AN INSIGHT ON MULTIPLE-PLAYS BANDITS IN THE STOCHASTIC SETTING

ALICIA Meeting

Claire Vernade & Paul Lagrée, Olivier Cappé

November 12, 2015

ALICIA – ANR Project

INTRODUCTION

OLD PROBLEM, VERY RECENT ADVANCES

Simple rule : you are allowed to pull several arms at each round.
feedback :

- Bandit feedback : unknown function of the super-action;
- Semi-Bandit feedback : individual rewards of each arm in the super-action, may be partially observed.

two possible settings :

- Adversarial : cf Sebastien's talk in April 2014,
- Stochastic : Today

MODELS IN LITERATURE

[Komiyama 15']

1. at each round $t = 1, \cdots, T$, select a set A_t of L elements among K arms
2. *feedback*: semi-bandit (the learner receives an observation for each item $i_k \in A_t$)
3. *reward*: linear (sum of L individual answers)

Thompson Sampling is proved to be optimal in the case of Bernoulli arms

[Kveton 15']

1. at each round $t = 1, \cdots, T$, select a list A_t of L elements among K arms
2. *feedback*: each item until the first click is observed, nothing after.
3. *reward* is binary: 1 if one click, 0 otherwise

Placing the worst items in the beginning of the list is better since more information is obtained without increasing regret.

CASCADE MODEL II

[Combes 15']

- Decreasing weights are added to positions (to avoid last effect)

- **Algorithm**: PIE(l)
  1. rank empirical means $\hat{\mu}_k$
  2. select list with L best $\hat{\mu}_k$
  3. with probability 1/2, explore on position l with any item whose KL-UCB is bigger than L-th empirical mean.

RANDOM SEMI-BANDIT MODEL

1. at each round $t = 1, \cdots, T$, choose a list A_t of L items among K possible arms.
2. the user rates every item $k \in A_t$ until she decides to leave at position Λ_t (random variable)
3. probability of scanning item in position l is modeled by $\kappa_l > 0$
4. *reward*: $\sum_{l=1}^{\Lambda_t} X_{t,l}$

RANDOM SEMI-BANDIT FEEDBACK

PROVING A LOWER-BOUND FOR A MULTIPLE-PLAYS MODEL

A process in 4 key steps using Emilie's arguments:

1. Compute the expectation of the log-likelihood of the observations : may vary according to your model;
2. Define a set of changes of measure $B(\theta)$ that modify the optimal arm : $a^*(\theta) \neq a^*(\lambda)$;
3. Appeal to Emilie's Theorem to get a lower-bound :

$$\lambda \in B(\theta),\ \liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}} \mathbb{E}[N_a(T)] \sum_{l=1}^{L} \kappa_l KL(\theta_{a(l)}, \lambda_{a(l)})}{\log(T)} \geq 1$$

4. Use it as a constraint to lower-bound your regret :

$$R(t) \geq \min_{c\geq 0} \sum_{a\in\mathcal{A}} c_a d_\theta(a) \quad (1)$$

$$\forall \lambda \in B(\theta),\quad \sum_{a\in\mathcal{A}} c_a \ s.t.\ \sum_{l=1}^{L} \kappa_l KL(\theta_{a(l)}, \lambda_{a(l)}) \geq 1 \quad (2)$$

A MORE EXPLICIT LOWER-BOUND

A more friendly lower-bound can be obtained looking for an even lower bound :

1. Relax some constraints of your problem : $B(\theta) \subset \cup_{i \notin a^*(\theta)} B_i(\theta)$ where in B_i, only parameter θ_i can be modified.
2. Remark that including suboptimal arms in your action leads to a regret $d_a(\theta)$ that verifies

$$d_a(\theta) \geq \sum_{k>L} \sum_{l=1}^{L} \mathbf{1}\{a(l) = k\} \kappa_l (\theta_L - \theta_k)$$

3. Combine previous ideas to lower-bound the optimal solution of the optimization problem.

RESULT...

Proceeding as described above, we obtain the following lower bound :

$$\liminf_{T\to\infty} \frac{R(T)}{\log(T)} \geq \sum_{k=l+1}^{L} \frac{\theta_L - \theta_k}{KL(\theta_k, \theta_L)}$$

Comments:

- No kappas : they are hidden,
- Same bound as Anatharam et al. for the L-set problem (unordered arms).
- Really different proof than Combes' one for their "Learning to rank" model.

One main question: **Can usual index policies be simply adapted for ranking arms ?**
Basic idea: Compute the indexes of each policies and rank them in order to get the chosen action (instead of simply pulling the best one).
Tested algorithms:

1. UCB and KL-UCB : rank upper bounds in decreasing order, draw the first L ones.
2. Thompson Sampling: rank posterior samples in decreasing order, draw the first L ones.

