
Variance-Aware Optimal Ranking in Log-Concave Random Utility Models

Diego Aloviseti

Politecnico di Milano, Milan, Italy
diego.aloviseti@polimi.it

Marco Mussi

Politecnico di Milano, Milan, Italy
marco.mussi@polimi.it

Alberto Maria Metelli

Politecnico di Milano, Milan, Italy
albertomaria.metelli@polimi.it

Abstract

We consider the problem of learning a utility-based ranking of k items from noisy feedback signals generated by a Random Utility Model (RUM) under the minimal assumption of log-concavity. We analyze two feedback types: Full-Ranking (FR), where the complete ordering of the sampled utilities is observed, and Winner-Only (WO), where the item with the maximum sampled utility is observed. Within the Probably Approximately Correct framework, we define the notion of ε -accuracy to measure the distance between the estimated and true rankings based on latent utility gaps. We present two algorithms named VURs (Variance-aware Utility Rankers) that recover the utility ranking without requiring exact knowledge of the noise distribution. Exploiting a common design principle, the proposed algorithms rely solely on structural regularity: in the FR setting, this amounts to enforcing an upper bound V on the noise variance, while the WO setting leverages the monotonicity of the reversed hazard rate ratio. By establishing matching information-theoretic lower bounds, we prove that the sample complexity of VURs is order-optimal, up to logarithmic factors, scaling as V/ε^2 for FR feedback and $V/(P_{\min}\varepsilon^2)$ for WO feedback. Our analysis identifies the minimum winning probability $P_{\min}$ as the fundamental informational bottleneck of winner-only signals and highlights this as the intrinsic difference in “learning cost” between FR and WO feedback in general log-concave RUMs.

1 Introduction

In many machine learning paradigms, a set of items is characterized by latent scalar values, inducing a strict ranking. In recommender systems, for example, products or media are evaluated based on underlying utility values representing user preferences [1]. A structurally identical challenge arises in Reinforcement Learning from Human Feedback (RLHF) [2]. Here, an agent generates a set of responses (e.g., actions or trajectories), each associated with a latent reward. Because humans struggle to assign consistent scalar rewards, feedback can be elicited as preferences or orderings over observed trajectories. Consequently, the true utility or reward values remain inaccessible. In these situations, be it a recommender system or an RLHF agent, an interesting question may be to recover the exact scalar values. Alternatively, one can focus on the observable signal and try to recover the ranking implied by the utility values. While not requiring exact recovery of the utilities, the successful completion of this task may be sufficient for downstream tasks such as serving the optimal recommendation or executing the optimal policy.

In this paper, we formalize and address this ranking recovery problem. We fix a finite set of items and model the feedback-generation process as a Random Utility Model (RUM), a standard paradigm for human preferences [3], including the well-studied Plackett-Luce (PL) model [4, 5]. Additionally, we assume that the noise distribution is log-concave (e.g., Gaussian, logistic), a standard assumption in the literature [6], ensuring several desirable properties needed for successful learning [7]. Each item is associated with a scalar utility, and a RUM draw is obtained by adding independent, identically distributed noise to the true utilities.

To model the partial observability of human feedback, we introduce the concept of a *feedback filter*, a map from RUM draws to observable (possibly partial) rankings. We analyze two variants: Full-Ranking (FR), where the complete ordering of sampled utilities is observed, and Winner-Only (WO), where the learner observes the item with the highest sampled utility. We evaluate the ranking recovery problem within the Probably Approximately Correct (PAC) framework [8], utilizing the novel notion of ε -accuracy. Unlike standard discrete distances that penalize any permutation error regardless of magnitude, such as Kendall tau [9] and maximum rank difference [10], our metric is anchored in the underlying utilities. Specifically, a ranking is ε -accurate if it correctly orders all pairs whose latent utility gap is at least ε . This acknowledges that items with near-identical utilities are not only statistically difficult to distinguish but are often interchangeable in downstream decision-making.

Our proposed solution to the ranking problem is the Variance-aware Utility Ranker (VUR), a unified framework comprising two algorithms for recovering the full utility ranking from FR and WO signals, respectively. In summary, our main contributions are:

- **Novel Distance Metric.** We introduce the notion of ε -accuracy through Definition 1. To the best of our knowledge, this is the first generalization of the ε -optimality metric used in best-item identification [11] to the problem of learning a full utility ranking.
- **Unified Framework for Diverse Feedback.** We design two (ε, δ) -PAC algorithms for FR and WO feedback, respectively. Owing to the novel structural regularity stated as Assumption (WO2), these rely on a common design principle that allows us to view them as part of a single, unified framework for learning the utility ranking across diverse feedback types.
- **Distribution-agnostic Algorithms.** We propose two algorithms that require no exact knowledge of the noise distribution. Instead, they rely solely on milder structural conditions, formalized through Assumption (V) in the FR setting, and Assumption (WO2) for WO feedback. The sample-complexity dependence V/ε^2 (in addition to $P_{\min}$ for the WO setting) is shown to be common across the two settings by Theorem 1 and Theorem 3, where ε is the accuracy parameter and V is a known upper bound on the noise variance.
- **Fundamental Limits of Learnability.** We establish information-theoretic lower bounds for the ranking problem under log-concave RUMs, demonstrating that our algorithms are order-optimal. Crucially, Theorem 4 identifies the minimum winning probability, $P_{\min}$, as the fundamental informational bottleneck of WO feedback.

The related works are discussed in Appendix A.

2 Preliminaries

Standard big-O notation is used throughout, where $\tilde{O}$ is used to suppress the logarithmic dependences. For any $k \in \mathbb{N}$, we denote $[k] = \{1, \dots, k\}$. The set of all permutations on $[k]$, i.e., the bijective functions $\sigma : [k] \rightarrow [k]$, is denoted by S_k . We think of any $\sigma \in S_k$ as the *ranking* that assigns position $\sigma(i)$ to i . The notation $(\Omega, \mathcal{F}, \mathbb{P})$ refers to a generic probability space, and $\mathbb{1}_A$ is the indicator function of any event $A \in \mathcal{F}$. The set of all probability measures on $\mathbb{R}$ is denoted by $\mathcal{P}(\mathbb{R})$. If f is a density function on $\mathbb{R}$, we define its support $\text{Supp}(f)$ as the set of all real numbers $x \in \mathbb{R}$ such that $f(x) > 0$. For a non-decreasing function $g : \mathbb{R} \rightarrow \mathbb{R}$, the *generalized inverse* is defined by:

$$g^{-1}(y) = \sup\{x \in \mathbb{R} : g(x) \leq y\}, y \in \mathbb{R}.$$

When well-defined, the inverse coincides with the generalized inverse; moreover, the generalized inverse is always non-decreasing.

Linear extensions and graphs. A *strict partial order* $<$ on the finite set $\mathcal{S}$ is a homogeneous binary relation that is irreflexive, asymmetric, and transitive. A general result by Szpilrajn [12] states that, given any strict partial order $<$ on $\mathcal{S}$, there exists a *linear extension* $<^*$, i.e., a total order such that $v <^* w$ whenever $v < w$. A generic directed graph is denoted by $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex

set and $\mathcal{E}$ is the edge set. A (directed) edge is of the form $v \rightarrow w$, where v and w are distinct vertices. A *tournament* is a directed graph with exactly one directed edge between any pair of distinct vertices. When G is acyclic, its *transitive closure* is the strict partial order “ $<$ ” on the vertex set satisfying: for any two vertices v and w , $v < w$ if and only if there exists a path in G that goes from v to w .

Log-concavity. A probability distribution $\nu \in \mathcal{P}(\mathbb{R})$ is called *log-concave* if $\nu(\lambda A + (1 - \lambda)B) \geq \nu(A)^\lambda \nu(B)^{1-\lambda}$ whenever $\lambda A + (1 - \lambda)B \subseteq \mathbb{R}$ is measurable. A log-concave distribution on $\mathbb{R}$ is either a Dirac measure or absolutely continuous with respect to the Lebesgue measure [13]. In this case, its density and cumulative distribution function (CDF) are log-concave [14].

3 Problem Formulation

We consider a collection of $k \in \mathbb{N}$ items, where $k \geq 2$, identified with the set $[k]$. These are associated with a *utility vector* $\mathbf{u} \in \mathbb{R}^k$, where the entry u_i is the utility of item i . The *utility ranking* corresponding to $\mathbf{u}$ is the unique permutation $\sigma_{\mathbf{u}} \in \mathcal{S}_k$ satisfying:¹

$$\sigma_{\mathbf{u}}(i) < \sigma_{\mathbf{u}}(j) \iff u_i > u_j \vee (u_i = u_j \wedge i < j).$$

We address the problem of approximating the utility ranking based on noisy observations named *feedback signals*. Such signals arise from a RUM, where the noise-generating distribution is assumed to be centered and absolutely continuous with respect to the Lebesgue measure with a log-concave density f . The latter assumption implies several desirable properties for ranking models [7] and is standard in the literature [6]. We will work under a bounded-variance condition. For a fixed $V > 0$, $\mathcal{N}_V$ is the set of all absolutely continuous centered log-concave distributions ν satisfying:

(V) The variance of distribution ν is bounded above by V .

