Multi-Player Bandits Revisited

Workshop on MAB and Learning Algorithms, May 2018, Rotterdam (NL)
PhD Student Day, June 2018, Vannes (Fr)

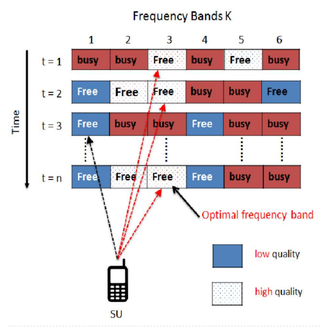
By: **Lilian Besson** – Lilian.Besson@CentraleSupelec.fr
SCEE Team, CentraleSupélec Rennes, IETR & SequeL Team, Inria Lille

1. Introduction & Goal

Goal: insert objects in a wireless network, keep a good *Quality of Service*.

- *Hypothesis*: object j choose channel $A_j(t) \in \{1, \dots, K\}$, to use to communicate at time t.
- *Idea*: use on-line **Machine Learning algorithms ?**
- *Not so easy*: each device takes its own decisions, without central control or communication, has light CPU/memory etc.
- $\Longrightarrow$ *Solution*: **Decentralized MAB algorithms !**

2. Model: Time/Frequency protocol



Fix-duration communication. Channels can be free or busy.

M players in the network

- K RF channels (of same bandwidth), *e.g.*, $K = 9$.
- Primary users ☎ create a background traffic.
- Channel k is busy with mean $\mu_1, \dots, \mu_K \in [0, 1]$.
- Sensing gives binary feedback $Y_{k,t} \sim \mathcal{B}(\mu_k)$.
- $1 \leq M \leq K$ secondary users 📱, $j \in \{1, \dots, M\}$.
- Base station 📡 replies if 📱 packet is received.
- Collision indic. $C^j(t) = \mathbb{1}(\exists j' \neq j, A^j(t) = A^{j'}(t))$.
- Non-stochastic $0/1$ reward $r^j(t) := Y_{A^j(t),t} \times C^j(t)$.

3. Previous works (since 2009)

Ideas: 1) use empirical means to learn the M best arms $(\mu^*_1, \dots, \mu^*_M)$, 2) and each user orthogonalize on this set.

- RhoRand: use an index policy (UCB) and random *ranks*, [Anandkumar et al, 2011]
- MEGA: use a simple ε-greedy algorithm (hard to tune, and not so efficient), [Avner & Mannor, 2015]
- Musical Chair: pure exploration then random hopping until convergence on M best arms, [Shamir et al, 2016]
- Some others: TDFS uses a time sharing, RhoLearn uses BayesUCB to learn the ranks, etc.

4. Our proposal

1) kl-UCB $>$ UCB_1 for best arms identification.
2) New algorithm MCTopM for orthogonalization:
   - Estimate the set of M best arm $\widehat{M^j}(t)$,
   - Randomly play in $\widehat{M^j}(t)$ until fixed on a good one (no collision: fixed on a "chair").

5. UCB_1 and kl-UCB indexes

Let $T^j_k(t)$ the selections of channel k for player j, $X^j_k(t)$ mean of free sensing.

$$\mathrm{UCB}_1 : g^j_k(t) := \frac{X^j_k(t)}{T^j_k(t)} + \sqrt{\frac{\log(t)}{2T^j_k(t)}}$$

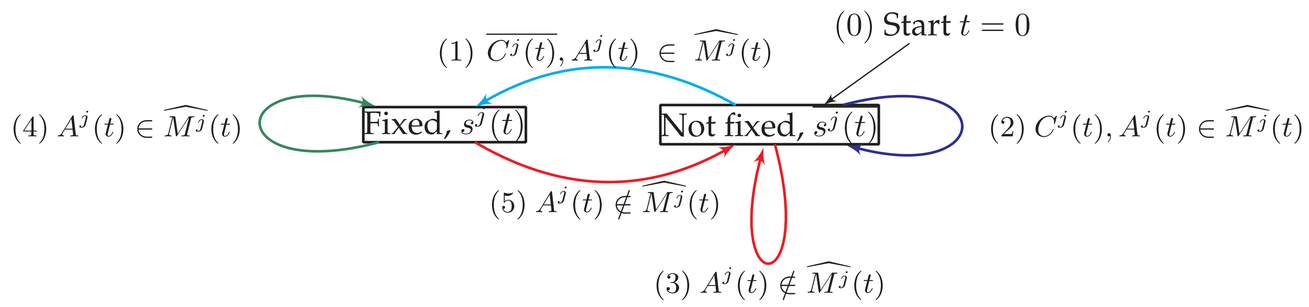
$$\text{kl-UCB} : g^j_k(t) := \sup_{q \in [0,1]} \left\{ q : \mathrm{kl}\left(\frac{X^j_k(t)}{T^j_k(t)}, q\right) \leq \frac{\log(t)}{T^j_k(t)} \right\}$$

[Garivier & Cappé, 2011], [Cappé et al, 2013]

