all the players
$\rightarrow$ Regret must scale as

$$\sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

Decentralized model: no communication between players
$\rightarrow$ SIC-MMAB scales as

$$\xcancel{M} \sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

Decentralized $\sim$ Centralized

How is it possible?

Observation: collision indicator in $\{0, 1\} \rightarrow$ a bit sent from one player to another

- Enable indirect communication between players
- Players exchange empirical means in binary
- Negligible communication cost
- *almost* centralized

```
Initialization: estimate $M$ and player rank $j$
for $p = 1, ..., \infty$ do
    Exploration: explore each arm $2^p$ rounds
    Communication: players exchange statistics using collisions
    if optimal arms found then enter exploitation phase
end
Exploitation phase: pull optimal arm until $T$
```

Toward a better model

Communication protocols abuse a loophole from the model.

Synchronisation: players all start at $\tau^j = 1$.
SIC-MMAB heavily depends on synchro.

Our claim: synchronisation assumption has to be removed
→ similar protocols not possible (?) in **dynamic model**

Dynamic Model

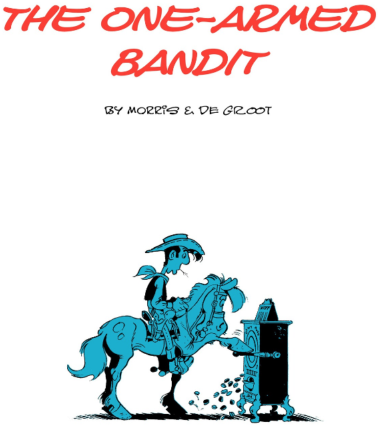
Setting:

- Players starting times τ^j: different and unknown
- Limited feedback: collision not observed, only the reward

DYN-MMAB: algorithm with logarithmic regret

- either sample uniformly at random (explore)
- or pull same arm until the end (exploit)

$\rightarrow$ simple algorithm, intricate analysis



Thank you!

Poster session: East Exhib. Hall B+C #11