UniRank: Unimodal Bandit Algorithm for Online Ranking

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont

To cite this version:

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont. UniRank: Unimodal Bandit Algorithm for Online Ranking. ICML 2022 - 39th International Conference on Machine Learning, Jul 2022, Baltimore, United States. pp.1-31. hal-03740981

HAL Id: hal-03740981
https://inria.hal.science/hal-03740981v1

Submitted on 2 Aug 2022

UniRank: Unimodal Bandit Algorithm for Online Ranking

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

Abstract

We tackle, in the multiple-play bandit setting, the online ranking problem of assigning L items to K predefined positions on a web page in order to maximize the number of user clicks. We propose a generic algorithm, UniRank, that tackles state-of-the-art click models. The regret bound of this algorithm is a direct consequence of the unimodality-like property of the bandit setting with respect to a graph where nodes are ordered sets of indistinguishable items. The main contribution of UniRank is its $\mathcal{O}\left(L/\Delta \log T\right)$ regret for T consecutive assignments, where Δ relates to the reward-gap between two items. This regret bound is based on the usually implicit condition that two items may not have the same attractiveness. Experiments against state-of-the-art learning algorithms specialized or not for different click models, show that our method has better regret performance than other generic algorithms on real life and synthetic datasets.

1. Introduction

We consider *Online Recommendation Systems* (ORS) which choose K relevant items among L potential ones ($L \geq K$), such as songs, ads or movies to be displayed on a website. The user feedbacks, such as listening time, clicks, rates, etc., reflecting the user's appreciation with respect to each displayed item, are collected after each recommendation. As these feedbacks are only available for the items which were actually presented to the user, this setting corresponds to an instance of the *multi-armed bandit problem* with *semi-bandit feedback* (Gai et al., 2012; Chen et al., 2013). Besides, some displayed items are not looked at and lead to a negative feedback while they would be appreciated by the user. It raises a specific challenge related to ranking: the attention toward a displayed item is impacted by its position. Numerous approaches have been proposed to handle this partial attention (Radlinski et al., 2008; Combes et al., 2015; Lagrée et al., 2016) referred to as *multiple-play bandit* or *online learning to rank*. Several models of partial attention, a.k.a. click models, are considered in the state of the art (Richardson et al., 2007; Craswell et al., 2008) and have been transposed to the bandit framework (Kveton et al., 2015a; Komiyama et al., 2017). In current paper, in the same line as (Zoghi et al., 2017; Lattimore et al., 2018) we propose an algorithm which handles multiple state-of-the-art click models.

The main contribution of our work is a new bandit algorithm, UniRank, dedicated to a generic online learning to rank setting. UniRank takes inspiration from unimodal bandit algorithms (Combes & Proutière, 2014; Gauthier et al., 2021b): we implicitly consider a graph $\mathcal{G}$ on the partitions of the item-set such that the considered bandit setting is unimodal w.r.t. $\mathcal{G}$, and UniRank chooses each recommendation in the $\mathcal{G}$-neighborhood of an elicited partition. Thanks to this restricted exploration, UniRank is the first algorithm dedicated to a generic setting with a $O(L/\Delta \log T)$ regret upper-bound, while previous state-of-the-art algorithms were suffering a $O(LK/\Delta \log T)$ regret. Note that this $O(L/\Delta \log T)$ upper-bound requires all items' attractiveness to be different, which is a usual assumption satisfied by real world applications. Otherwise, UniRank recovers the $O(LK/\Delta \log T)$ bound. From an application point of view, UniRank has several interesting features: it handles multiple state-of-the-art click models altogether; it is simple to implement and efficient in terms of computation time; it does not require the knowledge of the time horizon T; and it exhibits a smaller empirical regret than other generic algorithms by leaning on the *different attractiveness* property when this property is satisfied.

As an indirect contribution, UniRank demonstrates that unimodality is a key tool to analyze the intrinsic complexity of some combinatorial semi-bandit problems. We also demonstrate the flexibility of unimodal bandit algorithms and of the proof of their regret upper-bound. In particular, we extend (Combes & Proutière, 2014)'s analysis to a graph which is

[*]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. Correspondence to: Camille-Sovanneary Gauthier <camille-sovanneary.gauthier@louisvuitton.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1. Required click model and upper-bound on cumulative regret for T consecutive recommendations for some well-known recommender algorithms that chose K items among L. The exact definition of Δ is specific to each algorithm. The symbol * means that Assumption 3.1*, defined in Section 3.1, is satisfied. $\boldsymbol{\kappa}$ denotes the vector of observation-probabilities of PBM, and γ is the degree of the graph explored by the unimodal bandit algorithm.

Algorithm	**Click model**	**Regret**
UniRank (our algorithm)	CM*	$\mathcal{O}\left((L-K)/\Delta \log T\right)$
	PBM*, …	$\mathcal{O}\left(L/\Delta \log T\right)$
	PBM, CM, …	$\mathcal{O}\left(LK/\Delta \log T\right)$
TopRank (Lattimore et al., 2018)	PBM, CM, …	$\mathcal{O}\left(LK/\Delta \log T\right)$
CascadeKL-UCB (Kveton et al., 2015a)	CM	$\mathcal{O}\left((L-K)/\Delta \log T\right)$
GRAB (Gauthier et al., 2021b)	PBM*	$\mathcal{O}\left(L/\Delta \log T\right)$
PB-MHB (Gauthier et al., 2021a)	PBM ($\boldsymbol{\kappa}_1 = 0$)	unknown
PBM-PIE (Lagrée et al., 2016)	PBM ($\boldsymbol{\kappa}$ known)	$\mathcal{O}\left((L-K)/\Delta \log T\right)$
SAM (Sentenac et al., 2021)	Matching*	$\mathcal{O}\left(L \log L/\Delta \log T\right)$
OSUB (Combes & Proutière, 2014)	Unimodal	$\mathcal{O}\left(\gamma/\Delta \log T\right)$

unimodal in a weaker sense: (i) UniRank takes its decisions given an optimistic index which is not based on the expected reward but on the probability for an item to be more attractive than another one and (ii) some sub-optimal nodes in the graph have no better node in their neighborhood.

The paper is organized as follows: Section 2 presents the related work and Section 3 defines our target setting. We then introduce UniRank in Section 4, and theoretical guarantees and empirical performance are presented respectively in Section 5 and Section 6. We conclude in Section 7.

2. Related Work

Table 1 shows a comparison of the assumptions and the regret upper-bounds of the most related algorithms.

Several bandit algorithms are designed to handle the online learning to rank setting while the user follows one of the currently defined click models, namely the *position based model* (PBM) (Komiyama et al., 2015; Lagrée et al., 2016; Komiyama et al., 2017; Gauthier et al., 2021a;b) or the *cascading model* (CM) (Kveton et al., 2015a;b; Combes et al., 2015; Zong et al., 2016; Katariya et al., 2016; Li et al., 2016; Cheung et al., 2019). To the best of our knowledge, only the algorithms BatchRank (Zoghi et al., 2017), TopRank (Lattimore et al., 2018), and BubbleRank (Li et al., 2019a) handle users following a general model covering both behaviors. These three algorithms exhibit a regret upper-bound for T consecutive recommendations of at least $\mathcal{O}(LK/\Delta \log T)$, where Δ depends on the *attraction probability* $\boldsymbol{\theta}$ of items.

One ingredient of TopRank and BubbleRank is a statistic to compare two items independently of the position at which they are displayed. The algorithm we propose also makes use of this statistic. However, we define an exploration strategy which does not require the knowledge of the time-horizon T and which induces a $\mathcal{O}(L/\Delta \log T)$ regret upper-bound when items have strictly different attractiveness.

UniRank also builds upon an extension of the *unimodal bandit* setting (Combes & Proutière, 2014; Gauthier et al., 2021b). This setting assumes the knowledge of a graph $\mathcal{G}$ on the set $\mathcal{A}$ of bandit arms, such that the expected reward $\mu_{\boldsymbol{a}}$ associated to each arm $\boldsymbol{a}$ satisfies the following assumption:

Assumption 2.1 (Unimodality[1])**.** There exists a unique arm $\boldsymbol{a}^* \in \mathcal{A}$ with highest expected reward, and for any arm $\boldsymbol{a} \in \mathcal{A}$, either (i) $\boldsymbol{a} = \boldsymbol{a}^*$, or (ii) there exists $\boldsymbol{a}^+$ in the neighborhood $\mathcal{N}_{\mathcal{G}}(\boldsymbol{a})$ of $\boldsymbol{a}$ given $\mathcal{G}$ such that $\mu_{\boldsymbol{a}^+} > \mu_{\boldsymbol{a}}$.

The unimodal bandit algorithms are aware of $\mathcal{G}$, but ignore the weak order induced by the edges of $\mathcal{G}$. However, they rely on $\mathcal{G}$ to efficiently browse the arms up to the best one. Typically, the algorithm OSUB (Combes & Proutière, 2014) selects at each iteration t, an arm $\boldsymbol{a}(t)$ in the neighborhood $\mathcal{N}_{\mathcal{G}}(\tilde{\boldsymbol{a}}(t))$ given $\mathcal{G}$ of the current best arm $\tilde{\boldsymbol{a}}(t)$ (a.k.a. the *leader*). By restricting the exploration to this neighborhood, the regret suffered by OSUB scales in $\mathcal{O}(\gamma/\Delta \log T)$, where γ is the maximum degree of $\mathcal{G}$, to be compared with $\mathcal{O}(|\mathcal{A}|/\Delta \log T)$ if the arms were independent. OSUB is designed for the standard bandit setting and makes use of estimators of the expected reward of arms to select the leader and chose the arm to play. In comparison, UniRank extends OSUB's idea to the semi-bandit setting, relies on a new variant of the unimodality property (see Lemma 5.4), selects the leader and the recommended arm based on other statistics, and does not require the 'forced exploitation' step which consists in recommending the leader each γ-th iteration.

[1] Usually, the unimodality is defined as the existence of a strictly increasing path from any sub-optimal arm to $\boldsymbol{a}^*$. Assumption 2.1 is equivalent and we use it in this paper as it directly relates to the shape of the theoretical analysis.

Finally, (Gauthier et al., 2021b) also builds upon the unimodality framework to solve a learning to rank problem in the bandit setting. However, the corresponding algorithm (GRAB) is dedicated to the PBM click model. In this model, there is a natural statistic to look at to measure the quality of an item and a position: the probability of click when presenting item i in position k. This statistic is independent of the items at other positions. Within a CM model, such a statistic does not exist. Instead, we refer to a statistic related to the relative attractiveness of items i and j (which we denote $\hat{s}_{i,j}$). Secondly, in the PBM model, only weak assumptions are needed to guarantee a unique optimal recommendation, which is required to get the unimodality property. With a CM model, any recommendation including the K best items leads to the optimal reward. To recover the unicity of the best arm, our algorithm is not targeting the best reward, but the best ranking of items (which implies the best reward). However, when facing PBM model, our algorithm requires an assumption which is omitted by GRAB: the position are indexed from the most look-at position to the least one.

3. Learning to Rank in a Semi-Bandit Setting

We consider the following *online learning to rank (OLR) problem with clicks feedback*. For any integer n, let $[n]$ denote the set $\{1, \ldots, n\}$. A recommendation $\boldsymbol{a} = (a_1, \ldots, a_K)$ is a permutation of K distinct items among L, where a_k is the item displayed at position k and $\boldsymbol{a}([K]) := \{a_k : k \in [K]\}$ is the set of all displayed items. We denote $\mathcal{P}_K^L$ the set of such permutations. Throughout the paper, we will use the terms *permutation* and *recommendation* interchangeably to denote an element of $\mathcal{P}_K^L$.

An instance of our OLR problem is a tuple (L, K, ρ), where L is the number of available items, $K \leqslant L$ is the number of positions to display the items, and ρ is a function from $\mathcal{P}_K^L \times [K]$ to $(0, 1]$ such that for any recommendation $\boldsymbol{a}$ and position k, $\rho(\boldsymbol{a}, k)$ is the probability for a user to click on the item displayed at position k when recommending $\boldsymbol{a}$.

A recommendation algorithm is only aware of L and K and has to deliver T consecutive recommendations. At each iteration $t \in [T]$, the algorithm recommends a permutation $\boldsymbol{a}(t)$ and observes the values $c_{a_1(t)}(t), \ldots, c_{a_K(t)}(t)$, where for any position k, $c_{a_k(t)}(t)$ equals 1 whenever the user clicks on the item $a_k(t)$, and 0 otherwise. To keep notations simple, we also define $c_i(t) = 0$ for each undisplayed item $i \in [L] \setminus \boldsymbol{a}(t)([K])$. Note that the recommendation at time t is only based on previous recommendations and observations.

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

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

where the expectation is taken w.r.t. the recommendations from the algorithm and the clicks.

Illustration 3.1 (Click model PBM)**.** With the click model PBM, at each iteration t, the user looks at the position k with probability κ_k, independently of the displayed items $\boldsymbol{a}(t)$. Moreover, whenever she observes the position k, she clicks on the corresponding item $a_k(t)$ with probability $\theta_{a_k(t)}$, independently of her other actions. Overall, the clicks $c_{a_k(t)}(t)$ are independent and $\rho(\boldsymbol{a}(t), k) = \mathbb{E}\left[c_{a_k(t)}(t)\right] = \kappa_k \theta_{a_k(t)}$. Therefore, the optimal recommendation consists in displaying the item i with the ℓ-th highest value θ_i at the position k with the ℓ-th highest value θ_k. Hence, if $\theta_1 > \theta_2 > \cdots > \theta_K > \max_{K<k\leqslant L} \theta_k$ and $\kappa_1 > \kappa_2 > \cdots > \kappa_K$, $\mu^* = \sum_{k=1}^{K} \kappa_k \theta_k$.

3.1. Modeling Assumption

Up to now, an OLR problem assumes two main properties: (i) a click at a position is a random variable only conditioned by the recommendation and the position, and (ii) the expectation of the corresponding distribution is fixed. We now introduce the three assumptions required by UniRank, which are fulfilled by PBM and CM click models.

We first assume an order on items. Note that the existence of an order on item is a weak assumption by itself (we may chose any random order). The strength of this assumption derives from Assumptions 3.2 and 3.3 which enforces this order to relate with expected reward.

Assumption 3.1 (Strict weak order)**.** There exists a *preferential attachment* function $g : [L] \to \mathbb{R}$ on items, and for any pair of items (i, j),

- if $g(i) > g(j)$, item i is said *more attractive* than item j, which we denote $i \succ j$;
- if $g(i) = g(j)$, item i is said *equivalent* to item j, which we denote $i \sim j$.

Illustration 3.2 (Strict weak order with PBM)**.** With PBM, a typical choice for the function g is $g : i \mapsto \theta_i$.

Assumption 3.1 ensures the existence of a strict weak order $\succ$ on items: the items may be ranked by attractiveness, some items being equivalent. A typical example with $L = 4$ would be $1 \succ 2 \sim 3 \succ 4$, meaning item 1 is more attractive than any other item, and items 2 and 3 are equivalent and more attractive than item 4. Such situation may also be represented with an ordered partition: $(\{1\}, \{2, 3\}, \{4\})$, where if the subset E is listed before the subset F, then for any item $i \in E$ and any item $j \in F$, $i \succ j$. In the rest of the paper we will use either the preferential attachment function, or its associated strict weak order, or the corresponding ordered partition depending on the most appropriate representation.

The strongest results of the theoretical analysis require the slightly stronger assumption which ensures that the K best items are uniquely defined. This assumption is equivalent to any of both hypothesis: (i) the order $\succ$ is total on the K best items and the K-ith item is strictly more attractive than remaining $L-K$ items, and (ii) each of the K first subsets of the ordered partition is composed of only one item.

Assumption 3.1* (Strict total order on top-K items)**.** *There exists a* preferential attachment *function* $g : [L] \to \mathbb{R}$ *and a permutation* $\boldsymbol{a} \in \mathcal{P}_K^L$ *s.t.* $g(a_1) > g(a_2) > \cdots > g(a_K)$ *and for any item* $j \in [L] \setminus \boldsymbol{a}([K])$, $g(a_K) > g(j)$.

Our next assumption states that recommending the items according to the order $\succ$ associated to the preferential attachment leads to an optimal recommendation.

Definition 3.1 (Compatibility with a strict weak order)**.** Let $\succ$ be a strict weak order on items, and $\boldsymbol{a}$ be a recommendation. The recommendation $\boldsymbol{a}$ is *compatible* with $\succ$ if

1. for any position $k \in [K-1]$, either $a_k \succ a_{k+1}$ or $a_k \sim a_{k+1}$;
2. for any item $j \in [L] \setminus \boldsymbol{a}([K])$, either $a_K \succ j$ or $a_K \sim j$.

Assumption 3.2 (Optimal reward)**.** Any recommendation $\boldsymbol{a}$ compatible with $\succ$ is optimal, meaning $\mu_{\boldsymbol{a}} = \mu^*$.

Illustration 3.3 (Optimal reward with PBM)**.** With PBM, if the positions are ranked by decreasing observation probabilities and $g(i) = \theta_i$, this assumption means that the recommendation placing the k-th most attractive item at the k-th most observed position is optimal, which indeed is true.

Assumption 3.2 is of utmost importance for UniRank as it means that identifying a partition of the items coherent with $\succ$ is sufficient to ensure optimal recommendations.

Let us now consider the last assumption which regards the expectation of the random variable $c_i(t) - c_j(t)$.

Definition 3.2 (Expected click difference)**.** Let i and j be two items, and $\boldsymbol{a}$ a recommendation. The *probability of difference* and the *expected click difference* between items i and j w.r.t. the recommendation $\boldsymbol{a}$ are respectively:

$$\tilde{\delta}_{i,j}(\boldsymbol{a}) = \mathbb{P}_{\boldsymbol{a}(t)\sim\mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}\left[c_i(t) \neq c_j(t)\right] \text{ and}$$

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}\left[c_i(t) - c_j(t) \mid c_i(t) \neq c_j(t)\right],$$

where $(i,j)\circ\boldsymbol{a}$ is the permutation $\boldsymbol{a}$ such that items i and j have been swapped, and $\mathcal{U}(S)$ is the uniform distribution on the set S. If only i (respectively j) belongs to $\boldsymbol{a}$, $(i,j)\circ\boldsymbol{a}$ is the permutation $\boldsymbol{a}$ where item i is replaced by item j (resp. j by i). If neither i nor j belongs to $\boldsymbol{a}$, $(i,j)\circ\boldsymbol{a}$ is $\boldsymbol{a}$.

Algorithm 1 UniRank: Unimodal Bandit Algorithm for Online Ranking

```
Require: number of items $L$, number of positions $K$
for $t = 1, 2, \dots$ do
    compute the leader partition $\tilde{\boldsymbol{P}}(t)$
    $\boldsymbol{P}(t) \leftarrow \mathrm{argmax}_{\boldsymbol{P}\in\{\tilde{\boldsymbol{P}}(t)\}\cup\mathcal{N}(\tilde{\boldsymbol{P}}(t))}\, b_{\boldsymbol{P}}(t)$
    draw the recommendation $\boldsymbol{a}(t)$ uniformly at random
    in $\mathcal{A}(\boldsymbol{P}(t))$
    observe the clicks vector $\boldsymbol{c}(t)$
end for
```

Assumption 3.3 (Order identifiability)**.** The strict weak order $\succ$ on items is *identifiable*, meaning that for any couple of items (i,j) in $[L]^2$ s.t. $i \succ j$, and for any recommendation $\boldsymbol{a} \in \mathcal{P}_K^L$ s.t. at least one of both items is displayed, $\tilde{\delta}_{i,j}(\boldsymbol{a}) \neq 0$ and $\tilde{\Delta}_{i,j}(\boldsymbol{a}) > 0$.

