Worst-Case Regret Bounds for Combinatorial Bandits with Ranking Feedback

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 among n that are available. 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 which is sublinear in 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 $\tilde{\Theta}\left(\sqrt{\frac{n}{m}T}\right)$, which *matches* the *lower bound* up to logarithmic factors, 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". Our novel approach allows us to avoid the exploding burn-in term, inversely proportional to the no-choice probability, present in the prominent algorithmic strategies.

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

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Existing and new bounds for CMABs-R and MNL bandits. The "Feedback" 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 θ?
	Full-Ranking	Winner w/ No Choice	Top-m	Top-k	Weighted			
Saha and Gopalan (2019b)	✓	✗	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{1}{k}\sum_{v\in[\![k+1,n]\!]}\frac{1}{\Delta_{k,v}}\log T + \Delta'_{\max}\sum_{u\in[\![k]\!]}\sum_{v\in[\![k+1,n]\!]}\frac{\theta_u^2}{\Delta_{u,v}^2}\right)$	N/A	✗
Agrawal et al. (2019)	✗	✓	✗	✗	✓	N/A	$\tilde{\mathcal{O}}\left(\sqrt{\theta_0^{-1}nT}\right)$	✗
Saha and Gaillard (2025)	✗	✓	✗	✓	✓	N/A	$\tilde{\mathcal{O}}\left(\sqrt{\min\{\theta_0^{-1},n\}nT} + n^2k^2\theta_0^{-2}\right)$	✗
Ours	✓	✓	✓	✓	✗	$\mathcal{O}\left(\frac{1+\log(k/m)}{m}\sum_{u\in[\![k+1,n]\!]}\frac{1}{\Delta_{k,u}}\log T + \frac{kn^2}{m}\log T\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{n}{m}T} + \frac{kn^2}{m}\right)$	✓

of the ranking only (i.e., *top-m feedback*, Saha and Gopalan, 2019a). The goal is to identify the set of k best options. This setting takes the name of *Combinatorial Multi-Armed Bandits* with *Ranking feedback* (CMABs-R).[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 none 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 notion 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 $\theta_1 > \cdots > \theta_n$, 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

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

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

7.2, Lattimore and Szepesvári, 2020), it is possible to translate this term into a sublinear worst-case bound.[3] The issue with this result arises from the additional burn-in term: $\Delta'_{\max} \sum_{u \in [\![k]\!]} \sum_{v \in [\![k+1,n]\!]} \frac{\theta_u^2}{\Delta_{u,v}^2}$. 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, $u \in [\![k]\!]$, $v \in [\![k+1, n]\!]$, 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 $\Delta_{k,k+1} \to 0$, 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 $\theta_0 \to 0$. The connections with this setting are discussed in greater detail in Section 5. In both cases, the discussed approaches do not allow us 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 $\boldsymbol{\theta}$ 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 $\Omega\left(\sqrt{\frac{n}{m}T}\right)$ (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 $\tilde{\mathcal{O}}\left(\sqrt{\frac{n}{m}T}\right)$, matching the lower bound in all the dependencies up to logarithmic factors and proving for the first time that it is possible to achieve sublinear regret under *any* choice of the PL parameters. This provides 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 8) and proof strategies for the regret upper bound (see Lemma 17) 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, this approach allows us to achieve a sublinear regret bound of order $\tilde{\mathcal{O}}(\sqrt{nT})$ (Theorem 3), which holds for any value of the no-choice parameter θ_0 and matches the known worst-case lower bound (Chen and Wang, 2018, Theorem 1) up to logarithmic factors.

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

$$\theta_i > \theta_j \text{ for every } i \in S_{\boldsymbol{\theta}}^{(k)*}, j \in [\![n]\!] \setminus S_{\boldsymbol{\theta}}^{(k)*}, \tag{1}$$

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

[4] For 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.

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

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

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 $\mathrm{PL}((\theta_i)_{i \in S}) \in \Delta(\mathrm{Bij}(S, [\![|S|]\!]))$ such that:[5]

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

Thus, $R_t \sim \mathrm{PL}((\theta_i)_{i \in S_t})$ 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 $\max_{i \in [\![n]\!]} \theta_i = 1$. We denote by $\mathcal{E}_n^{(k,m)}$ the set of CMABs-R for fixed $n, k \in \mathbb{N}_{\geq 2}, m \in \mathbb{N}_{\geq 1}$ with $m \leq k \leq n$ and PL parameters $\boldsymbol{\theta}$ satisfying Equation (1) and $\max_{i \in [\![n]\!]} \theta_i = 1$. Finally, we use $r_{\boldsymbol{\theta}}^* \in \mathrm{Bij}([\![n]\!])$ for a bijection which orders the PL parameters of the corresponding instance $\nu = (n, \boldsymbol{\theta}, k, m) \in \mathcal{E}_n^{(k,m)}$: $1 = \theta_{r_{\boldsymbol{\theta}}^{*-1}(1)} \geq \cdots \geq \theta_{r_{\boldsymbol{\theta}}^{*-1}(n)} > 0$.

Learning Problem. The goal of the agent is to learn the set of top-k arms $S_{\boldsymbol{\theta}}^{(k)*}$. Observe that $S_{\boldsymbol{\theta}}^{(k)*} = \arg\max_{\substack{S \subseteq [\![n]\!] \\ |S|=k}} g_{\boldsymbol{\theta}}^{(k)}(S)$. 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 $H_{t-1} \coloneqq ((U_s, S_s, R_s^{(m)})_{s=1}^{t-1}, U_t)$. Here, U_t serves as a source of independent internal randomness for the policy at round t (e.g., $U_t \sim \mathrm{Unif}([0,1])$).[6] A *randomized policy* π assigns a conditional probability distribution to each possible history to determine the set of arms to pull: $S_t \sim \pi(\cdot \mid H_{t-1})$, where $S_t \in \{S \subseteq [\![n]\!] \text{ s.t. } |S| = k\}$. 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):

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

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, 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 $m \leq k \leq n$, randomized policy π and learning horizon $T \in \mathbb{N}_{\geq n}$, it holds:*

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \mathbb{E}\mathsf{Reg}_{\pi,\nu}^{(k)}(T) \geq C_1 \sqrt{\frac{n-k}{m} \min\left\{1, \frac{n-k}{k}\right\} T},$$

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

[5] We denote by $\mathrm{Bij}(A, B)$ the set of bijective functions which map set A onto set B. We write $\mathrm{Bij}(A)$ 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.

[6] We denote by $\mathrm{Unif}(\mathcal{X}) \in \Delta(\mathcal{X})$ 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 is a decomposition of the Kullback–Leibler divergence of the histories induced by two PL instances ν, ν' with parameters in $[1/2, 1]$, for the same randomized policy π, stopped at a stopping time $\tau \leq T$. We provide this in Lemma 4. By exploiting the properties of the PL distribution, we obtain a decomposition analogous to the one of (Lattimore and Szepesvári, 2020, Lemma 15.1) for stochastic MABs, but valid under top-m feedback in a combinatorial setting. The proof then follows the classical *parallel-games* argument for CMABs (Audibert et al., 2014): we construct k disjoint batches of roughly n/k arms, each containing one optimal arm. Differently from the original construction, where the sets of arms that the learner can play contain exactly one arm per batch, our learner may play several arms of the same batch in a round. We therefore control the number of pulls received by each batch with a separate stopping time, described in Appendix B. For a fixed batch we defined a base instance where all the arms in that batch have the same parameter and a hard instance where one arm in the batch is promoted to optimal. Finally, we apply Pinsker's inequality to the time-stopped number of pulls of the promoted arm, comparing the hard instance against the base instance, and average over all choices of the fixed batch and the promoted arm. □

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 $[\![n]\!]$ of cardinality k is $[\![n]\!]$ 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-R 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:

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

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):

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

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

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

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

Input: $T \in \mathbb{N}_{\geq 1},\ n \in \mathbb{N}_{\geq 2},\ k \in [\![2, n]\!],\ m \in [\![k]\!],\ \beta_{\delta,n,T} : [0, +\infty] \to [0, +\infty],\ \delta \in (0, 1]$
Initialize $P_1 \sim \mathrm{Unif}([\![n]\!])$, $\mathcal{A}_1 \leftarrow \{\}, \mathcal{U}_1 \leftarrow [\![n]\!]$,
  $N_i^{(p)} \leftarrow 0$ if $i \neq p$, $+\infty$ otherwise,
  $\hat{\mu}_i^{(p)} \leftarrow 1/2$ for every $i, p \in [\![n]\!]$
for $t \in [\![T]\!]$ **do**
  $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_t, P_{t+1} \leftarrow P_t$
  Let $\mathcal{P}_t \leftarrow \mathcal{A}_t \cup \{P_t\}$
  Sample $S_{\mathrm{exp},t} \sim \mathrm{Unif}(\{S \subseteq \mathcal{U}_t \setminus \{P_t\} \text{ s.t. } |S| = k - |\mathcal{P}_t|\})$
  Pull $S_t \leftarrow \mathcal{P}_t \cup S_{\mathrm{exp},t}$
  Observe $R_t^{(m)} : S_t \to [\![\min\{k, m+1\}]\!]$
  // 1. Update Statistics & Confidence Bounds
  $(\hat{\mu}_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t} \leftarrow$ RB-UPDATE$(\mathcal{P}_t, S_t, (\hat{\mu}_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t, p \in \mathcal{P}_t}, R_t^{(m)})$
  for $i \in \mathcal{U}_t, p \in \mathcal{P}_t$ **do**
    $\mathrm{UCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} + \beta_{\delta,n,T}(N_i^{(p)})$
    $\mathrm{LCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} - \beta_{\delta,n,T}(N_i^{(p)})$
  end for
  // 2. Update Unknown Arms
  $(A_i)_{i \in \mathcal{U}_t}, (B_i)_{i \in \mathcal{U}_t} \leftarrow$ COUNT-BA$(\mathcal{P}_t, \mathcal{U}_t, (\mathrm{UCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t}, (\mathrm{LCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t})$
  for $i \in \mathcal{U}_t$ **do**
    if $B_i \geq k - |\mathcal{A}_t|$ **then** ▷ Reject arm
      $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
    else if $A_i \geq |\mathcal{A}_t \cup \mathcal{U}_t| - k$ **then** ▷ Accept arm
      $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
    end if
  end for
  // 3. Pivot Update
  if $\mathcal{A}_t \neq \{\}$ **then** ▷ Stop pivot update
    for $i \in \mathcal{U}_t$ **do**
      if $\mathrm{LCB}_i^{(P_t)} > 1/2$ **then** ▷ Accept $\theta_i > \theta_{P_t}$
        $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
      end if
    end for
  else
    $p' \leftarrow$ NEW-PIVOT$(P_t, \mathcal{U}_t, (\mathrm{LCB}_i^{(P_t)})_{i \in \mathcal{U}_t})$
    if $p' \neq \perp$ **then**
      $P_{t+1} \leftarrow p'$
    end if
  end if
  if $|\mathcal{A}_{t+1}| = k$ **then** $\mathcal{U}_{t+1} \leftarrow \{\}$
end for

1. *Identification of suboptimal arms $j \notin S_{\boldsymbol{\theta}}^{(k)*}$.* 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 $\mathcal{U}_t \cup \mathcal{A}_t \subseteq [\![n]\!]$. $\mathcal{U}_t$ is the set of *unknown arms*, i.e., the set of arms for which we 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 9. 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 $\beta_{\delta,n,T}(0) \coloneqq +\infty$ and $\beta_{\delta,n,T}(+\infty) \coloneqq 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top-k regret suffered by* TOP-k-UIF*:*

$$\mathbb{E}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T) \leq \frac{C_2}{m}\left(kn(n-k) + (1+\ln(k/m)) \sum_{u \in [\![k+1,n]\!]} 1/\Delta_{k,u}\right)\ln(T),$$

where $C_2 \in \mathbb{R}_{>0}$ is an absolute constant and $\Delta_{k,u} \coloneqq \theta_{r^{*-1}_{\boldsymbol{\theta}}(k)} - \theta_{r^{*-1}_{\boldsymbol{\theta}}(u)}$ for all $u \in [\![k+1, n]\!]$.

Furthermore, the worst-case expected top-k regret is bounded by:

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \mathbb{E}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T) \leq C_3 \left(\frac{kn(n-k)}{m} \ln(nT) + \sqrt{\frac{n-k}{m} \min\left\{1, \frac{n-k}{k}\right\} T \ln(nT)(1+\ln(k/m))} \right),$$

where $C_3 \in \mathbb{R}_{>0}$ is an absolute constant. If $T \geq \frac{k^2 n^2 \max\{k,n-k\}}{m} \frac{\ln(nT)}{1+\ln(k/m)}$, the additive term is absorbed into the square-root term after changing the absolute constant.

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

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

since $\Delta_{k,k+1} \leq \Delta_{k,u}$ for all $u \in [\![k+1, n]\!]$, highlighting that the complexity of the instance depends on the hardness of differentiating the k-th arm from the $(k+1)$-th. Furthermore, as expected, the regret scales inversely with the informativeness of the feedback $m \in [\![k]\!]$.

In the worst case, for the sufficiently large horizons stated in Theorem 2, we obtain the upper bound $\tilde{\mathcal{O}}\left(\sqrt{\frac{n-k}{m} \min\left\{1, \frac{n-k}{k}\right\} T}\right)$, which matches the lower bound (Theorem 1) in every dependence up to logarithmic factors, for all $m \leq k \leq n$, 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-R 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. An MNL bandit with top-k objective is similar to a CMAB-R with top-1 feedback ($m = 1$). Indeed, the learner provides a subset $S_t \subseteq [\![n]\!]$ 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 $\nu_{\text{mnl}} = (n, \boldsymbol{\theta}, \theta_0, k)$ an MNL bandit with top-k objective, where $(n, \boldsymbol{\theta}, k, 1) \in \mathcal{E}_n^{(k,m)}$ and $\theta_0 \in (0, 1]$. Let $\mathcal{E}^{(k)}_{\text{mnl},n}$ 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-R setting (see Section 2). Saha and Gaillard (2025) summarize the existing results in the literature, showing how the regret bounds of the prominent approaches either rely on the assumption $\theta_0 = \max_{i \in [\![n]\!]} \theta_i = 1$ or scale as $\tilde{\mathcal{O}}\left(\sqrt{\theta_0^{-1} nT}\right)$, becoming vacuous when $\theta_0 \to 0$. They provide an upper bound $\tilde{\mathcal{O}}\left(\sqrt{\min\{\theta_0^{-1}, n\}nT} + n^2 k^2 \theta_0^{-2}\right)$, 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: $\hat{\theta}^{\text{ucb}}_{i,t} := \min_{j \in \{0\} \cup [\![n]\!]} \gamma^{\text{ucb}}_{i,j,t} \gamma^{\text{ucb}}_{j,0,t}$, where $\gamma^{\text{ucb}}_{i,j,t}$ is an optimistic estimate of $\gamma_{i,j} := \theta_i / \theta_j$. Letting $\theta_0 \to 0$ implies $\gamma_{i,0} \to +\infty$ for all $i \in [\![n]\!]$, 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 $j \in \{0\} \cup [\![n]\!]$ which provides the tightest upper confidence bound. Unfortunately, the estimate still relies on the comparisons against the no-choice item, and the scarcity

[7] The 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 $\tau_{ij,t} > C m^2 \theta^2_{\max} x$, which, following the notation used in this work, is translated to: $\tau_{ij,t} > C k^2 \theta_0^{-2} x$ where $x = 2 \ln T$.

of comparisons issue described in Section 4 still arises when $\theta_0 \approx 0$. Solving the problem requires a shift of paradigm. We address it through the pivoting approach in TOP-k-UIF, which 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 $\theta_0 \to 0$. 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 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 $\nu_{mnl} \in \mathcal{E}^{(k)}_{mnl,n}$ 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 $\beta_{\delta,n,T}(0) \coloneqq +\infty$ and $\beta_{\delta,n,T}(+\infty) \coloneqq 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top-k regret suffered by* MNL-TOP-k-UIF*:*

$$\mathbb{E}\mathsf{Reg}^{(k)}_{\text{MNL-TOP-}k\text{-UIF},\nu_{mnl}}(T) \leq C_4 \Big(kn(n-k) + \ln(k) \sum_{u \in [\![k+1,n]\!]} 1/\Delta_{k,u} \Big) \ln(T),$$

where $C_4 \in \mathbb{R}_{>0}$ is an absolute constant and $\Delta_{k,u} \coloneqq \theta_{r^{*-1}_{\boldsymbol{\theta}}(k)} - \theta_{r^{*-1}_{\boldsymbol{\theta}}(u)}$ for all $u \in [\![k+1, n]\!]$. Furthermore, the worst-case expected top-k regret is bounded by:

$$\sup_{\nu_{mnl} \in \mathcal{E}^{(k)}_{mnl,n}} \mathbb{E}\mathsf{Reg}^{(k)}_{\text{MNL-TOP-}k\text{-UIF},\nu_{mnl}}(T)$$
$$\leq C_5 \left(kn(n-k)\ln(nT) + \sqrt{(n-k)\min\left\{1, \frac{n-k}{k}\right\} T \ln(nT)\ln(k)} \right),$$

where $C_5 \in \mathbb{R}_{>0}$ is an absolute constant. If $T \geq k^2 n^2 \max\{k, n-k\} \frac{\ln(nT)}{\ln(k)}$, the additive term is absorbed into the square-root term after changing the absolute constant.

The upper bound proven in Theorem 3 matches the known worst-case lower bound (Chen and Wang, 2018, Theorem 1) for the MNL setting up to logarithmic terms for every value of the no-choice parameter θ_0. Indeed, the bound is sublinear in T (i.e., $\tilde{\mathcal{O}}(\sqrt{T})$) even for $\theta_0 \to 0$, proving that the dependence on θ_0^{-1} in the bounds presented before is not due to the intrinsic statistical complexity of the setting, but is rather a consequence 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 pointwise 95% confidence intervals for the mean over 20 independent runs.[9]

TOP-k-UIF vs RECMAXMINUCB (Saha and Gopalan, 2019b). We validate TOP-k-UIF against RECMAXMINUCB (Saha and Gopalan, 2019b) in a synthetic instance of a CMAB-R. To ensure the fairness of the comparison with the baseline, we set $k = m$. Let $n = 400$, $k = m = 2$, $T = 10^6$ and $\theta_1 = \theta_2 = 1$, $\theta_3 = \cdots = \theta_{12} = 0.9$, $\theta_i = 0.05$ for $i > 12$. We run $N = 20$ trials and average the cumulative regrets suffered by the two algorithms, which are reported in Figure 1. TOP-k-UIF markedly outperforms RECMAXMINUCB on this instance: its final mean cumulative regret is approximately 1.60×10^5, compared with 3.00×10^5 for the baseline, a reduction of about 47%.

Sensitivity of TOP-k-UIF to m. We fix a synthetic instance of a CMAB-R and only vary $m \in \{1, 2, 3, 6\}$. We test the subsequent performance variation of TOP-k-UIF. In particular, let $n = 16$, $k = 6$ and $T = 10^6$. The parameters $\boldsymbol{\theta}$ are in Appendix F. We run $N = 20$ trials for every value of

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

[9] `https://github.com/gianmarcogenalti/combinatorial-bandits-relative-feedback` provides the code to reproduce the simulations. Additional details are in Appendix F.

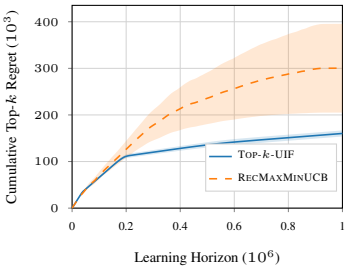

Figure 1: Top-k-UIF vs Saha and Gopalan (2019b) (20 trials, 95% CI)

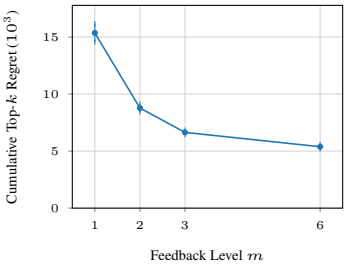

Figure 2: Top-k-UIF with varying m (20 trials, 95% CI)

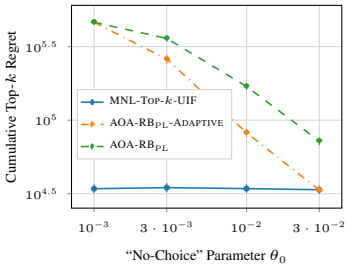

Figure 3: MNL-Top-k-UIF vs Saha and Gaillard (2025) (20 trials, 95% CI)

m. In Figure 2, we report the average cumulative regrets at the end of the learning horizon, over the N trials. The average regret decreases as m increases, with full-ranking feedback ($m = 6$) yielding the smallest regret.

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 $T = 10^6$. The parameters $\boldsymbol{\theta}$ are in Appendix F. We vary $\theta_0 \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}\}$, and we run $N = 20$ trials for every value of θ_0. In Figure 3, we report the average cumulative regrets over the N trials. At $\theta_0 \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$, MNL-Top-$k$-UIF incurs less mean regret than both baselines. Its regret remains approximately constant as θ_0 decreases, whereas both baselines incur sharply larger regret. In particular, this simulation shows that such previous approaches rely critically on the observability of the no-choice and that the dependence on θ_0 in their regret bounds is not merely an artifact of the proofs.

7 Conclusions

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 $\Omega\left(\sqrt{\frac{n-k}{m}\min\left\{1, \frac{n-k}{k}\right\}T}\right)$ for every $m \leq k$. In Section 4, we introduced the Top-k-UIF algorithm (Algorithm 1) for the setting under study, and in Theorem 2, we presented a corresponding regret upper bound of order $\tilde{\mathcal{O}}\left(\frac{kn(n-k)}{m} + \sqrt{\frac{n-k}{m}\min\left\{1, \frac{n-k}{k}\right\}T}\right)$, which matches the lower bound in all parameter dependencies up to logarithmic factors for sufficiently large learning horizons.

