Selfish Robustness and Equilibria in Multi-Player Bandits

Etienne Boursier

ENS Paris-Saclay

Vianney Perchet

ENSAE Paris
Criteo AI Lab

October 2020

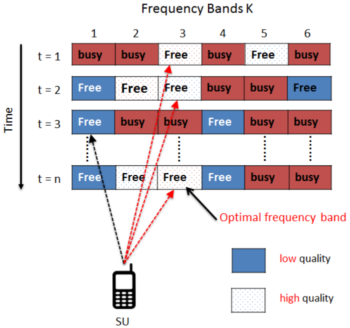
Motivation: Cognitive Radio

- licensed bands: Opportunistic Spectrum Access

arm $\leftrightarrow$ availability from primary users

- un-licensed bands: IoT communications

arm $\leftrightarrow$ background traffic



$\rightarrow$ what about **multiple devices**?

Stochastic bandits

K arms

Player

Pull arm 1

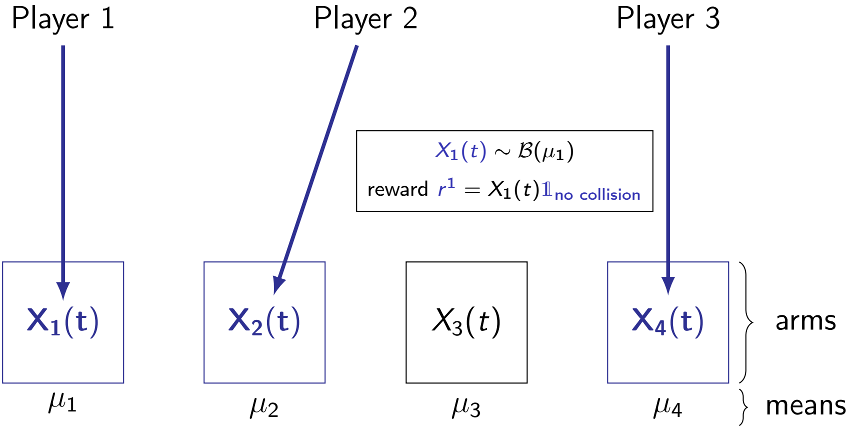
$X_1(t) \sim \mathcal{B}(\mu_1)$

reward $X_1(t)$

$\mathbf{X_1(t)}$	$X_2(t)$	$X_3(t)$	$X_4(t)$	arms
μ_1	μ_2	μ_3	μ_4	means

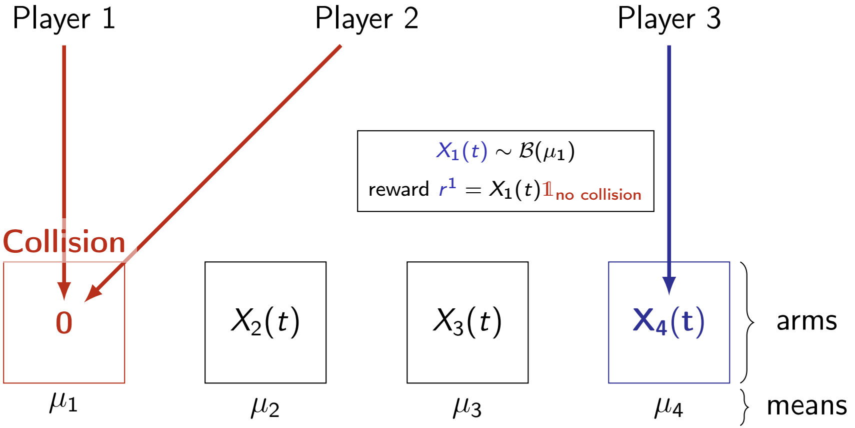
Stochastic bandits **[Multiplayer]**

K arms, M players



Stochastic bandits **[Multiplayer]**

K arms, M players



Model

M players pull arms $\pi^j(t)$; **Goal**: Maximize social welfare

Notation: $\mu_{(1)} > \mu_{(2)} > \ldots > \mu_{(K)}$

$$\textbf{Regret: } R_T = T\sum_{j=1}^{M}\mu_{(j)} - \sum_{j=1}^{M}\mathrm{Rew}_T^j$$

with $\mathrm{Rew}_T^j := \sum_{t=1}^T \mu^j_{\pi^j(t)} \mathbb{1}_{\text{no collision on } \pi^j(t)}$

Existing approaches: *Centralized case* or *Cooperative* players.

This paper: **selfish** players?

Centralized case

The benchmark

One single agent pulls M arms among K (combinatorial bandit)

Obviously: no collision.

