



HAL
open science

Parametric Graph for Unimodal Ranking Bandit

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

► **To cite this version:**

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, Boammani Aser Lompo. Parametric Graph for Unimodal Ranking Bandit: Supplementary Materials. ICML 2021 - Thirty-eighth International Conference on Machine Learning, Jul 2021, Virtual, Canada. pp.1-13. hal-03256621v1

HAL Id: hal-03256621

<https://hal.science/hal-03256621v1>

Submitted on 11 Jun 2021 (v1), last revised 23 Jun 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Parametric Graph for Unimodal Ranking Bandit Supplementary Materials

Camille-Sovanneary Gauthier^{*12} Romaric Gaudel^{*3} Elisa Fromont⁴⁵² Boammani Aser Lompo⁶

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
θ	VECTORS OF PROBABILITIES OF CLICK
θ_i	PROBABILITY OF CLICK ON ITEM i
κ	VECTORS OF PROBABILITIES OF VIEW
κ_k	PROBABILITY OF VIEW AT POSITION k
\mathcal{A}	SET OF BANDIT ARMS
\mathbf{a}	AN ARM IN \mathcal{A}
$\mathbf{a}(t)$	THE ARM CHOSEN AT ITERATION t
$\tilde{\mathbf{a}}(t)$	BEST ARM AT ITERATION t GIVEN THE PREVIOUS CHOICES AND FEEDBACKS (CALLED LEADER)
\mathbf{a}^*	BEST ARM
G	GRAPH CARRYING A PARTIAL ORDER ON \mathcal{A}
γ	MAXIMUM DEGREE OF G
$\mathcal{N}_G(\tilde{\mathbf{a}}(t))$	NEIGHBORHOOD OF $\tilde{\mathbf{a}}(t)$ GIVEN G
$\rho_{i,k}$	PROBABILITY OF CLICK ON ITEM i DISPLAYED AT POSITION k
$\mathbf{c}(t)$	CLICKS VECTOR AT ITERATION t
$r(t)$	REWARD COLLECTED AT ITERATION t , $r(t) = \sum_{k=1}^K c_k(t)$
$\mu_{\mathbf{a}}$	EXPECTATION OF $r(t)$ WHILE RECOMMENDING \mathbf{a} , $\mu_{\mathbf{a}} = \sum_{k=1}^K \rho_{\mathbf{a},k}$
μ^*	HIGHEST EXPECTED REWARD, $\mu^* = \max_{\mathbf{a} \in \mathcal{P}_K^L} \mu_{\mathbf{a}}$
$\Delta_{\mathbf{a}}$	GAP BETWEEN $\mu_{\mathbf{a}}$ AND μ^*
Δ_{min}	MINIMAL VALUE FOR $\Delta_{\mathbf{a}}$
Δ	GENERIC REWARD GAP BETWEEN ONE OF THE SUB-OPTIMAL ARMS AND ONE OF THE BEST ARMS

CONTINUED ON NEXT PAGE

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

Parametric Graph for Unimodal Ranking Bandit (Supplementary Materials)