This acts as a reliability constraint on the feedback signals: while keeping the exact noise distribution unknown, it ensures that the feedback remains a reliable, albeit noisy, reflection of the utility relationships. Throughout the rest of this paper, V is fixed and known to the proposed algorithms. Therefore, we drop the subscript and write $\mathcal{N}$, whose elements are referred to as *noise distributions*.

3.1 Feedback-generation Process

Given $\mathbf{u} \in \mathbb{R}^k$ and $\nu \in \mathcal{N}$, we let $N_1, \dots, N_k$ be i.i.d. random variables drawn from ν , and consider the *random utility vector* $U = (U_1, \dots, U_k)$ where $U_i = u_i + N_i$ for every $i \in [k]$. For $t \in \mathbb{N}$, we let $U^t = (U_1^t, \dots, U_k^t)$ be an independent copy of U . We will denote the stochastic process $(U^t)_{t \in \mathbb{N}}$ by $\text{RUM}(\mathbf{u}, \nu)$. Feedback signals arise by filtering the sampled random utility vectors through a *feedback filter*, which can be of the following two types.

- The *full-ranking* (FR) filter is the function $\varphi_{FR} : \mathbb{R}^k \rightarrow \mathcal{S}_k$, where $\varphi_{FR}(\mathbf{x})$ is the unique permutation σ defined by: $\sigma(i) < \sigma(j)$ if and only if $x_i > x_j$ or $x_i = x_j$ and $i < j$.
- The *winner-only* (WO) filter is the function $\varphi_{WO} : \mathbb{R}^k \rightarrow [k]$ given by $\varphi_{WO}(\mathbf{x}) := \min\{i \in [k] : x_i \geq x_j \forall j \neq i\}$.²

For a fixed feedback filter φ , we let $\varphi_t = \varphi(U^t)$ for every $t \in \mathbb{N}$. The *feedback filtration* is defined as $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}}$, where $\mathcal{F}_t$ is the σ -algebra generated by $\varphi_1, \dots, \varphi_t$.

3.2 The PAC Objective

Estimation of the utility ranking can be seen as a byproduct of utility estimation [15, 16]. In contrast, our approach focuses on the ranking itself, thus avoiding the need for explicit utility reconstruction. The intuition is as follows: if items i and j are misranked, the *utility gap* $\Delta_{ij} := u_i - u_j$ must be small. This is formalized by the following definition.

¹If $u_i = u_j$, the tie is broken by ordering the two items according to their indices.

²The minimum in the definition of φ_{WO} is used for tie-breaking. However, when the noise distribution is absolutely continuous with respect to the Lebesgue measure, ties are a zero-probability event.

Definition 1 (ε -accuracy). Let $\mathbf{u} \in \mathbb{R}^k$ with utility ranking σ_u , and let $\varepsilon > 0$. Given $\sigma \in \mathcal{S}_k$, we say that σ is an ε -accurate representation of σ_u if the following holds for any two distinct i and j :³

$$(\sigma(i) - \sigma(j))(\sigma_u(i) - \sigma_u(j)) < 0 \implies |\Delta_{ij}| < \varepsilon. \quad (1)$$

Saha and Gopalan [17] use a similar yet distinct notion, where the gap between PL parameters is used instead of the utility gap. To define the PAC objective, we first define an *algorithm* as any pair of the form $\mathcal{A} = (T, \hat{\sigma})$, where $T \in \mathbb{N}$ is an almost-surely finite stopping time with respect to the feedback filtration $\mathbb{F}$, and $\hat{\sigma}$ is an $\mathcal{F}_T$ -measurable random permutation of $[k]$. We can now formally define the PAC requirement.

Definition 2 (PAC objective). Let $\mathcal{A} = (T, \hat{\sigma})$ be an algorithm that learns a ranking based on the sequential observations φ_t generated by a fixed unknown $\text{RUM}(\mathbf{u}, \nu)$ with feedback filter φ . Given $\delta \in (0, 1)$ and $\varepsilon > 0$, we say that $\mathcal{A}$ is (ε, δ) -PAC if:

$$\mathbb{P}[\hat{\sigma} \text{ is an } \varepsilon\text{-accurate representation of } \sigma_u] \geq 1 - \delta.$$

If $\mathcal{A}$ is an (ε, δ) -PAC algorithm, we define its *sample complexity* as the expected value $\mathbb{E}[T]$, where the expectation is taken with respect to the probability measure induced by $\text{RUM}(\mathbf{u}, \nu)$.

4 Learning a Ranking Based on Full-Ranking Feedback Signals

In this section, we present FR-VUR (Full-Ranking Variance-aware Utility Ranker), an algorithm for learning the utility ranking based on full-ranking feedback signals. These arise from a fixed $\text{RUM}(\mathbf{u}, \nu)$ through the filter $\varphi = \varphi_{FR}$, and the noise distribution ν satisfies Assumption (V).

4.1 The Full-Ranking Resolution Factor

Our algorithm treats each FR observation as a set of outcomes of pairwise comparisons, thus adopting an approach known as *rank-breaking* [16]. For $i \neq j$, we define the *pairwise-preference probability*:

$$P_{ij} := \mathbb{P}[U_i > U_j].$$

As shown by Lemma 5 in Appendix C, independence of the noise implies that $P_{ij} > 1/2$ if and only if $u_i > u_j$. For every $t \in \mathbb{N}$, we define the event $E_{ij}^t := \{\varphi_t(i) < \varphi_t(j)\}$, the set of outcomes for which item i is ranked before item j at time t . Then, for every integer $n \geq 2$, we let:

$$P_{ij}^n := \frac{1}{n} \sum_{t=1}^n \mathbb{1}_{E_{ij}^t} \quad \text{and} \quad V_{ij}^n := \frac{1}{n(n-1)} \sum_{1 \leq t_1 < t_2 \leq n} (\mathbb{1}_{E_{ij}^{t_1}} - \mathbb{1}_{E_{ij}^{t_2}})^2.$$

The former is an unbiased estimator of P_{ij} , while the latter is the (unbiased) sample variance. For a given $\delta \in (0, 1)$, we define the *full-ranking time- n confidence parameter*:

$$\delta_n := \frac{6}{k(k-1)n^2\pi^2} \cdot \delta, \quad \forall n \in \mathbb{N}. \quad (2)$$

Concentration bounds for these estimators are given by Lemma 6 in Appendix C.2, which is based on [18]. These include a standard concentration inequality for the sample variance plus a more sophisticated bound stemming from an empirical version of Bernstein's inequality.

To approximate the utility ranking, our algorithm will rely on the following intuition: when $u_i - u_j \geq \varepsilon$, the ε -accuracy target can be achieved if the error in the estimate of P_{ij} does not obscure the true preference relation between i and j . To make this precise, we first derive a mapping between utility gaps and pairwise-preference probabilities. To this end, let $i, j \in [k]$ be any two items such that $u_i \geq u_j$. We define the function $\xi : [0, \infty) \rightarrow [1/2, 1]$ as:

$$\xi(t) := \int_{\mathbb{R}} F(x+t) d\nu(x), \quad (3)$$

³The left-hand side of the implication is equivalent to saying that the relative position of items i and j implied by σ is opposite to the one induced by σ_u .

where F is the CDF of ν . Then, we can rewrite the pairwise-preference probability P_{ij} as:

$$P_{ij} = \int_{\mathbb{R}} F(x + u_i - u_j) d\nu(x) = \xi(u_i - u_j). \quad (4)$$

The functioning of our algorithm will be enforced through the *full-ranking resolution function* $\eta : [0, \infty) \rightarrow [0, 1]$, defined by:

$$\eta(\varepsilon) := 2\xi(\varepsilon) - 1.$$

Remarkably, if we define $\eta^*(\varepsilon) : [0, \infty) \rightarrow [0, 1]$ as:

$$\eta^*(\varepsilon) := \min \left\{ \frac{2}{5}, \frac{\varepsilon}{2\sqrt{6V}} \right\}, \quad (5)$$

then we can lower bound $\eta(\varepsilon) \geq \eta^*(\varepsilon)$ for every $\varepsilon > 0$. This fact is proven by Lemma 7 in Appendix C.3 and is crucial for our algorithm to be distribution-agnostic.

4.2 The FR-VUR Algorithm

Given $\delta \in (0, 1)$, for every $n \in \mathbb{N}$, we let δ_n be the time- n confidence given by Equation (2). Then, for a fixed $\varepsilon > 0$, we let $\eta^*(\varepsilon)$ be defined by Equation (5). We present the (ε, δ) -PAC algorithm $\mathcal{A} = (T, \hat{\sigma})$, named FR-VUR (Full-Ranking Variance-aware Utility-Ranker), which learns an ε -accurate approximation of the utility ranking based on full-ranking observations. As noted earlier, its reliability hinges on an estimation strategy that leverages the lower bound $\eta^*(\varepsilon)$ to bypass the need for parametric noise assumptions. Pseudocode is given in Appendix B.2.

