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

Etienne Boursier\textsuperscript{1} \hspace{1cm} Vianney Perchet\textsuperscript{2, 3}

\textsuperscript{1} ENS Paris-Saclay, CMLA, Cachan, France
\textsuperscript{2} ENSAE, Palaiseau, France
\textsuperscript{3} Criteo AI Lab, Paris, France

NeurIPS 2019, Vancouver
Bandit game at round $t \in \{1, \ldots, T\}$

$K$ arms

Player

Pull arm 2

$X_1(t)$ \hspace{1cm} $X_2(t)$ \hspace{1cm} $X_3(t)$ \hspace{1cm} $X_4(t)$

$\mu_1$ \hspace{1cm} $\mu_2$ \hspace{1cm} $\mu_3$ \hspace{1cm} $\mu_4$

arms

means
Multiplayer Bandit game at round $t \in \{1, \ldots, T\}$

$K$ arms, $M$ players

Motivated by cognitive radio networks
Multiplayer Bandit game at round $t \in \{1, \ldots, T\}$
$K$ arms, $M$ players

Motivated by cognitive radio networks
What is the best possible algorithm?

Performance measured in **Regret**.

**Centralized model:** players share information.
→ Regret scales as

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

**Decentralized model:** no communication between players
→ Regret scales as

$$M \sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$ [Besson and Kaufmann, 2018]
What is the best possible algorithm?

Performance measured in Regret.

**Centralized model:** players share information.
→ Regret scales as
$$\sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

**Decentralized model:** no communication between players
→ SIC-MMAB scales as
$$\prod \sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$$

Decentralized $\sim$ Centralized
How is it possible?

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

- allows (coordinated) communication protocols between players
- players can share statistics
- problem becomes *almost* centralized

---

**Initialization Phase:** estimate \( M \) and player rank \( j \)

\[
\text{for } p = 1, \ldots, \infty \text{ do}
\]

- **Exploration phase:** explore each arm \( 2^p \) rounds
- **Communication phase:** players exchange statistics using collisions
  
  - if \textit{optimal arms found} then enter exploitation phase

\[
\text{end}
\]

**Exploitation phase:** pull optimal arm until \( T \)
Toward a *better* model

SIC-MMAB communication protocols abuse a loophole from the model.

**Synchronisation:** players start the game at the same time. SIC-MMAB depends on synchro.

**Our claim:** synchronisation assumption has to be removed → similar communications not possible in *dynamic model*
Dynamic Model

Setting:
- Players’ starting times differ
- Limited feedback: collision not observed, only the reward

DYN-MMAB: simple algorithm with logarithmic regret for dynamic model
- either pull arm chosen uniformly at random
- or pull same arm until the end (exploit)