
Combinatorial Bandits with Plackett-Luce Feedback: A Worst-Case Analysis

Cristiano Migali
Politecnico di Milano
cristiano.migali@polimi.it

Gianmarco Genalti
Politecnico di Milano
gianmarco.genalti@polimi.it

Alberto Maria Metelli
Politecnico di Milano
albertomaria.metelli@polimi.it

Marco Mussi
Politecnico di Milano
marco.mussi@polimi.it

Abstract

Combinatorial bandits with *ranking feedback* model a sequential decision-making problem in which the learner observes a *top- m* ranking of the set of k arms played in each round. The setting has been examined under the lens of *top- k regret minimization*, which accounts for the cost of pulling arms that are not among the best k . Existing works rely on the assumption that latent rankings are generated by a *Plackett-Luce* (PL) distribution and consider the special case of full-ranking feedback ($m = k$). They provide instance-dependent regret upper bounds which match the asymptotic logarithmic scaling in the learning horizon T , but suffer from a burn-in term which becomes $\Omega(T)$ for some choices of the PL parameters, preventing the derivation of sublinear worst-case bounds. In this work, we study the setting under the general top- m feedback ($m \leq k$) and provide a worst-case regret lower bound of order $\Omega(\sqrt{T})$. Then, by introducing a novel algorithmic strategy, we derive an instance-dependent regret upper bound which does not suffer from the exploding burn-in term and a corresponding worst-case bound of order $\tilde{O}(\sqrt{T})$, proving for the first time that it is possible to achieve sublinear worst-case regret w.r.t. the PL parameters. Moreover, we translate our algorithmic ideas to *multinomial logit* bandits, where the learner receives *winner feedback* ($m = 1$) and there is a non-zero probability of observing “no-choice”. The existing regret bounds suffer from an exploding burn-in term, inversely proportional to the no-choice probability, that we avoid through our novel approach.

1 Introduction

The latest breakthrough in the application of *reinforcement learning from human feedback* for the training of *large language models* produced a growing interest—both practical and theoretical—in the field of learning from human feedback. Human feedback is relative in nature. When provided with a set of options, it is difficult for a human to assign a numerical value to each option that reflects its intrinsic value. At the same time, a human can easily assess their quality by stating which option is preferred to which other. This motivates the growing body of literature in the area of learning from preferences, which encompasses fields like *learning-to-rank* (Liu, 2009), *dueling bandits* (Yue et al., 2009), or *preference-based reinforcement learning* (Wirth et al., 2017). This work studies a rather fundamental sequential decision-making problem in this context: a learner provides a set (*assortment*) of k different options among n that are available and observes a partial ranking over the provided options as feedback. The ranking is partial in the sense that the learner observes the first m options of the ranking only (i.e., *top- m feedback*, Saha and Gopalan, 2019a). The goal is to identify the set of

Table 1: Existing and new bounds for CMABs with ranking feedback and MNL bandits. The “Feed-back” column takes values “Full-Ranking” (i.e., $m = k$), “Winner w/ No Choice” (i.e., $m = 1$ with the fictitious item representing no-choice), “Top- m ” (i.e., the general case $m \leq k$). The “Objective” column takes values “Top- k ” (i.e., top- k objective) or “Weighted” (i.e., weighted objective).

Paper	Feedback		Objective			Instance-Dependent Upper Bound	Worst-Case Upper Bound	Sublinear for Worst-Case $\theta?$
	Full-Ranking	Winner w/ No Choice	Top- m	Top- k	Weighted			
Saha and Gopalan (2019b)	✓	✗	✗	✓	✗	$\tilde{O}\left(\frac{1}{k} \sum_{v \in [k+1, n]} \frac{1}{\Delta_{k,v}} \log T + \Delta'_{\max} \sum_{u \in [k]} \sum_{v \in [k+1, n]} \frac{\theta_u^2}{\Delta_{k,v}^2}\right)$	N/A	✗
Agrawal et al. (2019)	✗	✓	✗	✗	✓	N/A	$\tilde{O}\left(\sqrt{\theta_0^{-1} n T}\right)$	✗
Saha and Gaillard (2025)	✗	✓	✗	✓	✓	N/A	$\tilde{O}\left(\sqrt{\min\{\theta_0^{-1}, n\} n T} + n^2 k^2 \theta_0^{-2}\right)$	✗
Ours	✓	✓	✓	✓	✗	$\mathcal{O}\left(\frac{n}{m} \sum_{u \in [k+1, n]} \frac{1}{\Delta_{k,u}} \log T + \frac{n^2 k}{m} \log T\right)$	$\tilde{O}\left(\sqrt{\frac{n-k}{m} T}\right)$	✓

k best options. This setting takes the name of *Combinatorial Multi-Armed Bandits* (CMABs) with *ranking feedback*.¹

The problem has been extensively studied considering specific *distributions* of the *latent ranking* of the assortment, *feedback models*, i.e., the portion of the ranking observed by the learner, and *performance indices*. Regarding the *distribution* of the *latent ranking*, the ubiquitous assumption in the literature—which we also adopt in this work—is that the latent ranking is distributed according to a *Plackett-Luce* (PL) distribution (Luce, 1959; Plackett, 1975) which associates a parameter $\theta_i \in (0, 1]$ with each option $i \in [n]$, with $\boldsymbol{\theta} = (\theta_i)_{i \in [n]}$.² The parameter is proportional to the probability that the associated option appears in the top positions of the latent ranking. The studied *feedback models* instead include *full-ranking feedback* (e.g., Saha and Gopalan, 2019b), where the learner directly observes the latent ranking among items, and *winner feedback*, where instead only the best option of the assortment is observed, modeling the choice of one of the options in the assortment by a user. In particular, the literature on *Multinomial Logit* (MNL) bandits (Agrawal et al., 2019; Saha and Gaillard, 2025) considers *winner feedback* with “no-choice”, i.e., there is a non-zero probability that neither of the options is chosen from the assortment. Under the PL distribution assumption, this can be modeled through the addition of a fictitious option $i = 0$, associated with a parameter $\theta_0 \in (0, 1]$. Observe that both feedback models can be regarded as special cases of the top- m feedback, where $m = k$ for full-ranking feedback, and $m = 1$ for winner feedback. Finally, different *performance indices* are used in the literature to define the best assortment. Some works (e.g. Saha and Gopalan, 2019b) associate with each assortment $S \subseteq [n]$, $|S| = k$, a reward which corresponds to the average of the PL parameters: $g_{\boldsymbol{\theta}}^{(k)}(S) := 1/k \sum_{i \in S} \theta_i$ (see also *top- k objective*, Saha and Gaillard, 2025). Other works in the MNL bandits literature (Agrawal et al., 2019; Saha and Gaillard, 2025) consider instead a generalized MNL of reward, where each option is associated with a revenue value (see *weighted objective*, Saha and Gaillard, 2025). Both definitions of reward naturally induce a notion of regret. In particular, the notion of reward used in (Saha and Gopalan, 2019b) induces the so-called *top- k regret*, which is the one we study in this paper. Despite the existence of several works in this area, some fundamental questions still remain open. In particular, in Table 1, we summarize the regret upper bounds presented in the literature relevant to the setting under study. For the sake of clarity, we assume $\theta_1 > \dots > \theta_n$, i.e., that the labels reflect the true ordering of the PL parameters, and define $\Delta_{u,v} := \theta_u - \theta_v$ for all $u, v \in [n]$, $\Delta'_{\max} := 1/k (\sum_{u \in [k]} \theta_u - \sum_{u \in [n-k+1, n]} \theta_u)$. Saha and Gopalan (2019b) provide an instance-dependent bound on the top- k regret which is composed of two additive terms. The first term has logarithmic dependence on T and scales with the sum of the reciprocals of the gaps $\sum_{v \in [k+1, n]} \frac{1}{\Delta_{k,v}}$ as in stochastic *Multi-Armed Bandits* (MABs, Lattimore and Szepesvári, 2020). By following standard techniques of the MAB literature (e.g., Theorem 7.2, Lattimore and Szepesvári, 2020), it is possible to translate this term into a sublinear worst-case

¹We use the terms *ranking* and *relative feedback* interchangeably to refer to the same feedback mechanism.

²For $a, b \in \mathbb{N}_{\geq 0}$, we denote $\llbracket a, b \rrbracket := \{a, \dots, b\}$ and $\llbracket a \rrbracket := \llbracket 1, a \rrbracket$. When $b < a$, $\llbracket a, b \rrbracket = \{\}$.

bound.³ The issue with this result arises from the additional burn-in term: $\Delta'_{\max} \sum_{u \in \llbracket k \rrbracket} \sum_{v \in \llbracket k+1, n \rrbracket} \frac{\theta_u^2}{\Delta_{u,v}^2}$.

The numerator $\Delta'_{\max}$ accounts for the maximum instantaneous regret that the learner can incur in a round, while the terms $1/\Delta_{u,v}^2$ account for the sample complexity of telling apart arms u and v , $u \in \llbracket k \rrbracket$, $v \in \llbracket k+1, n \rrbracket$, dominated by $1/\Delta_{k,k+1}^2$. Such a burn-in term does not allow us to derive meaningful worst-case upper bounds, as it is possible to construct instances where $\Delta'_{\max}$ is fixed while $\Delta_{k,k+1} \rightarrow 0$, resulting in a linear regret bound. To better understand the problem, observe that, in a stochastic MAB, the burn-in term is upper bounded by an $\mathcal{O}(n)$ term (Lattimore and Szepesvári, 2020, Theorem 7.1). The other works (Agrawal et al., 2019; Saha and Gaillard, 2025) consider instead the MNL bandit setting. In this case, the bounds scale with θ_0^{-1} and become vacuous for $\theta_0 \rightarrow 0$. The connections with this setting are discussed in greater detail in Section 5. In both cases, it is not possible to derive sublinear worst-case regret bounds w.r.t. the PL parameters, motivating the following research question.

Research Question: Is it possible to derive **worst-case** upper bounds on the top- k regret of combinatorial bandits with ranking feedback which are **sublinear** in T under *any* choice of the PL parameters θ and general top- m feedback ($m \leq k$)?

Original Contribution. The contribution of this work is summarized as follows:

- In Section 3, we study the statistical complexity of the setting, deriving a worst-case lower bound on the top- k regret of order $\Omega(\sqrt{T})$ for $m \leq k/2$ (Theorem 1), suggesting that a worst-case sublinear regret bound could be achievable.
- In Section 4, we introduce the TOP- k -UIF algorithm and derive both an instance-dependent and a worst-case upper bound on its regret under general top- m feedback (Theorem 2), which are summarized in Table 1, where the first avoids the exploding burn-in term of (Saha and Gopalan, 2019b) and the latter has order $\tilde{\mathcal{O}}(\sqrt{T})$, proving for the first time that it is possible to achieve sublinear regret under *any* choice of the PL parameters and providing an affirmative answer to the research question. Furthermore, the TOP- k -UIF algorithm (Algorithm 1) relies on a uniform sampling scheme which requires ad hoc concentration results for the number of observed comparisons for each arm–pivot pair (see Lemma 7) and proof strategies for the regret upper bound (see Lemma 13) which are of independent interest.⁴
- In Section 5, we show how the ideas of the TOP- k -UIF algorithm can be seamlessly translated to MNL bandits, introducing the MNL-TOP- k -UIF algorithm (Algorithm 3) and showing that, also in this setting, it is possible to achieve a sublinear regret bound of order $\tilde{\mathcal{O}}(\sqrt{T})$ (Theorem 3), which holds for any value of the no-choice parameter θ_0 .

The rest of the manuscript is organized as follows: in Section 2, we formally introduce the setting; in Section 6, we provide numerical simulations supporting the validity of our theoretical results. Related works are discussed in Appendix A. The analyses of the regret lower and upper bounds are deferred to Appendices B and D, respectively.

2 Setting

A *CMAB with ranking feedback* (Saha and Gopalan, 2019b) is defined as a tuple $\nu = (n, \theta, k, m)$ where $n \in \mathbb{N}_{\geq 2}$ is the number of arms, $\theta \in (0, +\infty)^n$ is the vector of PL parameters, $k \in \llbracket 2, n \rrbracket$ is the cardinality of the set pulled in each round by the learner, and $m \in \llbracket k \rrbracket$ refers to the top- m feedback. Furthermore, as in (Saha and Gopalan, 2019b), we require the existence of a set of top- k arms $S_{\theta}^{(k)*} \subseteq \llbracket n \rrbracket$ ($|S_{\theta}^{(k)*}| = k$) such that:

$$\theta_i > \theta_j \text{ for every } i \in S_{\theta}^{(k)*}, j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}, \quad (1)$$

with the meaning that the top- k arms $S_{\theta}^{(k)*}$ are strictly better than the others according to the PL parameters. In each round, $t \in \mathbb{N}_{\geq 1}$, the learner:

³To be precise, the sublinear worst-case bound requires the derivation of a refined bound of the form $1/k \sum_{v \in \llbracket k+1, n \rrbracket} \min\{1/\Delta_{k,v} \log T, T\Delta_{k,v}\}$, which should be achievable through an adjustment of the analysis.

⁴For example, they can be used to prove a tight regret bound for a version of the SUCCESSIVE ELIMINATION algorithm (Even-Dar et al., 2006) with uniform sampling of alive arms instead of round-robin.

1. Pulls a set $S_t \subseteq \llbracket n \rrbracket$ with fixed cardinality $|S_t| = k$.
2. Observes a top- m ranking of the pulled arms. In particular, given a bijective function $R_t : S_t \rightarrow \llbracket k \rrbracket$ which ranks the arms from 1 to k , the observation consists of $R_t^{(m)} : S_t \rightarrow \llbracket \min\{k, m+1\} \rrbracket$ defined as $R_t^{(m)}(i) := \min\{R_t(i), m+1\}$ for all $i \in S_t$.
3. Gains expected top- k reward $g_{\theta}^{(k)}(S_t) = 1/k \sum_{i \in S_t} \theta_i$.

The latent ranking R_t of the pulled arms is generated according to the PL distribution. Given a finite non-empty set S and a tuple $(\theta_i)_{i \in S} \in (0, +\infty)^{|S|}$, a PL distribution over the set S with parameters $(\theta_i)_{i \in S}$ is defined as $\text{PL}((\theta_i)_{i \in S}) \in \Delta(\text{Bij}(S, \llbracket |S| \rrbracket))$ such that:⁵

$$\mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})}[R = r] := \prod_{u=1}^{|S|-1} \frac{\theta_{r^{-1}(u)}}{\sum_{v=u}^{|S|} \theta_{r^{-1}(v)}} \text{ for every } r \in \text{Bij}(S, \llbracket |S| \rrbracket).$$

Thus, $R_t \sim \text{PL}((\theta_i)_{i \in S_t})$ for every round $t \in \mathbb{N}_{\geq 1}$. Since the PL distribution is scale invariant, in what follows, analogously to what is done in (Saha and Gopalan, 2019b), we assume w.l.o.g. that $\max_{i \in \llbracket n \rrbracket} \theta_i = 1$. We denote by $\mathcal{E}_n^{(k,m)}$ the set of CMABs with ranking feedback for fixed $n, k \in \mathbb{N}_{\geq 2}, m \in \mathbb{N}_{\geq 1}$ with $m \leq k \leq n$ and PL parameters θ satisfying Equation (1) and $\max_{i \in \llbracket n \rrbracket} \theta_i = 1$. Finally, we use $r_{\theta}^* \in \text{Bij}(\llbracket n \rrbracket)$ for a bijection which orders the PL parameters of the corresponding instance $\nu = (n, \theta, k, m) \in \mathcal{E}_n^{(k,m)}$: $1 = \theta_{r_{\theta}^{*-1}(1)} \geq \dots \geq \theta_{r_{\theta}^{*-1}(n)} > 0$.

Learning Problem. The goal of the agent is to learn the set of top- k arms $S_{\theta}^{(k)*}$. Observe that $S_{\theta}^{(k)*} = \arg \max_{\substack{S \subseteq \llbracket n \rrbracket \\ |S|=k}} g_{\theta}^{(k)}(S)$. We now formally introduce a *policy* describing the behavior of

the agent and a corresponding *performance index* aligned with the learning objective. For every round $t \in \mathbb{N}_{\geq 1}$, we define the history prior to the agent's action as $H_{t-1} := ((U_s, S_s, R_s^{(m)})_{s=1}^{t-1}, U_t)$. Here, U_t serves as a source of independent internal randomness for the policy at round t (e.g., $U_t \sim \text{Unif}([0, 1])$).⁶ A *randomized policy* π assigns a conditional probability distribution to each possible history to determine the set of arms to pull: $S_t \sim \pi(\cdot | H_{t-1})$, where $S_t \in \{S \subseteq \llbracket n \rrbracket \text{ s.t. } |S| = k\}$. Given a learning horizon $T \in \mathbb{N}_{\geq 1}$, the performance of a policy π on an instance $\nu \in \mathcal{E}_n^{(k,m)}$ is evaluated according to its *expected top- k regret* (Saha and Gopalan, 2019b):

$$\text{EReg}_{\pi, \nu}^{(k)}(T) := \mathbb{E}_{\pi, \nu} \left[\sum_{t=1}^T \left(g_{\theta}^{(k)}(S_{\theta}^{(k)*}) - g_{\theta}^{(k)}(S_t) \right) \right],$$

where we use $\mathbb{E}_{\pi, \nu}[\cdot]$ and $\mathbb{P}_{\pi, \nu}[\cdot]$ to denote the expectation and the corresponding probability measure, respectively, which govern the stochastic process induced by the interaction between the randomized policy π and the instance ν .

3 Top- k Regret Lower Bound

In this section, we study the statistical complexity of the learning problem by providing a worst-case lower bound on the expected top- k regret under top- m feedback with $m \leq k/2$, which is presented below (Theorem 1). The formal derivation is deferred to Appendix B.

Theorem 1 (Worst-Case Lower Bound to the Top- k Regret). *For every $n, k \in \mathbb{N}_{\geq 2}, m \in \mathbb{N}_{\geq 1}$ with $m \leq k/2 \leq n/2$ and $n/k \in \mathbb{N}_{\geq 1}$, randomized policy π and learning horizon $T \in \mathbb{N}_{\geq 1}$ satisfying $T \geq \frac{1}{16} \frac{(n-k)^2}{nmk}$, it holds:*

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \text{EReg}_{\pi, \nu}^{(k)}(T) \geq C_1 \left(1 - \frac{k}{n} \right)^{3/2} \sqrt{\frac{n-k}{km}} T,$$

where $C_1 \in \mathbb{R}_{>0}$ is an absolute constant.

⁵We denote by $\text{Bij}(A, B)$ the set of bijective functions which map set A onto set B . We write $\text{Bij}(A)$ when $A = B$. We denote by $\Delta(\mathcal{X}, \mathcal{F})$ the set of probability measures defined over the measurable space $(\mathcal{X}, \mathcal{F})$. We write $\Delta(\mathcal{X})$ when the σ -algebra $\mathcal{F}$ is clear from the context.

⁶We denote by $\text{Unif}(\mathcal{X}) \in \Delta(\mathcal{X})$ the uniform distribution with support $\mathcal{X}$ which is either a finite set or a bounded subset of $\mathbb{R}$.

Proof Sketch. The first ingredient for the lower bound requires the derivation of a decomposition of the *Kullback-Leibler divergence* (KL, Kullback and Leibler, 1951) of the histories induced by two instances $\nu, \nu' \in \mathcal{E}_n^{(k,m)}$ for the same randomized policy π . We provide this in Lemma 4. In particular, by exploiting the properties of the PL distribution, we obtain a decomposition which is analogous to the one of (Lattimore and Szepesvári, 2020, Lemma 15.1), used for stochastic MABs, but holds under top- m feedback in a combinatorial setting when the PL parameters belong to $[1/2, 1]$. The proof then follows by standard lower-bound techniques from the MAB literature. In particular, we define a base instance $\nu^{(0,\epsilon)}$ whose top- k set is strictly separated by $\epsilon > 0$ and alternative instances $\nu^{(S)}$ which promote $k - 1$ arms in addition to a common best arm. We apply Pinsker’s inequality, let $\epsilon \rightarrow 0$, and finally optimize the value of Δ . $\square$

Observe that the lower bound is sublinear in the learning horizon T , suggesting that it is possible—as we prove in Section 4—to design algorithmic strategies that enjoy sublinear regret on every instance $\nu \in \mathcal{E}_n^{(k,m)}$. It is interesting to notice that, as expected, the complexity of the learning problem scales inversely with the informativeness of the feedback m . Furthermore, the lower bound depends on the difference $n - k$, instead of the usual scaling with plain n for stochastic MABs. This is due to the definition of the top- k objective. Indeed, the goal of the learner is to identify the set of k best arms. When $k = n$, the only subset of $\llbracket n \rrbracket$ of cardinality k is $\llbracket n \rrbracket$ itself, trivializing the learning problem.