6. MCTopM algorithm

$A^j(1) \sim \mathcal{U}(\{1, \dots, K\})$, $C^j(1) = s^j(1) = \text{False}$
for $t = 0, \dots, T-1$ **do**
  if $A^j(t) \notin \widehat{M^j}(t)$ **then** // (3) or (5)
    $A^j(t+1) \sim \mathcal{U}(\widehat{M^j}(t) \cap \left\{k : g^j_k(t-1) \leq g^j_{A^j(t)}(t-1)\right\})$
    $s^j(t+1) = \text{False}$ // arm with smaller UCB at $t-1$
  else if $C^j(t)$ and $\overline{s^j(t)}$ **then** // (2)
    $A^j(t+1) \sim \mathcal{U}\widehat{M^j}(t)$
    $s^j(t+1) = \text{False}$
  else // transition (1) or (4)
    $A^j(t+1) = A^j(t)$ // same arm
    $s^j(t+1) = \text{True}$ // on "chair"
  end
  Play arm $A^j(t+1)$, get new observations,
  Compute the indexes $g^j_k(t+1)$, and set $\widehat{M^j}(t+1)$ for next step.
end

7. Illustration of MCTopM



8. Theorems

Let $T_k(T) := \sum_{j=1}^M T^k_k(T)$ total selections of arm k, and $\mathcal{C}_k(T)$ counts collisions on arm k.

1. Regret: $R_T(\boldsymbol{\mu}, M) := \mathbb{E}_\mu \left[\sum_{t=1}^T \sum_{j=1}^M \mu^*_j - r^j(t) \right] = \left(\sum_{k=1}^M \mu^*_k \right) T - \mathbb{E}_\mu \left[\sum_{t=1}^T \sum_{j=1}^M r^j(t) \right]$

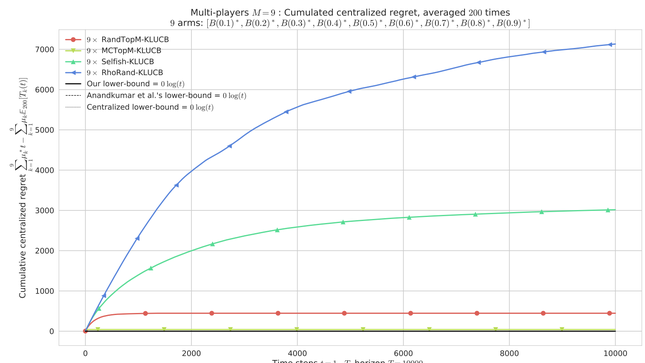
   $$R_T = \sum_{k \in M\text{-worst}} (\mu^*_M - \mu_k) \mathbb{E}_\mu[T_k(T)] + \sum_{k \in M\text{-best}} (\mu_k - \mu^*_M) \left(T - \mathbb{E}_\mu[T_k(T)]\right) + \sum_{k=1}^K \mu_k \mathbb{E}_\mu[\mathcal{C}_k(T)].$$

2. Lemma for upper-bound: only two terms to focus on (bad selections and collisions).

   $$\sum_{k \in M\text{-best}} (\mu_k - \mu^*_M) \left(T - \mathbb{E}_\mu[T_k(T)]\right) \leq (\mu^*_1 - \mu^*_M) \left(\sum_{k \in M\text{-worst}} \mathbb{E}_\mu[T_k(T)] + \sum_{k \in M\text{-best}} \mathbb{E}_\mu[C_k(T)] \right).$$

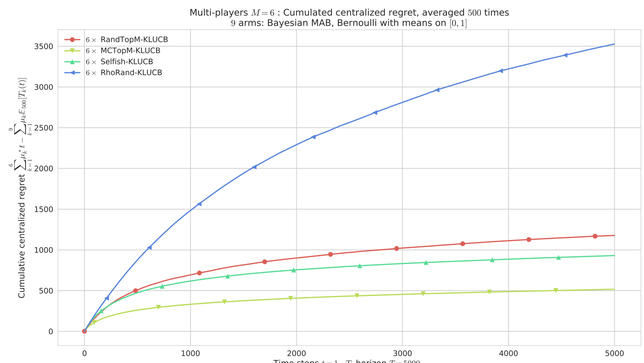
3. If all M players use MCTopM with kl-UCB, $\forall \boldsymbol{\mu}, \exists G_{M,\boldsymbol{\mu}}$, $R_T(\boldsymbol{\mu}, M) \leq G_{M,\boldsymbol{\mu}} \times \log(T) + o(\log T)$.

9.1. Illustration for $M = K$



Only RandTopM and MCTopM achieve constant regret in this saturated case (proved).

9.2. Illustration for $M < K$



RhoRand $<$ RandTopM $<$ Selfish $<$ MCTopM.

10. Conclusions

- Our algorithm MCTopM is uniformly better than any previous proposals.
- Great results: $\log(T)$ collisions, arm switches, bad selections and regret.
- Real-world implementation? $\hookrightarrow$ Yes, presented at ICT 2018!
- Future: learn M, arrival/departures of users, dynamic problems, jammers etc.

11. Main references

More on-line $\rightarrow$ **http://lbo.k.vu/JdD2018**

[BBM+17] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot (2017). *Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings*. In *12th EAI Conference on Cognitive Radio Oriented Wireless Network and Communication*.

[BK18] L. Besson and E. Kaufmann (April 2018). *Multi-Player Bandits Revisited*. In *Algorithmic Learning Theory*. Lanzarote, Spain. URL https://hal.inria.fr/hal-01629733.

Simulation code on GitHub.com/SMPyBandits/SMPyBandits, open source!

12. Thanks to ...

– Organizers of the *Workshop on MAB and Learning Algorithms*!
– *ADDI association* for the *PhD Students Day* 2018 !
– CentraleSupélec, IETR, Univ.Rennes 1, Inria (Lille) & CNRS
– My PhD advisors: Émilie Kaufmann, Christophe Moy.