Illustration 3.4 (Expected click difference with PBM)**.** With the click model PBM, if the positions are ranked by decreasing observation probabilities, for any recommendation $\boldsymbol{a}$, any position $k \in [K]$ and any position $\ell \in [L] \setminus \{k\}$, denoting i and j the items at respective positions k and ℓ, $\tilde{\delta}_{i,j}(\boldsymbol{a}) = \frac{1}{2}(\theta_i + \theta_j)(\kappa_k + \kappa_\ell) - 2\theta_i\theta_j\kappa_k\kappa_\ell$ and $\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\theta_i - \theta_j}{\theta_i + \theta_j} d_{i,j}(\boldsymbol{a})$, where $d_{i,j}(\boldsymbol{a}) > 1$. Therefore, if $g(i) = \theta_i$, Assumption 3.3 is fulfilled.

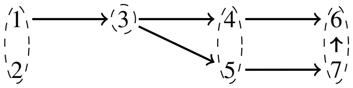
The expected click difference reflects the fact that an item leads to more clicks than another independently of the position of both items (other items being unchanged). Hence, Assumption 3.3 points out that when an item is more attractive than another one, it has a higher probability to be clicked upon, all other things being equal. This assumption is natural and ensures that the order on items may be recovered from the expected click difference, which is observed.

Finally, the following lemma, proven in Appendix D, states that both CM and PBM models fulfill our assumptions.

Lemma 3.1. *Let* (L, K, ρ) *be an online learning to rank problem with users following CM or PBM model with positions ranked by decreasing observation probabilities. Then Assumptions 3.1, 3.2, and 3.3 are fulfilled. Furthermore, Assumption 3.1** *is fulfilled if, for any top-*K *item* i *and any item* j *in* $[L] \setminus \{i\}$, *either* $i \succ j$ *or* $j \succ i$.

4. UniRank Algorithm

Our algorithm, UniRank, is detailed in Algorithm 1, and Figure 1 unfolds one of its iterations. This algorithm takes inspiration from the unimodal bandit algorithm OSUB (Combes & Proutière, 2014) by selecting at each iteration t an *arm to play* $\boldsymbol{P}(t)$ in the neighborhood of the current best one



$$
\begin{aligned}
\tilde{\boldsymbol{P}} &= \left(\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\right) = (\{1,2\},\{3\},\{4,5\},\{6,7\})\\
\mathcal{N}(\tilde{\boldsymbol{P}}) &= \{(\{1,2,3\},\{4,5\},\{6,7\})\,, && \text{\% merge of } \tilde{P}_1 \text{ and } \tilde{P}_2\\
&\quad\ (\{1,2\},\{3,4,5\},\{6,7\})\,, && \text{\% merge of } \tilde{P}_2 \text{ and } \tilde{P}_3\\
&\quad\ (\{1,2\},\{3\},\{4,5,6\},\{7\})\,, && \text{\% try 6}\\
&\quad\ (\{1,2\},\{3\},\{4,5,7\},\{6\})\} && \text{\% try 7}\\
\boldsymbol{P} &= (\{1,2\},\{3,4,5\},\{6,7\}) && \text{\% the } 2^{nd} \text{ neighbor wins}\\
\boldsymbol{a} &= (2,1,3,5)
\end{aligned}
$$

Figure 1. One iteration of UniRank with $L = 7$ items and $K = 4$ positions (t is omitted for clarity). Each arrow $i \to j$ in the top graph on items means the statistic $\hat{s}_{i,j}$ is non-negative. With these values, the leader partition (represented with dashed ellipses) is $\tilde{\boldsymbol{P}} = (\{1,2\},\{3\},\{4,5\},\{6,7\})$, where $\tilde{P}_4 = \{6,7\}$ gathers remaining items as the 3 first partitions contain more than K items. Then, we assume that $\max(\bar{\bar{s}}_{3,1}, \bar{\bar{s}}_{3,2}) > \max(\max(\bar{\bar{s}}_{4,3}, \bar{\bar{s}}_{5,3}),\ \max(\bar{\bar{s}}_{6,4}, \bar{\bar{s}}_{6,5}),\ \max(\bar{\bar{s}}_{7,4}, \bar{\bar{s}}_{7,5}),\ 0)$. Therefore, UniRank plays the optimistic partition $\boldsymbol{P} = (\{1,2\},\{3,4,5\},\{6,7\})$. Finally the recommendation $\boldsymbol{a}$ is obtained by concatenating a random permutation of $P_1 = \{1,2\}$ with a random permutation of 2 items from $P_2 = \{3,4,5\}$.

$\tilde{\boldsymbol{P}}(t)$ (a.k.a. the *leader*). However, UniRank's arms are not recommendations but sets of recommendations represented by ordered partitions. Hence, the recommendation $\boldsymbol{a}(t)$ is drawn uniformly at random in the subset $\mathcal{A}(\boldsymbol{P}(t))$ of recommendations compatible with $\boldsymbol{P}(t)$.

Let us now first define the notations used by UniRank and then present its concrete behaviour.

Statistic $\hat{s}_{i,j}(t)$ UniRank's choices are based on the statistic $\hat{s}_{i,j}(t)$ and the optimistic estimator of its expected value: the Kullback-Leibler-based one denoted $\bar{\bar{s}}_{i,j}(t)$. $\hat{s}_{i,j}(t)$ is the average value of $c_i(s) - c_j(s)$ for s in $[t-1]$, where we restrict ourselves to iterations at which items i and j are in the same subset of the played partition $\boldsymbol{P}(s) = \big(P_1(s), \ldots, P_{d(s)}(s)\big)$, and $c_i(s) \neq c_j(s)$. Specifically,

$$\hat{s}_{i,j}(t) := \frac{1}{T_{i,j}(t)} \sum_{s=1}^{t-1} O_{i,j}(s)\left(c_i(s) - c_j(s)\right),$$

where $O_{i,j}(s) := \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\} \mathbb{1}\{c_i(s) \neq c_j(s)\}$ denotes that the difference between items i and j is observable at iteration s, $T_{i,j}(t) := \sum_{s=1}^{t-1} O_{i,j}(s)$, and $\hat{s}_{i,j}(t) := 0$ when $T_{i,j}(t) = 0$. Note that $\hat{s}_{i,j}(t)$ is antisymmetric ($\hat{s}_{i,j}(t) = -\hat{s}_{j,i}(t)$) and $\hat{s}_{i,j}(t) > 0$ (equivalent to $\hat{s}_{j,i}(t) < 0$) indicates that i is probably more attractive than j.

The statistics $\hat{s}_{i,j}(t)$ are paired with their respective optimistic *indices*

$$\bar{\bar{s}}_{i,j}(t) := 2 * f\left(\frac{1 + \hat{s}_{i,j}(t)}{2}, T_{i,j}(t), \tilde{t}_{\tilde{\boldsymbol{P}}(t)}(t)\right) - 1,$$

where f is a function from $[0,1] \times \mathbb{N} \times \mathbb{N}$ to $[0,1]$ and $f(\hat{\mu}, N, t) := \sup\{\mu \in [\hat{\mu}, 1] : N \times \mathrm{kl}(\hat{\mu}, \mu) \leq \log(t) + 3\log(\log(t))\}$, with $\mathrm{kl}(p,q) := p \log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ the *Kullback-Leibler divergence* (KL) from a Bernoulli distribution of mean p to a Bernoulli distribution of mean q; $f(\hat{\mu}, N, t) := 0$ when $\hat{\mu} = 1$, $N = 0$, or $t = 0$; and $\tilde{t}_{\tilde{\boldsymbol{P}}(t)}(t)$ is the number of iterations the partition at which $\tilde{\boldsymbol{P}}$ has previously been the leader. This optimistic index is the one used for KL-based bandit algorithms, after a rescaling of $\hat{s}_{i,j}(t)$ to the interval $[0,1]$. Note that, unlike $\hat{s}_{i,j}(t)$, $\bar{\bar{s}}_{i,j}(t)$ is not antisymmetric, and $\bar{\bar{s}}_{j,i}(t) \geqslant 0$ while $\hat{s}_{i,j}(t) > 0$ indicates that it is unclear whether i is more attractive than j or not.

Leader Elicitation At each iteration, UniRank first builds a partition $\tilde{\boldsymbol{P}}(t) = (\tilde{P}_1(t), \ldots, \tilde{P}_{\tilde{d}}(t))$ using Algorithm 2 (see. Appendix C). This partition is *coherent* with $\hat{s}_{i,j}(t)$, meaning that for any couple of items (i,j) in $[L]^2$, if $\hat{s}_{i,j}(t) > 0$ then either i belongs to a subset $\tilde{P}_c(t)$ ranked before the subset of j, or there exists a cycle $(i_1, i_2, \ldots, i_N)$ such that $i_1 = i_N = i$, $i_2 = j$, and for any $n \in [N-1]$, $\hat{s}_{i_n, i_{n+1}}(t) > 0$. We also ensure that the $\tilde{d} - 1$ first subsets of $\tilde{\boldsymbol{P}}(t)$ gather at least K items: $\sum_{c=1}^{\tilde{d}-2} |\tilde{P}_c(t)| < K \leqslant \sum_{c=1}^{\tilde{d}-1} |\tilde{P}_c(t)|$. This means that the items in $\tilde{P}_{\tilde{d}}(t)$ are the ones which are never displayed by the recommendations in $\mathcal{A}(\tilde{\boldsymbol{P}}(t))$. Note that the subset $\tilde{P}_{\tilde{d}}(t)$ may be empty.

The partition $\tilde{\boldsymbol{P}}(t)$ is built by repeating the process of (i) identifying the smallest subset of items dominating all other items (meaning the items i for which $\hat{s}_{i,j}(t) > 0$ for any remaining item j), and (ii) removing this subset. A special care is taken to gather in the same subset remaining items as soon as the first subsets contain more than K items.

Optimistic Partition Elicitation The partition $\tilde{\boldsymbol{P}}(t)$ plays the role of leader, meaning that at each iteration, UniRank solves an exploration-exploitation dilemma and picks either $\tilde{\boldsymbol{P}}(t)$ or a permutation $\boldsymbol{P}(t)$ in the neighborhood $\mathcal{N}(\tilde{\boldsymbol{P}}(t))$ of $\tilde{\boldsymbol{P}}(t)$, where $\mathcal{N}(\tilde{\boldsymbol{P}}) :=$

$$\left\{\left(\tilde{P}_1, \ldots, \tilde{P}_{c-1}, \tilde{P}_c \cup \tilde{P}_{c+1}, \tilde{P}_{c+2}, \ldots \tilde{P}_{\tilde{d}}\right) : c \in [\tilde{d} - 2]\right\} \cup \left\{\left(\tilde{P}_1, \ldots, \tilde{P}_{\tilde{d}-2}, \tilde{P}_{\tilde{d}-1} \cup \{j\}, \tilde{P}_{\tilde{d}} \setminus \{j\}\right) : j \in \tilde{P}_{\tilde{d}}\right\}.$$

This neighborhood results either (i) from the merge of two consecutive subsets $\tilde{P}_c(t)$ and $\tilde{P}_{c+1}(t)$ of the partition $\tilde{\boldsymbol{P}}(t)$, or (ii) from the addition to $\tilde{P}_{\tilde{d}-1}(t)$ of an item j from the last subset. For each neighbor $\boldsymbol{P}$ of type (i), the optimistic index $b_{\boldsymbol{P}}(t)$ is $\max_{(i,j) \in \tilde{P}_c(t) \times \tilde{P}_{c+1}(t)} \bar{\bar{s}}_{j,i}(t)$ to reflect whether or not at least one of the items in $\tilde{P}_{c+1}(t)$ may potentially be more attractive than one of the items in $\tilde{P}_c(t)$. Similarly, for each neighbor $\boldsymbol{P}$ of type (ii), the optimistic index $b_{\boldsymbol{P}}(t)$ is $\max_{i \in \tilde{P}_{\tilde{d}-1}(t)} \bar{\bar{s}}_{j,i}(t)$.

Remark 4.1 (Recommendation chosen at random). Taking a random permutation is required to control the statistic $\hat{s}_{i,j}(t)$. Indeed, the theoretical analysis requires the probability for i to be ranked before j in the recommendation to be even. Overall, the aim is to identify a partition $\boldsymbol{P}^*$ such that any permutation in $\mathcal{A}(\boldsymbol{P}^*)$ is compatible with the unknown strict weak order on items.

5. Theoretical Analysis

The proof of the upper-bound on the regret of UniRank follows a similar path as the proof of OSUB (Combes & Proutière, 2014): (i) apply a standard bandit analysis to control the regret under the condition that the leader $\tilde{\boldsymbol{P}}(t)$ is an optimal partition, and (ii) upper-bound by $\mathcal{O}(\log\log T)$ the expected number of iterations such that $\tilde{\boldsymbol{P}}(t)$ is not an optimal partition. However, both steps differ from (Combes & Proutière, 2014). First, UniRank handles partitions instead of recommendations. Secondly, it builds upon $\hat{s}_{i,j}(t)$ instead of estimators of the expected reward. While $\hat{s}_{i,j}(t)$ is the average of dependent random variables with different expected values, these expected values are greater than some non-negative constant $\tilde{\Delta}_{i,j}$ when $i \succ j$, which is sufficient to lower-bound $\hat{s}_{i,j}(t)$ away from 0 as required by the proof of the regret upper-bound (see Appendices E.1, E.2, and E.3 for details). Finally, the proof is adapted to handle the fact that $T_{i,j}(t)$ randomly increases when we play items i and j due to the exploration-exploitation rule, which is unusual in the bandit literature. Up to our knowledge, this exploration-exploitation strategy and its analysis are new in the bandit community. We believe that it opens new perspectives for other semi-bandit settings.

Note that, as in (Sentenac et al., 2021) and (Gauthier et al., 2021b), we restrict the theoretical analysis to the setting where the order on top-items is total, meaning we use Assumption 3.1*. Without loss of generality, we also assume that $1 \succ 2 \succ \cdots \succ K \succ [L]\setminus[K]$ to shorten the notations. Hence the only partition $\boldsymbol{P}^*$ which is such that, any permutation $\boldsymbol{a}$ in $\mathcal{A}(\boldsymbol{P}^*)$ is compatible with the unknown strict order on items, is $(\{1\},\dots,\{K\},[L]\setminus[K])$.

We now propose the main theorem that upper-bounds the regret of UniRank.

Theorem 5.1 (Upper-bound on the regret of UniRank assuming a total order on top-K items)**.** *Let (L, K, ρ) be an OLR problem satisfying Assumptions 3.1*, 3.2, and 3.3 and such that $1 \succ 2 \succ \cdots \succ K \succ [L]\setminus[K]$. Denoting $\boldsymbol{P}^* = (\{1\},\dots,\{K\},[L]\setminus[K])$ the optimal partition associated to this order, when facing this problem, UniRank fulfills*

$$\forall k \in [L]\setminus\{1\},\ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\\ \exists c, P_c(t)=\{\min(k-1,K),k\}}\right\}\right] \leqslant \frac{16}{\tilde{\delta}^*_k \tilde{\Delta}_k^2} \log T + \mathcal{O}(\log\log T) \quad (2)$$

$$\textit{and } \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\tilde{\boldsymbol{P}}(t) \neq \boldsymbol{P}^*\}\right] = \mathcal{O}(\log\log T), \quad (3)$$

and hence

$$R(T) \leqslant \sum_{k=2}^{L} \frac{8\Delta_k}{\tilde{\delta}^*_k \tilde{\Delta}_k^2} \log T + \mathcal{O}(\log\log T) = \mathcal{O}\left(\frac{L}{\Delta}\log T\right),$$

where for any position $k > 1$, denoting $\ell := \min(k-1, K)$,
$\tilde{\delta}^*_k := \min_{\boldsymbol{P} \in \mathcal{N}(\boldsymbol{P}^*): \exists c, (\ell,k) \in P_c^2} \mathbb{P}_{\boldsymbol{a}(t) \sim \mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_\ell(t) \neq c_k(t)\right]$,
$\tilde{\Delta}_k := \min_{\boldsymbol{a} \in \mathcal{P}_K^L : \{\ell,k\} \cap \boldsymbol{a}([K]) \neq \varnothing} \tilde{\Delta}_{\ell,k}(\boldsymbol{a})$,
$\Delta_k := \mu_{(1,\dots,K)} - \mu_{(\ell,k)\circ(1,\dots,K)}$,
and $\Delta := \min_{k \in \{2,\dots,L\}} \tilde{\delta}^*_k \tilde{\Delta}_k^2 / \Delta_k$.

The first upper-bound (Equation (2)) controls the expected number of iterations at which UniRank explores while the leader is the optimal partition. Both types of exploration are covered: the merging of two consecutive subsets of $\tilde{\boldsymbol{P}}(t)$, and the addition of a sub-optimal arm to the last subset of the chosen partition $\boldsymbol{P}(t)$. The second upper-bound (Equation (3)) deals with the expected number of iterations at which the leader is not the optimal partition. Let us now express the same bounds while assuming one of the state-of-the-art click models.

Corollary 5.2 (Facing CM*)**.** *Under the hypotheses of Theorem 5.1, with the clik-model CM with probability θ_i to click on item i when it is observed, UniRank fulfills*

$$R(T) \leqslant \sum_{k=K+1}^{L} 16 \frac{\theta_K + \theta_k}{\theta_K - \theta_k} \log T + \mathcal{O}(\log\log T)$$
$$= \mathcal{O}\left((L-K)\frac{\theta_K + \theta_{K+1}}{\theta_K - \theta_{K+1}} \log T\right).$$

Corollary 5.3 (Facing PBM*)**.** *Under the hypotheses of Theorem 5.1, if the user follows PBM with the probability θ_i of clicking on item i when it is observed and the probability κ_k of observing the position k, then UniRank fulfills*

$$R(T) = \mathcal{O}\left(\frac{L}{\Delta}\log T\right),$$

where $\Delta := \min\{\frac{\theta_K - \theta_{K+1}}{\theta_K + \theta_{K+1}}, \min_{k \in \{2,\dots,K\}} \frac{((\kappa_{k-1}+\kappa_k)(\theta_{k-1}+\theta_k) - 4\kappa_{k-1}\kappa_k\theta_{k-1}\theta_k)(\theta_{k-1}-\theta_k)}{(\kappa_{k-1}-\kappa_k)(\theta_{k-1}+\theta_k)^2}\}$.

Note that the regret upper-bound reduces to $\mathcal{O}((L-K)/\Delta \log T)$ with CM since, with this model, the recommendation is optimal as soon as optimal items are displayed.

A more detailed version of these corollaries is given in the appendix, together with their proofs and Theorem 5.1's proof. These proofs builds upon the following pseudo-unimodality property.

Lemma 5.4 (Pseudo-unimodality assuming a total order on top-K items)**.** *Under the hypotheses of Theorem 5.1, for any ordered partition of the items* $\tilde{\boldsymbol{P}} = \left(\tilde{P}_1, \dots, \tilde{P}_{\tilde{d}}\right) \neq \boldsymbol{P}^*$*,*

- *either* $\exists c \in [\tilde{d}]$*, such that* $|P_c| > 1$ *and* $i^* \succ \operatorname{argmax}_{j \in P_c \setminus \{i^*\}} g(j)$*, where* $i^* = \operatorname{argmax}_{i \in P_c} g(i)$*;*
- *or* $\exists c \in [\tilde{d} - 1]$*,* $\exists (i,j) \in \tilde{P}_c \times \tilde{P}_{c+1}$*, such that* $j \succ i$*.*

The first alternative implies that the subset $\tilde{P}_c$ should be split, which will be discovered by recommending permutations compatible with either $\tilde{\boldsymbol{P}}$ or one of its neighbors. The second alternative implies that j should be in a subset ranked before the subset containing i, which will be discovered by recommending the permutation in the neighborhood of $\tilde{\boldsymbol{P}}$ which puts i and j in the same subset.