Stopping rule. The algorithm's stopping condition is given by the stopping time:

$$T := \min \left\{ n \in \mathbb{N} : \sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} < \frac{\eta^*(\varepsilon)}{4}, \text{ for all } i \neq j \right\}. \quad (6)$$

Taking advantage of the concentration properties of our estimators, the stopping condition ensures that $|P_{ij}^T - P_{ij}| < \eta^*(\varepsilon)/4$ with high probability. Although this may appear as an *additive* bound, it is instead a *multiplicative* requirement⁴ upon observing that $\max(P_{ij}^T, P_{ji}^T) \geq 1/2$. This implies:

$$|P_{ij}^T - P_{ij}| = |P_{ji}^T - P_{ji}| \leq \max(P_{ij}^T, P_{ji}^T) \cdot \frac{\eta^*(\varepsilon)}{2}.$$

To define the algorithm's output, we need some additional devices. For every $n \in \mathbb{N}$, we define the *pairwise-preference graph* $G_n = ([k], \mathcal{E}_n)$ as the directed graph given by:

$$i \rightarrow j \in \mathcal{E}_n \iff P_{ij}^n > \frac{1}{2} + \frac{\eta^*(\varepsilon)}{4}.$$

The *utility graph* is the tournament $G_u = ([k], \mathcal{E}_u)$ where the edge set is defined as follows: for any two $i, j \in [k]$, $i \rightarrow j \in \mathcal{E}_u$ if and only if either $u_i > u_j$, or $u_i = u_j$ and $i < j$. In particular, G_u is transitive, meaning that $i \rightarrow j$ and $j \rightarrow \ell$ imply $i \rightarrow \ell$.

Output recommendation. Based on G_T , the $\mathcal{F}_T$ -measurable random ranking $\hat{\sigma}$ is given as follows:

- If G_T contains a cycle, we choose $\hat{\sigma} \in \mathcal{S}_k$ arbitrarily.
- If G_T is a DAG, we let G^* be its transitive closure, which corresponds to a strict partial order of its vertices. Then, we let $\hat{\sigma} \in \mathcal{S}_k$ be any linear extension of this partial order.

In particular, when G_T contains no cycle, $i \rightarrow j \in \mathcal{E}_T$ implies $\hat{\sigma}(i) < \hat{\sigma}(j)$. To prove the PAC property given by Definition 2, we first observe that, with probability at least $1 - \delta$, G_T is a subgraph of the utility graph, G_u . This fact follows from concentration properties of our estimators, and is proven by Lemma 8 in Appendix C.4. Second, we note that the ε -accuracy target is met if, for any two items i and j with $u_i - u_j \geq \varepsilon$, we have $\hat{\sigma}(i) < \hat{\sigma}(j)$. By definition, this is the case if $i \rightarrow j$ is an edge of G_T . As demonstrated by Lemma 9 in Appendix C.4, this occurs with probability at least $1 - \delta$, thus proving that $\mathcal{A}$ is (ε, δ) -PAC.

⁴We use the term *multiplicative* bound to denote the fact that the bound itself depends on the quantity to estimate.

4.2.1 Sample Complexity Upper Bound

The next result provides an upper bound on the sample complexity of our algorithm, proven in Appendix C.5.

Theorem 1 (Sample complexity upper bound for FR-VUR). *Let $\mathbf{u} \in \mathbb{R}^k$ and $\nu \in \mathcal{N}$ be such that $\text{RUM}(\mathbf{u}, \nu)$ satisfies Assumption (V), and let $\mathcal{A} = (T, \hat{\sigma})$ be the FR-VUR algorithm. Then, for every $\delta \in (0, 1)$ and $\varepsilon > 0$, $\mathcal{A}$ is (ε, δ) -PAC with sample complexity:*

$$\mathbb{E}[T] = \mathcal{O} \left(\frac{P(1-P)}{\eta^*(\varepsilon)^2} \log \left(\frac{P(1-P)}{\eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right) + \frac{1}{\eta^*(\varepsilon)} \log \left(\frac{1}{\eta^*(\varepsilon)} \cdot \frac{k}{\delta} \right) \right), \quad (7)$$

where P is defined so that $P(1-P) = \max_{i \neq j} P_{ij}(1-P_{ij})$.

When $P(1-P)$ is non-zero, the upper bound evaluates to $\tilde{\mathcal{O}}(P(1-P)/\eta^*(\varepsilon)^2)$. The $P(1-P)$ dependence reflects our algorithm's use of an empirical Bernstein-type stopping rule, which adapts to the intrinsic variance of the pairwise comparisons and accelerates convergence when outcomes are highly skewed. To illustrate this advantage, consider the extreme case with distinct utilities and a vanishing noise variance, implying $P(1-P) = 0$. In this scenario, the sample complexity reduces to $\tilde{\mathcal{O}}(1/\eta^*(\varepsilon))$. As the variance upper bound V goes to zero, thus matching the true variance, $\eta^*(\varepsilon)$ saturates at $2/5$, and this residual term reduces to the purely logarithmic cost $\tilde{\mathcal{O}}(\log(k/\delta))$. In contrast, a Hoeffding-based approach would rely on the global variance upper bound $P(1-P) \leq 1/4$. For $\sqrt{V} \geq 5\varepsilon/(4\sqrt{6})$, so that $\eta^*(\varepsilon) = \varepsilon/(2\sqrt{6V})$, this would result in the worst-case sample complexity:

$$\mathbb{E}[T] = \mathcal{O} \left(\frac{V}{\varepsilon^2} \log \left(\frac{V}{\varepsilon^2} \cdot \frac{k}{\delta} \right) \right). \quad (8)$$

For a fixed $V > 0$, the right-hand side becomes $\tilde{\mathcal{O}}(V/\varepsilon^2)$, where the term $1/\varepsilon^2$ reflects the ε -accuracy objective. The V scaling captures the intuition that, if the noise variance increases, the observed rankings become less reliable. As we will establish in Theorem 2, the V/ε^2 dependence is information-theoretically optimal.

4.3 Information-theoretic Lower Bound

In this part, we demonstrate an information-theoretic worst-case lower bound matching Equation (8). Precisely, we address the “small-gap regime” $\varepsilon < \sqrt{6V}/(2\pi)$, which is the most interesting case.⁵ To do so, we identify a class of “hard” instances of the learning task, defined by a Gumbel noise paired with a utility vector with a utility gap of exactly ε . Our proof, provided in Appendix C.6, employs a change-of-measure argument following [19]. Let $\mathcal{A}^* = (T^*, \hat{\sigma}^*)$ be any ranking algorithm based on FR feedback signals, and assume that $\mathcal{A}^*$ is (ε, δ) -PAC for any $\delta \in (0, 1)$.

Theorem 2 (Lower bound for full-ranking feedback). *Let $\delta \in (0, 1/4]$. Then, there exists a noise-utility pair $(\nu, \mathbf{u})$ satisfying Assumption (V) such that the sample complexity of algorithm $\mathcal{A}^* = (T^*, \hat{\sigma}^*)$ satisfies:*

$$\mathbb{E}[T^*] = \Omega \left(\frac{V}{\varepsilon^2} \log \left(\frac{1}{\delta} \right) \right),$$

where the expectation is taken with respect to the probability measure induced by $\text{RUM}(\mathbf{u}, \nu)$.

This result establishes the V/ε^2 scaling as a fundamental learning cost of the ranking problem. In particular, Theorem 2 shows that the exact dependence on the noise variance is linear.

5 Learning a Ranking Based on Winner-Only Feedback Signals

In this section, we shift our focus to WO feedback, which amounts to choosing the feedback filter $\varphi = \varphi_{\text{WO}}$. As before, U^t is the random utility vector drawn at time t from a fixed $\text{RUM}(\mathbf{u}, \nu)$.

⁵The threshold $\sqrt{6V}/(2\pi) \approx 0.39\sqrt{V}$ is the one under which the Gumbel instances used in the proof satisfy $\varepsilon/\beta < 1/2$, where β is the scale parameter. Its precise value is immaterial: our results are asymptotic, thus justifying the restriction to the small-gap regime.

We introduce WO-VUR (Winner-Only Variance-aware Utility Ranker) as a direct counterpart to the FR-VUR algorithm. This naming is intentional: while being tailored to two distinct feedback types, our algorithms only differ in what kind of preference probabilities they estimate. In the FR case, these were *pairwise-preference probabilities*, while here, they will be *winning probabilities*. To emphasize this continuity, we deliberately maintain the notation for all key quantities (e.g., η , η^* , and $\mathcal{A}$) to formally signal that the fundamental mechanics remain invariant across both feedback models. As in the full-ranking case, the noise distribution satisfies Assumption (V).

5.1 Assumptions on the Noise Distribution

If an item has a vanishingly small probability of being the winner, the (ε, δ) -PAC objective becomes unattainable. Indeed, as we will show in our lower-bound analysis formalized by Theorem 4, the sample complexity of any PAC algorithm is inversely proportional to the minimum winning probability, $P_{\min}$. Therefore, to ensure well-posedness, we explicitly require that every item have a non-zero probability of being the winner. To do so, we fix a non-trivial interval (a, b) where possibly $a = -\infty$ and $b = \infty$, and assume that the following holds:

(WO1) The noise distribution ν has a density f whose support contains the interval (a, b) , and $u_i - u_j < b - a$ for any two $i, j \in [k]$.