Furthermore, in Section 5, we extended our approach to MNL bandits under the top-k objective through the MNL-Top-k-UIF algorithm (Algorithm 3). In Theorem 3, we proved a regret upper bound for MNL-Top-k-UIF of order $\tilde{\mathcal{O}}(\sqrt{nT})$, avoiding both the deteriorating regret rate and the exploding burn-in term in previous bounds as $\theta_0 \to 0$, and matching the known regret lower bound (Chen and Wang, 2018, Theorem 1) up to logarithmic factors.

Acknowledgments and Disclosure of Funding

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.

Xi Chen and Yining Wang. A note on a tight lower bound for capacitated MNL-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537, 2018.

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.

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.

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 Akrour, 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. In this setting, a Plackett-Luce model 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$_{\text{PL}}$-ADAPTIVE algorithm, which improves the dependence on the no-choice probability by moving it into a burn-in term which scales only logarithmically 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 bound the divergence between the distributions of the history up to a stopping time when a fixed randomized policy π interacts with two instances whose PL parameters lie in $[1/2, 1]$.

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

$$D_{\mathrm{KL}}(\mu \| \phi) \coloneqq \int_{\mathcal{X}} \ln\left(\frac{\mathrm{d}\mu}{\mathrm{d}\phi}\right) \mathrm{d}\mu.$$

In our setting, let $\mathcal{H}_t \coloneqq \sigma(H_t)$ be the σ-algebra induced by the random variable H_t for every $t \in [\![0, T]\!]$, as defined in Section 2. A *stopping time* τ takes values in $[\![0, T]\!]$ and satisfies $\{\tau \le t\} \in \mathcal{H}_t$ for every $t \in [\![0, T]\!]$: whether to stop by round t can be decided from H_t. We use the same stopping rule under both instances. The corresponding *stopped σ-algebra* is:

$$\mathcal{H}_\tau \coloneqq \{E \in \mathcal{H}_T \text{ s.t. } E \cap \{\tau \le t\} \in \mathcal{H}_t \text{ for every } t \in [\![0, T]\!]\}.$$

We denote the restriction of $\mathbb{P}_{\pi,\nu}$ to $\mathcal{H}_\tau$ by:

$$\mathbb{P}_{\pi,\nu}^{(\le\tau)} \coloneqq \mathbb{P}_{\pi,\nu}|_{\mathcal{H}_\tau},$$

and define the number of pulls of arm i through round t as $N_i(t) \coloneqq \sum_{s=1}^{t} \mathbb{I}[i \in S_s]$ for every $i \in [\![n]\!]$ and $t \in [\![0, T]\!]$.

With this notation, the following lemma bounds the divergence between stopped histories by expected arm counts, allowing us to compare instances under an adaptive stopping rule.

Lemma 4 (Divergence Decomposition at a Stopping Time)**.** *Let $n, k \in \mathbb{N}_{\ge 2}$ and $m \in \mathbb{N}_{\ge 1}$ with $m \le k \le n$. Consider two instances $\nu = (n, \boldsymbol{\theta}, k, m), \nu' = (n, \boldsymbol{\theta}', k, m) \in \mathcal{E}_n^{(k,m)}$ such that $\theta_i, \theta_i' \in [1/2, 1]$ for every $i \in [\![n]\!]$.*[10] *For every randomized policy π, learning horizon $T \in \mathbb{N}_{\ge 1}$, and stopping time $\tau \in [\![0, T]\!]$ with respect to $(\mathcal{H}_t)_{t=0}^T$, specified by the same rule under both instances, it holds:*

$$D_{\mathrm{KL}}\left(\mathbb{P}_{\pi,\nu}^{(\le\tau)} \,\middle\|\, \mathbb{P}_{\pi,\nu'}^{(\le\tau)}\right) \le 32\frac{m}{k} \sum_{i \in [\![n]\!]} (\theta_i - \theta_i')^2 \mathbb{E}_{\pi,\nu}[N_i(\tau)].$$

Proof. For every finite non-empty set S and $u \in [\![|S|]\!]$, we denote by $\mathrm{PL}^{(u)}((\theta_i)_{i\in S})$ the distribution of the top-u ranking $R^{(u)}$ when $R \sim \mathrm{PL}((\theta_i)_{i\in S})$. To apply the chain rule up to τ, let $\dagger$ be a fixed symbol outside the history space, and define:

$$\widetilde{H}_t^{(\le\tau)} \coloneqq \begin{cases} H_t & \text{if } t \le \tau \\ \dagger & \text{otherwise} \end{cases} \quad \text{for every } t \in [\![0, T]\!], \qquad \widetilde{H}^{(\le\tau)} \coloneqq (\widetilde{H}_t^{(\le\tau)})_{t=0}^T.$$

Each $\widetilde{H}_t^{(\le\tau)}$ is $\mathcal{H}_\tau$-measurable. Conversely, $\widetilde{H}^{(\le\tau)}$ determines τ and H_τ, and every $E \in \mathcal{H}_\tau$ is the union of the events $E \cap \{\tau = t\}$, each determined by H_t. Hence:

$$\sigma(\widetilde{H}_0^{(\le\tau)}, \ldots, \widetilde{H}_T^{(\le\tau)}) = \mathcal{H}_\tau.$$

Let $\mu_{\pi,\nu}^{(\le\tau)}$ and $\mu_{\pi,\nu'}^{(\le\tau)}$ be the distributions of $\widetilde{H}^{(\le\tau)}$ under $\mathbb{P}_{\pi,\nu}$ and $\mathbb{P}_{\pi,\nu'}$, respectively. All feedback probabilities are positive under both instances, so $\mu_{\pi,\nu}^{(\le\tau)} \ll \mu_{\pi,\nu'}^{(\le\tau)}$. Since $\mathcal{H}_\tau = \sigma(\widetilde{H}^{(\le\tau)})$, the

[10] The inequality also holds for parameter vectors $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [1/2, 1]^n$ without a strictly separated top-k set as required by Equation (1). For convenience, we state the lemma for instances in $\mathcal{E}_n^{(k,m)}$; the PL laws remain well defined without this separation, and we use such parameter tuples as auxiliary reference measures.

density of the first stopped law with respect to the second depends only on the stopped history $\widetilde{H}^{(\leq\tau)}$. It is obtained by evaluating the history density $\frac{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu}}{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu'}}$ at $\widetilde{H}^{(\leq\tau)}$. The change-of-variables formula therefore gives:

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mathbb{P}^{(\leq\tau)}_{\pi,\nu}\,\middle\|\,\mathbb{P}^{(\leq\tau)}_{\pi,\nu'}\right)\\
&\quad=\mathbb{E}_{\pi,\nu}\left[\ln\frac{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu}}{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu'}}(\widetilde{H}^{(\leq\tau)})\right]=\int\ln\left(\frac{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu}}{\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu'}}\right)\mathrm{d}\mu^{(\leq\tau)}_{\pi,\nu}\\
&\quad=D_{\mathrm{KL}}\left(\mu^{(\leq\tau)}_{\pi,\nu}\,\middle\|\,\mu^{(\leq\tau)}_{\pi,\nu'}\right).
\end{aligned}$$

We can thus apply the usual chain rule to $\widetilde{H}^{(\leq\tau)}$. Since $\{\tau\geq t\}\in\mathcal{H}_{t-1}$, only rounds through τ contribute a feedback divergence. The policy and its internal randomness contribute zero divergence because their conditional distributions are the same under both instances. Consequently:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}^{(\leq\tau)}_{\pi,\nu}\,\middle\|\,\mathbb{P}^{(\leq\tau)}_{\pi,\nu'}\right)&=D_{\mathrm{KL}}\left(\mu^{(\leq\tau)}_{\pi,\nu}\,\middle\|\,\mu^{(\leq\tau)}_{\pi,\nu'}\right)\\
&=\sum_{t=1}^{T}\mathbb{E}_{\pi,\nu}\left[\mathbb{I}[\tau\geq t]D_{\mathrm{KL}}\left(\mathrm{PL}^{(m)}((\theta_i)_{i\in S_t})\,\middle\|\,\mathrm{PL}^{(m)}((\theta'_i)_{i\in S_t})\right)\right]\\
&=\mathbb{E}_{\pi,\nu}\left[\sum_{t=1}^{\tau}D_{\mathrm{KL}}\left(\mathrm{PL}^{(m)}((\theta_i)_{i\in S_t})\,\middle\|\,\mathrm{PL}^{(m)}((\theta'_i)_{i\in S_t})\right)\right].
\end{aligned}\tag{2}$$

We now bound the KL divergence between the feedback distributions for a fixed set S. For every non-empty $A\subseteq S$, write $\Theta_A\coloneqq\sum_{i\in A}\theta_i$ and $\Theta'_A\coloneqq\sum_{i\in A}\theta'_i$. Suppose first that $m\leq k/2$. At position $u\in[\![m]\!]$, condition on the first $u-1$ ranked arms, and let A be the set of remaining arms. By the PL model, the next ranked arm is $i\in A$ with probability θ_i/Θ_A under ν and θ'_i/Θ'_A under ν'. Thus, its conditional ranking laws are $\mathrm{PL}^{(1)}((\theta_i)_{i\in A})$ and $\mathrm{PL}^{(1)}((\theta'_i)_{i\in A})$. Using $\ln x\leq x-1$ and completing the square, we obtain:

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mathrm{PL}^{(1)}((\theta_i)_{i\in A})\,\middle\|\,\mathrm{PL}^{(1)}((\theta'_i)_{i\in A})\right)\\
&\quad\leq\frac{\Theta'_A}{\Theta_A^2}\sum_{i\in A}\frac{\theta_i^2}{\theta'_i}-1\\
&\quad=\frac{\Theta'_A}{\Theta_A^2}\sum_{i\in A}\frac{(\theta_i-\theta'_i)^2}{\theta'_i}-\frac{(\Theta_A-\Theta'_A)^2}{\Theta_A^2}\\
&\quad\leq\frac{8k}{(k-m+1)^2}\sum_{i\in A}(\theta_i-\theta'_i)^2\leq\frac{32}{k}\sum_{i\in S}(\theta_i-\theta'_i)^2.
\end{aligned}\tag{3}$$

The last line follows from $\Theta'_A\leq k$, $\Theta_A\geq|A|/2\geq(k-m+1)/2$, $\theta'_i\geq1/2$, and $m\leq k/2$. The bound holds for every position and every preceding ranking. Therefore, applying the chain rule over the first m positions gives:

$$D_{\mathrm{KL}}\left(\mathrm{PL}^{(m)}((\theta_i)_{i\in S})\,\middle\|\,\mathrm{PL}^{(m)}((\theta'_i)_{i\in S})\right)\leq32\frac{m}{k}\sum_{i\in S}(\theta_i-\theta'_i)^2.\tag{4}$$

Suppose now that $m>k/2$. Let $(X_i)_{i\in S}$ be independent random variables with $X_i\sim\mathrm{Exp}(\theta_i)$.[11] Ordering the arms by increasing X_i gives a PL ranking: the probability that X_i is the minimum is θ_i/Θ_S, and the memoryless property gives the same selection rule at each subsequent position. The top-m ranking is thus a measurable function of $(X_i)_{i\in S}$. By the data-processing inequality and independence:

$$D_{\mathrm{KL}}\left(\mathrm{PL}^{(m)}((\theta_i)_{i\in S})\,\middle\|\,\mathrm{PL}^{(m)}((\theta'_i)_{i\in S})\right)\leq\sum_{i\in S}D_{\mathrm{KL}}\left(\mathrm{Exp}(\theta_i)\,\|\,\mathrm{Exp}(\theta'_i)\right)$$

[11] We use $\mathrm{Exp}(\lambda)\in\Delta([0,+\infty))$ to refer to the exponential distribution with rate $\lambda>0$, whose CDF is $F(x)=1-\exp(-\lambda x)$ for $x\geq0$ and $F(x)=0$ for $x<0$.

$$= \sum_{i \in S} \left(\ln \frac{\theta_i}{\theta'_i} + \frac{\theta'_i}{\theta_i} - 1 \right) \leq 2 \sum_{i \in S} (\theta_i - \theta'_i)^2.$$

For the last inequality, fix $a \in [1/2, 1]$ and let $f(b) \coloneqq \ln(a/b) + b/a - 1$ for $b \in [1/2, 1]$. Since $f(a) = f'(a) = 0$ and $f''(b) = 1/b^2 \leq 4$, Taylor's theorem gives $f(b) \leq 2(b-a)^2$. Using $m > k/2$, we conclude that:

$$D_{\mathrm{KL}} \left(\mathrm{PL}^{(m)}((\theta_i)_{i \in S}) \,\middle\|\, \mathrm{PL}^{(m)}((\theta'_i)_{i \in S}) \right) \leq 2 \sum_{i \in S} (\theta_i - \theta'_i)^2 \leq 4 \frac{m}{k} \sum_{i \in S} (\theta_i - \theta'_i)^2.$$

Thus, Equation (4) holds also when $m > k/2$, including $m = k$.

Finally, substituting Equation (4) into Equation (2) and interchanging the finite sums yields:

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

which proves the statement. □

We now derive the lower bound by constructing several finite games played in parallel, following the standard combinatorial-bandit approach of Audibert et al. (2014). The learner may pull several arms from the same game in one round. We account for this using a separate stopping time for each game.

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 $m \leq k \leq n$, randomized policy π and learning horizon $T \in \mathbb{N}_{\geq n}$, it holds:*

$$\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \mathbb{E}\mathsf{Reg}_{\pi,\nu}^{(k)}(T) \geq C_1 \sqrt{\frac{n-k}{m} \min \left\{ 1, \frac{n-k}{k} \right\} T},$$

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

Proof. If $n = k$, the regret and the lower bound are both zero. Assume henceforth that $d \coloneqq n - k \geq 1$. Fix a policy π and a horizon $T \geq n$. We first describe the hard instances, then relate their regret to stopped observations, and finally apply a change of measure to each game.

If $d \geq k$, let $g \coloneqq k$ and $s \coloneqq 1 + \lfloor d/k \rfloor \geq 2$. Construct g disjoint batches $\mathcal{B}_1, \ldots, \mathcal{B}_g$, each with s arms. This is possible because:

$$gs = k + k \left\lfloor \frac{d}{k} \right\rfloor \leq k + d = n.$$

There are $n - gs$ remaining arms, and no fixed optimal arms in this case. If instead $1 \leq d < k$, let $g \coloneqq d$ and $s \coloneqq 2$. Construct g disjoint binary batches and a disjoint set $\mathcal{F}$ of $k - d$ fixed optimal arms. Their total cardinality is $2d + (k - d) = n$. In the first case, write $\mathcal{F} \coloneqq \{\}$.

In both cases, $\mathcal{B}_a \subseteq [\![n]\!]$ and $|\mathcal{B}_a| = s$ for every $a \in [\![g]\!]$, with $\mathcal{B}_a \cap \mathcal{B}_{a'} = \{\}$ for distinct $a, a' \in [\![g]\!]$, and $\mathcal{F} \subseteq [\![n]\!] \setminus \bigcup_{a=1}^{g} \mathcal{B}_a$. Write $\mathcal{B}_a = \{b_{a,1}, \ldots, b_{a,s}\}$ for every $a \in [\![g]\!]$. We have $|\mathcal{F}| = k - g$, $2 \leq s \leq n - 1$, and $n \geq 2(s - 1)$. Define:

$$\Delta \coloneqq \frac{1}{64} \sqrt{\frac{ks}{mT}}. \tag{5}$$

If $d \geq k$, then $ks \leq n \leq T$, so $\Delta \leq 1/64$. If $d < k$, then $ks = 2k < 2n \leq 2T$, so $\Delta < \sqrt{2}/64 < 1/2$. Hence, $0 < \Delta \leq 1/2$ in both cases.

For every index vector $\boldsymbol{j} = (j_1, \ldots, j_g) \in ([\![s]\!])^g$, the coordinate j_a specifies the optimal arm b_{a,j_a} in batch a. These arms and the fixed arms in $\mathcal{F}$ have parameter 1, while all remaining arms are suboptimal with parameter $1 - \Delta$. Formally, define:

$$\theta_i^{\boldsymbol{j}} \coloneqq \begin{cases} 1 & \text{if } i \in \mathcal{F} \cup \{b_{a,j_a} \text{ s.t. } a \in [\![g]\!]\} \\ 1 - \Delta & \text{otherwise} \end{cases} \quad \text{for every } i \in [\![n]\!],$$

and let $\nu_{\boldsymbol{j}} \coloneqq (n, \boldsymbol{\theta}^{\boldsymbol{j}}, k, m)$. Since $|\mathcal{F}| + g = k$, there are exactly k arms with parameter 1. Thus, all parameters belong to $[1/2, 1]$, their maximum is 1, and the top-k separation is strict. In particular, every $\nu_{\boldsymbol{j}}$ belongs to $\mathcal{E}_n^{(k,m)}$ and its optimal set is:

$$S_{\boldsymbol{\theta}}^{(k)*} = \mathcal{F} \cup \{b_{a,j_a} \text{ s.t. } a \in [\![g]\!]\}.$$

Using the arm counts $N_i(t)$ defined above, set $N_{a,j}(t) \coloneqq N_{b_{a,j}}(t)$ and $N_{\mathcal{B}_a}(t) \coloneqq \sum_{j=1}^{s} N_{a,j}(t)$. Define the total number of omitted optimal arms over T rounds as:

$$M_{\boldsymbol{j}}(T) \coloneqq \sum_{a=1}^{g}(T - N_{a,j_a}(T)) + \sum_{i \in \mathcal{F}}(T - N_i(T)).$$

Since the pulled and optimal sets both have cardinality k, this also equals the total number of selected suboptimal arms. The regret accumulated over T rounds is therefore exactly:

$$\sum_{t=1}^{T}\left(g_{\boldsymbol{\theta}^{\boldsymbol{j}}}^{(k)}(S_{\boldsymbol{\theta}}^{(k)*}) - g_{\boldsymbol{\theta}^{\boldsymbol{j}}}^{(k)}(S_t)\right) = \frac{\Delta}{k} M_{\boldsymbol{j}}(T). \tag{6}$$

The learner can pull several arms from one batch in a round. For each batch $a \in [\![g]\!]$, we limit the total number of pulls used in the comparison by stopping at:

$$\tau_a \coloneqq \min\left\{T, \inf\{t \in \mathbb{N}_{\geq 1} \text{ s.t. } N_{\mathcal{B}_a}(t) \geq T\}\right\}, \tag{7}$$

with $\inf\{\} \coloneqq +\infty$. It is a stopping time because it depends only on the actions selected so far. At most s arms of a batch can be selected in one round. Hence:

$$N_{\mathcal{B}_a}(\tau_a) \leq T + s - 1, \qquad 0 \leq N_{a,j}(\tau_a) \leq T \text{ for every } j \in [\![s]\!]. \tag{8}$$

We next relate these stopped counts to the regret. Fix an alternative $\boldsymbol{j}$ and a batch a. If $N_{\mathcal{B}_a}(\tau_a) < T$, then necessarily $\tau_a = T$, so the stopped deficit $T - N_{a,j_a}(\tau_a)$ equals the final deficit $T - N_{a,j_a}(T)$. Otherwise:

$$T - N_{a,j_a}(\tau_a) \leq N_{\mathcal{B}_a}(\tau_a) - N_{a,j_a}(\tau_a) = \sum_{j \in [\![s]\!] \setminus \{j_a\}} N_{a,j}(\tau_a) \leq \sum_{j \in [\![s]\!] \setminus \{j_a\}} N_{a,j}(T).$$

Both final quantities are nonnegative, so the two cases imply:

$$T - N_{a,j_a}(\tau_a) \leq T - N_{a,j_a}(T) + \sum_{j \in [\![s]\!] \setminus \{j_a\}} N_{a,j}(T).$$

Summing over batches, the first term is at most $M_{\boldsymbol{j}}(T)$. The second counts only selected suboptimal arms within the batches, so it is also at most $M_{\boldsymbol{j}}(T)$. Combining this with Equation (6) and taking expectations gives:

$$\mathbb{E}\mathsf{Reg}_{\pi,\nu_{\boldsymbol{j}}}^{(k)}(T) \geq \frac{\Delta}{2k} \sum_{a=1}^{g}\left(T - \mathbb{E}_{\pi,\nu_{\boldsymbol{j}}}[N_{a,j_a}(\tau_a)]\right). \tag{9}$$

Fix $a \in [\![g]\!]$ and the distinguished indices $\boldsymbol{j}_{-a}$ in all other batches. Let $\nu_{\boldsymbol{j}_{-a}}^{(a,0)}$ be the auxiliary instance obtained by assigning parameter $1 - \Delta$ to every arm of batch a, while leaving all other parameters unchanged. For $j \in [\![s]\!]$, let $\nu_{\boldsymbol{j}_{-a}}^{(a,j)}$ increase only the parameter of $b_{a,j}$ from $1 - \Delta$ to 1. The auxiliary instance still has $k - 1 \geq 1$ arms with parameter 1, so its parameters remain normalized. Its top-k set need not be strictly separated, but it is used only as a reference measure. Comparing it with the s alternatives isolates the difficulty of identifying the optimal arm in batch a.

Define:

$$x_j \coloneqq \mathbb{E}_{\pi,\nu_{\boldsymbol{j}_{-a}}^{(a,0)}}[N_{a,j}(\tau_a)].$$

By Equation (8), $\sum_{j=1}^{s} x_j \leq T + s - 1$. We will use two consequences of this bound. Since $T \geq n \geq 2(s-1)$ and $T \geq n \geq s + 1$, respectively:

$$\sum_{j=1}^{s} x_j \leq \frac{3}{2}T, \qquad \frac{1}{s}\sum_{j=1}^{s} x_j \leq \frac{T + s - 1}{s} \leq \frac{2T}{s+1} \leq \frac{2}{3}T. \tag{10}$$

In the second inequality chain, $(T+s-1)/(sT) \le 2/(s+1)$ follows from $T \ge s+1$, and the last inequality uses $s \ge 2$.