5.1. Discussion

We gather here some remarks regarding the optimality of the theoretical results and their extension to a weak order.

Remark 5.1 (Optimality of UniRank's upper-bound). While deriving the exact lower-bound on the expected regret in this setting is out of the scope of our paper, we believe that this bound takes the form $\mathcal{O}\left(\sum_{k=2}^{K} \frac{\mu^* - \mu_k}{kl(\nu_k^*, \nu_k)} \log(T) + \sum_{k=K+1}^{L} \frac{\mu^* - \mu_k}{kl(\nu_k^*, \nu_k)} \log(T)\right)$, where for $k \in \{2, \dots, K\}$ (respectively $k \in \{K+1, \dots, L\}$), μ_k, ν_k^*, and ν_k result from the best partition and the best random variables to compare item k to item $k-1$ (resp. to item K).

In (Combes et al., 2015) (Propositions 1 and 2) and in (Lagrée et al., 2016) (Theorem 6) a bound with only the second sum is proven. The first sum is missing as both papers consider more restricting settings where the comparison between items $k \in \{2, \dots, K\}$ and $k-1$ is free in terms of regret: CM for (Combes et al., 2015), and PBM with $\boldsymbol{\kappa}$ known for (Lagrée et al., 2016).

We also believe that TopRank's upper-bound on the regret with our additional hypothesis either remains $\mathcal{O}(KL/\Delta \log(T))$ or reduces to $\mathcal{O}(L \log L/\Delta \log(T))$.[2] Indeed, while UniRank reduces the exploration by only comparing each item k to the item $\min(k-1, L)$, in the worst case scenario TopRank compares each item k to each item $k' \in \{1, \dots, \min(k-1, L)\}$ in order to conclude that k' should not be at one of the top-$\min(k-1, L)$ positions.

Remark 5.2 (Exploration not at the top). With CM model, exploring at the top reduces the regret: it leads to less exploration, while the instantaneous regret remains unchanged. However, this results does not hold with PBM (see Theorem 6 in (Lagrée et al., 2016) for details): while exploring at the top decreases the number of explorations, it also increases the regret per exploration; and the best trade-off depends on the values $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$.

Therefore, as in (Lagrée et al., 2016), we only explore through local changes in the recommendation. Note that these local changes are also more "user-friendly": as soon as the right leader has been identified, a sub-optimal item is always tried at the bottom of the recommendation, which is less surprising for users than a sub-optimal item displayed as the top recommendation.

Remark 5.3 (Upper-bound on the regret of UniRank assuming a weak order on items). If the order on the best items is not total, the proof of Theorem 5.1 may be adapted to get a $\mathcal{O}\left(LK/\Delta \log T\right)$ bound. Indeed, under the strict weak order assumption, there exists a set of optimal partitions, and therefore, any permutation compatible with a neighbor of any of these partitions may be recommended $\mathcal{O}\left(1/\Delta \log T\right)$ times. In the worst case scenario, K items are equivalent and strictly more attractive than the $L-K$ remaining items, and the set of the permutations compatible with a neighbor partition is composed of $K(L-K)$ permutations, which translates into a $\mathcal{O}\left(LK/\Delta \log T\right)$ regret bound. Note that (Lattimore et al., 2018) proves a $\Omega\left(LK/\Delta \log T\right)$ lower-bound on the regret assuming that the best items have the same attractiveness which means that the upper-bound of UniRank for this specific setting is optimal.

6. Experiments

In this section, we compare UniRank to TopRank (Lattimore et al., 2018), PB-MHB (Gauthier et al., 2021a), GRAB (Gauthier et al., 2021b), and CascadeKL-UCB (Kveton et al., 2015a). The experiments are conducted on the KDD Cup 2012 track 2 dataset, on the Yandex dataset (Yandex, 2013), and on a model with artificial parameters. We use the cumulative regret to evaluate the performance of each algorithm.

6.1. Experimental Settings

In order to evaluate our algorithm, we design six experiments inspired by the ones conducted in (Lattimore et al., 2018). The standard metric used is the expected cumulative regret (see Equation (1)), denoted as *regret*, which is the sum, over T consecutive recommendations, of the difference between the expected reward of the best answer and of the answer of a given ORS. The best algorithm is the one with the lowest regret. We use two click models for our experiments: the Position Based Model (PBM) and the Cascading Model (CM). To play according to those models,

[2] This second bound is proven in (Sentenac et al., 2021) for a matching problem handled with an algorithm similar to TopRank.

we extract the parameters of the chosen model from the KDD Cup 2012 track 2 (KDD for short) database and the Yandex database (Yandex, 2013), and we experiment with a set of parameters (denoted $\texttt{Simul}$) chosen to highlight the $\mathcal{O}\left(L/\Delta \log T\right)$ regret of UniRank: $L = 10$, $K = 5$, $\boldsymbol{\theta} = [0.1, 0.08, 0.06, 0.04, 0.02, 10^{-4}, 10^{-4}, 10^{-4}, 10^{-4}, 10^{-4}]$, and $\boldsymbol{\kappa} = [1, 0.9, 0.83, 0.78, 0.75]$.

Yandex database comes from fully anonymized real-life logs of actions toward the Yandex search engine. It contains 703 million items displayed among 65 million search queries and sharing 167 million hits (clicks). We consider the 10 most frequent queries in our experiments. We use the GPL3 Pyclick library (Chuklin et al., 2015) to infer the CM and PBM parameters of each query with the *expectation maximization* algorithm. Depending on the query, this leads to θ_i values ranging from 0.51 to 0.94, and κ_i values ranging from 0.71 to 1.00 when considering PBM and θ_i values ranging from 0.03 to 0.50 for CM.

We also extract parameters from the KDD dataset. Due to the type of data contained in this dataset, we can only extract parameters for the PBM model. This dataset consists of session logs of *soso.com*, a Tencent's search engine. It tracks clicks and displays of advertisements on a search engine result web-page, w.r.t. the user query. For each query, 3 positions are available for a various number of ads to display. Each of the 150M lines contains information about the search (UserId, QueryId…) and the ads displayed (AdId, Position, Click, Impression). We are looking for the best ads per query, namely the ones with a higher probability to be clicked. To follow previous works, instead of looking for the probability to be clicked per display, we target the probability to be clicked per session. This amounts to discarding the information *Impression*. We also filter the logs to restrict the analysis to (query, ad) couples with enough information: for each query, ads are excluded if they were displayed less than 1,000 times at any of the 3 possible positions. Then, we filter queries that have less than 5 ads satisfying the previous condition. We end up with 8 queries and from 5 to 11 ads per query. The overall process leads to θ_i values ranging from 0.004 to 0.149, and κ_k values ranging from 0.10 to 1.00, depending on the query.

Then we simulate the users' interactions given these parameters as it is commonly done in bandits settings. Similarly to (Lattimore et al., 2018), we look at the results averaged on the queries, while displaying K items among the L most attractive ones selected among all items possible for each query. With Yandex dataset, $K = 5$ and $L = 10$, while with KDD dataset $K = 3$ and L varies from 5 to 11. We run our experiments on an internal cluster to compute 20 independent sets of 10^7 consecutive recommendations for each of the 10 most frequent Yandex queries and each of the 8 KDD queries. It leads respectively to 200 games per setting and algorithm for Yandex and 160 games for KDD. As TopRank requires the knowledge of the horizon T, we test the impact of this parameter by setting it to the right value (10^7), to a too high value (10^{12}), and to a too small value (10^5) with doubling trick. To tune PB-MHB, we use the values recommended by (Gauthier et al., 2021a) for these datasets.

Table 2. Average computation time (in ms) per recommendation. For each top 10 query of Yandex dataset, 20 runs are performed assuming CM model and $L = 10$.

Algorithm	Computation Time (ms)
UniRank	1.0 ± 0.2
TopRank	0.7 ± 0.3
PB-MHB	13.9 ± 4.9
GRAB	0.9 ± 0.3
CascadeKL-UCB	0.9 ± 0.0

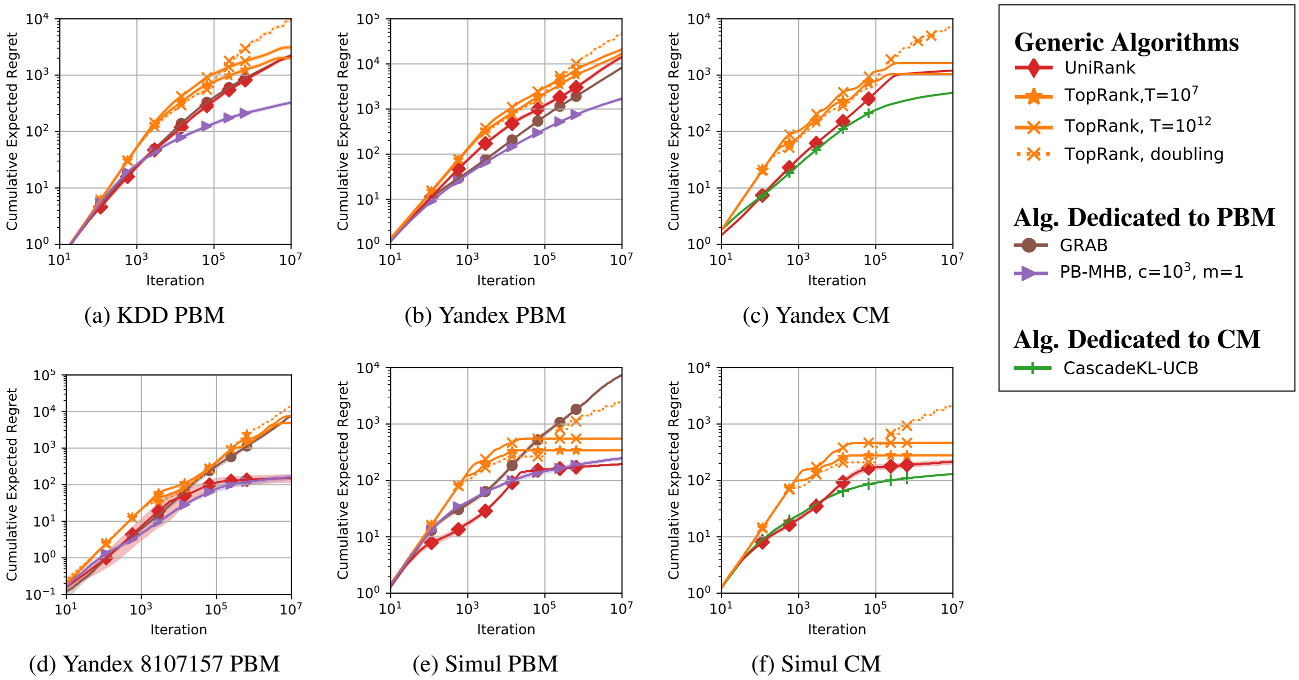
6.2. Results

Our results are shown in Figure 2. As expected, CascadeKL-UCB (respectively PB-MHB) outperforms other algorithms in the CM (resp. PBM) model for which it is designed. However, PB-MHB is computationally expensive (see Table 2) and lacks a theoretical analysis. Surprisingly, although GRAB is designed for PBM model, it suffers a high regret when confronted to the query 8107157 of Yandex and to Simul with PBM model.

Secondly, UniRank and TopRank enjoy a logarithmic regret in all settings and our algorithm UniRank outperforms TopRank for the models such that the K best items do not have the same attractiveness θ_i: query 8107157 of Yandex, simul PBM, and simul CM. When confronted to other models, UniRank has a regret strictly smaller than TopRank before the iteration $t = 10^6$, and smaller or equal to TopRank at the horizon. Moreover, as already explained, TopRank is aware of the horizon T and may stop (over)exploring early, as can be observed in the CM model after iteration 10^5. If TopRank targets a horizon $T = 10^{12}$ or uses the doubling trick it suffers a higher regret than UniRank.

Regarding the computational complexity, as shown in Table 2, PB-MHB is significantly slower with a computation time per recommendation ten times higher than any other algorithm. These other algorithms have a similar computation time of approximately 1 ms per recommendation.

Overall, as TopRank, UniRank is consistent over all settings, and require a reasonable computation time. Moreover, contrary to TopRank, (i) UniRank drastically decreases its regret by taking advantage of the differences of attractiveness between items, and (ii) UniRank does not require the



(a) KDD PBM

(b) Yandex PBM

(c) Yandex CM

(d) Yandex 8107157 PBM

(e) Simul PBM

(f) Simul CM

Figure 2. Cumulative regret w.r.t. iterations. $K = 5$ and $L = 10$ for Yandex and Simul models (b,c,e,f); $K = 3$ and $L \in \{5, \ldots, 11\}$ for KDD model (a); $K = 5$ and $L = 6$ for Yandex 8107157 (d) which corresponds to the parameters of the query 8107157 of Yandex. The plotted curves correspond to the average over 200, 160, or 20 independent sequences of recommendations (20 sequences per query). The (small) shaded areas depict the standard error of our regret estimates.

knowledge of the horizon T.

7. Conclusion

We have presented UniRank, a unimodal bandit algorithm for online ranking. The regret bound in $\mathcal{O}\left(L/\Delta \log T\right)$ of our algorithm, is a direct consequence of the unimodality-like property of the bandit setting with respect to a graph where nodes are ordered partitions of items. Even though the proof is inspired by OSUB (Combes & Proutière, 2014), the fact that UniRank handles partitions instead of recommendations, uses different estimators and builds upon an unusual exploration-exploitation strategy makes it original, and we believe that our theoretical analysis opens new perspectives for other semi-bandit settings. Experiments against state-of-the-art learning algorithms show that our method is consistent in all settings, enjoys a smaller regret than TopRank and GRAB on specific settings, and has a much smaller computation time than PB-MHB.

While in industrial applications, contextual information is also used to build recommendations (Li et al., 2019b; Chen et al., 2019; Ermis et al., 2020; Gampa & Fujita, 2021), in this paper we restricted ourselves to independent arms to simplify the presentation of the approach. However, the integration of unimodal bandit algorithms working on parametric spaces (Combes et al., 2020) should bridge the gap between both approaches.

Ethical Statement

Regarding the societal impact of the proposed approach, it is worth mentioning that the approach aims at identifying and recommending the most popular items. Therefore, the approach may increase the monopoly effects: the most attractive items are displayed more often, so their reputation increases, and then they may become even more attractive… However bandit algorithms continuously explore and therefore continuously offer an opportunity to less popular items to increase their reputation.

Acknowledgements

We thank the reviewers for their valuable comments towards clarification of the paper. This research was partially supported by the Inria Project Lab “Hybrid Approaches for Interpretable AI” (HyAIAI) and the network on the foundations of trustworthy AI, integrating learning, optimisation, and reasoning (TAILOR) financed by the EU’s Horizon 2020 research and innovation program under agreement 952215.

References

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. H. Top-k off-policy correction for a reinforce recommender system. In *Proc. of the 12th ACM Int. Conf. on Web Search and Data Mining*, WSDM '19, pp. 456–464, 2019.

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

Cheung, W. C., Tan, V., and Zhong, Z. A thompson sampling algorithm for cascading bandits. In *proc. of the 22nd Int. Conf. on Artificial Intelligence and Statistics*, AISTATS'19, 2019.

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

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

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

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

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

Ermis, B., Ernst, P., Stein, Y., and Zappella, G. Learning to rank in the position based model with bandit feedback. In *Proc. of the 29th ACM Int. Conf. on Information & Knowledge Management*, CIKM'20, pp. 2405–2412, 2020.

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

Gampa, P. and Fujita, S. Banditrank: Learning to rank using contextual bandits. In Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R. K., Reddy, P. K., Srivastava, J., and Chakraborty, T. (eds.), *Advances in Knowledge Discovery and Data Mining*, PAKDD'21, pp. 259–271. Springer International Publishing, 2021.

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

Gauthier, C.-S., Gaudel, R., and Fromont, E. Bandit algorithm for both unknown best position and best item display on web pages. In *19th International Symposium on Intelligent Data Analysis, Apr 2021, Porto (virtual), Portugal. pp.1-12*, IDA, 2021a.

Gauthier, C.-S., Gaudel, R., Fromont, E., and Lompo, B. A. Parametric graph for unimodal ranking bandit. In *Proc. of the 38th Int. Conf. on Machine Learning*, ICML'21, pp. 3630–3639, 2021b.

Katariya, S., Kveton, B., Szepesvári, C., and Wen, Z. DCM bandits: Learning to rank with multiple clicks. In *proc. of the 33rd Int. Conf. on Machine Learning*, ICML'16, 2016.

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

Komiyama, J., Honda, J., and Takeda, A. Position-based multiple-play bandit problem with unknown position bias. In *proc. of the 31st conf. on Neural Information Processing Systems*, NeurIPS'17, 2017.

Kveton, B., Szepesvári, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *proc. of the 32nd Int. Conf. on Machine Learning*, ICML'15, pp. 767–776, 2015a.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. Combinatorial cascading bandits. In *proc. of the 29th conf. on Neural Information Processing Systems*, NeurIPS'15, 2015b.

Lagrée, P., Vernade, C., and Cappé, O. Multiple-play bandits in the position-based model. In *proc. of the 30th conf. on Neural Information Processing Systems*, NeurIPS'16, 2016.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. Toprank: A practical algorithm for online stochastic ranking. In *proc. of the 32nd conf. on Neural Information Processing Systems*, NeurIPS'18, 2018.

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

Li, S., Wang, B., Zhang, S., and Chen, W. Contextual combinatorial cascading bandits. In *proc. of the 33rd Int. Conf. on Machine Learning*, ICML'16, 2016.

Li, S., Lattimore, T., and Szepesvári, C. Online learning to rank with features. In *Proc. of the 36th Int. Conf. on Machine Learning*, ICML'19, 2019b.

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

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

Sentenac, F., Yi, J., Calauzenes, C., Perchet, V., and Vojnovic, M. Pure exploration and regret minimization in matching bandits. In *Proc. of the 38th Int. Conf. on Machine Learning*, ICML'21, pp. 9434–9442, 2021.

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

Zoghi, M., Tunys, T., Ghavamzadeh, M., Kveton, B., Szepesvari, C., and Wen, Z. Online learning to rank in stochastic click models. In *proc. of the 34th Int. Conf. on Machine Learning*, ICML'17, 2017.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. In *proc. of the 32nd Conference on Uncertainty in Artificial Intelligence*, UAI '16, 2016.

A. Organisation of the Appendix

The appendix is organized as follows. After listing most of the notations used in the paper in Appendix B, we prove Lemma 3.1 in Appendix D. Then we prove some technical lemmas in Appendix E, which are required by the proof of Theorem 5.1 in Appendix F. Finally, we discuss the regret upper-bound of UniRank for some specific settings in Appendix G.

B. Notations

Table 4 summarizes the notations used throughout the paper and the appendix. Below are additional notations necessary for the proofs.

