Parametric Graph for Unimodal Ranking Bandit

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, Boammani Aser Lompo

Parametric Graph for Unimodal Ranking Bandit

Camille-Sovanneary Gauthier [*1 2] **Romaric Gaudel** [*3] **Elisa Fromont** [4 5 2] **Boammani Aser Lompo** [6]

Abstract

We tackle the online ranking problem of assigning L items to K positions on a web page in order to maximize the number of user clicks. We propose an original algorithm, easy to implement and with strong theoretical guarantees to tackle this problem in the Position-Based Model (PBM) setting, well suited for applications where items are displayed on a grid. Besides learning to rank, our algorithm, GRAB (for *parametric Graph for unimodal RAnking Bandit*), also learns the parameter of a compact graph over permutations of K items among L. The logarithmic regret bound of this algorithm is a direct consequence of the unimodality property of the bandit setting with respect to the learned graph. Experiments against state-of-the-art learning algorithms which also tackle the PBM setting, show that our method is more efficient while giving regret performance on par with the best known algorithms on simulated and real life datasets.

1. Introduction

Online Recommendation Systems are used to choose relevant items, such as songs, adds or movies on a website. At each call, they select K items among L potential ones, $K \leq L$. User feedbacks are collected for each displayed items, reflecting how relevant these choices are: listening time, clicks, rates, etc. These feedbacks are available only for the items which were presented to the user. This corresponds to an instance of the *multi-armed bandit problem* with *semi-bandit feedback* (Gai et al., 2012; Chen et al., 2013). Another problem, related to ranking, is to display the K chosen items at the right positions to maximize the user attention. Typical examples of such displays are (i) a list of news, visible one by one by scrolling; (ii) a list of products, arranged by rows; or (iii) advertisements spread in a web page. Numerous approaches have been proposed to jointly learn how to choose the best positions for the corresponding best items (Radlinski et al., 2008; Combes et al., 2015; Li et al., 2019) referred as *multiple-play bandit* or *online learning to rank* (OLR). To take into account the user behaviour while facing such a list of items, several models exist (Richardson et al., 2007; Craswell et al., 2008) and have been transposed to the bandit framework (Kveton et al., 2015a; Komiyama et al., 2017) such as the *Position-Based Model* (PBM) (Richardson et al., 2007). PBM allows to take into account displays where the best position is *a priori* unknown. This is typically the case when items are displayed on a grid and not in an ordered list. PBM assumes that the click probability on an item i at position k results from the product of two independent factors: the item relevance and its position's visibility. Items displayed at other positions do not impact the probability to consider the item i at position k. According to PBM, a user may give more than one feedbacks: she may click on all items relevant for her, e.g. when looking for product on commercial websites. PBM is also particularly interesting when the display is dynamic, as often on modern web pages, and may depend on the reading direction of the user (which varies from one country to another) and on the ever-changing layout of the page.

In this paper, we tackle an online learning to rank bandit setting, which mainly covers PBM click model, with an *unimodal bandits* point of view (Combes & Proutière, 2014). First, we expose a family of parametric graphs of degree $L-1$ over permutations, such that the PBM setting is unimodal w.r.t. one graph in this family. While the corresponding graph is unknown from the learner, graphs of this family enable an efficient exploration strategy of the set of potential recommendations. Secondly, we introduce a new bandit algorithm, GRAB, which learns online the appropriate graph in this family and bases its recommendations on the learned graph. From an application point of view, this algorithm has several interesting features: it is simple to implement and efficient in terms of computation time; it handles the PBM bandit setting without any knowledge on the impact of positions (contrarily to many competitors); and it empirically exhibits a regret on par with other theoretically proven

[*]Equal contribution [1]Louis Vuitton, F-75001 Paris, France [2]IRISA UMR 6074 / INRIA rba, F-35000 Rennes, France [3]Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France [4]Univ. Rennes 1, F-35000 Rennes, France [5] Institut Universitaire de France, M.E.S.R.I., F-75231 Paris [6]ENS Rennes, F-35000 Rennes, France. Correspondence to: Camille-Sovanneary Gauthier <camille-sovanneary.gauthier@louisvuitton.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Table 1. Settings and upper-bound on cumulative regret for state of the art algorithms. The main notations for the assumptions are given in Section 3. $\mathcal{N}_{\pi^*}(\boldsymbol{a}^*)$ is a set of recommendations in the neighborhood of the best recommendation. K_{max} is the maximum number of differences between two arms; see. Theorem 2 for a specific definition.

ALGORITHM	HANDLED SETTINGS	REGRET	Δ, ASSUMING $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_L$
GRAB (OUR ALGORITHM)	PBM	$\mathcal{O}\left(\frac{L}{\Delta}\log T\right)$	$\min_{\boldsymbol{a}\in\mathcal{N}_{\pi^*}(\boldsymbol{a}^*)} \mu^* - \mu_{\boldsymbol{a}}$
COMBUCB1 (KVETON ET AL., 2015B)	PBM	$\mathcal{O}\left(\frac{LK^2}{\Delta}\log T\right)$	$\min_{\boldsymbol{a}\in\mathcal{P}_K^L} \mu^* - \mu_{\boldsymbol{a}}$
PBM-PIE (LAGRÉE ET AL., 2016)	PBM WITH $\boldsymbol{\kappa}$ KNOWN	$\mathcal{O}\left(\frac{L-K}{\Delta}\log T\right)$	$\min_{i\in\{K+1,\dots,L\}} \mu^* - \mu_{\boldsymbol{a}[K:=i]}$
PMED-HINGE (KOMIYAMA ET AL., 2017)	PBM WITH $\kappa_1 \geqslant \cdots \geqslant \kappa_K$	$\mathcal{O}\left(c^*\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right)\log T\right)$	$\varnothing$
TOPRANK (LATTIMORE ET AL., 2018)	PBM WITH $\kappa_1 \geqslant \cdots \geqslant \kappa_K$, CM, ...	$\mathcal{O}\left(\frac{LK}{\Delta}\log T\right)$	$\min_{(j,i)\in[L]\times[K]:j>i} \frac{\theta_i-\theta_j}{\theta_i}$
OSUB (COMBES & PROUTIÈRE, 2014)	UNIMODAL	$\mathcal{O}\left(\frac{\gamma}{\Delta}\log T\right)$	$\min_{\boldsymbol{a}\in\mathcal{N}_G(\boldsymbol{a}^*)} \mu^* - \mu_{\boldsymbol{a}}$
KL-COMBUCB (THEOREM 2)	COMBINATORIAL	$\mathcal{O}\left(\frac{\lvert\mathcal{A}\rvert K_{max}^2}{\Delta}\log T\right)$	$\min_{\boldsymbol{a}\in\mathcal{A}} \mu^* - \mu_{\boldsymbol{a}}$

algorithms on both artificial and real datasets. In particular, we prove a $O(L/\Delta\log T)$ regret upper-bound for GRAB. The corresponding proof extends OSUB's proof (Combes & Proutière, 2014) both (i) to the context of a graph learned online, and (ii) to the combinatorial semi-bandit setting.

This paper is organized as follows: Section 2 presents the related work and Section 3 defines our target setting. We introduce GRAB and the hypotheses needed in Section 4. Theoretical guarantees and empirical performance are presented respectively in Section 5 and 6. We conclude in Section 7.

2. Related Work

A comparison of the assumptions and the regret upper-bounds of the related algorithms is shown in Table 1.

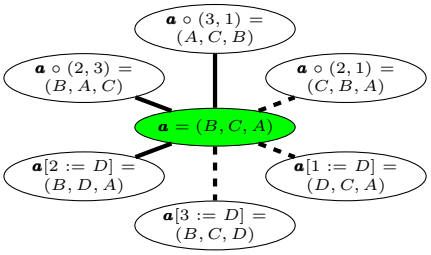
The Position-Based Model (PBM) (Richardson et al., 2007; Craswell et al., 2008) relies on two vectors of parameters: $\boldsymbol{\theta} \in [0,1]^L$ and $\boldsymbol{\kappa} \in [0,1]^K$, where θ_i is the probability for the user to click on item i when she observes this item, and κ_k is the probability for the user to observe position k. Several bandit algorithms are designed to handle PBM (Komiyama et al., 2015; Lagrée et al., 2016; Komiyama et al., 2017). However, each of them assumes some knowledge about the ranking of positions. (Komiyama et al., 2015) and (Lagrée et al., 2016) assume $\boldsymbol{\kappa}$ known beforehand. Thanks to this very strong assumption (that we do not make in this paper), the theoretical results from (Lagrée et al., 2016) depend on the $L-K$ worst items and their regret is expressed as $\mathcal{O}((L-K)/\Delta\log T)$. (Komiyama et al., 2017) and (Gauthier et al., 2021) propose respectively PMED and PB-MHB, the only approaches learning both $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ while recommending. However, PMED still requires the κ_k values to be organized in decreasing order. It derives a bound on the regret in $\mathcal{O}(c^*\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right)\log T)$, where $c^*\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right)$ only depends on $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ and is asymptotically optimal in this setting. Unfortunately, to the best of our knowledge, there is no known closed-form for $c^*\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right)$, which hinders the comparison to other algorithms, including ours. (Gauthier et al., 2021) has shown very good performances on PBM, but PB-MHB, based on Thomson sampling, does not have any theoretical guarantees. Other learning to rank algorithms such as TopRank (Lattimore et al., 2018) and BubbleRank (Li et al., 2019) cover many click models (including PBM). They exhibit a regret upper-bound for T iterations of $\mathcal{O}(LK/\Delta\log T)$, where Δ depends on the attraction probability $\boldsymbol{\theta}$ of items. They also assume that the best recommendation is the one displaying the items from the most attractive to the K-ith most attractive, which implies that the first position is the most-observed one, the second position is the second most-observed one, and so on.

Although the hypotheses taken by PMED and TopRank are often assumed by the approaches handling PBM setting. In this paper we tackle a full PBM setting where there is no a priori hypothesis on the ordering of positions. Our algorithm, GRAB, suffers a $\mathcal{O}(L/\Delta\log T)$ regret that we conjecture to be on par with the best theoretical results provided by PMED ($\mathcal{O}(c^*\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right)\log T)$).

Combinatorial algorithms (Gai et al., 2012; Chen et al., 2013) can also handle the PBM bandit setting. Typically, CombUCB1 (Kveton et al., 2015b) applied to PBM leads to an algorithm which suffers a $\mathcal{O}(LK^2/\Delta\log T)$ regret (see the appendix for more details), which is higher than the upper-bound on the regret of GRAB by a factor K^2. Note that the proof of the upper-bound on the regret of GRAB is based on the same reduction of the PBM bandit

$$\boldsymbol{\rho} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1.00 & 0.90 & 0.80 \\ 0.90 & 0.81 & 0.72 \\ 0.80 & 0.72 & 0.64 \\ 0.70 & 0.63 & 0.56 \end{bmatrix} \end{array}$$

$\boldsymbol{a} = (B, C, A)$
$\boldsymbol{\pi} = (1, 3, 2)$
$\Pi_{\boldsymbol{\rho}}(\boldsymbol{a}) = \{\boldsymbol{\pi}\}$



$$\begin{array}{ccccccccccc} \rho_{a_1,1} & = & 0.90 & \geqslant & \rho_{a_2,2} & = & 0.72 & < & \rho_{a_3,3} & = & 0.80 \\ \rho_{a_{\pi_1},\pi_1} & = & 0.90 & \geqslant & \rho_{a_{\pi_2},\pi_2} & = & 0.80 & \geqslant & \rho_{a_{\pi_3},\pi_3} & = & 0.72 \end{array}$$

$$\mu_{\boldsymbol{a}\circ(3,1)} = 1.00 + 0.72 + 0.72 = 2.44 > \mu_{\boldsymbol{a}} = 0.90 + 0.72 + 0.80 = 2.42$$

Figure 1. Assumption 1 in practice. To distinguish between items and positions, the 4 items are denoted A, B, C, and D. **On the left**: parameters and considered recommendation $\boldsymbol{a}$. We consider a matrix of probabilities of clicks $\boldsymbol{\rho}$ which corresponds to a PBM click model, and a sub-optimal recommendation $\boldsymbol{a}$. The corresponding set $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$ of appropriate rankings of positions is composed of a unique permutation $\boldsymbol{\pi}$. **On the right**: corresponding neighborhoods. Solid lines identify the neighborhood $\mathcal{N}_{\boldsymbol{\pi}}(\boldsymbol{a})$ used by GRAB, and both solid and dashed lines correspond to the neighborhood $\mathcal{N}_G(\boldsymbol{a})$ used by S-GRAB. Note that there is a recommendation better than $\boldsymbol{a}$ in both neighborhoods: $\boldsymbol{a} \circ (3,1) = (A, C, B)$.

to a combinatorial semi-bandit, but with two additional properties derived from the design of GRAB.

Finally, GRAB extends the *unimodal* bandit setting (Combes & Proutière, 2014) which assumes the existence of a known graph G carrying a partial order on the set of bandit arms denoted $\mathcal{A}$. The unimodal bandit algorithms are aware of G, but ignore the partial order induced by the edges of G. However, they rely on G to efficiently browse the arms up to the best one. Typically, the algorithm OSUB (Combes & Proutière, 2014) selects at each iteration t, an arm $\boldsymbol{a}(t)$ in the neighborhood $\mathcal{N}_G(\tilde{\boldsymbol{a}}(t))$ given G of the current best arm $\tilde{\boldsymbol{a}}(t)$ (a.k.a. the *leader*). By restricting the exploration to this neighborhood, the regret suffered by OSUB scales as $\mathcal{O}(\gamma/\Delta \log T)$, where γ is the maximum degree of G, to be compared with $\mathcal{O}(|\mathcal{A}|/\Delta \log T)$ if the arms were independent.

3. Learning to Rank in a Semi-Bandit Setting

We consider the following *online learning to rank (OLR) problem with clicks feedback* which encompasses the PBM setting. For any integer n, let $[n]$ denote the set $\{1, \dots, n\}$. An instance of our OLR problem is a tuple $(L, K, (\rho_{i,k})_{(i,k)\in[L]\times[K]})$, where L is the number of items to be displayed, $K \leqslant L$ is the number of positions to display the items, and for any item i and position k, $\rho_{i,k}$ is the probability for a user to click on item i when displayed at position k, independently of the items displayed at other positions. Under PBM click-model, there exists two vectors $\boldsymbol{\theta} \in \mathbb{R}^L$ and $\boldsymbol{\kappa} \in \mathbb{R}^K$, such that $\rho_{i,k} = \theta_i \kappa_k$ (i.e. $\boldsymbol{\rho}$ is of rank 1).

A recommendation algorithm is only aware of L and K and has to deliver T consecutive recommendations. At each iteration $t \in [T]$, the algorithm recommends a permutation $\boldsymbol{a}(t) = (a_1(t), \dots, a_K(t))$ of K distinct items among L, where $a_k(t)$ is the item displayed at position k. We denote $\mathcal{A} = \mathcal{P}_K^L$ the set of such permutations, which corresponds to the set of arms of the bandit setting. Throughout the paper, we will use the terms *permutation* and *recommendation* interchangeably to denote an element of $\mathcal{P}_K^L$. Thereafter, the algorithm observes the clicks vector $\boldsymbol{c}(t) \in \{0,1\}^K$, where for any position k, $c_k(t)$ equals 1 if the user clicks on item $a_k(t)$ displayed at position k, and 0 otherwise. Note that the recommendation at time t is only based on previous recommendations and observations.