Pull $M-1$ best empirical arms. **ucb** for the last one

- Finite regret from the $M-1$ best arms
- $\sum_{k>M} \frac{\mu_{(M)} - \mu_{(k)}}{\mathrm{kl}(\mu_{(M)}, \mu_{(k)})} \log(T)$ for the last one

$$\text{Regret} \leq \sum_{k>M} \frac{\mu_{(M)} - \mu_{(k)}}{\mathrm{kl}(\mu_{(M)}, \mu_{(k)})} \log(T) + o(\log(T))$$

Cooperative players – protocols

Different Sensings

After pull, reward $r_k(t) = X_k(t)(1 - \mathbb{1}_{\text{collision}})$, but agent observes

- Full sensing: $X_k(t) \in \{0, 1\}$ **and** $\mathbb{1}_{\text{collision}}$

  estimate μ_k and presence/absence of other agents

- No sensing: **Just** $r_k(t) \in \{0, 1\}$

  If $r_k(t) = 0$, collision or bad arm ?

- Stat. sensing: $X_k(t) \in \{0, 1\}$ **and** $r_k(t) \in \{0, 1\}$

  If $X_k(t) = 0$, collision or not ?

Emulate the centralized case

- Initialization: Estimate M, get "rank".

  Based on $\sharp$ collisions. **Finite cost**

  One player becomes the **leader**

  He will dictate the strategy to other players

- Explore/Exploit: Follow a centralized algorithm

  The leader makes all computations

- Communication: Collide on purpose to send a bit of info

  Report statistics to the leader/Get arm reco

  Almost costless: $\log^2 \log(T) = o(\log(T))$

- Regret: Same as centralized case

  With Full and Stat. sensing to observe collisions !

  No sensing: Extra **Multiplicative** factor M

- If all agents follow **scrupulously** the protocol !

Selfish players

Strategy/algo profile $(s', s_{-j}) := (s_1, \dots, \overbrace{s'}^{j}, \dots, s_M) \in \mathcal{S}^M$

Definition ε-Nash Equilibrium

$$\forall s' \in \mathcal{S}, \quad \mathbb{E}[\mathrm{Rew}^j_T(s', s_{-j})] \leq \mathbb{E}[\mathrm{Rew}^j_T(s)] + \varepsilon$$

- ε-gain from **unilateral** deviation

Definition (α, ε)-stability

For all $s' \in \mathcal{S}, i, j \in [M], \ell \in \mathbb{R}_+$:

$$\mathbb{E}[\mathrm{Rew}^i_T(s', s_{-j})] \leq \mathbb{E}[\mathrm{Rew}^i_T(s)] - \ell$$
$$\implies \mathbb{E}[\mathrm{Rew}^j_T(s', s_{-j})] \leq \mathbb{E}[\mathrm{Rew}^j_T(s)] + \varepsilon - \alpha\ell$$

- Cannot "hurt" someone else without "hurting" oneself
- ε-Nash equilibrium $\implies$ $(0, \varepsilon)$-stability

Existing protocols are not equilibria

- Communication: Selfish player can interfere

  By not communicating its statistics

  By improperly communicating its statistics

  By colliding while others are communicating (change bits)

- Fairness: Need strong symmetry/anonymity

  Algo a-priori fair **not** a-posteriori

  Selfish agent wants to the be the leader

- Omniscient selfish player

  Knows the values μ_k

  Knows the strategy of other players (the “normal” protocol)

Selfish-Robust MMAB

Statistic sensing: $X^t_{\pi^j(t)}$ and $r_j(t)$ observed

Emulate centralized independently

- Initialization: estimate M and get ranks
  - Small variant for robustness
- Explore/Exploit: blocks of size M:
  - pull $M - 1$ best empirical arms in a shifted way (no collision)
  - on remaining round $\begin{cases} \text{pull } M\text{-th best arm with probability } 1/2 \\ \text{explore at random otherwise} \end{cases}$
- Regret analysis. **M times** optimal regret
  - No collision if same empirical best arms ... all but finite number of times
- **Equilibrium !**
  - Estimating μ_k always possible.
  - Other players are occupying all but one of best $M - 1$ arms
  - Selfish can only spare its own regret

Selfish-Robust MMAB

Theoretical guarantees

Theorem (Selfish-Robust MMAB guarantees)

