Sequential ranking under random semi-bandit feedback

Hossein Vahabi , Paul Lagrée , Claire Vernade , Olivier Cappé

March 7, 2016

Abstract

In many web applications, a recommendation is not a single item suggested to a user but a list of possibly interesting contents that may be ranked in some contexts. The combinatorial bandit problem has been studied quite extensively these last two years and many theoretical results now exist : lower bounds on the regret or asymptotically optimal algorithms. However, because of the variety of situations that can be considered, results are designed to solve the problem for a specific reward structure such as the Cascade Model. The present work focuses on the problem of ranking items when the user is allowed to click on several items while scanning the list from top to bottom.

1 Introduction

In the context of online learning, the classical bandit framework refers to situations for which the learner is only allowed to observe the rewards of the action he took. This is actually to be compared to the full information framework in which after choosing one action among a pool of candidates, the learner observes the result of every action, including the non-chosen ones. The canonical application of online learning with bandit feedback is the case of online advertising. Suppose each time a user lands on a webpage, she is shown an ad among K in the current catalog. The goal of the ad displayer is to select the ad that has the highest probability of leading to a click, that is to maximize its *click-through rate* (CTR). Each time an ad is tried out, we can only record click/non-click on that ad and update our knowledge on its CTR but we cannot say anything about the other possible ads.

The same kind of process can be imagined in the context of recommendation. However, recommending items rarely boils down to showing a single one. In general, the user is presented a set of recommendations and she can click on many of them, giving larger feedback to the learner. Just as in the Cascade Model that we describe in section 3, we suppose that the user scans the list from the top but we assume that she has a limited patience and suddenly stops rating and leaves the system independently of what she was looking at. In that case, we are not strictly observing a bandit feedback: one multiple-action is made and we receive each individual reward instead of a unique response that would stand for the set of chosen items. We are not either in the case of full information since rewards corresponding to non-chosen options are not revealed. This framework is often called the *semi-bandit feedback*.

In this paper, we consider the following multiple recommendation scenario. A list of items is chosen by the learner. The user then scans the list sequentially: when observing an item, she can either rate it (0 or 1) or leave and stop rating the items in the list. Defined this way, the problem is not to find an optimal set of items but rather to find the best ordered list of items among K options.

We start by describing the bandit model we consider and especially the novel feedback structure that allows us to tackle problems of recommendation rather than information search. A large overview of the literature is provided next that aims at localizing the present work in a flourishing literature on combinatorial bandit problems. Technical sections come after: we present and show a lower bound for the bandit problem in our setting and then suggest efficient algorithms to solve the proposed problem.

2 Model

We consider the binary bandit problem with K arms. The parameters of the arms are the expectation of a Bernoulli distribution which lies in $\Theta = [0,1]^K$. A bandit model is then a tuple $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$. Without loss of generality, we shall always suppose that $\theta_1 > \ldots > \theta_K$. At each round t, the learner selects a list of L elements chosen among the K arms, which we materialize by a list of indices belonging to $\{1, \ldots, K\}$. The set of those lists is denoted by $\mathcal{A}$ and contains $L!/(K-L)!$ elements; the list chosen at time t will be denoted A_t. The user scans the list from the top to the end and gives a feedback for each observed item until she chooses to stop scanning, which occurs independently of previous recommendations. Let (Λ_t) be a sequence of i.i.d. random variables in $\{1, \ldots, L\}$ corresponding to the number of observed items at each round. Thus, playing action A_t, the learner receives a reward $r_{A_t}(t) = \sum_{l=1}^{\Lambda_t} X_{t,l}$ at round t, where $X_{t,l}$ is an independent draw with probability distribution $\mathcal{B}(\theta_{A_t(l)})$. The probability of scanning the item in position $l \in \{1, \ldots, L\}$ is modeled by κ_l, that is $\kappa_l = \mathbb{P}(\Lambda \geq l)$. We suppose that the learner knows when the user stops rating so that no ambiguity is left on the last unrated items. The items without feedback are unobserved and nothing can be inferred concerning the rewards that could have been earned.

Clearly, the optimal list of L arms is $a^* = (1, 2, \ldots, L)$, and the regret of the learner up to time T writes

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} r_{a^*} - r_{A_t} \\ \mathbb{E}_\theta[R(T)] &= \sum_{t=1}^{T} \sum_{l=1}^{L} \kappa_l(\theta_{a^*(l)} - \theta_{A_t(l)}) \\ &= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \left(\sum_{l=1}^{L} \kappa_l(\theta_l - \theta_{a(l)}) \right) \mathbb{1}\{A_t = a\} \end{aligned}$$

where the expectation is over the probability distribution of the bandit model. Taking the expectation of the above quantity with respect to Monte-Carlo repetitions, we obtain the expected regret

$$\mathbb{E}[R(T)] = \sum_{a \in \mathcal{A}} \left(\sum_{l=1}^{L} \kappa_l(\theta_l - \theta_{a(l)}) \right) \mathbb{E}[N_a(T)]$$

Let us introduce $\mu_a = \sum_{l=1}^{L} \kappa_l \theta_{a(l)}$ the reward expectation of arm $a \in \mathcal{A}$. Denoting μ_{a^*} by μ^*, the expected regret can be rewritten

$$\mathbb{E}[R(T)] = \sum_{a \in \mathcal{A}} (\mu^* - \mu_a) \mathbb{E}[N_a(T)]$$

3 Related Work

The closest work to ours is [7] that consider the problem of learning to rank items in a stochastic setting with semi-bandit feedback. The main difference with our work lies in the reward structure: in [7], the authors rely on the Cascade Model. This model comes from the Information Retrieval (IR) literature and suggests that the users of a search engine scan the proposed list from top to bottom and click on the first interesting item. In order to induce the necessity of ranking items, weights are sometimes attributed to positions in the chosen list: a click on position $i \in \{1, \dots, L\}$ gives reward $r(i) > r(i+1)$. Thus, order matters and the algorithm must be able to rank items according to their CTR in order to maximize reward. The authors suggest an algorithm based on KL-UCB ([10]) and prove a lower bound on the regret as well as an asymptotically optimal upper bound. Their ideas mainly come from very related previous works from same co-authors such as [5] and [8] as well as on the general analysis of lower bound for Markov Decision Processes [11] done by Graves & Lai in 1997.