4 The TOP- k -UIF Algorithm

In this section, we introduce the TOP- k -UIF algorithm (Top- k Uniform Interleaved Filter, Algorithm 1) for CMABs with ranking feedback and provide both an instance-dependent and a worst-case upper bound on its expected top- k regret (Theorem 2). The additional routines which are invoked in Algorithm 1 are presented and described in Appendix C.

Algorithmic Rationale. The algorithm is inspired by the INTERLEAVED FILTER 1 (IF1) algorithm (Yue et al., 2009) originally designed for dueling bandits. The algorithmic design revolves around the following idea: in order to exploit ranking feedback, we rely on a *pivotal arm* (P_t in the algorithm) to compare all the others. In the dueling bandits setting, there is a natural way of implementing this approach: the learner observes duels between the pivotal arm and any other arm and keeps track of the empirical frequency with which each arm beats the pivot. Then, arms are compared according to such frequency. We generalize the idea to the combinatorial setting with top- m feedback by means of *rank-breaking* (Khetan and Oh, 2016). Given a sequence of top- m rankings, the rank-breaking estimator is computed by tracking the following two quantities: the number of times an arm $i \in S_t$ and the pivotal arm P_t have different top- m ranks (i.e., $R_t^{(m)}(i) \neq R_t^{(m)}(P_t)$), and the number of times arm $i \in S_t$ precedes the pivotal arm in the top- m ranking (i.e., $R_t^{(m)}(i) < R_t^{(m)}(P_t)$). In particular, for an arm i and a pivot p , the rank-breaking estimator is informally the ratio:

$$\hat{\mu}_i^{(p)} := \frac{\#\{\text{observed } i\text{-vs-}p \text{ comparisons won by } i\}}{\#\{\text{observed } i\text{-vs-}p \text{ comparisons}\}}.$$

The rationale is that a partial ranking is broken into a list of pairwise comparisons $i \prec j$ whenever $R_t^{(m)}(i) < R_t^{(m)}(j)$, which then is processed exactly as in the dueling bandits case, that is, by computing the frequency of pairwise comparisons in which arm i beats the pivot P_t . The RB-UPDATE routine updates the estimates (Lines 9-13) exactly in this way. Under the PL distribution, such an estimator concentrates around a surprisingly well-behaved quantity (Lemma 5):

$$\mu_i^{(P_t)} := \frac{\theta_i}{\theta_i + \theta_{P_t}}.$$

Observe that, for every $P_t \in \llbracket n \rrbracket$, $\mu_i^{(P_t)}$ is increasing in θ_i , allowing the comparison of arms.

Now, as in the IF1 algorithm, we need to face the following challenges:

1. *Identification of suboptimal arms $j \notin S_\theta^{(k)*}$.* The IF1 algorithm eliminates all arms which are worse than the pivot with statistical confidence. Adapting this approach to the combinatorial setting requires care. The TOP- k -UIF algorithm keeps track of the following disjoint sets: $\mathcal{U}_t$ and $\mathcal{A}_t$, with $\mathcal{U}_t \cup \mathcal{A}_t \subseteq \llbracket n \rrbracket$. $\mathcal{U}_t$ is the set of *unknown arms*, i.e., the set of arms for which we

Algorithm 1 TOP- k -UIF (Top- k Uniform Interleaved Filter).

```
1: Input:  $T \in \mathbb{N}_{\geq 1}$ ,  $n \in \mathbb{N}_{\geq 2}$ ,  $k \in \llbracket 2, n \rrbracket$ ,  $m \in \llbracket k \rrbracket$ ,  $\beta_{\delta, n, T} : [0, +\infty] \rightarrow [0, +\infty]$ ,  $\delta \in (0, 1]$ 
2: Initialize  $P_1 \sim \text{Unif}(\llbracket n \rrbracket)$ ,  $\mathcal{A}_1 \leftarrow \{\}$ ,  $\mathcal{U}_1 \leftarrow \llbracket n \rrbracket$ ,
    $N_i^{(p)} \leftarrow 0$  if  $i \neq p$ ,  $+\infty$  otherwise,
    $\hat{\mu}_i^{(p)} \leftarrow 1/2$  for every  $i, p \in \llbracket n \rrbracket$ 
3: for  $t \in \llbracket T \rrbracket$  do
4:    $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t$ ,  $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_t$ ,  $P_{t+1} \leftarrow P_t$ 
5:   Let  $\mathcal{P}_t \leftarrow \mathcal{A}_t \cup \{P_t\}$ 
6:   Sample  $S_{\text{exp}, t} \sim \text{Unif}(\{S \subseteq \mathcal{U}_t \setminus \{P_t\} \text{ s.t. } |S| = k - |\mathcal{P}_t|\})$ 
7:   Pull  $S_t \leftarrow \mathcal{P}_t \cup S_{\text{exp}, t}$ 
8:   Observe  $R_t^{(m)} : S_t \rightarrow \llbracket \min\{k, m+1\} \rrbracket$ 
   // 1. Update Statistics & Confidence Bounds
9:    $(\hat{\mu}_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t} \leftarrow \text{RB-UPDATE}(\mathcal{P}_t, S_t, (\hat{\mu}_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, R_t^{(m)})$ 
10:  for  $i \in \mathcal{U}_t, p \in \mathcal{P}_t$  do
11:     $\text{UCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} + \beta_{\delta, n, T}(N_i^{(p)})$ 
12:     $\text{LCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} - \beta_{\delta, n, T}(N_i^{(p)})$ 
13:  end for
   // 2. Update Unknown Arms
14:   $(A_i)_{i \in \mathcal{U}_t}, (B_i)_{i \in \mathcal{U}_t} \leftarrow \text{COUNT-BA}(\mathcal{P}_t, \mathcal{U}_t, (\text{UCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t}, (\text{LCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t})$ 
15:  for  $i \in \mathcal{U}_t$  do
16:    if  $B_i \geq k - |\mathcal{A}_t|$  then ▷ Reject arm
17:       $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
18:    else if  $A_i \geq |\mathcal{A}_t \cup \mathcal{U}_t| - k$  then ▷ Accept arm
19:       $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
20:    end if
21:  end for
   // 3. Pivot Update
22:  if  $\mathcal{A}_t \neq \{\}$  then ▷ Stop pivot update
23:    for  $i \in \mathcal{U}_t$  do
24:      if  $\text{LCB}_i^{(P_t)} > 1/2$  then ▷ Accept  $\theta_i > \theta_{P_t}$ 
25:         $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
26:      end if
27:    end for
28:  else
29:     $p' \leftarrow \text{NEW-PIVOT}(P_t, \mathcal{U}_t, (\text{LCB}_i^{(P_t)})_{i \in \mathcal{U}_t})$ 
30:    if  $p' \neq \perp$  then
31:       $P_{t+1} \leftarrow p'$ 
32:    end if
33:  end if
34:  if  $|\mathcal{A}_{t+1}| = k$  then  $\mathcal{U}_{t+1} \leftarrow \{\}$ 
35: end for
```

cannot say with statistical confidence if they are optimal or not. $\mathcal{A}_t$ is the set of *accepted arms*, i.e., the set of arms for which we can say with statistical confidence that they are optimal. The first issue in translating the elimination approach from the dueling to the combinatorial setting is that, in the latter, the fact that an arm is statistically better than another is not sufficient to eliminate the inferior arm. Indeed, there are k arms in the optimal set. The issue is solved by the COUNT-BA routine (Line 14). The routine defines $i \prec j$ whenever an estimate relative to some pivotal arm $p \in \mathcal{P}_t$ —where the set of *pivotal arms* $\mathcal{P}_t$ includes both the pivot P_t and the accepted arms $\mathcal{A}_t$ —certifies that i is better than j , and then computes the transitive closure $\prec^+$. For each unknown arm $i \in \mathcal{U}_t$, B_i counts the unknown arms which precede i under $\prec^+$, while A_i counts those which follow i . The number of optimal arms which are still in the unknown set is $k - |\mathcal{A}_t|$. If there are $B_i \geq k - |\mathcal{A}_t|$ arms that are statistically better than i , then i is eliminated from $\mathcal{U}_t$ (Line 17) as it cannot be optimal.

2. *Exploration vs exploitation.* The IF1 algorithm handles the exploration by playing the pivotal arm in each round and the unknown arms in a round-robin fashion. TOP- k -UIF takes a slightly different route to adapt the approach to the combinatorial setting: in each round it pulls the pivot P_t and the accepted arms $\mathcal{A}_t$. It fills the remaining *exploration slots* in S_t by sampling uniformly (Line 6) a subset from the unknown arms without the pivot ($\mathcal{U}_t \setminus \{P_t\}$). As the number

of exploration slots varies throughout the execution of the algorithm, uniform sampling is an elegant way of distributing evenly the pulls among unknown arms. Every time we accept an unknown arm $i \in \mathcal{U}_t$ we sacrifice an exploration slot. TOP- k -UIF handles this trade-off by accepting items only when there is statistical confidence that they are optimal. In particular, we accept an arm $i \in \mathcal{U}_t$ when the number of unknown arms A_i which are worse than i with statistical confidence (computed by COUNT-BA) is greater than or equal to the number of suboptimal arms in the unknown set $|\mathcal{U}_t| - (k - |\mathcal{A}_t|) = |\mathcal{U}_t \cup \mathcal{A}_t| - k$ (Line 19, remember that $k - |\mathcal{A}_t|$ is the number of optimal arms in the unknown set). The acceptance rule, together with the rejection rule described in the previous point (Point 1), allows us to upper bound the number of times we play a suboptimal arm j and the number of times we do not play an optimal arm i , allowing us to recover the customary $1/\Delta$ dependence in the instance-dependent bound (Theorem 2).

3. *Update of the pivotal arm.* The choice of the pivot is a very delicate matter for two distinct reasons. The first—and most intuitive—is the fact that, on the one hand, playing a suboptimal arm as the pivot has a cost in terms of regret, encouraging updates towards better pivotal arms. On the other hand, every time we change the pivot, the rank-breaking estimates w.r.t. the old pivotal arms stop being updated, making them of no use. This is equivalent to restarting the algorithm. The second—and more subtle—matter is due to the type of feedback: top- m rankings. Indeed, in a top- m ranking, we observe a comparison between a pulled arm $i \in S_t$ and the pivot P_t if and only if at least one of the two is among the top- m arms of the latent ranking. If the pivot P_t is very bad (i.e., if $\theta_{P_t} \approx 0$), we observe very few comparisons with other bad items (i.e., $j \in \mathcal{U}_t$ s.t. $\theta_j \approx 0$), making it harder to eliminate them. We handle the problem as follows. As in IF1, when an unknown arm is superior to the current pivot with statistical confidence, it becomes the new pivot (see Line 29 and the NEW-PIVOT routine). Furthermore, the NEW-PIVOT routine is designed to guarantee that, in the first round $T_{\mathcal{A}}$ in which an arm is accepted ($\mathcal{A}_{T_{\mathcal{A}}+1} \neq \{\}$), the next pivot $P_{T_{\mathcal{A}}+1}$ is one of the accepted arms, unless the current pivot has already been accepted. This fact is formally proven in Lemma 8. Once we have an accepted arm, we stop updating the pivot and thus resetting the algorithm, preventing the regret incurred after each restart. Furthermore, we use all the accepted arms as pivotal arms ($\mathcal{P}_t = \mathcal{A}_t \cup \{P_t\}$, Line 5). This allows us to solve the problem associated with the scarcity of observations. Indeed, the last pivotal arm could have $\theta_{P_t} \approx 0$, leading to few observations with other very suboptimal arms. The algorithm handles this scenario by the additional acceptance routine at Line 25. The idea is that, when $\theta_{P_t} \approx 0$ and $P_t \in \mathcal{A}_t$, we still observe many comparisons with the unknown arms which are good ($\theta_i \approx 1$), which after a few rounds will be regarded as superior to the pivot with statistical confidence, and thus be accepted, becoming pivotal themselves.

Top- k Regret Upper Bound. The ideas in the previous paragraph collectively allow us to derive the following instance-dependent and worst-case upper bounds on the expected top- k regret, which we present in Theorem 2 below. The formal derivation of the results is deferred to Appendix D.

Theorem 2 (Upper Bounds to the Top- k Regret). *For every $n, k \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}_{\geq 1}$ with $m \leq k \leq n$, instance $\nu \in \mathcal{E}_n^{(k,m)}$ and learning horizon $T \in \mathbb{N}_{\geq n}$, by setting the algorithm parameters to $\beta_{\delta,n,T}(N) = \sqrt{\ln(2nT/\delta)/N}$ for $N \in \mathbb{N}_{\geq 1}$ (with $\beta_{\delta,n,T}(0) := +\infty$ and $\beta_{\delta,n,T}(+\infty) := 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top- k regret suffered by TOP- k -UIF:*

$$\text{EReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T) \leq C_2 \frac{L_\nu}{m} \left(nk + \sum_{u \in \llbracket k+1, n \rrbracket} 1/\Delta_{k,u} \right) \ln(T),$$

where $C_2 \in \mathbb{R}_{>0}$ is an absolute constant, L_ν is an upper bound to the number of pivots chosen by the TOP- k -UIF algorithm over the learning horizon (formally defined in Appendix D) and $\Delta_{k,u} := \theta_{r_\theta^{*-1}(k)} - \theta_{r_\theta^{*-1}(u)}$ for all $u \in \llbracket k+1, n \rrbracket$. If $k = n$, the regret is identically zero. If $k < n$, for a sufficiently large learning horizon satisfying $T \geq \frac{(nk)^2}{m(n-k)} L_{\mathcal{E}_n^{(k,m)}} \ln(2nT)$, the worst-case expected top- k regret is bounded by:

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \text{EReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T) \leq C_3 \sqrt{L_{\mathcal{E}_n^{(k,m)}} \frac{n-k}{m} T \ln(T)},$$

where $C_3 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\mathcal{E}_n^{(k,m)}} := \sup_{\nu \in \mathcal{E}_n^{(k,m)}} L_\nu \leq n$ (see Appendix D).

For $k < n$, the gap-dependent term in the instance-dependent bound is at most:

$$\mathcal{O}\left(L_\nu \frac{n-k}{m} \frac{1}{\Delta_{k,k+1}} \log(T)\right),$$

since $\Delta_{k,k+1} \leq \Delta_{k,u}$ for all $u \in \llbracket k+1, n \rrbracket$, highlighting that the complexity of the instance depends on the hardness of differentiating the k -th arm from the $(k+1)$ -th. L_ν and $L_{\mathcal{E}_n^{(k,m)}}$ control the expected number of pivotal arms chosen by the TOP- k -UIF algorithm over the learning horizon and are less than or equal to n . Furthermore, as expected, the regret scales inversely with the informativeness of the feedback $m \in \llbracket k \rrbracket$. In the worst case, we obtain the upper bound $\tilde{\mathcal{O}}\left(\sqrt{L_{\mathcal{E}_n^{(k,m)}} \frac{n-k}{m} T}\right)$, which matches the lower bound (Theorem 1) in the dependence on the learning horizon T and the informativeness of the feedback m in the regime $m \leq k/2$, providing—for the first time—a sublinear regret bound which holds for every possible choice of the PL parameters for an instance $\nu \in \mathcal{E}_n^{(k,m)}$.

5 Connection with Multinomial Logit Bandits

In this section, we discuss the connection between CMABs with ranking feedback and MNL bandits (Agrawal et al., 2019) under the top- k objective (Saha and Gaillard, 2025), showing how the algorithmic ideas of TOP- k -UIF translate seamlessly to this setting and allow us to solve an open problem in the literature. An MNL bandit with top- k objective is similar to a CMAB with ranking feedback, with $m = 1$. Indeed, the learner provides a subset $S_t \subseteq \llbracket n \rrbracket$ of arms, with $|S_t| = k$, and receives winner feedback (i.e., a top-1 ranking), with the addition that, with non-zero probability, no item is chosen from the set. This is modeled by adding a fictitious item $i = 0$, with corresponding PL parameter $\theta_0 \in (0, 1]$. Formally, the learner observes a top-1 ranking $R_t^{(1)}$ where the latent ranking is sampled from a PL distribution that involves the fictitious item: $R_t \sim \text{PL}((\theta_i)_{i \in S_t \cup \{0\}})$. We denote by $\nu_{\text{mnl}} = (n, \theta, \theta_0, k)$ an MNL bandit with top- k objective, where $(n, \theta, k, 1) \in \mathcal{E}_n^{(k,m)}$ and $\theta_0 \in (0, 1]$. Let $\mathcal{E}_{\text{mnl},n}^{(k)}$ be the set of MNL bandits with top- k objective. Randomized policies π and the expected top- k regret are defined exactly as in the CMAB with ranking feedback setting (see Section 2). Saha and Gaillard (2025) summarize the existing results in the literature, showing how all the regret bounds either rely on the assumption $\theta_0 = \max_{i \in \llbracket n \rrbracket} \theta_i = 1$ or scale as $\tilde{\mathcal{O}}\left(\sqrt{\theta_0^{-1} n T}\right)$, becoming vacuous when $\theta_0 \rightarrow 0$. They provide an upper bound $\tilde{\mathcal{O}}\left(\sqrt{\min\{\theta_0^{-1}, n\} n T} + n^2 k^2 \theta_0^{-2}\right)$, improving the regret rate when $T \gg \theta_0^{-1}$, but still suffering from an additional exploding burn-in term.⁷ Indeed, the algorithm proposed in (Saha and Gaillard, 2025) relies on an optimistic approach where upper confidence bounds on the PL parameters θ_j are defined as: $\hat{\theta}_{i,t}^{\text{ucb}} := \min_{j \in \{0\} \cup \llbracket n \rrbracket} \gamma_{i,j,t}^{\text{ucb}} \gamma_{j,0,t}^{\text{ucb}}$, where $\gamma_{i,j,t}^{\text{ucb}}$ is an optimistic estimate of $\gamma_{i,j} := \theta_i / \theta_j$. Letting $\theta_0 \rightarrow 0$ implies $\gamma_{i,0} \rightarrow +\infty$ for all $i \in \llbracket n \rrbracket$, making the optimistic estimates vacuous as they tend to infinity.

The problem arises from the fact that Saha and Gaillard (2025) implement a dynamic pivoting approach by solely modifying the way optimistic estimates are computed. In particular, for each arm i , they choose the arm $j \in \{0\} \cup \llbracket n \rrbracket$ which provides the tightest upper confidence bound. Unfortunately, the estimate still relies on the comparisons against the no-choice item, and the scarcity of comparisons issue described in Section 4 still arises when $\theta_0 \approx 0$. Solving the problem requires a shift of paradigm. Indeed, the pivoting approach in TOP- k -UIF relies on a direct modification of the set of pulled arms S_t by always playing the pivotal arm in each round. This requires care in the algorithmic design, as it can lead to a severe cost in terms of regret, but allows us to prevent the exploding burn-in term when $\theta_0 \rightarrow 0$. In Appendix C, we present the MNL-TOP- k -UIF algorithm (Algorithm 3) which adapts TOP- k -UIF to MNL bandits with top- k objective. The algorithm is almost identical, except that the pivot is initialized to $P_1 = 0$, an additional exploration slot is filled whenever $P_t = 0$, and an accepted real arm replaces item 0 as pivot if the first acceptance

⁷The burn-in term appears in the proof of (Saha and Gaillard, 2025, Theorem 5), despite being omitted from the statement of the theorem. In particular, this dependence arises from the inequality $\tau_{i,j,t} > C m^2 \theta_{\max}^2 x$, which, following the notation used in this work, is translated to: $\tau_{i,j,t} > C k^2 \theta_0^{-2} x$ where $x = 2 \ln T$.