SYMBOL	MEANING
$R(T)$	CUMULATIVE (PSEUDO-)REGRET, $R(T) = T\mu^* - \mathbb{E} \left[\sum_{t=1}^T \mu_{\mathbf{a}(t)} \right]$
$\Pi_{\rho}(\mathbf{a})$	SET OF PERMUTATIONS IN \mathcal{P}_K^K ORDERING THE POSITIONS S.T. $\rho_{a_{\pi_1}, \pi_1} \geq \rho_{a_{\pi_2}, \pi_2} \geq \dots \geq \rho_{a_{\pi_K}, \pi_K}$
π	ELEMENT OF $\Pi_{\rho}(\mathbf{a})$
$\tilde{\pi}$	ESTIMATION OF π
$\mathbf{a} \circ (\pi_k, \pi_{k+1})$	PERMUTATION SWAPPING ITEMS IN POSITIONS π_k AND π_{k+1}
$\mathbf{a}[\pi_K := i]$	PERMUTATION LEAVING \mathbf{a} THE SAME FOR ANY POSITION EXCEPT π_K FOR WHICH $\mathbf{a}[\pi_K := i]_{\pi_K} = i$
\mathcal{F}	RANKINGS OF POSITIONS RESPECTING Π_{ρ} , $\mathcal{F} = (\pi_{\mathbf{a}})_{\mathbf{a} \in \mathcal{P}_K^K}$ S.T. $\forall \mathbf{a} \in \mathcal{P}_K^K, \pi_{\mathbf{a}} \in \Pi_{\rho}(\mathbf{a})$
$T_{i,k}(t)$	NUMBER OF ITERATIONS S.T. ITEM i HAS BEEN DISPLAYED AT POSITION k , $T_{i,k}(t) = \sum_{s=1}^{t-1} \mathbb{1}\{a_k(s) = i\}$
$\tilde{T}_{\mathbf{a}}(t)$	NUMBER OF ITERATIONS S.T. THE LEADER WAS \mathbf{a} , $\tilde{T}_{\mathbf{a}}(t) \stackrel{def}{=} \sum_{s=1}^{t-1} \mathbb{1}\{\tilde{\mathbf{a}}(s) = \mathbf{a}\}$
$T_{\mathbf{a}}(t)$	NUMBER OF ITERATIONS S.T. THE CHOSEN ARM WAS \mathbf{a} , $T_{\mathbf{a}}(t) = \sum_{s=1}^{t-1} \mathbb{1}\{\mathbf{a}(s) = \mathbf{a}\}$
$T_{\tilde{\mathbf{a}}}^{\mathbf{a}}(t)$	NUMBER OF ITERATIONS S.T. THE LEADER WAS $\tilde{\mathbf{a}}$, THE CHOSEN ARM WAS \mathbf{a} , AND \mathbf{a} WAS CHOSEN BY THE ARGMAX ON $\sum_{k=1}^K b_{a_k, k}(t)$: $T_{\tilde{\mathbf{a}}}^{\mathbf{a}}(t) = \sum_{s=1}^{t-1} \mathbb{1}\{\tilde{\mathbf{a}}(s) = \tilde{\mathbf{a}}, \mathbf{a}(s) = \mathbf{a}, \tilde{T}_{\tilde{\mathbf{a}}}(s)/L \notin \mathbb{N}\}$
$\hat{\rho}_{i,k}(t)$	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 $\hat{\rho}_{i,k}(t)$, $b_{i,k}(t) = f(\hat{\rho}_{i,k}(t), T_{i,k}(t), \tilde{T}_{\tilde{\mathbf{a}}}(t) + 1)$
f	KULLBACK-LEIBLER INDEX FUNCTION, $f(\hat{\rho}, s, t) = \sup\{p \in [\hat{\rho}, 1] : s \times \text{kl}(\hat{\rho}, p) \leq \log(t) + 3 \log(\log(t))\}$,
$\text{kl}(p, q)$	KULLBACK-LEIBLER DIVERGENCE FROM A BERNOULLI DISTRIBUTION OF MEAN p TO A BERNOULLI DISTRIBUTION OF MEAN q , $\text{kl}(p, q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right)$
$B_{\mathbf{a}}(t)$	PSEUDO-SUM OF INDICES OF \mathbf{a} AT ITERATION t , $B_{\mathbf{a}}(t) = \sum_{k=1}^K b_{a_k, k}(t) - \sum_{k=1}^K b_{\tilde{a}_k(t), k}(t)$
$\mathcal{N}_{\pi^*}(\mathbf{a}^*)$	NEIGHBORHOOD OF THE BEST ARM
$K_{\mathbf{a}}$	(WITH COMBINATORIAL BANDIT SETTING) NUMBER OF ELEMENTS IN \mathbf{a} BUT NOT IN \mathbf{a}^* , $K_{\mathbf{a}} = \min_{\mathbf{a}^* \in \mathcal{A}: \mu_{\mathbf{a}^*} = \mu^*} \mathbf{a} \setminus \mathbf{a}^* $
K_{max}	(WITH COMBINATORIAL BANDIT SETTING) MAXIMAL NUMBER OF ELEMENTS IN A SUB-OPTIMAL ARM \mathbf{a} BUT NOT IN AN OPTIMAL ARM \mathbf{a}^* , $K_{max} = \max_{\mathbf{a} \in \mathcal{A}: \mu_{\mathbf{a}} \neq \mu^*} K_{\mathbf{a}}$
$c^*(\boldsymbol{\theta}, \boldsymbol{\kappa})$	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 $\tilde{\mathbf{a}}(t) = \tilde{\mathbf{a}} \neq \mathbf{a}^*$).