1. $\mathbb{E}[R_T] \leq M \sum_{k>M} \frac{\mu_{(M)}-\mu_{(k)}}{\mathrm{kl}(\mu_{(M)},\mu_{(k)})} \log(T) + \mathcal{O}\left(\frac{MK^3}{\mu_{(K)}} \log(T)\right)$,
2. *ε-Nash equilibrium and (ε, α)-stable with:*

$$\varepsilon = \frac{\mu_{(M)} - \mu_{(k)}}{\mathrm{kl}(\mu_{(M)}, \mu_{(k)})} \log(T) + \mathcal{O}\left(\frac{K^3 \mu_{(1)}}{\mu_{(K)}} \log(T)\right) \quad \textit{and} \quad \alpha = \frac{\mu_{(M)}}{\mu_{(1)}}$$

- **Optimal** without collision information [Besson and Kaufmann, 2019]
- α-stability. Collide with j by pulling 1 instead of M

No sensing - Impossibility

Only $r^j(t)$ observed

Th. There is **no** symmetric $o(T)-$Nash eq. s.t. $\mathbb{E}[R_T] = o(T)$

Proof.

- assume $\mu_1 > \mu_2 \ldots > \mu_K$ and $o(T)$ regret
- selfish player pulls arm 1 the whole time
  - others observe $(0, \mu_2, \ldots, \mu_K)$ and do not pull 1
- $\Omega(T)$-improvement for selfish player ■

Same arguments

- no $o(T)$-Nash eq. (non-symmetric) where $\mathbb{E}[R_T^j] = o(T)$

Reaching decentralized regret ?

Full sensing: Both $X_{\pi^j(t)}$ and $\mathbb{1}_{\text{collision}}$ observed

Th.: $\mathbb{E}[R_T] = \mathcal{O}\left(\sum_{k>M} \frac{1}{\mu_{(M)} - \mu_{(k)}} \log(T) + MK^2 \log(T)\right)$

Requires:

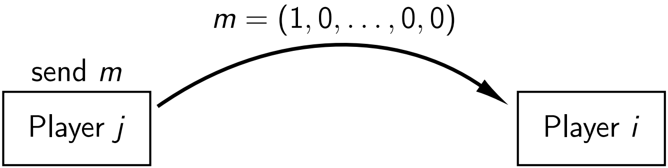
- A new "robust" initialization
  - Bi-partite leadership
- a new "robust" communication scheme.
  - Back and Forth messaging
- a new punishment protocol
  - Grim Trigger Strategies

Initialization

Bi-partite leadership

- Selfish players will try to be the leader
- Define **two leaders**
  - Each player reports statistics to both leaders
  - They check if statistics match & same updates
  - They **both** transmit recommendations to players
- Robust to **single** deviations
  - If s-selfish players : $s+1$ leaders
- **Fairness ?**
  - arms are exploited sequentially by all player (round robin)

Communication tricks
Back and forth

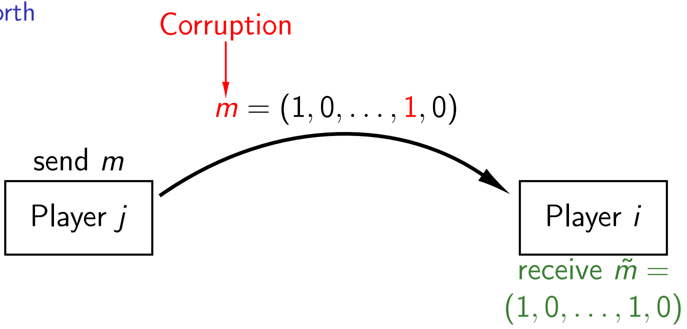


Communication

- j sends to i, $m_{i \to j} = (1, 0, \ldots, 0, 0)$ by pulling $(i, j, \ldots, j, j)$

Communication tricks

Back and forth

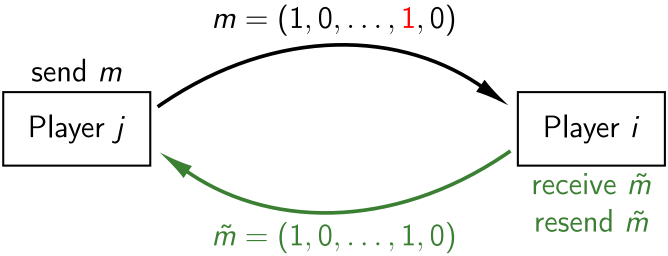


Communication

- j sends to i, $m_{i \to j} = (1, 0, \ldots, 0, 0)$ by pulling $(i, j, \ldots, j, j)$
- h can corrupt $m_{i \to j}$ by colliding $\to$ transform 0 in 1