While the individual clicks are observed, the reward of the algorithm is their total number $r(t) \overset{def}{=} \sum_{k=1}^K c_k(t)$. Let $\mu_{\boldsymbol{a}}$ denote the expectation of $r(t)$ when the recommendation is $\boldsymbol{a}(t) = \boldsymbol{a}$, and $\mu^* \overset{def}{=} \max_{\boldsymbol{a}\in\mathcal{P}_K^L} \mu_{\boldsymbol{a}}$ the highest expected reward. The aim of the algorithm is to minimize the *cumulative (pseudo-) regret*

$$R(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^T \mu_{\boldsymbol{a}(t)}\right], \tag{1}$$

where the expectation is taken w.r.t. the recommendations from the algorithm and the clicks. Note that for any recommendation $\boldsymbol{a} \in \mathcal{P}_K^L$, $\mu_{\boldsymbol{a}} = \sum_{k=1}^K \rho_{a_k,k}$.

3.1. Modeling Assumption

Apart from the independency of the clicks, the proposed algorithm assumes a relaxed version of unimodality. Here we present this assumption and state its relation with PBM. We first define the set of *appropriate rankings of positions*: for each recommendation $\boldsymbol{a} \in \mathcal{P}_K^L$, we denote $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a}) \subseteq \mathcal{P}_K^K$ the set of permutations $\boldsymbol{\pi}$ of the K positions such that $\rho_{a_{\pi_1},\pi_1} \geqslant \rho_{a_{\pi_2},\pi_2} \geqslant \cdots \geqslant \rho_{a_{\pi_K},\pi_K}$. Therefore, an appropriate ranking of positions orders the positions from the one with the highest probability of click to the one with the

lowest probability of click. See Figure 1 for an example.

With this notation, our assumption is the following:

Assumption 1 (Relaxed Unimodality)**.** *For any recommendation $\boldsymbol{a} \in \mathcal{P}_K^L$ and any ranking of positions $\boldsymbol{\pi} \in \Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$, if $\mu_{\boldsymbol{a}} \neq \mu^*$, then either there exists $k \in [K-1]$ such that*

$$\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a} \circ (\pi_k, \pi_{k+1})} \tag{2}$$

or there exists $i \in [L] \setminus \boldsymbol{a}([K])$ such that

$$\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a}[\pi_K := i]}, \tag{3}$$

where

- *$\boldsymbol{a} \circ (\pi_k, \pi_{k+1})$ is the permutation for which the items at positions π_k and π_{k+1} are swapped,*
- *$\boldsymbol{a}[\pi_K := i]$ is the permutation which is the same as $\boldsymbol{a}$ for any position $k \neq \pi_K$, and such that $\boldsymbol{a}[\pi_K := i]_{\pi_K} = i$,*
- *and $\boldsymbol{a}([K])$ is the set of items recommended by $\boldsymbol{a}$, namely $\boldsymbol{a}([K]) \overset{def}{=} \{a_1, \dots, a_K\}$.*

Assumption 1 relates to a natural property of standard click models: (i) for the optimal recommendation, the position with the k-th highest probability to be observed is the one displaying the k-th most attractive item, (ii) for a sub-optimal recommendation, swapping two consecutive items, given this order, leads to an increase of the expected reward. However, Assumption 1 considers the order based on the click probabilities $\rho_{a_k,k}$, not on the observation probabilities κ_k. Figure 1 gives an example of both orders and of the neighborhood associated to the ranking $\boldsymbol{\pi}$ defined after the order on click probabilities $\rho_{a_k,k}$.

While the existence of a better recommendation in the neighborhood defined given this order is less intuitive, it remains true for state of the art click models (PBM, the cascading model, and the dependent click model) and paves the way to an algorithm based on observed random variables. Note also that while there exists a better recommendation both in the neighborhood based on the order on observation probability and in the neighborhood based on the order on click probability, this is not true for any neighborhood based on any arbitrary order (as soon as $K \geqslant 4$).

Hereafter, Lemma 1, states that Assumption 1 is weaker than the PBM one. The proof of this Lemma is deferred to the appendix.

Lemma 1. *Let $(L, K, (\theta_i \kappa_k)_{(i,k) \in [L] \times [K]})$ be an online learning to rank problem with users following PBM, with positive probabilities of looking at a given position. Then Assumption 1 is fulfilled.*

3.2. Relation with Unimodality

Assumption 1 relates to the unimodality of the set of expected rewards $(\mu_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{P}_K^L}$. Let us first recall the definition of unimodality in (Combes & Proutière, 2014) and then express this relation.

Definition 1 (Unimodality)**.** *Let $\mathcal{A}$ be a set of arms, and $(\nu_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{A}}$ a set of reward distributions of respective expectations $(\mu_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{A}}$. Let $G = (V, E)$ be an undirected graph with vertices $V = \mathcal{A}$ and edges $E \subseteq V^2$. The set of expected rewards $(\mu_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{A}}$ is* unimodal w.r.t. *G, if and only if:*

1. *the set of expected rewards admits a unique best arm:* $\operatorname{argmax}_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}} = \{\boldsymbol{a}^*\}$*;*
2. *and from any arm $\boldsymbol{a} \neq \boldsymbol{a}^*$, there exists a path $(\boldsymbol{a}^0, \boldsymbol{a}^1, \dots, \boldsymbol{a}^n)$ in G such that $\boldsymbol{a}^0 = \boldsymbol{a}$, $\boldsymbol{a}^n = \boldsymbol{a}^*$, and $\forall i \in [n], \mu_{\boldsymbol{a}^i} > \mu_{\boldsymbol{a}^{i-1}}$.*

Note that the second property of unimodal sets of expected rewards is equivalent to the property stating that *from any sub-optimal arm $\boldsymbol{a}$, there exists an arm $\boldsymbol{a}' \in \mathcal{N}_G(\boldsymbol{a})$ such that $\mu_{\boldsymbol{a}'} > \mu_{\boldsymbol{a}}$*, where $\mathcal{N}_G(\boldsymbol{a})$ is the neighborhood of $\boldsymbol{a}$ in G.

Let's assume that there exists a unique recommendation $\boldsymbol{a}^*$ with maximum expected reward, and denote $\mathcal{F} = (\boldsymbol{\pi}_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{P}_K^L}$ a set of rankings of positions such that for any recommendation $\boldsymbol{a}$, $\boldsymbol{\pi}_{\boldsymbol{a}} \in \Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$. Then, by denoting $G_{\mathcal{F}} = (V, E_{\mathcal{F}})$ the directed graph with vertices $V = \mathcal{P}_K^L$ and edges

$$\begin{aligned} E_{\mathcal{F}} \overset{def}{=} & \left\{ (\boldsymbol{a}, \boldsymbol{a} \circ (\pi_{\boldsymbol{a}k}, \pi_{\boldsymbol{a}(k+1)})) : k \in [K-1] \right\} \\ & \cup \left\{ (\boldsymbol{a}, \boldsymbol{a}[\pi_{\boldsymbol{a}K} := i]) : i \in [L] \setminus \boldsymbol{a}([K]) \right\}, \end{aligned}$$

$(\mu_{\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{P}_K^L}$ is unimodal[1] with respect to $G_{\mathcal{F}}$. Note that this graph is unknown from the algorithm as it builds upon the unknown mapping $\mathcal{F}$. However, this mapping may be learned online, paving the way to an OSUB-like algorithm to explore the space of recommendations.

4. GRAB Algorithm

Our algorithm, GRAB, takes inspiration from the unimodal bandit algorithm OSUB (Combes & Proutière, 2014) by selecting at each iteration t an arm $\boldsymbol{a}(t)$ in the neighborhood of the current best arm (a.k.a. the leader). While in OSUB the neighborhood is known beforehand, here we learn it online. GRAB is described in Algorithm 1. This algorithm uses the following notations:

[1] While the definition of unimodality in (Combes & Proutière, 2014) involves an **undirected** graph, OSUB only requires a **directed** graph and the existence of a strictly increasing path from any sub-optimal arm to the optimal one.

At each iteration t, we denote

$$\hat{\rho}_{i,k}(t) \overset{def}{=} \frac{1}{T_{i,k}(t)} \sum_{s=1}^{t-1} \mathbb{1}\{a_k(s) = i\} c_k(s)$$

the average number of clicks obtained at position k when displaying item i at this position, where

$$T_{i,k}(t) \overset{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{a_k(s) = i\}$$

is the number of time item i has been displayed at position k; $\hat{\rho}_{i,k}(t) \overset{def}{=} 0$ when $T_{i,k}(t) = 0$.

We also denote $\tilde{\boldsymbol{a}}(t)$ the *leader*, meaning the recommendation with the best *pseudo average reward* $\bar{\mu}_{\boldsymbol{a}}(t) \overset{def}{=} \sum_{k=1}^{K} \hat{\rho}_{a_k,k}(t)$, and we note

$$\tilde{T}_{\boldsymbol{a}}(t) \overset{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{\tilde{\boldsymbol{a}}(s) = \boldsymbol{a}\}$$

the number of times the leader is $\boldsymbol{a}$ for iterations 1 to $t-1$.

Finally, the statistics $\hat{\rho}_{i,k}(t)$ are paired with their respective *indices*

$$b_{i,k}(t) \overset{def}{=} f\left(\hat{\rho}_{i,k}(t), T_{i,k}(t), \tilde{T}_{\tilde{\boldsymbol{a}}(t)}(t) + 1\right),$$

where $f(\hat{\rho}, s, t)$ stands for

$$\sup\{p \in [\hat{\rho}, 1] : s \times \mathrm{kl}(\hat{\rho}, p) \leq \log(t) + 3\log(\log(t))\},$$

with

$$\mathrm{kl}(p, q) \overset{def}{=} p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

the *Kullback-Leibler divergence* from a Bernoulli distribution of mean p to a Bernoulli distribution of mean q; $f(\hat{\rho}, s, t) \overset{def}{=} 1$ when $\hat{\rho} = 1$, $s = 0$, or $t = 0$.

At each iteration t, GRAB first identifies the leader $\tilde{\boldsymbol{a}}(t)$, and then recommends either $\tilde{\boldsymbol{a}}(t)$ every L-th iteration, or the best permutation in the inferred neighborhood, given the sum of indices $\sum_{k=1}^{K} b_{a_k,k}(t)$ (see Figure 1 for an example of a neighborhood). Each time an argmax is computed, the ties are randomly broken.

To finish the presentation of GRAB, let us now discuss its initialisation and its time-complexity.

Remark 1 (Initialisation)**.** The initialisation of the algorithm is handled through the default value of indices $b_{i,k}$: 1. This value ensures that any permutation is recommended at least once, as soon as it belongs to the neighborhood of an arm which is often the leader. If a permutation is not in such neighborhood, the theoretical analysis in Section 5 proves that this permutation is sub-optimal, hence it does not matter whether this permutation is explored at least once or not.

Algorithm 1 GRAB: parametric Graph for unimodal RAnking Bandit

Input: number of items L, number of positions K

1: **for** $t = 1, 2, \ldots$ **do**

2: $\tilde{\boldsymbol{a}}(t) \leftarrow \underset{\boldsymbol{a} \in \mathcal{P}_K^L}{\mathrm{argmax}} \sum_{k=1}^{K} \hat{\rho}_{a_k,k}(t)$

3: find $\tilde{\boldsymbol{\pi}}(t)$ s.t. $\hat{\rho}_{\tilde{a}_{\tilde{\pi}_1(t)}(t),\tilde{\pi}_1(t)}(t) \geqslant \hat{\rho}_{\tilde{a}_{\tilde{\pi}_2(t)}(t),\tilde{\pi}_2(t)}(t) \geqslant \cdots \geqslant \hat{\rho}_{\tilde{a}_{\tilde{\pi}_K(t)}(t),\tilde{\pi}_K(t)}(t)$

4: recommend

$$\boldsymbol{a}(t) = \begin{cases} \tilde{\boldsymbol{a}}(t) & \text{, if } \frac{\tilde{T}_{\tilde{\boldsymbol{a}}(t)}(t)}{L} \in \mathbb{N}, \\ \underset{\substack{\boldsymbol{a} \in \{\tilde{\boldsymbol{a}}(t)\} \\ \cup \mathcal{N}_{\tilde{\boldsymbol{\pi}}}(\tilde{\boldsymbol{a}}(t))}}{\mathrm{argmax}} \sum_{k=1}^{K} b_{a_k,k}(t) & \text{, otherwise} \end{cases}$$

where $\mathcal{N}_{\boldsymbol{\pi}}(\boldsymbol{a}) = \{\boldsymbol{a} \circ (\pi_k, \pi_{k+1}) : k \in [K-1]\} \cup \{\boldsymbol{a}[\pi_K := i] : i \in [L] \setminus \boldsymbol{a}([K])\}$

5: observe the clicks vector $\boldsymbol{c}(t)$

6: **end for**

Remark 2 (Algorithmic Complexity)**.** Even though the two optimization steps might seem costly, at each iteration the choice of a recommendation is done in a polynomial time w.r.t. L and K: first, the maximization at Line 2 is a *linear sum assignment problem* which is solvable in $\mathcal{O}\left(K^2(L + \log K)\right)$ time (Ramshaw & Tarjan, 2012); it is not required to scan the $L!/(L-K)!$ permutations of K distinct items among L. Secondly, the maximization at Line 4 is over a set of $L-1$ recommendations and is equivalent to the maximization of

$$B_{\boldsymbol{a}}(t) = \sum_{k=1}^{K} b_{a_k,k}(t) - \sum_{k=1}^{K} b_{\tilde{a}_k(t),k}(t)$$

which reduces to the sum of up to four $b_{a_k,k}(t)$ terms as we are looking at recommendations $\boldsymbol{a}$ in the neighborhood of the leader. Specifically, either

- $\boldsymbol{a} = \tilde{\boldsymbol{a}}(t)$ and $B_{\boldsymbol{a}}(t) = 0$,
- or $\boldsymbol{a} = \tilde{\boldsymbol{a}}(t) \circ (k, k')$ and $B_{\boldsymbol{a}}(t) = b_{\tilde{a}_{k'},k}(t) + b_{\tilde{a}_{k'},k'}(t) - b_{\tilde{a}_k,k}(t) - b_{\tilde{a}_{k'},k'}(t)$,
- or $\boldsymbol{a} = \tilde{\boldsymbol{a}}(t)[k := i]$ and $B_{\boldsymbol{a}}(t) = b_{i,k}(t) - b_{\tilde{a}_k,k}(t)$.

Hence, this maximization requires $\mathcal{O}(L)$ computation time. Overall, the computation time per iteration is a $\mathcal{O}\left(K^2(L + \log K)\right)$.

5. Theoretical Analysis

As already discussed in Section 2, the proof of the upper-bound on the regret of GRAB follows a similar path as the

proof of OSUB (Combes & Proutière, 2014): (1) apply standard bandit analysis to control the regret under the condition that the leader $\tilde{\boldsymbol{a}}(t)$ is the best arm $\boldsymbol{a}^*$, and (2) upper-bound the expected number of iterations such that $\tilde{\boldsymbol{a}}(t) \neq \boldsymbol{a}^*$ by a $\mathcal{O}(\log\log T)$. The inference of the rankings on positions adds up a third step (3) upper-bounding the expected number of iterations such that $\tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}}(t))$.