We anticipate that our proposed algorithm does not require exact knowledge of a and b . In the following, we denote by F the CDF of ν and define the Reversed Hazard Rate (RHR) $r : \text{Supp}(f) \rightarrow [0, \infty)$ as the function $r(x) := f(x)/F(x)$. For every $i \in [k]$, its *winning probability* is defined as:

$$P_i := \mathbb{P}[U_i > U_j, \forall j \in [k] \setminus \{i\}].$$

For any two indices $i, j \in [k]$ with $u_i \geq u_j$, the winning probabilities have the following monotonicity property: $P_i > P_j$ if and only if $u_i > u_j$ (see Lemma 12 in Appendix D.3). In general, the winning probability of an item depends on the magnitude of all utility gaps. However, in the PL model induced by a Gumbel noise with scale parameter $s > 0$, the utility gap between any two i and j is a function of the winning-probability ratio,⁶ that is:

$$u_i - u_j = s \log \left(\frac{P_i}{P_j} \right). \quad (9)$$

Exploiting this, Saha and Gopalan [17] use estimates of the probability ratios to approximate the ranking induced by the PL parameters. We aim to generalize this approach to a broader class of RUMs: rather than PL parameters, we establish a functional relation between winning-probability ratios and utilities through the sufficient condition on the noise distribution given by:⁷

(WO2) For every $t > 0$, the RHR ratio $r(x+t)/r(x)$ is a non-increasing function of x on $\text{Supp}(f)$.

Examples of log-concave distributions satisfying this assumption are reported in Table 1 in Appendix B, including the Gaussian and logistic distributions. To derive the desired relation under (WO2), consider two items i and j . Assuming that $u_i > u_j$, we have:

$$P_i - P_j = \int_{\text{Supp}(f)} \prod_{\ell \neq i} F(x + u_i - u_\ell) \left[1 - \frac{r(x + u_i - u_j)}{r(x)} \right] f(x) dx. \quad (10)$$

The product of CDFs is non-decreasing, while Assumption (WO2) ensures that the function in the brackets is also non-decreasing. Thus, a direct application of Harris' inequality [21] yields:

$$P_i - P_j \geq \int_{\text{Supp}(f)} \left(1 - \frac{r(x + u_i - u_j)}{r(x)} \right) f(x) dx \cdot P_i. \quad (11)$$

Now, let us define the functions $\zeta : (0, b - a) \rightarrow (0, 1)$ and $\xi : (0, b - a) \rightarrow (1, \infty)$ by setting:

$$\zeta(t) := 1 - \int_{\text{Supp}(f)} \frac{r(x+t)}{r(x)} f(x) dx \quad \text{and} \quad \xi(t) := \frac{1}{1 - \zeta(t)}. \quad (12)$$

⁶This property is typically referred to as Independence of Irrelevant Alternatives (IIA) [20].

⁷For $x \in \text{Supp}(f)$ and $t > 0$ such that $x + t$ lies outside of the support of f , we extend the definition by setting $r(x + t) := 0$.

Rearranging Equation (11) yields $P_i/P_j \geq \xi(u_i - u_j)$. By log-concavity, ξ is non-decreasing (Lemma 13 in Appendix D.4); taking the (generalized) inverse, we obtain the desired relation:

$$u_i - u_j \leq \xi^{-1} \left(\frac{P_i}{P_j} \right). \quad (13)$$

We define the subclass $\mathcal{N}_V(a, b) \subseteq \mathcal{N}$ as the set of all noise distributions for which Assumptions (WO1), (WO2), and (V) are satisfied. In Appendix A.1, we discuss the strengths and limitations implied by Assumption (WO2) by comparing our setting with the one considered by Saha and Gopalan [11].

5.2 The Winner-Only Resolution Factor

For every $i \in [k]$ and $t \in \mathbb{N}$, we define the event $E_i^t := \{\varphi_t = i\}$, consisting of those outcomes for which i is the winner at time t . For every $n \in \mathbb{N}$ and each $i \in [k]$, we let:

$$P_i^n := \frac{1}{n} \sum_{t=1}^n \mathbb{1}_{E_i^t} \text{ and } V_i^n := \frac{1}{n(n-1)} \sum_{1 \leq t_1 < t_2 \leq n} (\mathbb{1}_{E_i^{t_1}} - \mathbb{1}_{E_i^{t_2}})^2.$$

The former is an unbiased estimator for P_i , while the latter is the corresponding (unbiased) sample variance. Given a fixed $\delta \in (0, 1)$, we tune the *winner-only time- n confidence parameter* to be:

$$\delta_n := \frac{3}{\pi^2 n^2 k} \cdot \delta, \quad \forall n \in \mathbb{N}. \quad (14)$$

Concentration inequalities for the above estimators are given by Lemma 11 in Appendix D.2, which is based on the results of [18].

The FR resolution function defined in Section 4 was used in the design of FR-VUR to enforce a multiplicative constraint on the pairwise-preference probability estimates. As we aim to utilize the same rationale in the current setting, we overload notation and define the *winner-only resolution function* $\eta : (0, b - a) \rightarrow (0, 1)$ as:

$$\eta(t) := \frac{\xi(t) - 1}{\xi(t) + 1}, \quad (15)$$

where ξ is given by Equation (12). Remarkably, we can provide a lower bound on the resolution function which is uniform across all noise distributions for which Assumptions (WO1), (WO2), and (V) are satisfied. We define the function $\eta^* : (0, b - a) \rightarrow (0, 1)$ as:

$$\eta^*(\varepsilon) := \frac{1}{8e^3} \min \left(1, \frac{\varepsilon}{\sqrt{2V}} \right). \quad (16)$$

Then, $\eta(\varepsilon) \geq \eta^*(\varepsilon)$ for every $\varepsilon \in (0, b - a)$, as proven by Lemma 14 in Appendix D.5.

5.3 The WO-VUR Algorithm

Given $\delta \in (0, 1)$, for every $n \in \mathbb{N}$, we let δ_n be the winner-only time- n confidence parameter given by Equation (14). Then, for a fixed accuracy $\varepsilon \in (0, b - a)$, we let $\eta^*(\varepsilon)$ be the resolution-function lower bound given by Equation (16). We present and analyze the (ε, δ) -PAC algorithm WO-VUR $\mathcal{A} = (T, \hat{\sigma})$, where the overload of notation is intentional and aims to emphasize the continuity with the FR setting. Rather than estimating pairwise-preference probabilities, WO-VUR delivers its final recommendation based on the winning-probability estimates, computed as empirical frequencies. Remarkably, the noise distribution is allowed to be *any* element of $\mathcal{N}_V(a, b)$, thus making our algorithm distribution-agnostic. Pseudocode is given in Appendix B.2.

Stopping rule. The stopping condition is defined through the $\mathbb{F}$ -stopping time

$$T := \min \left\{ n \in \mathbb{N} : \sqrt{\frac{2V_i^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} < \frac{\eta^*(\varepsilon)}{2} \cdot P_i^n, \text{ for all } i \in [k] \right\}. \quad (17)$$

Given the estimates $P_1^n, \dots, P_k^n$ at time n , we let σ_n be the unique permutation of $[k]$ satisfying:

$$\sigma_n(i) < \sigma_n(j) \iff P_i^n > P_j^n \vee (P_i^n = P_j^n \text{ and } i < j).$$

Output recommendation. The recommendation is defined as the $\mathcal{F}_T$ -measurable permutation $\hat{\sigma} := \sigma_T$.

In order to show that $\mathcal{A}$ fulfills Definition 2, we suppose that:

$$|P_i^T - P_i| \leq \sqrt{\frac{2V_i^T \log(4/\delta_T)}{T}} + \frac{7 \log(4/\delta_T)}{3(T-1)}, \quad \forall i \in [k].$$

As shown by Lemma 11 in Appendix D.2, this is true with probability at least $1 - \delta$. Using the definition of T and the fact that $\eta^*(\varepsilon) < 1$, we have $|P_i^T - P_i| < P_i^T/2$ for every $i \in [k]$, highlighting the *multiplicative* nature of the bound. Now, suppose that items i and j are misranked at time T , thus implying $(P_i - P_j)(P_i^T - P_j^T) \leq 0$. W.l.o.g., we can assume that $u_i > u_j$, so we have:

$$0 < P_i - P_j < \frac{\eta^*(\varepsilon)}{2} (P_i^T + P_j^T) \leq \eta^*(\varepsilon) \cdot (P_i + P_j).$$

Rearranging these terms yields:

$$\frac{P_i}{P_j} < \frac{1 + \eta^*(\varepsilon)}{1 - \eta^*(\varepsilon)} \leq \frac{1 + \eta(\varepsilon)}{1 - \eta(\varepsilon)} = \xi(\varepsilon),$$

where $\eta(\varepsilon)$ is the resolution factor of the true noise distribution. On the other hand, since $u_i > u_j$, the Harris-based bound derived above gives $\xi(u_i - u_j) \leq P_i/P_j$, so that $\xi(u_i - u_j) < \xi(\varepsilon)$. As ξ is non-decreasing by Lemma 13, we conclude that $u_i - u_j < \varepsilon$. In other words, we have shown that: with probability at least $1 - \delta$, if i and j are misranked, then their absolute utility gap is bounded by ε . Hence, algorithm $\mathcal{A}$ is (ε, δ) -PAC.

5.3.1 Sample Complexity Upper Bound

The next result derives an upper bound on the sample complexity of the WO-VUR algorithm. Formal derivations are given in Appendix D.6.