On the contrary, recent works on the Cascade Model ([14],[15]) do not solve the ranking problem.They are still closely linked to this work because of the randomized size of the semi-bandit feedback they consider. In [14], the authors suggest and study a UCB-based algorithm for building such lists of items in order to maximize the number of clicks : the reward is 1 if the user clicks on at least one item so the best multiple arm to play is the set of the L-best arms where L is the size of the list. Concretely, they compare to the optimal strategy that chooses the L best arms – no matter how ordered – and prove that they obtain a $O(\log T)$ regret. The model studied in [15] is slightly different as they need to find lists for the reward of each item is 1. They also provide an algorithm to solve their problem and show a regret upper bound in $O(\log T)$

More generally, multiple-plays bandits – for which the agent draws a super-action at each round – have received a lot of attention recently both in the adversarial and in the stochastic setting. In the adversarial setting, primary work was presented in [18] for semi-bandit feedback and in [3] for more general bandit feedback. In the stochastic setting, an early work was done in order to propose algorithms for the multi-user channel allocation problem in cognitive radio ([9]). Then, more efficient algorithms were suggested successively in [4] and in the very recent work [6].

Even if the literature on combinatorial bandits seems to be quite recent, the simpler problem that consists in finding the $L-$best arms with semi-bandit feedback was already studied in 1987 by Anantharam et al. ([1]) who provided a lower-bound for this problem as well as an asymptotically optimal algorithm. Their contribution can be seen as the equivalent of the work of Lai & Robbins ([16]) for the classical multi-armed bandit with single play. However, the algorithm they proposed was not computationally efficient and the recent work by Komiyama et al. ([13]) suggest and analyze an efficient algorithm based on Thompson Sampling ([17]). They prove that Thompson Sampling is optimal for the problem of finding the best items when the order does not matter. Whether Thompson Sampling is optimal also for ranking items remains an open problem.

4 Lower Bound on the Regret

In this section, we state and prove a theorem giving a lower bound on the regret of a *uniformly efficient* algorithm for the sequential MP-MAB problem. The proof builds on ideas from ([16, 1, 12, 7]) .

4.1 Definitions and Notations

We recall that the set of arms – that is the lists of L elements chosen without replacement among K – is denoted by $\mathcal{A}$. We also denote $\overline{\mathcal{A}^*}$ the set of actions that contain at least one arm $k \notin a^*(\theta)$, or equivalently that is not a permutation of the arms of $a^*(\theta)$. The Kullback-Leibler divergence from p to q is denoted by $\mathrm{KL}(p, q)$ while $d(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$ is the binary relative entropy.

Definition 1. *A class $\mathcal{M}$ of bandit models is identifiable if it is of the form $\mathcal{M} = \mathcal{P}_1 \times \ldots \times \mathcal{P}_K$, where $\mathcal{P}_a$ is the set of possible distributions for arm a, and if for all a, $\mathcal{P}_a$ is such that*

$$\forall p, q \in \mathcal{P}_a \quad p \neq q \implies 0 < \mathrm{KL}(p, q) < \infty \tag{1}$$

Definition 2. *An algorithm $\mathcal{A}$ is said* uniformly efficient *if for all bandit model ν and for all $\alpha \in]0,1]$, the expected regret of algorithm $\mathcal{A}$ after T rounds is such that $R(T,\mathcal{A}) = o(T^\alpha)$. Then, for the problem of ranking, the unique optimal arm is $a^* = (1,2,\ldots,L)$ and any uniformly efficient algorithm must satisfy*

$$T - \mathbb{E}[N_{a^*}(T)] = o(T)$$
$$\mathbb{E}[N_a(T)] = o(T)\ ,\ \forall a \neq a^*$$

Definition 3. *For all bandit model $\theta \in \Theta$, we define an interesting set of changes of measure:*

$$B(\theta) = \{\lambda \in \Theta \mid \forall k \in a^*(\theta), \theta_k = \lambda_k \text{ and } \mu^*(\theta) < \mu^*(\lambda)\}\,. \tag{2}$$

For each suboptimal arm k, a subset of $B(\theta)$ will be used:

$$B_k(\theta) = \{\lambda \in \Theta \mid \forall j \neq k, \theta_j = \lambda_j \text{ and } \mu^*(\theta) < \mu^*(\lambda)\}\,. \tag{3}$$

Notice that this last definition implies that $\lambda_k > \theta_L$.