The first step differs from (Combes & Proutière, 2014), as we have to account for the semi-bandit feedback. We note that when the leader is the best arm, GRAB behaves as a Kullback-Leibler variation of CombUCB1 (Kveton et al., 2015b) that we call KL-CombUCB in the following (see the appendix for a complete definition of KL-CombUCB). We derive an upper-bound specific to KL-CombUCB which accounts for the fact that the maximization at Line 4 of Algorithm 1 can be reduced to the maximization over sums of at most 4 terms (see Remark 2). In the context of GRAB, this new result, expressed by Theorem 2, reduces the regret-bound by a factor K w.r.t. the standard upper-bound for CombUCB1.

The second part of the analysis is based on the fact that with high probability $\bar{\mu}_{\boldsymbol{a}}(t) > \bar{\mu}_{\boldsymbol{a}'}(t)$ if $\mu_{\boldsymbol{a}} > \mu_{\boldsymbol{a}'}$, which derives from the control of the deviation of each $\hat{\rho}_{i,k}(t)$. Here lies the second main difference with Combes & Proutière's analysis: we control the deviation of each individual $\hat{\rho}_{i,k}(t)$ while they control the deviation of $\hat{\mu}_{\boldsymbol{a}}(t) \overset{def}{=} (\sum_{s=1}^{t-1} \mathbb{1}\{\boldsymbol{a}(s) = \boldsymbol{a}\})^{-1} \sum_{s=1}^{t-1} \mathbb{1}\{\boldsymbol{a}(s) = \boldsymbol{a}\} r(s)$. Again, the analysis benefits from the small number of differences between recommendations in the neighborhood of the leader. Moreover, the analysis handles the fact that the neighborhoods may change from an iteration to another, while the neighborhoods are constant in Combes & Proutière's analysis. The corresponding result is expressed, in the following, by Lemma 2.

Finally, the number of iterations at which the inferred ranking on the positions is inappropriate is controlled by Lemma 3. The proof of this lemma is eased by the fact that the number of times the leader is played is at least proportional to the number of times it is the leader.

We now propose and prove the main theorem that upper-bounds the regret of GRAB. Its proof is given after the presentation of all the necessary theorems and lemmas.

Theorem 1 (Upper-Bound on the Regret of GRAB)**.** *Let $\left(L, K, (\rho_{i,k})_{(i,k)\in[L]\times[K]}\right)$ be an online learning to rank problem satisfying Assumption 1 and such that there exists a unique recommendation $\boldsymbol{a}^*$ with maximum expected reward. When facing this problem, GRAB fulfills:*

$$\forall \boldsymbol{a} \in \mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*),\ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t)=\boldsymbol{a}^*,\, \tilde{\boldsymbol{\pi}}(t)=\boldsymbol{\pi}^*,\, \boldsymbol{a}(t)=\boldsymbol{a}\right\}\right] \leqslant \frac{8}{\Delta_{\boldsymbol{a}}^2}\log T + \mathcal{O}(\log\log T), \quad (4)$$

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{a}}(t) \neq \boldsymbol{a}^*\}\right] = \mathcal{O}(\log\log T), \quad (5)$$

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}}(t))\}\right] = \mathcal{O}(1), \quad (6)$$

and hence

$$R(T) \leqslant \sum_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)} \frac{8}{\Delta_{\boldsymbol{a}}}\log T + \mathcal{O}(\log\log T) = \mathcal{O}\left(\frac{L}{\Delta_{min}}\log(T)\right), \quad (7)$$

where $\boldsymbol{\pi}^*$ is the unique ranking of positions in $\Pi_\rho(\boldsymbol{a}^*)$, $\Delta_{\boldsymbol{a}} \overset{def}{=} \mu^* - \mu_{\boldsymbol{a}}$, and $\Delta_{min} \overset{def}{=} \min_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)} \Delta_{\boldsymbol{a}}$.

The first upper-bound (Equation (4)) deals with the expected number of iterations at which GRAB recommends a sub-optimal permutation while the leader is the best permutation. It derives from Theorem 2 hereafter, which detailed proof is in the appendix.

Theorem 2 (New Upper-Bound on the Regret of KL-CombUCB)**.** *We consider a combinatorial semi-bandit setting. Let E be a set of elements and $\mathcal{A} \subseteq \{0,1\}^E$ be a set of arms, where each arm $\boldsymbol{a}$ is a subset of E. Let's assume that the reward when drawing the arm $\boldsymbol{a} \in \mathcal{A}$ is $\sum_{e\in\boldsymbol{a}} c_e$, where for each element $e \in E$, c_e is an independent draw of a Bernoulli distribution of mean $\rho_e \in [0,1]$. Therefore, the expected reward when drawing the arm $\boldsymbol{a} \in \mathcal{A}$ is $\mu_{\boldsymbol{a}} = \sum_{e\in\boldsymbol{a}} \rho_e$.*

When facing this bandit setting, KL-CombUCB (CombUCB1 equiped with Kullback-Leibler indices, see the appendix) fulfills

$$\forall \boldsymbol{a} \in \mathcal{A} \text{ s.t. } \mu_{\boldsymbol{a}} \neq \mu^*,\quad \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\boldsymbol{a}(t) = \boldsymbol{a}\}\right] \leqslant \frac{2K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}^2}\log T + \mathcal{O}(\log\log T),$$

and hence

$$R(T) \leqslant \sum_{\boldsymbol{a}\in\mathcal{A}:\mu_{\boldsymbol{a}}\neq\mu^*} \frac{2K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}}\log T + \mathcal{O}(\log\log T) = \mathcal{O}\left(\frac{|\mathcal{A}|K_{max}^2}{\Delta_{min}}\log(T)\right),$$

where $\mu^* \overset{def}{=} \max_{\boldsymbol{a}\in\mathcal{A}} \mu_{\boldsymbol{a}}$, $\Delta_{\boldsymbol{a}} \overset{def}{=} \mu^* - \mu_{\boldsymbol{a}}$, $\Delta_{min} \overset{def}{=} \min_{\boldsymbol{a}\in\mathcal{A}:\Delta_{\boldsymbol{a}}>0} \Delta_{\boldsymbol{a}}$, $K_{\boldsymbol{a}} \overset{def}{=} \min_{\boldsymbol{a}^*\in\mathcal{A}:\mu_{\boldsymbol{a}^*}=\mu^*} |\boldsymbol{a} \setminus \boldsymbol{a}^*|$ *is the smallest number of elements to remove from* $\boldsymbol{a}$ *to get an optimal arm, and* $K_{max} \overset{def}{=} \max_{\boldsymbol{a}\in\mathcal{A}:\mu_{\boldsymbol{a}}\neq\mu^*} K_{\boldsymbol{a}}$.

Secondly, the expected number of iterations at which the leader is not the optimal arm (Equation (5)) is controlled by Lemma 2, which detailed proof is in the appendix.

Lemma 2 (Upper-Bound on the Number of Iterations of GRAB for which $\tilde{\boldsymbol{a}}(t) \neq \boldsymbol{a}^*$)**.** *Under the hypotheses of Theorem 1 and using its notations,*

$$\forall \tilde{\boldsymbol{a}} \in \mathcal{P}_K^L \backslash \{\boldsymbol{a}^*\},\ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}\}\right] = \mathcal{O}\left(\log\log T\right).$$

Finally, the number of iterations at which the inferred ranking on the positions is inappropriate (Equation (6)) is controlled by Lemma 3, which detailed proof is in the appendix.

Lemma 3 (Upper-Bound on the Number of Iterations of GRAB for which $\boldsymbol{\pi}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})$)**.** *Under the hypothesises of Theorem 1 and using its notations,*

$$\forall \tilde{\boldsymbol{a}} \in \mathcal{P}_K^L,\ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)\right\}\right] = \mathcal{O}\left(1\right).$$

We assemble these results to get the proof of Theorem 1.

Proof of Theorem 1. First note that, since there is a unique optimal permutation, there is a unique appropriate ranking $\boldsymbol{\pi}^*$ of positions w.r.t. $\boldsymbol{a}^*$: $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a}^*) = \{\boldsymbol{\pi}^*\}$. Then, the proof is based on the following decomposition of the set $[T]$ of iterations:

$$[T] = \bigcup_{\substack{\boldsymbol{a}\in\{\boldsymbol{a}^*\}\\ \cup\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)}} \{t \in [T] : \tilde{\boldsymbol{a}}(t) = \boldsymbol{a}^*, \tilde{\boldsymbol{\pi}}(t) = \boldsymbol{\pi}^*, \boldsymbol{a}(t) = \boldsymbol{a}\}$$
$$\cup \{t \in [T] : \tilde{\boldsymbol{a}}(t) \neq \boldsymbol{a}^*\} \cup \{t \in [T] : \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}(t)\right)\}.$$

As for any recommendation $\boldsymbol{a}$, $\Delta_{\boldsymbol{a}} \leqslant K$, this decomposition leads to the inequality $R(T) \leqslant \sum_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)} \Delta_{\boldsymbol{a}} A_{\boldsymbol{a}} + KB + KC$, with

$$A_{\boldsymbol{a}} = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t) = \boldsymbol{a}^*, \tilde{\boldsymbol{\pi}}(t) = \boldsymbol{\pi}^*, \boldsymbol{a}(t) = \boldsymbol{a}\right\}\right],$$
$$B = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t) \neq \boldsymbol{a}^*\right\}\right],$$
$$C = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}(t)\right)\}\right].$$

The term $A_{\boldsymbol{a}}$ is smaller than the expected number of times the arm $\boldsymbol{a}$ is chosen by KL-CombUCB when it plays on the set of arms $\{\boldsymbol{a}^*\} \cup \mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)$. As any of these arms differs with $\boldsymbol{a}^*$ at at most two positions, Theorem 2 upper-bounds $A_{\boldsymbol{a}}$ by

$$\frac{8}{\Delta_{\boldsymbol{a}}^2} \log T + \mathcal{O}\left(\log\log T\right)$$

and hence $\sum_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)} \Delta_{\boldsymbol{a}} A_{\boldsymbol{a}} = \mathcal{O}\left(L/\Delta_{min} \log T\right)$ as $|\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)| = L - 1$.

Note that Theorem 5 of (Kveton et al., 2015b), upper-bounding the regret of CombUCB1, leads to a $\mathcal{O}\left(LK/\Delta \log T\right)$ bound[2] for $\sum_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}^*}(\boldsymbol{a}^*)} \Delta_{\boldsymbol{a}} A_{\boldsymbol{a}}$, which we reduce by a factor K by using Theorem 2.

From Lemma 2, the term B is upper-bounded by

$$B = \sum_{\tilde{\boldsymbol{a}}\in\mathcal{P}_K^L\backslash\{\boldsymbol{a}^*\}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}\right\}\right] = \mathcal{O}\left(\log\log T\right),$$

and we upper-bound the term C with Lemma 3:

$$C = \sum_{\tilde{\boldsymbol{a}}\in\mathcal{P}_K^L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)\}\right] = \mathcal{O}\left(1\right).$$

Finally, the regret of GRAB is upper-bounded by summing these three terms, which concludes the proof.

□

5.1. Discussion

Assumming $\theta_1 \geqslant \cdots \geqslant \theta_L$ and $\kappa_1 \geqslant \cdots \geqslant \kappa_K$, the detailed formula for the regret upper-bound (7) is $\sum_{k=1}^{K-1} \frac{8\log T}{(\kappa_k - \kappa_{k+1})(\theta_k - \theta_{k+1})} + \sum_{k=K+1}^{L} \frac{8\log T}{\kappa_K(\theta_K - \theta_k)}$, where the first sum corresponds to the set of neighbors of $\boldsymbol{a}^*$ which recommend the same items as $\boldsymbol{a}^*$, and the second sum relates to the set of neighbors of $\boldsymbol{a}^*$ which replace the 'last' item in $\boldsymbol{a}^*$. Hence, the number of displayed items does not impact the total number of terms, but the gaps $\Delta_{\boldsymbol{a}}$.

Note also that GRAB is, by design, robust to miss-specifications. Typically, GRAB would properly handle a matrix $\boldsymbol{\rho} = \boldsymbol{\theta}^T\boldsymbol{\kappa} + \mathcal{E}$, if $\max_{i,j} |\mathcal{E}_{i,j}|$ is smaller than half of the minimum gap between two entries of the matrix $\boldsymbol{\theta}^T\boldsymbol{\kappa}$.

However, if there is a set of optimal recommendations $\mathcal{A}^*$ (instead of a unique one), after convergence, the leader will be picked in that set at each iteration. So the neighborhood of each optimal recommendation will be explored, and we will get a regret bound in $\mathcal{O}(|\mathcal{A}^*|L)$. This behavior

[2] In this setting, the *ground set* is $E \overset{def}{=} \bigcup_{k\in[K]}\{(a_{\max(k-1,1)}, k), (a_k, k), (a_{\min(k+1,K)}, k)\} \cup \bigcup_{k\in[L]\backslash[K]}\{(a_k, K)\}$ and is of size $L + 2K - 2$, and any arm is composed of exactly K elements in E.

questions the applicability of unimodality to the *Cascading Model* (CM), as with this model there is at least $K!$ optimal recommendations. Moreover, while Assumption 1 is valid for CM and the *Dependent Click Model* (DCM), our setting also assumes the existence of the matrix $\boldsymbol{\rho}$, which is false for CM and DCM: in both settings the probability of clicking on item i in position ℓ depends on other displayed items.

6. Experiments

In this section, we compare GRAB to PMED (Komiyama et al., 2017), to TopRank (Lattimore et al., 2018), to PB-MHB (Gauthier et al., 2021), to ϵ_n-Greedy, to Static Graph for unimodal RAnking Bandit (S-GRAB), a simplified version of GRAB, and to KL-CombUCB, an adaptation of CombUCB1 (Kveton et al., 2015b) (see the appendix for details regarding S-GRAB and KL-CombUCB). The experiments are conducted on the Yandex dataset (Yandex, 2013) and on purely simulated data. We use the cumulative regret to evaluate the performance of each algorithm, where the cumulative regret is averaged over 20 independent runs of $T = 10^7$ iterations each. Code and data for replicating our experiments are available at `https://github.com/gaudel/ranking_bandits`.

6.1. Experimental Setting

We use two types of datasets: a *simulated* one for which we set the values for $\boldsymbol{\kappa}$ and $\boldsymbol{\theta}$ and a *real* one, where parameters are inferred from real life logs of Yandex search engine (Yandex, 2013). Let's remind that θ_i is the probability for the user to click on item i when it observes this item, and κ_k is the probability for the user to observe position k.

Simulated data allow us to test GRAB in extreme situations. We consider $L = 10$ items, $K = 5$ positions, and $\boldsymbol{\kappa} = [1, 0.75, 0.6, 0.3, 0.1]$. The range of values for $\boldsymbol{\theta}$ is either close to zero ($\boldsymbol{\theta}^- = [10^{-3}, 5.10^{-4}, 10^{-4}, 5.10^{-5}, 10^{-5}, 10^{-6}, \ldots, 10^{-6}]$), or close to one ($\boldsymbol{\theta}^+ = [0.99, 0.95, 0.9, 0.85, 0.8, 0.75, \ldots, 0.75]$).