Theorem 3 (Sample complexity upper bound for WO-VUR). *Let $\mathbf{u} \in \mathbb{R}^k$ and $\nu \in \mathcal{N}_V(a, b)$, so that $\text{RUM}(\mathbf{u}, \nu)$ satisfies Assumptions (WO1), (WO2), and (V), and let $\mathcal{A} = (T, \hat{\sigma})$ be the WO-VUR algorithm. Then, for every $\delta \in (0, 1)$ and $\varepsilon \in (0, b - a)$, $\mathcal{A}$ is (ε, δ) -PAC with sample complexity:*

$$\mathbb{E}[T] = \mathcal{O} \left(\frac{1}{P_{\min} \eta^*(\varepsilon)^2} \log \left(\frac{1}{P_{\min} \eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right) \right), \quad (18)$$

where we defined $P_{\min} := \min_{i \in [k]} P_i$. For $\varepsilon \leq \sqrt{2V}$, using the explicit expression for $\eta^*(\varepsilon)$ given by Equation (16), the sample complexity upper bound can be rewritten in terms of ε as:

$$\mathbb{E}[T] = \mathcal{O} \left(\frac{V}{P_{\min} \varepsilon^2} \log \left(\frac{V}{P_{\min} \varepsilon^2} \cdot \frac{k}{\delta} \right) \right). \quad (19)$$

The term $P_{\min}$ shows a fundamental limit of the WO setting. The utility gap between i and j is estimated based on the empirical frequencies of their wins. Thus, if these items are rarely observed as the winner, reliable estimation will require a large number of samples. Our Bernstein-type stopping rule results in the term $1/P_{\min}$. In contrast, a Hoeffding-based approach would rely on $P_i(1 - P_i) \leq 1/4$, thus resulting in the quadratic term⁸ $1/P_{\min}^2$. The $1/\varepsilon^2$ scaling given by Equation (19) reflects the ε -accuracy target, mirroring the full-ranking upper bound given by Theorem 1. Similarly, the dependence on V is the same as in the FR setting, thus confirming the linear dependence on the reliability of the feedback signals. The $\log(k)$ dependence results in a cost of lower order. As we shall see, Theorem 4 establishes the $V/(\varepsilon^2 P_{\min})$ scaling as information-theoretically optimal.

5.4 Information-theoretic Lower Bound

If $P_{\min} \approx 0$, observing the item with the least utility would take an arbitrary number of steps, so any algorithm would be ineffective. To obtain a meaningful lower bound, we analyze the sample complexity over all RUMs with $k \geq 3$ items and minimum winning probability $P_{\min} \in (0, 1/k)$.⁹

⁸This can be seen from Equation (46) in Appendix D.6.

⁹We deliberately exclude $P_{\min} = 1/k$. Indeed, this forces the trivial case $u_1 = \dots = u_k$. We also exclude $k = 2$: in that case $P_1 + P_2 = 1$, so the minimum winning probability is a function of the single utility gap, and $P_{\min}$ cannot be varied independently of ε .

Let $\mathcal{A}^* = (T^*, \hat{\sigma}^*)$ be an algorithm that outputs a ranking based on WO feedback signals, which we assume to be (ε, δ) -PAC for any $\delta \in (0, 1)$ and $\varepsilon \in (0, b - a)$. Formal derivations of the lower bound are given in Appendix D.7. The following result establishes the lower bound.

Theorem 4 (Lower bound for winner-only feedback). *Let $k \geq 3$, $P_{\min} \in (0, 1/k)$, and $\delta \in (0, 1/4]$ be fixed. Then, there exists a noise-utility pair $(\nu, \mathbf{u})$ satisfying Assumptions (WO1), (WO2) and (V) such that the sample complexity of algorithm $\mathcal{A}^* = (T^*, \hat{\sigma}^*)$ satisfies:*

$$\mathbb{E}[T^*] = \Omega\left(\frac{V}{P_{\min} \varepsilon^2} \log\left(\frac{1}{\delta}\right)\right),$$

where the expectation is taken with respect to the probability measure induced by $RUM(\mathbf{u}, \nu)$.

Our lower bound contrasts with the $\tilde{O}(k/\varepsilon^2)$ sample complexity derived by Saha and Gopalan [17] for the PL model. The discrepancy arises because their framework seeks an ε -optimal ranking, penalizing misclassification based on the difference between the PL parameters e^{u_ℓ} . If i and j have utilities $u_i, u_j \ll 0$, then $e^{u_i}, e^{u_j} \approx 0$. Thus, the distance metric used in [17] forgives the misranking of these two. This effectively circumvents the fundamental informational bottleneck of $P_{\min}$ established in Theorem 4, allowing for the claim of full ranking recovery without actually resolving the relative order of items with a low winning probability.

6 Open Challenges

We addressed the problem of recovering utility rankings from two types of feedback. A natural next step is to explore intermediate feedback models, such as top- m [17] or subset preferences [1]. Future research should investigate whether the variance-adaptive stopping rules central to the VURs can be generalized to accommodate other feedback structures while retaining order optimality. Our analysis of the winner-only feedback in Section 5 relied on Assumption (WO2) to derive a uniform lower bound on the resolution factor as a function of the variance upper bound and the accuracy parameter. While this assumption encompasses the Gaussian family and other common log-concave distributions, it excludes cases like the uniform or Laplace distributions. A significant research endeavor would be to identify alternative and, possibly, more relaxed conditions that preserve the existence of a strictly positive resolution factor across a broader class of noise distributions.

Acknowledgments

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

References

- [1] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 1475–1483, 2014.
- [2] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 4299–4307, 2017.
- [3] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.
- [4] Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.
- [5] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley New York, 1959.
- [6] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

- [7] Douglas E. Critchlow, Michael A. Fligner, and Joseph S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, 35(3):294–318, 1991.
- [8] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- [9] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.
- [10] Róbert Busa-Fekete, Balázs Szörényi, and Eyke Hüllermeier. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, pages 1701–1707, 2014.
- [11] 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.
- [12] Edward Szpilrajn. Sur l’extension de l’ordre partiel. *Fundamenta Mathematicae*, 16(1):386–389, 1930.
- [13] Christer Borell. Convex set functions in d -space. *Periodica Mathematica Hungarica*, 6(2): 111–136, 1975.
- [14] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.
- [15] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 2474–2482, 2012.
- [16] 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.
- [17] 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, 2019.
- [18] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Annual Conference on Learning Theory (COLT)*, 2009.
- [19] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1): 1–42, 2016.
- [20] Paramesh Ray. Independence of irrelevant alternatives. *Econometrica*, 41(5):987–991, 1973.
- [21] Theodore E. Harris. A lower bound for the critical probability in a certain percolation process. *Mathematical Proceedings of the Cambridge Philosophical Society*, 56(1):13–20, 1960.
- [22] Christian Wirth, Riad Akrou, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017.
- [23] Daniel McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, 53(3):S13–S29, 1980.
- [24] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126, 2010.
- [25] Yuxin Chen and Changho Suh. Spectral MLE: Top-K rank aggregation from pairwise comparisons. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 371–380. PMLR, 2015.

- [26] Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.
- [27] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.
- [28] Masrour Zoghi, Shimon Whiteson, Remi 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.
- [29] Hossein Azari Soufiani, William Z. Chen, David C. Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 2706–2714, 2013.
- [30] Tyler Lu and Craig Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 145–152. Omnipress, 2011.
- [31] Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of Mallows. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 1071–1079. PMLR, 2014.
- [32] Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in Markov decision processes with gap-dependent sample complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1253–1263, 2020.
- [33] Roxana A. Ion, Chris A. J. Klaassen, and Edwin R. van den Heuvel. Sharp inequalities of Bienaymé-Chebyshev and Gauss type for possibly asymmetric intervals around the mean. *arXiv preprint arXiv:2208.08813*, 2022.
- [34] Iosif Pinelis. Small deviations of real log-concave random variable. MathOverflow, 2023. URL <https://mathoverflow.net/q/458283>.

A Related Works

Our work relates to the existing literature along three primary dimensions: the way of generating feedback, the definition of the objectives and metrics, and the types of feedback used to achieve the target.

A standard paradigm in preference learning is inferring latent properties from human judgments, such as distinguishing “good” from “bad” trajectories in RLHF [2, 22] or predicting user choices in recommender systems [23, 24]. While the PL model [4, 5] is the ubiquitous choice for these tasks, it is a specific instance of the broader Random Utility Model (RUM) framework [3]. Our work adopts the RUM framework with log-concave noise, which provides a unifying lens for these models. Several works implicitly utilize log-concavity by considering the PL model [15–17, 25, 26], while others do this explicitly [1]. Remarkably, log-concavity of the noise ensures several desirable properties of the implied ranking distribution [7]. In this paper, we exploit this regularity to derive uniform lower bounds on the resolution factors.

The definition of “preference recovery” varies significantly across the literature. Multi-armed and dueling bandit frameworks often focus on identifying a single best item [11, 27, 28], a paradigm known as best-item identification. Many approaches focus on statistical parameter estimation, with particular attention devoted to the estimation of PL parameters, typically measured by the L_2 -norm [1, 6, 29]. When the goal is a ranking, metrics like Kendall-tau distance [9] or maximum rank difference [10] are standard. When the utility ranking is the target, it is possible to bypass utilities entirely by assuming distributions over permutations, such as the Mallows model [30, 31]. In contrast to these approaches, we operate within the PAC framework [8], defining success through the notion of ε -accuracy. This objective bridges the gap between pure parameter estimation and exact ordinal recovery, by incorporating utility information into the target ranking in an intuitive manner.