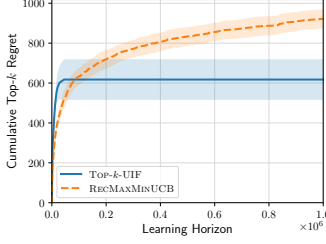


Figure 1: TOP- k -UIF vs Saha and Gopalan (2019b) (30 trials, mean $\pm$ SEM)

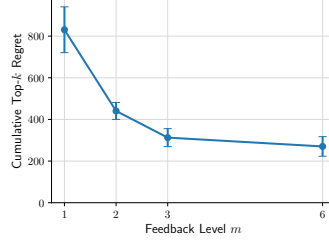


Figure 2: TOP- k -UIF with varying m (30 trials, mean $\pm$ SEM)

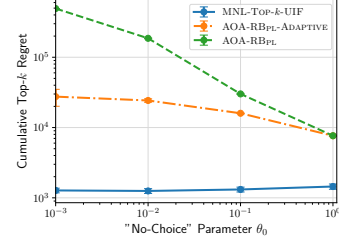


Figure 3: MNL-TOP- k -UIF vs Saha and Gaillard (2025) (30 trials, mean $\pm$ SEM)

does not otherwise trigger a pivot update. A derivation which is almost identical to the regret upper bounds for the TOP- k -UIF algorithm allows us to obtain the following result.⁸

Theorem 3 (Upper Bounds to the Top- k Regret for MNL Bandits). *For every $n, k \in \mathbb{N}_{\geq 2}$, with $k \leq n$, instance $\nu_{mnl} \in \mathcal{E}_{mnl,n}^{(k)}$ and learning horizon $T \in \mathbb{N}_{\geq n}$, by setting the algorithm parameters to $\beta_{\delta,n,T}(N) = \sqrt{\ln(2nT/\delta)/N}$ for $N \in \mathbb{N}_{\geq 1}$ (with $\beta_{\delta,n,T}(0) := +\infty$ and $\beta_{\delta,n,T}(+\infty) := 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top- k regret suffered by MNL-TOP- k -UIF:*

$$\text{EReg}_{\text{MNL-TOP-}k\text{-UIF}, \nu_{mnl}}^{(k)}(T) \leq C_4 L_{\nu_{mnl}} \left(nk + \sum_{u \in \llbracket k+1, n \rrbracket} 1/\Delta_{k,u} \right) \ln(T),$$

where $C_4 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\nu_{mnl}}$ is defined analogously to L_{ν} in the statement of Theorem 2 (see Appendix D) and $\Delta_{k,u} := \theta_{r_{\theta}^{*-1}(k)} - \theta_{r_{\theta}^{*-1}(u)}$ for all $u \in \llbracket k+1, n \rrbracket$. If $k = n$, the regret is identically zero. If $k < n$, for a sufficiently large learning horizon satisfying $T \geq \frac{(nk)^2}{n-k} L_{\mathcal{E}_{mnl,n}^{(k)}} \ln(2nT)$, the worst-case expected top- k regret is bounded by:

$$\sup_{\nu_{mnl} \in \mathcal{E}_{mnl,n}^{(k)}} \text{EReg}_{\text{MNL-TOP-}k\text{-UIF}, \nu_{mnl}}^{(k)}(T) \leq C_5 \sqrt{L_{\mathcal{E}_{mnl,n}^{(k)}} (n-k) T \ln(T)},$$

where $C_5 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\mathcal{E}_{mnl,n}^{(k)}} := \sup_{\nu_{mnl} \in \mathcal{E}_{mnl,n}^{(k)}} L_{\nu_{mnl}} \leq n+1$ (see Appendix D).

Observe that this is the first result which is sublinear in T (i.e., $\tilde{O}(\sqrt{T})$) even for $\theta_0 \rightarrow 0$, proving that the dependence on θ_0^{-1} in the existing bounds is not due to the intrinsic statistical complexity of the setting, but is rather an artifact of the algorithmic approaches used to tackle the problem.

6 Numerical Validation

In this section, we numerically validate the performance of TOP- k -UIF and MNL-TOP- k -UIF in synthetic simulation environments. Figures 1, 2, and 3 report the averaged cumulative regret with *Standard Error of the Mean* (SEM) error bands over 30 independent runs.

TOP- k -UIF vs REC-MAX-MIN-UCB (Saha and Gopalan, 2019b). We validate TOP- k -UIF against REC-MAX-MIN-UCB (Saha and Gopalan, 2019b) in a synthetic instance of a CMAB with ranking feedback. To ensure the fairness of the comparison with the baseline, we set $k = m$. Let $n = 10$, $k = m = 3$, $T = 10^6$ and $\theta_1 = \theta_2 = 1$, $\theta_3 = 0.65$, and $\theta_i = 0.55$ for $i > 3$. We run $N = 30$ trials and average the cumulative regrets suffered by the two algorithms, which are reported in Figure 1. TOP- k -UIF successfully identifies the correct set of actions before 10^5 rounds in all of the N trials, and suffers a significantly smaller regret than the baseline.

Sensitivity of TOP- k -UIF to m . We fix a synthetic instance of a CMAB with ranking feedback and only vary $m \in \{1, 2, 3, 6\}$. We test the subsequent performance variation of TOP- k -UIF. In

⁸The fictitious pivot $P_t = 0 \notin \llbracket n \rrbracket$ does not belong to the played set $S_t \subseteq \llbracket n \rrbracket$; the adaptation accounting for this difference is given in Appendix D.

particular, let $n = 16$, $k = 6$ and $T = 10^6$. We run $N = 30$ trials for every value of m . In Figure 2, we report the average cumulative regrets at the end of the learning horizon, over the N trials. As expected, the regret is noticeably larger as m gets smaller.

MNL-TOP- k -UIF vs AOA-RB_{PL}-Adaptive (Saha and Gaillard, 2025). We validate MNL-TOP- k -UIF against AOA-RB_{PL} and AOA-RB_{PL}-ADAPTIVE from (Saha and Gaillard, 2025) in a synthetic instance of an MNL bandit with top- k objective. Let $n = 30$, $k = 5$, $m = 1$ and $T = 10^6$. We vary $\theta_0 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and we run $N = 30$ trials for every value of θ_0 . In Figure 3, we report the average cumulative regrets over the N trials. MNL-TOP- k -UIF consistently outperforms both baselines, demonstrating performance consistency across the values of θ_0 . Indeed, it is the only algorithm whose performance does not degrade as $\theta_0 \rightarrow 0$, confirming empirically that the dependence of the baselines on θ_0 is not an analytical artifact.

7 Open Challenges

This paper studied the problem of combinatorial bandits with ranking feedback under general top- m feedback, providing, for the first time, sublinear worst-case regret bounds w.r.t. the PL parameters. In particular, in Theorem 1, we derived a worst-case lower bound on the expected top- k regret of order $\Omega\left(\sqrt{\frac{n-k}{km}T}\right)$ for $m \leq k/2$, while in Theorem 2 we presented a regret upper bound of order $\tilde{O}\left(\sqrt{L_{\mathcal{E}_n^{(k,m)}} \frac{n-k}{m} T}\right)$, which matches the lower bound w.r.t. T and m in this regime. The number of pivots over the learning horizon ($L_{\mathcal{E}_n^{(k,m)}}$) can be naturally upper bounded by n : $L_{\mathcal{E}_n^{(k,m)}} \leq n$, introducing an additional $\sqrt{n}$ factor in the upper bound. In the original dueling bandits paper, this term is upper bounded by an $\mathcal{O}(\log n)$ factor (Yue et al., 2009, Theorem 1). We wonder if it is possible to do the same in our setting due to the algorithmic similarity, allowing us to match the dependence w.r.t. $n - k$ in the lower bound. We conjecture that such a result, if possible, would require significantly different proof techniques. Analogously, when we rescale the instantaneous reward by a factor $1/k$ as in $g_{\theta}^{(k)}$ (see Section 2), the lower bounds for combinatorial bandits with bandit and semi-bandit feedback under constrained action space are $\Omega\left(\sqrt{(n-k)T}\right)$ and $\Omega\left(\sqrt{\frac{n-k}{k}T}\right)$, respectively (see Audibert et al., 2014, Table 1). Because of the analogy between these settings and CMABs with winner feedback ($m = 1$) and full-ranking feedback ($m = k$), respectively, we wonder whether the lower-bound analysis can be extended to the full range $m \leq k$ and, in particular, whether a full-ranking lower bound can be proved without the $\sqrt{1/k}$ factor suggested by extrapolating the dependence displayed in Theorem 1. By answering both questions positively, the setting would be closed up to logarithmic dependencies.

Acknowledgments

This publication was funded with the contribution of Ministero dell’Università e della Ricerca pursuant to D.D. n. 7206 of 17 April 2025 – BANDO FIS 2. Project FIS-2023-02598 (Starting Grant), title: “Unified Learning from Diverse Human Feedback” (HUmLrn). CUP: D53C25000710001.

References

- Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.
- Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.
- Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 360–368. PMLR, 2014.
- Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

- Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.
- Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 3rd edition, 2020.
- Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 1141–1154. PMLR, 2015.
- Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
- Robert D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.
- Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
- Robin L. Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975.
- Pantelimon G. Popescu, Sever S. Dragomir, Emil I. Slușanschi, and Octavian N. Stănășilă. Bounds for Kullback-Leibler divergence. *Electronic Journal of Differential Equations*, 2016(237):1–6, 2016.
- Aadirupa Saha and Pierre Gaillard. Finally rank-breaking conquers MNL bandits: Optimal and efficient algorithms for MNL assortment. In *International Conference on Learning Representations (ICLR)*, 2025.
- Aadirupa Saha and Aditya Gopalan. Active ranking with subset-wise preferences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 3312–3321. PMLR, 2019a.
- Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 983–993, 2019b.
- Aadirupa Saha and Aditya Gopalan. Best-item learning in random utility models with subset choices. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 4281–4291. PMLR, 2020.
- Tanguy Urvoy, Fabrice Clérot, Raphaël Féraud, and Sami Naamane. Generic exploration and K-armed voting bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pages 91–99. PMLR, 2013.
- Christian Wirth, Riad Akrou, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.
- John I. Yellott. The relationship between Luce’s choice axiom, Thurstone’s theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977.
- Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 241–248. Omnipress, 2011.
- Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.
- Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative upper confidence bound for the K-armed dueling bandit problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 10–18. PMLR, 2014.

A Related Works

Dueling Bandits. The K -armed *dueling bandits* problem has been introduced by Yue et al. (2009), modeling an online learning setting where, in each round, the agent suggests two options and observes *preference* feedback. Under the assumptions of *strong stochastic transitivity* and *stochastic triangle inequality*, the INTERLEAVED FILTER (Yue et al., 2009) algorithm attains a near-optimal regret bound. Yue and Joachims (2011) develop the BEAT-THE-MEAN algorithm which works under a relaxed version of the stochastic transitivity assumption (namely, *relaxed stochastic transitivity*). Broadening the scope of preference structures, Urvoy et al. (2013) propose the SAVAGE algorithm within a generic voting bandit framework, notably dropping strict transitivity requirements to focus on identifying the Copeland winner. Zoghi et al. (2014) adapt the renowned UCB algorithm for stochastic MABs to the dueling setting, developing the RUCB algorithm. Finally, establishing the fundamental limits of the problem, Komiyama et al. (2015) derive an asymptotic regret lower bound and introduce the RMED (Relative Minimum Empirical Divergence) algorithm, proving it to be asymptotically optimal.

Rank Breaking. Khetan and Oh (2016) study the problem of estimating the parameters of the PL model from incomplete rankings provided by different individuals, through a *rank-breaking* estimator. The idea of rank breaking consists of the disaggregation of incomplete rankings into a collection of pairwise comparisons which are processed independently. They provide a high-probability bound on the estimation error of the rank-breaking estimator which depends on the topology of the collected data, when represented in a comparison graph built by aggregating the pairwise comparisons. Azari Soufiani et al. (2014) consider a dataset of *Missing Not At Random* (MNAR) pairwise comparisons generated from the *rank-breaking* of incomplete rankings obtained by censoring some of the comparisons among the alternatives according to their position in the latent ranking. The goal is to show that, even if the pairwise comparisons are correlated, since some are obtained from the same incomplete ranking, it is still possible to process them as independent and maintain consistency under some assumptions on the MNAR censoring mechanisms. This allows us to reduce the computational load of the estimation, as working with pairwise comparisons requires lighter methods.

Online Learning from Ranking Feedback. There are several works which consider the problem of learning the best items in a set from the observation of *rankings* (potentially partial) over a subset (*assortment*) proposed by the learner. Some works are focused on the *pure exploration* setting. In particular, Saha and Gopalan (2019a) study an active (ϵ, δ) -PAC ranking problem in which a learner, from the observation of top- m Plackett-Luce rankings over the assortment, needs to minimize the sample complexity of learning an approximate ranking of the items. To this end, they design the BEAT-THE-PIVOT algorithm which first learns a good item that is then used as a pivot to compare all the others by means of rank-breaking. Saha and Gopalan (2020) consider a similar (ϵ, δ) -PAC best-item learning problem where latent rankings are generated by a Random Utility Model (RUM) which generalizes PL rankings. They identify a property of the noise distribution that induces the RUM which allows them to characterize the sample complexity of the problem. In the *regret minimization* setting instead, Saha and Gopalan (2019b) study the problem in which a learner is required to identify the best item by observing top- m PL rankings over a variable-size assortment. The notion of regret used in such a setting accounts for the cost of providing a suboptimal item in an assortment. In the same paper, they also study the setting in which the learner is constrained to play a subset of fixed cardinality and wants to learn the best assortment, i.e., the one associated with the highest PL parameters, under the same cardinality constraint.

Multinomial Logit Bandits. *Multinomial logit* bandits (Agrawal et al., 2019) model the problem in which a learner provides an assortment of items and observes the following kind of feedback: either one of the items is bought or none is bought (“no-choice”). Most of the works associate a revenue with each item, defining a notion of reward associated with an assortment which corresponds to the average revenue. The goal is to minimize the regret due to providing suboptimal assortments. Most of the works assume a Plackett-Luce model which describes with which probability an item in the assortment is bought. Agrawal et al. (2019) propose the MNL-UCB algorithm which relies on optimistic estimates of the PL parameters by repeatedly querying the same assortment. The regret bounds in (Agrawal et al., 2019) grow with the reciprocal of the probability of observing that no item is bought, becoming vacuous if such probability is allowed to go to 0. Saha and Gaillard (2025) present the AOA-RB_{PL}-ADAPTIVE algorithm, which improves the dependence on the no-choice probability by moving it into a burn-in term which does not scale with the learning horizon.

B Analysis for the Top- k Regret Lower Bound

We begin our analysis by deriving an analogue of Lattimore and Szepesvári (2020, Lemma 15.1) for the combinatorial setting with top- m feedback. Specifically, we establish an upper bound on the divergence between the distributions over the history H_T induced by playing a fixed randomized policy π on two distinct instances $\nu, \nu' \in \mathcal{E}_n^{(k,m)}$ whose PL parameters lie in $[1/2, 1]$.

To formally state this, we introduce the following measure-theoretic notation. Let μ and ϕ be two probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$. We write $\mu \ll \phi$ to indicate that μ is absolutely continuous with respect to ϕ . Under this condition, we denote the *Radon-Nikodym derivative* of μ with respect to ϕ by $\frac{d\mu}{d\phi}$, and define their *Kullback-Leibler* (KL) divergence as:

$$D_{\text{KL}}(\mu \parallel \phi) := \int_{\mathcal{X}} \ln \left(\frac{d\mu}{d\phi} \right) d\mu.$$

Furthermore, when $\mu, \phi \in \Delta(\mathcal{X}, \mathcal{F})$, for a random variable X , we denote the KL divergence between its distributions (pushforward measures) under μ and ϕ as:

$$D_{\text{KL}\mu \parallel \phi}(X) := D_{\text{KL}}(\mu \circ X^{-1} \parallel \phi \circ X^{-1}).$$

In our setting, we apply this to evaluate the divergence of the history H_T under the measures $\mathbb{P}_{\pi, \nu}$ and $\mathbb{P}_{\pi, \nu'}$.

Lemma 4 (Divergence Decomposition). *For every $n, k \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}_{\geq 1}$ with $m \leq k/2 \leq n/2$, instances $\nu, \nu' \in \mathcal{E}_n^{(k,m)}$ where $\nu = (n, \theta, k, m)$, $\nu' = (n, \theta', k, m)$, and $\theta_i, \theta'_i \in [1/2, 1]$ for every $i \in \llbracket n \rrbracket$, randomized policy π , and learning horizon $T \in \mathbb{N}_{\geq 1}$, it holds:*

$$D_{\text{KL}\mathbb{P}_{\pi, \nu} \parallel \mathbb{P}_{\pi, \nu'}}(H_T) \leq 32 \frac{m}{k} \sum_{i \in \llbracket n \rrbracket} (\theta_i - \theta'_i)^2 \mathbb{E}_{\pi, \nu}[N_i(T)],$$

where $N_i(T) := \sum_{t=1}^T \mathbb{I}[i \in S_t]$ for every $i \in \llbracket n \rrbracket$.

Proof. By the chain rule, since the randomized policy is the same under both probability measures $\mathbb{P}_{\pi, \nu}$ and $\mathbb{P}_{\pi, \nu'}$, we have:

$$D_{\text{KL}\mathbb{P}_{\pi, \nu} \parallel \mathbb{P}_{\pi, \nu'}}(H_T) = \sum_{t=1}^T \mathbb{E}_{\pi, \nu} \left[D_{\text{KLPL}((\theta_i)_{i \in S_t}) \parallel \text{PL}((\theta'_i)_{i \in S_t})}(R_t^{(m)}) \right].$$

To upper bound the KL divergence between the top- m ranking distributions we follow an approach analogous to (Saha and Gopalan, 2019b). Let $I_{u,t} := R_t^{-1}(u)$ and $S_t^{(u)} := \{I_{1,t}, \dots, I_{u,t}\}$ for $u \in \llbracket m \rrbracket$, with $S_t^{(0)} := \{\}$. Observe that, because of the chain rule of the KL divergence and the properties of the PL model, we can write the divergence of two top- m rankings as:

$$\begin{aligned} & D_{\text{KLPL}((\theta_i)_{i \in S_t}) \parallel \text{PL}((\theta'_i)_{i \in S_t})}(R_t^{(m)}) \\ &= \sum_{u=1}^m \mathbb{E}_{R_t \sim \text{PL}((\theta_i)_{i \in S_t})} \left[\ln \frac{\mathbb{P}_{R_t \sim \text{PL}((\theta_i)_{i \in S_t})} [I_{u,t} \mid S_t^{(u-1)}]}{\mathbb{P}_{R_t \sim \text{PL}((\theta'_i)_{i \in S_t})} [I_{u,t} \mid S_t^{(u-1)}]} \right]. \end{aligned}$$

By (Popescu et al., 2016, Theorem 1.4), we can upper bound the term inside the summation as:

$$\begin{aligned} & D_{\text{KLPL}((\theta_i)_{i \in S_t}) \parallel \text{PL}((\theta'_i)_{i \in S_t})}(R_t^{(m)}) \\ & \leq \sum_{u=1}^m \mathbb{E}_{R_t \sim \text{PL}((\theta_i)_{i \in S_t})} \left[\frac{\Theta'_{S_t \setminus S_t^{(u-1)}}}{\Theta_{S_t \setminus S_t^{(u-1)}}^2} \sum_{i \in S_t \setminus S_t^{(u-1)}} \frac{\theta_i^2}{\theta'_i} - 1 \right], \end{aligned}$$

where $\Theta_S := \sum_{i \in S} \theta_i$ and $\Theta'_S := \sum_{i \in S} \theta'_i$. We rewrite the inner term by completing the square w.r.t. θ'_i :

$$\frac{\Theta'_{S_t \setminus S_t^{(u-1)}}}{\Theta_{S_t \setminus S_t^{(u-1)}}^2} \sum_{i \in S_t \setminus S_t^{(u-1)}} \frac{\theta_i^2}{\theta'_i} - 1 = \frac{\Theta'_{S_t \setminus S_t^{(u-1)}}}{\Theta_{S_t \setminus S_t^{(u-1)}}^2} \left(\Theta'_{S_t \setminus S_t^{(u-1)}} \right)$$

$$\begin{aligned}
& + 2(\Theta_{S_t \setminus S_t^{(u-1)}} - \Theta'_{S_t \setminus S_t^{(u-1)}}) + \sum_{i \in S_t \setminus S_t^{(u-1)}} \frac{(\theta_i - \theta'_i)^2}{\theta'_i} \Big) - 1 \\
& = \frac{\Theta'_{S_t \setminus S_t^{(u-1)}}}{\Theta_{S_t \setminus S_t^{(u-1)}}^2} \sum_{i \in S_t \setminus S_t^{(u-1)}} \frac{(\theta_i - \theta'_i)^2}{\theta'_i} - \frac{(\Theta_{S_t \setminus S_t^{(u-1)}} - \Theta'_{S_t \setminus S_t^{(u-1)}})^2}{\Theta_{S_t \setminus S_t^{(u-1)}}^2} \\
& \leq 8 \frac{k}{(k-m+1)^2} \sum_{i \in S_t \setminus S_t^{(u-1)}} (\theta_i - \theta'_i)^2 \leq 32 \frac{1}{k} \sum_{i \in S_t} (\theta_i - \theta'_i)^2, \tag{2}
\end{aligned}$$

where Step (2) follows from the fact that $\theta_i, \theta'_i \in [1/2, 1]$ and $m \leq k/2$. Thus, since the bound is uniform for any $u \in \llbracket m \rrbracket$, we have:

$$\begin{aligned}
D_{\text{KL} \mathbb{P}_{\pi, \nu} \| \mathbb{P}_{\pi, \nu'}}(H_T) & \leq 32 \frac{m}{k} \mathbb{E}_{\pi, \nu} \left[\sum_{t=1}^T \sum_{i \in S_t} (\theta_i - \theta'_i)^2 \right] \\
& = 32 \frac{m}{k} \mathbb{E}_{\pi, \nu} \left[\sum_{i \in \llbracket n \rrbracket} \sum_{t=1}^T (\theta_i - \theta'_i)^2 \mathbb{I}[i \in S_t] \right] \\
& = 32 \frac{m}{k} \sum_{i \in \llbracket n \rrbracket} (\theta_i - \theta'_i)^2 \mathbb{E}_{\pi, \nu} [N_i(T)].
\end{aligned}$$

□

We now have all the required tools to derive the worst-case lower bound to the top- k regret. The construction is similar to (Audibert et al., 2014, Appendix B).