The null and alternative j differ in just one parameter by Δ. Lemma 4 gives:

$$D_{\mathrm{KL}}\left(\mathbb{P}^{(\le\tau_a)}_{\pi,\nu^{(a,0)}_{\boldsymbol{j}_{-a}}} \,\middle\|\, \mathbb{P}^{(\le\tau_a)}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}\right) \le 32\frac{m}{k}\Delta^2 x_j. \tag{11}$$

The random variable $N_{a,j}(\tau_a)$ is measurable with respect to $\mathcal{H}_{\tau_a}$ and takes values in $[0,T]$. Applying Pinsker's inequality, in the form of Lemma 24, to the two stopped laws therefore yields:

$$\begin{aligned}\mathbb{E}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}[N_{a,j}(\tau_a)] &\le x_j + T\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}^{(\le\tau_a)}_{\pi,\nu^{(a,0)}_{\boldsymbol{j}_{-a}}} \,\middle\|\, \mathbb{P}^{(\le\tau_a)}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}\right)} \\ &\le x_j + 4T\Delta\sqrt{\frac{m}{k}}\sqrt{x_j}. \end{aligned}\tag{12}$$

Summing over $j \in [\![s]\!]$, applying Cauchy–Schwarz, and using Equations (5) and (10), we obtain:

$$\begin{aligned}\sum_{j=1}^{s}\mathbb{E}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}[N_{a,j}(\tau_a)] &\le \sum_{j=1}^{s} x_j + 4T\Delta\sqrt{\frac{m}{k}}\sum_{j=1}^{s}\sqrt{x_j} \\ &\le \frac{2}{3}sT + 4T\Delta\sqrt{\frac{m}{k}}\sqrt{\frac{3}{2}sT} \\ &= \left(\frac{2}{3}+\frac{\sqrt{6}}{32}\right)sT < \frac{3}{4}sT.\end{aligned}$$

The last strict inequality follows from $\sqrt{6} < 8/3$. Consequently, for every fixed a and $\boldsymbol{j}_{-a}$:

$$\sum_{j=1}^{s}\left(T - \mathbb{E}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}[N_{a,j}(\tau_a)]\right) \ge \frac{1}{4}sT. \tag{13}$$

Finally, average Equation (9) over all s^g hard instances. For each fixed batch a, there are s^{g-1} choices of $\boldsymbol{j}_{-a}$, and Equation (13) applies to every one of them. Reordering the finite sums gives:

$$\begin{aligned}\frac{1}{s^g}\sum_{\boldsymbol{j}\in([\![s]\!])^g}\mathbb{E}\mathsf{Reg}^{(k)}_{\pi,\nu_{\boldsymbol{j}}}(T) &\ge \frac{\Delta}{2ks^g}\sum_{a=1}^{g}\sum_{\boldsymbol{j}_{-a}\in([\![s]\!])^{g-1}}\sum_{j=1}^{s}\left(T - \mathbb{E}_{\pi,\nu^{(a,j)}_{\boldsymbol{j}_{-a}}}[N_{a,j}(\tau_a)]\right) \\ &\ge \frac{\Delta}{2ks^g}\, g\, s^{g-1}\,\frac{1}{4}sT = \frac{\Delta g T}{8k} = \frac{g}{512k}\sqrt{\frac{ksT}{m}}.\end{aligned}$$

Since every hard instance belongs to $\mathcal{E}_n^{(k,m)}$, its supremum is at least this average. If $d \ge k$, then $g = k$ and $s > d/k$, yielding:

$$\sup_{\nu\in\mathcal{E}_n^{(k,m)}}\mathbb{E}\mathsf{Reg}^{(k)}_{\pi,\nu}(T) \ge \frac{1}{512}\sqrt{\frac{dT}{m}}.$$

If $1 \le d < k$, then $g = d$ and $s = 2$, yielding:

$$\sup_{\nu\in\mathcal{E}_n^{(k,m)}}\mathbb{E}\mathsf{Reg}^{(k)}_{\pi,\nu}(T) \ge \frac{\sqrt{2}}{512}\frac{d}{\sqrt{k}}\sqrt{\frac{T}{m}} \ge \frac{1}{512}\frac{d}{\sqrt{k}}\sqrt{\frac{T}{m}}.$$

Combining the cases proves the theorem with $C_1 = 1/512$:

$$\sup_{\nu\in\mathcal{E}_n^{(k,m)}}\mathbb{E}\mathsf{Reg}^{(k)}_{\pi,\nu}(T) \ge C_1\min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\}\sqrt{\frac{T}{m}}.$$

□

Algorithm 2 Subroutines for TOP-k-UIF.

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

function COUNT-BA($\mathcal{P}, \mathcal{U}, (\mathrm{UCB}_i^{(p)})_{i\in\mathcal{U}, p\in\mathcal{P}}, (\mathrm{LCB}_i^{(p)})_{i\in\mathcal{U}, p\in\mathcal{P}}$)
  Let $i \prec j$ if and only if $\mathrm{LCB}_i^{(p)} > \mathrm{UCB}_j^{(p)}$ for some $p \in \mathcal{P}$
  Let $\prec^+$ be the transitive closure of $\prec$
  for $i \in \mathcal{U}$ **do**
    $A_i \leftarrow |\{j \in \mathcal{U} \setminus \{i\} \text{ s.t. } i \prec^+ j\}|$
    $B_i \leftarrow |\{j \in \mathcal{U} \setminus \{i\} \text{ s.t. } j \prec^+ i\}|$
  end for
  return $(A_i)_{i\in\mathcal{U}}, (B_i)_{i\in\mathcal{U}}$
end function

function NEW-PIVOT(p, $\mathcal{U}, (\mathrm{LCB}_i^{(p)})_{i\in\mathcal{U}}$)
  $\mathcal{P}' \leftarrow \{i \in \mathcal{U} \text{ s.t. } \mathrm{LCB}_i^{(p)} > 1/2\}$
  if $\mathcal{P}' \neq \{\}$ **then**
    Sample $p' \sim \mathrm{Unif}(\arg\max_{i\in\mathcal{P}'} \mathrm{LCB}_i^{(p)})$
  else
    $p' \leftarrow \perp$
  end if
  return p'
end function

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).

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.

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

Input: $T \in \mathbb{N}_{\geq 1}$, $n \in \mathbb{N}_{\geq 2}$, $k \in [\![2, n]\!]$, $\beta_{\delta,n,T} : [0, +\infty] \to [0, +\infty]$, $\delta \in (0, 1]$
Initialize $P_1 \leftarrow 0$, $\mathcal{A}_1 \leftarrow \{\}$, $\mathcal{U}_1 \leftarrow [\![n]\!]$,
    $N_i^{(p)} \leftarrow 0$ if $i \neq p$, $+\infty$ otherwise, $\hat{\mu}_i^{(p)} \leftarrow 1/2$ for every $i, p \in [\![0, n]\!]$
for $t \in [\![T]\!]$ **do**
    $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t, \mathcal{U}_{t+1} \leftarrow \mathcal{U}_t$, $P_{t+1} \leftarrow P_t$
    Let $\mathcal{P}_t \leftarrow \mathcal{A}_t \cup \{P_t\}$
    Sample $S_{\text{exp},t} \sim \text{Unif}(\{S \subseteq \mathcal{U}_t \setminus \{P_t\} \text{ s.t. } |S| = k - |\mathcal{P}_t| + \mathbb{I}[0 \in \mathcal{P}_t]\})$
    Pull $S_t \leftarrow (\mathcal{P}_t \cup S_{\text{exp},t}) \setminus \{0\}$
    Observe $R_t^{(1)} : S_t \cup \{0\} \to [\![\min\{k, 2\}]\!]$
    // 1. Update Statistics & Confidence Bounds
    $(\hat{\mu}_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}$
        $\leftarrow$ RB-UPDATE$(\mathcal{P}_t, S_t \cup \{0\}, (\hat{\mu}_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}, (N_i^{(p)})_{i \in S_t \cup \{0\}, p \in \mathcal{P}_t}, R_t^{(1)})$
    for $i \in \mathcal{U}_t, p \in \mathcal{P}_t$ **do**
        $\text{UCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} + \beta_{\delta,n,T}(N_i^{(p)})$
        $\text{LCB}_i^{(p)} \leftarrow \hat{\mu}_i^{(p)} - \beta_{\delta,n,T}(N_i^{(p)})$
    end for
    // 2. Update Unknown Arms
    $(A_i)_{i \in \mathcal{U}_t}, (B_i)_{i \in \mathcal{U}_t} \leftarrow$ COUNT-BA$(\mathcal{P}_t, \mathcal{U}_t, (\text{UCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t}, (\text{LCB}_i^{(p)})_{i \in \mathcal{U}_t, p \in \mathcal{P}_t})$
    for $i \in \mathcal{U}_t$ **do**
        if $B_i \geq k - |\mathcal{A}_t|$ **then** ▷ Reject arm
            $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
        else if $A_i \geq |\mathcal{A}_t \cup \mathcal{U}_t| - k$ **then** ▷ Accept arm
            $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}$
            $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
        end if
    end for
    // 3. Pivot Update
    if $\mathcal{A}_t \neq \{\}$ **then** ▷ Stop pivot update after first accepted arm
        for $i \in \mathcal{U}_t$ **do**
            if $\text{LCB}_i^{(P_t)} > 1/2$ **then** ▷ Accept arms which are better than the (accepted) pivot
                $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_{t+1} \cup \{i\}$
                $\mathcal{U}_{t+1} \leftarrow \mathcal{U}_{t+1} \setminus \{i\}$
            end if
        end for
    else
        $p' \leftarrow$ NEW-PIVOT$(P_t, \mathcal{U}_t, (\text{LCB}_i^{(P_t)})_{i \in \mathcal{U}_t})$
        if $p' \neq \perp$ **then**
            $P_{t+1} \leftarrow p'$
        else if $P_t = 0$ and $\mathcal{A}_{t+1} \neq \{\}$ **then**
            Sample $P_{t+1} \sim \text{Unif}(\mathcal{A}_{t+1})$ ▷ Make an accepted arm the pivot
        end if
    end if
    if $|\mathcal{A}_{t+1}| = k$ **then**
        $\mathcal{U}_{t+1} \leftarrow \{\}$
    end if
end for

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

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 $\mathrm{LCB}_i^{(p)} > 1/2$. If such a condition is satisfied, the routine returns one of the unknown items with the highest $\mathrm{LCB}_i^{(p)}$ 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 $S_{\mathrm{exp},t}$. 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.

The appendix is organized as follows. First, we establish concentration results (Appendix D.1) and use them to show the correctness of Algorithm 1 (Appendix D.2). We then bound the regret suffered by the algorithm during an *epoch*, i.e., a portion of the horizon throughout which the pivot is fixed (Appendix D.3). Next, we combine the epoch bounds over the entire horizon, obtaining the final regret bound (Appendix D.4). Finally, Appendix D.5 adapts these results to MNL bandits.

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., $\hat{\mu}_{i,t}^{(p)}, N_{i,t}^{(p)}$). 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 $\nu \in \mathcal{E}_n^{(k,m)}$, $S \subseteq [\![n]\!]$, $|S| = k$, and distinct $i, p \in S$, we define:

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

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

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

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

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

Proof.

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

$$\begin{aligned}
\mathbb{P}_{R\sim\mathrm{PL}((\theta_i)_{i\in S})}\left[\mathcal{C}^{(p)}_u \cap \mathcal{C}^{(i)}_u \mid \mathcal{C}^{(p)}_{u-1} \cap \mathcal{C}^{(i)}_{u-1}\right] &= 1 - \mathbb{P}_{R\sim\mathrm{PL}((\theta_i)_{i\in S})}\left[\overline{\mathcal{C}^{(p)}_u} \cup \overline{\mathcal{C}^{(i)}_u} \mid \mathcal{C}^{(p)}_{u-1} \cap \mathcal{C}^{(i)}_{u-1}\right] \\
&= 1 - \mathbb{P}_{R\sim\mathrm{PL}((\theta_i)_{i\in S})}\left[I_u \in \{p, i\} \mid \mathcal{C}^{(p)}_{u-1} \cap \mathcal{C}^{(i)}_{u-1}\right] \\
&= 1 - \frac{\mathbb{E}_{R\sim\mathrm{PL}((\theta_i)_{i\in S})}\left[\frac{\theta_p+\theta_i}{\Theta_S - \sum_{v=1}^{u-1}\theta_{I_v}} \mathbb{I}_{\mathcal{C}^{(p)}_{u-1}} \mathbb{I}_{\mathcal{C}^{(i)}_{u-1}}\right]}{\mathbb{P}_{R\sim\mathrm{PL}((\theta_i)_{i\in S})}\left[\mathcal{C}^{(p)}_{u-1} \cap \mathcal{C}^{(i)}_{u-1}\right]} \qquad (14) \\
&\leq 1 - \frac{\theta_p + \theta_i}{k} \quad \text{for every } u \in [\![m]\!], \qquad (15)
\end{aligned}$$

   where Step (14) follows from the tower rule, conditioning on $I_1, \ldots, I_{u-1}$, and Step (15) follows from the fact that $\theta_j \leq 1$ for $j \in S$ and $|S| = k$. Then:

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

   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 $X_i \sim \mathrm{Gumbel}(\ln\theta_i, 1)$ for every $i \in S$,[12] and the bijection $R : S \to [\![k]\!]$ satisfies:

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

   then $R \sim \mathrm{PL}((\theta_i)_{i\in S})$ (under arbitrary tie-breaking, as ties occur with probability 0). Let $Y_i = e^{-X_i}$ for $i \in S$. Then $Y_i \sim \mathrm{Exp}(\theta_i)$, and:

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

   Let $M_{i,p} \coloneqq \min\{Y_i, Y_p\}$ 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 $Z_{i,p} \coloneqq +\infty$ otherwise. The event $R^{(m)}(i) \neq R^{(m)}(p)$ is precisely $\{M_{i,p} \leq Z_{i,p}\}$. By the exponential-race property, $M_{i,p} \sim \mathrm{Exp}(\theta_i + \theta_p)$ 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:

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

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

□

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

[12] We use $\mathrm{Gumbel}(u, 1) \in \Delta(\mathbb{R})$ to refer to the Gumbel distribution with location $u \in \mathbb{R}$ and unit scale, whose CDF is $F(x) = \exp(-\exp(-(x-u)))$ for $x \in \mathbb{R}$.

$s \leq t \leq T$, let:[13,14]

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

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 $i, p \in [\![n]\!]$, $i \neq p$, $t \in [\![T]\!]$, let:

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

We use the convention $\mu_i^{(i)} \coloneqq 1/2$. 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 $i, p \in [\![n]\!]$, $i \neq p$, $t \in [\![T+1]\!]$. Finally, for any $T \in \mathbb{N}_{\geq 1}$, $\delta \in (0, 1]$, define:

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

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 $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq n}$, $\delta \in (0, 1]$:*

$$\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| > \frac{\sqrt{3}}{2}\beta_{\delta,n,T}(N_{i,t}^{(p)}) \text{ for some } i, p \in [\![n]\!], t \in [\![T+1]\!]\right] \leq \delta.$$

Proof. Fix $t \in [\![T+1]\!]$ and $i, p \in [\![n]\!]$. 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:

$$|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| = \left|\frac{\sum_{s=1}^{t-1} W_{i,s}^{(p)}}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}} - \frac{\sum_{s=1}^{t-1} O_{i,s}^{(p)} \mu_i^{(p)}}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}}\right| \tag{16}$$
$$= \left|\frac{\sum_{s=1}^{t-1} O_{i,s}^{(p)}(W_{i,s}^{(p)} - \mu_i^{(p)})}{\sum_{s=1}^{t-1} O_{i,s}^{(p)}}\right|, \tag{17}$$

where Step (16) follows from Lemma 5, while Step (17) follows from $W_{i,s}^{(p)} = O_{i,s}^{(p)} W_{i,s}^{(p)}$. Thus:

$$|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| \geq \frac{\sqrt{3}}{2}\beta_{\delta,n,T}(N_{i,t}^{(p)}) \text{ iff}$$
$$\left|\sum_{s=1}^{t-1} O_{i,s}^{(p)}(W_{i,s}^{(p)} - \mu_i^{(p)})\right| \geq \frac{\sqrt{3}}{2}\beta_{\delta,n,T}(N_{i,t}^{(p)}) N_{i,t}^{(p)}. \tag{18}$$

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

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

[13] Given two tuples a, b, we denote by (a, b) their concatenation.

[14] Note 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.

The process:

$$M_{i,u}^{(p)} \coloneqq \sum_{s=1}^{u} O_{i,s}^{(p)}(W_{i,s}^{(p)} - \mu_i^{(p)}), \qquad u \in [\![0, T]\!],$$

is adapted to $(\mathcal{F}_{i,u}^{(p)})_{u=0}^{T}$. Moreover, for every $u \in [\![T]\!]$:

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

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 $\tau_{i,0}^{(p)} \coloneqq 0$ and, for every $q \in [\![T]\!]$, recursively define:

$$\tau_{i,q}^{(p)} \coloneqq \min\left\{u \in [\![\tau_{i,q-1}^{(p)} + 1, T]\!] \text{ s.t. } O_{i,u}^{(p)} = 1\right\},$$

with $\min\{\} \coloneqq T$. The variables $(\tau_{i,q}^{(p)})_{q=0}^{T}$ are nondecreasing bounded stopping times. By the Optional Sampling theorem (Klenke, 2020, Theorem 10.12), $(M_{i,\tau_{i,q}^{(p)}}^{(p)})_{q=0}^{T}$ is a martingale w.r.t. $(\mathcal{F}_{i,\tau_{i,q}^{(p)}}^{(p)})_{q=0}^{T}$. For every $q \in [\![T]\!]$, 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 during the horizon. Therefore, for every $N \in [\![T]\!]$, Azuma's inequality (Theorem 25) yields:

$$\begin{aligned}\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[|M_{i,\tau_{i,N}^{(p)}}^{(p)}| \geq \frac{\sqrt{3}}{2}\beta_{\delta,n,T}(N)N\right] &\leq 2\exp\left(-\frac{3}{2}N\beta_{\delta,n,T}^2(N)\right) \\ &= 2\left(\frac{\delta}{2nT}\right)^{3/2} \leq \frac{\delta}{n^2T}.\end{aligned}$$

The last inequality uses $\delta \leq 1$ and $T \geq n$. For any $t \in [\![T+1]\!]$ with $N_{i,t}^{(p)} = N \geq 1$, the N-th observed comparison occurs by round $t-1$, and $M_{i,\tau_{i,N}^{(p)}}^{(p)} = M_{i,t-1}^{(p)}$. Thus every deviation in Step (18) is covered by one of the events above. A union bound over at most n^2 ordered pairs and T positive observation counts proves the claim. □

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 $i, p \in [\![n]\!]$, $i \neq p$, $t \in [\![T]\!]$, let:

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

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

$$\begin{aligned}\tilde{O}_{i,t}^{(p)} &= \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[p \in \mathcal{P}_t, i \in S_t, R_t^{(m)}(i) \neq R_t^{(m)}(p) \mid \overline{H}_{0:t-1}\right] \\ &= \mathbb{I}\left[p \in \mathcal{P}_t\right]\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathbb{I}[i \in S_t]\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[R_t^{(m)}(i) \neq R_t^{(m)}(p) \mid \overline{H}_{0:t-1}, S_t\right] \mid \overline{H}_{0:t-1}\right] \\ &= \mathbb{I}[p \in \mathcal{P}_t]\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathbb{I}[i \in S_t]f_{i|S_t}^{(p)} \mid \overline{H}_{0:t-1}\right]. \qquad (19)\end{aligned}$$

Let $\tilde{N}_{i,t}^{(p)} \coloneqq \sum_{s=1}^{t-1}\tilde{O}_{i,s}^{(p)}$ for every $t \in [\![T+1]\!]$. Before stating the concentration results for the observation counts, we show that an arm with a larger PL parameter is at least as likely to be compared with the pivot before any arm has been accepted ($\mathcal{A}_t = \{\}$).