Real data contain the logs of actions toward the Yandex search engine: 65 million search queries and 167 million hits (clicks). Common use of this database in the bandit setting consists first in extracting from these logs the parameters of the chosen real model, and then in simulating users' interactions given these parameters (Lattimore et al., 2018). We use Pyclick library (Chuklin et al., 2015) to infer the PBM parameters of each query with the *expectation maximization* algorithm. This leads to θ_i values ranging from 0.070 to 0.936, depending on the query. Similarly to (Lattimore et al., 2018), we look at the results averaged on the 10 most frequent queries, while displaying $K = 5$ items among the $L = 10$ most attractive ones.

Among our opponents, TopRank and PMED require decreasing values of $\boldsymbol{\kappa}$ which may not be fulfilled by PBM. We pre-order them to fulfill these algorithms' requirements. Otherwise, $\boldsymbol{\kappa}$ is shuffled at the beginning of each sequence of recommendations. We also carefully tune the exploration hyper-parameter c of ε_n-greedy taking values ranging exponentially from 10^0 to 10^6. For PB-MHB, we use the hyper-parameters recommended in (Gauthier et al., 2021).

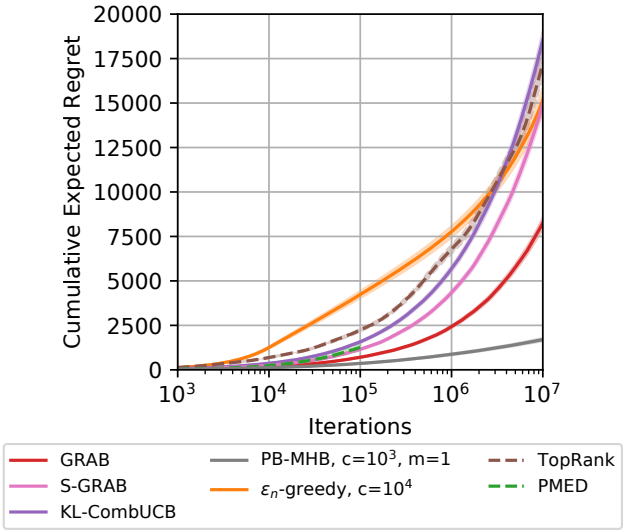
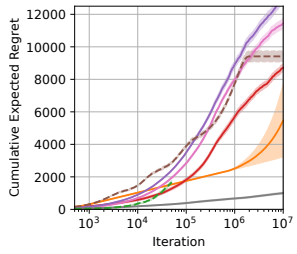


Figure 2. Cumulative regret w.r.t. iterations on Yandex dataset. The plotted curves correspond to the average over 200 independent sequences of recommendations (20 sequences per query). The shaded area depicts the standard error of our regret estimates.

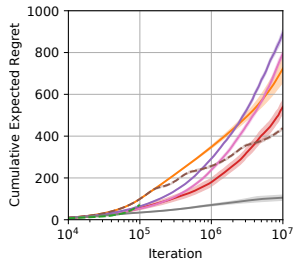
6.2. Results

Figure 2 shows the results for the algorithms on Yandex and Figure 3 on the simulated data. We measure the performance of each algorithm according to the cumulative regret (see Equation 1). It is the sum, over T consecutive recommendations, of the difference between the expected reward of the best answer and of the answer of a given recommender system. The best algorithm is the one with the lowest regret. We average the results of each algorithm over 20 independent sequences of recommendations per query or simulated setting. Although PMED theoretically yields an asymptotically optimal regret, we stop it at iteration $t = 10^5$ due to its heavy computation-time.

Ablation Study The two main ingredients of GRAB are the use of a graph to explore the set of recommendations, and the online inference of this graph. Without these ingredients, GRAB boils down to KL-CombUCB which recommends at each iteration the best permutation given the sum of indices $b_{i,k}$ and has a $\mathcal{O}\left(LK^2/\Delta \log T\right)$ regret. With only the first ingredient (namely a static graph of degree



(a) Parameter $\boldsymbol{\theta}$ close to 1



(b) Parameter $\boldsymbol{\theta}$ close to 0

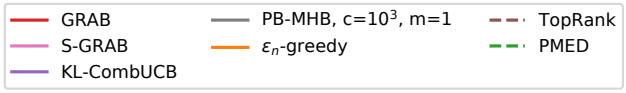


Figure 3. Cumulative regret w.r.t. iterations on simulated data. The plotted curves correspond to the average over 20 independent sequences of recommendations. The shaded area depicts the standard error of our regret estimates. For ϵ_n-Greedy, c is set to 10^5 when $\boldsymbol{\theta}$ is close to 0, and to 10^3 when $\boldsymbol{\theta}$ is close to 1.

$\Theta(LK)$), we get S-GRAB which regret is upper-bounded by $\mathcal{O}\left(LK/\Delta \log T\right)$, while GRAB's regret is upper-bounded by $\mathcal{O}\left(L/\Delta \log T\right)$ thanks to a set of graphs of degree $L-1$.

We want to assert the empirical impact of these ingredients. On Figures 2 and 3, we see that GRAB has a better regret than S-GRAB and KL-CombUCB in every settings. This confirms that the proposed graphs are relevant to explore the set of recommendations, and that GRAB quickly infer the appropriate graph in the family of potential ones.

Results Analysis Figure 2 compares the empirical regret of all algorithms on Yandex dataset. GRAB is the second best with a regret at $T = 10^7$ about two time smaller than the rest of the algorithms. Only PB-MHB yields a smaller regret, but PB-MHB is more than ten times slower to deliver a recommendation than GRAB and it does not have any theoretical guarantees.

Figure 3 shows our results on purely simulated data illustrating extreme settings even though these settings are less realistic. In both settings, GRAB is in the top-3 algorithms. PB-MHB is still the algorithm yielding the best regret. However, while TopRank provides better or similar result as

Table 2. Average computation time for sequences of 10^7 recommendations vs. all queries of Yandex dataset

ALGORITHM	(HOUR/MIN)	TRIAL (MS)
GRAB	2H24	0.9
S-GRAB	9H56	3.6
ε_n-GREEDY $c = 10^4$	1H13	0.4
PB-MHB $c = 10^3$, $m = 1$	44H50	16
KL-COMBUCB	2H03	0.7
PMED	474H13*	170
TOPRANK	9H29	3

* EXTRAPOLATION FROM 10^5 RECOMMENDATIONS.

GRAB at iteration 10^7, its regret is higher than the one of GRAB up to iteration $t = 4 \times 10^6$. TopRank only catchs-up GRAB at the end of the sequences of recommendations. We note that in the setting close to 1, TopRank manages to find the perfect order after 10^6 iterations. In this setting too, ε_n-greedy has better performance during the 10^6 first iterations, but suffers from its greedy behaviour during the last steps with a large variance.

Computation Time As shown in Table 2, the fastest algorithm is ε_n-greedy. KL-CombUCB and GRAB are two times slower. The exploration of S-GRAB multiplies its computation time by 4 compared to GRAB. TopRank is about three times slower than GRAB, and PB-MHB, despite its good regret is one of the slowest algorithm with PMED.

7. Conclusion

Our work targets the full PBM setting, which aims at recommending a ranking of K items among L and display them, without prior knowledge on the attractiveness of the positions. We learn online both the user preferences and their gaze habits. To solve this problem, we define a graph parametrized by rankings of positions, and we extend the unimodal bandit setting to this family of graphs. We also design GRAB, an algorithm that learns online the proper parametrization of the graph, and we prove a regret upper-bound in $O(L/\Delta \log T)$ for this algorithm which reduces by a factor K^2 (respectively K) the bound which would be obtained without the unimodal setting (resp. with the standard unimodal setting). On real and simulated data, GRAB quickly delivers good recommendations.

The extension of the unimodal setting is a promising tool which may benefit to recommendations to users with a more general behavior, or to other combinatorial semi-bandit scenarios. The integration of unimodal bandit algorithms working on parametric spaces (Combes et al., 2020) may also pave the way to efficient contextual recommendation systems handling larger sets of items and positions.

References

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *proc. of the 30th Int. Conf. on Machine Learning*, ICML'13, 2013.

Chuklin, A., Markov, I., and de Rijke, M. *Click Models for Web Search*. Morgan & Claypool Publishers, 2015.

Combes, R. and Proutière, A. Unimodal bandits: Regret lower bounds and optimal algorithms. In *proc. of the 31st Int. Conf. on Machine Learning, ICML'14*, 2014.

Combes, R., Magureanu, S., Proutière, A., and Laroche, C. Learning to rank: Regret lower bounds and efficient algorithms. In *proc. of the ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems*, 2015.

Combes, R., Proutière, A., and Fauquette, A. Unimodal bandits with continuous arms: Order-optimal regret without smoothness. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(1), May 2020.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. An experimental comparison of click position-bias models. In *proc. of the Int. Conf. on Web Search and Data Mining*, WSDM '08, 2008.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Trans. Netw.*, 20(5):1466–1478, October 2012.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *proc. of the 24th Annual Conf. on Learning Theory*, COLT'11, 2011.

Gauthier, C.-S., Gaudel, R., Fromont, E., and Lompo, B. A. Position-based multiple-play bandits with thompson sampling. In *proc. of the 19th Symposium on Intelligent Data Analysis*, IDA'21, 2021.

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *proc. of the 32nd Int. Conf. on Machine Learning*, ICML'15, 2015.

Komiyama, J., Honda, J., and Takeda, A. Position-based multiple-play bandit problem with unknown position bias. In *Advances in Neural Information Processing Systems 30*, NIPS'17, 2017.

Kveton, B., Szepesvári, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *proc. of the 32nd Int. Conf. on Machine Learning*, ICML'15, 2015a.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In *proc. of the 18th Int. Conf. on Artificial Intelligence and Statistics*, AISTATS'15, 2015b.

Lagrée, P., Vernade, C., and Cappé, O. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems 30*, NIPS'16, 2016.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems 31*, NIPS'18, 2018.

Li, C., Kveton, B., Lattimore, T., Markov, I., de Rijke, M., Szepesvári, C., and Zoghi, M. Bubblerank: Safe online learning to re-rank via implicit click feedback. In *proc. of the 35th Uncertainty in Artificial Intelligence Conference*, UAI'19, 2019.

Radlinski, F., Kleinberg, R., and Thorsten, J. Learning diverse rankings with multi-armed bandits. In *proc. of the 25th Int. Conf. on Machine Learning*, ICML'08, 2008.

Ramshaw, L. and Tarjan, R. E. On minimum-cost assignments in unbalanced bipartite graphs. Technical report, HP research labs, 2012.

Richardson, M., Dominowska, E., and Ragno, R. Predicting clicks: Estimating the click-through rate for new ads. In *proc. of the 16th International World Wide Web Conference*, WWW '07, 2007.

Yandex. Yandex personalized web search challenge. 2013. URL https://www.kaggle.com/c/yandex-personalized-web-search-challenge.

The appendix is organized as follows. We first list most of the notations used in the paper in Appendix A. Lemma 1 is proved in Appendix B. In Appendix C, we recall a Lemma from (Combes & Proutière, 2014) used by our own Lemmas and Theorems, and then in Appendices D to F we respectively prove Theorem 2, Lemma 2, and Lemma 3. In Appendix G we define KL-CombUCB and discuss its regret and its relation to GRAB. Finally in Appendix H we introduce and discuss S-GRAB.

A. Notations

The following table summarize the notations used through the paper and the appendix.

SYMBOL	MEANING
T	TIME HORIZON
t	ITERATION
L	NUMBER OF ITEMS
i	INDEX OF AN ITEM
K	NUMBER OF POSITIONS IN A RECOMMENDATION
k	INDEX OF A POSITION
$[n]$	SET OF INTEGERS $\{1,\dots,n\}$
$\mathcal{P}_K^L$	SET OF PERMUTATIONS OF K DISTINCT ITEMS AMONG L
$\boldsymbol{\theta}$	VECTORS OF PROBABILITIES OF CLICK
θ_i	PROBABILITY OF CLICK ON ITEM i
$\boldsymbol{\kappa}$	VECTORS OF PROBABILITIES OF VIEW
κ_k	PROBABILITY OF VIEW AT POSITION k
$\mathcal{A}$	SET OF BANDIT ARMS
$\boldsymbol{a}$	AN ARM IN $\mathcal{A}$
$\boldsymbol{a}(t)$	THE ARM CHOSEN AT ITERATION t
$\tilde{\boldsymbol{a}}(t)$	BEST ARM AT ITERATION t GIVEN THE PREVIOUS CHOICES AND FEEDBACKS (CALLED LEADER)
$\boldsymbol{a}^*$	BEST ARM
G	GRAPH CARRYING A PARTIAL ORDER ON $\mathcal{A}$
γ	MAXIMUM DEGREE OF G
$\mathcal{N}_G(\tilde{\boldsymbol{a}}(t))$	NEIGHBORHOOD OF $\tilde{a}(t)$ GIVEN G
$\rho_{i,k}$	PROBABILITY OF CLICK ON ITEM i DISPLAYED AT POSITION k
$\boldsymbol{c}(t)$	CLICKS VECTOR AT ITERATION t
$r(t)$	REWARD COLLECTED AT ITERATION t, $r(t)=\sum_{k=1}^K c_k(t)$
$\mu_{\boldsymbol{a}}$	EXPECTATION OF $r(t)$ WHILE RECOMMENDING $\boldsymbol{a}$, $\mu_{\boldsymbol{a}}=\sum_{k=1}^K \rho_{a_k,k}$
μ^*	HIGHEST EXPECTED REWARD, $\mu^*=\max_{\boldsymbol{a}\in\mathcal{P}_K^L}\mu_{\boldsymbol{a}}$
Δ_a	GAP BETWEEN μ_a AND μ^*
Δ_{min}	MINIMAL VALUE FOR Δ_a
Δ	GENERIC REWARD GAP BETWEEN ONE OF THE SUB-OPTIMAL ARMS AND ONE OF THE BEST ARMS
$R(T)$	CUMULATIVE (PSEUDO-)REGRET, $R(T)=T\mu^*-\mathbb{E}\left[\sum_{t=1}^T \mu_{\boldsymbol{a}(t)}\right]$
$\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$	SET OF PERMUTATIONS IN $\mathcal{P}_K^K$ ORDERING THE POSITIONS S.T. $\rho_{a_{\pi_1},\pi_1}\geqslant\rho_{a_{\pi_2},\pi_2}\geqslant\cdots\geqslant\rho_{a_{\pi_K},\pi_K}$
$\boldsymbol{\pi}$	ELEMENT OF $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$
$\tilde{\boldsymbol{\pi}}$	ESTIMATION OF $\boldsymbol{\pi}$
$\boldsymbol{a}\circ(\pi_k,\pi_{k+1})$	PERMUTATION SWAPPING ITEMS IN POSITIONS π_k AND π_{k+1}
$\boldsymbol{a}[\pi_K:=i]$	PERMUTATION LEAVING $\boldsymbol{a}$ THE SAME FOR ANY POSITION EXCEPT π_K FOR WHICH $\boldsymbol{a}[\pi_K:=i]_{\pi_K}=i$
$\mathcal{F}$	RANKINGS OF POSITIONS RESPECTING $\Pi_{\boldsymbol{\rho}}$, $\mathcal{F}=(\boldsymbol{\pi}_{\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{P}_K^L}$ S.T. $\forall\boldsymbol{a}\in\mathcal{P}_K^L,\boldsymbol{\pi}_{\boldsymbol{a}}\in\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$
$T_{i,k}(t)$	NUMBER OF ITERATIONS S.T. ITEM i HAS BEEN DISPLAYED AT POSITION k, $T_{i,k}(t)=\sum_{s=1}^{t-1}\mathbb{1}\{a_k(s)=i\}$
$\tilde{T}_{\boldsymbol{a}}(t)$	NUMBER OF ITERATIONS S.T. THE LEADER WAS $\boldsymbol{a}$, $\tilde{T}_{\boldsymbol{a}}(t)\overset{def}{=}\sum_{s=1}^{t-1}\mathbb{1}\{\tilde{\boldsymbol{a}}(s)=\boldsymbol{a}\}$
$T_{\boldsymbol{a}}(t)$	NUMBER OF ITERATIONS S.T. THE CHOSEN ARM WAS $\boldsymbol{a}$, $T_{\boldsymbol{a}}(t)=\sum_{s=1}^{t-1}\mathbb{1}\{\boldsymbol{a}(s)=\boldsymbol{a}\}$
$T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t)$	NUMBER OF ITERATIONS S.T. THE LEADER WAS $\tilde{\boldsymbol{a}}$, THE CHOSEN ARM WAS $\boldsymbol{a}$, AND $\boldsymbol{a}$ WAS CHOSEN BY THE ARGMAX ON $\sum_{k=1}^K b_{a_k,k}(t)$: $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t)=\sum_{s=1}^{t-1}\mathbb{1}\left\{\tilde{\boldsymbol{a}}(s)=\tilde{\boldsymbol{a}},\boldsymbol{a}(s)=\boldsymbol{a},\tilde{T}_{\tilde{\boldsymbol{a}}}(s)/L\notin\mathbb{N}\right\}$
$\hat{\rho}_{i,k}(t)$	ESTIMATION OF $\rho_{i,k}$ AT ITERATION t, $\hat{\rho}_{i,k}(t)=\frac{1}{T_{i,k}(t)}\sum_{s=1}^{t-1}\mathbb{1}\{a_k(s)=i\}c_k(s)$
$b_{i,k}(t)$	KULLBACK-LEIBLER INDEX OF $\hat{\rho}_{i,k}(t)$, $b_{i,k}(t)=f\left(\hat{\rho}_{i,k}(t),T_{i,k}(t),\tilde{T}_{\tilde{\boldsymbol{a}}(t)}(t)+1\right)$
f	KULLBACK-LEIBLER INDEX FUNCTION, $f(\hat{\rho},s,t)=\sup\{p\in[\hat{\rho},1]:s\times\mathrm{kl}(\hat{\rho},p)\leq\log(t)+3\log(\log(t))\}$,
$\mathrm{kl}(p,q)$	KULLBACK-LEIBLER DIVERGENCE FROM A BERNOULLI DISTRIBUTION OF MEAN p TO A BERNOULLI DISTRIBUTION OF MEAN q, $\mathrm{kl}(p,q)=p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)$
$B_{\boldsymbol{a}}(t)$	PSEUDO-SUM OF INDICES OF $\boldsymbol{a}$ AT ITERATION T, $B_{\boldsymbol{a}}(t)=\sum_{k=1}^K b_{a_k,k}(t)-\sum_{k=1}^K b_{\tilde{a}_k(t),k}(t)$