Theorem 1 (Worst-Case Lower Bound to the Top- k Regret). *For every $n, k \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}_{\geq 1}$ with $m \leq k/2 \leq n/2$ and $n/k \in \mathbb{N}_{\geq 1}$, randomized policy π and learning horizon $T \in \mathbb{N}_{\geq 1}$ satisfying $T \geq \frac{1}{16} \frac{(n-k)^2}{nmk}$, it holds:*

$$\sup_{\nu \in \mathcal{E}_n^{(k, m)}} \text{EReg}_{\pi, \nu}^{(k)}(T) \geq C_1 \left(1 - \frac{k}{n}\right)^{3/2} \sqrt{\frac{n-k}{km}} T,$$

where $C_1 \in \mathbb{R}_{>0}$ is an absolute constant.

Proof of Theorem 1. If $n = k$, the result follows from the non-negativity of the expected regret because the right-hand side of the lower bound is zero. Suppose, thus, that $n > k$. Fix $\Delta \in (0, 1/2]$ and $\epsilon \in (0, \Delta)$. For each $S \subseteq \llbracket 2, n \rrbracket$ with $|S| = k-1$, let:

$$\theta_j^{(S)} := \begin{cases} 1 & \text{if } j = 1 \\ 1/2 + \Delta & \text{if } j \in S \\ 1/2 & \text{otherwise} \end{cases} \quad \text{for every } j \in \llbracket n \rrbracket,$$

and let $\nu^{(S)} := (n, \theta^{(S)}, k, m)$. Define the base instance $\nu^{(0, \epsilon)} := (n, \theta^{(0, \epsilon)}, k, m)$ by:

$$\theta_j^{(0, \epsilon)} := \begin{cases} 1 & \text{if } j = 1 \\ 1/2 + \epsilon & \text{if } j \in \llbracket 2, k \rrbracket \\ 1/2 & \text{otherwise} \end{cases} \quad \text{for every } j \in \llbracket n \rrbracket.$$

Both $\nu^{(S)}$ and $\nu^{(0, \epsilon)}$ belong to $\mathcal{E}_n^{(k, m)}$: their parameters are normalized, belong to $[1/2, 1]$, and have a strict top- k separation. In particular, the optimal set for $\nu^{(S)}$ is $\{1\} \cup S$.

Let $N_i(T) := \sum_{t=1}^T \mathbb{I}[i \in S_t]$ and $N_S(T) := \sum_{i \in S} N_i(T)$. The exact instantaneous regret on $\nu^{(S)}$ gives:

$$\text{EReg}_{\pi, \nu^{(S)}}^{(k)}(T) = \frac{1}{k} \mathbb{E}_{\pi, \nu^{(S)}} \left[\sum_{t=1}^T \left(\frac{1}{2} \mathbb{I}[1 \notin S_t] + \Delta \sum_{i \in S} \mathbb{I}[i \notin S_t] \right) \right]$$

$$\geq \Delta T - \frac{\Delta}{k} \mathbb{E}_{\pi, \nu^{(S)}} [N_1(T) + N_S(T)], \quad (3)$$

where the inequality follows from $\Delta \leq 1/2$. Apply Lemma 14 to the random variable $(N_1(T) + N_S(T))/(kT) \in [0, 1]$, with the alternative and base laws as μ and ϕ , respectively. Together with Lemma 4, Equation (3) yields:

$$\begin{aligned} \text{EReg}_{\pi, \nu^{(S)}}^{(k)}(T) &\geq \Delta T - \frac{\Delta}{k} \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_1(T) + N_S(T)] \\ &\quad - 4T\Delta \sqrt{\frac{m}{k} (\Delta^2 \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_S(T)] + \epsilon^2(k-1)T)}. \end{aligned} \quad (4)$$

Choose pairwise-disjoint sets $S_1, \dots, S_{n/k} \subseteq \llbracket 2, n \rrbracket$ of cardinality $k-1$, which exist because $(n/k)(k-1) = n - n/k \leq n-1$, and let:

$$S^* \in \arg \min_{S \in \{S_1, \dots, S_{n/k}\}} \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_S(T)].$$

Their disjointness implies:

$$\begin{aligned} \frac{n}{k} \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_{S^*}(T)] &\leq \sum_{u=1}^{n/k} \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_{S_u}(T)] \\ &\leq kT - \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_1(T)]. \end{aligned}$$

Consequently, using $N_1(T) \leq T$:

$$\begin{aligned} \Delta T - \frac{\Delta}{k} \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_1(T) + N_{S^*}(T)] &\geq \Delta T \left(1 - \frac{1}{k}\right) \left(1 - \frac{k}{n}\right), \\ \mathbb{E}_{\pi, \nu^{(0, \epsilon)}} [N_{S^*}(T)] &\leq \frac{k^2 T}{n}. \end{aligned}$$

Since the supremum over $\mathcal{E}_n^{(k, m)}$ is at least the regret on $\nu^{(S^*)}$, Equation (4) gives, for every $\epsilon \in (0, \Delta)$:

$$\begin{aligned} \sup_{\nu \in \mathcal{E}_n^{(k, m)}} \text{EReg}_{\pi, \nu}^{(k)}(T) &\geq \Delta T \left(1 - \frac{1}{k}\right) \left(1 - \frac{k}{n}\right) \\ &\quad - 4T\Delta \sqrt{\frac{m}{k} \left(\Delta^2 \frac{k^2 T}{n} + \epsilon^2(k-1)T\right)}. \end{aligned}$$

Letting $\epsilon \rightarrow 0$, we obtain:

$$\sup_{\nu \in \mathcal{E}_n^{(k, m)}} \text{EReg}_{\pi, \nu}^{(k)}(T) \geq \Delta T \left(1 - \frac{1}{k}\right) \left(1 - \frac{k}{n}\right) - 4T\Delta^2 \sqrt{\frac{mkT}{n}}.$$

The right-hand side is maximized at $\Delta^* := (1 - 1/k)(1 - k/n)/(8\sqrt{mkT/n})$. The horizon assumption implies $\sqrt{mkT/n} \geq (n-k)/(4n)$ and, therefore, $\Delta^* \leq (k-1)/(2k) \leq 1/2$. Thus:

$$\begin{aligned} \sup_{\nu \in \mathcal{E}_n^{(k, m)}} \text{EReg}_{\pi, \nu}^{(k)}(T) &\geq \frac{1}{16} \left(\frac{k-1}{k}\right)^2 \left(1 - \frac{k}{n}\right)^{3/2} \sqrt{\frac{n-k}{km}} T \\ &\geq \frac{1}{64} \left(1 - \frac{k}{n}\right)^{3/2} \sqrt{\frac{n-k}{km}} T, \end{aligned}$$

which proves the statement with $C_1 = 1/64$. $\square$

C Details of TOP- k -UIF Pseudocode

This appendix provides additional details on the pseudocode of the TOP- k -UIF algorithm (Algorithm 1), together with the pseudocode of all the invoked subroutines. Furthermore, we describe the adaptation of TOP- k -UIF to MNL bandits under the top- k objective, obtaining the MNL-TOP- k -UIF algorithm (Algorithm 3).

Algorithm 2 Subroutines for TOP- k -UIF.

```

1: function RB-UPDATE( $\mathcal{P}, S, (\hat{\mu}_i^{(p)})_{i \in S, p \in \mathcal{P}}, (N_i^{(p)})_{i \in S, p \in \mathcal{P}}, R^{(m)}$ )
2:   for  $i \in S, p \in \mathcal{P}$  do
3:     if  $R^{(m)}(p) \neq R^{(m)}(i)$  then
4:        $N_i^{(p)} \leftarrow N_i^{(p)} + 1$ 
5:        $\hat{\mu}_i^{(p)} \leftarrow \frac{N_i^{(p)} - 1}{N_i^{(p)}} \hat{\mu}_i^{(p)} + \frac{\mathbb{I}[R^{(m)}(i) < R^{(m)}(p)]}{N_i^{(p)}}$ 
6:     end if
7:   end for
8:   return  $(\hat{\mu}_i^{(p)})_{i \in S, p \in \mathcal{P}}, (N_i^{(p)})_{i \in S, p \in \mathcal{P}}$ 
9: end function

10: function COUNT-BA( $\mathcal{P}, \mathcal{U}, (\text{UCB}_i^{(p)})_{i \in \mathcal{U}, p \in \mathcal{P}}, (\text{LCB}_i^{(p)})_{i \in \mathcal{U}, p \in \mathcal{P}}$ )
11:   Let  $i \prec j$  iff  $\text{LCB}_i^{(p)} > \text{UCB}_j^{(p)}$  for some  $p \in \mathcal{P}$ 
12:   Let  $\prec^+$  be the transitive closure of  $\prec$ 
13:   for  $i \in \mathcal{U}$  do
14:      $A_i \leftarrow |\{j \in \mathcal{U} \setminus \{i\} \text{ s.t. } i \prec^+ j\}|$ 
15:      $B_i \leftarrow |\{j \in \mathcal{U} \setminus \{i\} \text{ s.t. } j \prec^+ i\}|$ 
16:   end for
17:   return  $(A_i)_{i \in \mathcal{U}}, (B_i)_{i \in \mathcal{U}}$ 
18: end function

19: function NEW-PIVOT( $p, \mathcal{U}, (\text{LCB}_i^{(p)})_{i \in \mathcal{U}}$ )
20:    $\mathcal{P}' \leftarrow \{i \in \mathcal{U} \text{ s.t. } \text{LCB}_i^{(p)} > 1/2\}$ 
21:   if  $\mathcal{P}' \neq \{\}$  then
22:     Sample  $p' \sim \text{Unif}(\arg \max_{i \in \mathcal{P}'} \text{LCB}_i^{(p)})$ 
23:   else
24:      $p' \leftarrow \perp$ 
25:   end if
26:   return  $p'$ 
27: end function

```

Remark. Under arbitrary realizations of the underlying rankings R_t which violate the confidence bounds, the TOP- k -UIF and MNL-TOP- k -UIF algorithms could otherwise accept more than k items or reject more than $n - k$, making $S_{\text{exp}, t}$ ill-defined (see Line 6). Formally, on every history, both algorithms accept an arm only if $|\mathcal{A}_{t+1}| < k$, reject one only if $|\mathcal{A}_{t+1} \cup \mathcal{U}_{t+1}| > k$, and play $\mathcal{A}_{t+1}$ in every remaining round once $|\mathcal{A}_{t+1}| = k$. For clarity, we omit these guards and the terminal behavior from the displayed pseudocode, as they do not alter the algorithms on the good event in which all confidence bounds hold.

C.1 Subroutines

In this section we describe the subroutines in Algorithm 2.

The RB-UPDATE subroutine updates the rank-breaking estimates for all the pulled arms as explained in Section 4. In particular, it takes as input the set of pivotal arms $\mathcal{P}$ and the set of pulled arms S , together with the rank-breaking estimates and the number of observed comparisons w.r.t. each pivot $p \in \mathcal{P}$. The last input is the top- m ranking $R^{(m)}$. Every time $R^{(m)}(p) \neq R^{(m)}(i)$ for a pivotal arm $p \in \mathcal{P}$ and a pulled arm $i \in S$, we update the corresponding rank-breaking estimate depending on whether $R^{(m)}(p) < R^{(m)}(i)$ or vice versa.

The COUNT-BA subroutine takes as input a set of pivotal arms $\mathcal{P}$, a set of unknown arms $\mathcal{U}$, together with the lower and upper confidence bounds for the rank-breaking estimates of each unknown arm $i \in \mathcal{U}$ w.r.t. each pivot arm $p \in \mathcal{P}$. Whenever there is at least one pivotal arm $p \in \mathcal{P}$ that allows us to conclude that item $i \in \mathcal{U}$ is superior to item $j \in \mathcal{U}$ with statistical confidence ($\text{LCB}_i^{(p)} > \text{UCB}_j^{(p)}$), the routine records the directed relation $i \prec j$. It then computes its transitive closure $\prec^+$ and returns the number A_i of unknown arms reachable from i and the number B_i of unknown arms from which

Algorithm 3 MNL-TOP- k -UIF (Multinomial Top- k Uniform Interleaved Filter).

```
1: Input:  $T \in \mathbb{N}_{\geq 1}$ ,  $n \in \mathbb{N}_{\geq 2}$ ,  $k \in \llbracket 2, n \rrbracket$ ,  $\beta_{\delta, n, T} : [0, +\infty] \rightarrow [0, +\infty]$ ,  $\delta \in (0, 1]$ 
2: Initialize  $P_1 \leftarrow 0$ ,  $\mathcal{A}_1 \leftarrow \{\}$ ,  $\mathcal{U}_1 \leftarrow \llbracket n \rrbracket$ ,
    $N_i^{(p)} \leftarrow 0$  if  $i \neq p, +\infty$  otherwise,  $\hat{\mu}_i^{(p)} \leftarrow 1/2$  for every  $i, p \in \llbracket 0, n \rrbracket$ 
3: for  $t \in \llbracket T \rrbracket$  do
4:    $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t$ ,  $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_t$ ,  $P_{t+1} \leftarrow P_t$ 
5:   Let  $\mathcal{P}_t \leftarrow \mathcal{A}_t \cup \{P_t\}$ 
6:   Sample  $S_{\text{exp}, t} \sim \text{Unif}(\{S \subseteq \mathcal{U}_t \setminus \{P_t\} \text{ s.t. } |S| = k - |\mathcal{P}_t| + \mathbb{I}[0 \in \mathcal{P}_t]\})$ 
7:   Pull  $S_t \leftarrow (\mathcal{P}_t \cup S_{\text{exp}, t}) \setminus \{0\}$ 
8:   Observe  $R_t^{(1)} : S_t \cup \{0\} \rightarrow \llbracket \min\{k, 2\} \rrbracket$ 
   // 1. Update Statistics & Confidence Bounds
9:    $(\hat{\mu}_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}$ ,  $(N_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}$ 
      $\leftarrow \text{RB-UPDATE}(\mathcal{P}_t, S_t \cup \{0\}, (\hat{\mu}_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}, R_t^{(1)})$ 
10:  for  $i \in \mathcal{U}_t, p \in \mathcal{P}_t$  do
11:     $\text{UCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} + \beta_{\delta, n, T}(N_i^{(p)})$ 
12:     $\text{LCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} - \beta_{\delta, n, T}(N_i^{(p)})$ 
13:  end for
   // 2. Update Unknown Arms
14:   $(A_i)_{i \in \mathcal{U}_t}, (B_i)_{i \in \mathcal{U}_t} \leftarrow \text{COUNT-BA}(\mathcal{P}_t, \mathcal{U}_t, (\text{UCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t}, (\text{LCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t})$ 
15:  for  $i \in \mathcal{U}_t$  do
16:    if  $B_i \geq k - |\mathcal{A}_t|$  then ▷ Reject arm
17:       $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
18:    else if  $A_i \geq |\mathcal{A}_t \cup \mathcal{U}_t| - k$  then ▷ Accept arm
19:       $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}$ 
20:       $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
21:    end if
22:  end for
   // 3. Pivot Update
23:  if  $\mathcal{A}_t \neq \{\}$  then ▷ Stop pivot update after first accepted arm
24:    for  $i \in \mathcal{U}_t$  do
25:      if  $\text{LCB}_i^{(P_t)} > 1/2$  then ▷ Accept arms which are better than the (accepted) pivot
26:         $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}$ 
27:         $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$ 
28:      end if
29:    end for
30:  else
31:     $p' \leftarrow \text{NEW-PIVOT}(P_t, \mathcal{U}_t, (\text{LCB}_i^{(P_t)})_{i \in \mathcal{U}_t})$ 
32:    if  $p' \neq \perp$  then
33:       $P_{t+1} \leftarrow p'$ 
34:    else if  $P_t = 0$  and  $\mathcal{A}_{t+1} \neq \{\}$  then
35:      Sample  $P_{t+1} \sim \text{Unif}(\mathcal{A}_{t+1})$  ▷ Make an accepted arm the pivot
36:    end if
37:  end if
38:  if  $|\mathcal{A}_{t+1}| = k$  then
39:     $\mathcal{U}_{t+1} \leftarrow \{\}$ 
40:  end if
41: end for
```

i is reachable. Constructing the directed graph and running a depth-first search from every node requires $\mathcal{O}(n^3)$ time per round.

The NEW-PIVOT subroutine returns an element different from $\perp$ whenever there is at least one unknown arm $i \in \mathcal{U}$ which is better than the pivot p with statistical confidence. This happens whenever $\text{LCB}_i^{(p)} > 1/2$. If such a condition is satisfied, the routine returns one of the unknown

items with the highest $\text{LCB}_i^{(p)}$ w.r.t. the pivotal arm p , guaranteeing that, if some item is accepted, the new pivot belongs to the set of the accepted arms.

C.2 Adapting TOP- k -UIF to MNL Bandits with Top- k Objective

In this section we provide the pseudocode of MNL-TOP- k -UIF, which adapts the TOP- k -UIF algorithm to the MNL bandit setting with top- k objective.

We initialize the pivotal arm to 0: $P_1 = 0$. The algorithm is almost identical to TOP- k -UIF; we only need to ensure that, when $P_t = 0$, instead of playing the pivot, we fill an additional exploration slot through $S_{\text{exp},t}$. Furthermore, whenever we run the rank-breaking update, we consider all the comparisons against the no-choice fictitious item by passing $S_t \cup \{0\}$ to the RB-UPDATE routine. If the first acceptance occurs while the fictitious item is still the pivot and NEW-PIVOT returns $\perp$, we set the next pivot to an accepted arm. This fallback is necessary because, unlike an ordinary pivot, item 0 does not belong to $\mathcal{U}_t$ and hence cannot itself be accepted.

D Analysis for the Top- k Regret Upper Bound

In this appendix, we provide the proofs of the results stated in Section 4.

By convention, given a variable X defined in the pseudocode of Algorithm 1, we add the subscript t (X_t) to refer to its value *at the beginning* of round t (e.g., $\hat{\mu}_{i,t}^{(p)}$, $N_{i,t}^{(p)}$, etc.). In what follows, we fix $n, k \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}_{\geq 1}$, $m \leq k \leq n$, and a learning horizon $T \in \mathbb{N}_{\geq 1}$.

D.1 Concentrations

We begin with some concentration results. Let us start by defining quantities of interest for the concentration of the estimator $\hat{\mu}_i^{(p)}$ and the number of observed comparisons $N_i^{(p)}$. For every $\nu \in \mathcal{E}_n^{(k,m)}$, $S \subseteq \llbracket n \rrbracket$, $|S| = k$, and distinct $i, p \in S$, we define:

- $f_{i|S}^{(p)} := \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [R^{(m)}(i) \neq R^{(m)}(p)]$.
- $\mu_{i|S}^{(p)} := \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [R^{(m)}(i) < R^{(m)}(p) \mid R^{(m)}(i) \neq R^{(m)}(p)]$.

Due to the nature of the PL distribution, the following results hold.

Lemma 5. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $S \subseteq \llbracket n \rrbracket$, $|S| = k$, $i, p \in S$, $i \neq p$:*

1. $f_{i|S}^{(p)} \geq \frac{m}{4k}(\theta_i + \theta_p)$.
2. $\mu_{i|S}^{(p)} := \frac{\theta_i}{\theta_i + \theta_p} = \mu_{i|S}^{(p)}$. Furthermore, for any $j \in S \setminus \{p\}$:

$$\begin{aligned} |\mu_i^{(p)} - \mu_j^{(p)}| &\geq \frac{\theta_p}{4} |\theta_i - \theta_j|, \\ |\mu_i^{(p)} - 1/2| &\geq \frac{1}{4} |\theta_i - \theta_p|. \end{aligned}$$

Proof.