Definition B.1 (Specific notations to count events and observations)**.** The proofs are based on the concentration of the statistic $\hat{s}_{i,j}(t)$ which is the average over $T_{i,j}(t)$ observations. The number $T_{i,j}(t)$ itself is a sum: the sum of the random variables $\mathbb{1}\{c_i(s) \neq c_j(s)\} \mid \exists c, (i,j) \in P_c(s)^2$, where s is in $[t]$. To discuss the concentration of this sum, for any iteration t in $[T]$, we denote $t_{i,j}(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\}$ the number of iterations at which the random variable is observed.

Definition B.2 (Recommended subset)**.** Let (L, K, ρ) be an online learning to rank problem, $\boldsymbol{P}$ be an ordered partition of $[L]$ in d subsets, and $c \in [d]$ the index of one of these subsets. The subset P_c is *recommended* (denoted $\mathrm{Rec}(P_c)$) if the recommendations compatible with $\boldsymbol{P}$ include some items from P_c. More specifically, the subset P_c is *recommended* if $|\bigcup_{\ell \in [c-1]} P_\ell| < K$.

Definition B.3 (Expectations on clicks)**.** let i and j be two different items.

We denote

$$\tilde{\delta}_{i,j} := \min_{\boldsymbol{P}:\exists c,(i,j)\in P_c^2 \wedge \mathrm{Rec}(P_c)} \mathbb{P}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) \neq c_j(t)\right]$$

the smallest probability for $c_i(t)$ to be different from $c_j(t)$ while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$ (and may potentially be clicked upon). If we assume $1 \succ 2 \succ \cdots \succ L$, we also denote

$$\tilde{\delta}^*_i := \min_{\boldsymbol{P}\in\mathcal{N}((\{1\},\dots,\{K\},\{K+1,\dots,L\})):\exists c,(\min(i-1,K),i)\in P_c^2} \mathbb{P}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_{\min(i-1,K)}(t) \neq c_i(t)\right]$$

the smallest probability for $c_{\min(i-1,K)}(t)$ to be different from $c_i(t)$ while both items $\min(i-1,K)$ and i are in the same subset of the chosen partition $\boldsymbol{P}(t)$ (and may potentially be clicked upon), and $\boldsymbol{P}(t)$ is in the neighborhood of the optimal partition $\boldsymbol{P}^* = (\{1\},\dots,\{K\},\{K+1,\dots,L\})$.

If $i \succ j$, we denote

$$\tilde{\Delta}_{i,j} := \min_{\boldsymbol{P}:\exists c,(i,j)\in P_c^2 \wedge \mathrm{Rec}(P_c)} \mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) - c_j(t) \mid c_i(t) \neq c_j(t)\right] = \min_{\boldsymbol{a}\in\mathcal{P}_K^L:\{i,j\}\cap\boldsymbol{a}([K])\neq\varnothing} \tilde{\Delta}_{i,j}(\boldsymbol{a}),$$

the smallest expected difference of clicks between items i and j while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$ (and may potentially be clicked upon).

Symmetrically, if $j \succ i$, we denote

$$\tilde{\Delta}_{i,j} := \max_{\boldsymbol{P}:\exists c,(i,j)\in P_c^2 \wedge \mathrm{Rec}(P_c)} \mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) - c_j(t) \mid c_i(t) \neq c_j(t)\right] = \max_{\boldsymbol{a}\in\mathcal{P}_K^L:\{i,j\}\cap\boldsymbol{a}([K])\neq\varnothing} \tilde{\Delta}_{i,j}(\boldsymbol{a}),$$

the greatest expected difference of clicks between items i and j while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$ (and may potentially be clicked upon).

Lemma E.1 in Appendix E.1 ensures the proper definition of these notations under Assumptions 3.1, 3.2, and 3.3, and states that $\tilde{\delta}_{i,j} = \tilde{\delta}_{j,i} > 0$ and $\tilde{\Delta}_{i,j} = -\tilde{\Delta}_{j,i} > 0$ if $i \succ j$.

Definition B.4 (Reward gap)**.** Let (L, K, ρ) be an OLR problem satisfying Assumption 3.2 and such that the order on items is a total order. Without loss of generality, let us assume that $1 \succ 2 \succ \cdots \succ L$.. Denoting $\boldsymbol{P}^* = (\{1\},\dots,\{K\},\{K+1,\dots,L\})$ the optimal partition associated to this order and taking $c \geqslant 2$, the *reward gap* of item c is

$$\begin{aligned}\Delta_c := {} & \rho\left(\boldsymbol{a}^*, \min(c-1,K)\right) + \rho\left(\boldsymbol{a}^*, c\right) \\ & - \rho\left((\min(c-1,K), c)\circ\boldsymbol{a}^*, \min(c-1,K)\right) - \rho\left((\min(c-1,K), c)\circ\boldsymbol{a}^*, c\right)\end{aligned}$$

Symbol	Meaning
T	Time horizon
t	iteration
L	number of items
i	index of an item
K	number of positions in a recommendation
k	index of a position
$[n]$	set of integers $\{1, \dots, n\}$
$\mathcal{P}_K^L$	set of permutations of K distinct items among L
$\boldsymbol{\theta}$	vectors of probabilities of click
θ_i	probability of click on item i
$\boldsymbol{\kappa}$	vectors of probabilities of view
κ_k	probability of view at position k
$\mathcal{A}$	set of bandit arms
$\boldsymbol{a}$	an arm in $\mathcal{A}$
$\boldsymbol{a}(t)$	the arm chosen at iteration t
a_k	item displayed at position K in the recommendation $\boldsymbol{a}$
$\boldsymbol{a}^*$	best arm
ρ	function from $\mathcal{P}_K^L \times [K]$ to $[0,1]$ giving the probability of click
$\rho(\boldsymbol{a}, k)$	probability of click on the item displayed at position k when recommending $\boldsymbol{a}$
$\boldsymbol{c}(t)$	clicks vector at iteration t
$c_i(t)$	clicks on item i at iteration t
$r(t)$	reward collected at iteration t, $r(t) = \sum_{i=1}^{L} c_i(t)$
$\mu_{\boldsymbol{a}}$	expectation of $r(t)$ while recommending $\boldsymbol{a}$, $\mu_{\boldsymbol{a}} = \mathbb{E}[r(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}]$
μ^*	highest expected reward, $\mu^* = \max_{\boldsymbol{a} \in \mathcal{P}_K^L} \mu_{\boldsymbol{a}}$
Δ	generic reward gap between one of the sub-optimal arms and one of the best arms
Δ_c	reward gap while exchanging items $\min(c-1, K)$ and c in the optimal recommendation,
$\tilde{\delta}_{i,j}$	smallest probability for $c_i(t)$ to be different from $c_j(t)$ while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$
$\tilde{\delta}_k^*$	smallest probability for $c_{\min(k-1,K)}(t)$ to be different from c_k, while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$ and $\boldsymbol{P}(t)$ is in the neighborhood of the optimal partition
$\tilde{\Delta}_{i,j}$	smallest (respectively highest) expected difference of click between items i and j if $i \succ j$ (resp. $j \succ i$) while both items are in the same subset of the chosen partition $\boldsymbol{P}(t)$
$R(T)$	cumulative (pseudo-)regret, $R(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{\boldsymbol{a}(t)}\right]$
$\succ$	strict weak order
$(i,j) \circ \boldsymbol{a}$	permutation swapping items i and j in recommendation $\boldsymbol{a}$
$\boldsymbol{P}$	ordered partition of items representing a subset of recommendations, $\boldsymbol{P} = (P_1, \dots, P_d)$
P_c	c^{th} part of $\boldsymbol{P}$ such as $\bigcup_{c=1}^{d} P_c = [L]$, and $P_c \cap P_{c'}$ is empty when $c \neq c'$
$\mathcal{A}(\boldsymbol{P})$	set of recommendations $\boldsymbol{a}$ agreeing with $\boldsymbol{P}$
$\tilde{\boldsymbol{P}}(t)$	best partition at iteration t given the previous choices and feedbacks (called leader)
$\boldsymbol{P}^*$	partition such that any permutation $\boldsymbol{a}$ in $\mathcal{A}(\boldsymbol{P}^*)$ is compatible with the strict weak order on items.
$\mathcal{G}$	graph carrying a partial order on the partitions of items
$\mathcal{N}(\tilde{\boldsymbol{P}})$	Neighborhood in $\mathcal{G}$ of the partition $\boldsymbol{P}$, $\mathcal{N}(\tilde{\boldsymbol{P}}) := \left\{ \left(\tilde{P}_1(t), \dots, \tilde{P}_{c-1}(t), \tilde{P}_c(t) \cup \tilde{P}_{c+1}(t), \tilde{P}_{c+2}(t), \dots \tilde{P}_{\tilde{d}}(t) \right) : c \in [\tilde{d} - 2] \right\}$ $\cup \left\{ \left(\tilde{P}_1(t), \dots, \tilde{P}_{\tilde{d}-1}(t) \cup \{j\}, \tilde{P}_{\tilde{d}-1}(t) \setminus \{j\}, \tilde{P}_{\tilde{d}}(t) \right) : j \in \tilde{P}_{\tilde{d}}(t) \right\}.$
$t_{i,j}(t)$	number of iterations at which items i and j have been gathered in the same subset of items $P_c(s)$, $t_{i,j}(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\}$
$T_{i,j}(t)$	number of iterations at which items i and j have been gathered in the same subset of items $P_c(s)$ and lead to a different click value, $T_{i,j}(t) = \sum_{s=1}^{t-1} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\} \mathbb{1}\{c_i(s) \neq c_j(s)\}$
$\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$	number of time a permutation $\tilde{\boldsymbol{P}}$ as been the leader, $\tilde{t}_{\tilde{\boldsymbol{P}}}(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{P}}(s) = \tilde{\boldsymbol{P}}\right\}$
$\tilde{\delta}_{i,j}(\boldsymbol{a})$	probability of difference, $\tilde{\delta}_{i,j}(\boldsymbol{a}) = \mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a}, (i,j) \circ \boldsymbol{a}\})}\left[c_i \neq c_j\right]$
$\tilde{\Delta}_{i,j}(\boldsymbol{a})$	expected click difference, $\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a}, (i,j) \circ \boldsymbol{a}\})}\left[c_i - c_j \mid c_i \neq c_j\right]$
$\hat{s}_{i,j}(t)$	UniRank's main statistic to infer that $i \succ j$, $\hat{s}_{i,j}(t) := \frac{1}{T_{i,j}(t)} \sum_{s=1}^{t-1} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\} (c_i(s) - c_j(s))$
$\bar{\bar{s}}_{j,i}(t)$	Kullback-Leibler based optimistic estimator, $\bar{\bar{s}}_{j,i}(t) := 2 * f\left(\frac{1 + \hat{s}_{i,j}(t)}{2}, T_{i,j}(t), \tilde{t}_{\tilde{\boldsymbol{P}}}(t)\right) - 1$
f	Kullback-Leibler index function, $f(\hat{\mu}, T, t) := \inf\{\mu \in [0, \hat{\mu}] : T \times \mathrm{kl}(\hat{\mu}, \mu) \leq \log(t) + 3\log(\log(t))\}$,
$\mathrm{kl}(p,q)$	Kullback-Leibler divergence from a Bernoulli distribution of mean p to a Bernoulli distribution of mean q, $\mathrm{kl}(p,q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right)$
$\mathcal{U}(S)$	uniform distribution on the set S
c	(in PB-MHB) parameter controlling size of the step in the Metropolis Hasting inference

Table 4. Summary of the notations.

Algorithm 2 Elicitation of the leader partition $\tilde{\boldsymbol{P}}(t)$

```
Require: number of items $L$, number of positions $K$, iteration index $t$, statistics $\hat{s}_{i,j}(t)$
$\tilde{d} \leftarrow 1$; $R \leftarrow [L]$; $n \leftarrow L$
repeat
    for each $i \in R$, $S_i \leftarrow |\{j \in R : \hat{s}_{i,j}(t) > 0\}|$
    sort items in $R$ by $S_i$: $S_{i_1} > S_{i_2} > \cdots > S_{i_n}$
    $\ell \leftarrow \min\{\ell \in [n] : \forall k < \ell, \forall k' \geqslant \ell : \hat{s}_{i_k,i_{k'}}(t) > 0\}$
    $B \leftarrow \{i_1, \dots, i_{\ell-1}\}$; $\tilde{P}_{\tilde{d}}(t) \leftarrow B$
    $\tilde{d} \leftarrow \tilde{d} + 1$; $R \leftarrow R \setminus B$; $n \leftarrow |R|$
until $\left|\bigcup_{\tilde{c}=1}^{\tilde{d}} \tilde{P}_{\tilde{c}}(t)\right| \geqslant K$
$\tilde{d} \leftarrow \tilde{d} + 1$ ; $\tilde{P}_{\tilde{d}}(t) \leftarrow R$
return $\tilde{\boldsymbol{P}}(t)$
```

Note that for $c \leqslant K$, $\Delta_c := \rho(\boldsymbol{a}^*, c-1) + \rho(\boldsymbol{a}^*, c) - \rho((c-1,c) \circ \boldsymbol{a}^*, c-1) - \rho((c-1,c) \circ \boldsymbol{a}^*, c)$, and for $c \geqslant K+1$, $\Delta_c = \rho(\boldsymbol{a}^*, K) - \rho((K,c) \circ \boldsymbol{a}^*, K)$.

C. Algorithm for the Elicitation of the leader partition $\tilde{\boldsymbol{P}}(t)$

D. Proof of Lemma 3.1 (PBM and CM Fulfills Assumptions 3.1, 3.2, and 3.3)

For both CM and PBM click models, we note θ_i the click probability of item i. For PBM we have κ_k the probability that a user see the position k.

Proof. Let us begin with some preliminary remarks.

First, with PBM model, the positions are ranked by decreasing observation probability, meaning that $\kappa_{a_1} \geqslant \kappa_{a_2} \geqslant \cdots \geqslant \kappa_{a_K}$.

Secondly, by definition, $\rho(k, \boldsymbol{a}) > 0$ for any position k and recommendation $\boldsymbol{a}$, which implies that:

- $\min_i \theta_i > 0$ and $\max_i \theta_i < 1$ in CM model;
- $\kappa_K > 0$ in PBM model.

Let us now prove that Assumptions 3.1, 3.2 and 3.3 are fulfilled by PBM and CM click models with the strict weak order $\succ$ defined by $i \succ j \iff \theta_i > \theta_j$.

By definition of $\succ$, Assumption 3.1 is fulfilled taking the the preferential attachment function $g : i \mapsto \theta_i$, and Assumption 3.1* is fulfilled as soon as $\theta_i \neq \theta_j$ for any item i in top-K items and any item $j \neq i$.

For Assumption 3.2, we have to prove that having $\boldsymbol{a}$ compatible with $\succ$ is optimal, meaning $\mu_{\boldsymbol{a}} = \mu^*$.

Let $\boldsymbol{a}$ be a permutation compatible with $\succ$.

In the case of CM, $\mu_{\boldsymbol{a}} = 1 - \sum_{k=1}^{K}(1 - \theta_{a_k})$. In order to maximize $\mu_{\boldsymbol{a}}$, one has to select the K higher values of $\boldsymbol{\theta}$. As $\boldsymbol{a}$ is compatible with $\succ$, which is defined based on values θ_i, it satisfies this property. Hence, CM fulfills Assumption 3.2.

For PBM, $\mu_{\boldsymbol{a}} = \sum_{k=1}^{K} \theta_{a_k} \kappa_k$. As the series $(\kappa_k)_{k \in [K]}$ is non-increasing, $\mu_{\boldsymbol{a}}$ is maximized if $(\theta_k)_{k \in [K]}$ is also non-increasing and if $\theta_K \geqslant \max_{k \geqslant K+1} \theta_k$. These properties are ensured by the fact that $\boldsymbol{a}$ is compatible with $\succ$ and that $\succ$ is defined based on values θ_i. Hence, PBM fulfills Assumption 3.2.

We now prove that CM and PBM fulfill Assumption 3.3. Let i and j be two distinct items such that $i \succ j$ and $\boldsymbol{a} \in \mathcal{P}_K^L$ be a recommendation such that at least one of both items is displayed.

First, $\mathbb{E}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a}, (i,j) \circ \boldsymbol{a}\})}\left[c_i(t) \neq c_j(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}'\right]$ is non-null with PBM model as $c_i(t)$ and $c_j(t)$ are independent and as at least one of the four variables $c_i(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}$, $c_i(t) \mid \boldsymbol{a}(t) = (i,j) \circ \boldsymbol{a}$, $c_j(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}$, $c_j(t) \mid \boldsymbol{a}(t) = (i,j) \circ \boldsymbol{a}$ has an expectation which is non-zero and strictly smaller than 1 (due to $\kappa_K > 0$ and $\theta_i > \theta_j$).

Similarly, $\mathbb{E}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}\left[c_i(t) \neq c_j(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}'\right]$ is non-null with CM model as at most one of both items can be clicked at each iteration and the shown item has non-zero probability to be clicked (by definition of ρ).

Then, we consider $\tilde{\Delta}_{i,j}(\boldsymbol{a})$ as

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}(c_i = 1, c_j = 0) - \mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}(c_i = 0, c_j = 1)}{\mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}(c_i = 1, c_j = 0) + \mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}(c_i = 0, c_j = 1)}$$

We want to control the sign of $\tilde{\Delta}_{i,j}(\boldsymbol{a})$, which is also the sign of its numerator, as its denominator (noted $D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}$) is non-negative.

The recommendation $\boldsymbol{a}'$ is drawn uniformly in $\{\boldsymbol{a}, (i,j) \circ \boldsymbol{a}\}$ thus

$$\mathbb{P}_{\boldsymbol{a}' \sim \mathcal{U}(\{\boldsymbol{a},(i,j)\circ\boldsymbol{a}\})}(c_i = 1, c_j = 0) = \frac{1}{2}\mathbb{P}_{\boldsymbol{a}}(c_i = 1, c_j = 0) + \frac{1}{2}\mathbb{P}_{(i,j)\circ\boldsymbol{a}}(c_i = 1, c_j = 0).$$

When considering a CM click model, we have $\mathbb{P}_{\boldsymbol{a}}(c_i = 1, c_j = 0) = \prod_{p=1}^{k-1}(1 - \theta_{a_p})\theta_i$ and $\mathbb{P}_{\boldsymbol{a}}(c_i = 0, c_j = 1) = \prod_{p=1}^{l-1}(1 - \theta_{a_p})\theta_j$ when i and j $\in \boldsymbol{a}$.

In that case, we have:

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}\prod_{p=1}^{k-1}(1-\theta_{a_p})\theta_i + \frac{1}{2}\prod_{p=1}^{l-1}(1-\theta_{a_p})\theta_i - \left(\frac{1}{2}\prod_{p=1}^{l-1}(1-\theta_{a_p})\theta_j + \frac{1}{2}\prod_{p=1}^{k-1}(1-\theta_{a_p})\theta_j\right)}{D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}}$$

which can be simplified in:

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}\left(\prod_{p=1}^{k-1}(1-\theta_{a_p}) + \prod_{p=1}^{l-1}(1-\theta_{a_p})\right)(\theta_i - \theta_j)}{D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}}.$$

Since $\max_i \theta_i < 1$, $\prod_{p=1}^{k-1}(1-\theta_{a_p}) + \prod_{p=1}^{l-1}(1-\theta_{a_p}) > 0$, thus the sign of $\tilde{\Delta}_{i,j}(\boldsymbol{a})$ is the sign of $(\theta_i - \theta_j)$ and $\tilde{\Delta}_{i,j}(\boldsymbol{a}) > 0 \iff \theta_i > \theta_j \iff i \succ j$.