Finally, we distinguish works by the density of the feedback signal. In line with the present paper, some works consider winner-only feedback signals [11, 17], which are generalized through the notion of top- m feedback. For full-ranking feedback, a common technique is rank-breaking, which decomposes permutations into pairwise comparisons modeled via a comparison graph [15, 16, 26].

A.1 Additional Comments on the Winner-Only Setting

We compare our approach to the Minimum Advantage Ratio introduced by Saha and Gopalan [11]. While both frameworks seek to lower-bound the winning-probability ratio P_i/P_j in terms of the utility gap Δ_{ij} , the methodologies and objectives differ in two key aspects. First, Saha and Gopalan [11] focus on identifying the ε -best item, whereas we target the recovery of the full utility ranking. Under WO feedback, the latter becomes a significantly more challenging problem. This is due to the informational bottleneck represented by the minimum winning probability $P_{\min}$, as established by our lower bound analysis formalized through Theorem 4. Second, their approach relies on a distribution-specific constant c , which is derived via a so-called “variational lower bound” formalized through Lemma 2 in [11]. This constant must be computed for each noise model and provided to the algorithm as a parameter.

In contrast, our approach leverages the structural regularity of log-concave distributions. Owing to Assumption (WO2), applying Harris’ inequality allows us to derive a lower bound on the resolution factor that is distribution-agnostic and expressed solely as a function of the variance upper bound V and the accuracy parameter ε . We acknowledge that our assumption is more restrictive than the variational requirement in [11], as it excludes distributions like the uniform or Laplace. However, it enables the design of algorithms that do not require exact knowledge of the noise distribution, effectively capturing the “cost of agnosticism” through the universal dependence on V .

B Variance-aware Utility Rankers

B.1 Comparison Table

Below is a comparison table listing several examples of common log-concave distributions. For each of them, we provide explicit conditions ensuring that the mean is zero and Assumption (V) is satisfied. If these two are met, the given distribution is allowed by the FR-VUR algorithm presented in Section 4. In other words, provided that the noise distribution ν of the feedback-generation process

$\text{RUM}(\mathbf{u}, \nu)$ is any of those for which the above two conditions are verified, the theoretical guarantees derived for FR-VUR will hold.

Similarly, we say, for each of the listed distributions, if Assumption (WO2) is also met (see Appendix D.1). In this case, the distribution is allowed by the WO-VUR algorithm as well, and the sample complexity analysis of Section 5 applies. We recall that successful learning in the WO setting requires that the largest utility gap is smaller than $b - a$, where $(a, b) \subseteq \text{Supp}(f)$, and f is the density of the noise distribution. Therefore, it is implicitly assumed that, given an absolutely continuous noise distribution ν , the utility vector is chosen so that the above condition is met.

Noise distribution	Centeredness	Bounded-variance (V)	FR-VUR	WO-VUR
Gumbel(μ, s)	$\mu = -s\gamma$	$s \leq \sqrt{6V}/\pi$	✓	✓
Normal(μ, σ^2)	$\mu = 0$	$\sigma \leq \sqrt{V}$	✓	✓
Logistic(μ, s)	$\mu = 0$	$s \leq \sqrt{3V}/\pi$	✓	✓
Laplace(μ, s)	$\mu = 0$	$s \leq \sqrt{V/2}$	✓	✗
Uniform(c, d)	$c = -d$	$d - c \leq \sqrt{12V}$	✓	✗

Table 1: The parameter γ in the centeredness condition of the Gumbel distribution is the Euler-Mascheroni constant, whose value is approximately 0.58. All distributions are allowed by the FR-VUR algorithm, while the uniform and the Laplace distributions do not exhibit the required monotonicity of the RHR ratio needed for successful learning in the WO setting.

B.2 Pseudocode for Variance-aware Utility-Rankers

Below is the pseudocode for the VUR algorithms presented in Sections 4 and 5, respectively.

Algorithm 1 Full-Ranking Variance-aware Utility Ranker

Input: Items $[k]$, confidence $\delta \in (0, 1)$, accuracy $\varepsilon > 0$, variance bound $V > 0$

```

1: Initialize:  $n \leftarrow 0$ ,  $G \leftarrow ([k], \emptyset)$ ,  $\eta^*(\varepsilon) \leftarrow \min(\frac{2}{5}, \frac{\varepsilon}{2\sqrt{6V}})$ 
2:    $P_{ij} \leftarrow 0$ ,  $V_{ij} \leftarrow 0$ , and  $I_{ij} \leftarrow \eta^*(\varepsilon)/4$  for every  $i \neq j$  in  $[k]$ 
3: repeat
4:    $n \leftarrow n + 1$ 
5:   Observe full ranking  $\sigma_n$  from  $\text{RUM}(\mathbf{u}, \nu)$  and set  $\delta_n \leftarrow \frac{6\delta}{k(k-1)n^2\pi^2}$ 
6:   for each pair  $i \neq j$  do
7:      $P_{ij} \leftarrow \frac{n-1}{n}P_{ij} + \frac{1}{n}\mathbb{1}\{i \text{ is ranked before } j \text{ in } \sigma_n\}$ 
8:     if  $n > 1$  then
9:        $V_{ij} \leftarrow \frac{n}{n-1}P_{ij}(1 - P_{ij})$ 
10:       $I_{ij} \leftarrow \sqrt{\frac{2V_{ij} \ln(4/\delta_n)}{n}} + \frac{7 \ln(4/\delta_n)}{3(n-1)}$ 
11:     end if
12:   end for
13: until  $I_{ij} < \eta^*(\varepsilon)/4$  for every  $i \neq j$ 
14: for  $i, j$  such that  $1 \leq i < j \leq k$  do
15:   if  $P_{ij} > 1/2 + \eta^*(\varepsilon)/4$  then
16:     Add  $i \rightarrow j$  to  $G$ 
17:   else if  $P_{ji} > 1/2 + \eta^*(\varepsilon)/4$  then
18:     Add  $j \rightarrow i$  to  $G$ 
19:   end if
20: end for
21: if  $G$  is acyclic then
22:    $G \leftarrow$  transitive closure of  $G$  and  $\hat{\sigma} \leftarrow$  topological sort of  $G$ 
23: else
24:    $\hat{\sigma} \leftarrow$  arbitrary permutation of  $[k]$ 
25: end if
Output:  $\hat{\sigma}$ 

```

Algorithm 2 Winner-Only Variance-aware Utility Ranker

Input: Items $[k]$, confidence $\delta \in (0, 1)$, accuracy $\varepsilon \in (0, b - a)$, variance bound $V > 0$

```
1: Initialize:  $n \leftarrow 0$ ,  $\eta^*(\varepsilon) \leftarrow \min\left(\frac{1}{8e^3}, \frac{\varepsilon}{8e^3\sqrt{2V}}\right)$ 
2:    $P_i \leftarrow 1$ ,  $V_i \leftarrow 1/4$ , and  $I_i \leftarrow \eta^*(\varepsilon)/2$  for every  $i \in [k]$ 
3: repeat
4:    $n \leftarrow n + 1$ 
5:   Observe winner  $w_n$  from  $\text{RUM}(\mathbf{u}, \nu)$  and set  $\delta_n \leftarrow \frac{3\delta}{k\pi^2 n^2}$ 
6:   for each  $i \in [k]$  do
7:      $P_i \leftarrow \frac{n-1}{n}P_i + \frac{1}{n}\mathbb{1}\{w_n = i\}$ 
8:     if  $n > 1$  then
9:        $V_i \leftarrow \frac{n}{n-1}P_i(1 - P_i)$ 
10:       $I_i \leftarrow \sqrt{\frac{2V_i \ln(4/\delta_n)}{n}} + \frac{7 \ln(4/\delta_n)}{3(n-1)}$ 
11:    end if
12:  end for
13: until  $I_i < P_i \cdot \eta^*(\varepsilon)/2$  for every  $i \in [k]$ 
14:  $\hat{\sigma} \leftarrow$  unique permutation  $\sigma \in \mathcal{S}_k$  such that  $\sigma(i) < \sigma(j) \iff P_i > P_j \vee (P_i = P_j \text{ and } i < j)$ 
Output:  $\hat{\sigma}$ 
```

C Technical Results and Proofs for Section 4

C.1 The Pairwise-preference Probabilities

We begin by showing that the pairwise-preference probabilities P_{ij} behave consistently with the underlying utilities.

Lemma 5. *Let $P_{ij} = \mathbb{P}[U_i > U_j]$ be the pairwise-preference probability defined in Section 4. Then, $P_{ij} > 1/2$ if and only if $u_i > u_j$.*

Proof. We denote by f and F the density and the CDF of ν , respectively. Additionally, we let (a, b) be an interval fully contained in $\text{Supp}(f)$.