1. If $m = k$, then the complete ranking distinguishes i and p , so $f_{i|S}^{(p)} = 1 \geq (\theta_i + \theta_p)/4$. Suppose, thus, that $m \leq k - 1$. Let $I_u = R^{-1}(u)$ for $u \in \llbracket k \rrbracket$ be the u -th element in $R \sim \text{PL}((\theta_i)_{i \in S})$. Let $\mathcal{C}_u^{(i)} := \{i \notin \{I_1, \dots, I_u\}\}$ for every $i \in S$, $u \in \llbracket 0, k \rrbracket$. Finally, let $\Theta_S := \sum_{i \in S} \theta_i$. Then $f_{i|S}^{(p)} = 1 - \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\mathcal{C}_m^{(p)} \cap \mathcal{C}_m^{(i)}]$. Observe that $\mathcal{C}_u^{(i)} \cap \mathcal{C}_v^{(i)} = \mathcal{C}_{\max\{u,v\}}^{(i)}$ for every $u, v \in \llbracket 0, k \rrbracket$. Furthermore:

$$\begin{aligned} \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\mathcal{C}_u^{(p)} \cap \mathcal{C}_u^{(i)} \mid \mathcal{C}_{u-1}^{(p)} \cap \mathcal{C}_{u-1}^{(i)}] &= 1 - \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\overline{\mathcal{C}_u^{(p)}} \cup \overline{\mathcal{C}_u^{(i)}} \mid \mathcal{C}_{u-1}^{(p)} \cap \mathcal{C}_{u-1}^{(i)}] \\ &= 1 - \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [I_u \in \{p, i\} \mid \mathcal{C}_{u-1}^{(p)} \cap \mathcal{C}_{u-1}^{(i)}] \end{aligned}$$

$$= 1 - \frac{\mathbb{E}_{R \sim \text{PL}((\theta_i)_{i \in S})} \left[\frac{\theta_p + \theta_i}{\Theta_S - \sum_{v=1}^{u-1} \theta_{tv}} \mathbb{I}_{\mathcal{C}_{u-1}^{(p)}} \mathbb{I}_{\mathcal{C}_{u-1}^{(i)}} \right]}{\mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\mathcal{C}_{u-1}^{(p)} \cap \mathcal{C}_{u-1}^{(i)}]} \quad (5)$$

$$\leq 1 - \frac{\theta_p + \theta_i}{k} \text{ for every } u \in \llbracket m \rrbracket, \quad (6)$$

where Step (5) follows from the tower rule, conditioning on $I_1, \dots, I_{u-1}$, and Step (6) follows from the fact that $\theta_j \leq 1$ for $j \in S$ and $|S| = k$. Then:

$$\begin{aligned} f_{i|S}^{(p)} &= 1 - \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\mathcal{C}_m^{(p)} \cap \mathcal{C}_m^{(i)}] \\ &= 1 - \prod_{u=1}^m \mathbb{P}_{R \sim \text{PL}((\theta_i)_{i \in S})} [\mathcal{C}_u^{(p)} \cap \mathcal{C}_u^{(i)} \mid \mathcal{C}_{u-1}^{(p)} \cap \mathcal{C}_{u-1}^{(i)}] \\ &\geq 1 - \left(1 - \frac{\theta_p + \theta_i}{k} \right)^m \\ &\geq 1 - \exp \left(-\frac{m}{k} (\theta_p + \theta_i) \right), \end{aligned}$$

where the last line follows from the standard inequality $(1-x)^m \leq e^{-mx}$ for $x \in [0, 1]$, with $x = \frac{\theta_p + \theta_i}{k}$. Finally, the result follows from the fact that $1 - \exp(-x) \geq x/4$ for $x \in [0, 2]$.

2. Yellott (1977) shows that, if $X_i \sim \text{Gumbel}(\ln \theta_i, 1)$ for every $i \in S$, where $\text{Gumbel}(u, 1) \in \Delta(\mathbb{R})$ denotes the Gumbel distribution with CDF $F(x) = \exp(-\exp(-(x-u)))$, and the bijection $R : S \rightarrow \llbracket k \rrbracket$ satisfies:

$$X_{R^{-1}(1)} \geq \dots \geq X_{R^{-1}(k)},$$

then $R \sim \text{PL}((\theta_i)_{i \in S})$ (under arbitrary tie-breaking, as ties occur with probability 0). Let $Y_i = e^{-X_i}$ for $i \in S$. Then $Y_i \sim \text{Exp}(\theta_i)$, where $\text{Exp}(\lambda) \in \Delta(\mathbb{R})$ denotes the exponential distribution with CDF $F(x) = 1 - \exp(-\lambda x)$, and:

$$Y_{R^{-1}(1)} \leq \dots \leq Y_{R^{-1}(k)}.$$

Let $M_{i,p} := \min\{Y_i, Y_p\}$ and $Z_{i,p}$ be the m -th smallest variable among $(Y_j)_{j \in S \setminus \{i,p\}}$ if $m \leq k-2$, and $Z_{i,p} := +\infty$ otherwise. The event $R^{(m)}(i) \neq R^{(m)}(p)$ is precisely $\{M_{i,p} \leq Z_{i,p}\}$. By the exponential-race property, $M_{i,p} \sim \text{Exp}(\theta_i + \theta_p)$ and the indicator $\mathbb{I}[Y_i < Y_p]$ is independent of $M_{i,p}$, with probability $\theta_i/(\theta_i + \theta_p)$. Moreover, $Z_{i,p}$ is independent of (Y_i, Y_p) . Consequently:

$$\mu_{i|S}^{(p)} = \mathbb{P}[Y_i < Y_p \mid M_{i,p} \leq Z_{i,p}] = \frac{\theta_i}{\theta_i + \theta_p}.$$

The additional inequalities in the statement are direct consequences of the expression of $\mu_i^{(p)}$. $\square$

Before delving into the actual concentration results, we need to define a notion of extended history for the TOP- k -UIF algorithm which induces the required natural filtrations. For any $s, t \in \mathbb{N}_{\geq 0}$ with $s \leq t \leq T$, let:^{9,10}

$$\overline{H}_{s:t} := \begin{cases} ((P_u, \mathcal{A}_u, \mathcal{U}_u, S_u, R_u^{(m)})_{u=1}^t, (P_{t+1}, \mathcal{A}_{t+1}, \mathcal{U}_{t+1})) & \text{if } s = 0 \\ ((S_s, R_s^{(m)}), (P_u, \mathcal{A}_u, \mathcal{U}_u, S_u, R_u^{(m)})_{u=s+1}^t, (P_{t+1}, \mathcal{A}_{t+1}, \mathcal{U}_{t+1})) & \text{if } s > 0 \end{cases}.$$

Observe that $\overline{H}_{s:t} = (\overline{H}_{s:u}, \overline{H}_{u+1:t})$ for every $0 \leq s \leq u < t \leq T$. Furthermore, for all $i, p \in \llbracket n \rrbracket$, $i \neq p$, $t \in \llbracket T \rrbracket$, let:

$$O_{i,t}^{(p)} := \mathbb{I} \left[p \in \mathcal{P}_t, i \in S_t, R_t^{(m)}(i) \neq R_t^{(m)}(p) \right],$$

⁹Given two tuples a, b , we denote by (a, b) their concatenation.

¹⁰Note that this specific formulation is fully compatible with the general history H_{t-1} introduced in Section 2, as the algorithm-specific state variables $(P_t, \mathcal{A}_t, \mathcal{U}_t)$ can be formally expressed as functions of the generic internal randomness.

$$W_{i,t}^{(p)} := \mathbb{I} \left[p \in \mathcal{P}_t, i \in S_t, R_t^{(m)}(i) < R_t^{(m)}(p) \right].$$

We use the convention $\mu_i^{(i)} := 1/2$. Observe that, whenever $N_{i,t}^{(p)} > 0$:

$$N_{i,t}^{(p)} = \sum_{s=1}^{t-1} O_{i,s}^{(p)} \text{ and } \hat{\mu}_{i,t}^{(p)} = \frac{1}{N_{i,t}^{(p)}} \sum_{s=1}^{t-1} W_{i,s}^{(p)},$$

for every $i, p \in \llbracket n \rrbracket, i \neq p, t \in \llbracket T+1 \rrbracket$. Finally, for any $T \in \mathbb{N}_{\geq 1}, \delta \in (0, 1]$, define:

$$\beta_{\delta,n,T}(N) := \begin{cases} +\infty & \text{if } N = 0 \\ \sqrt{\frac{1}{N} \ln \left(\frac{2nT}{\delta} \right)} & \text{if } 0 < N < +\infty \\ 0 & \text{if } N = +\infty \end{cases}.$$

This function computes the *confidence bonus* given the number of observations. We now have the required notation for the concentration of $\hat{\mu}_i^{(p)}$.

Lemma 6 (Concentration of $\hat{\mu}_i^{(p)}$). *For every $\nu \in \mathcal{E}_n^{(k,m)}, T \in \mathbb{N}_{\geq 1}, \delta \in (0, 1]$:*

$$\mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| > \beta_{\delta,n,T}(N_{i,t}^{(p)}) \text{ for some } i, p \in \llbracket n \rrbracket, t \in \llbracket T+1 \rrbracket \right] \leq \delta.$$

Proof. Fix $t \in \llbracket T+1 \rrbracket$ and $i, p \in \llbracket n \rrbracket$. Because of the initialization of $\hat{\mu}_i^{(p)}$ and $N_i^{(p)}$ and the definition of $\beta_{\delta,n,T}$, the claim holds when $i = p$ or $N_{i,t}^{(p)} = 0$. Assume $i \neq p$ and $N_{i,t}^{(p)} \geq 1$. Then:

$$|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| = \left| \frac{\sum_{s=1}^{t-1} W_{i,s}^{(p)}}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}} - \frac{\sum_{s=1}^{t-1} O_{i,s}^{(p)} \mu_{i|S_s}^{(p)}}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}} \right| \quad (7)$$

$$= \left| \frac{\sum_{s=1}^{t-1} O_{i,s}^{(p)} (W_{i,s}^{(p)} - \mu_{i|S_s}^{(p)})}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}} \right|, \quad (8)$$

where Step (7) follows from Lemma 5, while Step (8) follows from $W_{i,s}^{(p)} = O_{i,s}^{(p)} W_{i,s}^{(p)}$. Thus:

$$\begin{aligned} |\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| &\geq \beta_{\delta,n,T}(N_{i,t}^{(p)}) \text{ iff} \\ \left| \sum_{s=1}^{t-1} O_{i,s}^{(p)} (W_{i,s}^{(p)} - \mu_{i|S_s}^{(p)}) \right| &\geq \beta_{\delta,n,T}(N_{i,t}^{(p)}) N_{i,t}^{(p)}. \end{aligned} \quad (9)$$

Consider the filtration $\mathbb{F}_i^{(p)} := (\mathcal{F}_{i,u}^{(p)})_{u=0}^T$, where:

$$\mathcal{F}_{i,u}^{(p)} := \begin{cases} \sigma(\bar{H}_{0,u}, S_{u+1}, O_{i,u+1}^{(p)}) & \text{if } u < T \\ \sigma(\bar{H}_{0,T}) & \text{if } u = T \end{cases}.$$

The process:

$$M_{i,u}^{(p)} := \sum_{s=1}^u O_{i,s}^{(p)} (W_{i,s}^{(p)} - \mu_{i|S_s}^{(p)}), \quad u \in \llbracket 0, T \rrbracket,$$

is $\mathbb{F}_i^{(p)}$ -adapted. Moreover, for every $u \in \llbracket T \rrbracket$:

$$\begin{aligned} \mathbb{E}_{\text{TOP-}k\text{-UIF}, \nu} [M_{i,u}^{(p)} \mid \mathcal{F}_{i,u-1}^{(p)}] &= M_{i,u-1}^{(p)} + O_{i,u}^{(p)} \left(\mathbb{E}_{\text{TOP-}k\text{-UIF}, \nu} [W_{i,u}^{(p)} \mid \mathcal{F}_{i,u-1}^{(p)}] - \mu_{i|S_u}^{(p)} \right) \\ &= M_{i,u-1}^{(p)} + O_{i,u}^{(p)} (O_{i,u}^{(p)} - 1) \mu_{i|S_u}^{(p)} = M_{i,u-1}^{(p)}, \end{aligned}$$

where the second equality follows from Lemma 5 and the definition of $W_{i,u}^{(p)}$. Hence, $(M_{i,u}^{(p)})_{u=0}^T$ is a martingale.

Let $\tau_{i,0}^{(p)} := 0$ and, for every $q \in \llbracket T \rrbracket$, recursively define:

$$\tau_{i,q}^{(p)} := \min \left\{ u \in \llbracket \tau_{i,q-1}^{(p)} + 1, t-1 \rrbracket \text{ s.t. } O_{i,u}^{(p)} = 1 \right\},$$

with $\min\{\cdot\} := t - 1$. The variables $(\tau_{i,q}^{(p)})_{q=0}^T$ are nondecreasing bounded stopping times. By the Optional Sampling theorem (Klenke, 2020, Theorem 10.12), $(M_{i,\tau_{i,q}^{(p)}}^{(p)})_{q=0}^N$ is a martingale w.r.t. $(\mathcal{F}_{i,\tau_{i,q}^{(p)}}^{(p)})_{q=0}^N$. For every $q \in \llbracket N \rrbracket$, Lemma 5 and $W_{i,\tau_{i,q}^{(p)}}^{(p)} \in \{0, 1\}$ give:

$$M_{i,\tau_{i,q}^{(p)}}^{(p)} - M_{i,\tau_{i,q-1}^{(p)}}^{(p)} \in \{-\mu_i^{(p)}, 0, 1 - \mu_i^{(p)}\} \subseteq [-\mu_i^{(p)}, 1 - \mu_i^{(p)}].$$

Here the value 0 applies when the q -th observed comparison does not occur before t . Therefore, for every $N \in \llbracket T \rrbracket$, Azuma's inequality (Theorem 15) yields:

$$\begin{aligned} \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu} \left[|M_{i,\tau_{i,N}^{(p)}}^{(p)}| \geq \beta_{\delta,n,T}(N)N \right] &\leq 2 \exp(-2N\beta_{\delta,n,T}^2(N)) \\ &= 2 \left(\frac{\delta}{2nT} \right)^2 \leq \frac{\delta}{(nT)^2}. \end{aligned}$$

On the event $\{N_{i,t}^{(p)} = N\}$, we have $M_{i,\tau_{i,N}^{(p)}}^{(p)} = M_{i,t-1}^{(p)}$, and Step (9) is contained in the event bounded above. The proof follows by a union bound over at most n^2 ordered pairs and at most T^2 choices of (t, N) . $\square$

In order to state the concentration result for $N_i^{(p)}$, we first need to define the predictable process around which the number of observed comparisons concentrates. For every $i, p \in \llbracket n \rrbracket$, $i \neq p$, $t \in \llbracket T \rrbracket$, let:

$$\tilde{O}_{i,t}^{(p)} := \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[O_{i,t}^{(p)} \mid \overline{H}_{0:t-1} \right].$$

Observe that, because of the tower rule and the properties of conditional expectation:

$$\begin{aligned} \tilde{O}_{i,t}^{(p)} &= \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu} \left[p \in \mathcal{P}_t, i \in S_t, R_t^{(m)}(i) \neq R_t^{(m)}(p) \mid \overline{H}_{0:t-1} \right] \\ &= \mathbb{I}[p \in \mathcal{P}_t] \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\mathbb{I}[i \in S_t] \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu} \left[R_t^{(m)}(i) \neq R_t^{(m)}(p) \mid \overline{H}_{0:t-1}, S_t \right] \mid \overline{H}_{0:t-1} \right] \\ &= \mathbb{I}[p \in \mathcal{P}_t] \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\mathbb{I}[i \in S_t] f_{i|S_t}^{(p)} \mid \overline{H}_{0:t-1} \right]. \end{aligned} \quad (10)$$

Let $\tilde{N}_{i,t}^{(p)} := \sum_{s=1}^{t-1} \tilde{O}_{i,s}^{(p)}$ for every $t \in \llbracket T+1 \rrbracket$. We are ready to state and prove the concentration result for $N_i^{(p)}$.

Lemma 7 (Concentration of $N_i^{(p)}$). *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, $\delta \in (0, 1]$, $\lambda > 0$:*

$$\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu} \left[\begin{array}{ll} N_{i,t}^{(p)} \geq \frac{e^\lambda - 1}{\lambda} \tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) & \text{for some } i, p \in \llbracket n \rrbracket, i \neq p, \\ \text{or} & \\ N_{i,t}^{(p)} \leq \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) & \text{and } t \in \llbracket T+1 \rrbracket \end{array} \right] \leq \delta.$$

Proof. For $\lambda \in \mathbb{R}$, let:

$$M_{i,u}^{(p)}(\lambda) := \exp \left(\sum_{s=1}^u (\lambda O_{i,s}^{(p)} - \tilde{O}_{i,s}^{(p)}(e^\lambda - 1)) \right),$$

for every $u \in \llbracket 0, T \rrbracket$. Then, for every $u \in \llbracket T \rrbracket$:

$$\begin{aligned} &\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[M_{i,u}^{(p)}(\lambda) \mid \overline{H}_{0:u-1} \right] \\ &= M_{i,u-1}^{(p)}(\lambda) \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\exp(\lambda O_{i,u}^{(p)} - \tilde{O}_{i,u}^{(p)}(e^\lambda - 1)) \mid \overline{H}_{0:u-1} \right] \\ &= M_{i,u-1}^{(p)}(\lambda) \exp \left(-\tilde{O}_{i,u}^{(p)}(e^\lambda - 1) \right) \left(1 + \tilde{O}_{i,u}^{(p)}(e^\lambda - 1) \right) \\ &\leq M_{i,u-1}^{(p)}(\lambda), \end{aligned} \quad (11)$$

where the equality uses that, conditionally on $\bar{H}_{0:u-1}$, $O_{i,u}^{(p)} \sim \text{Be}(\tilde{O}_{i,u}^{(p)})$,¹¹ and Step (11) follows from $1 + x \leq e^x$. Thus $(M_{i,u}^{(p)}(\lambda))_{u=0}^T$ is a supermartingale w.r.t. the natural filtration induced by the histories $(\sigma(\bar{H}_{0:u}))_{u=0}^T$. Furthermore, $M_{i,0}^{(p)}(\lambda) = 1$. Now, by the maximal inequality (Lattimore and Szepesvári, 2020, Theorem 3.9):

$$\begin{aligned} & \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[M_{i,u}^{(p)}(\lambda) \geq \epsilon \text{ for some } u \in \llbracket 0, T \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right] \\ & \leq \sum_{\substack{i, p \in \llbracket n \rrbracket \\ i \neq p}} \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[M_{i,u}^{(p)}(\lambda) \geq \epsilon \text{ for some } u \in \llbracket 0, T \rrbracket \right] \\ & \leq \sum_{\substack{i, p \in \llbracket n \rrbracket \\ i \neq p}} \frac{\mathbb{E}_{\text{TOP-}k\text{-UIF}, \nu} \left[M_{i,0}^{(p)}(\lambda) \right]}{\epsilon} \leq \frac{n^2}{\epsilon} \text{ for every } \epsilon > 0. \end{aligned}$$

Let $\epsilon = (2n/\delta)^2$. Thus, if $\lambda > 0$:

$$\begin{aligned} \frac{\delta}{2} & \geq \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[M_{i,u}^{(p)}(\lambda) \geq (2n/\delta)^2 \text{ for some } u \in \llbracket 0, T \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right] \\ & = \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[\lambda N_{i,t}^{(p)} - (e^\lambda - 1) \tilde{N}_{i,t}^{(p)} \geq 2 \ln(2n/\delta) \text{ for some } t \in \llbracket T+1 \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right] \\ & = \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[N_{i,t}^{(p)} \geq \frac{e^\lambda - 1}{\lambda} \tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda} \ln(2n/\delta) \text{ for some } t \in \llbracket T+1 \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right]. \end{aligned}$$

Analogously:

$$\begin{aligned} \frac{\delta}{2} & \geq \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[M_{i,u}^{(p)}(-\lambda) \geq (2n/\delta)^2 \text{ for some } u \in \llbracket 0, T \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right] \\ & = \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} \left[N_{i,t}^{(p)} \leq \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda} \ln(2n/\delta) \text{ for some } t \in \llbracket T+1 \rrbracket, i, p \in \llbracket n \rrbracket, i \neq p \right]. \end{aligned}$$

This concludes the proof. $\square$

The concentration results stated in this section induce the following good events:

$$\begin{aligned} \mathcal{G}_{\text{conf}, T}^{(\delta)} & := \left\{ |\hat{\mu}_{i,t}^{(p)} - \mu_{i,t}^{(p)}| \leq \beta_{\delta, n, T}(N_{i,t}^{(p)}) \text{ for every } i, p \in \llbracket n \rrbracket, t \in \llbracket T+1 \rrbracket \right\}, \\ \mathcal{G}_{\text{pulls}}^{(\delta, \lambda)} & := \left\{ \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) < N_{i,t}^{(p)} < \frac{e^\lambda - 1}{\lambda} \tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) \right. \\ & \quad \left. \text{for every } i, p \in \llbracket n \rrbracket, i \neq p, t \in \llbracket T+1 \rrbracket \right\}, \end{aligned}$$

for every $T \in \mathbb{N}_{\geq 1}$, $\delta \in (0, 1]$, $\lambda > 0$.