Lemma 7 (Ordering of Observation Probabilities)**.** *Fix a history $\overline{H}_{0:t-1}$ with $\mathcal{A}_t = \{\}$ and $P_t = p$. For $i, j \in \mathcal{U}_t \setminus \{p\}$ with $i \neq j$ and $\theta_i \geq \theta_j$:*

$$\tilde{O}_{i,t}^{(p)} \geq \tilde{O}_{j,t}^{(p)}.$$

Proof. Condition on $\overline{H}_{0:t-1}$. Since $\mathcal{A}_t = \{\}$:

$$\mathcal{P}_t = \{p\}, \qquad S_t \setminus \{p\} \sim \mathrm{Unif}\left(\{S \subseteq \mathcal{U}_t \setminus \{p\} \text{ s.t. } |S| = k-1\}\right).$$

Independently of S_t, draw independent $Y_h \sim \mathrm{Exp}(\theta_h)$ for $h \in \mathcal{U}_t \cup \{p\}$. For each $S \subseteq \mathcal{U}_t \cup \{p\}$, let $Y_S^{(m)}$ be the m-th smallest of $(Y_h)_{h\in S}$, with $Y_S^{(m)} = +\infty$ if $m > |S|$. If $S_t \cap \{i,j\} = \{\}$, then $O_{i,t}^{(p)} = O_{j,t}^{(p)} = 0$. For the other cases:

1. $|S_t \cap \{i,j\}| = 1$. For every $S \subseteq \mathcal{U}_t \setminus \{p,i,j\}$ with $|S| = k-2$:

$$\mathbb{P}_{\text{Top-}k\text{-UIF},\nu}[S_t = S \cup \{p,i\} \mid \overline{H}_{0:t-1}] = \mathbb{P}_{\text{Top-}k\text{-UIF},\nu}[S_t = S \cup \{p,j\} \mid \overline{H}_{0:t-1}].$$

   By the exponential representation of PL rankings (see the proof of Lemma 5), for $h \in \{i,j\}$ the comparison is observed if and only if $\min\{Y_h, Y_p\} \le Y_S^{(m)}$. Conditioning on $Y_S^{(m)}$ and using the independence of Y_h, Y_p, and $(Y_a)_{a\in S}$ gives:

$$\begin{aligned} f_{h|S\cup\{p,h\}}^{(p)} &= 1 - \mathbb{P}_{Y_a\sim\mathrm{Exp}(\theta_a)}[Y_h > Y_S^{(m)},\ Y_p > Y_S^{(m)}] \\ &= 1 - \mathbb{E}_{Y_a\sim\mathrm{Exp}(\theta_a)}\left[\mathbb{P}_{Y_a\sim\mathrm{Exp}(\theta_a)}[Y_h > Y_S^{(m)},\ Y_p > Y_S^{(m)} \mid Y_S^{(m)}]\right] \\ &= \mathbb{E}_{Y_a\sim\mathrm{Exp}(\theta_a)}\left[1 - e^{-\theta_h Y_S^{(m)}} e^{-\theta_p Y_S^{(m)}}\right] = \mathbb{E}_{Y_a\sim\mathrm{Exp}(\theta_a)}\left[1 - e^{-(\theta_h+\theta_p)Y_S^{(m)}}\right]. \end{aligned}$$

   Subtracting the expressions for $h = i$ and $h = j$ therefore yields:

$$f_{i|S\cup\{p,i\}}^{(p)} - f_{j|S\cup\{p,j\}}^{(p)} = \mathbb{E}_{Y_h\sim\mathrm{Exp}(\theta_h)}\left[e^{-(\theta_j+\theta_p)Y_S^{(m)}} - e^{-(\theta_i+\theta_p)Y_S^{(m)}}\right] \ge 0.$$

2. $\{i,j\} \subseteq S_t$. For every $S \subseteq \mathcal{U}_t \cup \{p\}$ with $\{p,i,j\} \subseteq S$ and $|S| = k$:

$$\begin{aligned} f_{i|S}^{(p)} - f_{j|S}^{(p)} &= \left(1 - \mathbb{P}_{Y_h\sim\mathrm{Exp}(\theta_h)}[Y_i > Y_S^{(m)},\ Y_p > Y_S^{(m)}]\right) \\ &\quad - \left(1 - \mathbb{P}_{Y_h\sim\mathrm{Exp}(\theta_h)}[Y_j > Y_S^{(m)},\ Y_p > Y_S^{(m)}]\right) \\ &= \mathbb{P}_{Y_h\sim\mathrm{Exp}(\theta_h)}[Y_i \le Y_S^{(m)} < \min\{Y_j, Y_p\}] \\ &\quad - \mathbb{P}_{Y_h\sim\mathrm{Exp}(\theta_h)}[Y_j \le Y_S^{(m)} < \min\{Y_i, Y_p\}], \end{aligned}$$

   where the cases in which all three arrival times Y_i, Y_j, Y_p exceed $Y_S^{(m)}$ cancel. To compare the two probabilities, pair each realization in the first event with the realization obtained by exchanging Y_i and Y_j, keeping all other arrival times $(Y_h)_{h\in S\setminus\{i,j\}}$ fixed. This exchange preserves $Y_S^{(m)}$ and maps the first event bijectively onto the second. In the first event, $Y_i \le Y_S^{(m)} < Y_j$, so $Y_i < Y_j$. The ratio of the joint density at this realization to the joint density at the exchanged realization is:

$$\frac{\theta_i\theta_j e^{-\theta_i Y_i - \theta_j Y_j}}{\theta_i\theta_j e^{-\theta_i Y_j - \theta_j Y_i}} = e^{(\theta_i-\theta_j)(Y_j - Y_i)} \ge 1.$$

   The density factors for all other arrival times $(Y_h)_{h\in S\setminus\{i,j\}}$ cancel by independence. Thus every realization in the first event has at least as much density as its counterpart in the second. Exchanging two coordinates preserves volume, so integrating over these paired realizations shows that the first probability is at least the second, and $f_{i|S}^{(p)} \ge f_{j|S}^{(p)}$.

Summing over the sampled sets in Equation (19) yields $\tilde{O}_{i,t}^{(p)} \ge \tilde{O}_{j,t}^{(p)}$. □

The following lemma controls the deviations of the observation counts $N_{i,t}^{(p)}$ from their predictable counterparts $\tilde{N}_{i,t}^{(p)}$.

Lemma 8 (Concentration of Observation Counts)**.** *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\ge 1}$, $\delta \in (0,1]$, and $\lambda > 0$:*

$$\mathbb{P}_{\text{Top-}k\text{-UIF},\nu}\left[\begin{array}{ccc} N_{i,t}^{(p)} \ge & \frac{e^\lambda - 1}{\lambda}\tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda}\ln\left(\frac{2n}{\delta}\right) & \text{for some } i,p \in [\![n]\!],\ i \ne p, \\ \text{or} & & \text{and } t \in [\![T+1]\!] \\ N_{i,t}^{(p)} \le & \frac{1-e^{-\lambda}}{\lambda}\tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda}\ln\left(\frac{2n}{\delta}\right) & \end{array}\right] \le \delta.$$

Proof. We prove Lemma 8. For $\lambda \in \mathbb{R}$, let:

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

for every $u \in [\![0,T]\!]$. Then, for every $u \in [\![T]\!]$:

$$\begin{aligned}
&\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathcal{M}_{i,u}^{(p)}(\lambda) \mid \overline{H}_{0:u-1}\right] \\
&\quad= \mathcal{M}_{i,u-1}^{(p)}(\lambda)\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu}\left[\exp(\lambda O_{i,u}^{(p)} - \tilde{O}_{i,u}^{(p)}(e^{\lambda}-1)) \mid \overline{H}_{0:u-1}\right] \\
&\quad= \mathcal{M}_{i,u-1}^{(p)}(\lambda)\exp\left(-\tilde{O}_{i,u}^{(p)}(e^{\lambda}-1)\right)\left(1+\tilde{O}_{i,u}^{(p)}(e^{\lambda}-1)\right) \\
&\quad\le \mathcal{M}_{i,u-1}^{(p)}(\lambda), \qquad (20)
\end{aligned}$$

where the equality uses that, conditionally on $\overline{H}_{0:u-1}$, $O_{i,u}^{(p)} \sim \mathrm{Be}(\tilde{O}_{i,u}^{(p)})$,[15] and Step (20) follows from $1+x \le e^x$. Thus $(\mathcal{M}_{i,u}^{(p)}(\lambda))_{u=0}^{T}$ is a supermartingale w.r.t. the natural filtration induced by the histories $(\sigma(\overline{H}_{0:u}))_{u=0}^{T}$. Furthermore, $\mathcal{M}_{i,0}^{(p)}(\lambda) = 1$. Now, by the maximal inequality (Lattimore and Szepesvári, 2020, Theorem 3.9):

$$\begin{aligned}
&\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathcal{M}_{i,u}^{(p)}(\lambda) \ge \epsilon \text{ for some } u \in [\![0,T]\!], i,p \in [\![n]\!], i \ne p\right] \\
&\quad\le \sum_{\substack{i,p\in[\![n]\!]\\ i\ne p}} \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathcal{M}_{i,u}^{(p)}(\lambda) \ge \epsilon \text{ for some } u \in [\![0,T]\!]\right] \\
&\quad\le \sum_{\substack{i,p\in[\![n]\!]\\ i\ne p}} \frac{\mathbb{E}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathcal{M}_{i,0}^{(p)}(\lambda)\right]}{\epsilon} \le \frac{n^2}{\epsilon} \text{ for every } \epsilon > 0.
\end{aligned}$$

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

$$\begin{aligned}
\frac{\delta}{2} &\ge \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[\mathcal{M}_{i,u}^{(p)}(\lambda) \ge (2n/\delta)^2 \text{ for some } u \in [\![0,T]\!], i,p \in [\![n]\!], i \ne p\right] \\
&= \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[\lambda N_{i,t}^{(p)} - (e^{\lambda}-1)\tilde{N}_{i,t}^{(p)} \ge 2\ln(2n/\delta) \text{ for some } t \in [\![T+1]\!], i,p \in [\![n]\!], i \ne p\right] \\
&= \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[N_{i,t}^{(p)} \ge \frac{e^{\lambda}-1}{\lambda}\tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda}\ln(2n/\delta) \text{ for some } t \in [\![T+1]\!], i,p \in [\![n]\!], i \ne p\right]. \qquad (21)
\end{aligned}$$

By computations analogous to Equation (21), with $-\lambda$ in place of λ, we obtain:

$$\mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[N_{i,t}^{(p)} \le \frac{1-e^{-\lambda}}{\lambda}\tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda}\ln(2n/\delta) \text{ for some } t \in [\![T+1]\!],\ i,p \in [\![n]\!],\ i \ne p\right] \le \frac{\delta}{2}.$$

A union bound over the two tails proves the claim.

□

Lemmas 6 and 8 control the estimation errors and the deviations of the observation counts from their predictable counterparts, respectively, and induce the following good events:

$$\mathcal{G}_{\text{conf},T}^{(\delta)} \coloneqq \left\{|\hat{\mu}_{i,t}^{(p)} - \mu_i^{(p)}| \le \frac{\sqrt{3}}{2}\beta_{\delta,n,T}(N_{i,t}^{(p)}) \text{ for every } i,p \in [\![n]\!], t \in [\![T+1]\!]\right\},$$

$$\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)} \coloneqq \left\{\frac{1-e^{-\lambda}}{\lambda}\tilde{N}_{i,t}^{(p)} - \frac{2}{\lambda}\ln\left(\frac{2n}{\delta}\right) < N_{i,t}^{(p)} < \frac{e^{\lambda}-1}{\lambda}\tilde{N}_{i,t}^{(p)} + \frac{2}{\lambda}\ln\left(\frac{2n}{\delta}\right)\right.$$

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

$$\text{for every } i, p \in [\![n]\!], i \neq p, t \in [\![T+1]\!] \Bigg\},$$

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

Because the lower confidence bound in Algorithm 1 is defined by subtracting $\beta_{\delta,n,T}(N_{i,t}^{(p)})$ from $\hat{\mu}_{i,t}^{(p)}$, under $\mathcal{G}_{\text{conf},T}^{(\delta)}$ we have:

$$\text{LCB}_{i,t}^{(p)} > \frac{1}{2} \text{ implies } \mu_i^{(p)} - \frac{1}{2} > \left(1 - \frac{\sqrt{3}}{2}\right) \beta_{\delta,n,T}(N_{i,t}^{(p)}) > \frac{1}{8} \beta_{\delta,n,T}(N_{i,t}^{(p)}). \tag{22}$$

We will use this implication to show that the next pivot's pairwise win probability against the current pivot exceeds $1/2$ by at least a constant fraction of the corresponding excess for a best arm. This will control the regret accumulated across further pivot changes once the pivot parameter is at least $1/2$.

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 9. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in [\![T+1]\!]$:*

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 $\mathcal{G}_{conf,T}^{(\delta)}$:

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

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 9.1 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 9.1.

We next prove Lemma 9.3 by induction. It holds at $t = 1$ because $\mathcal{A}_1 = \{\}$ and $\mathcal{U}_1 = [\![n]\!]$. Suppose it holds at the beginning of round t. Since $\mathcal{A}_t$ and $\mathcal{U}_t$ are disjoint, the induction hypothesis gives:

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

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:

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

An optimal unknown arm has at most $|\mathcal{U}_t \cap S_{\boldsymbol{\theta}}^{(k)*}| - 1$ higher-parameter unknown arms, so the rejection rule cannot reject it. Likewise, a suboptimal unknown arm has at most $|\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}| - 1$

lower-parameter unknown arms, so the first acceptance rule cannot accept it. Finally, when $\mathcal{A}_t \neq \{\}$, Lemma 9.1 and the induction hypothesis give $P_t \in \mathcal{A}_t \subseteq S_{\boldsymbol{\theta}}^{(k)*}$. Thus $\mathrm{LCB}_{i,t+1}^{(P_t)} > 1/2$ implies $\theta_i > \theta_{P_t}$ under $\mathcal{G}_{\mathrm{conf},T}^{(\delta)}$, so the second acceptance rule also accepts only optimal arms. This proves Lemma 9.3 at round $t+1$.

Lemma 9.2 now follows from Lemma 9.3, the disjointness of $\mathcal{A}_t$ and $\mathcal{U}_t$, and $|S_{\boldsymbol{\theta}}^{(k)*}| = k$. □

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 Line 34 of Algorithm 1 and the structural guarantees established in Lemma 9.

Lemma 10. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in [\![T+1]\!]$, under the good event $\mathcal{G}_{conf,T}^{(\delta)}$, if $|\mathcal{A}_t| = k$, then $\mathcal{U}_t = \{\}$. As a consequence, when $|\mathcal{A}_t| = k$, we have $\mathcal{A}_t = S_{\boldsymbol{\theta}}^{(k)*} = \mathcal{A}_t \cup \mathcal{U}_t$.*

Proof. Line 34 of Algorithm 1 sets $\mathcal{U}_t = \{\}$ whenever the preceding round produces $|\mathcal{A}_t| = k$. Lemma 9 then gives $\mathcal{A}_t \subseteq S_{\boldsymbol{\theta}}^{(k)*} \subseteq \mathcal{A}_t \cup \mathcal{U}_t$, and the conclusion follows. □

For any $t \in [\![T+1]\!]$, let $\mathcal{U}_t' \coloneqq \mathcal{U}_t \setminus \{P_t\}$ and $\mathcal{A}_t' \coloneqq \mathcal{A}_t \setminus \{P_t\}$.

Consequently, whenever a suboptimal arm remains unknown, an optimal unknown arm distinct from the pivot is available for comparisons.

Corollary 11. *For every $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, and $t \in [\![T]\!]$, under the good event $\mathcal{G}_{conf,T}^{(\delta)}$, the existence of $j \in \mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}$ implies the existence of $i \in \mathcal{U}_t' \cap S_{\boldsymbol{\theta}}^{(k)*}$.*

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

If $\mathcal{A}_t = \{\}$, Lemma 9 gives $|\mathcal{U}_t \cap S_{\boldsymbol{\theta}}^{(k)*}| = k$. 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 10, $|\mathcal{A}_t| < k$, so Lemma 9 yields $|\mathcal{U}_t \cap S_{\boldsymbol{\theta}}^{(k)*}| = k - |\mathcal{A}_t| \geq 1$. Moreover, Lemma 9.1 gives $P_t \in \mathcal{A}_t$, and hence $\mathcal{U}_t' \cap S_{\boldsymbol{\theta}}^{(k)*} = \mathcal{U}_t \cap S_{\boldsymbol{\theta}}^{(k)*} \neq \{\}$. □

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

We now use correctness to bound regret in terms of the arms that remain unknown. We split the horizon into epochs with fixed pivots and analyze each epoch separately. The main step is to bound the gaps between PL parameters using the confidence radii, and then sum the resulting instantaneous pseudo-regret bounds. This gives both instance-dependent and worst-case bounds for an epoch. We also exploit the absence of accepted arms in epochs that end before the horizon to sharpen these bounds. Appendix D.4 will combine these estimates over the entire horizon.

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:

$$\mathsf{Reg}_{\pi,\nu}^{(k)}(s,t) \coloneqq \sum_{u=s}^{t} \mathsf{IstReg}_{\pi,\nu}^{(k)}(u),$$

and set $\mathsf{Reg}_{\pi,\nu}^{(k)}(s,t) \coloneqq 0$ when $s > t$. Here:

$$\mathsf{IstReg}_{\pi,\nu}^{(k)}(t) \coloneqq g_{\boldsymbol{\theta}}^{(k)}(S_{\boldsymbol{\theta}}^{(k)*}) - \sum_{\substack{S \subseteq [\![n]\!] \\ |S|=k}} g_{\boldsymbol{\theta}}^{(k)}(S)\pi(S \mid H_{t-1})$$

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

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

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

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

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

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

under both the good events $\mathcal{G}^{(\delta)}_{\text{conf},T}$ and $\mathcal{G}^{(\delta,\lambda)}_{\text{pulls}}$. We say that an epoch l is *not final* if $l < L(T)$, and *high* or *low* according as $\theta_{P^{(l)}} \geq 1/2$ or $\theta_{P^{(l)}} < 1/2$.

We denote the corresponding sets of epoch indices by:

$$\mathfrak{L}_{\text{high}}(T) \coloneqq \{l \in \mathfrak{L}(T) \text{ s.t. } \theta_{P^{(l)}} \geq 1/2\}, \qquad \mathfrak{L}_{\text{low}}(T) \coloneqq \mathfrak{L}(T) \setminus \mathfrak{L}_{\text{high}}(T),$$
$$\mathfrak{L}_{\text{final}}(T) \coloneqq \{L(T)\}, \qquad \mathfrak{L}_{\neg\text{final}}(T) \coloneqq \mathfrak{L}(T) \setminus \mathfrak{L}_{\text{final}}(T).$$

For high and low epochs that are not final, respectively, we use:

$$\mathfrak{L}_{\text{high},\neg\text{final}}(T) \coloneqq \mathfrak{L}_{\text{high}}(T) \cap \mathfrak{L}_{\neg\text{final}}(T),$$
$$\mathfrak{L}_{\text{low},\neg\text{final}}(T) \coloneqq \mathfrak{L}_{\text{low}}(T) \cap \mathfrak{L}_{\neg\text{final}}(T).$$

Let $i^*_1 \coloneqq r^{*-1}_{\boldsymbol{\theta}}(1)$ denote a fixed best arm, so that $\theta_{i^*_1} = 1$.

We first record how pivot changes and acceptances delimit these epochs. In particular, no arms have been accepted during an epoch that is not final.

Lemma 12 (Comparison Counts at the Beginning of an Epoch)**.** *For every $\delta \in (0,1]$, under $\mathcal{G}^{(\delta)}_{conf,T}$, all pivots are distinct and $N^{(P^{(l)})}_{i,T_{l-1}+1} = 0$ for every epoch $l \in \mathfrak{L}(T)$ and $i \neq P^{(l)}$. Moreover, every epoch l that is not final ($l < L(T)$) has $\mathcal{A}_t = \{\}$ for all $t \in [\![T_{l-1}+1, T_l]\!]$, and ends with a pivot change to an arm whose lower confidence bound is greater than $1/2$.*

Proof. Under $\mathcal{G}^{(\delta)}_{\text{conf},T}$, a lower confidence bound greater than $1/2$ identifies an arm with parameter strictly larger than the current pivot's parameter. Once any arm has been accepted, Lemma 9 puts the pivot in the accepted set, and the algorithm never changes it again. Thus every pivot change strictly increases its parameter, the pivots are distinct, and any epoch containing a round with nonempty accepted set is final.

Before any arm has been accepted ($\mathcal{A}_t = \{\}$), only comparison counts with the current pivot are updated. An arm selected as the next pivot has neither been pivotal nor accepted before, so its comparison counts are still zero. This also applies when it is selected in the first acceptance round, because that round updates counts with the preceding pivotal set. □

Every played set contains at most $\min\{k, n-k\}$ suboptimal arms. Since $0 < \theta_i \leq 1$, each suboptimal arm contributes at most $1/k$ to instantaneous regret. Averaging over the played set therefore gives, for every $t \in [\![T]\!]$:

$$\mathsf{IstReg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(t) \leq \frac{\min\{k, n-k\}}{k}. \tag{23}$$

We use this bound during each burn-in interval. We will also use the elementary inequality:

$$\frac{n \min\{k, n-k\}}{k} \leq 2(n-k). \tag{24}$$

Indeed, if $n - k \geq k$, the left-hand side is $n \leq 2(n-k)$; otherwise, it is $n(n-k)/k \leq 2(n-k)$.

We now bound the instantaneous pseudo-regret in terms of the gaps of the arms that remain unknown.

Lemma 13 (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 $\mathcal{G}_{conf,T}^{(\delta)}$:*

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

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

Proof. If $\mathcal{U}_t' = \{\}$, Lemmas 9 and 10 imply that the played set is $S_{\boldsymbol{\theta}}^{(k)*}$, so the result is immediate. Suppose $\mathcal{U}_t' \neq \{\}$.

If $P_t \in S_{\boldsymbol{\theta}}^{(k)*}$, every arm in $\mathcal{P}_t$ is optimal, and the number of uniformly sampled arms is $|\mathcal{U}_t' \cap S_{\boldsymbol{\theta}}^{(k)*}|$. Lemma 26 therefore gives the exact identity:

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

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 $P_t \notin S_{\boldsymbol{\theta}}^{(k)*}$. Lemma 9.1 implies $\mathcal{A}_t = \{\}$, so $\mathcal{P}_t = \{P_t\}$, $\mathcal{U}_t' \cap S_{\boldsymbol{\theta}}^{(k)*} = S_{\boldsymbol{\theta}}^{(k)*}$, and $k-1$ arms are sampled uniformly from $\mathcal{U}_t'$. Direct expansion gives:

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

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

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

This proves the result in the second case as well. □

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 13 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:

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

and the following predictable process:

$$\tilde{Q}_{s:t} \coloneqq \begin{cases} \sum_{u=s}^{t} \tilde{q}_u & \text{if } s \leq t \\ 0 & \text{if } s > t \end{cases}, \qquad \tilde{q}_t \coloneqq \begin{cases} \dfrac{k - |\mathcal{P}_t|}{|\mathcal{U}_t'|} & \text{if } \mathcal{U}_t' \neq \{\} \\ 0 & \text{if } \mathcal{U}_t' = \{\} \end{cases} \quad \text{with } s,t \in [\![T]\!],$$

which records the sum of conditional sampling probabilities over rounds s through t. Observe that $\tilde{q}_t$ is the probability of pulling an arm in $\mathcal{U}_t'$ as remarked in the proof of Lemma 13.

We first lower bound the predictable comparison count $\tilde{N}_{a,t+1}^{(p)}$ in terms of $\tilde{Q}_{T_{\text{start}}:t}$, then use the burn-in condition to obtain a bound on the observed count $N_{a,t}^{(p)}$.