CONTINUED ON NEXT PAGE

SYMBOL	MEANING
$\mathcal{N}_{\pi^*}(a^*)$	NEIGHBORHOOD OF THE BEST ARM
$K_{\boldsymbol{a}}$	(WITH COMBINATORIAL BANDIT SETTING) NUMBER OF ELEMENTS IN $\boldsymbol{a}$ BUT NOT IN $\boldsymbol{a}^*$, $K_{\boldsymbol{a}} = \min_{\boldsymbol{a}^* \in \mathcal{A}: \mu_{\boldsymbol{a}^*} = \mu^*} \lvert \boldsymbol{a} \setminus \boldsymbol{a}^* \rvert$
K_{max}	(WITH COMBINATORIAL BANDIT SETTING) MAXIMAL NUMBER OF ELEMENTS IN A SUB-OPTIMAL ARM $\boldsymbol{a}$ BUT NOT IN AN OPTIMAL ARM a^*, $K_{max} = \max_{\boldsymbol{a} \in \mathcal{A}: \mu_{\boldsymbol{a}} \neq \mu^*} K_{\boldsymbol{a}}$
$c^*(\boldsymbol{\theta}, \boldsymbol{\kappa})$	COEFFICIENT IN THE REGRET BOUND OF PMED
c	(IN ε_n-GREEDY) PARAMETER CONTROLLING THE PROBABILITY OF EXPLORATION
c	(IN PB-MHB) PARAMETER CONTROLLING SIZE OF THE STEP IN THE METROPOLIS HASTING INFERENCE
m	(IN PB-MHB) NUMBER OF STEP IN THE METROPOLIS HASTING INFERENCE

B. Proof of Lemma 1 (PBM Fulfills Assumption 1)

Proof of Lemma 1. Let $(L, K, (\rho_{i,k})_{(i,k)\in[L]\times[K]})$ be an online learning to rank (OLR) problem with users following PBM, with positive probabilities of looking at a given position. Therefore, there exists $\boldsymbol{\theta} \in [0,1]^L$ and $\boldsymbol{\kappa} \in (0,1]^K$ such that for any item i and any position k, $\rho_{i,k} = \theta_i \kappa_k$.

Let $\boldsymbol{a} \in \mathcal{P}_K^L$ be a recommendation, and let $\boldsymbol{\pi} \in \Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$ be an appropriate ranking of positions. One of the four following properties is satisfied:

$$\exists k \in [K-1] \text{ s.t. } \theta_{a_{\pi_k}} < \theta_{a_{\pi_{k+1}}}, \tag{8}$$

$$\exists k \in [K-1] \text{ s.t. } \kappa_{\pi_k} < \kappa_{\pi_{k+1}}, \tag{9}$$

$$\exists i \in [L] \setminus \boldsymbol{a}([K]) \text{ s.t. } \theta_{a_{\pi_K}} < \theta_i, \tag{10}$$

$$\begin{cases} \forall k \in [K-1], \theta_{a_{\pi_k}} \geqslant \theta_{a_{\pi_{k+1}}} \\ \forall k \in [K-1], \kappa_{\pi_k} \geqslant \kappa_{\pi_{k+1}} \\ \forall i \in [L] \setminus \boldsymbol{a}([K]), \theta_{a_{\pi_K}} \geqslant \theta_i \end{cases}. \tag{11}$$

Let prove, by considering each of these properties one by one, that $\boldsymbol{a}$ is either one of the best arms, or $\boldsymbol{a}$ fulfills either Property (2) or Property (3) of Assumption 1.

If Property (8) is satisfied and $\theta_{a_{\pi_k}} = 0$, then by definition of $\boldsymbol{\pi}$ and $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$, $0 = \theta_{a_{\pi_k}} \kappa_{\pi_k} \geqslant \theta_{a_{\pi_{k+1}}} \kappa_{\pi_{k+1}} > 0$ which is absurd.

Therefore, If Property (8) is satisfied, $\frac{\theta_{a_{\pi_{k+1}}}}{\theta_{a_{\pi_k}}} > 1$.

Note that by definition of $\boldsymbol{\pi}$ and $\Pi_{\boldsymbol{\rho}}(\boldsymbol{a})$, and as $\rho_{i,k} = \theta_i \kappa_k$, $\theta_{a_{\pi_k}} \kappa_{\pi_k} \geqslant \theta_{a_{\pi_{k+1}}} \kappa_{\pi_{k+1}}$.

Hence $\kappa_{\pi_k} \geqslant \frac{\theta_{a_{\pi_{k+1}}}}{\theta_{a_{\pi_k}}} \kappa_{\pi_{k+1}} > \kappa_{\pi_{k+1}}$, and

$$\begin{aligned} \mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a} \circ (\pi_k, \pi_{k+1})} &= \theta_{a_{\pi_k}} \kappa_{\pi_k} + \theta_{a_{\pi_{k+1}}} \kappa_{\pi_{k+1}} - \left(\theta_{a_{\pi_{k+1}}} \kappa_{\pi_k} + \theta_{a_{\pi_k}} \kappa_{\pi_{k+1}} \right) \\ &= \left(\theta_{a_{\pi_k}} - \theta_{a_{\pi_{k+1}}} \right) \left(\kappa_{\pi_k} - \kappa_{\pi_{k+1}} \right) \\ &< 0, \end{aligned}$$

meaning $\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a} \circ (\pi_k, \pi_{k+1})}$, which corresponds to Property (2) of Assumption 1.

Similarly, if Property (9) is satisfied, then Property (2) of Assumption 1 is fulfilled.

If Property (10) is satisfied,

$$\begin{aligned} \mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}[\pi_K := i]} &= \theta_{a_{\pi_K}} \kappa_{\pi_K} - \theta_i \kappa_{\pi_K} \\ &= \left(\theta_{a_{\pi_K}} - \theta_i \right) \kappa_{\pi_K} \\ &< 0. \end{aligned}$$

Hence $\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a}[\pi_K:=i]}$, which corresponds to Property (3) of Assumption 1.

Finally, if Property (11) is satisfied, $\mu_{\boldsymbol{a}} = \mu^*$.

Overall, either $\boldsymbol{a}$ is one of the best arms, or $\boldsymbol{a}$ fulfills Property (2) of Assumption 1, or $\boldsymbol{a}$ fulfills Property (3) of Assumption 1, which concludes the proof.

□

C. Preliminary to the Analysis of GRAB

The analysis of GRAB requires a control of the number of high deviations, as expressed by Lemma $B.1$ of (Combes & Proutière, 2014). Let us recall this lemma, which we denote Lemma 4 in current paper.

Lemma 4 (Lemma B.1 of (Combes & Proutière, 2014))**.** *Let $i \in [L]$, $k \in [K]$, $\epsilon > 0$. Define $\mathcal{F}(T)$ the σ-algebra generated by $(\boldsymbol{c}(t))_{t\in[T]}$. Let $\Lambda \subseteq \mathbb{N}$ be a random set of instants. Assume that there exists a sequence of random sets $(\Lambda(s))_{s\geq 1}$ such that (i) $\Lambda \subseteq \bigcup_{s\geq 1} \Lambda(s)$, (ii) for all $s \geqslant 1$ and all $t \in \Lambda(s)$, $T_{i,k}(t) \geq \epsilon s$, (iii) $|\Lambda(s)| \leqslant 1$, and (iv) the event $t \in \Lambda(s)$ is $\mathcal{F}_t$-measurable. Then for all $\delta > 0$,*

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\{t \in \Lambda, |\hat{\rho}_{i,k}(t) - \rho_{i,k}| \geqslant \delta\}\right] \leqslant \frac{1}{\epsilon\delta^2}$$

D. Proof of Theorem 2 (Upper-bound on the Regret of KL-CombUCB)

Proof of Theorem 2. Let $\boldsymbol{a} \in \mathcal{A}$ be a sub-optimal arm. Let $\boldsymbol{a}^* \in \mathcal{A}$ be an optimal arm such that $|\boldsymbol{a} \setminus \boldsymbol{a}^*| = K_{\boldsymbol{a}}$.

We denote $\bar{K}_{\boldsymbol{a}} \overset{def}{=} |\boldsymbol{a}^* \setminus \boldsymbol{a}|$, $T_{\boldsymbol{a}}(t) \overset{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{\boldsymbol{a}(s) = \boldsymbol{a}\}$ the number of time the arm $\boldsymbol{a}$ has been drawn, and $T_e(t) \overset{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{e \in \boldsymbol{a}(s)\}$ the number of time the element e was in the drawn arm.

Let decompose the expected number of iterations at which the permutation $\boldsymbol{a}$ is recommended:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\boldsymbol{a}(t) = \boldsymbol{a}\}\right] \leqslant & \sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\boldsymbol{a}(t) = \boldsymbol{a}, |\hat{\rho}_e(t) - \rho_e| \geqslant \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}\right\}\right] \\
& + \sum_{e\in\boldsymbol{a}^*\setminus\boldsymbol{a}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{b_e(t) \leqslant \rho_e\}\right] \\
& + \mathbb{E}\left[\sum_{t=|E|}^{T} \mathbb{1}\left\{\boldsymbol{a}(t) = \boldsymbol{a}, \forall e \in \boldsymbol{a}\setminus\boldsymbol{a}^*, |\hat{\rho}_e(t) - \rho_e| < \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}, \forall e \in \boldsymbol{a}^*\setminus\boldsymbol{a}, b_e(t) > \rho_e\right\}\right] \\
& + |E|.
\end{aligned}$$

The proof consists in upper-bounding each term on the right-hand side.

First Term Let $e \in \boldsymbol{a} \setminus \boldsymbol{a}^*$, and denote $A_e = \left\{t \in [T] : \boldsymbol{a}(t) = \boldsymbol{a}, |\hat{\rho}_e(t) - \rho_e| \geqslant \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}\right\}$.

$A_e \subseteq \bigcup_{s\in\mathbb{N}} \Lambda_k(s)$, where $\Lambda_k(s) \overset{def}{=} \{t \in A_e : T_{\boldsymbol{a}}(t) = s\}$. For any integer value s, $|\Lambda_k(s)| \leqslant 1$ as $T_{\boldsymbol{a}}(t)$ increases for each $t \in A_e$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda_k(s)$, $T_e(n) \geqslant T_{\boldsymbol{a}}(n) = s$. Then, by Lemma 4

$$\begin{aligned}\mathbb{E}\left[|A_e|\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in A_e\}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{t \in A_e, |\hat{\rho}_e(t) - \rho_e| \geqslant \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}\right\}\right] \\ &\leqslant \frac{4K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}^2}.\end{aligned}$$

Hence, $\sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\boldsymbol{a}(t) = \boldsymbol{a}, |\hat{\rho}_e(t) - \rho_e| \geqslant \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}\right\}\right] = \sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} \mathbb{E}\left[|A_e|\right] \leqslant \frac{4K_{\boldsymbol{a}}^3}{\Delta_{\boldsymbol{a}}^2}$.

Second Term Let $e \in \boldsymbol{a}^* \setminus \boldsymbol{a}$, and denote $B_e \overset{def}{=} \{t \in [T] : b_e(t) \leqslant \rho_e\}$.

By Theorem 10 of (Garivier & Cappé, 2011), $\mathbb{E}\left[|B_e|\right] = O(\log\log T)$, so $\sum_{e\in\boldsymbol{a}^*\setminus\boldsymbol{a}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{b_e(t) \leqslant \rho_e\}\right] = \mathcal{O}(\bar{K}_{\boldsymbol{a}} \log\log T)$.

Third Term Let note $C \overset{def}{=} \left\{t \in [T] \setminus [|E|] : \boldsymbol{a}(t) = \boldsymbol{a}, \forall e \in \boldsymbol{a} \setminus \boldsymbol{a}^*, |\hat{\rho}_e(t) - \rho_e| < \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}, \forall e \in \boldsymbol{a}^* \setminus \boldsymbol{a}, b_e(t) > \rho_e\right\}$.