D.2 Correctness

In this section, we prove that, under the good event $\mathcal{G}_{\text{conf}, T}^{(\delta)}$, TOP- k -UIF accepts only arms $i \in S_{\boldsymbol{\theta}}^{(k)*}$ and rejects only arms $i \notin S_{\boldsymbol{\theta}}^{(k)*}$. In particular, the following result holds.

Lemma 8. *For every $\nu \in \mathcal{E}_n^{(k, m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in \llbracket T+1 \rrbracket$:*

1. $P_t \in \mathcal{A}_t \cup \mathcal{U}_t$. Moreover, if $\mathcal{A}_t \neq \{\}$, then $P_t \in \mathcal{A}_t$.

Furthermore, under the good event $\mathcal{G}_{\text{conf}, T}^{(\delta)}$:

2. $k - |\mathcal{A}_t| = |\mathcal{U}_t \cap S_{\boldsymbol{\theta}}^{(k)*}|$, $|\mathcal{U}_t \cup \mathcal{A}_t| - k = |\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}|$.

¹¹We denote by $\text{Be}(\mu) \in \Delta(\{0, 1\})$ the Bernoulli distribution with mean $\mu \in [0, 1]$.

$$3. \mathcal{A}_t \subseteq S_{\theta}^{(k)*} \subseteq \mathcal{A}_t \cup \mathcal{U}_t.$$

Proof. Remember that, by convention, any variable X_t defined in the pseudocode is evaluated at the beginning of round t .

We first prove Lemma 8.1 by induction. It holds at $t = 1$ because $P_1 \in \mathcal{U}_1$. Suppose first that $\mathcal{A}_t \neq \{\}$. Then $P_t \in \mathcal{A}_t$ by the induction hypothesis, and the algorithm neither changes the pivot nor removes accepted arms, so $P_{t+1} \in \mathcal{A}_{t+1}$.

Suppose instead that $\mathcal{A}_t = \{\}$. Then $\mathcal{P}_t = \{P_t\}$, and the direct relation used by COUNT-BA is transitive: if $i \prec j$ and $j \prec h$, then

$$\text{LCB}_{i,t+1}^{(P_t)} > \text{UCB}_{j,t+1}^{(P_t)} \geq \text{LCB}_{j,t+1}^{(P_t)} > \text{UCB}_{h,t+1}^{(P_t)}.$$

Thus its transitive closure equals the direct relation. After the call to NEW-PIVOT, if the routine returns a candidate, its maximality gives $\text{LCB}_{h,t+1}^{(P_t)} \leq \text{LCB}_{P_{t+1},t+1}^{(P_t)}$ for every $h \in \mathcal{U}_t$. If it returns $\perp$, then $P_{t+1} = P_t$, $\text{LCB}_{P_{t+1},t+1}^{(P_t)} = \text{UCB}_{P_{t+1},t+1}^{(P_t)} = 1/2$, and every $h \in \mathcal{U}_t$ has $\text{LCB}_{h,t+1}^{(P_t)} \leq 1/2$. In either case, no arm precedes P_{t+1} , so P_{t+1} is not rejected. If no arm is accepted in round t , it follows that $P_{t+1} \in \mathcal{U}_{t+1}$. Otherwise, let i be an accepted arm. We have $\text{LCB}_{P_{t+1},t+1}^{(P_t)} \geq \text{LCB}_{i,t+1}^{(P_t)}$, so P_{t+1} reaches every arm reached by i and satisfies the same acceptance threshold. Hence $P_{t+1} \in \mathcal{A}_{t+1}$. This proves Lemma 8.1.

We next prove Lemma 8.3 by induction. It holds at $t = 1$ because $\mathcal{A}_1 = \{\}$ and $\mathcal{U}_1 = \llbracket n \rrbracket$. Suppose it holds at the beginning of round t . Since $\mathcal{A}_t$ and $\mathcal{U}_t$ are disjoint, the induction hypothesis gives:

$$k - |\mathcal{A}_t| = |\mathcal{U}_t \cap S_{\theta}^{(k)*}|, \quad |\mathcal{A}_t \cup \mathcal{U}_t| - k = |\mathcal{U}_t \setminus S_{\theta}^{(k)*}|.$$

Under $\mathcal{G}_{\text{conf},T}^{(\delta)}$, every direct relation $i \prec j$ found in round t by COUNT-BA implies $\theta_i > \theta_j$; the same is therefore true for every pair in its transitive closure. Consequently:

$$B_{i,t+1} \leq |\{j \in \mathcal{U}_t \text{ s.t. } \theta_j > \theta_i\}|, \quad A_{i,t+1} \leq |\{j \in \mathcal{U}_t \text{ s.t. } \theta_j < \theta_i\}|.$$

An optimal unknown arm has at most $|\mathcal{U}_t \cap S_{\theta}^{(k)*}| - 1$ higher-parameter unknown arms, so the rejection rule cannot reject it. Likewise, a suboptimal unknown arm has at most $|\mathcal{U}_t \setminus S_{\theta}^{(k)*}| - 1$ lower-parameter unknown arms, so the first acceptance rule cannot accept it. Finally, when $\mathcal{A}_t \neq \{\}$, Lemma 8.1 and the induction hypothesis give $P_t \in \mathcal{A}_t \subseteq S_{\theta}^{(k)*}$. Thus $\text{LCB}_{i,t+1}^{(P_t)} > 1/2$ implies $\theta_i > \theta_{P_t}$ under $\mathcal{G}_{\text{conf},T}^{(\delta)}$, so the second acceptance rule also accepts only optimal arms. This proves Lemma 8.3 at round $t + 1$.

Lemma 8.2 now follows from Lemma 8.3, the disjointness of $\mathcal{A}_t$ and $\mathcal{U}_t$, and $|S_{\theta}^{(k)*}| = k$. $\square$

To conclude the analysis of the algorithm's correctness, we provide a final technical lemma to guarantee that once the target number of k arms has been accepted, no unknown arms remain. This property bounds the maximum size of the active sets and is a direct, critical consequence of the structural guarantees established in Lemma 8.

Lemma 9. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in \llbracket T + 1 \rrbracket$, under the good event $\mathcal{G}_{\text{conf},T}^{(\delta)}$, if $|\mathcal{A}_t| = k$, then $\mathcal{U}_t = \{\}$. As a consequence, when $|\mathcal{A}_t| = k$, we have $\mathcal{A}_t = S_{\theta}^{(k)*} = \mathcal{A}_t \cup \mathcal{U}_t$.*

Proof. Line 34 of Algorithm 1 sets $\mathcal{U}_t = \{\}$ whenever the preceding round produces $|\mathcal{A}_t| = k$. Lemma 8 then gives $\mathcal{A}_t \subseteq S_{\theta}^{(k)*} \subseteq \mathcal{A}_t \cup \mathcal{U}_t$, and the conclusion follows. $\square$

For any $t \in \llbracket T + 1 \rrbracket$, let $\mathcal{U}'_t := \mathcal{U}_t \setminus \{P_t\}$ and $\mathcal{A}'_t := \mathcal{A}_t \setminus \{P_t\}$.

Corollary 10. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in \llbracket T \rrbracket$, under the good event $\mathcal{G}_{\text{conf},T}^{(\delta)}$, the existence of $j \in \mathcal{U}_t \setminus S_{\theta}^{(k)*}$ implies the existence of $i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}$.*

Proof. We analyze the cases $\mathcal{A}_t = \{\}$ and $\mathcal{A}_t \neq \{\}$ separately.

If $\mathcal{A}_t = \{\}$, Lemma 8 gives $|\mathcal{U}_t \cap S_{\theta}^{(k)*}| = k$. Because $k \geq 2$, even if P_t is optimal, the set $\mathcal{U}'_t = \mathcal{U}_t \setminus \{P_t\}$ contains at least $k - 1 \geq 1$ optimal arms.

If $\mathcal{A}_t \neq \{\}$, the assumed suboptimal unknown arm implies $\mathcal{U}_t \neq \{\}$. By the contrapositive of Lemma 9, $|\mathcal{A}_t| < k$, so Lemma 8 yields $|\mathcal{U}_t \cap S_{\theta}^{(k)*}| = k - |\mathcal{A}_t| \geq 1$. Moreover, Lemma 8.1 gives $P_t \in \mathcal{A}_t$, and hence $\mathcal{U}'_t \cap S_{\theta}^{(k)*} = \mathcal{U}_t \cap S_{\theta}^{(k)*} \neq \{\}$. $\square$

D.3 Upper Bound to the Top- k Pseudo-Regret for a Fixed Pivot

For any $s, t \in \mathbb{N}_{\geq 1}$, $s \leq t$, we define the *top- k pseudo-regret* suffered by a randomized policy π on an instance $\nu \in \mathcal{E}_n^{(k,m)}$ as:

$$\text{PReg}_{\pi, \nu}^{(k)}(s, t) := \sum_{u=s}^t \text{IPReg}_{\pi, \nu}^{(k)}(u),$$

and set $\text{PReg}_{\pi, \nu}^{(k)}(s, t) := 0$ when $s > t$. Here:

$$\text{IPReg}_{\pi, \nu}^{(k)}(t) := g_{\theta}^{(k)}(S_{\theta}^{(k)*}) - \sum_{\substack{S \subseteq [n] \\ |S|=k}} g_{\theta}^{(k)}(S) \pi(S | H_{t-1})$$

is the *top- k instantaneous pseudo-regret*. Observe that, because of the tower rule:

$$\text{EReg}_{\pi, \nu}^{(k)}(T) = \mathbb{E}_{\pi, \nu} \left[\text{PReg}_{\pi, \nu}^{(k)}(1, T) \right].$$

Let $T_0 := 0$. Recursively, whenever $T_{l-1} < T$, define:

$$T_l := \min\{t \in [T] \text{ s.t. } t \geq T_{l-1} + 1 \text{ and } (t = T \text{ or } P_{t+1} \neq P_{T_{l-1}+1})\}.$$

These times split the horizon into epochs with fixed pivot $P^{(l)} := P_{T_{l-1}+1} = P_t$ for $T_{l-1} + 1 \leq t \leq T_l$. Let $L(T) := \min\{l \in \mathbb{N}_{\geq 1} \text{ s.t. } T_l = T\}$. The goal of this section is to upper bound the top- k pseudo-regret suffered by TOP- k -UIF during an epoch $l \in [L(T)]$, inside the learning horizon:

$$\text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1} + 1, \min\{T_l, T\}),$$

under both the good events $\mathcal{G}_{\text{conf}, T}^{(\delta)}$ and $\mathcal{G}_{\text{pulls}}^{(\delta, \lambda)}$. Let us start with an upper bound to the instantaneous pseudo-regret under $\mathcal{G}_{\text{conf}, T}^{(\delta)}$.

Lemma 11 (Upper Bound to the Top- k Instantaneous Pseudo-Regret). *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, $\delta \in (0, 1]$, $t \in [T]$, under the good event $\mathcal{G}_{\text{conf}, T}^{(\delta)}$:*

$$\text{IPReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(t) \leq \sum_{i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}} \frac{1}{k|\mathcal{U}'_t|} \left[\sum_{\substack{j \in \mathcal{U}_t \setminus S_{\theta}^{(k)*} \\ \theta_j \geq \theta_{P_t}}} (\theta_i - \theta_{P_t}) + \sum_{\substack{j \in \mathcal{U}_t \setminus S_{\theta}^{(k)*} \\ \theta_j < \theta_{P_t}}} (\theta_i - \theta_j) \right],$$

where the right-hand side is understood to be zero if $\mathcal{U}'_t = \{\}$.

Proof. If $\mathcal{U}'_t = \{\}$, Lemmas 8 and 9 imply that the played set is $S_{\theta}^{(k)*}$, so the result is immediate. Suppose $\mathcal{U}'_t \neq \{\}$.

If $P_t \in S_{\theta}^{(k)*}$, every arm in $\mathcal{P}_t$ is optimal, and the number of uniformly sampled arms is $|\mathcal{U}'_t \cap S_{\theta}^{(k)*}|$. Lemma 16 therefore gives the exact identity:

$$\text{IPReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(t) = \frac{1}{k|\mathcal{U}'_t|} \sum_{i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}} \sum_{j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*}} (\theta_i - \theta_j).$$

Splitting the inner sum according to whether $\theta_j \geq \theta_{P_t}$ and using $\theta_i - \theta_j \leq \theta_i - \theta_{P_t}$ in the first part yields the claimed bound.

Suppose instead that $P_t \notin S_{\theta}^{(k)*}$. Lemma 8.1 implies $\mathcal{A}_t = \{\}$, so $\mathcal{P}_t = \{P_t\}$, $\mathcal{U}'_t \cap S_{\theta}^{(k)*} = S_{\theta}^{(k)*}$, and $k-1$ arms are sampled uniformly from $\mathcal{U}'_t$. Direct expansion gives:

$$k|\mathcal{U}'_t| \text{IPReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(t) = (|\mathcal{U}'_t \setminus S_{\theta}^{(k)*}| + 1) \sum_{i \in S_{\theta}^{(k)*}} \theta_i - |\mathcal{U}'_t| \theta_{P_t} - (k-1) \sum_{j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*}} \theta_j.$$

Multiplying the claimed upper bound by $k|\mathcal{U}'_t|$, and subtracting $k|\mathcal{U}'_t| \text{IPReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(t)$, gives:

$$(k-1) \sum_{\substack{j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*} \\ \theta_j \geq \theta_{P_t}}} (\theta_j - \theta_{P_t}) + \sum_{\substack{j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*} \\ \theta_j < \theta_{P_t}}} (\theta_{P_t} - \theta_j) \geq 0.$$

This proves the result in the second case as well. $\square$

We proceed with the analysis by upper bounding the gaps $\theta_i - \theta_{P_t}$ and $\theta_i - \theta_j$ in the instantaneous regret upper bound provided by Lemma 11 with the confidence radii, exploiting how unknown arms and the pivot are updated by TOP- k -UIF. To this end, we define the following burn-in term:

$$t_n^{(k,m,\delta,\lambda)} := \frac{64}{1-e^{-\lambda}} \frac{nk}{m} \ln \left(\frac{2n}{\delta} \right),$$

and the following predictable process:

$$\tilde{Q}_{s:t} := \begin{cases} \sum_{u=s}^t \tilde{q}_u & \text{if } s \leq t \\ 0 & \text{if } s > t \end{cases}, \quad \tilde{q}_t := \begin{cases} \frac{k - |\mathcal{P}_t|}{|\mathcal{U}'_t|} & \text{if } \mathcal{U}'_t \neq \{\} \\ 0 & \text{if } \mathcal{U}'_t = \{\} \end{cases}, \quad s, t \in \llbracket T \rrbracket,$$

which we exploit in the result below for the upper bounds to hold. Observe that $\tilde{q}_t$ is the probability of pulling an arm in $\mathcal{U}'_t$ as remarked in the proof of Lemma 11.

Lemma 12 (Upper Bound on Gaps via Confidence Radii). *Let $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, $t \in \llbracket T \rrbracket$, $\delta \in (0, 1]$, $\lambda > 0$. If there exist $T_{\text{start}} \in \llbracket T \rrbracket$ and $p \in \llbracket n \rrbracket$ s.t. $t - T_{\text{start}} + 1 \geq t_n^{(k,m,\delta,\lambda)}$ and $p \in \mathcal{P}_u$ for all $u \in \llbracket T_{\text{start}}, t \rrbracket$, then, under both good events $\mathcal{G}_{\text{conf}, T}^{(\delta)}$ and $\mathcal{G}_{\text{pulls}}^{(\delta, \lambda)}$:*

1. *If $\theta_p \geq 1/2$, for any $i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}$, $j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*}$:*

$$\theta_i - \theta_j \leq 64 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \beta_{\delta, n, T}(\tilde{Q}_{T_{\text{start}}:t}).$$

2. *If, additionally, $p = P_u$ for all $u \in \llbracket T_{\text{start}}, t \rrbracket$, then, for any $i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}$ such that $\theta_i > \theta_p$ and either $\theta_i \geq 1/2$ or $\theta_p \geq 1/2$:*

$$\theta_i - \theta_p \leq 8 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \beta_{\delta, n, T}(\tilde{Q}_{T_{\text{start}}:t}).$$

Proof. For every $u \in \llbracket T_{\text{start}}, t \rrbracket$, Corollary 10 gives $\mathcal{U}'_u \cap S_{\theta}^{(k)*} \neq \{\}$ in Lemma 12.1, since $j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*} \subseteq \mathcal{U}_u \setminus S_{\theta}^{(k)*}$. In Lemma 12.2, the same conclusion follows because i remains unknown and $i \neq p = P_u$. Lemma 8 therefore gives $k - |\mathcal{P}_u| \geq 1$. Since $|\mathcal{U}'_u| \leq n$, it follows that $\tilde{q}_u \geq 1/n$. Then:

$$\tilde{Q}_{T_{\text{start}}:t} = \sum_{u=T_{\text{start}}}^t \tilde{q}_u \geq \frac{1}{n}(t - T_{\text{start}} + 1) \geq \frac{1}{n} t_n^{(k,m,\delta,\lambda)} \quad (12)$$

$$= \frac{64}{1-e^{-\lambda}} \frac{k}{m} \ln \left(\frac{2n}{\delta} \right) \geq 2 \geq 2\tilde{q}_t, \quad (13)$$

where Step (13) follows since $\tilde{q}_t$ is a probability. The above inequality implies:

$$\tilde{Q}_{T_{\text{start}}:t-1} = \tilde{Q}_{T_{\text{start}}:t} - \tilde{q}_t \geq \frac{1}{2} \tilde{Q}_{T_{\text{start}}:t}.$$

Then, for any $a \in \mathcal{U}'_{t-1}$:

$$\tilde{N}_{a,t}^{(p)} = \sum_{u=1}^{t-1} \tilde{O}_{a,u}^{(p)} \geq \sum_{u=T_{\text{start}}}^{t-1} \tilde{O}_{a,u}^{(p)} \geq \frac{m}{4k}(\theta_a + \theta_p) \sum_{u=T_{\text{start}}}^{t-1} \mathbb{P}_{\text{TOP-}k\text{-UIF}, \nu} [a \in S_u \mid \bar{H}_{0:u-1}] \quad (14)$$

$$\geq \frac{m}{4k}(\theta_a + \theta_p) \sum_{u=T_{\text{start}}}^{t-1} \tilde{q}_u = \frac{m}{4k}(\theta_a + \theta_p) \tilde{Q}_{T_{\text{start}}:t-1} \geq \frac{m}{8k}(\theta_a + \theta_p) \tilde{Q}_{T_{\text{start}}:t}, \quad (15)$$

where Step (14) follows from Equation (10) and Lemma 5. Furthermore, by combining Equations (12) and (15):

$$\frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) = \frac{1 - e^{-\lambda}}{32\lambda} \frac{m}{k} \frac{1}{n} t_n^{(k,m,\delta,\lambda)} \leq \frac{1 - e^{-\lambda}}{4\lambda} \frac{m}{8k} \tilde{Q}_{T_{\text{start}}:t} \leq \frac{1}{4(\theta_a + \theta_p)} \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{a,t}^{(p)}. \quad (16)$$