4.2 Related works on lower-bounds for MAB problems

For a classical MAB problem $\nu = (\nu_1,\ldots,\nu_K)$ with one only optimal arm ν^* with mean μ^* and one observation at each step, Lai & Robbins gave a lower-bound on the regret that comes from a lower bound on the number of suboptimal draws

$$\mu_a < \mu* \implies \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{KL}(\nu_a,\nu^*)}.$$

This bound was later generalized by [2] to families that depend on multiple parameters. Closer to our work, [1] consider a multiple play bandit problem in the semi-bandit setting that consists in drawing L different arms at each round. They give a lower bound on the regret that is based on a lower-bound on the number of pulls of the worst arms : for each worst arm j – i.e. that does not belong to the L best ones –,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_j(T)]}{\log(T)} \geq \frac{1}{\mathrm{KL}(\nu_j,\nu_L)}.$$

This last bound is closely related to our problem even if it seems not to carry all the complexity of the ranking aspect related to sequential rewards. Surprisingly, the lower-bound of [7] as well as ours does not show any additional term that would stand for the complexity of ranking the optimal arms. In fact, for the "learning to rank" problem in [7] where rewards follow the weighted Cascade Model with convex decreasing weights $(r(l))_{i=1,\ldots,L}$ on the positions, they obtain the following lower-bound :

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

The end of the section is dedicated to the detailed proof of the lower-bound on the regret of our model for sequences of feedback whose length do not depend on the actual items proposed in the list. The proof uses ideas from [7] and [6] but is not strictly built on the same architecture and relies on generic lower-bound results from [12].

4.3 Lower-bound on the sequential MP-MAB problem

The main result of this section is stated in the following theorem.

Theorem 4. *Simplifying the bound of Theorem 5, the regret of any uniformly efficient algorithm can be lower-bounded by*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R(T)]}{\log T} \geq \sum_{k=L+1}^{K} \frac{\theta_L - \theta_k}{\mathrm{KL}(\theta_k, \theta_L)}.$$

This theorem is actually a more explicit version of the very generic result on lower-bounds given by Graves & Lai ([11]) for our own bandit model.

Theorem 5. *The regret of any uniformly efficient algorithm is lower bounded as follows*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R(T)]}{\log T} \geq c(\theta).$$

where

$$c(\theta) = \inf_{c\succeq 0} \sum_{a\in\mathcal{A}} d_a(\theta) c_a$$

$$s.t \inf_{\lambda\in B(\theta)} \sum_{a\in\mathcal{A}} K_a(\theta,\lambda) c_a \geq 1$$

This result comes from a more general work on controlled Markov chains and is pointed out by [6] as a very powerful tool to obtain lower-bounds for bandit problem. Interestingly, it is possible to provide an original proof of this theorem using results from [12]. The following proposition allows us to prove Theorem 5 and for completeness we give a full proof of it in Appendix. The main problem-dependent step of the proof relies in the computation of the expectation of the log-likelihood ratio of the observations. The result of it provides the numerator of the following expression.

Proposition 6. *For any uniformly efficient algorithm*

$$\forall \lambda \in B(\theta), \qquad \liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}} K_a(\theta,\lambda)\mathbb{E}[N_a(T)]}{\log(T)} \geq 1$$

where $K_a(\theta,\lambda) = \sum_{l=1}^{L} \kappa_l \,\mathrm{KL}(\theta_{a(l)}, \lambda_{a(l)})$.

Proof. (Theorem 5) In order to prove the lower bound on the regret, one needs to show that the uniform efficiency of any algorithm requires a certain amount of suboptimal pulls that avoid any confusion with the optimal option. We proceed in two steps. First, we rewrite the regret lower bounding some of the numerous terms in order to get a more simple expression. Then, we lower bound each of the remaining terms using changes of measure arguments.

Step 1: $\overline{\mathcal{A}^*}$ denotes the set of actions containing at least one suboptimal arm. Suppressing the terms due to actions in $\mathcal{A}^*$, the regret can be written as follows

$$\begin{aligned}\mathbb{E}[R(T)] &= \sum_{a\in\mathcal{A}} \left(\sum_{l=1}^{L} \kappa_l(\theta_{a^*(l)} - \theta_{a(l)})\right) \mathbb{E}[N_a(T)] \\ &= \sum_{a\in\mathcal{A}} \mathbb{E}[N_a(T)] \left(\sum_{l=1}^{L} \kappa_l(\theta_l - \theta_{a(l)})\right) \\ &= \sum_{a\in\mathcal{A}} d_a(\theta)\mathbb{E}[N_a(T)].\end{aligned}$$

In the last line, we denote the positive regret associated to each suboptimal action $d_a(\theta) := \sum_{l=1}^{L} \kappa_l(\theta_l - \theta_{a(l)})$.

Step 2: We can use the previous decomposition to lower bound the regret as follows

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R(T)]}{\log T} \geq \sum_{a\in\mathcal{A}} d_a(\theta) \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)}.$$

According to Proposition 6, for all $\lambda \in B(\theta)$,

$$\sum_{a\in\mathcal{A}} K_a(\theta,\lambda) \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq 1.$$

This result means that the regret is actually lower bounded by a sum of non-negative quantities that satisfy the above constraint provided by Proposition 6. The solution of the optimization problem stated in the theorem is then effectively a lower bound of this sum. □

4.4 Simplified lower-bound : proof of Theorem 4

Following the ideas of [5], we can simplify the lower-bound of Theorem 5 in order to obtain the more explicit version of Theorem 4.

It can be noticed that this lower-bound is actually the same as the one of [1] even if the combinatorial structure of the problem seems more complex here. Of course, we do not show that this lower-bound is in fact the optimal one as long as we do not provide an algorithm whose regret is asymptotically upper-bounded by the same quantity. Developments and experiments of the next Sections show that it can be attained in practice and thus that it is eventually the optimal lower-bound for our problem.