Now if $i \notin \boldsymbol{a}$ then $\mathbb{P}_{\boldsymbol{a}}(c_i = 1, c_j = 0) = 0$ as the position is not seen. We have:

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}(\prod_{p=1}^{l-1}(1-\theta_{a_p}))(\theta_i - \theta_j)}{D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}}$$

which leads to the same conclusion as the previous case. By symmetry, we have the same conclusion with j $\notin \boldsymbol{a}$.

Now with a PBM click model, we have $\mathbb{P}_{\boldsymbol{a}}(c_i = 1, c_j = 0) = \kappa_k\theta_i(1 - \kappa_l\theta_j)$ as $c_i = 1$ and $c_j = 0$ are independant events.

Thus, we have:

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}\kappa_k\theta_i(1-\kappa_l\theta_j) + \frac{1}{2}\kappa_l\theta_i(1-\kappa_k\theta_j) - \left(\frac{1}{2}\kappa_l\theta_j(1-\kappa_k\theta_i) + \frac{1}{2}\kappa_k\theta_j(1-\kappa_l\theta_i)\right)}{D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}}$$

which can be simplified in:

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}(\kappa_k + \kappa_l)(\theta_i - \theta_j)}{D_{\tilde{\Delta}_{i,j}(\boldsymbol{a})}}$$

As κ_k or κ_l is positive if i or j is presented, similarly to the CM case we have $\tilde{\Delta}_{i,j}(\boldsymbol{a}) > 0 \iff \theta_i > \theta_j \iff i \succ j$.

This proof can be extended to i or $j \notin \boldsymbol{a}$ by taking $\kappa_k = 0$ when $k > K$.

We can conclude that both CM and PBM fulfills Assumption 3.3. □

E. Technical Lemmas Required by the Proof of Theorem 5.1

In this section, we gather technical Lemmas required to prove the regret upper-bound of UniRank. These lemmas regard the concentration away from zero of the statistic $\hat{s}_{i,j}(t)$ (Appendices E.1 and E.2), and the sufficient optimism brought by $\bar{\bar{s}}_{j,i}(t)$ (Appendix E.3).

E.1. Minimum Expected Click Difference

Assumption 3.3 builds upon $\tilde{\Delta}_{i,j}(\boldsymbol{a})$ which measures the difference of attractiveness between i and j while all other items are at fixed positions. In the theoretical analysis of UniRank, we handle situations where other items may also change in position thanks to the following Lemma.

Lemma E.1 (Minimum expected click difference)**.** *Let (L, K, ρ) be an OLR problem satisfying Assumptions 3.2 and 3.3 with $\succ$ the order on items, and let i and j be two items such that $i \succ j$. Then, for any partition of items $\boldsymbol{P}$, if there exists c such that $(i, j) \in P_c^2$ and $\mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) \neq c_j(t)\right] \neq 0$, then $\mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) - c_j(t) \mid c_i(t) \neq c_j(t)\right] > 0$ and therefore*

$$\tilde{\delta}_{i,j} > 0 \qquad \text{and} \qquad \tilde{\Delta}_{i,j} > 0.$$

Symmetrically, if $j \succ i$, for any partition of items $\boldsymbol{P}$, if there exists c such that $(i, j) \in P_c^2$ and $\mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) \neq c_j(t)\right] \neq 0$, then $\mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) - c_j(t) \mid c_i(t) \neq c_j(t)\right] < 0$ and therefore

$$\tilde{\delta}_{i,j} > 0 \qquad \text{and} \qquad \tilde{\Delta}_{i,j} < 0.$$

Proof. The proof consists in writing $\mathbb{E}_{\boldsymbol{a}(t)\sim\mathcal{U}(\mathcal{A}(\boldsymbol{P}))}\left[c_i(t) \neq c_j(t)\right] \neq 0$ two times as a sum other $\boldsymbol{a}(t) \in \mathcal{U}(\mathcal{A}(\boldsymbol{P}))$, and in reindexing one of both sums by $(i, j) \circ \boldsymbol{a}(t) \in \mathcal{U}(\mathcal{A}(\boldsymbol{P}))$. Then, adding the terms of both sums we get a sum of terms $\tilde{\Delta}_{i,j}(\boldsymbol{a})$ which by assumption 3.3 are positive. Hence this sum is positive, which concludes the proof. □

E.2. Upper-bound on the Number of High Deviations for Variables with Lower-Bounded Mean

The Proof of Theorem 5.1 requires the control of the expected number of high deviations of the statistic $\hat{s}_{i,j}(t)$. We control this expectation through Lemma E.4 which derives from the application of Lemmas E.2 and E.3 to $\hat{s}_{i,j}(t)$ and $\hat{T}_{i,j}(t)$. Hereafter, we express and prove the three lemmas. Note that Lemmas E.2 and E.3 are extensions of Lemmas 4.3 and B.1 of (Combes & Proutière, 2014) to a setting where the handled statistic is a mixture of variables following different laws of bounded expectation.

Lemma E.2 (Concentration bound with lower-bounded mean)**.** *Let $(X_t^a)_{t\geqslant 1}$ with $a \in \mathcal{R}$, be $|\mathcal{R}| < \infty$ independent sequences of independent random variables bounded in $[0, B]$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_t$ be an increasing sequence of $\sigma-$fields of $\mathcal{F}$ such that for each t, $\sigma((X_1^a)_{a\in\mathcal{R}}, \dots, (X_t^a)_{a\in\mathcal{R}}) \subset \mathcal{F}_t$ and for $s > t$ and a a recommendation, X_s^a is independent from $\mathcal{F}_t$. Consider $|\mathcal{R}|$ previsible sequences $(\epsilon_t^a)_{t\geq 1}$ of Bernoulli variables (for all $t > 0$, ϵ_t^a is $\mathcal{F}_{t-1} - mesurable$) such that for all $t > 0$, $\sum_i \epsilon_t^a \in \{0, 1\}$. Let $\delta > 0$ and for every $t \in \{1, \dots, n\}$ let*

$$S(t) = \sum_{s=1}^{t} \sum_i \epsilon_s^i (X_s^i - \mathbb{E}[X_s^i]), \qquad T(t) = \sum_{s=1}^{t} \sum_i \epsilon_s^i, \qquad \hat{\mu}(t) = \frac{S(t)}{N(t)}.$$

Define $\phi \in \{t_0, \dots, T+1\}$ a $\mathcal{F}$-stopping time such that either $T(\phi) \geqslant s$ or $\phi = T + 1$.

Then

$$\mathbb{P}\left(S(\phi) \geqslant T(\phi)\delta, \phi \leqslant T\right) \leqslant \exp(-\frac{2n\delta^2}{B^2}).$$

Proof. Let $\lambda > 0$, and define $G_t = \exp(\lambda(S(t) - \delta T(t)))\mathbb{1}\{t \leqslant T\}$. We have that:

$$\begin{aligned}\mathbb{P}(S(\phi) \geqslant T(\phi)\delta, \phi \leqslant T) &= \mathbb{P}(\exp(\lambda(S(\phi) - \delta T(\phi))\mathbb{1}\{\phi \leqslant T\} \geqslant 1)\\ &= \mathbb{P}(g_\phi \geqslant 1)\\ &\leqslant \mathbb{E}[G_\phi].\end{aligned}$$

Next we provide an upper bound for $\mathbb{E}[G_\phi]$. We define the following quantities:

$$Y_s^i = \varepsilon_s^i(\lambda(X_s^i - \mathbb{E}[X_s^i]) - \lambda^2 B^2/8)$$

$$\tilde{G}_t = \exp\left(\sum_{s=1}^{t}\sum_i Y_s^i\right)\mathbb{1}\{t \leqslant T\}.$$

Taking $\lambda = 4\delta/B^2$, G_t can be written:

$$G_t = \tilde{G}_t \exp(-T(t)(\lambda\delta - \lambda^2 B^2/8) = \tilde{G}_t \exp(-2T(t)\delta^2/B^2).$$

As $T(t) \geqslant n$ if $\phi \leqslant T$ we can upper bound G_ϕ by:

$$G_\phi = \tilde{G}_\phi \exp(-2T(\phi)\delta^2/B^2) \leqslant \tilde{G}_\phi \exp(-2n\delta^2/B^2).$$

It is noted that the above inequality holds even when $\phi = T + 1$, since $G_{T+1} = \tilde{G}_{T+1} = 0$. Hence:

$$\mathbb{E}[G_\phi] \leqslant \mathbb{E}[\tilde{G}_\phi]\exp(-2n\delta^2/B^2)$$

We prove that $\left(\tilde{G}_t\right)_t$ is a super-martingale. We have that $\mathbb{E}[\tilde{G}_{T+1} \mid \mathcal{F}_T] = 0 \leqslant \tilde{G}_T$. For $s \leqslant T - 1$, since B_{t+1} is $\mathcal{F}$ measurable:

$$\mathbb{E}[\tilde{G}_{t+1} \mid \mathcal{F}_t] = \tilde{G}_t((1 - \sum_i \varepsilon_{t+1}^i) + \sum_i \varepsilon_{t+1}^i \mathbb{E}[\exp(Y_{t+1}^i)]).$$

As proven in (Hoeffding, 1963)[eq. 4.16] since $X_{t+1}^i \in [0, B]$:

$$\mathbb{E}[\exp(\lambda(X_{t+1}^i - \mathbb{E}[X_{t+1}^i])] \leqslant \exp(\lambda^2 B^2/8),$$

so $\mathbb{E}[\exp(Y_{t+1}^i)] \leqslant 1$ and $\left(\tilde{G}_t\right)_t$ is a super-martingale: $\mathbb{E}[\tilde{G}_{t+1} \mid \mathcal{F}_t] \leqslant \tilde{G}_t$. Since $\phi \leqslant T + 1$ almost surely, and $\left(\tilde{G}_t\right)_t$ is a supermartingale, Doob's optional stopping theorem yields: $\mathbb{E}[\tilde{G}_\phi] \leqslant \mathbb{E}[\tilde{G}_0] = 1$, and so

$$\begin{aligned}\mathbb{P}(S(\phi) \geqslant T(\phi)\delta, \phi \leqslant T) &\leqslant \mathbb{E}[G_\phi]\\ &\leqslant \mathbb{E}[\tilde{G}_\phi]\exp(-2n\delta^2/B^2)\\ &\leqslant \exp(-2n\delta^2/B^2),\end{aligned}$$

which concludes the proof □

Lemma E.3 (Expected number of large deviation with lower-bounded mean)**.** *Let (L, K, ρ) be an OLR problem, $\mathcal{F}_t$ the natural σ-algebra generated by the OLR problem, and $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{Z}}$ the corresponding filtration. We denote $O_t := (\boldsymbol{a}(1), \boldsymbol{c}(1), \dots, \boldsymbol{a}(t-1), \boldsymbol{c}(t-1))$ the set of random values observed up to time t − 1. Let $Z_t \in [0, B]$ and $B_t \in \{0, 1\}$ be two $\mathcal{F}_{t-1}$-measurable random variables, $\Lambda \subseteq \mathbb{N}$ be a random set of instants, and $\varepsilon > 0$. For any $t \in \mathbb{Z}$, we denote $S(t) := \sum_{s=0}^{t} B_s Z_s$ and $T(t) := \sum_{s=0}^{t} B_s$. If for any $t > 0$, $\mathbb{E}[Z_t \mid O_t, B_t = 1] \geqslant \delta$ and there exists a sequence of* *random sets $(\Lambda(n))_{n>0}$ such that (i) $\Lambda \subseteq \bigcup_{n>0} \Lambda(n)$, (ii) for all $n > 0$ and all $t \in \Lambda(n)$, $T(t) \geqslant \varepsilon n$, (iii) $|\Lambda(n)| \leqslant 1$, and (iv) the event $t \in \Lambda(n)$ is $\mathcal{F}$-measurable. Then*

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\{t \in \Lambda : S(t) < \frac{\delta}{2} T(t)\}\right] \leq \frac{2B^2}{\epsilon\delta^2}$$

Proof. Let $T \in \mathbb{N}$. For all $n \in \mathbb{N}$, $|\Lambda(n)| \leqslant 1$, we define Φ_n as $T+1$ if $\Lambda(n) \cap [T]$ is empty and $\{\Phi_n\} = \Lambda(n)$ otherwise. Since $\Lambda \subseteq \bigcup_{n>0} \Lambda(n)$, we have

$$\sum_{t=1}^{T} \mathbb{1}\left\{t \in \Lambda : S(t) < \frac{\delta}{2} T(t)\right\} \leqslant \sum_{n \geqslant 1} \mathbb{1}\left\{S(\Phi_n) < \frac{\delta}{2} T(\Phi_n), \Phi_n \leqslant T\right\}.$$

Taking expectations,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{t \in \Lambda : S(t) < \frac{\delta}{2} T(t)\right\}\right] \leqslant \sum_{n \geqslant 1} \mathbb{P}\left[S(\Phi_n) < \frac{\delta}{2} T(\Phi_n), \Phi_n \leqslant T\right]$$

For any $t \in \mathbb{N}$, denote $S'(t) := \sum_{s=0}^{t} B_s(Z_s - \mathbb{E}[Z_s \mid 0_s, B_s = 1])$. As for any $s \in \mathbb{N}$, $\mathbb{E}[Z_s \mid 0_s, B_s = 1] > \delta$, $S'(t) < S(t) - T(t)\delta$. Therefore, for any $n \in \mathbb{N}$

$$\mathbb{P}\left[S(\Phi_n) < \frac{\delta}{2} T(\Phi_n), \Phi_n \leqslant T\right] \leqslant \mathbb{P}\left[S'(\Phi_n) < -\frac{\delta}{2} T(\Phi_n), \Phi_n \leqslant T\right]$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{t \in \Lambda : S(t) < \frac{\delta}{2} T(t)\right\}\right] \leqslant \sum_{n \geqslant 1} \mathbb{P}\left[S'(\Phi_n) < -\frac{\delta}{2} T(\Phi_n), \Phi_n \leqslant T\right]$$

By Lemma E.2, since Φ_n is a stopping time upper bounded by $T+1$, and $T(\Phi_n) \geqslant \varepsilon n$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{t \in \Lambda : S(t) < \frac{\delta}{2} T(t)\right\}\right] \leqslant \sum_{n \geqslant 1} \exp\left(-\frac{\varepsilon n \delta^2}{2B^2}\right) \leqslant \frac{2B^2}{\varepsilon \delta^2},$$

where the last inequality drives from the $\sum_{n \geqslant 1} \exp(-nw) \leqslant \int_0^{+\infty} \exp(-uw)\, du = \frac{1}{w}$.

This upper-bound is valid for any T, which concludes the proof. □

Lemma E.4 (Expected number of large deviation for our statistics)**.** *Let (L, K, ρ) be an OLR problem satisfying Assumptions 3.2 and 3.3 with $\succ$ the order on items, and let i and j be two items. If there exists a sequence of random sets $(\Lambda(n))_{n>0}$ such that (i) $\Lambda \subseteq \bigcup_{n>0} \Lambda(n)$, (ii) for all $n > 0$ and all $t \in \Lambda(n)$, $t_{i,j}(t+1) \geqslant \varepsilon n$, (iii) $|\Lambda(n)| \leqslant 1$, and (iv) the event $t \in \Lambda(n)$ is $\mathcal{F}$-measurable. Then,*

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in \Lambda, T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2} t_{i,j}(t)\right\}\right] = \mathcal{O}(1) \tag{4}$$

and

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in \Lambda, \frac{\hat{s}_{i,j}(t)}{\tilde{\Delta}_{i,j}} < \frac{1}{2}\right\}\right] = \mathcal{O}(1),$$

meaning

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\left\{t\in\Lambda, \hat{s}_{i,j}(t) < \frac{\tilde{\Delta}_{i,j}}{2}\right\}\right] = \mathcal{O}(1) \qquad , \textit{if } i \succ j; \tag{5}$$

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\left\{t\in\Lambda, \hat{s}_{i,j}(t) > \frac{\tilde{\Delta}_{i,j}}{2}\right\}\right] = \mathcal{O}(1) \qquad , \textit{if } j \succ i. \tag{6}$$

Proof. Let assume $i \succ j$. We first prove Claim (4) and then prove Claim (5) using Claim (4).

For any $t \leqslant 1$, we define both following $\mathcal{F}_{t-1}$-measurable random variables

$$Z_t := \mathbb{1}\left\{c_i(t) \neq c_j(t)\right\} \qquad\qquad B_t := \mathbb{1}\left\{\exists c, (i,j)\in P_c(t)^2\right\},$$

and we denote $O_t := (\boldsymbol{a}(1), \boldsymbol{c}(1), \dots, \boldsymbol{a}(t-1), \boldsymbol{c}(t-1))$ the set of random values observed up to time $s-1$. Note that $T_{i,j}(t+1) = \sum_{s=1}^{t} B_s Z_s$, $t_{i,j}(t+1) = \sum_{s=1}^{t} B_s$, and $\mathbb{E}\left[Z_t \mid 0_t, B_t = 1\right] > \tilde{\delta}_{i,j}$ by Lemma E.1.

Therefore by Lemma E.3

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\{t\in\Lambda : T_{i,j}(t+1) < \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t+1)\}\right] \leqslant \frac{2}{\epsilon\tilde{\delta}_{i,j}^2},$$

meaning

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\{t\in\Lambda : T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\}\right] \leqslant 1 + \frac{2}{\epsilon\tilde{\delta}_{i,j}^2} = \mathcal{O}(1),$$

which corresponds to Claim (4).

Let now prove Claim (5) using the following decomposition

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{t\in\Lambda, \hat{s}_{i,j}(t) < \frac{\tilde{\Delta}_{i,j}}{2}\right\}\right] \leqslant\ & \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{t\in\Lambda, \hat{s}_{i,j}(t) < \frac{\tilde{\Delta}_{i,j}}{2}T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\right\}\right]\\ &+ \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{t\in\Lambda, \hat{s}_{i,j}(t) < \frac{\tilde{\Delta}_{i,j}}{2}, T_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\right\}\right],\end{aligned}$$

Where the first right-hand side term is smaller than $\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\left\{t\in\Lambda, T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\right\}\right]$ and therefore is a $\mathcal{O}(1)$. We control the second term by applying again Lemma E.3.

For any $t \leqslant 1$, we define both following $\mathcal{F}_{t-1}$-measurable random variables

$$Z_t := c_i(t) - c_j(t) \qquad\qquad B_t := \mathbb{1}\left\{\exists c, (i,j)\in P_c(t)^2, c_i(t) \neq c_j(t)\right\},$$

Note that $Z_t \in [-1, 1]$, $\hat{s}_{i,j}(t+1)T_{i,j}(t+1) = \sum_{s=1}^{t} B_s Z_s$, $T_{i,j}(t+1) = \sum_{s=1}^{t} B_s$, and $\mathbb{E}\left[Z_t \mid 0_t, B_t = 1\right] > \tilde{\Delta}_{i,j}$ by Lemma E.1 as $i \succ j$.