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

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 $\mathbf{a} \in \mathcal{P}_K^K$ be a recommendation, and let $\pi \in \Pi_{\rho}(\mathbf{a})$ be an appropriate ranking of positions. One of the four following

properties is satisfied:

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

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

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

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

Let prove, by considering each of these properties one by one, that \mathbf{a} is either one of the best arms, or \mathbf{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 $\Pi_{\boldsymbol{\rho}}(\mathbf{a})$, $0 = \theta_{a_{\pi_k}} \kappa_{\pi_k} \geq \theta_{a_{\pi_{k+1}}} \kappa_{\pi_{k+1}} > 0$ 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 $\Pi_{\boldsymbol{\rho}}(\mathbf{a})$, and as $\rho_{i,k} = \theta_i \kappa_k$, $\theta_{a_{\pi_k}} \kappa_{\pi_k} \geq \theta_{a_{\pi_{k+1}}} \kappa_{\pi_{k+1}}$.

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

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

meaning $\mu_{\mathbf{a}} < \mu_{\mathbf{a} \circ (\pi_k, \pi_{k+1})}$, 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,

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

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

Finally, if Property (10) is satisfied, $\mu_{\mathbf{a}} = \mu^*$.

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

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 $(\mathbf{c}(t))_{t \in [T]}$. 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 $s \geq 1$ and all $t \in \Lambda(s)$, $T_{i,k}(t) \geq \epsilon s$, (iii) $|\Lambda(s)| \leq 1$, and (iv) the event $t \in \Lambda(s)$ is \mathcal{F}_t -measurable. Then for all $\delta > 0$,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 , $\mathbf{a}(t) = \mathbf{a}$ is chosen as $\sum_{e \in \mathbf{a}} b_e(t) = \max_{\mathbf{a}' \in \mathcal{A}} \sum_{e \in \mathbf{a}'} b_e(t)$.

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

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

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

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

Regret upper-bound Overall,

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

and

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

which concludes the proof. \square

E. Proof of Lemma 2 (Upper-bound on the Number of Iterations of GRAB for which

$$\tilde{\mathbf{a}}(t) = \tilde{\mathbf{a}} \neq \mathbf{a}^*)$$

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

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

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

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

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

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

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

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

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

and

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

Hence,

$$\mathbb{E}[\mathbb{1}\{\tilde{\mathbf{a}}(t) = \tilde{\mathbf{a}}\}] \leq \sum_{\mathbf{a}^+ \in \mathcal{N}^+} \mathbb{E}[|A_{\mathbf{a}^+}|] + \mathbb{E}[|B|].$$

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

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

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

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

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

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

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

Bound on $\mathbb{E}[|B|]$ We first split B in two parts: $B = B^{t_0} \cup B_{t_0}^T$, where $B^{t_0} \stackrel{def}{=} \{t \in B : \tilde{T}_{\tilde{\mathbf{a}}}(t) \leq t_0\}$, $B_{t_0}^T \stackrel{def}{=} \{t \in B : \tilde{T}_{\tilde{\mathbf{a}}}(t) > t_0\}$, 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 $L \geq 2$) as $\epsilon < 1/(2N)$.

We also define

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

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

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

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

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

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

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

So, there exists a recommendation \mathbf{a} such that $T_{\mathbf{a}}^{\tilde{\mathbf{a}}}(t) \geq \epsilon \tilde{T}_{\tilde{\mathbf{a}}}(t) + 1$. Let denote s' the first iteration such that $T_{\mathbf{a}}^{\tilde{\mathbf{a}}}(s') \geq \epsilon \tilde{T}_{\tilde{\mathbf{a}}}(t) + 1$. At this iteration, $T_{\mathbf{a}}^{\tilde{\mathbf{a}}}(s') = T_{\mathbf{a}}^{\tilde{\mathbf{a}}}(s' - 1) + 1$, meaning that $\tilde{\mathbf{a}}(s' - 1) = \tilde{\mathbf{a}}, \mathbf{a}(s' - 1) = \mathbf{a}, \tilde{T}_{\tilde{\mathbf{a}}}(s' - 1)/L \notin \mathbb{N}$, and