Proof. The detailed proof is provided in Appendix. We only give the sketch of our analysis here.

The first argument of the proof is that when we release constraints, we actually widen the feasible set and this allows us to possibly reach a lower-bound of the wanted result. Concretely, instead of considering all possible changes of measure in the constraint equation, we only allow λ to lie in $B_k(\theta)$ for $k \notin a^*(\theta)$. We obtain the following relaxed optimization problem :

$$c(\theta) = \inf_{c \succeq 0} \sum_{a \neq a^*(\theta)} d_a(\theta) c_a \tag{4}$$

$$s.t\ \forall k \notin a^*(\theta),\ \forall \lambda \in B_k(\theta), \sum_{a \in \mathcal{A}} K_a(\theta, \lambda) c_a \geq 1. \tag{5}$$

In the above problem, the constraints are still not really explicit since each of them must be valid for all $\lambda \in B_k(\theta)$ even in the worst case : for $k \notin a^*(\theta)$, one must satisfy

$$\inf_{\lambda \in B_k(\theta)} \sum_{a \in \mathcal{A}} K_a(\theta, \lambda) c_a \geq 1.$$

Fortunately, computing this infimum is quite easy on the given sets $B_k(\theta)$ and returns an explicit optimization problem

$$\begin{aligned} c(\theta) = \inf_{c \succeq 0} & \sum_{a \neq a^*(\theta)} d_a(\theta) c_a \\ s.t\ & \forall k \notin a^*(\theta), \\ & \sum_{a \neq a^*(\theta)} c_a \sum_{l=1}^{L} \mathbb{1}\{a(l) = k\} \kappa_l \operatorname{KL}(\theta_k, \theta_L) \geq 1. \end{aligned}$$

It remains to rewrite the argument using Lemma 9 to obtain

$$\begin{aligned} & \sum_{a \neq a^*(\theta)} d_a(\theta) c_a \\ & \geq \sum_{k \notin a^*(\theta)} (\theta_L - \theta_k) \sum_{a \neq a^*(\theta)} c_a \sum_{l=1}^{L} \kappa_l \mathbb{1}\{a(l) = k\} \\ & \geq \sum_{k \notin a^*(\theta)} \frac{\theta_L - \theta_k}{\operatorname{KL}(\theta_k, \theta_L)} \end{aligned}$$

□

5 Algorithm

In this section, we propose extensions of Thompson sampling, UCB and KL-UCB to the semi-bandit feedback problem. The TS-like version (RSF-TS) is given in Algorithm .

Algorithm 1 Random semi-bandit feedback Thompson sampling (RSF-TS)

Require: number of arms K, number of positions L
  for $k = 1, \dots, K$ **do**
    $\alpha_k, \beta_k \leftarrow 1, 1$
  end for
  for $t = 1, \dots, N$ **do**
    for $k = 1, \dots, K$ **do**
      $\theta_k(t) \sim \text{Beta}(\alpha_k, \beta_k)$
    end for
    A_t = top-L arms ordered by decreasing $\theta_k(t)$
    for $l = 1, \dots, \Lambda_t$ **do**
      if $X_{t,A_t(l)} = 1$ **then**
        $\alpha_{A_t(l)} \leftarrow \alpha_{A_t(l)} + 1$
      else
        $\beta_{A_t(l)} \leftarrow \beta_{A_t(l)} + 1$
      end if
    end for
  end for

Each of these algorithms reuses upper-confidence based indexes and choses the action by pulling the arms with the best indices in decreasing order. Concretely, we set

- in RSF-UCB, for $k \in 1, ..., K$, $u_k(t) = \hat{\mu}_i(t) + \sqrt{(2\log t)/(3N_i(t))}$;
- in RSF-KL-UCB, for $k \in 1, ..., K$, $u_k(t) = \sup_{q \in [\hat{\mu}_i(t), 1]} \{q | N_i(t) d(\hat{\mu}_i(t), q) \leq \log t\}$ as proposed in [10];
- and in RSF-TS, for $k \in 1, ..., K$, $u_k(t) \sim \pi_k, t$ is a sample from the posterior distribution over the mean of arm k. Recall that here only Bernoulli bandits are considered so that the posterior p_k, t is always a Beta distribution.

6 Experiments

In this section we evaluate the empirical properties of RSF-TS and its variants in several scenarios. Our performance are compared to those obtained by the PIE(L) algorithm of [7]. [1].

6.1 On the optimality of PIE(L) for a multiple-click model

6.2 Synthetic experiment

We consider a simple scenario with 6 Bernoulli distributions with respective expected rewards 0.6, 0.58, 0.55, 0.45, 0.45 and 0.4. We fix $L = 3$. In this

[1]The source code of the simulations is available at https://github.com/plagree/random-semi-bandit.

section, results are averaged over $10K$ runs for a number of rounds $T = 10^5$.

Influence of the vector κ on the regret First, we evaluate the effect of the probabilities of observation κ when setting various values. The simulation results are shown in Figure 1. Unsurprisingly, the graph comfort the lower bound regret analysis. The different values for κ have no influence on the asymptotic behavior. Low values for κ_2 and κ_3 provoke a delay on the round when the algorithm regret and the lower bound become parallel.