We also define $A := \Lambda \cap \left\{t\in\mathbb{N} : T_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\right\}$ and for any $n \in \mathbb{N}$, $A(n) := \Lambda(n) \cap \left\{t\in\mathbb{N} : T_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)\right\}$. Then, (i) as $\Lambda \subseteq \bigcup_{n>0}\Lambda(n)$, $A \subseteq \bigcup_{n>0}A(n)$, (ii) for all $n > 0$ and all $t \in A(n)$, $T_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2}\varepsilon n$, (iii) $|A(n)| \leqslant |\Lambda(n)| \leqslant 1$, and (iv) the event $t \in A(n)$ is $\mathcal{F}$-measurable. Therefore by Lemma E.3

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\{t\in A : \hat{s}_{i,j}(t+1)T_{i,j}(t+1) < \frac{\tilde{\Delta}_{i,j}}{2}T_{i,j}(t+1)\}\right] \leqslant \frac{8}{\tilde{\delta}_{i,j}\varepsilon\tilde{\Delta}_{i,j}^2},$$

meaning

$$\mathbb{E}\left[\sum_{t\geq 1}\mathbb{1}\left\{\substack{t\in\Lambda,\hat{s}_{i,j}(t)<\frac{\tilde{\Delta}_{i,j}}{2},\\ T_{i,j}(t)\geqslant\frac{\tilde{\delta}_{i,j}}{2}t_{i,j}(t)}\right\}\right]\leqslant 1+\frac{8}{\tilde{\delta}_{i,j}\varepsilon\tilde{\Delta}_{i,j}^2}=\mathcal{O}(1).$$

Overall, $\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left\{t\in\Lambda,\hat{s}_{i,j}(t)<\frac{\tilde{\Delta}_{i,j}}{2}\right\}\right]=\mathcal{O}(1)+\mathcal{O}(1)=\mathcal{O}(1)$ which corresponds to Claim (5).

Other claims are proved symmetrically. □

E.3. Upper-Bound on the Number of Lower-Estimations of an Optimistic Estimator

This section presents two results aiming at upper-bounding the number of iterations at which $\tilde{\Delta}_{j,i}$ is lower-estimated by $\bar{\bar{s}}_{j,i}(t)$ if $j \succ i$. These new results are extensions of Lemma 9 and Theorem 10 of (Garivier & Cappé, 2011) to a setting where the handled statistic is a mixture of variables following different laws of bounded expectation.

Lemma E.5. *Let X be a random variable taking value in $[0,1]$ and let $\mu \leq \mathbb{E}[X]$. then for all $\lambda < 0$,*

$$\mathbb{E}[\exp(\lambda X)] \leq 1-\mu+\mu\exp(\lambda),$$

Proof. The function f : $[0,1] \xrightarrow{\mathbb{R}}$ defined by $f(x)=\exp(\lambda x)-x(\exp(\lambda)-1)-1$ is convex and such that $f(0)=f(1)=0$, hence $f(x)\leq 0$ for all $x\in[0,1]$. Consequently,

$$\mathbb{E}[\exp(\lambda X)]\leq\mathbb{E}[X(\exp(\lambda)-1)+1]=\mathbb{E}[X](\exp(\lambda)-1)+1$$

As $\lambda<0$ and $\mu\leq\mathbb{E}[X]$, we have $\mathbb{E}[X](\exp(\lambda)-1)\leq\mu(\exp(\lambda)-1)$ and

$$\mathbb{E}[\exp(\lambda X)]\leq\mu(\exp(\lambda)-1)+1$$

□

Lemma E.6. *Let $(X_t^a)_{t\geqslant 1}$ with $a\in\mathcal{R}$, be $|\mathcal{R}|<\infty$ independent sequences of independent random variables bounded in $[0,1]$ defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ with common expectations $\mu^a=\mathbb{E}[X_t^a]$ of minimal value $\mu=\min_{a\in\mathcal{R}}\mu^a$. Let $\mathcal{F}_t$ be an increasing sequence of $\sigma-$fields of $\mathcal{F}$ such that for each t, $\sigma((X_1^a)_{a\in\mathcal{R}},\dots,(X_t^a)_{a\in\mathcal{R}})\subset\mathcal{F}_t$ and for $s>t$ and a a recommendation, X_s^a is independent from $\mathcal{F}_t$. Consider $|\mathcal{R}|$ previsible sequences $(\epsilon_t^a)_{t\geq 1}$ of Bernoulli variables (for all $t>0$, ϵ_t^a is $\mathcal{F}_{t-1}-mesurable$) such that for all $t>0$, $\sum_i\epsilon_t^a\in\{0,1\}$. Let $\delta>0$ and for every t let*

$$S(t)=\sum_{s=1}^{t}\sum_i\epsilon_s^iX_s^i,\qquad N(t)=\sum_{s=1}^{t}\sum_i\epsilon_s^i,\qquad\hat{\mu}(t)=\frac{S(t)}{N(t)}$$

$$u(t)=\max\{q>\hat{\mu}(t):N(t)d(\hat{\mu}(t),q)\leq\delta\}$$

Then

$$\mathbb{P}(u(t)<\mu)\leq e\lceil\delta\log(t)\rceil\exp(-\delta)$$

Proof. For every $\lambda<0$, by Lemma E.5, it holds that $\log(\mathbb{E}[\exp(\lambda X_1^a)])\leq\log(1-\mu+\mu\exp(\lambda))=\phi_\mu(\lambda)$ for all a. Let $W_0^\lambda=1$ and for $t\geq 1$,

$$W_t^\lambda=\exp(\lambda S(t)-N(t)\phi_\mu(\lambda))$$

$(W_t^\lambda)_{t\geq}0$ is a super-martingale relative to $(\mathcal{F}_t)_{t\geq 0}$. In fact,

$$\mathbb{E}[\exp(\lambda\{S(t+1)-S(t)\})|\mathcal{F}_t]=\mathbb{E}[\exp(\lambda\sum_i\epsilon_{t+1}^iX_{t+1}^i)|\mathcal{F}_t]$$

As $(X_t^i)_t$ are independent sequences, we can rewrite :

$$\mathbb{E}[\exp(\lambda\{S(t+1)-S(t)\})|\mathcal{F}_t]=\prod_i\mathbb{E}[\exp(\lambda\epsilon_{t+1}^iX_{t+1}^i)|\mathcal{F}_t]=\prod_i\exp(\epsilon_{t+1}^i\log(\mathbb{E}[\exp(\lambda X_{t+1}^i)|\mathcal{F}_t]))$$

$$= \exp(\sum_i \epsilon_{t+1}^i \log(\mathbb{E}[\exp(\lambda X_1^i)|\mathcal{F}_t])) \leq \exp(\sum_i \epsilon_{t+1}^i \phi_\mu(\lambda)) = \exp(\{N(t+1) - N(t)\}\phi_\mu(\lambda))$$

which can be rewritten as

$$\mathbb{E}[\exp(\lambda S(t+1) - N(t+1)\phi_\mu(\lambda))|\mathcal{F}_t] \leq \exp(\lambda S(t) - N(t)\phi_\mu(\lambda))$$

The rest of the proof follows (Garivier & Cappé, 2011). Using the ”peeling trick”: the interval $\{1, \dots, t\}$ of possible values for $N(t)$ is divided into slices $\{t_{k-1}+1, \dots, t_k\}$ of geometrically increasing size. Each slice is treated independently. We assume that $\delta > 1$ and we construct the slicing as follow : $t_0 = 0$ and for $k \in \mathbb{N}^*$, $t_k = \lfloor (1+\eta)^k \rfloor$, with $\eta = 1/(\delta - 1)$. Let $D = \lceil \frac{\log t}{\log 1+\eta} \rceil$ be the first interval such that $t_D \geq t$ and A_k the event $\{t_{k-1} \leq N(t) \leq t_k\} \cap \{u(t) < \mu\}$. We have :

$$\mathbb{P}(u(t) < \mu) \leq \mathbb{P}(\bigcup_{k=1}^{D} A_k) \leq \sum_{k=1}^{D} \mathbb{P}(A_k)$$

Note that by definition of $u(t)$, we have $u(t) < \mu$ if and only if $\hat{\mu}(t) < \mu$ and $N(t)d(\hat{\mu}(t), \mu) > \delta$. Let s be the smallest integer such that $\delta/(s+1) \leq d(0, \mu)$. If $N(t) \leq s$, then

$$N(t)d(\hat{\mu}, \mu) \leq sd(\hat{\mu}, \mu) \underset{\text{as } \hat{\mu} \leq \mu}{\leq} sd(0, \mu) \underset{\text{by definition of } s}{<} \delta.$$

Thus, we can’t have $\hat{\mu} < \mu$ and $N(t)d(\hat{\mu}, \mu) > \delta$ and $\mathbb{P}(u(t) < \mu) = 0$. We have for all k such that $t_k \leq s$, $\mathbb{P}(A_k) = 0$ and we have $u(t) > \mu$ when $N(t) \in \{t_{k-1}+1, \dots, t_k\}$ and $t_k \leq s$.

Now lets see how $u(t)$ can be upper bounded by μ when $N(t) > s$. For k such that $t_k \geq s$, we note $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$ and we take $z < \mu$ such as $d(z, \mu) = \delta/(1+\eta)^k$ and $x \in]0, \mu[$ such that $d(x, \mu) = \delta/N(t)$. We define $\lambda(x) = \log(x(1-\mu)) - log(\mu(1-x)) < 0$ so that we can rewrite $d(x, \mu)$ as $d(x, \mu) = \lambda(x)x - \phi_\mu(\lambda(x))$.

- with $N(t) > \tilde{t}_{k-1}$, we have $d(z, \mu) = \frac{\delta}{(1+\mu)^k} \geq \frac{\delta}{(1+\mu)N(t)}$
- with $N(t) \leq t_k$, we have $d(\hat{\mu}(t), \mu) > \frac{\delta}{N(t)} > \frac{\delta}{(1+\eta)^k} = d(z, \mu)$. As $\hat{\mu} < \mu$, we have $\hat{\mu}(t) \leq z$

Hence on the event $\{\tilde{t}_{k-1} < N(t) \leq t_k\} \cap \{\hat{\mu}(t) < \mu\} \cap \{d(\hat{\mu}(t), \mu)\}$ it holds that $\lambda(z)\hat{\mu}(t) - \phi_\mu(\lambda(z)) \geq \lambda(z)z - \phi_\mu(\lambda(z)) = d(z, \mu) \geq \frac{\delta}{(1+\eta)N(t)}$

It leads to :

$$\begin{aligned}
\{\tilde{t}_{k-1} < N(t) \leq t_k\} \cap \{u(t) < \mu\} &\subset \{\lambda(z)\hat{\mu}(t) - \phi_\mu(\lambda(z)) \geq \frac{\delta}{(1+\eta)N(t)}\} \\
&\subset \{\lambda(z)S(t) - N(t)\phi_\mu(\lambda(z)) \geq \frac{\delta}{(1+\eta)}\} \\
&\subset \{W_n^\lambda(z) > \exp\left(\frac{\delta}{(1+\eta)}\right)\}
\end{aligned}$$

As $(W_t^\lambda)_{t \geq 0}$ is a supermartingale, $\mathbb{E}[W_n^{\lambda(z)}] \leq \mathbb{E}[W_n^{\lambda(z)}] = 1$, and the Markov inequality yields :

$$\mathbb{P}(\{\tilde{t}_{k-1} < N(t) \leq t_k\} \cap \{u(t) < \mu\}) \leq \mathbb{P}\left(W_n^\lambda(z) > \exp\left(\frac{\delta}{(1+\eta)}\right)\right) \leq \exp\left(-\frac{\delta}{(1+\eta)}\right)$$

As $\eta = 1/(\delta - 1)$, $D = \lceil \frac{\log n}{\log 1+\eta} \rceil$ and $\log(1 + 1/(\delta - 1)) \geq 1/\delta$, we obtain :

$$\mathbb{P}(u(t) < \mu) \leq \left\lceil \frac{\log n}{\log\left(1 + \frac{1}{\delta - 1}\right)} \right\rceil \exp(-\delta + 1) \leq e\lceil \delta \log(t) \rceil \exp(-\delta)$$

□

F. Proof of Theorem 5.1 (Upper-Bound on the Regret of UniRank Assuming a Total Order on Items)

Before proving the regret upper-bound of UniRank, we prove Lemmas F.1 and F.2 which are respectively bounding the exploration when the leader is the optimal one, and the number of iterations at which the leader is sub-optimal. Finally, the regret upper-bound of UniRank is given in Appendix F.3.

F.1. Upper-Bound on the Number of Sub-Optimal Merges of UniRank when the Leader is the Optimal Partition

Lemma F.1 (Upper-bound on the number of sub-optimal merges of UniRank when the leader is the optimal partition)**.** *Under the hypotheses of Theorem 5.1, for any position $c \in \{2, \dots, L\}$ UniRank fulfills*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\\ \exists c', P_{c'}(t)=\{\min(c-1,K),c\}}\right\}\right] \leqslant \frac{16}{\tilde{\delta}^*_c \tilde{\Delta}^2_{\min(c-1,K),c}} \log T + \mathcal{O}\left(\log\log T\right).$$

Proof. Let $c \in \{2, \dots, L\}$ be a position, and denote i (respectively j) the item $\min(c-1, K)$ (resp. c). We aim at upper-bounding the number of iterations such that the leader $\tilde{\boldsymbol{P}}(t)$ is the optimal partition $\boldsymbol{P}^*$, and either the subsets $\boldsymbol{P}^*_{c-1} = \{i\}$ and $\boldsymbol{P}^*_c = \{j\}$ are merged in the chosen partition $\boldsymbol{P}(t)$, or $j \in \boldsymbol{P}^*_{K+1}(t)$ is added to the subset $\boldsymbol{P}^*_K = \{i\}$ in the chosen partition $\boldsymbol{P}(t)$. Both situations require $\bar{\bar{s}}_{j,i}(t)$ to be positive.

Let decompose this number of iterations:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\\ \exists c', P_{c'}(t)=\{\min(c-1,K),c\}}\right\}\right] &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\},\\ \bar{\bar{s}}_{j,i}(t)\geqslant 0}\right\}\right]\\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\},\\ \hat{s}_{j,i}(t)>\frac{\tilde{\Delta}_{j,i}}{2}}\right\}\right]\\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\},\\ T^*_j(t)<\frac{\tilde{\delta}^*_j}{2}t^*_j(t)}\right\}\right]\\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\substack{\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\},\\ T^*_j(t)\geqslant\frac{\tilde{\delta}^*_j}{2}t^*_j(t),\ \hat{s}_{j,i}(t)\leqslant\frac{\tilde{\Delta}_{j,i}}{2},\\ \bar{\bar{s}}_{j,i}(t)\geqslant 0}\right\}\right],
\end{aligned}$$

where $t^*_j(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right\} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\}$,
and $T^*_j(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right\} \mathbb{1}\left\{\exists c, (i,j) \in P_c(s)^2\right\} \mathbb{1}\{c_i(s) \neq c_j(s)\}$.

Let bound the first term in the right-hand side.

Denote $\Lambda = \left\{t : \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*,\ \exists c', P_{c'}(t) = \{i,j\}\right\}$ the set of iterations at which $\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*$ and both items i and j are gathered in a subset of $\boldsymbol{P}(t)$. We decompose that set as $\Lambda \subseteq \bigcup_{s\in\mathbb{N}} \Lambda(s)$, with $\Lambda(s) := \{t \in \Lambda : t_{i,j}(t) = s\}$. $|\Lambda(s)| \leqslant 1$ as $t_{i,j}(t)$ increases for each $t \in \Lambda$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda(s)$, $t_{i,j}(n) \geqslant t_{i,j}(n) = s$.

Note also that with the current hypothesis on the order, $i \succ j$, hence by Lemma E.4,

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in \Lambda, \hat{s}_{j,i}(t) > \frac{\tilde{\Delta}_{j,i}}{2}\right\}\right] = \mathcal{O}(1).$$

The second term is bounded similarly with the same set Λ but with a different decomposition: $\Lambda \subseteq \bigcup_{s\in\mathbb{N}} \Lambda(s)$, with $\Lambda(s) := \{t \in \Lambda : t^*_j(t) = s\}$. $|\Lambda(s)| \leqslant 1$ as $t^*_j(t)$ increases for each $t \in \Lambda$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda(s)$, $t^*_j(n) \geqslant t^*_j(n) = s$.

Therefore, the same proof as the one used in Lemma E.4 gives

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in \Lambda, T^*_j(t) < \frac{\tilde{\delta}^*_j}{2} t^*_j(t)\right\}\right] = \mathcal{O}(1)$$

It remains to upper-bound the third term.

Let note $C := \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*,\ \exists c', P_{c'}(t) = \{i,j\},\ T_j^*(t) \geqslant \frac{\tilde{\delta}_j^*}{2} t_j^*(t),\ \hat{s}_{j,i}(t) \leqslant \frac{\tilde{\Delta}_{j,i}}{2},\ \bar{\bar{s}}_{j,i}(t) \geqslant 0\right\}$.

Let $t \in C$.

By Pinsker's inequality and as $\bar{\bar{s}}_{j,i}(t) \geqslant 0$,

$$\begin{aligned}\frac{1}{2} &\leqslant \frac{\bar{\bar{s}}_{j,i}(t) + 1}{2} \\ &\leqslant \frac{\hat{s}_{j,i}(t) + 1}{2} + \sqrt{\frac{\log(\tilde{t}_{\boldsymbol{P}^*}(t)) + 3\log(\log(\tilde{t}_{\boldsymbol{P}^*}(t)))}{2T_{i,j}(t)}} \\ &\leqslant \frac{\tilde{\Delta}_{j,i}}{4} + \frac{1}{2} + \sqrt{\frac{\log(\tilde{t}_{\boldsymbol{P}^*}(t))) + 3\log(\log(\tilde{t}_{\boldsymbol{P}^*}(t))))}{2T_{i,j}(t)}}.\end{aligned}$$

Hence, $T_{i,j}(t) \leqslant \frac{8\log(\tilde{t}_{\boldsymbol{P}^*}(t)))+24\log(\log(\tilde{t}_{\boldsymbol{P}^*}(t))))}{\tilde{\Delta}_{i,j}^2}$ as $\tilde{\Delta}_{i,j} = -\tilde{\Delta}_{j,i} > 0$ given Lemma E.1. Then, by definition of C and as (i) $\tilde{t}_{\boldsymbol{P}^*}(t) \leqslant t \leqslant T$, (ii) $T_j^*(t) \leqslant T_{i,j}(t)$, and (iii) $\tilde{\delta}_j^* \geqslant \tilde{\delta}_{i,j} > 0$ given Lemma E.1, $t_j^*(t) \leqslant \frac{2T_j^*(t)}{\tilde{\delta}_j^*} \leqslant \frac{2T_{i,j}(t)}{\tilde{\delta}_j^*} \leqslant \frac{16\log(T)+48\log(\log(T))}{\tilde{\delta}_j^* \tilde{\Delta}_{i,j}^2}$.

Therefore, $C \subseteq \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*,\ \exists c', P_{c'}(t) = \{i,j\},\ t_j^*(t) \leqslant \frac{16\log(T)+48\log(\log(T))}{\tilde{\delta}_j^* \tilde{\Delta}_{i,j}^2}\right\}$, and

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\begin{smallmatrix}\tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\}, \\ T_j^*(t) \geqslant \frac{\tilde{\delta}_j^*}{2} t_j^*(t),\ \hat{s}_{j,i}(t) \leqslant \frac{\tilde{\Delta}_{j,i}}{2}, \\ \bar{\bar{s}}_{j,i}(t) \geqslant 0\end{smallmatrix}\right\}\right] &= \mathbb{E}\left[|C|\right] \\ &\leqslant \mathbb{E}\left[\left|\left\{\begin{smallmatrix}t \in [T]:\ \tilde{\boldsymbol{P}}(t)=\boldsymbol{P}^*,\ \exists c', P_{c'}(t)=\{i,j\}, \\ t_j^*(t) \leqslant \frac{16\log(T)+48\log(\log(T))}{\tilde{\delta}_j^* \tilde{\Delta}_{i,j}^2}\end{smallmatrix}\right\}\right|\right] \\ &\leqslant \frac{16\log(T)+48\log(\log(T))}{\tilde{\delta}_j^* \tilde{\Delta}_{i,j}^2},\end{aligned}$$

which concludes the proof.