Lemma 14 (Comparison Counts on a Fixed-Pivot Interval)**.** *Let $\nu \in \mathcal{E}_n^{(k,m)}$, $1 \leq T_{start} \leq t \leq T$, $\delta \in (0,1]$, and $\lambda > 0$. Suppose $p \in \mathcal{P}_u$ throughout $u \in [\![T_{start}, t]\!]$. For every $a \in \mathcal{U}_t' \setminus \{p\}$, we have:*

$$\tilde{N}_{a,t+1}^{(p)} \geq \frac{m}{4k}(\theta_a + \theta_p)\tilde{Q}_{T_{start}:t}. \tag{25}$$

If $\tilde{q}_u \geq 1/n$ throughout the interval and $t - T_{start} + 1 \geq t_n^{(k,m,\delta,\lambda)}$, then every $a \in \mathcal{U}'_{t-1} \setminus \{p\}$ with $\theta_a + \theta_p \geq 1/2$ satisfies, under $\mathcal{G}_{pulls}^{(\delta,\lambda)}$:

$$N_{a,t}^{(p)} > \frac{1 - e^{-\lambda}}{32\lambda} \frac{m}{k} \tilde{Q}_{T_{start}:t}. \tag{26}$$

Proof. For $a \in \mathcal{U}'_t \setminus \{p\}$, Equation (19), Lemma 5, and uniform sampling give:

$$\tilde{N}_{a,t+1}^{(p)} \geq \sum_{u=T_{\text{start}}}^{t} \tilde{O}_{a,u}^{(p)} \geq \frac{m}{4k}(\theta_a + \theta_p) \sum_{u=T_{\text{start}}}^{t} \tilde{q}_u.$$

This proves Equation (25). For the second claim, $\tilde{q}_u \geq 1/n$ and the burn-in condition give:

$$\tilde{Q}_{T_{\text{start}}:t} = \sum_{u=T_{\text{start}}}^{t} \tilde{q}_u \geq \frac{1}{n}(t - T_{\text{start}} + 1) \geq \frac{1}{n} t_n^{(k,m,\delta,\lambda)} \tag{27}$$

$$= \frac{64}{1 - e^{-\lambda}} \frac{k}{m} \ln\left(\frac{2n}{\delta}\right) \geq 2 \geq 2\tilde{q}_t, \tag{28}$$

where Step (28) 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 $a \in \mathcal{U}'_{t-1}$:

$$\tilde{N}_{a,t}^{(p)} = \sum_{u=1}^{t-1} \tilde{O}_{a,u}^{(p)} \geq \sum_{u=T_{\text{start}}}^{t-1} \tilde{O}_{a,u}^{(p)} \geq \frac{m}{4k}(\theta_a + \theta_p) \sum_{u=T_{\text{start}}}^{t-1} \mathbb{P}_{\text{TOP-}k\text{-UIF},\nu}\left[a \in S_u \mid \overline{H}_{0:u-1}\right] \tag{29}$$

$$\geq \frac{m}{4k}(\theta_a + \theta_p) \sum_{u=T_{\text{start}}}^{t-1} \tilde{q}_u = \frac{m}{4k}(\theta_a + \theta_p) \tilde{Q}_{T_{\text{start}}:t-1} \geq \frac{m}{8k}(\theta_a + \theta_p) \tilde{Q}_{T_{\text{start}}:t}, \tag{30}$$

where Step (29) follows from Equation (19) and Lemma 5. Furthermore, by combining Equations (27) and (30):

$$\frac{2}{\lambda} \ln\left(\frac{2n}{\delta}\right) = \frac{1 - e^{-\lambda}}{32\lambda} \frac{m}{k} \frac{1}{n} t_n^{(k,m,\delta,\lambda)} \leq \frac{1 - e^{-\lambda}}{4\lambda} \frac{m}{8k} \tilde{Q}_{T_{\text{start}}:t} \leq \frac{1}{4(\theta_a + \theta_p)} \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{a,t}^{(p)}. \tag{31}$$

Since $\theta_a + \theta_p \geq 1/2$, the lower count bound in $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ now yields:

$$N_{a,t}^{(p)} > \frac{1 - e^{-\lambda}}{\lambda} \tilde{N}_{a,t}^{(p)} - \frac{2}{\lambda} \ln(2n/\delta)$$

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

This proves Equation (26). □

Using these predictable counts, the next lemma bounds the gaps in the regret decomposition once enough rounds have elapsed to satisfy the burn-in condition.

Lemma 15 (Upper Bound on Gaps via Confidence Radii)**.** *Let $\nu \in \mathcal{E}_n^{(k,m)}$, $T \in \mathbb{N}_{\geq 1}$, $t \in [\![T]\!]$, $\delta \in (0, 1]$, $\lambda > 0$. If there exist $T_{start} \in [\![T]\!]$ and $p \in [\![n]\!]$ such that $t - T_{start} + 1 \geq t_n^{(k,m,\delta,\lambda)}$ and $p \in \mathcal{P}_u$ for all $u \in [\![T_{start}, t]\!]$, then, under both good events $\mathcal{G}_{conf,T}^{(\delta)}$ and $\mathcal{G}_{pulls}^{(\delta,\lambda)}$:*

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

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

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

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

Proof. For every $u \in [\![T_{\text{start}}, t]\!]$, Corollary 11 gives $\mathcal{U}'_u \cap S^{(k)*}_{\boldsymbol{\theta}} \neq \{\}$ in Lemma 15.1, since $j \in \mathcal{U}'_t \setminus S^{(k)*}_{\boldsymbol{\theta}} \subseteq \mathcal{U}_u \setminus S^{(k)*}_{\boldsymbol{\theta}}$. In Lemma 15.2, the same conclusion follows because i remains unknown and $i \neq p = P_u$. Lemma 9 therefore gives $k - |\mathcal{P}_u| \geq 1$. Since $|\mathcal{U}'_u| \leq n$, it follows that $\tilde{q}_u \geq 1/n$. Lemma 14 therefore applies.

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

$$\text{LCB}^{(p)}_{i,t} \leq \text{UCB}^{(p)}_{f,t}, \tag{32}$$

otherwise we would have $A_{i,t} \geq |\mathcal{U}_{t-1} \setminus S^{(k)*}_{\boldsymbol{\theta}}| = |\mathcal{A}_{t-1} \cup \mathcal{U}_{t-1}| - k$ by Lemma 9, implying $i \in \mathcal{A}_t$. Analogously, there must exist $g \in \mathcal{U}_{t-1} \cap S^{(k)*}_{\boldsymbol{\theta}}$ such that:

$$\text{UCB}^{(p)}_{j,t} \geq \text{LCB}^{(p)}_{g,t}. \tag{33}$$

Thus:

$$\begin{aligned}
\theta_i - \theta_j &\leq \frac{4}{\theta_p}(\mu^{(p)}_i - \mu^{(p)}_j) \leq 8(\text{UCB}^{(p)}_{i,t} - \text{LCB}^{(p)}_{j,t}) && (34)\\
&= 8(\text{UCB}^{(p)}_{i,t} - \text{LCB}^{(p)}_{i,t} + \text{UCB}^{(p)}_{j,t} - \text{LCB}^{(p)}_{j,t} + \text{LCB}^{(p)}_{i,t} - \text{UCB}^{(p)}_{j,t})\\
&\leq 8(2\beta_{\delta,n,T}(N^{(p)}_{i,t}) + 2\beta_{\delta,n,T}(N^{(p)}_{j,t}) + \text{UCB}^{(p)}_{f,t} - \text{LCB}^{(p)}_{g,t}) && (35)\\
&\leq 16(\beta_{\delta,n,T}(N^{(p)}_{i,t}) + \beta_{\delta,n,T}(N^{(p)}_{j,t}) + \beta_{\delta,n,T}(N^{(p)}_{f,t}) + \beta_{\delta,n,T}(N^{(p)}_{g,t}))\\
&\qquad + 8(\mu^{(p)}_f - \mu^{(p)}_g) && (36)\\
&< 16(\beta_{\delta,n,T}(N^{(p)}_{i,t}) + \beta_{\delta,n,T}(N^{(p)}_{j,t}) + \beta_{\delta,n,T}(N^{(p)}_{f,t}) + \beta_{\delta,n,T}(N^{(p)}_{g,t})), && (37)
\end{aligned}$$

where Step (34) follows from Lemma 5, Step (35) follows from Equations (32) and (33), Step (36) follows from the good event $\mathcal{G}^{(\delta)}_{\text{conf},T}$, and Step (37) follows from $\theta_f < \theta_g$, since $f \notin S^{(k)*}_{\boldsymbol{\theta}}$ and $g \in S^{(k)*}_{\boldsymbol{\theta}}$. If $a = p$, then $N^{(p)}_{a,t} = +\infty$ and the corresponding confidence bonus is zero. For every $a \in \{i, j, f, g\} \setminus \{p\}$, Lemma 14, using $\theta_p \geq 1/2$, implies that Equation (26) holds. 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:

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

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 $\mathcal{G}^{(\delta)}_{\text{conf},T}$, Lemma 5 gives:

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

Since either $\theta_p \geq 1/2$ or $\theta_i \geq 1/2$, Lemma 14 gives:

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

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

□

Now, we combine Lemmas 13 and 15 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 $i \in [\![n]\!]$, define the auxiliary last-active time:

$$T_{\mathcal{U},i} \coloneqq \max\{t \in [\![T]\!] \text{ s.t. } i \in \mathcal{U}_t\} \in [\![T]\!].$$

This last-active time lets us sum the preceding gap bounds only over rounds in which each arm remains unknown, yielding regret bounds for an entire epoch.

The following lemma sums confidence-radius bounds on the gaps over an interval. We keep the multiplier b in the gap bound $b\beta_{\delta,n,T}(\tilde{Q}_{s:u})$ explicit so that the same calculation can be used for each type of epoch.

Lemma 16 (Regret from Confidence-Radius Gap Bounds)**.** *Let $\nu \in \mathcal{E}_n^{(k,m)}$, $1 \le s \le t \le T$, $\delta \in (0,1]$, $\tau > 0$, and $b > 0$. Suppose $P_u = P_s$ throughout $u \in [\![s,t]\!]$ and $\mathcal{G}_{conf,T}^{(\delta)}$ holds. For every $u \in [\![s,t]\!]$ with $u - s + 1 \ge \tau$, suppose that, for every $i \in \mathcal{U}'_u \cap S_{\boldsymbol{\theta}}^{(k)*}$ and $j \in \mathcal{U}_u \setminus S_{\boldsymbol{\theta}}^{(k)*}$, the gaps in Lemma 13 satisfy:*

$$\theta_i - \theta_{P_u} \le b\beta_{\delta,n,T}(\tilde{Q}_{s:u}) \quad \text{if } \theta_j \ge \theta_{P_u},$$
$$\theta_i - \theta_j \le b\beta_{\delta,n,T}(\tilde{Q}_{s:u}) \quad \text{if } \theta_j < \theta_{P_u}.$$

Then:

$$\mathsf{Reg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(s,t) \le \frac{\min\{k, n-k\}}{k}\tau + \frac{4b}{k}\sqrt{\ln(2nT/\delta)} \sum_{\substack{j \in [\![n]\!] \\ j \notin S_{\boldsymbol{\theta}}^{(k)*}}} \mathbb{I}[\min\{t, T_{\mathcal{U},j}\} - s + 1 \ge \tau]\sqrt{\tilde{Q}_{s:\min\{t,T_{\mathcal{U},j}\}}}. \tag{38}$$

Proof. For $u - s + 1 \ge \tau$, the assumed gap bounds and Lemma 13 give, whenever $\mathcal{U}'_u \ne \{\}$:

$$\mathsf{lstReg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(u) \le \frac{b}{k}\beta_{\delta,n,T}(\tilde{Q}_{s:u}) \sum_{j \in \mathcal{U}_u \setminus S_{\boldsymbol{\theta}}^{(k)*}} \frac{|\mathcal{U}'_u \cap S_{\boldsymbol{\theta}}^{(k)*}|}{|\mathcal{U}'_u|}. \tag{39}$$

As in the proof of Lemma 13,

$$|\mathcal{U}'_u \cap S_{\boldsymbol{\theta}}^{(k)*}| = k - |\mathcal{A}'_u| - \mathbb{I}[P_s \in S_{\boldsymbol{\theta}}^{(k)*}]. \tag{40}$$

For every nonzero summand, Lemma 10 gives $|\mathcal{A}_u| < k$. If $\mathcal{A}_u = \{\}$, then $\mathcal{A}'_u = \{\}$ and $|\mathcal{P}_u| = 1$, so:

$$k - |\mathcal{A}'_u| - \mathbb{I}[P_s \in S_{\boldsymbol{\theta}}^{(k)*}] \le k \le 2(k-1) = 2(k - |\mathcal{P}_u|).$$

If $\mathcal{A}_u \ne \{\}$, Lemma 9 gives $P_s = P_u \in \mathcal{A}_u \subseteq S_{\boldsymbol{\theta}}^{(k)*}$, whence $\mathcal{P}_u = \mathcal{A}_u$ and:

$$k - |\mathcal{A}'_u| - \mathbb{I}[P_s \in S_{\boldsymbol{\theta}}^{(k)*}] = k - |\mathcal{A}_u| \le 2(k - |\mathcal{P}_u|).$$

Substitution into Equation (39) yields:

$$\mathsf{lstReg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(u) \le \frac{2b}{k}\tilde{q}_u\beta_{\delta,n,T}(\tilde{Q}_{s:u})|\mathcal{U}_u \setminus S_{\boldsymbol{\theta}}^{(k)*}|. \tag{41}$$

The same bound holds with value zero if $\mathcal{U}'_u = \{\}$. We bound the burn-in regret by Equation (23), interchange the remaining sums, and use the fact that unknown sets only shrink:

$$\mathsf{Reg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(s,t) \le \frac{\min\{k, n-k\}}{k}\tau + \frac{2b}{k} \sum_{\substack{j \in [\![n]\!] \\ j \notin S_{\boldsymbol{\theta}}^{(k)*}}} \mathbb{I}[\min\{t, T_{\mathcal{U},j}\} - s + 1 \ge \tau] \sum_{u=s}^{\min\{t,T_{\mathcal{U},j}\}} \tilde{q}_u\beta_{\delta,n,T}(\tilde{Q}_{s:u}). \tag{42}$$

We take $\tilde{q}_u\beta_{\delta,n,T}(\tilde{Q}_{s:u}) = 0$ when $\tilde{q}_u = 0$. Finally, (Orabona, 2019, Lemma 4.13) gives, for every $v \in [\![s,t]\!]$,

$$\sum_{u=s}^{v} \tilde{q}_u\beta_{\delta,n,T}(\tilde{Q}_{s:u}) \le 2\sqrt{\ln(2nT/\delta)}\sqrt{\tilde{Q}_{s:v}}.$$

Substitution proves Equation (38). □

Lemma 17 (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 $l \in \mathfrak{L}(T)$, under both good events $\mathcal{G}_{conf,T}^{(\delta)}$ and $\mathcal{G}_{pulls}^{(\delta,\lambda)}$:*

1. *If $l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T)$, the following instance-dependent and worst-case bounds hold. We use the gap notation $\Delta_{k,r^*_{\boldsymbol{\theta}}(j)} = \theta_{r^{*-1}_{\boldsymbol{\theta}}(k)} - \theta_j$ of Theorem 2, and write $\Delta_{1,r^*_{\boldsymbol{\theta}}(P^{(l)})} = 1 - \theta_{P^{(l)}}$ for the gap from the fixed best arm to the pivot:*

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1,T_l) &\leq \frac{128}{1-e^{-\lambda}}\frac{k(n-k)}{m}\ln(2nT/\delta)\\ &+ \frac{2^{19}\lambda}{1-e^{-\lambda}}\frac{\ln(2nT/\delta)}{m}\sum_{j\in\mathcal{U}_{T_{l-1}+1}\setminus S^{(k)*}_{\boldsymbol{\theta}}}\min\left\{\frac{1}{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}},\frac{1}{\Delta_{1,r^*_{\boldsymbol{\theta}}(P^{(l)})}}\right\},\end{aligned} \tag{43}$$

and:

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1,T_l) &\leq \frac{128}{1-e^{-\lambda}}\frac{k(n-k)}{m}\ln\left(\frac{2n}{\delta}\right)\\ &+ \frac{2^8}{\sqrt{km}}\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\ln\left(\frac{2nT}{\delta}\right)}\sum_{\substack{j\in\mathcal{U}_{T_{l-1}+1}\setminus S^{(k)*}_{\boldsymbol{\theta}}\\ \min\{T_l,T_{\mathcal{U},j}\}-T_{l-1}\geq t_n^{(k,m,\delta,\lambda)}}}\sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}.\end{aligned} \tag{44}$$

2. *If $l \in \mathfrak{L}_{\mathrm{low},\neg\mathrm{final}}(T)$, then:*

$$\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1,T_l) \leq \frac{2^7k+2^{15}\lambda}{1-e^{-\lambda}}\frac{n-k}{m}\ln\left(\frac{2nT}{\delta}\right). \tag{45}$$

3. *If $l \in \mathfrak{L}_{\mathrm{final}}(T)$ (so $T_l = T$):*

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1,T_l) &\leq 2\frac{\min\{k,n-k\}}{k}t_n^{(k,m,\delta,\lambda)}\\ &+ \frac{2^{20}\lambda}{1-e^{-\lambda}}\ln\left(\frac{2nT}{\delta}\right)\frac{1}{m}\sum_{j\in[\![n]\!]\setminus S^{(k)*}_{\boldsymbol{\theta}}}\frac{1}{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}},\end{aligned}$$

and:

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

Proof. We use Equation (23) to bound regret during each burn-in interval, while Lemmas 13 and 15 bound the regret after the burn-in period.

Case 1: High epochs that are not final. We first derive the high-pivot bounds using the burn-in condition $t - T_{l-1} \geq t_n^{(k,m,\delta,\lambda)}$. These calculations apply whenever $\theta_{P^{(l)}} \geq 1/2$, including in a final epoch. We then use the fact that the epoch is not final to obtain the arm/pivot-gap minimum in Part 1. Lemma 16, with $s = T_{l-1}+1$, $t = T_l$, $\tau = t_n^{(k,m,\delta,\lambda)}$, and $b = 64\sqrt{32\lambda/(1-e^{-\lambda})}\sqrt{k/m}$, gives:

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

$$\sum_{j\in[\![n]\!]\setminus S_{\boldsymbol{\theta}}^{(k)*}} \mathbb{I}[\min\{T_l, T_{\mathcal{U},j}\} \geq T_{l-1} + t_n^{(k,m,\delta,\lambda)}] \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}}. \tag{46}$$

We first derive the instance-dependent bound. For every $j \notin S_{\boldsymbol{\theta}}^{(k)*}$ whose indicator in Equation (46) is one, we have $j \in \mathcal{U}_{\min\{T_l,T_{\mathcal{U},j}\}}$, and Corollary 11 provides an arm $i_j \in \mathcal{U}'_{\min\{T_l,T_{\mathcal{U},j}\}} \cap S_{\boldsymbol{\theta}}^{(k)*}$. If $\theta_j \geq \theta_{P^{(l)}}$, Lemma 15.2 bounds $\theta_{i_j} - \theta_{P^{(l)}}$; if $\theta_j < \theta_{P^{(l)}}$, then $j \in \mathcal{U}'_{\min\{T_l,T_{\mathcal{U},j}\}}$ and Lemma 15.1 bounds $\theta_{i_j} - \theta_j$. In both cases:

$$\theta_{r_{\boldsymbol{\theta}}^{*-1}(k)} - \theta_j \leq 64\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\frac{k}{m}}\beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}). \tag{47}$$

Therefore:

$$\sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}} \leq 64\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\frac{k}{m}}\frac{1}{\theta_{r_{\boldsymbol{\theta}}^{*-1}(k)} - \theta_j}. \tag{48}$$

Substitution into Equation (46) gives:

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1, T_l) \leq{}& \frac{\min\{k, n-k\}}{k} t_n^{(k,m,\delta,\lambda)} \\ &+ \frac{2^{19}\lambda}{1-e^{-\lambda}} \ln\left(\frac{2nT}{\delta}\right)\frac{1}{m} \sum_{j\in[\![n]\!]\setminus S_{\boldsymbol{\theta}}^{(k)*}} \frac{1}{\theta_{r_{\boldsymbol{\theta}}^{*-1}(k)} - \theta_j}.\end{aligned} \tag{49}$$

For the worst-case bound, we return to Equation (46). Expanding $\tilde{Q}$, swapping sums, and using $|\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}| \leq |\mathcal{U}'_t|$ gives:

$$\sum_{j\notin S_{\boldsymbol{\theta}}^{(k)*}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}} = \sum_{t=T_{l-1}+1}^{T_l} \tilde{q}_t |\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}| \leq k(T_l - T_{l-1}). \tag{50}$$

If the pivot is optimal, every unknown suboptimal arm belongs to $\mathcal{U}'_t$, so $|\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}| \leq |\mathcal{U}'_t|$. If the pivot is suboptimal, Lemma 9 gives $\mathcal{A}_t = \{\}$ and all k optimal arms are unknown. Thus $\mathcal{U}'_t = \mathcal{U}_t \setminus \{P_t\}$ and:

$$|\mathcal{U}_t \setminus S_{\boldsymbol{\theta}}^{(k)*}| = |\mathcal{U}_t| - k \leq |\mathcal{U}_t| - 1 = |\mathcal{U}'_t|.$$

Cauchy–Schwarz and $\tilde{q}_t \leq 1$ therefore give:

$$\sum_{j\notin S_{\boldsymbol{\theta}}^{(k)*}} \sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}} \leq \min\left\{\sqrt{k(n-k)}, n-k\right\}\sqrt{T_l - T_{l-1}}. \tag{51}$$

Substituting Equation (51) into Equation (46), after dropping the indicators, gives:

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1, T_l) \leq{}& \frac{\min\{k, n-k\}}{k} t_n^{(k,m,\delta,\lambda)} \\ &+ 2^8\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\frac{1}{\sqrt{m}}\min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\}\sqrt{(T_l - T_{l-1})\ln\left(\frac{2nT}{\delta}\right)}.\end{aligned} \tag{52}$$

We now use the fact that the epoch is not final. The epoch ends with a pivot change to a strictly better arm, so $\theta_{P^{(l)}} < 1$. Lemma 12 also gives $\mathcal{A}_t = \{\}$ and $i_1^* \in \mathcal{U}_t$ for every $T_{l-1} + 1 \leq t \leq T_l$.
The burn-in contribution is bounded, using Equations (23) and (24), by:

$$t_n^{(k,m,\delta,\lambda)}\frac{\min\{k, n-k\}}{k}$$

$$= \frac{64}{1-e^{-\lambda}} \frac{n \min\{k, n-k\}}{m} \ln\left(\frac{2n}{\delta}\right) \leq \frac{128}{1-e^{-\lambda}} \frac{k(n-k)}{m} \ln\left(\frac{2n}{\delta}\right). \quad (53)$$

Since the unknown sets only shrink, the sum in Equation (46) can be restricted to $j \in \mathcal{U}_{T_{l-1}+1} \setminus S_{\boldsymbol{\theta}}^{(k)*}$. Every nonzero term satisfies:

$$\min\{T_l, T_{\mathcal{U},j}\} - T_{l-1} \geq t_n^{(k,m,\delta,\lambda)}. \quad (54)$$

Applying Lemma 15.2 to the fixed best arm at round T_l, whose burn-in condition follows from Equation (54), gives:

$$\sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}} \leq \sqrt{\frac{\tilde{Q}_{T_{l-1}+1:T_l}}{\ln(2nT/\delta)}} \leq 8\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\frac{k}{m}}\frac{1}{1-\theta_{P^{(l)}}}. \quad (55)$$

Both estimates bound the same sampling term: Equation (48) uses the gap of arm j, while Equation (55) uses the gap between the best arm and the pivot. Taking their minimum gives:

$$\sqrt{\frac{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}{\ln(2nT/\delta)}} \leq 64\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\frac{k}{m}} \min\left\{\frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}, \frac{1}{1-\theta_{P^{(l)}}}\right\}. \quad (56)$$

Substituting Equation (56) into Equation (46), dropping the indicators, and using Equation (53) proves Equation (43). Restricting Equation (46) to its nonzero terms and using the same burn-in bound proves Equation (44).