If $u_i = u_j$, then, by symmetry, we have $P_{ij} = 1/2$. If $u_i > u_j$, then

$$\begin{aligned} P_{ij} &= \int_{\mathbb{R}} \mathbb{P}[U_j < U_i \mid U_i = x] f(x - u_i) dx \\ &= \int_{\mathbb{R}} F(x - u_j) f(x - u_i) dx \\ &\stackrel{(i)}{=} \int_{\mathbb{R}} F(y + u_i - u_j) f(y) dy, \end{aligned}$$

where we used the change of variable $y = x - u_i$ for (i). Now consider the function $g : [0, \infty) \rightarrow \mathbb{R}$ defined by

$$g(t) := \int_{\mathbb{R}} F(y + t) f(y) dy.$$

Taking the first derivative, we find

$$g'(t) = \int_{\mathbb{R}} f(y + t) f(y) dy.$$

Since $(a, b) \subset \text{Supp}(f)$, for $t < b - a$, there exists a subinterval of $\text{Supp}(f)$ where both $f(y + t)$ and $f(y)$ are strictly positive. Thus,

$$g'(t) > 0 \quad \text{for every } t \in [0, b - a).$$

Moreover, $g'(t) \geq 0$ for every $t \geq 0$, so g is non-decreasing on the whole half-line. Together with the fact that $g(0) = 1/2$, this implies $g(t) > 1/2$ for every $t \in (0, b - a)$ and, by monotonicity, $g(t) \geq g(t_0) > 1/2$ for every $t \geq b - a$ and any fixed $t_0 \in (0, b - a)$. Since $P_{ij} = g(u_i - u_j)$ and $u_i - u_j > 0$, we conclude that $P_{ij} > 1/2$.

To prove the converse implication, suppose $P_{ij} > 1/2$. As noted above, we cannot have $u_i = u_j$, so either $u_i < u_j$ or $u_i > u_j$. The former would imply $P_{ji} > 1/2$, but since $P_{ij} + P_{ji} = 1$, this yields a contradiction. $\square$

C.2 Concentration Bounds

The following result provides anytime-concentration bounds for the unbiased estimators introduced in Section 4. Our inequalities are based on [18].

Lemma 6. Fix $\delta \in (0, 1)$ and let δ_n be the time- n confidence parameter defined by (2). Then, with probability at least $1 - \delta$, the following hold simultaneously for every pair (i, j) with $i \neq j$ and every $n \in \mathbb{N} \setminus \{1\}$:

$$|P_{ij} - P_{ij}^n| \leq \sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)}, \quad (20)$$

$$\left| \sqrt{V_{ij}^n} - \sqrt{P_{ij}(1 - P_{ij})} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}}. \quad (21)$$

Proof. For a fixed pair (i, j) with $1 \leq i < j \leq k$ and $n \geq 2$, we define the events

$$A_{ij}^n := \left\{ |P_{ij} - P_{ij}^n| \leq \sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} \right\}$$

and

$$B_{ij}^n := \left\{ \left| \sqrt{V_{ij}^n} - \sqrt{P_{ij}(1 - P_{ij})} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}} \right\}.$$

By Theorem 4 and Theorem 10 in [18], we have $\mathbb{P}[A_{ij}^n] \geq 1 - \delta_n$ and $\mathbb{P}[B_{ij}^n] \geq 1 - \delta_n$. By taking a union bound, we obtain

$$\begin{aligned} \mathbb{P} \left[\bigcap_{n \in \mathbb{N}} \left(\bigcap_{1 \leq i < j \leq k} (A_{ij}^n \cap B_{ij}^n) \right) \right] &= 1 - \mathbb{P} \left[\bigcup_{n \in \mathbb{N}} \left(\bigcup_{1 \leq i < j \leq k} (A_{ij}^n)^c \cup (B_{ij}^n)^c \right) \right] \\ &\geq 1 - \frac{k(k-1)}{2} \sum_{n=1}^{\infty} 2\delta_n \\ &= 1 - \frac{6\delta}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \\ &\stackrel{(i)}{=} 1 - \delta, \end{aligned}$$

where (i) follows from the well-known fact that $\sum_{n \in \mathbb{N}} 1/n^2 = \pi^2/6$. To conclude, it suffices to observe that, for any two distinct items i and j , one has

$$|P_{ij} - P_{ij}^n| = |P_{ji} - P_{ji}^n| \quad \text{and} \quad V_{ij}^n = V_{ji}^n.$$

$\square$

C.3 Uniform Lower Bound on the Resolution Function

In this section, we formally derive a uniform lower bound on the FR resolution function that holds across all noise distributions satisfying Assumption (V).

Lemma 7. Let $\nu \in \mathcal{N}$ be a noise distribution satisfying the bounded-variance condition (V). Then

$$\eta(t) \geq \eta^*(t) \quad \text{for every } t \in [0, \infty),$$

where $\eta^*(\varepsilon)$ is defined by (5).

Proof. Since $\nu \in \mathcal{N}$, it is absolutely continuous w.r.t. the Lebesgue measure; we denote by f its density. For $t \geq 0$, we define the function

$$\zeta(t) := 1 - \frac{\int_{\mathbb{R}} F(x-t)f(x) dx}{\int_{\mathbb{R}} F(x+t)f(x) dx}. \quad (22)$$

Using the definition of η , we have

$$\eta(t) = \frac{\zeta(t)}{2 - \zeta(t)} \geq \frac{\zeta(t)}{2}, \quad (23)$$

where we used the fact that ζ is $[0, 1]$ -valued. If we define X and Y as any two independent random variables distributed according to ν , we can rewrite

$$\begin{aligned} \int_{\mathbb{R}} F(x-t)f(x) dx &= \int_{\mathbb{R}} \mathbb{P}[Y < x-t] f(x) dx \\ &= \int_{\mathbb{R}} \mathbb{P}[Y < X-t \mid X=x] f(x) dx \\ &= \mathbb{P}[Y - X < -t]. \end{aligned}$$

Similarly,

$$\int_{\mathbb{R}} F(x+t)f(x) dx = \mathbb{P}[Y - X < t].$$

If we set $Z := Y - X$, then (22) implies

$$\zeta(t) = 1 - \frac{\mathbb{P}[Y - X < -t]}{\mathbb{P}[Y - X < t]} = 1 - \frac{H(-t)}{H(t)},$$

where H is the CDF of Z .

By this, a lower bound on η can be obtained by upper-bounding the ratio $H(-t)/H(t)$. The random variable Z is log-concave and symmetric, so its mean, mode, and median coincide and are equal to zero. In particular, $H(-t) = \mathbb{P}[Z \geq t]$ for every $t \geq 0$. Denoting by σ the standard deviation of Z , Lemma 18 implies

$$H(-t) \leq \begin{cases} \frac{1}{2} - \frac{t}{2\sqrt{3}\sigma} & \text{if } 0 \leq t < 2\sigma/\sqrt{3} \\ \frac{2\sigma^2}{9t^2} & \text{if } t \geq 2\sigma/\sqrt{3} \end{cases}.$$

We now consider two cases.

Case 1: If $t \in (0, 2\sigma/\sqrt{3})$, then

$$\frac{H(-t)}{1 - H(-t)} \leq \frac{\sqrt{3}\sigma - t}{\sqrt{3}\sigma + t}.$$

Using the fact that $t \leq \sqrt{3}\sigma$ and $\sigma^2 \leq 2V$, we obtain

$$1 - \frac{H(-t)}{1 - H(-t)} \geq \frac{t}{\sqrt{6V}}.$$

Case 2: If $t \geq 2\sigma/\sqrt{3}$, then

$$\frac{2\sigma^2}{9t^2} \leq \frac{1}{6}.$$

Thus, we have

$$1 - \frac{H(-t)}{1 - H(-t)} \geq \frac{4}{5}.$$

Using (23), we conclude that

$$\eta(t) \geq \min\left(\frac{2}{5}, \frac{t}{2\sqrt{6V}}\right) \quad \forall t \in [0, \infty).$$

□

C.4 The PAC Property

In this section, we prove two results showing that the FR-VUR algorithm presented in Section 4 is (ε, δ) -PAC.

Lemma 8. *With probability at least $1 - \delta$, the pairwise-preference graph G_T is a subgraph of the utility graph.*

Proof. With probability at least $1 - \delta$, the inequality given by (20) holds for all $i \neq j$ and $n \geq 2$. In particular, taking $n = T$, this further implies

$$|P_{ij}^T - P_{ij}| < \frac{\eta^*(\varepsilon)}{4} \quad \forall i \neq j.$$

If $i \rightarrow j \in \mathcal{E}_T$, i.e., there is an edge from i to j in G_T , then

$$P_{ij} = P_{ij}^T + (P_{ij} - P_{ij}^T) > \frac{1}{2} + \frac{\eta^*(\varepsilon)}{4} - \frac{\eta^*(\varepsilon)}{4} = \frac{1}{2}.$$

Since $P_{ij} > 1/2$ if and only if $u_i > u_j$ (see Lemma 5), we conclude that $i \rightarrow j$ is an edge of G_u . $\square$

Lemma 9. *With probability at least $1 - \delta$, if i and j are any two items such that $u_i - u_j \geq \varepsilon$, then $\hat{\sigma}(i) < \hat{\sigma}(j)$.*