□

F.2. Upper-Bound on the Expected Number of Iterations at which the Leader is not the Optimal Partition

Lemma F.2 (Upper-bound on the expected number of iterations at which the leader is not the optimal partition)**.** *Under the hypotheses of Theorem 5.1, UniRank fulfills*

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

Proof. Let $\tilde{\boldsymbol{P}} \neq \boldsymbol{P}^*$ be an ordered partition of items of size d, and let upper-bound the expected number of iterations at which $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$ by $\mathcal{O}\left(\log\log T\right)$. As there is a finite number of partitions, this will conclude the proof.

In this proof, for any couple of items (i,j) we denote $\tilde{t}_{i,j}(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \exists c, (i,j) \in P_c(s)^2\right\}$ the number of iterations at which both items have been gathered in the same subset of $\boldsymbol{P}(s)$ while the leader was $\tilde{\boldsymbol{P}}$. For each partition $\boldsymbol{P}$ in the neighborhood $\mathcal{N}\left(\tilde{\boldsymbol{P}}\right)$, we also denote $t_{\boldsymbol{P}}(t) := \sum_{s=1}^{t-1} \mathbb{1}\left\{\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(t) = \boldsymbol{P}\right\}$ the number of iterations at which $\boldsymbol{P}$ has been chosen while the leader was $\tilde{\boldsymbol{P}}$.

The proof depends on the difference between $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}^*$. By Lemma 5.4,

- either $\exists c \in [\tilde{d}]$, such that $|P_c| > 1$ and $i^* \succ \operatorname{argmax}_{j \in P_c \setminus \{i^*\}} g(j)$, where $i^* = \operatorname{argmax}_{i \in P_c} g(i)$;

- or $\exists c \in [\tilde{d}-1]$, $\exists (i,j) \in \tilde{P}_c \times \tilde{P}_{c+1}$, such that $j \succ i$.

We first upper-bound the expected number of iterations at which $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$ under the first condition, and then prove a similar upper-bound under the second condition.

Assume that there exists $c \in [\tilde{d}]$**, such that** $|P_c| > 1$ **and** $i \succ \operatorname{argmax}_{j \in P_c \setminus \{i^*\}} g(j)$**, where** $i = \operatorname{argmax}_{i \in P_c} g(i)$**.** Let t be an iteration such that $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$. By Assumption 3.3 and by design of the algorithm, if for each item $j \in \tilde{P}_c \setminus \{i\}$, the sign of $\hat{s}_{i,j}(t)$ would be the same as the sign of $\tilde{\Delta}_{i,j} > 0$, then i would be alone in $\tilde{P}_c(t)$. So $\hat{s}_{i,j}(t) \leqslant 0$ for at least one item $j \in \tilde{P}_c \setminus \{i\}$. Let control the number of iteration at which this is true by considering the following decomposition:

$$\left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}\right\} \subseteq \bigcup_{j \in \tilde{P}_c \setminus \{i\}} A_{i,j} \cup B_{i,j} \cup C_{i,j},$$

where

$$A_{i,j} := \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2} t_{i,j}(t)\right\},$$

$$B_{i,j} := \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \frac{\hat{s}_{i,j}(t)}{\tilde{\Delta}_{i,j}} < \frac{1}{2}\right\},$$

and

$$C_{i,j} := \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, T_{i,j}(t) \geqslant \frac{\tilde{\delta}_{i,j}}{2} t_{i,j}(t), \frac{\hat{s}_{i,j}(t)}{\tilde{\Delta}_{i,j}} \geqslant \frac{1}{2}, \hat{s}_{i,j}(t) \leqslant 0, \right\}.$$

Let j be an item in $\tilde{P}_c \setminus \{i\}$, and let first upper-bound the expected size of $A_{i,j}$ and $B_{i,j}$, and then the expected size of $C_{i,j}$.

Note that at each iteration such that $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$, i and j are in the same subset of the partition $\boldsymbol{P}(t)$, therefore $\tilde{t}_{i,j}(t) = \tilde{t}_{\tilde{\boldsymbol{P}}}(t)$.

Denote $\Lambda = \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}\right\}$ the set of iterations at which $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$, and decompose that set as $\Lambda \subseteq \bigcup_{s \in \mathbb{N}} \Lambda(s)$, with $\Lambda(s) := \{t \in \Lambda : \tilde{t}_{\tilde{\boldsymbol{P}}}(t) = s\}$. $|\Lambda(s)| \leqslant 1$ as $\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$ increases for each $t \in \Lambda$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda(s)$, $t_{i,j}(n) \geqslant \tilde{t}_{i,j}(n) = \tilde{t}_{\tilde{\boldsymbol{P}}}(t) = s$.

Then by Lemma E.4

$$\mathbb{E}\left[|A_{i,j}|\right] = \mathbb{E}\left[\sum_{t \geq 1} \mathbb{1}\left\{t \in \Lambda, T_{i,j}(t) < \frac{\tilde{\delta}_{i,j}}{2} t_{i,j}(t)\right\}\right] = \mathcal{O}(1)$$

and

$$\mathbb{E}\left[|B_{i,j}|\right] = \mathbb{E}\left[\sum_{t \geq 1} \mathbb{1}\left\{t \in \Lambda, \frac{\hat{s}_{i,j}(t)}{\tilde{\Delta}_{i,j}} < \frac{1}{2}\right\}\right] = \mathcal{O}(1).$$

Let now upper-bound the expected size of $C_{i,j}$.

As $i \succ j$, $\tilde{\Delta}_{i,j} > 0$.

Let $t \in C_{i,j}$. As $\hat{s}_{i,j}(t) \leqslant 0$, $t \leqslant T$, and $\tilde{t}_{\tilde{\boldsymbol{P}}}(t) = \tilde{t}_{i,j}(t) \leqslant t_{i,j}(t) \leqslant \frac{2}{\tilde{\delta}_{i,j}} T_{i,j}(t)$,

$$0 \geqslant \hat{s}_{i,j}(t) \geqslant \frac{\tilde{\Delta}_{i,j}}{2} > 0,$$

which is absurd. Hence, $C_{i,j} = \varnothing$, and $\mathbb{E}\left[|C_{i,j}|\right] = 0$.

Overall, if there exists $c \in [\tilde{d}]$, such that $|P_c| > 1$ and $i \succ \operatorname{argmax}_{j \in P_c \setminus \{i\}} g(j)$, where $i = \operatorname{argmax}_{i \in P_c} g(i)$,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\{\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}\}\right] &\leqslant \sum_{j \in \tilde{P}_c \setminus \{i\}} \mathbb{E}\left[|A_{i,j}|\right] + \mathbb{E}\left[|B_{i,j}|\right] + \mathbb{E}\left[|C_{i,j}|\right] \\ &= \mathcal{O}(1) + \mathcal{O}(1) + 0 \\ &= \mathcal{O}(1)\end{aligned}$$

Assume that there exists $c \in [\tilde{d} - 1]$**, and** $(i, j) \in \tilde{P}_c \times \tilde{P}_{c+1}$**, such that** $j \succ i$**.** By design of UniRank, each neighbor of $\tilde{\boldsymbol{P}}$ takes one of both forms:

1. $\left(\tilde{P}_1, \dots, \tilde{P}_{c-1}, \tilde{P}_c \cup \tilde{P}_{c+1}, \tilde{P}_{c+2}, \dots \tilde{P}_{\tilde{d}}\right)$,
2. $\left(\tilde{P}_1, \dots, \tilde{P}_{\tilde{d}-2}, \tilde{P}_{\tilde{d}-1} \cup \{j\}, \tilde{P}_{\tilde{d}} \setminus \{j\}\right)$.

Let $\boldsymbol{P}$ be such neighbor. In the first scenario we denote $i(\boldsymbol{P}) := \operatorname{argmin}_{i \in P_c} g(i)$ and $j(\boldsymbol{P}) := \operatorname{argmax}_{j \in P_{c+1}} g(j)$. In the second scenario we denote $i(\boldsymbol{P}) := \operatorname{argmin}_{i \in P_{\tilde{d}-1}} g(i)$, and $j(\boldsymbol{P})$ the item j. Finally, we denote $\mathcal{N}^+$ the set of neighbors $\boldsymbol{P}$ of $\tilde{\boldsymbol{P}}$ such that $j(\boldsymbol{P}) \succ i(\boldsymbol{P})$, and $\mathcal{N}^-$ its complement $\{\tilde{\boldsymbol{P}}\} \cup \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$.

It is also worth noting that with current hypothesis on $\tilde{\boldsymbol{P}}$,

- $|\mathcal{N}^+| + |\mathcal{N}^-| = \left|\mathcal{N}\left(\tilde{\boldsymbol{P}}\right)\right| + 1 \leqslant L$;
- $\mathcal{N}^+$ is non-empty (due to current assumption on $\tilde{\boldsymbol{P}}$);
- for each partition $\boldsymbol{P} \in \mathcal{N}(\tilde{\boldsymbol{P}})$, $t_{\boldsymbol{P}}(t) = \tilde{t}_{i(\boldsymbol{P}),j(\boldsymbol{P})}(t)$;
- by design of the algorithm, at each iteration t such that $\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}$, $\hat{s}_{i(\boldsymbol{P}),j(\boldsymbol{P})}(t) > 0$ for each partition $\boldsymbol{P} \in \mathcal{N}(\tilde{\boldsymbol{P}})$ as $i(\boldsymbol{P})$ is in a subset before $j(\boldsymbol{P})$ in $\tilde{\boldsymbol{P}}$.

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

$$A_{\boldsymbol{P}^+} = \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, t_{\boldsymbol{P}^+}(t) \geqslant \varepsilon \tilde{t}_{\tilde{\boldsymbol{P}}}(t)\right\},$$

$$B = \left\{t : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \forall \boldsymbol{P} \in \mathcal{N}^+, t_{\boldsymbol{P}^+}(t) < \varepsilon \tilde{t}_{\tilde{\boldsymbol{P}}}(t)\right\},$$

$$\text{and } \varepsilon := \frac{1}{\left|\mathcal{N}\left(\tilde{\boldsymbol{P}}\right)\right| + 1} \geqslant \frac{1}{L}.$$

Hence,

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

Bound on $\mathbb{E}\left[|A_{\boldsymbol{P}^+}|\right]$ Let $\boldsymbol{P} \in \mathcal{N}^+$ be a permutation.

First, let's t be in $A_{\boldsymbol{P}^+}$. Note that $\tilde{\Delta}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)} < 0$, as $j(\boldsymbol{P}^+) \succ i(\boldsymbol{P}^+)$. Therefore, as $\hat{s}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(t) > 0$, $\hat{s}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(t) > \frac{\tilde{\Delta}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}}{2}$, and thus $E\left[|A_{\boldsymbol{P}^+}|\right] = \mathbb{E}\left[\sum_{t \geq 1} \mathbb{1}\left\{t \in A_{\boldsymbol{P}^+}, \hat{s}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(t) > \frac{\tilde{\Delta}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}}{2}\right\}\right]$.

Secondly, let's decompose $A_{\boldsymbol{P}^+}$ as $A_{\boldsymbol{P}^+} \subseteq \bigcup_{s \in \mathbb{N}} \Lambda(s)$, with $\Lambda(s) := \{t \in A_{\boldsymbol{P}^+} : \tilde{t}_{\tilde{\boldsymbol{P}}}(t) = s\}$. $|\Lambda(s)| \leqslant 1$ as $\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$ increases for each $t \in A_{\boldsymbol{P}^+}$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda(s)$, $t_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(n) \geqslant \tilde{t}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(n) = t_{\boldsymbol{P}^+}(n) \geqslant \varepsilon \tilde{t}_{\tilde{\boldsymbol{P}}}(t) = \varepsilon s$.

Thus, as $j(\boldsymbol{P}^+) \succ i(\boldsymbol{P}^+)$, by Lemma E.4

$$\mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in A_{\boldsymbol{P}^+}, \hat{s}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(t) > \frac{\tilde{\Delta}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}}{2}\right\}\right] = \mathcal{O}(1).$$

Overall, $E\left[|A_{\boldsymbol{P}^+}|\right] = \mathbb{E}\left[\sum_{t\geq 1} \mathbb{1}\left\{t \in A_{\boldsymbol{P}^+}, \hat{s}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}(t) > \frac{\tilde{\Delta}_{i(\boldsymbol{P}^+),j(\boldsymbol{P}^+)}}{2}\right\}\right] = \mathcal{O}\left(1\right).$

Bound on $\mathbb{E}\left[|B|\right]$ We first split B in two parts: $B = B^{t_0} \cup B^T_{t_0}$, where $B^{t_0} := \{t \in B : \tilde{t}_{\tilde{\boldsymbol{P}}}(t) \leqslant t_0\}$, $B^T_{t_0} := \{t \in B : \tilde{t}_{\tilde{\boldsymbol{P}}}(t) > t_0\}$, and t_0 is chosen as small as possible to satisfy a constraint required later on in the proof. Namely, $t_0 = \max_{\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}})\setminus\mathcal{N}^+} \inf\left\{s : \sqrt{\frac{\log(s)+3\log(\log(s))}{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\varepsilon s-1)}} < \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{8}\right\}$, with $t_0 = 0$ if $\mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$ is empty. Note that t_0 only depends on $\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}$ and $\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}$ for $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$.

We also define

- $D_{\boldsymbol{P}^-} := \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(t) = \boldsymbol{P}^-, T_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(t) < \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2} t_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(t)\right\}$, for each $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$
- $E_{\boldsymbol{P}^-} := \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(t) = \boldsymbol{P}^-, \hat{s}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t) < \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{2}\right\}$, for each $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$
- $F_{\boldsymbol{P}^+} := \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \frac{\bar{\bar{s}}_{j(\boldsymbol{P}),i(\boldsymbol{P})}(t)+1}{2} \leqslant \frac{\tilde{\Delta}_{j(\boldsymbol{P}),i(\boldsymbol{P})}+1}{2}\right\}$ for each $\boldsymbol{P}^+ \in \mathcal{N}^+$.

Let $t \in B^T_{t_0}$. We have

$$\tilde{t}_{\tilde{\boldsymbol{P}}}(t) = \sum_{\boldsymbol{P}^+ \in \mathcal{N}^+} t_{\boldsymbol{P}^+}(t) + \sum_{\boldsymbol{P}^- \in \mathcal{N}^-} t_{\boldsymbol{P}^-}(t),$$

and by definition of B, $t_{\boldsymbol{P}^+}(t) < \varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$ for each $\boldsymbol{P}^+ \in \mathcal{N}^+$. So, there exists $\boldsymbol{P}^- \in \mathcal{N}^-$ such that $t_{\boldsymbol{P}^-}(t) > \varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$ (otherwise, $\tilde{t}_{\tilde{\boldsymbol{P}}}(t) = \sum_{\boldsymbol{P}^+ \in \mathcal{N}^+} t_{\boldsymbol{P}^+}(t) + \sum_{\boldsymbol{P}^- \in \mathcal{N}^-} t_{\boldsymbol{P}^-}(t) < (|\mathcal{N}^+| + |\mathcal{N}^-|)\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t) = \tilde{t}_{\tilde{\boldsymbol{P}}}(t)$, which is absurd).

Let's elicit an iteration $\psi(t)$ with specific properties. We denote s' the first iteration such that $t_{\boldsymbol{P}^-}(s') \geqslant \varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)$. At this iteration, $t_{\boldsymbol{P}^-}(s') = \lceil\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)\rceil$, and $t_{\boldsymbol{P}^-}(s') = t_{\boldsymbol{P}^-}(s'-1)+1$, meaning that $\tilde{\boldsymbol{P}}(s'-1) = \tilde{\boldsymbol{P}}$ and $\boldsymbol{P}(s'-1) = \boldsymbol{P}^-$, and $t_{\boldsymbol{P}^-}(s'-1) = \lceil\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)\rceil - 1$. Therefore, the set $\{s \in [t] : \tilde{\boldsymbol{P}}(s) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(s) = \boldsymbol{P}^-, t_{\boldsymbol{P}^-}(s) = \lceil\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)\rceil - 1\}$ is non-empty. We define $\psi(t)$ as the minimum on this set

$$\psi(t) := \min\{s \in [t] : \tilde{\boldsymbol{P}}(s) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(s) = \boldsymbol{P}^-, t_{\boldsymbol{P}^-}(s) = \lceil\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)\rceil - 1\}.$$

Let prove by contradiction that $\psi(t) \in \bigcup_{\boldsymbol{P}^- \in \mathcal{N}^-}(D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}) \cup \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}$. Assume that $\psi(t) \notin \bigcup_{\boldsymbol{P}^- \in \mathcal{N}^-}(D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}) \cup \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}$. The partition $\boldsymbol{P}^-$ is in $\mathcal{N}^-$, so either $\boldsymbol{P}^- = \tilde{\boldsymbol{P}}$ or $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$.

The set $\mathcal{N}^+$ is non-empty, so there exists a partition $\boldsymbol{P}^+ \in \mathcal{N}^+$. As $\psi(t) \notin \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}$, $\frac{\bar{\bar{s}}_{j(\boldsymbol{P}),i(\boldsymbol{P})}(\psi(t))+1}{2} > \frac{\tilde{\Delta}_{j(\boldsymbol{P}),i(\boldsymbol{P})}+1}{2}$, where $\tilde{\Delta}_{j(\boldsymbol{P}),i(\boldsymbol{P})} > 0$ as $j(\boldsymbol{P}) \succ i(\boldsymbol{P})$. Thus $\bar{\bar{s}}_{j(\boldsymbol{P}),i(\boldsymbol{P})}(\psi(t)) > 0$ and $\boldsymbol{P}(\psi(t)) = \boldsymbol{P}^-$ cannot be $\tilde{\boldsymbol{P}}$ by design of UniRank. Therefore $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$.

Thus, either $\mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$ is empty and we get a contradiction, or $i(\boldsymbol{P}^-)$ and $j(\boldsymbol{P}^-)$ are properly defined, and, by design of UniRank, $\bar{\bar{s}}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) \geqslant 0$. Moreover, since $\tilde{\boldsymbol{P}}(\psi(t)) = \tilde{\boldsymbol{P}}^-$ and $\psi(t) \notin D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}$, $T_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) \geqslant \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2} t_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t))$ and $\hat{s}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(\psi(t)) \geqslant \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{2}$.

Therefore,

$$\begin{aligned} T_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) &\geqslant \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2} t_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) \geqslant \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2} \tilde{t}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) \\ &= \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2} t_{\boldsymbol{P}^-}(\psi(t)) = \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2}(\lceil\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t)\rceil - 1) \geqslant \frac{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}}{2}(\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t) - 1) \end{aligned}$$

and by Pinsker's inequality and the fact that $\psi(t) \leqslant t$ and $\tilde{t}_{\tilde{\boldsymbol{P}}}(s)$ is non-decreasing in s, and $\tilde{t}_{\tilde{\boldsymbol{P}}}(t) > t_0$,

$$\begin{aligned}
\frac{1}{2} \geqslant \frac{-\bar{\bar{s}}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) + 1}{2} &\geqslant \frac{-\hat{s}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t)) + 1}{2} - \sqrt{\frac{\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(\psi(t))) + 3\log(\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(\psi(t))))}{2T_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t))}} \\
&= \frac{\hat{s}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(\psi(t)) + 1}{2} - \sqrt{\frac{\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(\psi(t))) + 3\log(\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(\psi(t))))}{2T_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\psi(t))}} \\
&\geqslant \frac{1}{2} + \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{4} - \sqrt{\frac{\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(t)) + 3\log(\log(\tilde{t}_{\tilde{\boldsymbol{P}}}(t)))}{\tilde{\delta}_{j(\boldsymbol{P}^-),i(\boldsymbol{P}^-)}(\varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t) - 1)}} \\
&\geqslant \frac{1}{2} + \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{4} - \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{8} \\
&= \frac{1}{2} + \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{8}
\end{aligned}$$

which contradicts the fact that $\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)} > 0$.