Communication tricks

Back and forth

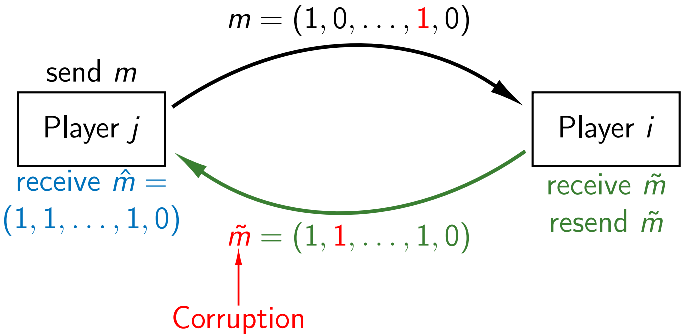


Communication

- j sends to i, $m_{i \to j} = (1, 0, \ldots, 0, 0)$ by pulling $(i, j, \ldots, j, j)$
- h can corrupt $m_{i \to j}$ by colliding $\to$ transform 0 in 1

Communication tricks

Back and forth

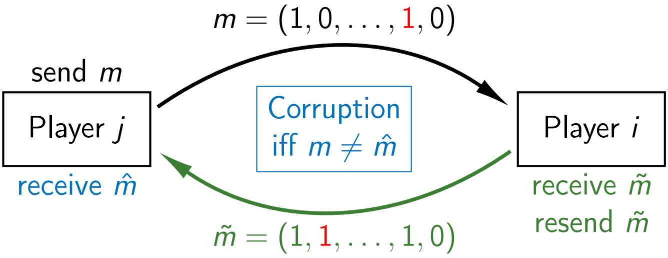


Communication

- j sends to i, $m_{i \to j} = (1, 0, \ldots, 0, 0)$ by pulling $(i, j, \ldots, j, j)$
- h can corrupt $m_{i \to j}$ by colliding $\to$ transform 0 in 1

Communication tricks

Back and forth



Communication

- j sends to i, $m_{i \to j} = (1, 0, \ldots, 0, 0)$ by pulling $(i, j, \ldots, j, j)$
- h can corrupt $m_{i \to j}$ by colliding $\to$ transform 0 in 1

Communication tricks

Punishment

Grim Trigger: Malicious player detected $\rightarrow$ punish until T. **How?**

- **1st idea:** sample any arm with probability $\frac{1}{K}$.
  - Selfish player gains $\mu_{(1)}(1 - 1/K)^{M-1}$
  - not enough, can be bigger than $\sum \mu_j / M$

Communication tricks

Punishment

Grim Trigger: Malicious player detected $\rightarrow$ punish until T. **How?**

- **1st idea:** sample any arm with probability $\frac{1}{K}$.
  - Selfish player gains $\mu_{(1)}(1-1/K)^{M-1}$
  - not enough, can be bigger than $\sum \mu_j/M$
- **2nd idea:** sample arm k with proba $\approx 1 - \left(\gamma \frac{\sum_{j=1}^{M} \mu_j}{M \mu_k}\right)^{\frac{1}{M-1}}$.
  - Selfish player gains $\approx \gamma \frac{\sum_{j=1}^{M} \mu_j}{M}$ on k.
  - **Relative loss** $1-\gamma$
  - Perfect! (for us). Admissible value: $\gamma = (1-\frac{1}{K})^{M-1}$

SIC-GT

Theoretical Guarantees

Theorem (SIC-GT guarantees)

1. $\mathbb{E}[R_T] = \mathcal{O}\left(\sum_{k>M} \frac{\log(T)}{\mu_{(M)}-\mu_{(k)}} + MK^2\log(T)\right)$
2. *ε-Nash equilibrium and (α, ε) stable with:*

$$\varepsilon = \mathcal{O}\left(\sum_{k>M} \frac{\log(T)}{\mu_{(M)}-\mu_{(k)}} + K^2\log(T) + \frac{K\log(T)}{\alpha^2\mu_{(K)}}\right)$$

$$\text{and } 2\alpha = 1 - (1-1/K)^{M-1}$$

Heterogeneous setting

Impossibility result

Heterogeneous : μ_k^j different among players

Find best matching: $W^* = \max_{\sigma} \sum_{i=1}^{M} \mu^i_{\sigma(i)}$

Theorem (Heterogeneous Full sensing)

There is no $o(T)$-Nash equilibrium such that $\mathbb{E}[R_T] = o(T)$.

Theorem [Zhou, 1990]There is no symmetric, Pareto optimal and strategy-proof random assignment algorithm.