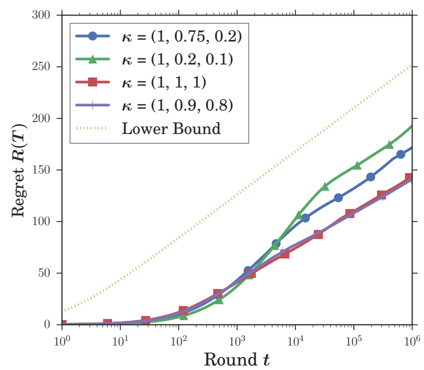

Figure 1: Impact of κ. the results in log-scale allowing to compare the growing rate of the regret with respect to the theoretic lower-bound.

Figure 2 shows the effectiveness of several algorithms. The PIE(L) algorithm used in our framework seems to have a quite high regret that asymptotically shows an optimal behavior with respect to the proven lower-bound.

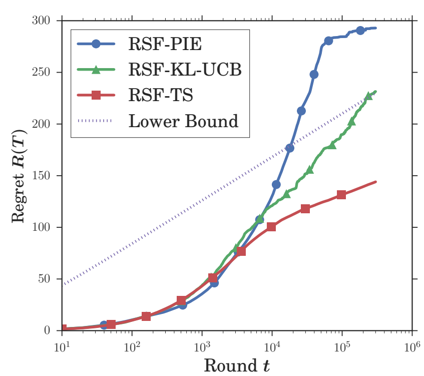

Figure 2: Regret of RSF-TS, RSF-KL-UCB and RSF-UCB.

6.3 Search advertising

We used a publicly available dataset provided for KDD Cup 2012 track 2 [2]. The dataset involves session logs of soso.com, a search engine owned by Tencent. It consists in ads that were inserted among the results of the search engine. Each of the $150M$ lines from the log contains the user, the query she typed, an ad, a position (1, 2 or 3) at which it was displayed and a binary reward (click/no-click). First, for every query, we excluded ads that were not displayed at least $1,000$ times at every position. We also filtered queries that had less than 5 ads satisfying previous constraints. As a result, we obtained 8 queries with 5 up to 11 ads. For each query q, we computed the matrix $M_q \in \mathbb{R}^{K\times L}$ (where K is the number of ads and L the number of positions) and filled it with the average click-through-rate (CTR). We computed the SVD of M_q matrix and discarded every singular value but the largest. Thus, we obtained a decomposition $M_q = \theta_q \kappa_q^T$ where $\theta_q \in \mathbb{R}^K$ and $\kappa_q \in \mathbb{R}^L$. We emphasize that we normalized κ_q such that the probability to observe an ad in first position is 1. Finally, each ad a was converted into a Bernoulli distribution of expectation $\theta_{q,a}$. Table 1 reports some statistics about the resulting bandit model.

We conducted a serie of $N = 250$ simulations over this dataset. At the beginning of each run, a query was randomly selected together with corresponding probabilities of scanning positions and arm expectations. Even though rewards were still simulated, this scenario was more realistic since the values of the parameters were extracted from a real-world dataset.

#ads (K)	**#records**	**min θ**	**max θ**
5	$216,565$	0.016	0.077
5	$68,179$	0.031	0.050
6	$435,951$	0.025	0.067
6	$110,071$	0.023	0.069
6	$147,214$	0.004	0.148
8	$122,218$	0.108	0.146
11	$1,228,004$	0.022	0.149
11	$391,951$	0.022	0.084

Table 1: Statistics on the resulting queries.

We show the results of the simulations in Figure 3. We display the evolution of the CTR for 3 competitive algorithms: RSF-PIE(L), RSF-KL-UCBand RSF-TS. Note that we do not show the CTR until $t > 500$ to avoid unrelevant measures. Similarly to the previous scenario, RSF-TSslightly outperforms the other algorithms. Furthermore, we emphasize that the CTR rapidly grows to a near optimal value which is an important characteristic for production use.

[2] http://www.kddcup2012.org/

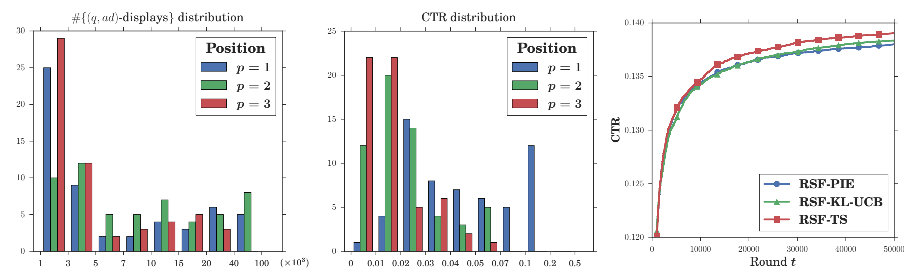

Figure 3: CTR of RSF-TS, RSF-KL-UCB and RSF-PIE(L).

7 Conclusion

The main idea of this work is to study a reward model that would be independent of the content of the recommendation. The now well understood multiple-plays problems with semi-bandits feedback has been studied in the Cascading Model ([14, 15, 7]) where the total reward is either 1 or 0 depending on the items proposed in each round. We suggest that in some situations the user may be allowed to click on several items stopping only because she gets tired of rating. We study the lower-bound on the regret associated with such a bandit model and propose a generic algorithm to solve this problem showing that optimal performances can be attained. Even if empirical evaluation are quite convincing, the statistical analysis of RSF-TS remains to be done.

Acknowledgements

References

[1] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

[2] Apostolos N Burnetas and Michaël N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[3] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[4] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[5] Richard Combes, Marc Lelarge, Alexandre Proutiere, and M Sadegh Talebi. Stochastic and adversarial combinatorial bandits. *arXiv preprint arXiv:1502.03475*, 2015.