Let $t \in C$.

At each step of the initialization phase, the algorithm removes at least one element e of the set $\tilde{E}$ of unseen elements. Therefore, the initialization lasts at most $|E|$ iterations. Hence, at iteration t, $\boldsymbol{a}(t) = \boldsymbol{a}$ is chosen as $\sum_{e\in\boldsymbol{a}} b_e(t) = \max_{\boldsymbol{a}'\in\mathcal{A}} \sum_{e\in\boldsymbol{a}'} b_e(t)$.

Then, by Pinsker's inequality and the fact that $t \leqslant T$, and $T_e(t) \geqslant T_{\boldsymbol{a}}(t)$ for any e in $\boldsymbol{a}$,

$$\begin{aligned}0 &\leqslant \sum_{e\in\boldsymbol{a}} b_e(t) - \sum_{e\in\boldsymbol{a}^*} b_e(t) \\ &= \sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} b_e(t) - \sum_{e\in\boldsymbol{a}^*\setminus\boldsymbol{a}} b_e(t) \\ &\leqslant \sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} \hat{\rho}_e(t) + \sqrt{\frac{\log(t) + 3\log(\log(t))}{2T_e(t)}} - \sum_{e\in\boldsymbol{a}^*\setminus\boldsymbol{a}} b_e(t) \\ &< \sum_{e\in\boldsymbol{a}\setminus\boldsymbol{a}^*} \rho_e + \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}} + \sqrt{\frac{\log(T) + 3\log(\log(T))}{2T_{\boldsymbol{a}}(t)}} - \sum_{e\in\boldsymbol{a}^*\setminus\boldsymbol{a}} \rho_e \\ &\leqslant \sum_{e\in\boldsymbol{a}} \rho_e - \sum_{e\in\boldsymbol{a}^*} \rho_e + K_{\boldsymbol{a}}\frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}} + K_{\boldsymbol{a}}\sqrt{\frac{\log(T) + 3\log(\log(T))}{2T_{\boldsymbol{a}}(t)}} \\ &= -\Delta_{\boldsymbol{a}} + \frac{2\Delta_{\boldsymbol{a}}}{2} + K_{\boldsymbol{a}}\sqrt{\frac{\log(T) + 3\log(\log(T))}{2T_{\boldsymbol{a}}(t)}}. \\ &= -\frac{\Delta_{\boldsymbol{a}}}{2} + K_{\boldsymbol{a}}\sqrt{\frac{\log(T) + 3\log(\log(T))}{2T_{\boldsymbol{a}}(t)}}.\end{aligned}$$

Hence, $T_{\boldsymbol{a}}(t) < K_{\boldsymbol{a}}^2 \frac{2\log(T)+6\log(\log(T))}{\Delta_{\boldsymbol{a}}^2}$. Therefore, $C \subseteq \left\{t \in [T] \setminus [|E|] : \boldsymbol{a}(t) = \boldsymbol{a}, T_{\boldsymbol{a}}(t) < K_{\boldsymbol{a}}^2 \frac{2\log(T)+6\log(\log(T))}{\Delta_{\boldsymbol{a}}^2}\right\}$, and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=|E|}^{T} \mathbb{1}\left\{\boldsymbol{a}(t)=\boldsymbol{a}, \forall e \in \boldsymbol{a} \setminus \boldsymbol{a}^*, |\hat{\rho}_e(t)-\rho_e| < \frac{\Delta_{\boldsymbol{a}}}{2K_{\boldsymbol{a}}}, \forall e \in \boldsymbol{a}^* \setminus \boldsymbol{a}, b_e(t) > \rho_e\right\}\right] \\
&= \mathbb{E}\left[|C|\right] \\
&\leqslant \mathbb{E}\left[\left|\left\{t \in [T] \setminus [|E|] : \boldsymbol{a}(t)=\boldsymbol{a}, T_{\boldsymbol{a}}(t) < K_{\boldsymbol{a}}^2 \frac{2\log(T)+6\log(\log(T))}{\Delta_{\boldsymbol{a}}^2}\right\}\right|\right] \\
&\leqslant K_{\boldsymbol{a}}^2 \frac{2\log(T)+6\log(\log(T))}{\Delta_{\boldsymbol{a}}^2}.
\end{aligned}$$

Regret upper-bound Overall,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\boldsymbol{a}(t)=\boldsymbol{a}\}\right] &\leqslant \frac{4K_{\boldsymbol{a}}^3}{\Delta_{\boldsymbol{a}}^2} + \mathcal{O}(\bar{K}_{\boldsymbol{a}} \log\log T) + K_{\boldsymbol{a}}^2 \frac{2\log(T)+6\log(\log(T))}{\Delta_{\boldsymbol{a}}^2} + |E| \\
&= \frac{2K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}^2}\log(T) + \mathcal{O}\left(\left(\bar{K}_{\boldsymbol{a}} + \frac{K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}^2}\right)\log\log T\right)
\end{aligned}$$

and

$$\begin{aligned}
R(T) &= \sum_{\boldsymbol{a}\in\mathcal{A}:\mu_{\boldsymbol{a}}\neq\mu^*} \Delta_{\boldsymbol{a}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\boldsymbol{a}(t)=\boldsymbol{a}\}\right] \\
&\leqslant \sum_{\boldsymbol{a}\in\mathcal{A}:\mu_{\boldsymbol{a}}\neq\mu^*} \frac{2K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}}\log(T) + \mathcal{O}\left(\left(\bar{K}_{\boldsymbol{a}}\Delta_{\boldsymbol{a}} + \frac{K_{\boldsymbol{a}}^2}{\Delta_{\boldsymbol{a}}}\right)\log\log T\right) \\
&= \mathcal{O}\left(\frac{|\mathcal{A}|K_{max}^2}{\Delta_{\min}}\log T\right),
\end{aligned}$$

which concludes the proof.

□

E. Proof of Lemma 2 (Upper-bound on the Number of Iterations of GRAB for which $\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}} \neq \boldsymbol{a}^*$)

Proof of Lemma 2. Let $\tilde{\boldsymbol{a}} \in \mathcal{P}_K^L \setminus \{\boldsymbol{a}^*\}$ and prove that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}\}\right] = \mathcal{O}\left(\log\log T\right)$.

The proof requires notations related to the neighborhood of $\tilde{\boldsymbol{a}}$. Let $\mathcal{N} \overset{def}{=} \bigcup_{\boldsymbol{\pi}\in\mathcal{P}_K^K} \mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})$ be the set of all the potential neighbors of $\tilde{\boldsymbol{a}}$. By definition of the neighborhoods,

$$\mathcal{N} = \left\{\tilde{\boldsymbol{a}} \circ (k,k') : k,k' \in [K]^2, k > k'\right\} \cup \left\{\tilde{\boldsymbol{a}}[k := i] : k \in [K], i \in [L] \setminus \tilde{\boldsymbol{a}}([K])\right\},$$

and its size is $N = K(2L-K-1)/2$. As $\tilde{\boldsymbol{a}}$ is sub-optimal, and due to Assumption 1, for any appropriate ranking of positions $\boldsymbol{\pi} \in \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})$, there exists a recommendation $\boldsymbol{a}^+$ with a strictly better expected reward than $\tilde{\boldsymbol{a}}$ in the neighborhood $\mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})$. We denote

$$\mathcal{N}^+ \overset{def}{=} \bigcup_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})} \left\{\boldsymbol{a}^+ \in \mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}}) : \mu_{\boldsymbol{a}^+} = \max_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})} \mu_{\boldsymbol{a}}\right\}$$

the set of such recommendations. We also chose $\epsilon < \min\{1/(2N), 1/L\}$ and note

$$\delta \overset{def}{=} \min_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})} \min_{\boldsymbol{a}\in\mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})\cup\{\tilde{\boldsymbol{a}}\}\setminus\mathcal{N}^+} \left(\max_{\boldsymbol{a}'\in\mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})} \mu_{\boldsymbol{a}'} - \mu_{\boldsymbol{a}}\right).$$

To bound $\mathbb{E}\left[\mathbb{1}\{\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}\}\right]$, we use the decomposition $\{t\in[T]:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}\}\subseteq\bigcup_{\boldsymbol{a}^+\in\mathcal{N}^+}A_{\boldsymbol{a}^+}\cup B$ where for any permutation $\boldsymbol{a}^+\in\mathcal{N}^+$,

$$A_{\boldsymbol{a}^+}=\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},T_{\boldsymbol{a}^+}(t)\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)\}$$

and

$$B=\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\forall\boldsymbol{a}^+\in\mathcal{A}+,T_{\boldsymbol{a}^+}(t)<\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)\}.$$

Hence,

$$\mathbb{E}\left[\mathbb{1}\{\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}\}\right]\leqslant\sum_{\boldsymbol{a}^+\in\mathcal{A}+}\mathbb{E}\left[|A_{\boldsymbol{a}^+}|\right]+\mathbb{E}\left[|B|\right].$$

Bound on $\mathbb{E}\left[|A_{\boldsymbol{a}^+}|\right]$ Let $\boldsymbol{a}^+$ be a permutation in $\mathcal{N}^+$ and denote $\mathcal{K}^+$ the set of positions for which $\boldsymbol{a}^+$ and $\tilde{\boldsymbol{a}}$ disagree: $\mathcal{K}^+=\left\{k\in[K]:a_k^+\neq\tilde{a}_k\right\}$. The permutation $\boldsymbol{a}^+$ is in the neighborhood of $\tilde{\boldsymbol{a}}$, so either $\boldsymbol{a}^+=\tilde{\boldsymbol{a}}\circ(k,k')$ or $\boldsymbol{a}^+=\boldsymbol{a}[k:=i]$, with k and k' in $[K]$, and i in $[L]$. Overall, $|\mathcal{K}^+|\leqslant 2$.

By the design of the algorithm and by definition of ϵ, we have that $\forall t\in A_{\boldsymbol{a}^+}$, $T_{\tilde{\boldsymbol{a}}}(t)\geqslant\tilde{T}_{\tilde{\boldsymbol{a}}}(t)/L>\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)$. Moreover, at the considered iterations $\tilde{\boldsymbol{a}}$ is the leader, so

$$\begin{aligned}
A_{\boldsymbol{a}^+}&\subseteq\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\tilde{T}_{\tilde{\boldsymbol{a}}}(t)<\frac{1}{\epsilon}\right\}\cup\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\min\{T_{\tilde{\boldsymbol{a}}}(t),T_{\boldsymbol{a}^+}(t)\}\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)\geqslant 1,\sum_{\ell}\hat{\rho}_{\tilde{a}_\ell,\ell}(t)\geqslant\sum_{\ell}\hat{\rho}_{a^+_\ell,\ell}(t)\right\}\\
&\subseteq\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\tilde{T}_{\tilde{\boldsymbol{a}}}(t)<\frac{1}{\epsilon}\right\}\cup\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\min\{T_{\tilde{\boldsymbol{a}}}(t),T_{\boldsymbol{a}^+}(t)\}\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t),\sum_{k\in\mathcal{K}^+}\hat{\rho}_{\tilde{a}_k,k}(t)\geqslant\sum_{k\in\mathcal{K}^+}\hat{\rho}_{a^+_k,k}(t)\right\}\\
&\subseteq\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\tilde{T}_{\tilde{\boldsymbol{a}}}(t)<\frac{1}{\epsilon}\right\}\\
&\quad\cup\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\min\{T_{\tilde{\boldsymbol{a}}}(t),T_{\boldsymbol{a}^+}(t)\}\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t),\exists k\in\mathcal{K}^+,|\hat{\rho}_{\tilde{a}_k,k}(t)-\rho_{\tilde{a}_k,k}|\geqslant\frac{\delta}{2|\mathcal{K}^+|}\text{ or }|\hat{\rho}_{a^+_k,k}(t)-\rho_{a^+_k,k}|\geqslant\frac{\delta}{2|\mathcal{K}^+|}\right\}\\
&\subseteq\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\tilde{T}_{\tilde{\boldsymbol{a}}}(t)<\frac{1}{\epsilon}\right\}\cup\bigcup_{k\in\mathcal{K}^+}\bigcup_{i\in\left\{\tilde{a}_k,a^+_k\right\}}\Lambda_{i,k},
\end{aligned}$$

with $\Lambda_{i,k}\overset{def}{=}\left\{t:\tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}},\min\{T_{\tilde{\boldsymbol{a}}}(t),T_{\boldsymbol{a}^+}(t)\}\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t),|\hat{\rho}_{i,k}(t)-\rho_{i,k}|\geqslant\frac{\delta}{2|\mathcal{K}^+|}\right\}$.

Fix k in $\mathcal{K}^+$ and i in $\left\{\tilde{a}_k,a^+_k\right\}$. $\Lambda_{i,k}\subseteq\bigcup_{s\in\mathbb{N}}\Lambda_{i,k}(s)$, with $\Lambda_{i,k}(s)\overset{def}{=}\{t\in\Lambda_{i,k}:\tilde{T}_{\tilde{\boldsymbol{a}}}(t)=s\}$. $|\Lambda_{i,k}(s)|\leqslant 1$ as $\tilde{T}_{\tilde{\boldsymbol{a}}}(t)$ increases for each $t\in\Lambda_{i,k}$. Note that for each $s\in\mathbb{N}$ and $n\in\Lambda_{i,k}(s)$, $T_{i,k}(n)\geqslant\min\{T_{\boldsymbol{a}}(n),T_{\boldsymbol{a}^+}(n)\}\geqslant\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(n)=\epsilon s$. Then, by Lemma 4

$$\begin{aligned}
\mathbb{E}\left[|\Lambda_{i,k}|\right]&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{t\in\Lambda_{i,k}\}\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{t\in\Lambda_{i,k},|\hat{\rho}_{i,k}(t)-\rho_{i,k}|>\frac{\delta}{2|\mathcal{K}^+|}\right\}\right]\\
&\leqslant\frac{4|\mathcal{K}^+|^2}{\epsilon\delta^2}
\end{aligned}$$

Hence, $\mathbb{E}\left[|A_{\boldsymbol{a}^+}|\right]\leqslant\frac{1}{\epsilon}+\sum_{k\in\mathcal{K}^+}\sum_{i\in\left\{\tilde{a}_k,a^+_k\right\}}\mathbb{E}\left[|\Lambda_{i,k}|\right]\leqslant\frac{1}{\epsilon}+\frac{8|\mathcal{K}^+|^3}{\epsilon\delta^2}$.

Bound on $\mathbb{E}\left[|B|\right]$ We first split B in two parts: $B=B^{t_0}\cup B^T_{t_0}$, where $B^{t_0}\overset{def}{=}\{t\in B:\tilde{T}_{\tilde{\boldsymbol{a}}}(t)\leqslant t_0\}$, $B^T_{t_0}\overset{def}{=}\{t\in B:\tilde{T}_{\tilde{\boldsymbol{a}}}(t)>t_0\}$, and t_0 is chosen as small as possible to satisfy three constraints required in the rest of the proof.

Namely, $t_0=\max\left\{\frac{1}{\epsilon},(1+N)(1-\frac{1}{L}-\epsilon N)^{-1},\inf\left\{t:2\sqrt{\frac{\log(t+1)+3\log(\log(t+1))}{2\epsilon t}}<\frac{\delta}{8}\right\}\right\}$. Note that t_0 only depends on K, L and δ, and that $(1-\frac{1}{L}-\epsilon N)>0$ (assuming $L\geqslant 2$) as $\epsilon<1/(2N)$.