Overall, we always get a contradiction, so, for any $t \in B_{t_0}^T$, $\psi(t) \in \bigcup_{\boldsymbol{P}^- \in \mathcal{N}^-}(D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}) \cup \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}$.

Hence, $B_{t_0}^T \subseteq \bigcup_{n \in \bigcup_{\boldsymbol{P}^- \in \mathcal{N}^-}(D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}) \cup \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}} B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}$. Let n be in $\bigcup_{\boldsymbol{P}^- \in \mathcal{N}^-}(D_{\boldsymbol{P}^-} \cup E_{\boldsymbol{P}^-}) \cup \bigcup_{\boldsymbol{P}^+ \in \mathcal{N}^+} F_{\boldsymbol{P}^+}$. For any t in $B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}$, there exists a partition $\boldsymbol{P}^- \in \mathcal{N}^-$ such that $t_{\boldsymbol{P}^-}(n) = \lceil \varepsilon\tilde{t}_{\tilde{\boldsymbol{P}}}(t) - 1 \rceil$, and $t_{\boldsymbol{P}^-}(n+1) = t_{\boldsymbol{P}^-}(n) + 1$. So $|B_{t_0}^T \cap \{t \in [T] : \psi(t) = n\}| \leqslant L$ and

$$\mathbb{E}\left[|B|\right] \leqslant t_0 + \mathbb{E}\left[|B_{t_0}^T|\right] \leqslant t_0 + |\mathcal{N}^-|(\mathbb{E}\left[|D|\right] + \mathbb{E}\left[|E|\right] + \mathbb{E}\left[|F|\right]).$$

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 $\mathbb{E}\left[|D_{\boldsymbol{P}^-}|\right]$ and $\mathbb{E}\left[|E_{\boldsymbol{P}^-}|\right]$ Let $\boldsymbol{P}^- \in \mathcal{N}(\tilde{\boldsymbol{P}}) \setminus \mathcal{N}^+$ The upper-bound on $\mathbb{E}\left[|D_{\boldsymbol{P}^-}|\right]$ and $\mathbb{E}\left[|E_{\boldsymbol{P}^-}|\right]$ are obtained through Lemma E.4. Let $\Lambda := \left\{t \in [T] : \tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}, \boldsymbol{P}(t) = \boldsymbol{P}^-\right\}$ and let use the decomposition $\Lambda \subseteq \bigcup_{s \in \mathbb{N}} \Lambda(s)$, where $\Lambda(s) := \{t \in \Lambda : t_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t) = s\}$. $|\Lambda(s)| \leqslant 1$ as $t_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t)$ increases for each $t \in \Lambda$. Note that for each $s \in \mathbb{N}$ and $n \in \Lambda(s)$, $t_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(n) \geqslant t_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(n) = s$. Then, by Lemma E.4, as $i(\boldsymbol{P}^-) \succ j(\boldsymbol{P}^-)$

$$\mathbb{E}\left[|D_{\boldsymbol{P}^-}|\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in \Lambda : T_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t) < \frac{\tilde{\delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{2} t_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t)\}\right] = \mathcal{O}(1)$$

and

$$\mathbb{E}\left[|E_{\boldsymbol{P}^-}|\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{t \in \Lambda : \hat{s}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}(t) < \frac{\tilde{\Delta}_{i(\boldsymbol{P}^-),j(\boldsymbol{P}^-)}}{2}\}\right] = \mathcal{O}(1).$$

Bound on $\mathbb{E}\left[|F_{\boldsymbol{P}^+}|\right]$ By Lemma E.6, for each partition $\boldsymbol{P}^+ \in \mathcal{N}^+$, $\mathbb{E}\left[|F_{\boldsymbol{P}^+}|\right] = O(\log(\log(T)))$.

Overall $\mathbb{E}\left[\mathbb{1}\{\tilde{\boldsymbol{P}}(t) = \tilde{\boldsymbol{P}}\}\right] \leqslant |\mathcal{N}^+|\mathcal{O}(1) + t_0 + |\mathcal{N}^-|\left(\left(|\mathcal{N}^-| - 1\right)\mathcal{O}(1) + (|\mathcal{N}^-| - 1)\mathcal{O}(1) + |\mathcal{N}^+|\mathcal{O}(\log\log T)\right) = \mathcal{O}\left(\log\log T\right)$, which concludes the proof. □

F.3. Final Step of the Proof of Theorem 5.1 (Upper-Bound on the Regret of UniRank Assuming a Total Order on Items)

The proof of Theorem 5.1 from Lemmas F.1 and F.2 is mainly based on an appropriate decomposition of the regret.

Proof of Theorem 5.1. The upper-bound on the expected number of iterations at which UniRank explores while the leader is the optimal partition is given by Lemma F.1.

The upper-bound on the expected number of iterations at which the leader is not the optimal partition is given by Lemma F.2.

Let now consider the impact of these upper-bounds on the regret of UniRank.

Let remind that $P_c^* = \{c\}$ for $c \in [K]$, $d^* = K+1$, and $P_{K+1}^* = [L] \setminus [K]$. Therefore, $\mu^* = \mu_{a^*} = \sum_{k=1}^K \rho(\boldsymbol{a}^*, k)$, where $\boldsymbol{a}^* := (1, 2, \dots, K)$.

Let first upper-bound the regret suffered at iteration t while the the leader is the optimal partition:

$$\begin{aligned}
R_t^* &= \mu^* - \mathbb{E}_{\boldsymbol{a}(t)}\left[\mu_{\boldsymbol{a}(t)} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\\
&= \sum_{k=1}^K \rho(\boldsymbol{a}^*, k) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k) \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\\
&= \sum_{k=1}^K \mathbb{P}\left(a_k(t) = k \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\left(\rho(\boldsymbol{a}^*, k) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k) \mid a_k(t) = k, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right)\\
&\quad + \sum_{k=2}^K \mathbb{P}\left(a_{k-1}(t) = k \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\left(\rho(\boldsymbol{a}^*, k-1) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k-1) \mid a_{k-1}(t) = k, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right)\\
&\quad + \sum_{k=2}^K \mathbb{P}\left(a_k(t) = k-1 \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\left(\rho(\boldsymbol{a}^*, k) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k) \mid a_k(t) = k-1, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right)\\
&\quad + \sum_{\ell=K+1}^L \mathbb{P}\left(a_K(t) = \ell \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\left(\rho(\boldsymbol{a}^*, k) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), K) \mid a_K(t) = \ell, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right)
\end{aligned}$$

Let's focus on the first right hand-side term. As the probability of click at position k only depends on the set of items in positions 1 to $k-1$, and as under the condition $a_k(t) = k \wedge \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*$, $\boldsymbol{a}(t)$ and $\boldsymbol{a}^*$ have the same set of items in positions 1 to $k-1$, $\rho(\boldsymbol{a}^*, k) = \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k) \mid a_k(t) = k, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]$. Hence that term is equal to 0.

Let now take a look at the second term. By design of UniRank, as $a_{k-1}(t) = k \wedge \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*$, there exists c' such that $P_c(t) = \{k-1, k\}$, and

$$\begin{aligned}
\mathbb{P}\left(a_{k-1}(t) = k \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) &= \mathbb{P}\left(a_{k-1}(t) = k, P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\\
&= \frac{1}{2}\mathbb{P}\left(P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right).
\end{aligned}$$

Similarly, the third term corresponds to the existence of c' such that $P_c(t) = \{k-1, k\}$, and

$$\mathbb{P}\left(a_k(t) = k-1 \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) = \frac{1}{2}\mathbb{P}\left(P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right).$$

By summing both terms, we have to handle

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left(P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \cdot\\
&\qquad\left(\rho(\boldsymbol{a}^*, k-1) + \rho(\boldsymbol{a}^*, k) - \mathbb{E}_{\boldsymbol{a}(t)}\left[\rho(\boldsymbol{a}(t), k-1) + \rho(\boldsymbol{a}(t), k) \mid a_{k-1}(t) = k, a_k(t) = k-1, \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right),
\end{aligned}$$

which is equal to $\frac{1}{2}\mathbb{P}\left(P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\Delta_k$, where

$$\Delta_k := \rho(\boldsymbol{a}^*, k-1) + \rho(\boldsymbol{a}^*, k) - \rho((k-1, k) \circ \boldsymbol{a}^*, k-1) - \rho((k-1, k) \circ \boldsymbol{a}^*, k),$$

as the probability of click at any position k' only depends on the set of items in positions 1 to $k'-1$.

Finally, following the same argumentation, the last term is equal to $\frac{1}{2}\mathbb{P}\left(P_c(t) = \{K, \ell\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right)\Delta_\ell$, where $\Delta_\ell := \rho(\boldsymbol{a}^*, K) - \rho((K, \ell) \circ \boldsymbol{a}^*, K)$.

Overall

$$\begin{aligned}
R_t^* &= \sum_{k=2}^{K} \frac{1}{2} \mathbb{P}\left(P_c(t) = \{k-1, k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \Delta_k \\
&\quad + \sum_{\ell=K+1}^{L} \frac{1}{2} \mathbb{P}\left(P_c(t) = \{K, \ell\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \Delta_\ell \\
&= \sum_{k=2}^{L} \frac{1}{2} \mathbb{P}\left(P_c(t) = \{\min(k-1, K), k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \Delta_k .
\end{aligned}$$

Let finally upper-bound the overall regret.

$$\begin{aligned}
R(T) &= \sum_{t=1}^{T} \mu^* - \mathbb{E}_{\boldsymbol{a}(t)}\left[\mu_{\boldsymbol{a}(t)}\right] \\
&= \sum_{t=1}^{T} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) \neq \boldsymbol{P}^*\right) \left(\mu^* - \mathbb{E}_{\boldsymbol{a}(t)}\left[\mu_{\boldsymbol{a}(t)} \mid \tilde{\boldsymbol{P}}(t) \neq \boldsymbol{P}^*\right]\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \left(\mu^* - \mathbb{E}_{\boldsymbol{a}(t)}\left[\mu_{\boldsymbol{a}(t)} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right]\right) \\
&\leqslant \sum_{t=1}^{T} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) \neq \boldsymbol{P}^*\right) K \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \sum_{k=2}^{L} \frac{1}{2} \mathbb{P}\left(P_c(t) = \{\min(k-1, K), k\} \mid \tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*\right) \Delta_k \\
&\leqslant \mathcal{O}\left(\log\log T\right) \\
&\quad + \sum_{t=1}^{T} \sum_{k=2}^{L} \frac{1}{2} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*, P_c(t) = \{\min(k-1, K), k, \}\right) \Delta_k \\
&= \mathcal{O}\left(\log\log T\right) \\
&\quad + \sum_{k=2}^{L} \frac{\Delta_k}{2} \sum_{t=1}^{T} \mathbb{P}\left(\tilde{\boldsymbol{P}}(t) = \boldsymbol{P}^*, P_c(t) = \{\min(k-1, K), k, \}\right) \\
&\leqslant \mathcal{O}\left(\log\log T\right) \\
&\quad + \sum_{k=2}^{L} \frac{\Delta_k}{2} \left(\frac{16}{\tilde{\delta}_k^* \tilde{\Delta}_k^2} \log T + \mathcal{O}\left(\log\log T\right)\right) \\
&= \sum_{k=2}^{L} \frac{8\Delta_k}{\tilde{\delta}_k^* \tilde{\Delta}_k^2} \log T + \mathcal{O}\left(\log\log T\right) \\
&= \mathcal{O}\left(\frac{L}{\Delta} \log T\right),
\end{aligned}$$

where for any index $k \geqslant 2$

$$\tilde{\Delta}_k := \tilde{\Delta}_{\min(k-1,K),k} \qquad \textit{and} \qquad \Delta := \min_{k \in \{2,\dots,K\}} \frac{\tilde{\delta}_k^* \tilde{\Delta}_k^2}{8\Delta_k},$$

which concludes the proof.

□

G. UniRank's Theoretical Results While Facing State-of-the-Art Click Models

Here, we prove Corollaries 5.2 and 5.3 and then discuss the relationship between our upper-bounds and the known lower bounds.

G.1. Proof of Corollary 5.2 (Upper-Bound on the Regret of UniRank when Facing CM* Click Model)

Corollary G.1 is a more precise version of Corollary 5.2. Its proof consists in identifying the gaps $\tilde{\delta}_k^*$, $\tilde{\Delta}_k$, and Δ_k, where k is the index of an item.

Corollary G.1 (Facing CM* click model)**.** *Under the hypotheses of Theorem 5.1, if the user follows CM with probability θ_i to click on item i when it is observed, then for any index $k \geqslant 2$,*

$$\begin{aligned}
\tilde{\delta}_k^* &= (\theta_{k-1} + \theta_k - \theta_{k-1}\theta_k) \prod_{\ell=1}^{k-2} (1 - \theta_\ell) && \textit{if } k \leqslant K, \\
\tilde{\delta}_k^* &= \frac{1}{2} (\theta_K + \theta_k) \prod_{\ell=1}^{K-1} (1 - \theta_\ell) && \textit{if } k \geqslant K+1, \\
\tilde{\Delta}_k &\geqslant \frac{\theta_{\min(K,k-1)} - \theta_k}{\theta_{\min(K,k-1)} + \theta_k}, \\
\Delta_k &= 0 && \textit{if } k \leqslant K, \\
\Delta_k &= (\theta_K - \theta_k) \prod_{\ell=1}^{K-1} (1 - \theta_\ell) && \textit{if } k \geqslant K+1.
\end{aligned}$$

Hence, UniRank fulfills

$$\begin{aligned}
R(T) &\leqslant \sum_{k=K+1}^{L} 16 \frac{\theta_K + \theta_k}{\theta_K - \theta_k} \log T + \mathcal{O}(\log \log T) \\
&= \mathcal{O}\left((L-K) \frac{\theta_K + \theta_{K+1}}{\theta_K - \theta_{K+1}} \log T \right).
\end{aligned}$$

Proof of Corollary G.1. Values $\tilde{\delta}_k^*$ and Δ_k derive from a straightforward computation given CM model.

Let us prove the lower-bound on $\tilde{\Delta}_k$. Let i and j be two items such that $i \neq j$. Let $\boldsymbol{a}$ be a recommendation such that $\mathbb{P}(c_i(t) \neq c_j(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}) > 0$.

Without loss of generality, assume i appears in $\boldsymbol{a}$ in position k, and if j appears in $\boldsymbol{a}$, it is in a position $\ell > k$. Then

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{A \frac{1+B}{2} (\theta_i - \theta_j)}{A \frac{1+B}{2} (\theta_i + \theta_j) - AB\theta_i\theta_j} \geqslant \frac{\theta_i - \theta_j}{\theta_i + \theta_j},$$

with $A := \prod_{c=1}^{k-1} (1 - \theta_{a_c})$ and $B := \prod_{c=k+1}^{\ell-1} (1 - \theta_{a_c})$ if j appears in $\boldsymbol{a}$ and 0 otherwise.

Hence the lower-bounding values for $\tilde{\Delta}_k$, by noting that the term A is lower-bounded by $\prod_{\ell=1}^{K-1} (1 - \theta_\ell)$.

Regarding the last formula in Lemma G.1, it derives from the fact that $\frac{\theta_K + \theta_k}{\theta_K - \theta_k}$ is maximized when θ_k is maximized, meaning $k = K+1$. □

G.2. Proof of Corollary 5.3 (Upper-Bound on the Regret of UniRank when Facing PBM* Click Model)

Corollary G.2 is a more precise version of Corollary 5.3. Its proof consists in identifying the gaps $\tilde{\delta}_k^*$, $\tilde{\Delta}_k$, and Δ_k, where k is the index of an item.

Corollary G.2 (Facing PBM* click model)**.** *Under the hypotheses of Theorem 5.1, if the user follows PBM with the probability θ_i of clicking on item i when it is observed and the probability κ_k of observing the position k, then for any index*

$k \geqslant 2$,

$$\begin{aligned}
\tilde{\delta}_k^* &= \frac{1}{2}\left(\theta_{k-1}+\theta_k\right)\left(\kappa_{k-1}+\kappa_k\right) - 2\theta_{k-1}\theta_k\kappa_{k-1}\kappa_k && \textit{if } k \leqslant K,\\
\tilde{\delta}_k^* &= \frac{1}{2}\left(\theta_K+\theta_k\right)\kappa_K && \textit{if } k \geqslant K+1,\\
\tilde{\Delta}_k &\geqslant \frac{\theta_{\min(K,k-1)}-\theta_k}{\theta_{\min(K,k-1)}+\theta_k}, &&\\
\Delta_k &= \left(\theta_{k-1}-\theta_k\right)\left(\kappa_{k-1}-\kappa_k\right) && \textit{if } k \leqslant K,\\
\Delta_k &= \left(\theta_K-\theta_k\right)\kappa_K && \textit{if } k \geqslant K+1.
\end{aligned}$$

Hence, UniRank fulfills

$$\begin{aligned}
R(T) &\leqslant \sum_{k=2}^{K} \frac{8(\kappa_{k-1}-\kappa_k)(\theta_{k-1}+\theta_k)^2}{\tilde{\delta}_k^*(\theta_{k-1}-\theta_k)}\log T + \sum_{k=K+1}^{L} 16\frac{\theta_K+\theta_k}{\theta_K-\theta_k}\log T + \mathcal{O}\left(\log\log T\right)\\
&= \mathcal{O}\left(\frac{L}{\Delta}\log T\right),
\end{aligned}$$

where $\Delta := \min\{\min_{k\in\{2,\dots,K\}} \frac{\tilde{\delta}_k^*(\theta_{k-1}-\theta_k)}{(\kappa_{k-1}-\kappa_k)(\theta_{k-1}+\theta_k)^2}, \min_{k\in\{K+1,\dots,L\}} \frac{\theta_K-\theta_k}{\theta_K+\theta_k}\}$.

Proof of Corollary G.2. Values $\tilde{\delta}_k^*$ and Δ_k derive from a straightforward computation given PBM model.

Let us prove the lower-bound on $\tilde{\Delta}_k$. Let i and j be two items such that $i \neq j$. Let $\boldsymbol{a}$ be a recommendation such that $\mathbb{P}(c_i(t) \neq c_j(t) \mid \boldsymbol{a}(t) = \boldsymbol{a}) > 0$.

If both i and j appear in $\boldsymbol{a}$, denote $k < \ell$ these positions. Then

$$\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\frac{1}{2}(\kappa_k+\kappa_\ell)(\theta_i-\theta_j)}{\frac{1}{2}(\kappa_k+\kappa_\ell)(\theta_i+\theta_j) - 2\kappa_k\kappa_\ell\theta_i\theta_j} \geqslant \frac{\theta_i-\theta_j}{\theta_i+\theta_j}.$$

If only one of both items i and j appears in $\boldsymbol{a}$ then $\tilde{\Delta}_{i,j}(\boldsymbol{a}) = \frac{\theta_i-\theta_j}{\theta_i+\theta_j}$.

Hence for any index $k \geqslant 2$, $\tilde{\Delta}_k \geqslant \frac{\theta_{\min(K,k-1)}-\theta_k}{\theta_{\min(K,k-1)}+\theta_k}$. □