[6] Richard Combes, Marc Lelarge, Alexandre Proutiere, and M. Sadegh Talebi. Stochastic and adversarial combinatorial bandits. *CoRR*, abs/1502.03475, 2015.

[7] Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 231–244. ACM, 2015.

[8] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. *arXiv preprint arXiv:1405.5096*, 2014.

[9] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, pages 1–9. IEEE, 2010.

[10] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. *Conference on Learning Theory*, pages 359–376, 2011.

[11] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[12] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identication in multi-armed bandit models. *Journal of Machine Learning Research*, 2015.

[13] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[14] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits : Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[15] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *arXiv preprint arXiv:1507.04208*, 2015.

[16] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[17] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[18] Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for adversarial bandit problems with multiple plays. In *Algorithmic Learning Theory*, pages 375–389. Springer, 2010.

A Appendix

A.1 Lemmas

We call $\mathcal{A}$ the set of all possible actions, that is the set of all sequences of L elements chosen among K options. We have $|\mathcal{A}| = K!/(K-L)!$. We recall that at each round s, the random variable Λ_s is equal to the length of the observed sequence of feedback and it is independent of the chosen action.

Lemma 7. *Let $\theta = (\theta_1, \dots, \theta_K)$ and $\lambda = (\lambda_1, \dots, \lambda_K)$ be two bandit models such that the distributions of all arms in θ and λ are mutually absolutely continuous. Let σ be a stopping time with respect to $(\mathcal{F}_t)$ such that $(\sigma < +\infty)$ a.s. under both models. Let $\mathcal{E} \in \mathcal{F}_\sigma$ be an event such that $0 < \mathbb{P}_\theta(\mathcal{E}) < 1$. Then one has*

$$\sum_{a \in \mathcal{A}} K_a(\theta, \lambda) \mathbb{E}_\theta[N_a(T)] \geq d(\mathbb{P}_\theta(\mathcal{E}), \mathbb{P}_\lambda(\mathcal{E}))$$

where $K_a(\theta, \lambda) = \sum_{l=1}^{L} \kappa_l \, \mathrm{KL}(\theta_{a(l)}, \lambda_{a(l)})$.

Proof. Let us denote by X_s the reward obtained at time s, following the selection of arm A_s. Thus, X_s is a vector of length Λ_s.

The log-likelihood ratio of the observations up to time t under a bandit algorithm is defined by

$$\begin{aligned} L_t &:= \sum_{s=1}^{t} \log \frac{f(X_s; \theta \mid \mathcal{F}_{s-1}, \Lambda_s)}{f(X_s; \lambda \mid \mathcal{F}_{s-1}, \Lambda_s)} \\ &= \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} \mathbb{1}\{A_s = a\} \sum_{l=1}^{L} \mathbb{1}\{\Lambda_s \geq l\} \log \frac{f_{a(l)}(X_{s,l}; \theta)}{f_{a(l)}(X_{s,l}; \lambda)} \end{aligned}$$

We proceed in two steps : we first justify that the expectation of the log-likelihood ratio is lower-bounded by the binary entropy $d(\mathbb{P}_\theta(\mathcal{E}), \mathbb{P}_\lambda(\mathcal{E}))$ and then we show that the same expectation can be rewritten in a more convenient form that gives the desired result. We follow the ideas of the proof of Lemma 1 in Appendix A.1 of [12].

Step 1 : Using the conditional Jensen's inequality and Lemma 8, we have :

$$\begin{aligned} \mathbb{P}_\lambda(\mathcal{E}) &= \mathbb{E}_\theta[\mathbb{E}_\theta[\exp(-L_t)|\mathbb{1}_\mathcal{E}]\mathbb{1}_\mathcal{E}] \\ &\geq \mathbb{E}_\theta[\exp(-\mathbb{E}_\theta[L_t|\mathcal{E}])\mathbb{1}_\mathcal{E}] \\ &= \exp(-\mathbb{E}_\theta[L_t|\mathcal{E}])\mathbb{P}_\theta(\mathcal{E}). \end{aligned}$$

Writing the same for $\overline{\mathcal{E}}$ yields

$$\mathbb{P}_\lambda(\overline{\mathcal{E}}) \geq \exp(-\mathbb{E}_\theta[L_t|\overline{\mathcal{E}}])\mathbb{P}_\theta(\overline{\mathcal{E}}).$$

Finally, using the formula of total expectation, one gets

$$\begin{aligned}\mathbb{E}_\theta[L_t] &= \mathbb{E}_\theta[L_t|\mathcal{E}]\mathbb{P}_\theta(\mathcal{E}) + \mathbb{E}_\theta[L_t|\overline{\mathcal{E}}]\mathbb{P}_\theta(\overline{\mathcal{E}}) \\ &\geq \mathbb{P}_\theta(\mathcal{E})\log\left(\frac{\mathbb{P}_\theta(\mathcal{E})}{\mathbb{P}_\lambda(\mathcal{E})}\right) + \mathbb{P}_\theta(\overline{\mathcal{E}})\log\left(\frac{\mathbb{P}_\theta(\overline{\mathcal{E}})}{\mathbb{P}_\lambda(\overline{\mathcal{E}})}\right) \\ &= d(\mathbb{P}_\theta(\mathcal{E}), \mathbb{P}_\lambda(\mathcal{E}))\end{aligned}$$

