Parametric Graph for Unimodal Ranking Bandit
Supplementary Materials

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

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 $\{1, \ldots, n\}$
$\mathcal{P}_K^L$	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

CONTINUED ON NEXT PAGE

*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 38th International Conference on Machine Learning, PMLR 139, 2021.

SYMBOL	MEANING
$R(T)$	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 $\boldsymbol{a}$, $\tilde{T}_{\boldsymbol{a}}(t) \stackrel{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{\tilde{\boldsymbol{a}}(s) = \boldsymbol{a}\}$
$T_{\boldsymbol{a}}(t)$	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, $B_{\boldsymbol{a}}(t) = \sum_{k=1}^{K} b_{a_k,k}(t) - \sum_{k=1}^{K} b_{\tilde{a}_k(t),k}(t)$
$\mathcal{N}_{\pi^*}(a^*)$	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

References to Theorems

Lemma 1 (PBM Fulfills Assumption 1)**.**

Theorem 1 (Upper-Bound on the Regret of GRAB)**.**

Theorem 2 (Upper-Bound on the Regret of KL-CombUCB)**.**

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

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

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:

$$\exists k \in [K-1] \text{ s.t. } \theta_{a_{\pi_k}} < \theta_{a_{\pi_{k+1}}}, \tag{7}$$

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

$$\exists i \in [L] \setminus \boldsymbol{a}([K]) \text{ s.t. } \theta_{a_{\pi_K}} < \theta_i, \tag{9}$$

$$\begin{cases} \forall k \in [K-1], \theta_{a_{\pi_k}} \geqslant \theta_{a_{\pi_{k+1}}} \\ \forall k \in [K-1], \kappa_{\pi_k} \geqslant \kappa_{\pi_{k+1}} \\ \forall i \in [L] \setminus \boldsymbol{a}([K]), \theta_{a_{\pi_K}} \geqslant \theta_i \end{cases}. \tag{10}$$

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 (7) 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 (7) 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 (8) is satisfied, then Property (2) of Assumption 1 is fulfilled.

If Property (9) 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 (10) 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

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

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

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), $\mathbb{E}[|B_e|] = O(\log \log T)$, 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 $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}\}\right] = \mathcal{O}(\log\log T)$.

The proof requires notations related to the neighborhood of $\tilde{\boldsymbol{a}}$. Let $\mathcal{N} \stackrel{def}{=} \bigcup_{\boldsymbol{\pi}\in\mathcal{P}_K^K} \mathcal{N}_{\boldsymbol{\pi}}(\tilde{\boldsymbol{a}})$ 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

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

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

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

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 $n \in \Lambda_{i,k}(s)$, $T_{i,k}(n) \geqslant \min\left\{T_{\boldsymbol{a}}(n), T_{\boldsymbol{a}^+}(n)\right\} \geqslant \epsilon \tilde{T}_{\tilde{\boldsymbol{a}}}(n) = \epsilon s$. 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: $B = B^{t_0} \cup B_{t_0}^T$, where $B^{t_0} \overset{def}{=} \{t \in B : \tilde{T}_{\tilde{\boldsymbol{a}}}(t) \leqslant t_0\}$, $B_{t_0}^T \overset{def}{=} \{t \in B : \tilde{T}_{\tilde{\boldsymbol{a}}}(t) > t_0\}$, 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 $F \overset{def}{=} \{t \in [T] : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)\}$.

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

$T^{\tilde{\boldsymbol{a}}}_{\boldsymbol{a}}(s'-1) \geqslant \epsilon \tilde{T}_{\tilde{\boldsymbol{a}}}(t)$. Therefore, the set $\{s \in [t] : \tilde{\boldsymbol{a}}(s) = \tilde{\boldsymbol{a}}, T^{\tilde{\boldsymbol{a}}}_{\boldsymbol{a}(s)}(s) \geqslant \epsilon \tilde{T}_{\tilde{\boldsymbol{a}}}(t), \tilde{T}_{\tilde{\boldsymbol{a}}}(s)/L \notin \mathbb{N}\}$ is non-empty. We define $\psi(t)$ as the minimum on this set

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

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

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

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 $\Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)$ contains all the permutations of K elements, the set $\{t : \tilde{\boldsymbol{a}}(t) = \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\pi}}(t) \notin \Pi_{\boldsymbol{\rho}}\left(\tilde{\boldsymbol{a}}\right)\}$ 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,

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

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 $n \in \Lambda_{\tilde{\boldsymbol{\pi}},k}(s)$, $T_{\tilde{a}_{\tilde{\pi}_k},\tilde{\pi}_k}(n) \geqslant$

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 $t = t_0, t_0 + 1, \dots$ **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

$T_{\tilde{\boldsymbol{a}}}(n) \geqslant \tilde{T}_{\tilde{\boldsymbol{a}}}(n)/L = s/L$. 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,

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

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.

G.1. KL-CombUCB for Generic Setting

CombUCB1 (Kveton et al., 2015) 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., 2015), 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.

Algorithm 3 KL-ComUCB1 (applied to PBM)

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

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

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

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 $K_{max} = 2$.

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

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

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

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

end for

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: $\{\tilde{\boldsymbol{a}}(t)\}\cup\mathcal{N}_{\tilde{\boldsymbol{\pi}}(t)}\left(\tilde{\boldsymbol{a}}(t)\right)$ 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., 2015) 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

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

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

References

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.

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.

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, 2015.

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