Case 2: Low epochs that are not final. By Lemma 12, $\mathcal{A}_t = \{\}$ throughout the epoch, so the fixed best arm satisfies $i_1^* \in \mathcal{U}_t$ for every $T_{l-1} + 1 \leq t \leq T_l$.
Since $\theta_{P^{(l)}} < 1/2$ 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 13 is strictly bounded by 1. We can thus uniformly upper bound every gap in Lemma 13 by $2(\theta_{i_1^*} - \theta_{P^{(l)}})$.

Because $i_1^* \in \mathcal{U}_t' \cap S_{\boldsymbol{\theta}}^{(k)*}$, we can apply Lemma 15.2 directly to i_1^* and the pivot $P^{(l)}$, under the burn-in condition $t - T_{l-1} \geq t_n^{(k,m,\delta,\lambda)}$. Multiplying the result by 2, we obtain:

$$2(\theta_{i_1^*} - \theta_{P^{(l)}}) \leq 16\sqrt{\frac{32\lambda}{1-e^{-\lambda}}}\sqrt{\frac{k}{m}}\beta_{\delta,n,T}(\tilde{Q}_{T_{l-1}+1:t}). \quad (57)$$

By Lemma 16, using Equation (57) with $b = 16\sqrt{32\lambda/(1-e^{-\lambda})}\sqrt{k/m}$, and inverting the best-arm gap bound as in Equation (55), we obtain the following upper bound on the contribution after burn-in:

$$\frac{2^{14}\lambda}{1-e^{-\lambda}} \frac{n-k}{m(1-\theta_{P^{(l)}})} \ln\left(\frac{2nT}{\delta}\right) \leq \frac{2^{15}\lambda}{1-e^{-\lambda}} \frac{n-k}{m} \ln\left(\frac{2nT}{\delta}\right). \quad (58)$$

Here $1 - \theta_{P^{(l)}} > 1/2$. Adding the burn-in bound in Equation (53) proves Equation (45).

Case 3: Final epochs. Here $l = L(T)$ and $T_l = T$. If $\theta_{P^{(l)}} \geq 1/2$, Equations (49) and (52) imply Part 3. Suppose $\theta_{P^{(l)}} < 1/2$. We split the epoch into two phases at:

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

Thus, the first phase is empty if $T_{\text{best}} = T_{l-1}$, while otherwise it consists of rounds $T_{l-1}+1 \leq t \leq T_{\text{best}}$, during which $i_1^* \in \mathcal{U}_t$. It includes the round whose end accepts i_1^*, if acceptance occurs within the epoch. Other arms may already have been accepted during this phase.

We therefore use Lemma 15.2 under the condition $t - T_{l-1} \geq t_n^{(k,m,\delta,\lambda)}$. Applying Lemma 16 with the same coefficient as in the low-epoch case and upper endpoint T_{best} gives:

$$\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1, T_{\text{best}}) \le \frac{\min\{k, n-k\}}{k} t_n^{(k,m,\delta,\lambda)} + \frac{2^{15}\lambda}{1-e^{-\lambda}} \frac{n-k}{m} \ln\left(\frac{2nT}{\delta}\right). \tag{59}$$

For the worst-case bound, the predictable-count argument in Equations (51) and (50), with T_{best} as the upper endpoint, gives:

$$\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T_{l-1}+1, T_{\text{best}}) \le \frac{\min\{k, n-k\}}{k} t_n^{(k,m,\delta,\lambda)} + 2^6 \sqrt{\frac{32\lambda}{1-e^{-\lambda}}} \frac{1}{\sqrt{m}} \min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\} \sqrt{(T_{\text{best}} - T_{l-1}) \ln\left(\frac{2nT}{\delta}\right)}. \tag{60}$$

If $T_{\text{best}} + 1 \le T_l$, then i_1^* has been accepted and belongs to $\mathcal{A}_t \subseteq \mathcal{P}_t$ for every round t with $T_{\text{best}} + 1 \le t \le T_l$. Since the pivot has not changed within the epoch, Lemma 9 also gives $P^{(l)} \in \mathcal{A}_t \subseteq S_{\boldsymbol{\theta}}^{(k)*}$. Therefore, every suboptimal arm j satisfies $\theta_j < \theta_{P^{(l)}}$, and the first inner sum in Lemma 13 is empty. Because $\theta_{i_1^*} = 1 \ge 1/2$, Lemma 15.1, with $p = i_1^*$ and $T_{\text{start}} = T_{\text{best}} + 1$, bounds every remaining gap $\theta_i - \theta_j$ when $t - T_{\text{best}} \ge t_n^{(k,m,\delta,\lambda)}$. Applying Lemma 16 with $b = 64\sqrt{32\lambda/(1-e^{-\lambda})}\sqrt{k/m}$ on $[\![T_{\text{best}}+1, T_l]\!]$, with $\tilde{Q}_{T_{\text{best}}+1:t}$ in place of $\tilde{Q}_{T_{l-1}+1:t}$, gives the instance-dependent and worst-case bounds of Equations (49) and (52), with burn-in contribution at most $\frac{\min\{k,n-k\}}{k} t_n^{(k,m,\delta,\lambda)}$ and phase length $T_l - T_{\text{best}}$. If $T_{\text{best}} + 1 > T_l$, the second phase is empty.

The two phases contribute at most $2\frac{\min\{k,n-k\}}{k} t_n^{(k,m,\delta,\lambda)}$ in burn-in regret. Since every positive top-k gap is at most one, the remaining terms satisfy:

$$2^{15}(n-k) + 2^{19} \sum_{j \notin S_{\boldsymbol{\theta}}^{(k)*}} \frac{1}{\theta_{r_{\boldsymbol{\theta}}^{*-1}(k)} - \theta_j} \le 2^{20} \sum_{j \notin S_{\boldsymbol{\theta}}^{(k)*}} \frac{1}{\theta_{r_{\boldsymbol{\theta}}^{*-1}(k)} - \theta_j},$$
$$2^6\sqrt{T_{\text{best}} - T_{l-1}} + 2^8\sqrt{T_l - T_{\text{best}}} \le 2^9\sqrt{T_l - T_{l-1}}. \tag{61}$$

Combining the phase bounds with Equation (61) proves Part 3.

□

D.4 From Epoch Bounds to Regret over the Entire Horizon

We now pass from regret bounds for a single epoch to bounds over the entire horizon. The main difficulty is that changing the pivot starts new comparison counts $N_{i,t}^{(P^{(l)})}$, so an arm that remains unknown can contribute to the regret in several epochs.

Low epochs that are not final have only logarithmic regret in T/δ by Lemma 17.2. Since there are at most n epochs, their total regret is absorbed into the additive term.

For high epochs that are not final, our conservative confidence bounds prevent a pivot change with too few comparisons. Together with observation ordering, this ensures that the gap between the best arm and the pivot contracts by a constant factor at each change. For the instance-dependent bound, Lemma 17.1, Equation (43), bounds the term associated with an unknown suboptimal arm j in epoch l by:

$$\tilde{\mathcal{O}}\left(\frac{1+\log(1/\delta)}{m} \min\left\{\frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}, \frac{1}{1-\theta_{P^{(l)}}}\right\}\right).$$

The reciprocal pivot-gap terms can then be summed geometrically until they reach the reciprocal of the arm's gap. Beyond that point, these comparison-count lower bounds ensure that the arm is rejected before it can contribute to the regret in many more epochs. Together, these facts bound the total contribution of j to the regret bound over these epochs by:

$$\tilde{\mathcal{O}}\left(\frac{1+\log(1/\delta)}{m} \frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}\right),$$

as shown in Lemma 22 after multiplication by the common factor in the epoch bound.

For the worst-case bound, Lemma 17.1, Equation (44), instead gives a term for j of order:

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{1+\log(1/\delta)}{km}}\sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}\right).$$

The comparisons needed for a pivot change force $\tilde{Q}_{T_{l-1}+1:T_l} = \sum_{t=T_{l-1}+1}^{T_l} \tilde{q}_t$, where $\tilde{q}_t = (k-1)/(|\mathcal{U}_t|-1)$, to grow geometrically across sufficiently separated epochs after burn-in and once $1-\theta_{P^{(l)}} < 1/8$. Lemma 23 sums the resulting square-root terms, also accounting for the earlier epochs and the removal of j.

Finally, we combine these results with the bounds for low epochs that are not final and for the final epoch in Lemma 17 to prove Theorem 2.

To quantify the improvement at a pivot change, we compare $\mu^{(P^{(l)})}_{P^{(l+1)}} - 1/2$ with $\mu^{(P^{(l)})}_{i_1^*} - 1/2$, the differences between the respective probabilities of beating the current pivot and $1/2$. The smaller the first difference, the more comparisons the next pivot needs to pass the confidence test. Observation ordering gives the best arm at least as many predictable comparisons. Together with count concentration and conservative confidence bounds, this forces the first difference to be at least a constant fraction of the second. We use the best arm's confidence bound before the triggering update and the next pivot's bound after it, since both arms may pass the test in the same round.

Lemma 18 (Improvement at a Pivot Change)**.** *Suppose $2 \le k < n$, $T \ge n$, $\delta \in (0,1]$, and both good events $\mathcal{G}^{(\delta)}_{conf,T}$ and $\mathcal{G}^{(\delta,\lambda)}_{pulls}$ (with $\lambda = 1$) hold. For every $l \in \mathfrak{L}_{\neg\mathrm{final}}(T)$:*

$$\mu^{(P^{(l)})}_{P^{(l+1)}} - \frac{1}{2} \ge \frac{1}{64}\left(\mu^{(P^{(l)})}_{i_1^*} - \frac{1}{2}\right).$$

Proof. If $P^{(l+1)} = i_1^*$, the claim is immediate. Suppose otherwise. Lemma 12 applies under $\mathcal{G}^{(\delta)}_{\mathrm{conf},T}$. The current pivot cannot be a best arm, as the pivot changes to a strictly better arm. During this epoch, i_1^* remains unknown by Lemma 9. The chosen arm $P^{(l+1)} \in \mathcal{U}_{T_l}$ was also unknown in every earlier round of the epoch, since unknown sets only shrink. By Lemma 12, no arms are accepted during this epoch, which is not final, and the comparison counts with $P^{(l)}$ start at zero.

Since $P^{(l)}$ was neither pivotal nor accepted before this epoch, Equation (19) gives zero initial predictable comparison counts as well. Summing Lemma 7 over the epoch therefore gives:

$$\tilde{N}^{(P^{(l)})}_{P^{(l+1)},T_l+1} \le \tilde{N}^{(P^{(l)})}_{i_1^*,T_l+1}.$$

Applying the upper and lower count bounds in $\mathcal{G}^{(\delta,\lambda)}_{\mathrm{pulls}}$ (with $\lambda = 1$) at $T_l + 1$ yields:

$$\begin{aligned}
N^{(P^{(l)})}_{P^{(l+1)},T_l+1} &< (e-1)\tilde{N}^{(P^{(l)})}_{P^{(l+1)},T_l+1} + 2\ln\left(\frac{2n}{\delta}\right) \\
&\le (e-1)\tilde{N}^{(P^{(l)})}_{i_1^*,T_l+1} + 2\ln\left(\frac{2n}{\delta}\right) \\
&< eN^{(P^{(l)})}_{i_1^*,T_l+1} + 2(e+1)\ln\left(\frac{2n}{\delta}\right) \\
&< 3N^{(P^{(l)})}_{i_1^*,T_l} + 10\ln\left(\frac{2nT}{\delta}\right).
\end{aligned}$$

The last inequality uses $e < 3$, the fact that a count increases by at most one per round, $\ln\left(\frac{2n}{\delta}\right) \le \ln\left(\frac{2nT}{\delta}\right)$, and $\ln\left(\frac{2nT}{\delta}\right) \ge \ln 18 > 3/2$.

Before the triggering round T_l, the best arm's lower confidence bound is at most $1/2$. Indeed, it is $-\infty$ at the beginning of the epoch because its comparison count is zero. At a later round, a lower confidence bound greater than $1/2$ at the preceding update would already have caused a pivot change.

We distinguish two cases according to the next pivot's post-update count.

Case 1: $N^{(P^{(l)})}_{P^{(l+1)},T_l+1} \geq 20\ln\left(\frac{2nT}{\delta}\right)$. The preceding count comparison gives:

$$N^{(P^{(l)})}_{i_1^*,T_l} > \frac{1}{6} N^{(P^{(l)})}_{P^{(l+1)},T_l+1}.$$

Hence, using the best arm's pre-update lower confidence bound and the event $\mathcal{G}^{(\delta)}_{\text{conf},T}$:

$$\begin{aligned}\mu^{(P^{(l)})}_{i_1^*} - \frac{1}{2} &\leq \left(1 + \frac{\sqrt{3}}{2}\right) \beta_{\delta,n,T}(N^{(P^{(l)})}_{i_1^*,T_l}) \\ &< 2\sqrt{6}\,\beta_{\delta,n,T}(N^{(P^{(l)})}_{P^{(l+1)},T_l+1}) < 5\beta_{\delta,n,T}(N^{(P^{(l)})}_{P^{(l+1)},T_l+1}).\end{aligned}$$

Case 2: $N^{(P^{(l)})}_{P^{(l+1)},T_l+1} < 20\ln\left(\frac{2nT}{\delta}\right)$. The next pivot's confidence radius is greater than $1/\sqrt{20}$. Since $\mu^{(P^{(l)})}_{i_1^*} - 1/2 \leq 1/2$, we again have:

$$\mu^{(P^{(l)})}_{i_1^*} - \frac{1}{2} < 5\beta_{\delta,n,T}(N^{(P^{(l)})}_{P^{(l+1)},T_l+1}).$$

The selected arm satisfies $\text{LCB}^{(P^{(l)})}_{P^{(l+1)},T_l+1} > 1/2$, so $\mathcal{G}^{(\delta)}_{\text{conf},T}$ gives:

$$\mu^{(P^{(l)})}_{P^{(l+1)}} - \frac{1}{2} > \frac{1}{8}\beta_{\delta,n,T}(N^{(P^{(l)})}_{P^{(l+1)},T_l+1}) \geq \frac{1}{40}\left(\mu^{(P^{(l)})}_{i_1^*} - \frac{1}{2}\right).$$

This proves the stated $1/64$ bound. □

Let us now use Lemma 18 to control the gaps of consecutive high epochs that are not final. We show that, when the pivot parameter is at least $1/2$, each pivot change reduces the difference between the PL parameter of a best arm and that of the current pivot, $1 - \theta_{P^{(l)}}$, by a constant factor.

Lemma 19 (Contraction for High Epochs That Are Not Final)**.** *Suppose $2 \leq k < n$, $T \geq n$, $\delta \in (0,1]$, and both good events $\mathcal{G}^{(\delta)}_{conf,T}$ and $\mathcal{G}^{(\delta,\lambda)}_{pulls}$ (with $\lambda = 1$) hold. At each epoch l that is not final and has $\theta_{P^{(l)}} \geq 1/2$, the next pivot satisfies:*

$$1 - \theta_{P^{(l+1)}} \leq \frac{95}{96}(1 - \theta_{P^{(l)}}).$$

Proof. If $P^{(l+1)} = i_1^*$, the contraction is immediate. Otherwise, Lemma 18 and the expression of the pairwise preferences give:

$$\begin{aligned}\theta_{P^{(l+1)}} - \theta_{P^{(l)}} &\geq \frac{1}{64}(1 - \theta_{P^{(l)}})\frac{\theta_{P^{(l+1)}} + \theta_{P^{(l)}}}{1 + \theta_{P^{(l)}}} \\ &\geq \frac{1}{64}(1 - \theta_{P^{(l)}})\frac{2\theta_{P^{(l)}}}{1 + \theta_{P^{(l)}}} \geq \frac{1}{96}(1 - \theta_{P^{(l)}}).\end{aligned}$$

Here $\theta_{P^{(l+1)}} > \theta_{P^{(l)}}$ follows from $\mathcal{G}^{(\delta)}_{\text{conf},T}$, and $2\theta_{P^{(l)}}/(1 + \theta_{P^{(l)}}) \geq 2/3$ follows from $\theta_{P^{(l)}} \geq 1/2$. Rearranging proves the contraction.

□

An epoch $l \in \mathfrak{L}_{\neg\text{final}}(T)$ ends with a pivot change, so the next pivot $P^{(l+1)}$ must pass the confidence test $\text{LCB}^{(P^{(l)})}_{P^{(l+1)},T_l+1} > 1/2$ against the current pivot $P^{(l)}$.

Even the easiest candidate to certify, the best arm, needs enough comparisons. Its difference from $1/2$, $(1 - \theta_{P^{(l)}})/(2(1 + \theta_{P^{(l)}})) = \mu^{(P^{(l)})}_{i_1^*} - 1/2$, is the largest, so every prospective next pivot has a true margin no larger. Under $\mathcal{G}^{(\delta)}_{\text{conf},T}$, the estimation error is at most $\sqrt{3}/2$ times the confidence radius, whereas the test subtracts the full radius. To pass, the true margin must exceed the remaining fraction

$1-\sqrt{3}/2$ of that radius. Thus the radius must already be at most a constant times the best arm's margin. Since radii decrease as the inverse square root of the comparison count, any successful pivot change requires comparison counts growing with $1/(1-\theta_{P^{(l)}})^2$. The following lemma converts this test into lower bounds on the comparison counts $N_{a,T_l+1}^{(P^{(l)})}$ of arms $a \in \mathcal{U}_{T_l} \setminus \{P^{(l)}\}$ and on the sum of sampling probabilities $\tilde{Q}_{T_{l-1}+1:T_l}$ during the epoch.

Lemma 20 (Comparison Counts in Epochs That Are Not Final)**.** *Let $\nu \in \mathcal{E}_n^{(k,m)}$, $T \geq n$, $\delta \in (0,1]$, and $l \in \mathfrak{L}_{\neg\text{final}}(T)$. Suppose both good events $\mathcal{G}_{conf,T}^{(\delta)}$ and $\mathcal{G}_{pulls}^{(\delta,\lambda)}$ (with $\lambda = 1$) hold and $1-\theta_{P^{(l)}} < 1/8$. Then $\theta_{P^{(l)}} < 1$ and:*

$$(e-1)\tilde{Q}_{T_{l-1}+1:T_l} > \frac{\ln(2nT/\delta)}{32(1-\theta_{P^{(l)}})^2}, \tag{62}$$

and every $a \in \mathcal{U}_{T_l} \setminus \{P^{(l)}\}$ satisfies:

$$N_{a,T_l+1}^{(P^{(l)})} > \frac{m\ln(2nT/\delta)}{256ek(1-\theta_{P^{(l)}})^2} - 2\ln(2n/\delta). \tag{63}$$

Proof. We first work under $\mathcal{G}_{\text{conf},T}^{(\delta)}$ and $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ with an arbitrary $\lambda > 0$, and specialize to $\lambda = 1$ at the end. Lemma 12 gives $\theta_{P^{(l+1)}} > \theta_{P^{(l)}}$, hence $\theta_{P^{(l)}} < 1$. At round T_l, the event $\mathcal{G}_{\text{conf},T}^{(\delta)}$ and $\text{LCB}_{P^{(l+1)},T_l+1}^{(P^{(l)})} > 1/2$ give:

$$\begin{aligned}\frac{1-\theta_{P^{(l)}}}{2} &\geq \frac{1-\theta_{P^{(l)}}}{2(1+\theta_{P^{(l)}})} = \mu_{i_1^*}^{(P^{(l)})} - \frac{1}{2} \geq \mu_{P^{(l+1)}}^{(P^{(l)})} - \frac{1}{2} \\ &> \frac{1}{8}\beta_{\delta,n,T}(N_{P^{(l+1)},T_l+1}^{(P^{(l)})}) \geq \frac{1}{8}\sqrt{\frac{\ln(2nT/\delta)}{N_{P^{(l+1)},T_l+1}^{(P^{(l)})}}}.\end{aligned}$$

After rearranging, we obtain:

$$N_{P^{(l+1)},T_l+1}^{(P^{(l)})} > \frac{\ln(2nT/\delta)}{16(1-\theta_{P^{(l)}})^2}. \tag{64}$$

For $u \in [\![T_{l-1}+1, T_l]\!]$ and $a \in \mathcal{U}_u \setminus \{P^{(l)}\}$, Equation (19) and uniform sampling give:

$$\tilde{O}_{a,u}^{(P^{(l)})} \leq \mathbb{P}_{\text{Top-}k\text{-UIF},\nu}\big[a \in S_u \mid \overline{H}_{0:u-1}\big] = \tilde{q}_u. \tag{65}$$

Lemma 12 and Equation (19) give zero initial observed and predictable counts with $P^{(l)}$. Since $P^{(l+1)} \in \mathcal{U}_{T_l}$, summing Equation (65) yields:

$$\tilde{N}_{P^{(l+1)},T_l+1}^{(P^{(l)})} = \sum_{u=T_{l-1}+1}^{T_l} \tilde{O}_{P^{(l+1)},u}^{(P^{(l)})} \leq \sum_{u=T_{l-1}+1}^{T_l} \tilde{q}_u = \tilde{Q}_{T_{l-1}+1:T_l}.$$

Applying the upper count bound in $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ proves Equation (66):

$$\frac{e^\lambda - 1}{\lambda}\tilde{Q}_{T_{l-1}+1:T_l} > \frac{\ln(2nT/\delta)}{16(1-\theta_{P^{(l)}})^2} - \frac{2}{\lambda}\ln(2n/\delta). \tag{66}$$

If the epoch is high, Lemma 14, Equation (25), gives $\tilde{N}_{a,T_l+1}^{(P^{(l)})} \geq (m/(8k))\tilde{Q}_{T_{l-1}+1:T_l}$ for every $a \in \mathcal{U}_{T_l} \setminus \{P^{(l)}\}$. Thus a lower bound on sampling also controls every surviving arm's predictable comparison count. Applying the lower count bound in $\mathcal{G}_{\text{pulls}}^{(\delta,\lambda)}$ proves Equation (67):

$$N_{a,T_l+1}^{(P^{(l)})} > \frac{1-e^{-\lambda}}{\lambda}\frac{m}{8k}\tilde{Q}_{T_{l-1}+1:T_l} - \frac{2}{\lambda}\ln(2n/\delta). \tag{67}$$

When $\lambda = 1$ and $1-\theta_{P^{(l)}} < 1/8$, we have

$$2\ln(2n/\delta) \leq \frac{\ln(2nT/\delta)}{32(1-\theta_{P^{(l)}})^2}.$$

Substituting into Equation (66) proves Equation (62). Here $\theta_{P^{(l)}} > 7/8$, so Equation (67) applies. Combining the two bounds and using $(1-e^{-1})/(e-1) = 1/e$ proves Equation (63). □

For the instance-dependent bound, we use the explicit sum over suboptimal arms in Lemma 17.1. For an unknown suboptimal arm j, once the pivot gap falls below $\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}$, its term in this sum is capped at $1/\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}$. The contraction in Lemma 19 therefore does not by itself limit how often the capped term can recur.