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

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

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

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

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

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

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

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

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

which contradicts the fact that \mathbf{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 $B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}$, $T_{\tilde{\mathbf{a}}(n)}^{\tilde{\mathbf{a}}}(n) = \lceil \epsilon \tilde{T}_{\tilde{\mathbf{a}}}(t) \rceil$ and $\tilde{T}_{\tilde{\mathbf{a}}}(t+1) = \tilde{T}_{\tilde{\mathbf{a}}}(t) + 1$. So $|B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}| < 1/\epsilon + 1$. Overall,

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

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

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

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

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

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

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

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

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

Proof of Theorem 3. Let $\tilde{\mathbf{a}}$ be a K -permutation of L items. If $\Pi_{\rho}(\tilde{\mathbf{a}})$ contains all the permutations of K elements, the set $\{t : \tilde{\mathbf{a}}(t) = \tilde{\mathbf{a}}, \tilde{\pi}(t) \notin \Pi_{\rho}(\tilde{\mathbf{a}})\}$ 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 $k' \neq k$: $\delta \stackrel{def}{=} \min\{\rho_{\tilde{a}_k,k} - \rho_{\tilde{a}_{k'},k'} : (k, k') \in [K]^2, \rho_{\tilde{a}_k,k} - \rho_{\tilde{a}_{k'},k'} > 0\}$. The gap δ is the minimum on a finite set, so $\delta > 0$.

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

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

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

Let $\tilde{\pi} \in \mathcal{P}_K^K$ be a ranking of positions, and $k \in [K]$ be a rank. $\Lambda_{\tilde{\pi},k} \subseteq \bigcup_{s \in \mathbb{N}} \Lambda_{\tilde{\pi},k}(s)$, with $\Lambda_{\tilde{\pi},k}(s) \stackrel{def}{=} \{t \in \Lambda_{\tilde{\pi},k} : \tilde{T}_{\tilde{\mathbf{a}}}(t) = s\}$. $|\Lambda_{\tilde{\pi},k}(s)| \leq 1$ as $\tilde{T}_{\tilde{\mathbf{a}}}(t)$ increases for each $t \in \Lambda_{\tilde{\pi},k}$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda_{\tilde{\pi},k}(s)$, $T_{\tilde{a}_{\tilde{\pi}_k}, \tilde{\pi}_k}(n) \geq$

Algorithm 2 KL-ComUCB1 (generic version)

Input: set of elements E , set of arms \mathcal{A}
 $t \leftarrow 1$
while $\{e \in E : T_e(t) = 0\} \neq \emptyset$ **do**
 $\tilde{E} \leftarrow \{e \in E : T_e(t) = 0\}$
 $\tilde{\mathcal{A}} \leftarrow \{\mathbf{a} \in \mathcal{A} : \mathbf{a} \cap \tilde{E} \neq \emptyset\}$

 recommend $\mathbf{a}(t) = \operatorname{argmax}_{\mathbf{a} \in \tilde{\mathcal{A}}} \sum_{e \in \mathbf{a}} b_e(t)$

 observe the weights $[w_e(t) : e \in \mathbf{a}]$
 $t \leftarrow t + 1$
end while
 $t_0 \leftarrow t$
for $t = t_0, t_0 + 1, \dots$ **do**

 recommend $\mathbf{a}(t) = \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}} \sum_{e \in \mathbf{a}} b_e(t)$

 observe the weights $[w_e(t) : e \in \mathbf{a}]$
end for
 $T_{\tilde{\mathbf{a}}}(n) \geq \tilde{T}_{\tilde{\mathbf{a}}}(n)/L = s/L$. Then, by Lemma 4

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

Hence,

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

 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 \mathbf{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 $\mathbf{a} \in \mathcal{A}$ and receives the reward $\sum_{e \in \mathbf{a}} w_e$, where \mathbf{w} is an independent draw of a distribution ν on $[0, 1]^E$. Given these assumptions, CombUCB1 chooses an arm $\mathbf{a}(t)$ 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 $t = 1, 2, \dots, L$ **do**

recommend $\mathbf{a}(t) = (((t-1)\%L) + 1, (t\%L) + 1, \dots, ((t+K-2)\%L) + 1)$