We also define

- $D \overset{def}{=} \bigcup_{(\boldsymbol{a},k)\in(\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}\setminus\mathcal{N}^+)\times[K]} D_{\boldsymbol{a},k}$, where $D_{\boldsymbol{a},k} \overset{def}{=} \left\{t\in[T] : \tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}, \boldsymbol{a}(t)=\boldsymbol{a}, |\hat{\rho}_{a_k,k}(t)-\rho_{a_k,k}| \geqslant \frac{\delta}{8}\right\}$,
- $E \overset{def}{=} \bigcup_{(\boldsymbol{a}^+,k)\in\mathcal{N}^+\times[K]} E_{\boldsymbol{a}^+,k}$, where $E_{\boldsymbol{a}^+,k} \overset{def}{=} \{t\in[T] : \tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}, b_{a_k^+,k}(t) \leqslant \rho_{a_k^+,k}\}$,
- and $F \overset{def}{=} \{t\in[T] : \tilde{\boldsymbol{a}}(t)=\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})\}$.

Let $t\in B_{t_0}^T$. By construction, GRAB forces itself to select $\left\lceil \frac{\tilde{T}_{\tilde{\boldsymbol{a}}}(t)}{L}\right\rceil$ times the leader $\tilde{\boldsymbol{a}}$ between iterations 1 and $t-1$. So,

$$\tilde{T}_{\tilde{\boldsymbol{a}}}(t) = \left\lceil \frac{\tilde{T}_{\tilde{\boldsymbol{a}}}(t)}{L}\right\rceil + \sum_{\boldsymbol{a}\in\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}} T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t)$$

where $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t) = \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{a}}(s)=\tilde{\boldsymbol{a}}, \boldsymbol{a}(s)=\boldsymbol{a}, \tilde{T}_{\tilde{\boldsymbol{a}}}(s)/L \notin \mathbb{N}\right\}$ is the number of times arm $\boldsymbol{a}\in\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}$ has been played **normally** (i.e not forced) while $\tilde{\boldsymbol{a}}$ was leader, up to time $t-1$. Let prove by contradiction that there is at least one recommendation $\boldsymbol{a}$ that has been selected **normally** more than $\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1$ times, namely $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1$.

Assume that for each recommendation $\boldsymbol{a}$ in $\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}$, $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t) < \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1$. Then

$$\begin{aligned}\tilde{T}_{\tilde{\boldsymbol{a}}}(t) &= \left\lceil \frac{\tilde{T}_{\tilde{\boldsymbol{a}}}(t)}{L}\right\rceil + \sum_{\boldsymbol{a}\in\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}} T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t)\\ &< 1 + \frac{\tilde{T}_{\tilde{\boldsymbol{a}}}(t)}{L} + N(\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1).\end{aligned}$$

Therefore $\tilde{T}_{\tilde{\boldsymbol{a}}}(t)(1-\frac{1}{L}-N\epsilon) < 1+N$, which contradicts $t\in B_{t_0}^T$.

So, there exists a recommendation $\boldsymbol{a}$ such that $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(t) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1$. Let denote s' the first iteration such that $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(s') \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1$. At this iteration, $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(s') = T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(s'-1)+1$, meaning that $\tilde{\boldsymbol{a}}(s'-1)=\tilde{\boldsymbol{a}}$, $\boldsymbol{a}(s'-1)=\boldsymbol{a}$, $\tilde{T}_{\tilde{\boldsymbol{a}}}(s'-1)/L \notin \mathbb{N}$, and $T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(s'-1) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)$. Therefore, the set $\{s\in[t] : \tilde{\boldsymbol{a}}(s)=\tilde{\boldsymbol{a}}, T_{\boldsymbol{a}(s)}^{\tilde{\boldsymbol{a}}}(s) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t), \tilde{T}_{\tilde{\boldsymbol{a}}}(s)/L \notin \mathbb{N}\}$ is non-empty. We define $\psi(t)$ as the minimum on this set

$$\psi(t) \overset{def}{=} \min\left\{s\in[t] : \tilde{\boldsymbol{a}}(s)=\tilde{\boldsymbol{a}}, T_{\boldsymbol{a}(s)}^{\tilde{\boldsymbol{a}}}(s) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t), \tilde{T}_{\tilde{\boldsymbol{a}}}(s)/L \notin \mathbb{N}\right\}.$$

We note $\boldsymbol{a}$ the recommendation $\boldsymbol{a}(\psi(t))$ at iteration $\psi(t)$. We have $\boldsymbol{a}\notin\mathcal{N}^+$ since for any recommendation $\boldsymbol{a}^+\in\mathcal{N}^+$, $T_{\boldsymbol{a}^+}^{\tilde{\boldsymbol{a}}}(\psi(t)) \leqslant T_{\boldsymbol{a}^+}^{\tilde{\boldsymbol{a}}}(t) \leqslant T_{\boldsymbol{a}^+}(t) < \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)$. Let $\boldsymbol{a}^+$ be one of the best recommendations in $\mathcal{N}_{\tilde{\boldsymbol{\pi}}(\psi(t))}(\tilde{\boldsymbol{a}})\cup\{\tilde{\boldsymbol{a}}\}$, meaning $\mu_{\boldsymbol{a}^+} = \max_{\boldsymbol{a}'\in\mathcal{N}_{\tilde{\boldsymbol{\pi}}(\psi(t))}(\tilde{\boldsymbol{a}})\cup\{\tilde{\boldsymbol{a}}\}} \mu_{\boldsymbol{a}'}$, and let $\mathcal{K}$ denote the set of positions for which $\boldsymbol{a}$ and $\boldsymbol{a}^+$ disagree. As both recommendations are in $\mathcal{N}_{\tilde{\boldsymbol{\pi}}(\psi(t))}(\tilde{\boldsymbol{a}})\cup\{\tilde{\boldsymbol{a}}\}$, $|\mathcal{K}| \leqslant 4$.

Let prove by contradiction that $\psi(t)\in D\cup E\cup F$. Assume that $\psi(t)\notin D\cup E\cup F$.

Since $\psi(t)\notin F$, $\tilde{\boldsymbol{\pi}}(\psi(t))$ belongs to $\Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})$ and hence $\boldsymbol{a}^+$ is in $\mathcal{N}^+$ and $\sum_k \rho_{a_k^+,k} - \sum_k \rho_{a_k,k} = \mu_{\boldsymbol{a}^+} - \mu_{\boldsymbol{a}} \geqslant \delta$.

Moreover, since $\psi(t)\notin D\cup E$, for each position $k\in[K]$, $|\hat{\rho}_{a_k,k}(\psi(t)) - \rho_{a_k,k}| < \frac{\delta}{8}$, and $b_{a_k^+,k}(\psi(t)) > \rho_{a_k^+,k}$.

Finally, $T_{\boldsymbol{a}}(\psi(t)) \geqslant T_{\boldsymbol{a}}^{\tilde{\boldsymbol{a}}}(\psi(t)) \geqslant \epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t) \geqslant 1$, and therefore $b_{a_k,k}(\psi(t))$ and $\hat{\rho}_{a_k,k}(\psi(t))$ are properly defined for any position $k\in[K]$.

Then, by Pinsker's inequality and the fact that $\psi(t) \leqslant t$, $\tilde{T}_{\tilde{\boldsymbol{a}}}(s)$ is non-decreasing in s, and $T_{\boldsymbol{a}}(\psi(t)) \geqslant \epsilon \tilde{T}_{\tilde{\boldsymbol{a}}}(t)$,

$$\begin{aligned}
\sum_k b_{a_k,k}(\psi(t)) - \sum_k b_{a_k^+,k}(\psi(t)) &= \sum_{k\in\mathcal{K}} b_{a_k,k}(\psi(t)) - b_{a_k^+,k}(\psi(t)) \\
&\leqslant \sum_{k\in\mathcal{K}} \hat{\rho}_{a_k,k}(\psi(t)) + \sqrt{\frac{\log(\tilde{T}_{\tilde{\boldsymbol{a}}}(\psi(t))+1) + 3\log(\log(\tilde{T}_{\tilde{\boldsymbol{a}}}(\psi(t))+1))}{2T_{\boldsymbol{a}}(\psi(t))}} - b_{a_k^+,k}(\psi(t)) \\
&< \sum_{k\in\mathcal{K}} \rho_{a_k,k} + \frac{\delta}{8} + \sqrt{\frac{\log(\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1) + 3\log(\log(\tilde{T}_{\tilde{\boldsymbol{a}}}(t)+1))}{2\epsilon\tilde{T}_{\tilde{\boldsymbol{a}}}(t)}} - \rho_{a_k^+,k} \\
&\leqslant \sum_{k\in\mathcal{K}} \rho_{a_k,k} + \frac{\delta}{8} + \frac{\delta}{8} - \rho_{a_k^+,k} \\
&\leqslant \sum_k \rho_{a_k,k} - \sum_k \rho_{a_k^+,k} + |\mathcal{K}| \cdot 2\frac{\delta}{8} \\
&\leqslant -\delta + 8\frac{\delta}{8} \\
&= 0,
\end{aligned}$$

which contradicts the fact that $\boldsymbol{a}$ is played at iteration $\psi(t)$. So $\psi(t) \in D \cup E \cup F$.

Overall, for any $t \in B_{t_0}^T$, $\psi(t) \in D \cup E \cup F$. So, $B_{t_0}^T \subseteq \bigcup_{n\in D\cup E\cup F} B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}$. Let n be in $D \cup E \cup F$. For any t in $B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}$, $T_{\boldsymbol{a}(n)}^{\tilde{\boldsymbol{a}}}(n) = \lceil \epsilon \tilde{T}_{\tilde{\boldsymbol{a}}}(t) \rceil$ and $\tilde{T}_{\tilde{\boldsymbol{a}}}(t+1) = \tilde{T}_{\tilde{\boldsymbol{a}}}(t) + 1$. So $|B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}| < 1/\epsilon + 1$. Overall,

$$\mathbb{E}\left[|B|\right] \leqslant t_0 + \mathbb{E}\left[|B_{t_0}^T|\right] \leqslant t_0 + (1/\epsilon + 1)\left(\mathbb{E}\left[|D|\right] + \mathbb{E}\left[|E|\right] + \mathbb{E}\left[|F|\right]\right).$$

It remains to upper-bound $\mathbb{E}\left[|D|\right]$, $\mathbb{E}\left[|E|\right]$, and $\mathbb{E}\left[|F|\right]$ to conclude the proof.

Bound on $\mathbb{E}\left[|D|\right]$ The upper-bound on $\mathbb{E}\left[|D|\right]$ is obtained with the same strategy as the last step in the proof of the upper-bound on $\mathbb{E}\left[|A_{\boldsymbol{a}^+}|\right]$. Let $\boldsymbol{a}$ be a recommendation in $\mathcal{N} \cup \{\tilde{\boldsymbol{a}}\} \setminus \mathcal{N}^+$, and $k \in [K]$ be a position. $D_{\boldsymbol{a},k} \subseteq \bigcup_{s\in\mathbb{N}} \Lambda_{\boldsymbol{a},k}(s)$, where $\Lambda_{\boldsymbol{a},k}(s) \overset{def}{=} \{t \in D_{\boldsymbol{a},k} : T_{\boldsymbol{a}}(t) = s\}$. $|\Lambda_{\boldsymbol{a},k}(s)| \leqslant 1$ as $T_{\boldsymbol{a}}(t)$ increases for each $t \in D_{\boldsymbol{a},k}$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda_{\boldsymbol{a},k}(s)$, $T_{a_k,k}(n) \geqslant T_{\boldsymbol{a}}(n) = s$. Then, by Lemma 4

$$\begin{aligned}
\mathbb{E}\left[|D_{\boldsymbol{a},k}|\right] &\leq \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{t \in D_{\boldsymbol{a},k}\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left\{t \in D_{\boldsymbol{a},k}, |\hat{\rho}_{a_k,k}(t) - \rho_{a_k,k}| \geqslant \frac{\delta}{8}\right\}\right] \\
&\leqslant \frac{64}{\delta^2}
\end{aligned}$$

Hence, $\mathbb{E}\left[|D|\right] \leq \sum_{(\boldsymbol{a},k)\in(\mathcal{N}\cup\{\tilde{\boldsymbol{a}}\}\setminus\mathcal{N}^+)\times[K]} \mathbb{E}\left[|D_{\boldsymbol{a},k}|\right] \leqslant \frac{64(N+1)K}{\delta^2}$.

Bound on $\mathbb{E}\left[|E|\right]$ By Theorem 10 of (Garivier & Cappé, 2011), $\mathbb{E}\left[|E_{\boldsymbol{a}^+,k}|\right] = O(\log(\log(T)))$, so $\mathbb{E}\left[|E|\right] \leqslant \sum_{(\boldsymbol{a}^+,k)\in\mathcal{N}^+\times[K]} \mathbb{E}\left[|E_{\boldsymbol{a}^+,k}|\right] = O(|\mathcal{N}^+|K\log(\log(T)))$.

Bound on $\mathbb{E}\left[|F|\right]$ By Lemma 3, $\mathbb{E}\left[|F|\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)\right\}\right] = \mathcal{O}\left(1\right)$.

Overall $\mathbb{E}\left[\mathbb{1}\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}\}\right] \leqslant \frac{|\mathcal{K}^+|}{\epsilon} + \frac{8|\mathcal{K}^+|^3|\mathcal{N}^+|}{\epsilon\delta^2} + t_0 + \left(\frac{1}{\epsilon}+1\right)\frac{64(N+1)K}{\delta^2} + \mathcal{O}\left(\frac{|\mathcal{N}^+|K}{\epsilon}\log\log T\right) + \mathcal{O}(1) = \mathcal{O}\left(\frac{|\mathcal{N}^+|K}{\epsilon}\log\log T\right)$, which concludes the proof. □

F. Proof of Lemma 3 (Upper-bound on the Number of Iterations of GRAB for which $\tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})$)

Proof of Theorem 3. Let $\tilde{\boldsymbol{a}}$ be a K-permutation of L items. If $\Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})$ contains all the permutations of K elements, the set $\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})\}$ is empty.

Otherwise, let denote δ the smallest non-zero gap between the probability of click at position k and the probability of click at position $k' \neq k$: $\delta \overset{def}{=} \min\left\{\rho_{\tilde{a}_k,k} - \rho_{\tilde{a}_{k'},k'} : (k,k') \in [K]^2, \rho_{\tilde{a}_k,k} - \rho_{\tilde{a}_{k'},k'} > 0\right\}$. The gap δ is the minimum on a finite set, so $\delta > 0$.