1. Let $j \in \mathcal{U}'_t \setminus S_{\theta}^{(k)*}$. Because of how the algorithm updates unknown arms, the fact that $i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}$ implies the existence of $f \in \mathcal{U}_{t-1} \setminus S_{\theta}^{(k)*}$ such that:

$$\text{LCB}_{i,t}^{(p)} \leq \text{UCB}_{f,t}^{(p)}, \quad (17)$$

otherwise we would have $A_{i,t} \geq |\mathcal{U}_{t-1} \setminus S_{\theta}^{(k)*}| = |\mathcal{A}_{t-1} \cup \mathcal{U}_{t-1}| - k$ by Lemma 8, implying $i \in \mathcal{A}_t$. Analogously, there must exist $g \in \mathcal{U}_{t-1} \cap S_{\theta}^{(k)*}$ such that:

$$\text{UCB}_{j,t}^{(p)} \geq \text{LCB}_{g,t}^{(p)}. \quad (18)$$

Thus:

$$\theta_i - \theta_j \leq \frac{4}{\theta_p} (\mu_i^{(p)} - \mu_j^{(p)}) \leq 8(\text{UCB}_{i,t}^{(p)} - \text{LCB}_{j,t}^{(p)}) \quad (19)$$

$$\begin{aligned} &= 8(\text{UCB}_{i,t}^{(p)} - \text{LCB}_{i,t}^{(p)} + \text{UCB}_{j,t}^{(p)} - \text{LCB}_{j,t}^{(p)} + \text{LCB}_{i,t}^{(p)} - \text{UCB}_{j,t}^{(p)}) \\ &\leq 8(2\beta_{\delta,n,T}(N_{i,t}^{(p)}) + 2\beta_{\delta,n,T}(N_{j,t}^{(p)}) + \text{UCB}_{f,t}^{(p)} - \text{LCB}_{g,t}^{(p)}) \end{aligned} \quad (20)$$

$$\begin{aligned} &\leq 16(\beta_{\delta,n,T}(N_{i,t}^{(p)}) + \beta_{\delta,n,T}(N_{j,t}^{(p)}) + \beta_{\delta,n,T}(N_{f,t}^{(p)}) + \beta_{\delta,n,T}(N_{g,t}^{(p)})) \\ &\quad + 8(\mu_f^{(p)} - \mu_g^{(p)}) \end{aligned} \quad (21)$$

$$< 16(\beta_{\delta,n,T}(N_{i,t}^{(p)}) + \beta_{\delta,n,T}(N_{j,t}^{(p)}) + \beta_{\delta,n,T}(N_{f,t}^{(p)}) + \beta_{\delta,n,T}(N_{g,t}^{(p)})), \quad (22)$$

where Step (19) follows from Lemma 5, Step (20) follows from Equations (17) and (18), Step (21) follows from the good event $\mathcal{G}_{\text{conf},T}^{(\delta)}$, and Step (22) follows from $\theta_f < \theta_g$, since $f \notin S_{\theta}^{(k)*}$ and $g \in S_{\theta}^{(k)*}$. If $a = p$, then $N_{a,t}^{(p)} = +\infty$ and the corresponding confidence bonus is zero. For every $a \in \{i, j, f, g\} \setminus \{p\}$, Equations (16) and (15), together with $\theta_p \geq 1/2$, imply under $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ that:

$$N_{a,t}^{(p)} > \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{a,t}^{(p)} - \frac{2}{\lambda} \ln \left(\frac{2n}{\delta} \right) \geq \frac{1 - e^{-\lambda}}{2\lambda} \tilde{N}_{a,t}^{(p)} \geq \frac{1 - e^{-\lambda}}{32\lambda} \frac{m}{k} \tilde{Q}_{T_{\text{start}}:t}.$$

Hence, by the monotonicity of $\beta_{\delta,n,T}$:

$$\theta_i - \theta_j < 64 \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sqrt{\frac{k}{m}} \beta_{\delta,n,T}(\tilde{Q}_{T_{\text{start}}:t}).$$

2. In this case, we must have:

$$\text{LCB}_{i,t}^{(p)} \leq 1/2.$$

Otherwise, if $\mathcal{A}_{t-1} \neq \{\}$, we would have $i \in \mathcal{A}_t$ (contradicting $i \in \mathcal{U}'_t$), while, if $\mathcal{A}_{t-1} = \{\}$, the NEW-PIVOT routine would return an arm different from p in round $t - 1$ (contradicting $P_t = p$). Then, under the good event $\mathcal{G}_{\text{conf},T}^{(\delta)}$, Lemma 5 gives:

$$\theta_i - \theta_p \leq 4(\mu_i^{(p)} - 1/2) \leq 8\beta_{\delta,n,T}(N_{i,t}^{(p)}).$$

Analogously to the argument for Lemma 12.1, since either $\theta_p \geq 1/2$ or $\theta_i \geq 1/2$, we have under $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ that:

$$N_{i,t}^{(p)} > \frac{1 - e^{-\lambda}}{32\lambda} \frac{m}{k} \tilde{Q}_{T_{\text{start}}:t}.$$

The result follows again by the monotonicity of $\beta_{\delta,n,T}$.

□

Now, we combine Lemmas 11 and 12 to obtain both an instance-dependent and a worst-case upper bound on the top- k pseudo-regret suffered during an epoch. To this end, for every $i \in \llbracket n \rrbracket$, define the auxiliary last-active time:

$$T_{\mathcal{U},i} := \max\{t \in \llbracket T \rrbracket \text{ s.t. } i \in \mathcal{U}_t\} \in \llbracket T \rrbracket.$$

Lemma 13 (Upper Bound to the Top- k Pseudo-Regret for a Fixed Pivot). *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, $\delta \in (0, 1]$, $\lambda > 0$, and $l \in \llbracket L(T) \rrbracket$, under both good events $\mathcal{G}_{\text{conf},T}^{(\delta)}$ and $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$:*

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T_{l-1} + 1, \min\{T_l, T\}) &\leq 2t_n^{(k,m,\delta,\lambda)} \\ &+ \frac{2^{20}\lambda}{1 - e^{-\lambda}} \ln\left(\frac{2nT}{\delta}\right) \frac{1}{m} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}, \end{aligned}$$

and:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T_{l-1} + 1, \min\{T_l, T\}) &\leq 2t_n^{(k,m,\delta,\lambda)} \\ &+ 2^9 \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sqrt{\frac{n-k}{m}} \sqrt{(\min\{T_l, T\} - T_{l-1}) \ln\left(\frac{2nT}{\delta}\right)}. \end{aligned}$$

Proof.

Case 1: $\theta_{P^{(l)}} \geq 1/2$. Then, by Lemmas 11 and 12:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T_{l-1} + 1, \min\{T_l, T\}) &\leq \lceil t_n^{(k,m,\delta,\lambda)} \rceil - 1 \\ &+ \mathbb{I}[\min\{T_l, T\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}] \sum_{t=T_{l-1} + \lceil t_n^{(k,m,\delta,\lambda)} \rceil}^{\min\{T_l, T\}} \\ &\sum_{j \in \mathcal{U}_t \setminus S_{\theta}^{(k)*}} \sum_{i \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}} \frac{2^6}{\sqrt{km}|\mathcal{U}'_t|} \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}) \end{aligned} \quad (23)$$

$$\leq t_n^{(k,m,\delta,\lambda)} + \frac{2^6 \mathbb{I}[\min\{T_l, T\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}]}{\sqrt{km}} \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}}$$

$$\begin{aligned} &\sum_{t=T_{l-1} + \lceil t_n^{(k,m,\delta,\lambda)} \rceil}^{\min\{T_l, T\}} \frac{|\mathcal{U}'_t \cap S_{\theta}^{(k)*}|}{|\mathcal{U}'_t|} \beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}) \mathbb{I}[j \in \mathcal{U}_t] \\ &= t_n^{(k,m,\delta,\lambda)} + \frac{2^6 \mathbb{I}[\min\{T_l, T\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}]}{\sqrt{km}} \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \end{aligned} \quad (24)$$

$$\begin{aligned} &\sum_{t=T_{l-1} + \lceil t_n^{(k,m,\delta,\lambda)} \rceil}^{\min\{T_l, T\}} \frac{k - |\mathcal{A}'_t| - \mathbb{I}[P^{(l)} \in S_{\theta}^{(k)*}]}{|\mathcal{U}'_t|} \beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}) \mathbb{I}[j \in \mathcal{U}_t] \\ &\leq t_n^{(k,m,\delta,\lambda)} + \frac{2^6 \mathbb{I}[\min\{T_l, T\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}]}{\sqrt{km}} \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \end{aligned}$$

$$\begin{aligned} &\sum_{t=T_{l-1} + \lceil t_n^{(k,m,\delta,\lambda)} \rceil}^{\min\{T_l, T\}} \frac{2(k - |\mathcal{P}_t|)}{|\mathcal{U}'_t|} \beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}) \mathbb{I}[j \in \mathcal{U}_t] \\ &\leq t_n^{(k,m,\delta,\lambda)} + \frac{2^7}{\sqrt{km}} \sqrt{\frac{32\lambda}{1 - e^{-\lambda}}} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \mathbb{I}[\min\{T_l, T_{\mathcal{U},j}, T\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}] \end{aligned} \quad (25)$$

$$\begin{aligned}
& \sum_{t=T_{l-1}+1}^{\min\{T_l, T_{\mathcal{U},j}, T\}} \tilde{q}_t \beta_{\delta, n, T}(\tilde{Q}_{T_{l-1}+1:t}) \\
& \leq t_n^{(k, m, \delta, \lambda)} + \frac{2^8}{\sqrt{km}} \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \ln\left(\frac{2nT}{\delta}\right) \\
& \quad \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \mathbb{I}[\min\{T_l, T_{\mathcal{U},j}, T\} \geq T_{l-1} + t_n^{(k, m, \delta, \lambda)}] \\
& \quad \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}}}{\ln(2nT/\delta)}}. \tag{26}
\end{aligned}$$

where Step (23) follows by applying the upper bound to the gaps developed in Lemma 12 to both the sums in Lemma 11, using Lemma 12.2 for the left sum and Lemma 12.1 for the right sum, and Step (24) follows by the expression for $|\mathcal{U}'_t \cap S_{\theta}^{(k)*}|$ derived at the beginning of the proof of Lemma 11. For Step (25), the presence of $j \in \mathcal{U}_t$ and Lemma 9 give $|\mathcal{A}_t| < k$. If $\mathcal{A}_t = \{\}$, then $\mathcal{A}'_t = \{\}$ and $|\mathcal{P}_t| = 1$, so:

$$k - |\mathcal{A}'_t| - \mathbb{I}[P^{(l)} \in S_{\theta}^{(k)*}] \leq k \leq 2(k-1) = 2(k - |\mathcal{P}_t|).$$

If $\mathcal{A}_t \neq \{\}$, Lemma 8 gives $P_t \in \mathcal{A}_t \subseteq S_{\theta}^{(k)*}$, whence $\mathcal{P}_t = \mathcal{A}_t$ and:

$$k - |\mathcal{A}'_t| - \mathbb{I}[P^{(l)} \in S_{\theta}^{(k)*}] = k - |\mathcal{A}_t| \leq 2(k - |\mathcal{P}_t|).$$

Finally, Step (26) follows from (Orabona, 2019, Lemma 4.13).

We first derive the instance-dependent bound. For every $j \notin S_{\theta}^{(k)*}$ whose indicator in Equation (26) is one, we have $j \in \mathcal{U}_{\min\{T_l, T_{\mathcal{U},j}, T\}}$, and Corollary 10 provides an arm $i_j \in \mathcal{U}'_{\min\{T_l, T_{\mathcal{U},j}, T\}} \cap S_{\theta}^{(k)*}$. If $\theta_j \geq \theta_{P^{(l)}}$, Lemma 12.2 bounds $\theta_{i_j} - \theta_{P^{(l)}}$; if $\theta_j < \theta_{P^{(l)}}$, then $j \in \mathcal{U}'_{\min\{T_l, T_{\mathcal{U},j}, T\}}$ and Lemma 12.1 bounds $\theta_{i_j} - \theta_j$. In both cases:

$$\theta_{r_{\theta}^{*-1}(k)} - \theta_j \leq 64 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \beta_{\delta, n, T}(\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}}).$$

Therefore:

$$\sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}}}{\ln(2nT/\delta)}} \leq 64 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}.$$

Substitution into Equation (26) gives:

$$\begin{aligned}
& \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1}+1, \min\{T_l, T\}) \leq t_n^{(k, m, \delta, \lambda)} \\
& + \frac{2^{19}\lambda}{1-e^{-\lambda}} \ln\left(\frac{2nT}{\delta}\right) \frac{1}{m} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}.
\end{aligned}$$

For the worst-case bound, we return to Equation (26) and apply the Cauchy-Schwarz inequality to the sum over the suboptimal arms. Since $|\llbracket n \rrbracket \setminus S_{\theta}^{(k)*}| = n - k$, we have:

$$\begin{aligned}
& \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \mathbb{I}[\min\{T_l, T_{\mathcal{U},j}, T\} \geq T_{l-1} + t_n^{(k, m, \delta, \lambda)}] \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}}}{\ln(2nT/\delta)}} \\
& \leq \sqrt{n-k} \sqrt{\sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}}}{\ln\left(\frac{2nT}{\delta}\right)}}. \tag{27}
\end{aligned}$$

To bound the sum of the predictable processes inside the square root, we unpack the definition of $\tilde{Q}_{s:t}$, swap the sums, and use $|\mathcal{U}_t \setminus S_{\theta}^{(k)*}| \leq |\mathcal{U}'_t|$. This inequality is immediate when

P_t is optimal; when P_t is suboptimal, Lemma 8 gives $\mathcal{A}_t = \{\}$ and $\mathcal{U}_t$ contains all $k \geq 2$ optimal arms, so it also holds. Thus:

$$\begin{aligned} \sum_{j \in \llbracket n \rrbracket \setminus S_{\theta}^{(k)*}} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}, T\}} &= \sum_{t=T_{l-1}+1}^{\min\{T_l, T\}} \tilde{q}_t |\mathcal{U}_t \setminus S_{\theta}^{(k)*}| \\ &\leq \sum_{t=T_{l-1}+1}^{\min\{T_l, T\}} k = k(\min\{T_l, T\} - T_{l-1}). \end{aligned} \quad (28)$$

Substituting Equation (28) back into Equation (27), and then plugging the result into Equation (26), the $\sqrt{k}$ cancels out cleanly and the logarithmic term simplifies, yielding the final gap-independent worst-case bound:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1} + 1, \min\{T_l, T\}) &\leq t_n^{(k,m,\delta,\lambda)} \\ &\quad + 2^8 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{n-k}{m}} \sqrt{(\min\{T_l, T\} - T_{l-1}) \ln \left(\frac{2nT}{\delta} \right)}. \end{aligned}$$

Case 2: $\theta_{P^{(l)}} < 1/2$. Let $i_1^* := r_{\theta}^{*-1}(1)$ denote the best overall arm, which by definition has $\theta_{i_1^*} = 1$. We split the epoch into two phases at

$$T_{\text{best}} := \max\{T_{l-1}, \min\{T_l, T_{\mathcal{U}, i_1^*}, T\}\}.$$

Thus, the first phase is empty if $T_{\text{best}} = T_{l-1}$, while otherwise T_{best} is its final round and i_1^* remains in the unknown set throughout that phase.

If $T_{\text{best}} + 1 \leq \min\{T_l, T\}$, then i_1^* has been accepted and belongs to $\mathcal{A}_t \subseteq \mathcal{P}_t$ for every round t in the second phase. Since the pivot has not changed within the epoch, Lemma 8 also gives $P^{(l)} \in \mathcal{A}_t \subseteq S_{\theta}^{(k)*}$. Therefore, every suboptimal arm j satisfies $\theta_j < \theta_{P^{(l)}}$, and the first inner sum in Lemma 11 is empty. Because $\theta_{i_1^*} = 1 \geq 1/2$, Lemma 12.1, with $p = i_1^*$ and $T_{\text{start}} = T_{\text{best}} + 1$, bounds every remaining gap $\theta_i - \theta_j$ after the corresponding burn-in. Applying Steps (23)–(26) on $\llbracket T_{\text{best}} + 1, \min\{T_l, T\} \rrbracket$, with $\tilde{Q}_{T_{\text{best}}+1:t}$ in place of $\tilde{Q}_{T_{l-1}+1:t}$, yields:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{\text{best}} + 1, \min\{T_l, T\}) &\leq t_n^{(k,m,\delta,\lambda)} \\ &\quad + \frac{2^{19}\lambda}{1-e^{-\lambda}} \ln \left(\frac{2nT}{\delta} \right) \frac{1}{m} \sum_{j \notin S_{\theta}^{(k)*}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}, \\ \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{\text{best}} + 1, \min\{T_l, T\}) &\leq t_n^{(k,m,\delta,\lambda)} \\ &\quad + 2^8 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{n-k}{m}} \sqrt{(\min\{T_l, T\} - T_{\text{best}}) \ln \left(\frac{2nT}{\delta} \right)}. \end{aligned}$$

If $T_{\text{best}} + 1 > \min\{T_l, T\}$, this phase is empty. We now bound the first phase, where $t \in \llbracket T_{l-1} + 1, T_{\text{best}} \rrbracket$.

For $t \leq T_{\text{best}}$, because $\theta_{P^{(l)}} < 1/2$ and $\theta_{i_1^*} = 1$, we have $2(\theta_{i_1^*} - \theta_{P^{(l)}}) > 1$. Furthermore, since all PL parameters are bounded in $(0, 1]$, any gap $\theta_i - \theta_j$ or $\theta_i - \theta_{P^{(l)}}$ appearing in the instantaneous regret bound of Lemma 11 is strictly bounded by 1. We can thus uniformly upper bound every gap in Lemma 11 by $2(\theta_{i_1^*} - \theta_{P^{(l)}})$.

Because $i_1^* \in \mathcal{U}'_t \cap S_{\theta}^{(k)*}$ is still alive during this phase, we can apply Lemma 12.2 directly to i_1^* and the pivot $P^{(l)}$, provided the burn-in phase is satisfied (which we track via the indicator function). Multiplying the result by 2, we obtain:

$$2(\theta_{i_1^*} - \theta_{P^{(l)}}) \leq 16 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}).$$

Substituting this uniform bound into Lemma 11 and applying Steps (23)–(26) with upper endpoint T_{best} gives:

$$\text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1} + 1, T_{\text{best}}) \leq t_n^{(k,m,\delta,\lambda)} + \frac{2^6}{\sqrt{km}} \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \ln \left(\frac{2nT}{\delta} \right)$$

$$\sum_{j \in [n] \setminus S_{\theta}^{(k)*}} \mathbb{I}[\min\{T_{\text{best}}, T_{\mathcal{U},j}\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}] \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_{\text{best}}, T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}}. \quad (29)$$

For the worst-case bound, Cauchy–Schwarz and Equation (28), with T_{best} in place of $\min\{T_l, T\}$, give:

$$\begin{aligned} & \sum_{j \notin S_{\theta}^{(k)*}} \mathbb{I}[\min\{T_{\text{best}}, T_{\mathcal{U},j}\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}] \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_{\text{best}}, T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}} \\ & \leq \sqrt{\frac{k(n-k)(T_{\text{best}} - T_{l-1})}{\ln(2nT/\delta)}}. \end{aligned}$$

Substitution into Equation (29) yields:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1} + 1, T_{\text{best}}) & \leq t_n^{(k,m,\delta,\lambda)} \\ & + 2^6 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{n-k}{m}} \sqrt{(T_{\text{best}} - T_{l-1}) \ln\left(\frac{2nT}{\delta}\right)}. \end{aligned}$$

For the instance-dependent bound, the indicator guarantees the burn-in condition. Since $\beta_{\delta,n,T}(Q) = \sqrt{\ln(2nT/\delta)/Q}$ for $Q > 0$, the displayed gap inequality gives:

$$\sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_{\text{best}}, T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}} \leq 8 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \sqrt{\frac{k}{m}} \frac{1}{\theta_{i_1^*} - \theta_{P^{(l)}}}.$$

Since $1 \leq 2(\theta_{i_1^*} - \theta_{P^{(l)}})$ and $0 < \theta_{r_{\theta}^{*-1}(k)} - \theta_j \leq 1$, we have:

$$\frac{1}{\theta_{i_1^*} - \theta_{P^{(l)}}} \leq 2 \leq \frac{2}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}.$$

Substitution into Equation (29) gives:

$$\begin{aligned} \text{PReg}_{\text{TOP-}k\text{-UIF}, \nu}^{(k)}(T_{l-1} + 1, T_{\text{best}}) & \leq t_n^{(k,m,\delta,\lambda)} \\ & + \frac{2^{15}\lambda}{1-e^{-\lambda}} \ln\left(\frac{2nT}{\delta}\right) \frac{1}{m} \sum_{j \in [n] \setminus S_{\theta}^{(k)*}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}. \end{aligned}$$

In the case $\theta_{P^{(l)}} \geq 1/2$, the entire epoch contributes one burn-in term and one regret term. In the case $\theta_{P^{(l)}} < 1/2$, each of the two phases contributes at most one burn-in term and one regret term; the first-phase constants satisfy $2^{15} \leq 2^{19}$ and $2^6 \leq 2^8$. Summing the two phases therefore gives the factor 2 in both bounds stated in the lemma. $\square$

D.4 Proof of Theorem 2

The goal of this section is to use the results developed in the rest of the appendix in order to prove Theorem 2. For each T , the expectation below is taken for TOP- k -UIF run with horizon T and $\delta = 1/T$, and $\mathcal{G}_{\text{conf},T}^{(\delta)}$ denotes the corresponding event. We define the following instance-dependent and worst-case upper bounds to the number of pivots under the good event:

$$\begin{aligned} L_{\nu} &:= \sup_{T \in \mathbb{N}_{\geq 1}} \mathbb{E}_{\text{TOP-}k\text{-UIF}, \nu} \left[\mathbb{I}_{\mathcal{G}_{\text{conf},T}^{(\delta)}} L(T) \right], \\ L_{\mathcal{E}_n^{(k,m)}} &:= \sup_{\nu \in \mathcal{E}_n^{(k,m)}} L_{\nu}. \end{aligned}$$

Together with Lemma 13, these expressions allow us to derive regret upper bounds over the entire horizon $[T]$. Under the good event $\mathcal{G}_{\text{conf},T}^{(\delta)}$, because of the NEW-PIVOT routine (see Algorithm 2), we have $\theta_{P^{(1)}} < \dots < \theta_{P^{L(T)}}$, implying that the pivots are all distinct. Thus, we can write:

$$L(T) = \sum_{i \in [n]} \mathbb{I}[P_t = i \text{ for some } t \in [T]].$$

At this point, we clearly have:

$$L_\nu \leq n \text{ for every } \nu \in \mathcal{E}_n^{(k,m)} \text{ and thus } L_{\mathcal{E}_n^{(k,m)}} \leq n. \quad (30)$$

We are ready to prove Theorem 2.