EXPERIMENTAL RESULTS - κ IMPACT

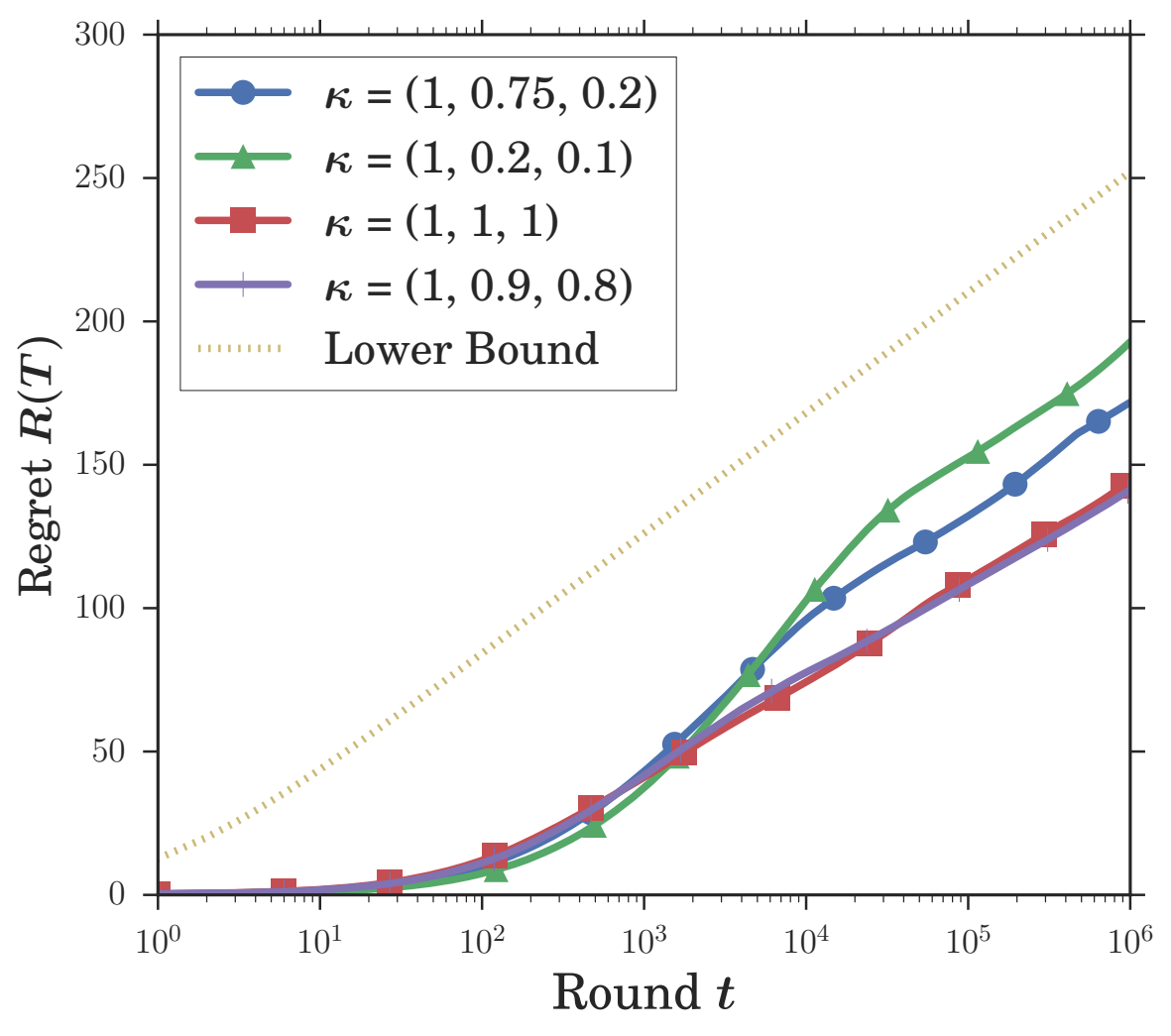

Figure 1: Regret in logarithmic scale for various κ

EXPERIMENTAL RESULTS - ALGORITHMS

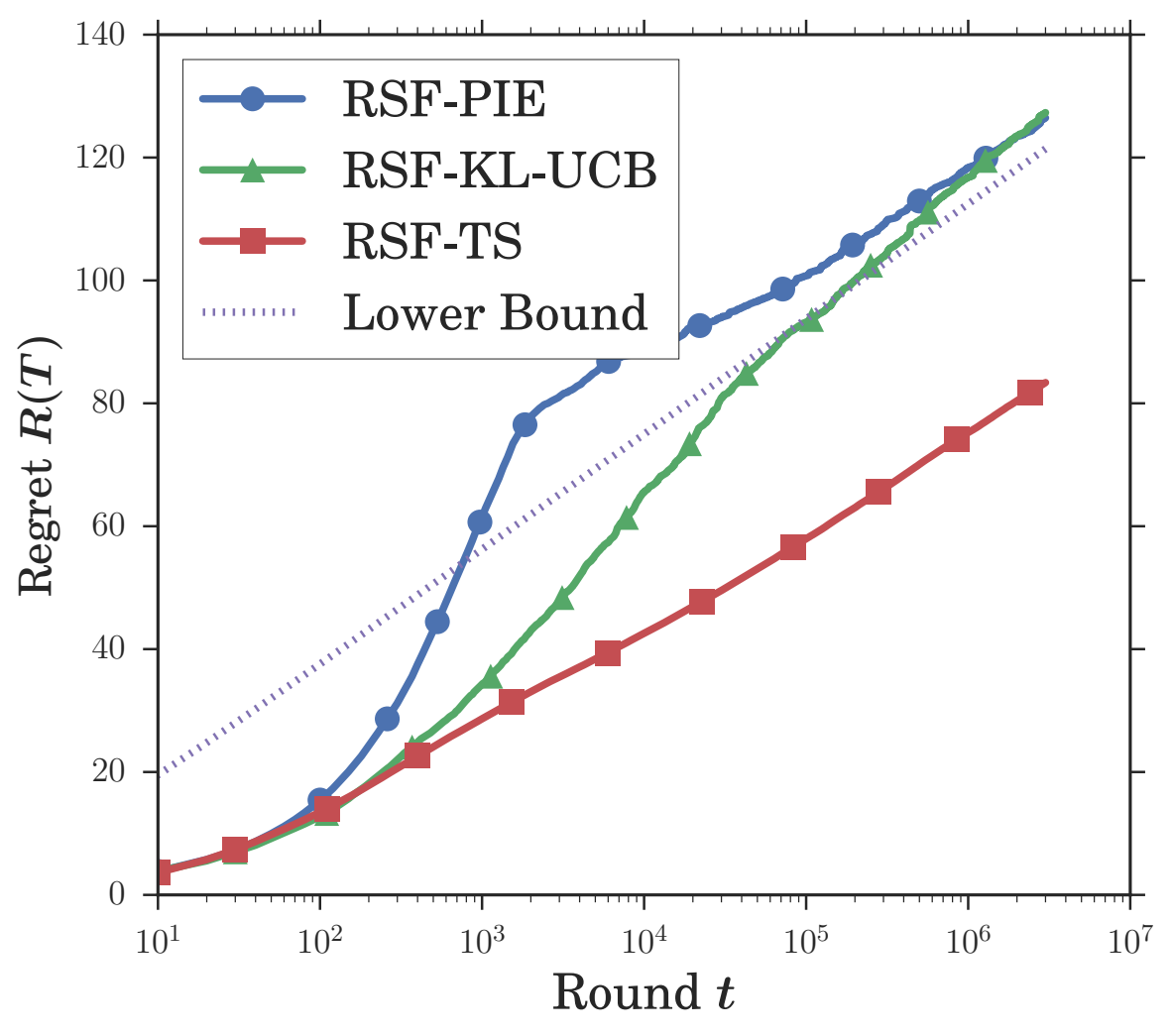

Figure 2: Regret in logarithmic scale for various algorithms

FUTURE WORK

WHAT REMAINS TO BE DONE

Several questions remain :

- Can we analyze any of the proposed algorithms ?
- Is it worth it after all that has already been done ?
- What if we did not know whether the user dislikes or do not see the proposed items (lack of observations) ?

Questions?

REFERENCES I

J. Komiyama, J. Honda and H. Nakagawa
Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays.
Proceedings of the 32nd International Conference on Machine Learning, 2015.

B. Kveton, C. Szepesvari, Z. Wen and A. Ashkan
Cascading Bandits : Learning to Rank in the Cascade Model.
Proceedings of the 32nd International Conference on Machine Learning, 2015.

R. Combes, M. Lelarge, A. Proutieres and M.S. Sedegh
Stochastic and Adversarial Combinatorial Bandits.
arXiv preprint arXiv:1502.03475, 2015.

References II


R. Combes, S. Magureanu, A. Proutieres and C. Laroche
Learning to Rank: Regret Lower Bounds and Efficient Algorithms.
SIGMETRICS, 2015.