We next show that, once the pivot gap is sufficiently small relative to the arm's gap, the arm is rejected before the next epoch. If j survives to round T_l, Lemma 20, Equation (63), gives lower comparison-count bounds growing with $m/[k(1-\theta_{P^{(l)}})^2]$ for j and the optimal arms. Counts on the inverse-square scale $1/\Delta^2_{k,r^*_{\boldsymbol{\theta}}(j)}$ suffice to certify that all k optimal arms are better than j, triggering its rejection. Comparing the two inverse-square scales explains the factor $\sqrt{m/k}$ in the condition below.

Lemma 21 (Rejection Before a Pivot Change with Small Pivot Gap)**.** *Suppose $2 \le k < n$, $T \ge n$, $\delta \in (0,1]$, and both good events $\mathcal{G}^{(\delta)}_{conf,T}$ and $\mathcal{G}^{(\delta,\lambda)}_{pulls}$ (with $\lambda = 1$) hold. Let l be a high epoch that is not final and $j \in \mathcal{U}_{T_{l-1}+1} \setminus S^{(k)*}_{\boldsymbol{\theta}}$. If*

$$1 - \theta_{P^{(l)}} < \frac{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}}{1024}\sqrt{\frac{m}{k}}, \tag{68}$$

then $j \notin \mathcal{U}_{T_l+1}$: arm j is rejected no later than the last round of the epoch.

Proof. Fix l and j satisfying Equation (68). If $j \notin \mathcal{U}_{T_l}$, the conclusion already holds. Otherwise, j remains unknown throughout the epoch, and we show that it is rejected in round T_l. All k optimal arms are also unknown, since no arm has been accepted in an epoch that is not final (Lemmas 9 and 12).

Throughout the epoch, $\mathcal{A}_u = \{\}$ and $\mathcal{P}_u = \{P^{(l)}\}$. Equation (68) gives $\theta_{P^{(l)}} > 1 - \Delta_{k,r^*_{\boldsymbol{\theta}}(j)} \ge \theta_j$, so $j \ne P^{(l)}$. By Lemma 12, $0 < 1 - \theta_{P^{(l)}}$ and all comparison counts with $P^{(l)}$ start at zero, as do their predictable counterparts by Equation (19).

Equation (68) gives $1 - \theta_{P^{(l)}} < 1/8$, so Lemma 20 applies.

For each $a \in (S^{(k)*}_{\boldsymbol{\theta}} \cup \{j\}) \setminus \{P^{(l)}\}$, Lemma 20, Equation (63), and Equation (68) give:

$$\begin{aligned} N^{(P^{(l)})}_{a,T_l+1} &> \frac{m\ln(2nT/\delta)}{256ek(1-\theta_{P^{(l)}})^2} - 2\ln(2nT/\delta) \\ &> \frac{1024\ln(2nT/\delta)}{\Delta^2_{k,r^*_{\boldsymbol{\theta}}(j)}}. \end{aligned} \tag{69}$$

Consequently,

$$\beta_{\delta,n,T}(N^{(P^{(l)})}_{a,T_l+1}) < \frac{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}}{32}, \qquad a \in S^{(k)*}_{\boldsymbol{\theta}} \cup \{j\}. \tag{70}$$

This also holds for $a = P^{(l)}$ when it is optimal, since $N^{(P^{(l)})}_{P^{(l)},T_l+1} = +\infty$ and $\beta_{\delta,n,T}(+\infty) = 0$.

For every $i \in S^{(k)*}_{\boldsymbol{\theta}}$, Lemma 5 and $\theta_{P^{(l)}} \ge 1/2$ give:

$$\mu_i^{(P^{(l)})} - \mu_j^{(P^{(l)})} \ge \frac{\theta_i - \theta_j}{8} \ge \frac{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}}{8}. \tag{71}$$

The same bound holds for $i = P^{(l)}$ by the convention $\mu^{(P^{(l)})}_{P^{(l)}} = 1/2$. Combining $\mathcal{G}^{(\delta)}_{\text{conf},T}$ with Equations (70) and (71) yields:

$$\begin{aligned} &\text{LCB}^{(P^{(l)})}_{i,T_l+1} - \text{UCB}^{(P^{(l)})}_{j,T_l+1} \\ &\quad \ge \frac{\Delta_{k,r^*_{\boldsymbol{\theta}}(j)}}{8} - 2\beta_{\delta,n,T}(N^{(P^{(l)})}_{i,T_l+1}) - 2\beta_{\delta,n,T}(N^{(P^{(l)})}_{j,T_l+1}) > 0. \end{aligned} \tag{72}$$

Thus all $k = k - |\mathcal{A}_{T_l}|$ optimal arms, which belong to $\mathcal{U}_{T_l}$, satisfy the strict comparison against j. Algorithm 1 therefore rejects j in round T_l (Line 17), so $j \notin \mathcal{U}_{T_l+1}$.

□

Combining the contraction of pivot gaps with this rejection property bounds the total contribution of each suboptimal arm across all high epochs that are not final.

Lemma 22 (Sum of Individual Arm Regret Bounds)**.** *Suppose $2 \le k < n$, $T \ge n$, and $\delta \in (0,1]$. Under both good events $\mathcal{G}_{conf,T}^{(\delta)}$ and $\mathcal{G}_{pulls}^{(\delta,\lambda)}$ (with $\lambda = 1$), for an absolute constant $C_{10} > 0$, the following bound holds:*

$$\begin{aligned}\sum_{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T)} \sum_{j \in \mathcal{U}_{T_{l-1}+1} \setminus S_{\boldsymbol{\theta}}^{(k)*}} \min\left\{ \frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}, \frac{1}{1-\theta_{P^{(l)}}} \right\} \\ \le C_{10}(1+\ln(k/m)) \sum_{j \in [\![n]\!] \setminus S_{\boldsymbol{\theta}}^{(k)*}} \frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}.\end{aligned} \tag{73}$$

Proof. Fix $j \notin S_{\boldsymbol{\theta}}^{(k)*}$. We bound its contribution to Equation (73) by splitting the high epochs that are not final into three ranges according to their pivot gap. By Lemma 19, the positive gaps of consecutive high epochs that are not final contract by a factor at most $95/96$. First, consider $1-\theta_{P^{(l)}} \ge \Delta_{k,r_{\boldsymbol{\theta}}^*(j)}$. Here the minimum in Equation (73) equals $1/(1-\theta_{P^{(l)}})$. For consecutive epochs in this range, contraction gives $1/(1-\theta_{P^{(l)}}) \le (95/96)/(1-\theta_{P^{(l+1)}})$, and the last reciprocal is at most $1/\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}$. Summing backwards and dropping the restriction $j \in \mathcal{U}_{T_{l-1}+1}$ therefore gives:

$$\sum_{\substack{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T) \\ 1-\theta_{P^{(l)}} \ge \Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}} \frac{1}{1-\theta_{P^{(l)}}} \le \frac{1}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}} \sum_{h=0}^{\infty} (95/96)^h = \frac{96}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}.$$

Second, consider epochs with:

$$\frac{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}{1024}\sqrt{\frac{m}{k}} \le 1-\theta_{P^{(l)}} < \Delta_{k,r_{\boldsymbol{\theta}}^*(j)}.$$

The minimum in Equation (73) equals $1/\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}$, and the ratio of the range's upper and lower bounds is $1024\sqrt{k/m}$. The capped term is the same in each of these epochs, so it remains to count how many epochs fit in this gap range. By Lemma 19, each additional epoch reduces the pivot gap by a factor at most $95/96$. Thus the number of these epochs is at most:

$$1 + \frac{\ln(1024\sqrt{k/m})}{\ln(96/95)} = \mathcal{O}(1+\log(k/m)).$$

Finally, consider $1-\theta_{P^{(l)}} < \frac{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}{1024}\sqrt{\frac{m}{k}}$. The minimum in Equation (73) again equals $1/\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}$. At most one such epoch satisfies $j \in \mathcal{U}_{T_{l-1}+1}$: Lemma 21 removes j by that epoch's last update, and unknown sets only shrink. The three ranges therefore contribute at most:

$$\frac{98 + \ln(1024\sqrt{k/m})/\ln(96/95)}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}} \le \frac{C_{10}}{\Delta_{k,r_{\boldsymbol{\theta}}^*(j)}}(1+\ln(k/m)),$$

for an absolute constant $C_{10} > 0$, since $k/m \ge 1$. Summing this bound over all $j \notin S_{\boldsymbol{\theta}}^{(k)*}$ and interchanging the finite sums over j and l proves the statement. □

For the worst-case bound, the next lemma uses Lemma 19 to sum the square-root terms in Lemma 17.1, Equation (44), across epochs.

Lemma 23 (Worst-Case Summation across High Epochs)**.** *Suppose $2 \le k < n$, $T \ge n$, and $\delta \in (0,1]$. Under both good events $\mathcal{G}_{conf,T}^{(\delta)}$ and $\mathcal{G}_{pulls}^{(\delta,\lambda)}$ (with $\lambda = 1$), for every $j \in [\![n]\!] \setminus S_{\boldsymbol{\theta}}^{(k)*}$, there is an absolute constant $C_{11} > 0$ such that:*

$$\begin{aligned}\sum_{\substack{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T) \\ T_l - T_{l-1} \ge t_n^{(k,m,\delta,1)}}} \sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}} \\ \le C_{11} \sqrt{(1+\ln(k/m)) \sum_{\substack{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T) \\ T_l - T_{l-1} \ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}}.\end{aligned} \tag{74}$$

Proof. Fix $j \in [\![n]\!] \setminus S_{\boldsymbol{\theta}}^{(k)*}$. We first bound the nonzero terms in Equation (74) for which $1 - \theta_{P^{(l)}} \geq 1/8$ or $T_{\mathcal{U},j} < T_l$. Denote their epoch indices by $\mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)$.

There are only an absolute constant number of high epochs with $1 - \theta_{P^{(l)}} \geq 1/8$. To see this, let $l_1 < \cdots < l_r$ be their indices, assuming $r \geq 1$. Since epoch l_1 is high, the gap between the best arm and its pivot satisfies $1 - \theta_{P^{(l_1)}} \leq 1/2$. Iterating Lemma 19 between epochs l_1 and l_r, and using $l_r - l_1 \geq r - 1$, gives:

$$\frac{1}{8} \leq 1 - \theta_{P^{(l_r)}} \leq \left(\frac{95}{96}\right)^{l_r - l_1} (1 - \theta_{P^{(l_1)}}) \leq \frac{1}{2}\left(\frac{95}{96}\right)^{r-1}.$$

Rearranging gives $r \leq 1 + \ln 4 / \ln(96/95)$.

A nonzero term with $T_{\mathcal{U},j} < T_l$ requires $T_{l-1} + 1 \leq T_{\mathcal{U},j} < T_l$. Since the last-active time $T_{\mathcal{U},j}$ is fixed, it can belong to at most one epoch. Together with the preceding bound on the number of epochs with $1 - \theta_{P^{(l)}} \geq 1/8$, this gives $|\mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)| \leq 2 + \ln 4 / \ln(96/95)$. Cauchy–Schwarz gives:

$$\begin{aligned}
&\sum_{l \in \mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)} \sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}} \\
&\quad \leq \sqrt{2 + \frac{\ln 4}{\ln(96/95)}} \sqrt{\sum_{l \in \mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}}.
\end{aligned} \tag{75}$$

We now consider the remaining nonzero terms, for which $T_{\mathcal{U},j} \geq T_l$ and $1 - \theta_{P^{(l)}} < 1/8$. Denote their epoch indices by $\mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)$. Thus $\min\{T_l, T_{\mathcal{U},j}\} = T_l$, and it suffices to bound the full-epoch sum $\tilde{Q}_{T_{l-1}+1:T_l}$.

For an epoch $l \in \mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)$, Lemma 15.2, applied to i_1^* with $T_{\text{start}} = T_{l-1} + 1$, $p = P^{(l)}$, $t = T_l$, and $\lambda = 1$, bounds $\tilde{Q}_{T_{l-1}+1:T_l}$ from above after rearranging its gap bound. The burn-in condition holds by assumption, and $i_1^* \in \mathcal{U}_{T_l} \setminus \{P^{(l)}\}$ by Lemma 12. This upper bound follows because the epoch stops after certifying one improvement: a larger sampling budget would already let the best arm pass the test, without needing to resolve the smaller differences. Conversely, due to our conservative confidence bounds, the epoch cannot end until enough evidence has accumulated to certify an improvement, which gives the lower bound. Combining this upper bound with Lemma 20, Equation (62), there are absolute constants $c, C > 0$ such that:

$$c\frac{\ln(2nT/\delta)}{(1 - \theta_{P^{(l)}})^2} \leq \tilde{Q}_{T_{l-1}+1:T_l} \leq C\frac{k}{m}\frac{\ln(2nT/\delta)}{(1 - \theta_{P^{(l)}})^2}. \tag{76}$$

Here $\tilde{Q}_{T_{l-1}+1:T_l}$ sums the conditional probabilities $\tilde{q}_t = (k-1)/(|\mathcal{U}_t| - 1)$ of selecting any fixed arm other than the pivot that remains unknown throughout the epoch. Thus it measures how often such an arm is sampled. Ranking feedback determines which selections yield observed comparisons with the pivot.

We want these sums to grow geometrically so that we can sum their square roots. As the pivot gap decreases, both bounds in Equation (76) increase. However, the upper bound is $(C/c)(k/m)$ times the lower bound, so they need not guarantee growth between consecutive epochs. We therefore compare epochs at least h pivot changes apart. By Lemma 19, sufficiently many changes make the later lower bound at least twice the earlier upper bound. The later sum must then be at least twice the earlier sum.

Choose

$$h = \max\left\{1, \left\lceil \frac{\ln(2(C/c)(k/m))}{2\ln(96/95)} \right\rceil \right\} = \mathcal{O}(1 + \log(k/m)). \tag{77}$$

Then $(C/c)(k/m)(95/96)^{2h} \leq 1/2$. For two epoch indices $l, l' \in \mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)$ with $l < l'$, Lemma 19 gives $1 - \theta_{P^{(l')}} \leq (95/96)^{l'-l}(1 - \theta_{P^{(l)}})$. Combining this with Equation (76) gives, whenever $l' \geq l + h$:

$$\tilde{Q}_{T_{l-1}+1:T_l} \leq C\frac{k}{m}\frac{\ln(2nT/\delta)}{(1 - \theta_{P^{(l)}})^2}$$

$$\leq \frac{C}{c}\frac{k}{m}\left(\frac{95}{96}\right)^{2(l'-l)}\tilde{Q}_{T_{l'-1}+1:T_{l'}} \leq \frac{1}{2}\tilde{Q}_{T_{l'-1}+1:T_{l'}}. \tag{78}$$

For each $a \in [\![h]\!]$, let $\mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)$ contain the epoch indices $l \in \mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)$ satisfying $l \equiv a \pmod{h}$. Thus $\mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)$ is the group of remaining epoch indices with remainder a modulo h (with $a = h$ representing remainder zero). Two distinct indices in the same group differ by at least h, so Equation (78) makes successive square-root terms grow by a factor of at least $\sqrt{2}$. Fix $a \in [\![h]\!]$ for which $\mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T) \neq \{\}$, so this group contains at least one epoch index. Write $\mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T) = \{l_1, \ldots, l_r\}$ with $l_1 < \cdots < l_r$ and $r \geq 1$. Iterating Equation (78) from l_s to l_r gives, for every $s \in [\![r]\!]$:

$$\sqrt{\tilde{Q}_{T_{l_s-1}+1:T_{l_s}}} \leq 2^{-(r-s)/2}\sqrt{\tilde{Q}_{T_{l_r-1}+1:T_{l_r}}}.$$

Summing the resulting geometric series backwards, and bounding its last term by the square root of the sum of squared terms, gives:

$$\begin{aligned}
&\sum_{l \in \mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)} \sqrt{\tilde{Q}_{T_{l-1}+1:T_l}} \\
&\leq \sqrt{\tilde{Q}_{T_{l_r-1}+1:T_{l_r}}} \sum_{s=1}^{r} 2^{-(r-s)/2} \\
&\leq \sqrt{\tilde{Q}_{T_{l_r-1}+1:T_{l_r}}} \sum_{u=0}^{\infty} 2^{-u/2} = \frac{\sqrt{\tilde{Q}_{T_{l_r-1}+1:T_{l_r}}}}{1-1/\sqrt{2}} \\
&\leq \frac{1}{1-1/\sqrt{2}} \sqrt{\sum_{l \in \mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)} \tilde{Q}_{T_{l-1}+1:T_l}}.
\end{aligned}$$

If $\mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T) = \{\}$, both sums in the bound are zero, so the bound also holds for this group. Summing over $a \in [\![h]\!]$ and applying Cauchy–Schwarz gives:

$$\begin{aligned}
&\sum_{a=1}^{h} \sum_{l \in \mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)} \sqrt{\tilde{Q}_{T_{l-1}+1:T_l}} \\
&\leq \frac{1}{1-1/\sqrt{2}} \sum_{a=1}^{h} \sqrt{\sum_{l \in \mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)} \tilde{Q}_{T_{l-1}+1:T_l}} \\
&\leq \frac{\sqrt{h}}{1-1/\sqrt{2}} \sqrt{\sum_{a=1}^{h} \sum_{l \in \mathfrak{L}_{j,a}^{(k,m,\delta,h)}(T)} \tilde{Q}_{T_{l-1}+1:T_l}}.
\end{aligned} \tag{79}$$

The exceptional set $\mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)$ and the remaining set $\mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)$ partition the indices of the nonzero summands in Equation (74). Adding the bounds in Equations (75) and (79), and then applying Cauchy–Schwarz, gives:

$$\begin{aligned}
&\sum_{\substack{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T) \\ T_l - T_{l-1} \geq t_n^{(k,m,\delta,1)}}} \sqrt{\tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}} \\
&\leq \sqrt{2 + \frac{\ln 4}{\ln(96/95)}} \sqrt{\sum_{l \in \mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\sqrt{h}}{1-1/\sqrt{2}} \sqrt{\sum_{l \in \mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)} \tilde{Q}_{T_{l-1}+1:T_l}} \\
&\leq \sqrt{2 + \frac{\ln 4}{\ln(96/95)} + \frac{h}{(1-1/\sqrt{2})^2}} \\
&\quad \cdot \sqrt{\sum_{l \in \mathfrak{L}_{j,\mathrm{exc}}^{(k,m,\delta)}(T)} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}} + \sum_{l \in \mathfrak{L}_{j,\mathrm{rem}}^{(k,m,\delta)}(T)} \tilde{Q}_{T_{l-1}+1:T_l}} \\
&= \sqrt{2 + \frac{\ln 4}{\ln(96/95)} + \frac{h}{(1-1/\sqrt{2})^2}} \\
&\quad \cdot \sqrt{\sum_{\substack{l \in \mathfrak{L}_{\mathrm{high},\neg\mathrm{final}}(T) \\ T_l - T_{l-1} \geq t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l, T_{\mathcal{U},j}\}}}.
\end{aligned}
$$

By Equation (77), $h = \mathcal{O}(1 + \log(k/m))$, so the prefactor is $\mathcal{O}(\sqrt{1 + \log(k/m)})$, proving Equation (74). □

We now combine these results to prove Theorem 2. We sum the contributions of the low epochs that are not final, the high epochs that are not final, and the final epoch on the good events, and then bound the contribution of event failures to the expected regret.

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 $\beta_{\delta,n,T}(0) \coloneqq +\infty$ and $\beta_{\delta,n,T}(+\infty) \coloneqq 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top-k regret suffered by* Top-k-UIF*:*

$$
\mathbb{E}\mathsf{Reg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(T) \leq \frac{C_2}{m} \left(kn(n-k) + (1 + \ln(k/m)) \sum_{u \in [\![k+1,n]\!]} 1/\Delta_{k,u} \right) \ln(T),
$$

where $C_2 \in \mathbb{R}_{>0}$ is an absolute constant and $\Delta_{k,u} \coloneqq \theta_{r^{*-1}_{\boldsymbol{\theta}}(k)} - \theta_{r^{*-1}_{\boldsymbol{\theta}}(u)}$ for all $u \in [\![k+1, n]\!]$.