Step 2 : Now we rewrite the expectation of the log-likelihood ratio using the tower-property :

$$\begin{aligned}\mathbb{E}_\theta[L_t] =& \mathbb{E}_\theta\left[\mathbb{E}\left[\sum_{s=1}^{t}\sum_{l=1}^{L}\mathbb{1}\{\Lambda_s \geq l\}\right.\right. \\ &\left.\left.\log\frac{d\nu_{a(l)}(\theta)}{d\nu_{a(l)}(\lambda)}(X_{s,l}|\mathcal{F}_{t-1})\right|\mathcal{F}_{t-1}\right]\right]\end{aligned}$$

As the action A_s at time s is determined by the past observations and actions, the central conditional expectation can be scattered into action-related parts :

$$\begin{aligned}\mathbb{E}_\theta[L_t] = \mathbb{E}_\theta&\left[\sum_{s=1}^{t}\sum_{l=1}^{L}\mathbb{1}\{\Lambda_s \geq l\}\sum_{a\in\mathcal{A}}\mathbb{1}\{A_s = a\}\right. \\ &\left.\log\frac{d\nu_{a(l)}(\theta)}{d\nu_{a(l)}(\lambda)}(X_{s,l}|\mathcal{F}_{t-1})\right]\end{aligned}$$

Using the independence of Λ_s with respect to the sequence of feedback at each round, we obtain

$$\begin{aligned}\mathbb{E}_\theta[L_t] = \sum_{s=1}^{t}\sum_{l=1}^{L}\kappa_l\mathbb{E}_\theta&\left[\sum_{a\in\mathcal{A}}\mathbb{1}\{A_s = a\}\right. \\ &\left.\log\frac{d\nu_{a(l)}(\theta)}{d\nu_{a(l)}(\lambda)}(X_{s,l}|\mathcal{F}_{t-1})\right]\end{aligned}$$

and finally, rewriting the above sum using the notations $N_a(t)$ and $K_a(\theta,\lambda)$ previously introduced, we obtain the result

$$\mathbb{E}_\theta[L_t] = \sum_{a\in\mathcal{A}}K_a(\theta,\lambda)\mathbb{E}_\theta[N_a(t)]$$

□

Lemma 8. *Let σ be any stopping time with respect to $(\mathcal{F}_t)$. For every event $A \in \mathcal{F}_\sigma$,*

$$\mathbb{P}_{\nu'}(A) = \mathbb{E}_\nu[\mathbb{1}\{A\}\exp(-L_\sigma)]$$

A full proof of Lemma 8 can be found in the paper [12].

A.2 Proof of Proposition 6

Proof. We denote by $a^*(\theta)$ (resp. $a^*(\lambda)$) the optimal action under bandit model parameterized by θ (resp. λ).

Let λ be in $B(\theta)$ and $\mathcal{E}_T$ be the event

$$\mathcal{E}_T = \left(N_{a^*(\theta)}(T) \leq T - \sqrt{T}\right)$$

The event $\mathcal{E}_T$ is not very likely to hold under bandit model parameterized by θ since the optimal arm $a^*(\theta)$ should be pulled in order of $T - O(\log T)$. On the opposite, $\mathcal{E}_T$ is very likely to hold under model parameterized by λ because $a^*(\theta)$ is not an optimal action anymore. Thus, it should be chosen little compared to $a^*(\lambda)$.

Markov inequality gives

$$\mathbb{P}_\theta(\mathcal{E}_T) = \mathbb{P}_\theta\left(\sum_{a \neq a^*(\theta)} N_a(T) \geq \sqrt{T}\right) \leq \frac{\sum_{a \neq a^*(\theta)} \mathbb{E}_\theta[N_a(T)]}{\sqrt{T}}$$

$$\mathbb{P}_\lambda(\mathcal{E}_T^c) = \mathbb{P}_\lambda\left(N_{a^*(\theta)} \geq T - \sqrt{T}\right) \leq \frac{\mathbb{E}_\lambda[N_{a^*(\theta)}(T)]}{T - \sqrt{T}}$$
$$\leq \frac{\sum_{a \neq a^*(\lambda)} \mathbb{E}_\lambda[N_a(T)]}{T - \sqrt{T}}$$

From definition 2, one obtains $\mathbb{P}_\theta(\mathcal{E}_T) \xrightarrow[T\to\infty]{} 0$ and $\mathbb{P}_\lambda(\mathcal{E}_T^c) \xrightarrow[T\to\infty]{} 0$. Therefore, using the definition of binary entropy, we get

$$\frac{d(\mathbb{P}_\theta(\mathcal{E}_T), \mathbb{P}_\lambda(\mathcal{E}_T))}{\log T} \sim_{T\to\infty} \frac{1}{\log T} \log\left(\frac{1}{\mathbb{P}_\lambda(\mathcal{E}_T^c)}\right)$$
$$\geq \frac{1}{\log T} \log\left(\frac{T - \sqrt{T}}{\sum_{a \neq a^*(\lambda)} \mathbb{E}_\lambda[N_a(T)]}\right)$$

The RHS rewrites

$$1 + \frac{\log(1 - \frac{1}{\sqrt{T}})}{\log T} - \frac{\log(\sum_{a \neq a^*(\lambda)} \mathbb{E}_\lambda[N_a(T)])}{\log T} \xrightarrow[T\to\infty]{} 1$$

Here, we used again definition 2 which tells us that $\sum_{a \neq a^*(\lambda)} \mathbb{E}_\lambda[N_a(T)] = o(T^\alpha)$ for any $\alpha \in]0, 1]$.