By definition of $\tilde{\pi}(t)$, $\hat{\rho}_{\tilde{a}_{\tilde{\pi}_1(t)}(t),\tilde{\pi}_1(t)}(t) \geqslant \hat{\rho}_{\tilde{a}_{\tilde{\pi}_2(t)}(t),\tilde{\pi}_2(t)}(t) \geqslant \cdots \geqslant \hat{\rho}_{\tilde{a}_{\tilde{\pi}_K(t)}(t),\tilde{\pi}_K(t)}(t)$, so,

$$\begin{aligned}
\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})\} &= \bigcup_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^K} \bigcup_{k \in [K-1]} \left\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) = \tilde{\boldsymbol{\pi}}, \rho_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k} < \rho_{\tilde{a}_{\tilde{\pi}_{k+1}},\tilde{\pi}_{k+1}}\right\} \\
&\subseteq \bigcup_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^K} \bigcup_{k \in [K-1]} \left\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) = \tilde{\boldsymbol{\pi}}, \substack{|\hat{\rho}_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}(t) - \rho_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}| > \frac{\delta}{2} \\ \text{or } |\hat{\rho}_{\tilde{a}_{\tilde{\pi}_{k+1}},\tilde{\pi}_{k+1}}(t) - \rho_{\tilde{a}_{\tilde{\pi}_{k+1}},\tilde{\pi}_{k+1}}| > \frac{\delta}{2}}\right\} \\
&= \bigcup_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^K} \bigcup_{k \in [K]} \Lambda_{\tilde{\boldsymbol{\pi}},k},
\end{aligned}$$

with $\Lambda_{\tilde{\boldsymbol{\pi}},k} \overset{def}{=} \left\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) = \tilde{\boldsymbol{\pi}}, |\hat{\rho}_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}(t) - \rho_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}| > \frac{\delta}{2}\right\}$, for any ranking of positions $\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^L$ and any rank $k \in [K]$.

Let $\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^L$ be a ranking of positions, and $k \in [K]$ be a rank. $\Lambda_{\tilde{\boldsymbol{\pi}},k} \subseteq \bigcup_{s \in \mathbb{N}} \Lambda_{\tilde{\boldsymbol{\pi}},k}(s)$, with $\Lambda_{\tilde{\boldsymbol{\pi}},k}(s) \overset{def}{=} \{t \in \Lambda_{\tilde{\boldsymbol{\pi}},k} : \tilde{T}_{\tilde{\boldsymbol{a}}}(t) = s\}$. $|\Lambda_{\tilde{\boldsymbol{\pi}},k}(s)| \leqslant 1$ as $\tilde{T}_{\tilde{\boldsymbol{a}}}(t)$ increases for each $t \in \Lambda_{\tilde{\boldsymbol{\pi}},k}$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda_{\tilde{\boldsymbol{\pi}},k}(s)$, $T_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}(n) \geqslant T_{\tilde{\boldsymbol{a}}}(n) \geqslant \tilde{T}_{\tilde{\boldsymbol{a}}}(n)/L = s/L$. Then, by Lemma 4

$$\begin{aligned}
\mathbb{E}\left[|\Lambda_{\tilde{\boldsymbol{\pi}},k}|\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in \Lambda_{\tilde{\boldsymbol{\pi}},k}\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{t \in \Lambda_{\tilde{\boldsymbol{\pi}},k}, |\hat{\rho}_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}(t) - \rho_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}| > \frac{\delta}{2}\right\}\right] \\
&\leqslant \frac{4L}{\delta^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}(\tilde{\boldsymbol{a}})\}\right] &\leqslant \sum_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}_K^K} \sum_{k \in [K]} \mathbb{E}\left[\Lambda_{\tilde{\boldsymbol{\pi}},k}\right] \\
&\leqslant \frac{4LKK!}{\delta^2} \\
&= \mathcal{O}\left(LKK!\right),
\end{aligned}$$

which concludes the proof.

□

G. KL-CombUCB and its Application to PBM Setting

In this section we first define the generic combinatorial semi-bandit algorithm KL-CombUCB and we compare two upper-bounds on its regret. Then, we present the application of KL-CombUCB to PBM setting and discuss its relation to GRAB.

Algorithm 2 KL-ComUCB1 (generic version)

Input: set of elements E, set of arms $\mathcal{A}$
  $t \leftarrow 1$
  while $\{e \in E : T_e(t) = 0\} \neq \varnothing$ **do**
    $\tilde{E} \leftarrow \{e \in E : T_e(t) = 0\}$
    $\tilde{\mathcal{A}} \leftarrow \{\boldsymbol{a} \in \mathcal{A} : \boldsymbol{a} \cap \tilde{E} \neq \varnothing\}$
    recommend $\boldsymbol{a}(t) = \operatorname{argmax}_{\boldsymbol{a} \in \tilde{\mathcal{A}}} \sum_{e \in \boldsymbol{a}} b_e(t)$
    observe the weights $[w_e(t) : e \in \boldsymbol{a}]$
    $t \leftarrow t + 1$
  end while
  $t_0 \leftarrow t$
  for $t = t_0, t_0 + 1, \ldots$ **do**
    recommend $\boldsymbol{a}(t) = \operatorname{argmax}_{\boldsymbol{a} \in \mathcal{A}} \sum_{e \in \boldsymbol{a}} b_e(t)$
    observe the weights $[w_e(t) : e \in \boldsymbol{a}]$
  end for

G.1. KL-CombUCB for Generic Setting

CombUCB1 (Kveton et al., 2015b) is a bandit algorithm handling the following combinatorial setting. Let E be a set of elements and $\mathcal{A} \subseteq \{0,1\}^E$ be a set of arms, where each arm $\boldsymbol{a}$ is a subset of E. Following the terminology used in (Kveton et al., 2015b), E is the *ground set* and $\mathcal{A}$ the *feasible set*. At each iteration, the bandit algorithm chooses a subset of elements $\boldsymbol{a} \in \mathcal{A}$ and receives the reward $\sum_{e \in \boldsymbol{a}} w_e$, where $\boldsymbol{w}$ is an independent draw of a distribution ν on $[0,1]^E$. Given these assumptions, CombUCB1 chooses an arm $\boldsymbol{a}(t)$ at each iteration, aiming at minimizing the total regret defined as usual.

We denote $\rho_e \overset{def}{=} \mathbb{E}_{\boldsymbol{w} \sim \nu}[w_e]$ the expected reward associated to element e, $\mu_{\boldsymbol{a}} \overset{def}{=} \mathbb{E}_{\boldsymbol{w} \sim \nu}\left[\sum_{e \in \boldsymbol{a}} w_e\right] = \sum_{e \in \boldsymbol{a}} \rho_e$ the expected reward when choosing the arm $\boldsymbol{a} \in \mathcal{A}$, and $\mu^* \overset{def}{=} \max_{\boldsymbol{a} \in \mathcal{A}} \mu_{\boldsymbol{a}}$ the best expected reward. We also denote $\Delta_{\boldsymbol{a}} \overset{def}{=} \mu^* - \mu_{\boldsymbol{a}}$ the gap between the best expected reward and the reward of an arm $\boldsymbol{a}$, and $\Delta_{min} \overset{def}{=} \min_{\boldsymbol{a} \in \mathcal{A} : \Delta_{\boldsymbol{a}} > 0} \Delta_{\boldsymbol{a}}$ the smallest gap of a suboptimal arm. Finally, $K \overset{def}{=} \max_{\boldsymbol{a} \in \mathcal{A}} |\boldsymbol{a}|$ denotes the maximum size of an arm (meaning the maximum number of chosen elements), $K_{\boldsymbol{a}} \overset{def}{=} \min_{\boldsymbol{a}^* \in \mathcal{A} : \mu_{\boldsymbol{a}^*} = \mu^*} |\boldsymbol{a} \setminus \boldsymbol{a}^*|$ is the smallest number of elements to remove from $\boldsymbol{a}$ to get an optimal arm, and $K_{max} \overset{def}{=} \max_{\boldsymbol{a} \in \mathcal{A} : \mu_{\boldsymbol{a}} \neq \mu^*} K_{\boldsymbol{a}}$ is its lager value.

In our paper, we use the Kullback-Leibler variation of CombUCB1 which chooses the arm based on the index $b_e(t)$ (defined hereafter) instead of the usual confidence upper-bound derived from the Hoeffding's inequality. The corresponding algorithm (KL-CombUCB) also assumes that the weight-vector $\boldsymbol{w}(t)$ is in $\{0,1\}^E$. KL-CombUCB is depicted by Algorithm 2 which uses the following notations. At each iteration t, we denote

$$\hat{\rho}_e(t) \overset{def}{=} \frac{1}{T_e(t)} \sum_{s=1}^{t-1} \mathbb{1}\{e \in \boldsymbol{a}(s)\} w_e(s)$$

the average number of clicks obtained by the element e, where

$$T_e(t) \overset{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{e \in \boldsymbol{a}(s)\}$$

is the number of times element e has been selected; $\hat{\rho}_e(t) \overset{def}{=} 0$ when $T_e(t) = 0$. The statistics $\hat{\rho}_e(t)$ are paired with their respective *indices*

$$b_e(t) \overset{def}{=} f\left(\hat{\rho}_e(t), T_e(t), t\right),$$

where $f(\hat{\rho}, s, t)$ stands for

$$\sup\{p \in [\hat{\rho}, 1] : s \times \mathrm{kl}(\hat{\rho}, p) \leq \log(t) + 3\log(\log(t))\},$$

Algorithm 3 KL-ComUCB1 (applied to PBM)

Input: number of items L, number of positions K
  for $t = 1, 2, \dots, L$ **do**
    recommend $\boldsymbol{a}(t) = (((t-1)\%L)+1, (t\%L)+1, \dots, ((t+K-2)\%L)+1)$
    observe the clicks-vector $\boldsymbol{c}(t)$
  end for
  for $t = L+1, L+2, \dots$ **do**
    recommend $\boldsymbol{a}(t) = \underset{\boldsymbol{a} \in \mathcal{P}_K^L}{\operatorname{argmax}} \sum_{k=1}^{K} b_{a_k,k}(t)$
    observe the clicks-vector $\boldsymbol{c}(t)$
  end for

with

$$\mathrm{kl}(p,q) \overset{def}{=} p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

the *Kullback-Leibler divergence* from a Bernoulli distribution of mean p to a Bernoulli distribution of mean q; $f(\hat{\rho}, s, t) \overset{def}{=} 1$ when $\hat{\rho} = 1$, $s = 0$, or $t = 0$.

Kveton et al. prove that the regret of CombUCB1 is upper-bounded by $\mathcal{O}\left(|E|K/\Delta_{min} \log T\right)$, and a similar proof would lead to the same upper-bound for KL-CombUCB. In our paper we prove in Theorem 2 a completely different regret upper-bound for KL-CombUCB: $\mathcal{O}\left(|\mathcal{A}|K_{max}^2/\Delta_{min} \log T\right)$. For most combinatorial bandit settings, this new bound is useless since $|\mathcal{A}| \gg |E|$, and $K_{max} \approx K$. However, the analysis of GRAB involves an application of KL-CombUCB to a setting where the new bound is smaller than the standard one as $|\mathcal{A}| = |E| - 1$ and $K_{max} = 2$.

G.2. KL-CombUCB Applied to PBM Setting

In the experiments (Section 6), we apply KL-CombUCB to PBM bandit setting by choosing the *ground set* $E = [L] \times [K]$, the *feasible set* $\Theta = \{\{(a_k, k) : k \in [K]\} : \boldsymbol{a} \in \mathcal{P}_K^L\}$, and the *expected weights* $\rho_{(i,k)} = \theta_i \kappa_k$ for any "element" $(i,k) \in E$. Note that the observed weights of the generic setting correspond to the clicks-vector in the PBM setting.

The corresponding algorithm, depicted by Algorithm 3, recommends at each iteration t the best permutation given the indices $b_{i,k}(t)$ defined for GRAB. This optimization problem is a *linear sum assignment problem* which is solvable in $\mathcal{O}\left(K^2(L + \log K)\right)$ time (Ramshaw & Tarjan, 2012). Note the close relationship with GRAB:

- both algorithms solve a linear sum assignment problem, they only differ from the metric to optimize: $\sum_{k=1}^{K} \hat{\rho}_{a_k,k}(t)$ for GRAB vs. $\sum_{k=1}^{K} b_{a_k,k}(t)$ for KL-CombUCB;
- both algorithms recommend the best permutation $\boldsymbol{a}$ regarding $\sum_{k=1}^{K} b_{a_k,k}(t)$, they only differ from the considered set of permutations: $\{\tilde{\boldsymbol{a}}(t)\} \cup \mathcal{N}_{\tilde{\boldsymbol{\pi}}(t)}(\tilde{\boldsymbol{a}}(t))$ for GRAB vs. $\mathcal{P}_K^L$ for KL-CombUCB.

By considering a larger set of permutations, KL-ComUCB1 suffers a $\mathcal{O}(LK^2/\Delta_{min} \log T)$ regret (by applying (Kveton et al., 2015b) bound), which is higher than the upper-bound on the regret of GRAB by a factor K^2.

H. S-GRAB: OSUB on a Static Graph

The algorithm S-GRAB, depicted in Algorithm 4, is similar to GRAB except that it explores a static graph $G = (E, V)$ defined by

$$V \overset{def}{=} \mathcal{P}_K^L,$$
$$E \overset{def}{=} \left\{(\boldsymbol{a}, \boldsymbol{a} \circ (k,k')) : k, k' \in [K]^2, k > k'\right\} \cup \left\{(\boldsymbol{a}, \boldsymbol{a}[k := i]) : k \in [K], i \in [L] \setminus \boldsymbol{a}([K])\right\}.$$

This graph is chosen to ensure that with PBM setting any sub-optimal recommendation has a strictly better recommendation

Algorithm 4 S-GRAB: Static Graph for unimodal RAnking Bandit

Input: number of items L, number of positions K

$\gamma \leftarrow K(2L - K - 1)/2$

for $t = 1, 2, \dots$ **do**

$$\tilde{\boldsymbol{a}}(t) \leftarrow \operatorname*{argmax}_{\boldsymbol{a} \in \mathcal{P}_K^L} \sum_{k=1}^{K} \hat{\rho}_{a_k,k}(t)$$

$$\text{recommend } \boldsymbol{a}(t) = \begin{cases} \tilde{\boldsymbol{a}}(t) & \text{, if } \frac{\tilde{T}_{\tilde{\boldsymbol{a}}(t)}(t)}{\gamma+1} \in \mathbb{N}, \\ \operatorname*{argmax}_{\boldsymbol{a} \in \{\tilde{\boldsymbol{a}}(t)\} \cup \mathcal{N}_G(\tilde{\boldsymbol{a}}(t))} \sum_{k=1}^{K} b_{a_k,k}(t) & \text{, otherwise} \end{cases}$$

where $\mathcal{N}_G(\boldsymbol{a}) = \left\{\boldsymbol{a} \circ (k, k') : k, k' \in [K]^2, k > k'\right\} \cup \{\boldsymbol{a}[k := i] : k \in [K], i \in [L] \setminus \boldsymbol{a}([K])\}$

observe the clicks vector $\boldsymbol{c}(t)$

end for

in its neighborhood given G. This graph is fixed and does not require the knowledge of a mapping $\mathcal{P}$, but its degree is also about K times larger than the degree of the graphs handled by GRAB.

As for GRAB, any recommendation in the neighborhood of the leader given G differs with the leader at, at most two positions. Therefore a proof similar to the one of Theorem 1 ensures that S-GRAB's regret is upper-bounded by $\mathcal{O}\left(LK/\Delta_{min} \log T\right)$. This regret upper-bound is higher than GRAB's one by a factor K due to the larger size of the considered neighborhoods. However, this regret remains smaller than KL-CombUCB's one by a factor K thanks to the bounded number of differences between the leader and the arm played.