Furthermore, the worst-case expected top-k regret is bounded by:

$$
\begin{aligned}
&\sup_{\nu \in \mathcal{E}_n^{(k,m)}} \mathbb{E}\mathsf{Reg}^{(k)}_{\text{Top-}k\text{-UIF},\nu}(T) \\
&\qquad \leq C_3 \left(\frac{kn(n-k)}{m} \ln(nT) + \sqrt{\frac{n-k}{m} \min\left\{1, \frac{n-k}{k}\right\} T \ln(nT)(1 + \ln(k/m))} \right),
\end{aligned}
$$

where $C_3 \in \mathbb{R}_{>0}$ is an absolute constant. If $T \geq \frac{k^2 n^2 \max\{k, n-k\}}{m} \frac{\ln(nT)}{1+\ln(k/m)}$, the additive term is absorbed into the square-root term after changing the absolute constant.

Proof. If $k = n$, every played set contains every arm and the regret is identically zero. Suppose $k < n$, and set $\delta = 1/T$ and $\lambda = 1$. By Lemmas 6 and 8, the failure probabilities of $\mathcal{G}^{(\delta)}_{\mathrm{conf},T}$ and $\mathcal{G}^{(\delta,\lambda)}_{\mathrm{pulls}}$ are at most $1/T$ each. Since total pseudo-regret lies in $[0, T]$, these failures contribute at most 2 to expected regret. We therefore work under both good events below.

Lemma 12 gives $L(T) \leq n$. Thus Lemma 17.2, together with the burn-in terms in Parts 1 and 3 of that lemma, bounds the total regret of low epochs that are not final and all burn-in contributions by $\mathcal{O}(kn(n-k)\log T/m)$, using $T \geq n$. This term also absorbs the failure contribution.

For the instance-dependent bound, apply Lemma 22 to the arm sums in Lemma 17.1, Equation (43) and add the final-epoch bound in Lemma 17.3. Together with the preceding additive bound, this

gives:

$$\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(1,T) \le C\frac{\ln T}{m}\left(kn(n-k) + (1+\ln(k/m))\sum_{u=k+1}^{n}\frac{1}{\Delta_{k,u}}\right),$$

for an absolute constant $C > 0$. Taking expectations proves the first claim.

For the worst-case bound, apply Lemma 23 to the arm sums in Lemma 17.1, Equation (44). Epochs shorter than the burn-in contribute only to the preceding additive bound. Expanding $\tilde{Q}$ and interchanging the sums, Lemma 9 gives $\tilde{q}_t|\mathcal{U}_t \setminus S^{(k)*}_{\boldsymbol{\theta}}| \le \tilde{q}_t|\mathcal{U}'_t| \le k$. Since the epochs are disjoint and $\tilde{q}_t \le 1$, we obtain:

$$\sum_{j\notin S^{(k)*}_{\boldsymbol{\theta}}}\sum_{\substack{l\in\mathfrak{L}_{\text{high},\neg\text{final}}(T)\\ T_l - T_{l-1}\ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}} \le kT, \tag{80}$$

$$\sum_{\substack{l\in\mathfrak{L}_{\text{high},\neg\text{final}}(T)\\ T_l - T_{l-1}\ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}} \le T \qquad \text{for every } j \notin S^{(k)*}_{\boldsymbol{\theta}}. \tag{81}$$

Applying Cauchy–Schwarz across the $n-k$ suboptimal arms and then Equation (80) gives:

$$\begin{aligned}&\sum_{j\in[\![n]\!]\setminus S^{(k)*}_{\boldsymbol{\theta}}}\sqrt{\sum_{\substack{l\in\mathfrak{L}_{\text{high},\neg\text{final}}(T)\\ T_l - T_{l-1}\ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}}\\ &\le \sqrt{(n-k)\sum_{j\in[\![n]\!]\setminus S^{(k)*}_{\boldsymbol{\theta}}}\sum_{\substack{l\in\mathfrak{L}_{\text{high},\neg\text{final}}(T)\\ T_l - T_{l-1}\ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}} \le \sqrt{k(n-k)T}.\end{aligned} \tag{82}$$

Alternatively, applying Equation (81) to each arm gives:

$$\sum_{j\in[\![n]\!]\setminus S^{(k)*}_{\boldsymbol{\theta}}}\sqrt{\sum_{\substack{l\in\mathfrak{L}_{\text{high},\neg\text{final}}(T)\\ T_l - T_{l-1}\ge t_n^{(k,m,\delta,1)}}} \tilde{Q}_{T_{l-1}+1:\min\{T_l,T_{\mathcal{U},j}\}}} \le (n-k)\sqrt{T}. \tag{83}$$

By Lemma 23 and Lemma 17.1, Equation (44), the common left-hand side of Equations (82) and (83) is multiplied by $\sqrt{\ln(2nT^2)(1+\ln(k/m))/(km)}$, up to an absolute constant. Using the smaller of these two bounds, the contribution of high epochs that are not final, excluding burn-in, is at most:

$$\mathcal{O}\left(\frac{1}{\sqrt{m}}\min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\}\sqrt{T\log(2nT^2)(1+\log(k/m))}\right).$$

Adding the final-epoch bound from Lemma 17.3 and the previously bounded contributions of low epochs, burn-in, and event failures gives, after taking expectations, for an absolute constant $C > 0$:

$$\begin{aligned}&\mathbb{E}\mathsf{Reg}^{(k)}_{\text{TOP-}k\text{-UIF},\nu}(T)\\ &\quad\le C\left(\frac{kn(n-k)}{m}\ln(nT) + \frac{1}{\sqrt{m}}\min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\}\sqrt{T\ln(nT)(1+\ln(k/m))}\right).\end{aligned}$$

Here we used $\ln(2nT^2) \le 3\ln(nT)$.

Finally, the condition $T \ge \frac{k^2n^2\max\{k,n-k\}}{m}\frac{\ln(nT)}{1+\ln(k/m)}$ ensures:

$$\frac{kn(n-k)}{m}\ln(nT) \le \min\left\{\sqrt{n-k}, \frac{n-k}{\sqrt{k}}\right\}\sqrt{\frac{T\ln(nT)(1+\ln(k/m))}{m}}.$$

Thus the additive term can be absorbed into the square-root term under the stated horizon condition.

□

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 $\nu_{mnl} \in \mathcal{E}^{(k)}_{mnl,n}$ 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 $\beta_{\delta,n,T}(0) \coloneqq +\infty$ and $\beta_{\delta,n,T}(+\infty) \coloneqq 0$) and $\delta = 1/T$, we obtain the following instance-dependent upper bound to the expected top-k regret suffered by* MNL-TOP-k-UIF*:*

$$\mathbb{E}\mathsf{Reg}^{(k)}_{\text{MNL-TOP-}k\text{-UIF},\nu_{mnl}}(T) \leq C_4\Big(kn(n-k) + \ln(k) \sum_{u \in [\![k+1,n]\!]} 1/\Delta_{k,u}\Big) \ln(T),$$

where $C_4 \in \mathbb{R}_{>0}$ is an absolute constant and $\Delta_{k,u} \coloneqq \theta_{r^{*-1}_{\boldsymbol{\theta}}(k)} - \theta_{r^{*-1}_{\boldsymbol{\theta}}(u)}$ for all $u \in [\![k+1, n]\!]$. Furthermore, the worst-case expected top-k regret is bounded by:

$$\sup_{\nu_{mnl} \in \mathcal{E}^{(k)}_{mnl,n}} \mathbb{E}\mathsf{Reg}^{(k)}_{\text{MNL-TOP-}k\text{-UIF},\nu_{mnl}}(T)$$
$$\leq C_5 \left(kn(n-k)\ln(nT) + \sqrt{(n-k)\min\left\{1, \frac{n-k}{k}\right\} T \ln(nT)\ln(k)} \right),$$

where $C_5 \in \mathbb{R}_{>0}$ is an absolute constant. If $T \geq k^2 n^2 \max\{k, n-k\}\frac{\ln(nT)}{\ln(k)}$, the additive term is absorbed into the square-root term after changing the absolute constant.

Proof. 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:

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

Consequently, Lemma 5 continues to hold with m/k replaced by $1/(k+1)$.

The MNL analogues of $\mathcal{G}^{(\delta)}_{\text{conf},T}$ and $\mathcal{G}^{(\delta,\lambda)}_{\text{pulls}}$ each have failure probability at most δ, by the proofs of Lemmas 6 and 8. Their union bounds still involve only n^2 distinct estimator pairs (i,p) with $i \in [\![n]\!]$, $p \in [\![0,n]\!]$, and $i \neq p$.

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 9 applies to the arms in $[\![n]\!]$. The induction proving Lemma 9 also establishes the following pivot invariant. At the first acceptance, if $P_t = 0$ and NEW-PIVOT returns no candidate, Algorithm 3 selects $P_{t+1} \in \mathcal{A}_{t+1}$; thereafter, the pivot remains accepted.

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

Define:

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

Every arm in $\mathcal{U}'_t$ is pulled with conditional probability $\widetilde{q}_t$. When $P_t \in [\![n]\!]$, 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:

$$\mathsf{Ist}\mathsf{Reg}^{(k)}_{\text{MNL-TOP-}k\text{-UIF},\nu_{\mathrm{mnl}}}(t) = \frac{1}{k|\mathcal{U}_t|} \sum_{i \in \mathcal{U}_t \cap S^{(k)*}_{\boldsymbol{\theta}}} \sum_{j \in \mathcal{U}_t \setminus S^{(k)*}_{\boldsymbol{\theta}}} (\theta_i - \theta_j).$$

Splitting the inner sum according to $\theta_j \geq \theta_0$ or $\theta_j < \theta_0$ yields the analogue of Lemma 13, since in the former case $\theta_i - \theta_j \leq \theta_i - \theta_0$.

Every epoch after the initial one has a real pivot, $P_t \in [\![n]\!]$, so the low-pivot phase split in the proof of Lemma 17 applies unchanged.

Lemmas 14 and 15 apply with m/k replaced by $1/(k+1)$, with the predictable process defined from the probabilities $\widetilde{q}_t$, and with burn-in:

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

Set $\delta = 1/T$ and $\lambda = 1$, and suppose $k < n$, since otherwise regret is zero. We use the gap notation of Theorem 2 for the real arms. For epochs with real pivots, the observation identity above gives $f^{(p)}_{i|S_t\cup\{0\}} \geq (\theta_i + \theta_p)/(4k)$, and Lemma 7 follows by the same argument with $m = 1$ and item 0 included in every comparison set. The sampling and update rules and the fresh counts of Lemma 12 are unchanged. Thus, under both $\mathcal{G}^{(\delta)}_{\text{conf},T}$ and $\mathcal{G}^{(\delta,\lambda)}_{\text{pulls}}$, Lemmas 17–23 apply to these epochs with $m = 1$.

At the transition from pivot 0 to an accepted real arm, we may have $\theta_{P_{T_1+1}} \leq \theta_0$, so Lemma 19 does not apply to this transition. We therefore bound the initial epoch separately. Throughout this epoch, $i_1^* \in \mathcal{U}_t$. Apply Lemma 16 with the preceding observation and regret identities and $\tau = \bar{t}_n^{(k,1/T,1)}$. If $\theta_0 \geq 1/2$, Lemma 15 bounds both types of gaps with $b = 64\sqrt{32/(1-e^{-1})}\sqrt{k+1}$. If $\theta_0 < 1/2$, every gap is at most $2(1-\theta_0)$, and Lemma 15.2, applied to i_1^*, gives $b = 16\sqrt{32/(1-e^{-1})}\sqrt{k+1}$. Inverting these gap bounds gives the instance-dependent estimate. For the worst-case estimate, interchange the arm and round sums and apply Cauchy–Schwarz. This yields:

$$\begin{aligned}\mathsf{Reg}^{(k)}_{\text{MNL-Top-}k\text{-UIF},\nu_{\text{mnl}}}(T_0+1,T_1) &= \mathcal{O}\left(n(n-k)\log T\right)\\ &+ \mathcal{O}\left(\min\left\{\sum_{u=k+1}^{n}\frac{\log T}{\Delta_{k,u}},\, \min\left\{\sqrt{n-k},\frac{n-k}{\sqrt{k}}\right\}\sqrt{(T_1-T_0)\log T}\right\}\right).\end{aligned}$$

Since $\widetilde{q}_t = k/|\mathcal{U}_t|$ during the initial epoch, we have $\widetilde{q}_t|\mathcal{U}_t \setminus S^{(k)*}_{\boldsymbol{\theta}}| \leq k$ and $\widetilde{q}_t \leq 1$. We also used $(k+1)/k \leq 3/2$ and $T \geq n$.

Now sum as in the proof of Theorem 2, adding the initial-epoch bound above. There are at most $n+1$ epochs. Lemma 23 applies to the high epochs that are not final with real pivots, using the MNL burn-in $\bar{t}_n^{(k,1/T,1)}$ and $m = 1$ (since $(k+1)/k \leq 3/2$), giving the factor $\sqrt{1+\ln(k)}$. Since $k \geq 2$, the inequality $1 + \ln(k) \leq (1 + 1/\ln 2)\ln(k)$ allows us to replace this factor by $\sqrt{\ln(k)}$ after changing the absolute constant. We use the separate square-root bounds for the initial and final epochs, counting the epoch only once if they coincide. Adding these contributions changes only the absolute constant. The total burn-in and low-epoch additive contribution remains $\mathcal{O}(kn(n-k)\log T)$, and event failures contribute at most 2 to expected regret. The same summation and horizon-absorption arguments therefore prove both claims with $m = 1$, uniformly in θ_0. □

E Auxiliary Results

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

E.1 Information Theory

Lemma 24 (Pinsker's Inequality for Bounded Variables)**.** *Let μ and ϕ be two probability measures on a measurable space $(\mathcal{X},\mathcal{F})$, and let $X : \mathcal{X} \to [0,1]$ be a random variable. Then, it holds that:*

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

Proof. Let $\|\mu - \phi\|_{\mathrm{TV}} \coloneqq \sup_{A\in\mathcal{F}} |\mu(A) - \phi(A)|$ 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:

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

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 $\|\phi - \mu\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\phi \| \mu)}$, yields the stated result. □

E.2 Concentration Inequalities

Theorem 25 (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 $(c_t)_{t=1}^{+\infty} \subseteq (0, +\infty)$ such that:*

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

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

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

E.3 General Probability

Lemma 26. *Let $\mathcal{U}$ be a finite set, $c \in [\![|\mathcal{U}|]\!]$. Then, for every $i \in \mathcal{U}$:*

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

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:

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

□

F Numerical Simulations

In this appendix, we report additional details on the numerical validation reported in Section 6.

F.1 Parameters of the Instances

In this section, we report the PL parameters $\boldsymbol{\theta}$ of the instances ν used in the numerical simulations.

- TOP-k-UIF *vs* RECMAXMINUCB. For $\varepsilon \in \{0.05, 0.1, 0.2\}$, we use $n = 400$, $k = m = 2$, and the PL parameters:

$$\theta_i = \begin{cases} 1 & i \in \{1, 2\} \\ 1 - \varepsilon & i \in \{3, \dots, 12\} \\ 0.05 & i \in \{13, \dots, 400\} \end{cases}.$$

  Figure 1 uses $\varepsilon = 0.1$.
- *Sensitivity of* TOP-k-UIF *to* m. The PL parameters are defined as $\boldsymbol{\theta} = (1.0, 0.93, 0.82, 0.78, 0.74, 0.7, 0.6, 0.57, 0.54, 0.51, 0.48, 0.45, 0.3, 0.25, 0.2, 0.15)$.
- MNL-TOP-k-UIF *vs Saha and Gaillard (2025).* The PL parameters are $\boldsymbol{\theta} = (1.00, 0.95, 0.90, 0.85, 0.80, 0.79, 0.78, 0.77, 0.76, 0.75, 0.50, 0.47, 0.44, 0.41, 0.38, 0.35, 0.32, 0.29, 0.26, 0.23, 0.20, 0.18, 0.16, 0.14, 0.12, 0.10, 0.08, 0.06, 0.04, 0.02)$.

F.2 Sensitivity to the Boundary Gap

Table 2 reports the final mean cumulative regret and pointwise 95% confidence intervals for the three boundary gaps, using 20 independent trials per algorithm and gap. TOP-k-UIF incurs less mean regret than RECMAXMINUCB for each tested gap.

Table 2: Final mean cumulative top-k regret, with pointwise 95% confidence intervals over 20 trials per algorithm and gap. Values are in units of 10^3; $n = 400$, $k = m = 2$, and $T = 10^6$.

Boundary gap ε	TOP-k-UIF	RECMAXMINUCB
0.05	141.1 $[135.0, 147.3]$	292.9 $[220.4, 365.3]$
0.1	160.3 $[153.7, 166.8]$	300.3 $[205.1, 395.6]$
0.2	160.4 $[156.0, 164.8]$	346.3 $[256.5, 436.1]$

F.3 Algorithm Parameters

Both TOP-k-UIF and MNL-TOP-k-UIF use the confidence radius $\beta_{\delta,n,T}(s) = \sqrt{\ln(2nT/\delta)/s}$ for $s > 0$, with $\delta = 0.05$, $\beta_{\delta,n,T}(0) = +\infty$, and $\beta_{\delta,n,T}(+\infty) = 0$. We set $T = 10^6$ and use 20 independent repetitions per setting. The baselines use $\alpha = 0.51$ for RECMAXMINUCB and $x = 2 \ln T$ for both AOA-RB variants. All error bars and bands show pointwise 95% Student-t confidence intervals for the mean, computed as $\overline{R} \pm t_{0.975,N-1}\, s/\sqrt{N}$, where s is the sample standard deviation across the $N = 20$ independent runs.

F.4 Computational Resources

The primary suite comprises 360 independent trials with $T = 10^6$. We used Python 3.12.3, NumPy 1.26.4, SciPy 1.13.1, and Experiments W/O Stress (EWS; `https://github.com/m1gwings/experiments-wo-stress`), a Python library for reproducible parallel numerical experiments, analysis, and plotting. The experiments ran on an Ubuntu 26.04 machine equipped with two AMD EPYC 9135 processors, 32 physical cores, 64 logical CPUs, and approximately 1 TB of RAM. We used 48 parallel workers, with one numerical-library thread per worker.

The primary simulations and analysis took 24.6 minutes of elapsed time and 19.26 hours of cumulative simulation-worker time.

The three ranking-gap settings, feedback-level study, and MNL study comprise 440 full trials in total. Their invocations, including analysis and figure export, took 42.9 minutes of summed elapsed time and 26.39 hours of simulation-worker time. Table 3 reports per-trial timings. Diagnostic validation used the same instance and algorithm specifications for 18 trials of 10^4 rounds with four workers, taking 5.5 seconds of elapsed time and 18.2 seconds of simulation-worker time.

Table 3: Simulation-task times in seconds. Each MNL row summarizes 80 trials: 20 for each of the four outside weights. The feedback-level row summarizes 80 trials: 20 for each $m \in \{1, 2, 3, 6\}$.

Experiment	Algorithm	Trials	Mean (s)	Range (s)
Ranking ($\varepsilon = 0.05$)	RECMAXMINUCB	20	474.3	429.3–516.6
Ranking ($\varepsilon = 0.05$)	TOP-k-UIF	20	204.9	169.6–241.0
Ranking ($\varepsilon = 0.1$)	RECMAXMINUCB	20	812.6	792.3–826.6
Ranking ($\varepsilon = 0.1$)	TOP-k-UIF	20	246.2	237.5–253.2
Ranking ($\varepsilon = 0.2$)	RECMAXMINUCB	20	472.1	418.5–584.0
Ranking ($\varepsilon = 0.2$)	TOP-k-UIF	20	131.6	88.9–163.4
Feedback level	TOP-k-UIF	80	93.1	65.5–135.5
MNL	MNL-TOP-k-UIF	80	262.2	238.6–273.3
MNL	AOA-RB$_{\mathrm{PL}}$-Adaptive	80	130.1	120.0–134.9
MNL	AOA-RB$_{\mathrm{PL}}$	80	116.6	85.9–124.4

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the problem of sequential decision-making from top-m ranking feedback and accurately summarize the theoretical contributions, specifically the lower and upper bounds provided. These claims directly map to the main theoretical results presented in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 2 specifies the PL model and normalized top-k objective. Theorems 2 and 3 give explicit additive terms and the horizon conditions for the near-optimal leading rate.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems fully state their operating assumptions within the main text. Furthermore, we provide rigorous proofs in the appendices (specifically Appendix B for the lower bound and Appendix D for the upper bounds).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental parameters, including learning horizons and exact Plackett-Luce parameters used in the numerical simulations, are fully disclosed in Section 6 and Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, experiment configurations, and reproduction instructions are hosted in the repository linked in Section 6.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 and Appendix F provide full experimental setups including learning horizons, number of items, cardinality of the pulled sets, trials run, and the specific Plackett-Luce parameters of the instances.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are explicitly reported in the plotted results. The cumulative regrets plotted in Figures 1, 2, and 3 report the mean values averaged across 20 independent runs with 95% Student-t confidence intervals, as specified in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F.4 states the exact computing resources utilized.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This theoretical machine learning research uses standard synthetic multi-armed bandit settings and conforms with all ethical standards established in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper performs foundational mathematical research. There are no direct paths to specific positive or negative societal impacts to discuss beyond general machine learning theory.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper consists exclusively of mathematical derivations and experiments run on purely synthetic data. It does not release any models, weights, or curated datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: No pre-existing datasets, proprietary models, or external code assets were used or adapted in this paper. All evaluation was conducted using fully simulated synthetic environments designed for the problem.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The research develops algorithmic strategies for bandit bounds and provides reproduction scripts. It does not introduce or release any new general-purpose software or dataset assets requiring documentation beyond the provided code.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research relies completely on mathematics and computer simulations of multi-armed bandit settings. There were no human subjects or crowdsourcing involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Because the research is theoretical and does not involve human study participants or crowdsourcing, IRB approvals or equivalents are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core methodology consists of mathematical modeling and algorithmic design. We used LLMs as proof assistants and to review the proofs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.