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

end for

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

recommend $\mathbf{a}(t) = \operatorname{argmax}_{\mathbf{a} \in \mathcal{P}_K^L} \sum_{k=1}^K b_{a_k, k}(t)$

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

end for

We denote $\rho_e \stackrel{def}{=} \mathbb{E}_{\mathbf{w} \sim \nu} [w_e]$ the expected reward associated to element e , $\mu_{\mathbf{a}} \stackrel{def}{=} \mathbb{E}_{\mathbf{w} \sim \nu} [\sum_{e \in \mathbf{a}} w_e] = \sum_{e \in \mathbf{a}} \rho_e$ the expected reward when choosing the arm $\mathbf{a} \in \mathcal{A}$, and $\mu^* \stackrel{def}{=} \max_{\mathbf{a} \in \mathcal{A}} \mu_{\mathbf{a}}$ the best expected reward. We also denote $\Delta_{\mathbf{a}} \stackrel{def}{=} \mu^* - \mu_{\mathbf{a}}$ the gap between the best expected reward and the reward of an arm \mathbf{a} , and $\Delta_{min} \stackrel{def}{=} \min_{\mathbf{a} \in \mathcal{A}: \Delta_{\mathbf{a}} > 0} \Delta_{\mathbf{a}}$ the smallest gap of a suboptimal arm. Finally, $K \stackrel{def}{=} \max_{\mathbf{a} \in \mathcal{A}} |\mathbf{a}|$ denotes the maximum size of an arm (meaning the maximum number of chosen elements), $K_{\mathbf{a}} \stackrel{def}{=} \min_{\mathbf{a}^* \in \mathcal{A}: \mu_{\mathbf{a}^*} = \mu^*} |\mathbf{a} \setminus \mathbf{a}^*|$ is the smallest number of elements to remove from \mathbf{a} to get an optimal arm, and $K_{max} \stackrel{def}{=} \max_{\mathbf{a} \in \mathcal{A}: \mu_{\mathbf{a}} \neq \mu^*} K_{\mathbf{a}}$ is its larger 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 $\mathbf{w}(t)$ is in $\{0, 1\}^E$. KL-CombUCB is depicted by Algorithm 2 which uses the following notations. At each iteration t , we denote

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

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

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

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

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

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

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

with

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

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

Kveton et al. prove that the regret of CombUCB1 is upper-bounded by $\mathcal{O}(|E|K/\Delta_{min} \log T)$, 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: $\mathcal{O}(|\mathcal{A}|K_{max}^2/\Delta_{min} \log T)$. 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 $t = 1, 2, \dots$ **do**

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

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

where $\mathcal{N}_G(\mathbf{a}) = \{\mathbf{a} \circ (k, k') : k, k' \in [K]^2, k > k'\} \cup \{\mathbf{a}[k := i] : k \in [K], i \in [L] \setminus \mathbf{a}([K])\}$
 observe the clicks vector $\mathbf{c}(t)$

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* $\Theta = \{(a_k, k) : k \in [K]\} : \mathbf{a} \in \mathcal{P}_K^L$, 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 $\mathcal{O}(K^2(L + \log K))$ 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 \mathbf{a} regarding $\sum_{k=1}^K b_{a_k, k}(t)$, they only differ from the considered set of permutations: $\{\tilde{\mathbf{a}}(t)\} \cup \mathcal{N}_{\tilde{\pi}(t)}(\tilde{\mathbf{a}}(t))$ 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., 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

$$V \stackrel{\text{def}}{=} \mathcal{P}_K^L,$$

$$E \stackrel{\text{def}}{=} \{(\mathbf{a}, \mathbf{a} \circ (k, k')) : k, k' \in [K]^2, k > k'\} \cup \{(\mathbf{a}, \mathbf{a}[k := i]) : k \in [K], i \in [L] \setminus \mathbf{a}([K])\}.$$

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 $\mathcal{O}(LK/\Delta_{\min} \log T)$. 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.