- Assume that $\mu_1^{\text{selfish}} = \frac{1}{2} > \mu_k^{\text{selfish}}$ and
- Optimal matching σ^*
  - Arm 1 is allocated to another player (not selfish)
  - Total utility $W^* = \max_{\sigma^*} \sum_{i=1}^{M} \mu^i_{\sigma^*(i)}$
- Best matching $\hat{\sigma}$ giving arm 1 to selfish
  - Total utility $\hat{W} = \sum_{i=1}^{M} \mu^i_{\hat{\sigma}(i)}$
- Non strategy-proof if $W^* \leq \hat{W} + \frac{1}{3}$ (or $\leq \hat{W} + \frac{1}{2} - \eta$)
  - Report/act as if $\mu_1^{\text{selfish}} = 1$ and $\mu_k^{\text{selfish}} = 0$
  - "Optimal" allocation **becomes** $\hat{W}$
- Regret $R_T \simeq (W^* - \hat{W})T$.

Random Serial Dictatorship

(RSD) Symmetric & strategy-proof [Abdulkadiroglu and Sonmez, 1998]

- Choose dictator ordering σ at random
- $\sigma(1)$ chooses her preferred arm, $\sigma(2)$ her preferred remaining...
- **Not efficient** (i.e., welfare max)

$$\textbf{RSD-regret: } R_T^{\mathrm{RSD}} = T\mathbb{E}_\sigma\left[\sum_{k=1}^{M} \mu_{\pi_\sigma(k)}^{\sigma(k)}\right] - \sum_{j=1}^{M} \mathrm{Rew}_T^j$$

where $\pi_\sigma(k)$ = arm attributed to $\sigma(k)$ when order of dictators is σ.

RSD-GT

Description

- Initialization: estimate M and attribute ranks (order σ)
- Exploration: pull all arms
  - End when M-best arms identified
  - Signal it to others and exploit
- Exploitation: M blocks
  - Block k, order is $\sigma_0^k \circ \sigma$ where $\sigma_0 = \text{cycle}\,(1, \ldots, M)$.
  - **Cycles over permutations.**

    No benefit from initialization rank and σ (robustness)
- **Malicious behavior** detected $\rightarrow$ punishment protocol
  - δ-heterogeneous: for all j, k: $\mu_k^j \in [(1-\delta)\mu_k, (1+\delta)\mu_k]$

    Needed for punishment (selfish player unidentified)

RSD-GT

$\Delta = \min_{j,k<M} \mu_k^j - \mu_{k+1}^j$, $2r = 1 - \left(\frac{1+\delta}{1-\delta}\right)^2 (1-1/K)^{M-1}$

Theorem (δ-heterogeneous)

1. $\mathbb{E}[R_T^{\mathrm{RSD}}] = \mathcal{O}\left(\frac{MK}{\Delta^2}\log(T) + MK^2\log(T)\right)$
2. *ε-Nash equilibrium and (α, ε)-stable with*

$$\varepsilon = \mathcal{O}\left(\frac{K\log(T)}{\Delta^2} + K^2\log(T) + \frac{K\log(T)}{(1-\delta)r^2\mu_{(K)}}\right)$$

$$\alpha = \min\left(r\left(\frac{1+\delta}{1-\delta}\right)^3 \frac{\sqrt{\log(T)} - 4M}{\sqrt{\log(T)} + 4M}; \ \frac{1}{(1+\delta)}\frac{\Delta}{\mu_{(1)}}; \ \frac{(1-\delta)}{(1+\delta)}\frac{\mu_{(M)}}{\mu_{(1)}}\right)$$

- For stability, **random inspections** during exploitation
  - Selfish misreports μ_k^{selfish} to hurt j (if $\mu_{(1)}^{\text{selfish}}$ still available)
  - With proba $\frac{\sqrt{\log(T)}}{T}$, check if other players are well behaving

Recap

- **Upsides**
  - robust algorithms for many settings
  - impossibility result for no sensing and heterogeneous settings
  - centralized like regret still achievable
- **Downsides**
  - Rely on strong assumption: synchronicity - stationarity
  - Players arrive and leave in "real life"

    Bottleneck: stream-Evaluation of M
  - Coalitions of selfish players (using the same providers)

Thank you!

References

Abdulkadiroglu, A. and Sonmez, T. (1998).
Random serial dictatorship and the core from random endowments in house allocation problems.
Econometrica, 66(3):689.

Besson, L. and Kaufmann, E. (2019).
Lower bound for multi-player bandits: Erratum for the paper multi-player bandits revisited.

Boursier, E. and Perchet, V. (2019).
SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits.
NeurIPS.

Rosenski, J., Shamir, O., and Szlak, L. (2016).
Multi-player bandits–a musical chairs approach.
In *International Conference on Machine Learning*, pages 155–163.

Zhou, L. (1990).
On a conjecture by Gale about one-sided matching problems.
Journal of Economic Theory, 52(1):123–135.