Now, for all $\lambda \in B_k(\theta)$ Why $B_k(\theta)$?, lemma 7 applied with event $\mathcal{E}_T$ gives

$$\liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}} K_a(\theta,\lambda)\mathbb{E}_\theta[N_a(T)]}{\log T} \tag{6}$$
$$= \liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}_k} K_a(\theta,\lambda)\mathbb{E}_\theta[N_a(T)]}{\log T} \tag{7}$$
$$\geq \liminf_{T\to\infty} \frac{d(\mathbb{P}_\theta(\mathcal{E}_T), \mathbb{P}_\lambda(\mathcal{E}_T))}{\log T} = 1 \tag{8}$$

where equality 7 is due to the elimination of all null terms in the sum that appears when arm k is not selected in action a. □

A.3 Inequality for suboptimal pulls

Suppose arm k is drawn in position $l \le L$. The overall regret of that action can be minimized by pulling arms $1, ..., L-1$, which are the best possible ones, in decreasing order. Acting so leads to a null regret on first $l-1$ positions and negative regret on last $L-l$ positions. The same reasoning can be done for multiple suboptimal pulls. We summarize this property in the following lemma.

Lemma 9. *Let a be an action containing S suboptimal arms $k_1, \ldots, k_S$ in positions $l_1, \ldots, l_S$, we have $d_a(\theta) \ge \sum_{i=1}^{S} \kappa_{l_i}(\theta_L - \theta_{k_i})$. Or, more generally,*

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

Proof. The regret of action a containing 1 suboptimal arm and optimally completed can be decomposed as follows.

$$\begin{aligned}
d_a(\theta) &= 0 + \kappa_l(\theta_l - \theta_k) + \overbrace{\sum_{j=l+1}^{L} \kappa_j (\theta_j - \theta_{j-1})}^{\text{negative rest}} \\
&= \kappa_l(\theta_l - \theta_{l+1} + \theta_{l+1} - \theta_{l+2} + ... + \theta_L - \theta_k) + \\
&\sum_{j=l+1}^{L} \kappa_j (\theta_j - \theta_{j-1}) \\
&= \kappa_l(\theta_L - \theta_k) + \sum_{j=l+1}^{L} (\theta_{j-1} - \theta_j)(\kappa_l - \kappa_j) \\
&> \kappa_l(\theta_L - \theta_k).
\end{aligned}$$

The result for multiple suboptimal plays can be obtained by decomposing in the same way both the positive regret incurred by each suboptimal arm and the negative rest resulting from the shift of optimal arms. □

Notice that this means that pulling arm k in position $l \le L$ is always worst than pulling it in position L. Moreover, if the action a contains more than one suboptimal arm, the sharpest lower-bound is obtained by considering the *worst* suboptimal arm in the list.

A.4 Proof of Theorem 4

In order to prove the simplified lower-bound of Theorem 4 we basically have two arguments :

1. a lower-bound on $c(\theta)$ can be obtained by widening the feasible set, that is by relaxing some constraints;

2. the Lemma 9 can be used to lower-bound the objective function of the problem.

The constant $c(\theta)$ is defined by

$$c(\theta) = \inf_{c \succeq 0} \sum_{a \neq a^*(\theta)} d_a(\theta) c_a \tag{9}$$

$$s.t \inf_{\lambda \in B(\theta)} \sum_{a \in \mathcal{A}} K_a(\theta, \lambda) c_a \geq 1. \tag{10}$$

We begin by relaxing some constraints : we only allow the change of measure λ to belong to the sets $B_k(\theta)$ defined in Section 4, Equation (3) :

$$c(\theta) = \inf_{c \succeq 0} \sum_{a \neq a^*(\theta)} d_a(\theta) c_a \tag{11}$$

$$s.t\ \forall k \notin a^*(\theta),\ \forall \lambda \in B_k(\theta), \sum_{a \in \mathcal{A}} K_a(\theta, \lambda) c_a \geq 1. \tag{12}$$

The $K - L$ constraints (12) only let one parameter move and must be true for any value satisfying the definition of the corresponding set $B_k(\theta)$. In practice, for each k, the parameter λ_k must be set to at least θ_L. Consequently, these constraints may then be rewritten

$$\forall k \notin a^*(\theta) \sum_{a \neq a^*(\theta)} c_a \sum_{l=1}^{L} \mathbb{1}\{a(l) = k\} \kappa_l \, \mathrm{KL}(\theta_k, \theta_L) \geq 1$$

where we used the fact that the hardest constraint to satisfy is when $\lambda_k = \theta_L$, which allows to reach the infimum of the above sum. Notice that the above sum actually does not contain all coefficients of suboptimal actions c_a: only appear coefficients of actions containing the suboptimal arm k in any position. Before going on, remark that we end up with a set of constraints that do not concern actions containing no suboptimal arm. Consequently, those coefficients can be set to any value such as 0.

Finally, it remains to lower bound the objective function of the optimization problem (9). To do so, the actions containing suboptimal arms must be carefully treated because they are those which induce the highest regret. Then, because of Lemma 9, we can lower bound the objective function as follows

$$\sum_{a \neq a^*(\theta)} d_a(\theta) c_a$$

$$\geq \sum_{k \notin a^*(\theta)} (\theta_L - \theta_k) \sum_{a \neq a^*(\theta)} c_a \sum_{l=1}^{L} \kappa_l \mathbb{1}\{a(l) = k\} \tag{13}$$

$$\geq \sum_{k \notin a^*(\theta)} \frac{\theta_L - \theta_k}{\mathrm{KL}(\theta_k, \theta_L)} \tag{14}$$

where inequality (13) is a due to Lemma 9 and the last inequality is obtained by plugging in the constraints of the optimization problem previously rewritten.