Theorem 2 (Upper Bounds to the Top- k Regret). *For every $n, k \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}_{\geq 1}$ with $m \leq k \leq n$, instance $\nu \in \mathcal{E}_n^{(k,m)}$ and learning horizon $T \in \mathbb{N}_{\geq n}$, by setting the algorithm parameters to $\beta_{\delta,n,T}(N) = \sqrt{\ln(2nT/\delta)/N}$ for $N \in \mathbb{N}_{\geq 1}$ (with $\beta_{\delta,n,T}(0) := +\infty$ and $\beta_{\delta,n,T}(+\infty) := 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top- k regret suffered by TOP- k -UIF:*

$$\text{EReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T) \leq C_2 \frac{L_\nu}{m} \left(nk + \sum_{u \in \llbracket k+1, n \rrbracket} 1/\Delta_{k,u} \right) \ln(T),$$

where $C_2 \in \mathbb{R}_{>0}$ is an absolute constant, L_ν is an upper bound to the number of pivots chosen by the TOP- k -UIF algorithm over the learning horizon (formally defined in Appendix D) and $\Delta_{k,u} := \theta_{r_\theta^{*-1}(k)} - \theta_{r_\theta^{*-1}(u)}$ for all $u \in \llbracket k+1, n \rrbracket$. If $k = n$, the regret is identically zero. If $k < n$, for a sufficiently large learning horizon satisfying $T \geq \frac{(nk)^2}{m(n-k)} L_{\mathcal{E}_n^{(k,m)}} \ln(2nT)$, the worst-case expected top- k regret is bounded by:

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \text{EReg}_{\text{TOP-}k\text{-UIF},\nu}^{(k)}(T) \leq C_3 \sqrt{L_{\mathcal{E}_n^{(k,m)}} \frac{n-k}{m} T \ln(T)},$$

where $C_3 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\mathcal{E}_n^{(k,m)}} := \sup_{\nu \in \mathcal{E}_n^{(k,m)}} L_\nu \leq n$ (see Appendix D).

Proof of Theorem 2. Let us set $\lambda = 1$. The instance-dependent bound on the expected regret follows directly from the first result in Lemma 13 and the definition of L_ν through algebraic manipulation. Because $\delta = 1/T$, a union bound shows that the expected regret incurred when either $\mathcal{G}_{\text{conf},T}^{(\delta)}$ or $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ fails is at most 2. For the worst-case upper bound, we apply Cauchy-Schwarz and Jensen inequalities to the upper bound in the second result of Lemma 13, in order to reduce to a telescoping sum:

$$\begin{aligned} \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} & \left[\mathbb{I}_{\mathcal{G}_{\text{conf},T}^{(\delta)} \cap \mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}} \sum_{l=1}^{L(T)} \sqrt{\min\{T_l, T\} - T_{l-1}} \right] \\ & \leq \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\mathbb{I}_{\mathcal{G}_{\text{conf},T}^{(\delta)} \cap \mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}} \sqrt{L(T) \sum_{l=1}^{L(T)} (\min\{T_l, T\} - T_{l-1})} \right] \end{aligned} \quad (31)$$

$$\begin{aligned} & = \sqrt{T} \mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\sqrt{\mathbb{I}_{\mathcal{G}_{\text{conf},T}^{(\delta)} \cap \mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}} L(T)} \right] \\ & \leq \sqrt{T} \sqrt{\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu} \left[\mathbb{I}_{\mathcal{G}_{\text{conf},T}^{(\delta)} \cap \mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}} L(T) \right]} \leq \sqrt{L_{\mathcal{E}_n^{(k,m)}} T}, \end{aligned} \quad (32)$$

where Step (31) follows by applying the Cauchy-Schwarz inequality to the vector $(\sqrt{\min\{T_l, T\} - T_{l-1}})_{l=1}^{L(T)}$, and Step (32) follows from Jensen's inequality ($\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$). As above, the statement follows by algebraic manipulation. $\square$

D.5 Adaptation to MNL Bandits

This subsection proves Theorem 3 by adapting the comparison-probability, concentration, and epoch-regret arguments to the fictitious no-choice item 0.

Theorem 3 (Upper Bounds to the Top- k Regret for MNL Bandits). *For every $n, k \in \mathbb{N}_{\geq 2}$, with $k \leq n$, instance $\nu_{\text{mnl}} \in \mathcal{E}_{\text{mnl},n}^{(k)}$ and learning horizon $T \in \mathbb{N}_{\geq n}$, by setting the algorithm parameters to $\beta_{\delta,n,T}(N) = \sqrt{\ln(2nT/\delta)/N}$ for $N \in \mathbb{N}_{\geq 1}$ (with $\beta_{\delta,n,T}(0) := +\infty$ and $\beta_{\delta,n,T}(+\infty) := 0$) and*

$\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top- k regret suffered by MNL-TOP- k -UIF:

$$\text{EReg}_{\text{MNL-TOP-}k\text{-UIF}, \nu_{\text{mnl}}}^{(k)}(T) \leq C_4 L_{\nu_{\text{mnl}}} \left(nk + \sum_{u \in \llbracket k+1, n \rrbracket} 1/\Delta_{k,u} \right) \ln(T),$$

where $C_4 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\nu_{\text{mnl}}}$ is defined analogously to L_ν in the statement of Theorem 2 (see Appendix D) and $\Delta_{k,u} := \theta_{r_\theta^{*-1}(k)} - \theta_{r_\theta^{*-1}(u)}$ for all $u \in \llbracket k+1, n \rrbracket$. If $k = n$, the regret is identically zero. If $k < n$, for a sufficiently large learning horizon satisfying $T \geq \frac{(nk)^2}{n-k} L_{\mathcal{E}_{\text{mnl},n}}^{(k)} \ln(2nT)$, the worst-case expected top- k regret is bounded by:

$$\sup_{\nu_{\text{mnl}} \in \mathcal{E}_{\text{mnl},n}^{(k)}} \text{EReg}_{\text{MNL-TOP-}k\text{-UIF}, \nu_{\text{mnl}}}^{(k)}(T) \leq C_5 \sqrt{L_{\mathcal{E}_{\text{mnl},n}}^{(k)} (n-k) T \ln(T)},$$

where $C_5 \in \mathbb{R}_{>0}$ is an absolute constant, $L_{\mathcal{E}_{\text{mnl},n}}^{(k)} := \sup_{\nu_{\text{mnl}} \in \mathcal{E}_{\text{mnl},n}^{(k)}} L_{\nu_{\text{mnl}}} \leq n+1$ (see Appendix D).

Proof of Theorem 3. The observed winner is distributed according to a PL model on $S_t \cup \{0\}$, which contains $k+1$ items. For distinct $i, p \in S_t \cup \{0\}$, winner feedback gives:

$$\begin{aligned} f_{i|S_t \cup \{0\}}^{(p)} &= \mathbb{P}_{R \sim \text{PL}((\theta_h)_{h \in S_t \cup \{0\}})} \left[R^{(1)}(i) \neq R^{(1)}(p) \right] \\ &= \frac{\theta_i + \theta_p}{\sum_{h \in S_t \cup \{0\}} \theta_h} \geq \frac{\theta_i + \theta_p}{k+1}, \\ \mu_{i|S_t \cup \{0\}}^{(p)} &= \mathbb{P}_{R \sim \text{PL}((\theta_h)_{h \in S_t \cup \{0\}})} \left[R^{(1)}(i) < R^{(1)}(p) \mid R^{(1)}(i) \neq R^{(1)}(p) \right] \\ &= \frac{\theta_i}{\theta_i + \theta_p}. \end{aligned}$$

Consequently, Lemma 5 continues to hold with m/k replaced by $1/(k+1)$; its two inequalities concerning $\mu_i^{(p)}$ are unchanged.

The concentration proofs also extend to pivots in $\llbracket 0, n \rrbracket$. There are at most $n(n+1)$ relevant ordered pairs (i, p) , with $i \in \llbracket n \rrbracket$ and $p \in \llbracket 0, n \rrbracket$. For a fixed pair and a fixed positive observation count, the tail probability in the proof of Lemma 6 is $\delta^2/(2n^2T^2)$. A union bound over the relevant pairs and at most T observation counts is therefore at most $\delta^2(n+1)/(2nT) \leq \delta$. Similarly, each tail in the proof of Lemma 7 is at most $\delta^2(n+1)/(4n) \leq \delta/2$. Hence the same confidence radius $\beta_{\delta,n,T}$ defines MNL analogues of $\mathcal{G}_{\text{conf},T}^{(\delta)}$ and $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$, each with failure probability at most δ .

Under the MNL confidence event, every relation $i \prec j$ returned by COUNT-BA implies $\theta_i > \theta_j$, also when the comparison pivot is 0. Thus the acceptance and rejection argument in Lemma 8 applies to the arms in $\llbracket n \rrbracket$. It remains to verify the pivot invariant. Before the first acceptance there is only one pivotal item, so the direct relation used by COUNT-BA is transitive. If an arm is accepted and NEW-PIVOT returns p' , the maximality of $\text{LCB}_{p'}^{(P_t)}$ implies that p' reaches every arm reached by the accepted arm, has no predecessor, and is itself accepted. If NEW-PIVOT returns $\perp$ and $P_t \in \llbracket n \rrbracket$, the same argument with $\text{LCB}_{P_t}^{(P_t)} = 1/2$ shows that P_t is accepted. If instead $P_t = 0$, the fallback in Algorithm 3 chooses an accepted arm in $\llbracket n \rrbracket$. If no arm is accepted, the maximal-LCB candidate returned by NEW-PIVOT has no predecessor and therefore belongs to $\mathcal{U}_{t+1}$. When the routine returns $\perp$, the same holds for P_t if $P_t \in \llbracket n \rrbracket$, whereas $P_{t+1} = 0$ if $P_t = 0$. Induction from $P_1 = 0$ consequently gives:

$$\mathcal{A}_t = \{\} \text{ implies } P_t = 0 \text{ or } P_t \in \mathcal{U}_t, \quad \mathcal{A}_t \neq \{\} \text{ implies } P_t \in \mathcal{A}_t \subseteq S_\theta^{(k)*}.$$

Moreover, if $\mathcal{A}_t \neq \{\}$ and $\text{LCB}_i^{(P_t)} > 1/2$, then the MNL confidence event gives $\theta_i > \theta_{P_t}$; since $P_t \in \mathcal{A}_t \subseteq S_\theta^{(k)*}$, the arm i accepted by this rule also belongs to $S_\theta^{(k)*}$.

Define:

$$c_t := k - |\mathcal{P}_t| + \mathbb{I}[0 \in \mathcal{P}_t], \quad \tilde{q}_t := \begin{cases} c_t/|\mathcal{U}'_t| & \text{if } \mathcal{U}'_t \neq \{\}, \\ 0 & \text{if } \mathcal{U}'_t = \{\}. \end{cases}$$

Every arm in $\mathcal{U}_t$ is pulled with conditional probability $\tilde{q}_t$. When $P_t \in \llbracket n \rrbracket$, this is exactly the exploration probability used in the proof of Theorem 2. When $P_t = 0$, the pivot invariant gives $\mathcal{A}_t = \{\}$, $c_t = k$, and S_t is a uniformly sampled k -subset of $\mathcal{U}_t$. In this case:

$$\text{IPReg}_{\text{MNL-Top-}k\text{-UIF}, \nu_{\text{mnl}}}^{(k)}(t) = \frac{1}{k|\mathcal{U}_t|} \sum_{i \in \mathcal{U}_t \cap S_{\theta}^{(k)*}} \sum_{j \in \mathcal{U}_t \setminus S_{\theta}^{(k)*}} (\theta_i - \theta_j).$$

Splitting the inner sum according to $\theta_j \geq \theta_0$ or $\theta_j < \theta_0$ yields the analogue of Lemma 11, since in the former case $\theta_i - \theta_j \leq \theta_i - \theta_0$.

In the initial epoch with $P_t = 0$, any first acceptance triggers either a pivot $p' \in \llbracket n \rrbracket$ returned by NEW-PIVOT or the fallback in Algorithm 3, which also sets $P_{t+1} \in \llbracket n \rrbracket$. That epoch therefore ends in the acceptance round, so its post-acceptance phase is empty. Every subsequent epoch satisfies $P_t \in \llbracket n \rrbracket$, and the low-pivot phase split in the proof of Lemma 13 applies unchanged.

The remainder of the proof of Lemma 12 applies with m/k replaced by $1/(k+1)$, with the predictable process defined from the probabilities $\tilde{q}_t$, and with burn-in:

$$\bar{t}_n^{(k, \delta, \lambda)} := \frac{64}{1 - e^{-\lambda}} n(k+1) \ln \left(\frac{2n}{\delta} \right).$$

Thus every occurrence of $\sqrt{k/m}$ in its conclusions is replaced by $\sqrt{k+1}$. Repeating the proof of Lemma 13 gives, for absolute constants $C_6, C_7 \in \mathbb{R}_{>0}$:

$$\begin{aligned} \text{PReg}_{\text{MNL-Top-}k\text{-UIF}, \nu_{\text{mnl}}}^{(k)}(T_{l-1} + 1, T_l) &\leq C_6 n(k+1) \ln \left(\frac{2nT}{\delta} \right) \\ &\quad + C_6 \frac{k+1}{k} \ln \left(\frac{2nT}{\delta} \right) \sum_{j \notin S_{\theta}^{(k)*}} \frac{1}{\theta_{r_{\theta}^{*-1}(k)} - \theta_j}, \\ \text{PReg}_{\text{MNL-Top-}k\text{-UIF}, \nu_{\text{mnl}}}^{(k)}(T_{l-1} + 1, T_l) &\leq C_7 n(k+1) \ln \left(\frac{2nT}{\delta} \right) \\ &\quad + C_7 \sqrt{\frac{k+1}{k}} (n-k) \sqrt{T_l - T_{l-1}} \sqrt{\ln \left(\frac{2nT}{\delta} \right)}. \end{aligned}$$

The predictable-count sum used in Equation (28) satisfies:

$$\sum_{j \notin S_{\theta}^{(k)*}} \tilde{Q}_{T_{l-1}+1: \min\{T_l, T_{\mathcal{U}, j}\}} = \sum_{t=T_{l-1}+1}^{T_l} \tilde{q}_t |\mathcal{U}_t \setminus S_{\theta}^{(k)*}| \leq k(T_l - T_{l-1}),$$

because $\tilde{q}_t |\mathcal{U}_t \setminus S_{\theta}^{(k)*}| \leq c_t \leq k$. This inequality and Cauchy–Schwarz give the factor $\sqrt{(k+1)(n-k)(T_l - T_{l-1})/k}$ in the second displayed bound.

For each T , let $\mathcal{G}_{\text{conf}, T}^{\text{mnl}}$ denote the MNL analogue of the estimator-confidence event $\mathcal{G}_{\text{conf}, T}^{(\delta)}$ for the run with $\delta = 1/T$, and define:

$$L_{\nu_{\text{mnl}}} := \sup_{T \in \mathbb{N}_{\geq 1}} \mathbb{E}_{\text{MNL-Top-}k\text{-UIF}, \nu_{\text{mnl}}} \left[\mathbb{I}_{\mathcal{G}_{\text{conf}, T}^{\text{mnl}}} L(T) \right], \quad L_{\mathcal{E}_{\text{mnl}, n}^{(k)}} := \sup_{\nu_{\text{mnl}} \in \mathcal{E}_{\text{mnl}, n}^{(k)}} L_{\nu_{\text{mnl}}}.$$

Whenever NEW-PIVOT changes the pivot, the confidence event implies that its PL parameter strictly increases. Item 0 occurs only in the initial epoch and is never selected again. The fallback from item 0 can occur only at the first acceptance, after which the pivot is fixed. Hence every $i \in \llbracket n \rrbracket$ occurs as a pivot at most once and:

$$L_{\nu_{\text{mnl}}} \leq n+1, \quad L_{\mathcal{E}_{\text{mnl}, n}^{(k)}} \leq n+1.$$

Finally, $(k+1)/k \leq 3/2$ and $n(k+1) \leq 3nk/2$ because $k \geq 2$. Setting $\delta = 1/T$, summing the two displayed estimates over the epochs, applying Cauchy–Schwarz to $(\sqrt{T_l - T_{l-1}})_{l=1}^{L(T)}$, and adding at most 2 for the failure of the two confidence events gives the instance-dependent bound and a worst-case bound containing:

$$C_6 L_{\mathcal{E}_{\text{mnl}, n}^{(k)}} nk \ln(2nT) + C_7 \sqrt{L_{\mathcal{E}_{\text{mnl}, n}^{(k)}}} (n-k) T \ln(2nT).$$

Under the condition in Theorem 3:

$$L_{\mathcal{E}_{\text{mnl},n}^{(k)}} nk \ln(2nT) \leq \sqrt{L_{\mathcal{E}_{\text{mnl},n}^{(k)}} (n-k)T \ln(2nT)},$$

so the first term is absorbed into the second after changing the absolute constant. If $k = n$, every played assortment contains all arms in $\llbracket n \rrbracket$ and the regret is identically zero. $\square$

E Auxiliary Results

In this appendix, we report the auxiliary results required for the proofs presented throughout this manuscript.

E.1 Information Theory

Lemma 14 (Pinsker's Inequality for Bounded Variables). *Let μ and ϕ be two probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$, and let $X : \mathcal{X} \rightarrow [0, 1]$ be a random variable. Then, it holds that:*

$$\mathbb{E}_\mu[X] - \mathbb{E}_\phi[X] \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\phi \| \mu)}.$$

Proof. Let $\|\mu - \phi\|_{\text{TV}} := \sup_{A \in \mathcal{F}} |\mu(A) - \phi(A)|$ denote the total variation distance between the measures μ and ϕ . By the variational definition of the total variation distance for a random variable X bounded in $[0, 1]$, and the symmetry of the total variation distance, we have:

$$\mathbb{E}_\mu[X] - \mathbb{E}_\phi[X] \leq \sup_{X \in [0,1]^{\mathcal{X}}} (\mathbb{E}_\mu[X] - \mathbb{E}_\phi[X]) = \|\mu - \phi\|_{\text{TV}} = \|\phi - \mu\|_{\text{TV}}.$$

Applying Pinsker's inequality (see e.g. Lattimore and Szepesvári, 2020, Chapter 14), which bounds the total variation distance by the Kullback-Leibler divergence as $\|\phi - \mu\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\phi \| \mu)}$, yields the stated result. $\square$

E.2 Concentration Inequalities

Theorem 15 (Azuma's Inequality). *Let $(X_t)_{t=0}^{+\infty}$ be a martingale w.r.t. the filtration $(\mathcal{F}_t)_{t=0}^{+\infty}$. Assume there are predictable processes $(A_t)_{t=1}^{+\infty}$ and $(B_t)_{t=1}^{+\infty}$ w.r.t. $(\mathcal{F}_t)_{t=0}^{+\infty}$, and constants $(c_t)_{t=1}^{+\infty} \subseteq (0, +\infty)$ s.t.*

$$A_t \leq X_t - X_{t-1} \leq B_t \text{ and } B_t - A_t \leq c_t \text{ a.s. for every } t \in \mathbb{N}_{\geq 1}.$$

Then, for every $n \in \mathbb{N}_{\geq 1}$ and $\epsilon > 0$:

$$\mathbb{P}[|X_n - X_0| \geq \epsilon] \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^n c_t^2}\right).$$

E.3 General Probability

Lemma 16. *Let $\mathcal{U}$ be a finite set, $c \in \llbracket |\mathcal{U}| \rrbracket$. Then, for every $i \in \mathcal{U}$:*

$$\mathbb{P}_{S \sim \text{Unif}(\{S' \subseteq \mathcal{U} \text{ s.t. } |S'|=c\})} [i \in S] = \frac{c}{|\mathcal{U}|}.$$

Proof. Observe that:

$$|\{S' \subseteq \mathcal{U} \text{ s.t. } i \in S', |S'| = c\}| = |\{S'' \cup \{i\} \text{ s.t. } S'' \subseteq \mathcal{U} \setminus \{i\}, |S''| = c-1\}| = \binom{|\mathcal{U}|-1}{c-1}.$$

Then:

$$\mathbb{P}_{S \sim \text{Unif}(\{S' \subseteq \mathcal{U} \text{ s.t. } |S'|=c\})} [i \in S] = \frac{\binom{|\mathcal{U}|-1}{c-1}}{\binom{|\mathcal{U}|}{c}} = \frac{c}{|\mathcal{U}|}.$$

$\square$