Proof. With probability at least $1 - \delta$, inequality (20) holds for all $i \neq j$ and $n \geq 2$. In this regime, the condition of Lemma 8 is satisfied, so, taking $n = T$,

$$|P_{ij}^T - P_{ij}| < \frac{\eta^*(\varepsilon)}{4} \quad \forall i \neq j.$$

Now let i and j be any two items such that $|u_i - u_j| \geq \varepsilon$, say $u_i \geq u_j + \varepsilon$. If $\eta(\varepsilon)$ denotes the value of the resolution function computed for the true noise distribution, then Lemma 7 ensures $\eta(\varepsilon) \geq \eta^*(\varepsilon)$. By (4) and the monotonicity of ξ , $P_{ij} \geq \xi(\varepsilon)$. Using the definition of $\eta(\varepsilon)$, we have

$$P_{ij} \geq \frac{1}{2} + \frac{\eta(\varepsilon)}{2} \geq \frac{1}{2} + \frac{\eta^*(\varepsilon)}{2}.$$

Hence,

$$P_{ij}^T = P_{ij} + (P_{ij}^T - P_{ij}) > \frac{1}{2} + \frac{\eta^*(\varepsilon)}{2} - \frac{\eta^*(\varepsilon)}{4} = \frac{1}{2} + \frac{\eta^*(\varepsilon)}{4}.$$

This implies $i \rightarrow j \in \mathcal{E}_T$. Since G_T is a DAG, from the definition of $\hat{\sigma}$, it follows that $\hat{\sigma}(i) < \hat{\sigma}(j)$. $\square$

C.5 Sample Complexity Upper Bound

In this section, we prove Theorem 1. To do so, we first define $\mathcal{E}$ as the set of outcomes for which the following hold simultaneously for every $n \in \mathbb{N} \setminus \{1\}$ and any two distinct i and j :

$$|P_{ij}^n - P_{ij}| \leq \sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)}, \quad (24)$$

$$\left| \sqrt{V_{ij}^n} - \sqrt{P_{ij}(1 - P_{ij})} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}}, \quad (25)$$

where δ_n is the one given by (2).

If (25) holds, then multiplying both sides of $\sqrt{V_{ij}^n} \leq \sqrt{P_{ij}(1 - P_{ij})} + \sqrt{2 \log(2/\delta_n)/(n-1)}$ by $\sqrt{2 \log(4/\delta_n)/n}$ gives

$$\sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} \leq \sqrt{\frac{2P_{ij}(1 - P_{ij}) \log(4/\delta_n)}{n}} + \frac{2 \log(4/\delta_n)}{n-1},$$

where we used $\log(2/\delta_n) \leq \log(4/\delta_n)$ and $n(n-1) \geq (n-1)^2$. Adding $7 \log(4/\delta_n)/(3(n-1))$ to both sides, we obtain

$$\sqrt{\frac{2V_{ij}^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} \leq \sqrt{\frac{2P_{ij}(1-P_{ij}) \log(4/\delta_n)}{n}} + \frac{13 \log(4/\delta_n)}{3(n-1)}, \quad (26)$$

whose left-hand side is exactly the quantity appearing in the stopping rule (6). Thus, if $\mathcal{E}$ occurs, the stopping rule is satisfied if the following hold simultaneously for every $i \neq j$:

$$\sqrt{\frac{2P_{ij}(1-P_{ij}) \log(4/\delta_n)}{n}} < \frac{\eta^*(\varepsilon)}{8}, \quad (27)$$

$$\frac{13 \log(4/\delta_n)}{3(n-1)} < \frac{\eta^*(\varepsilon)}{8}. \quad (28)$$

Condition (27) holds for every $i \neq j$ if the stronger requirement below is satisfied:

$$n > \max_{i \neq j} (P_{ij}(1-P_{ij})) \cdot \frac{128 \log(4/\delta_n)}{\eta^*(\varepsilon)^2}.$$

In particular, it holds for n_0 , which we define as the minimum positive integer n such that

$$n \geq \max \left\{ 2, p(1-p) \cdot \frac{128 \log(4/\delta_n)}{\eta^*(\varepsilon)^2} \right\},$$

where $p \in [0, 1]$ is chosen so that

$$p(1-p) = \max_{i \neq j} (P_{ij}(1-P_{ij})).$$

Using the explicit expression of δ_n given by (2), we have

$$n_0 \leq p(1-p) \cdot \frac{128 \log(4/\delta_{n_0})}{\eta^*(\varepsilon)^2} + 1 \leq \max \left(p(1-p) \cdot \frac{256}{\eta^*(\varepsilon)^2} \log \left(\frac{2kn_0\pi}{\sqrt{6\delta}} \right), 2 \right).$$

By Lemma 17, there exists a universal constant C_0 such that

$$n_0 \leq \max \left(\frac{C_0 p(1-p)}{\eta^*(\varepsilon)^2} \log \left(\frac{C_0 p(1-p)}{\eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right), 2 \right), \quad (29)$$

where $\log(0)$ is understood as $-\infty$.

Now we consider the condition given by (28), and define n_1 as the minimum positive integer n such that

$$n \geq \left\lceil \frac{104 \log(4/\delta_n)}{3\eta^*(\varepsilon)} + 1 \right\rceil.$$

Such an integer exists, since the right-hand side grows logarithmically in n , while the left-hand side grows linearly. By minimality, the inequality above fails at $n_1 - 1$, and since δ_n is non-increasing in n , this yields

$$n_1 \leq \frac{104 \log(4/\delta_{n_1})}{3\eta^*(\varepsilon)} + 2.$$

It follows that

$$n_1 \leq \max \left(\frac{208 \log(4/\delta_{n_1})}{3\eta^*(\varepsilon)}, 4 \right).$$

Using the definition of δ_n , Lemma 17 ensures that there exists a constant $C_1 > 0$ such that

$$n_1 \leq \max \left(\frac{C_1}{\eta^*(\varepsilon)} \log \left(\frac{C_1}{\eta^*(\varepsilon)} \cdot \frac{k}{\delta} \right), 4 \right). \quad (30)$$

If we let $n_2 := n_0 + n_1$, then for every $n \geq n_2$, the conditions given by (27) and (28) are both satisfied.

Now, the sample complexity can be decomposed as

$$\mathbb{E}[T] = \mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}}] + \mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}^c}]. \quad (31)$$

Using the tail-sum formula for the expectation, we can rewrite

$$\mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}^c}] = \sum_{n=1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n].$$

If $n > n_2$, then necessarily $\mathbb{1}_{\mathcal{E}^c} = 1$. Indeed, we defined n_2 in such a way that, if $\mathcal{E}$ occurs, then the stopping condition is met for all $n \geq n_2 + 1$. Therefore, we have

$$\sum_{n=n_2+1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n] = \sum_{n=n_2+1}^{\infty} \mathbb{P}[T \geq n].$$

If $T \geq n$, then the stopping condition was not satisfied at time $n - 1$. Thus, at least one of (24) and (25) is violated for at least one pair (i, j) with $i \neq j$. By a union bound over the $k(k - 1)/2$ unordered pairs and the two conditions, this occurs with probability at most $k(k - 1)\delta_{n-1}$. Hence, using the definition of δ_n given by (2),

$$\sum_{n=n_2+1}^{\infty} \mathbb{P}[T \geq n] \leq k(k - 1) \sum_{n=n_2}^{\infty} \delta_n = \sum_{n=n_2}^{\infty} \frac{6\delta}{\pi^2 n^2} \leq C_2,$$

for a universal constant $C_2 > 0$. It follows that

$$\sum_{n=1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n] \leq n_2 \cdot \mathbb{P}[\mathcal{E}^c] + C_2.$$

Using the decomposition given by (31), this implies

$$\mathbb{E}[T] \leq n_2 + C_2.$$

To prove the claim, we use the upper bounds given by (29) and (30) together with the definition of $\eta^*(\varepsilon)$. $\square$

C.6 Information-theoretic Lower Bound

We derive an information-theoretic lower bound for the ranking problem in the FR setting, thus proving Theorem 2. To this end, we fix a variance upper bound $V > 0$ and consider the centered Gumbel distribution ν with scale parameter $\beta = \sqrt{6V}/\pi$ and location parameter $\mu = -\beta\gamma$, where γ is the Euler–Mascheroni constant. This ensures that ν is centered and that its variance is equal to V . We fix an accuracy level ε in the small-gap regime $\varepsilon < \sqrt{6V}/(2\pi)$, which is equivalent to requiring $\tilde{\varepsilon} := \varepsilon/\beta \in (0, 1/2)$. Then, consider the utility vectors $\mathbf{u}$ and $\mathbf{v}$ given by

$$u_1 = v_2 = \varepsilon$$

and

$$u_\ell = v_m = 0 \quad \text{for every } \ell \in [k] \setminus \{1\} \text{ and } m \in [k] \setminus \{2\}.$$

Since the location parameter shifts all sampled utilities by the same constant, the induced ranking law does not depend on μ , and we can consider $\text{RUM}(\mathbf{u}, \nu)$ and $\text{RUM}(\mathbf{v}, \nu)$. These are seen as two instances of a multi-armed bandit problem, where the number of arms is $n = 1$. At each round t , the learner plays the only available arm. The observed reward is the feedback signal, consisting of a full ranking of items $1, \dots, k$. The *ranking distributions* P and Q correspond to the reward distributions, and are given by

$$P(\sigma) = \mathbb{P}[U_{(i)} = U_{\sigma^{-1}(i)} \ \forall i \in [k] \mid \text{RUM}(\mathbf{u}, \nu)]$$

and

$$Q(\sigma) = \mathbb{P}[U_{(i)} = U_{\sigma^{-1}(i)} \ \forall i \in [k] \mid \text{RUM}(\mathbf{v}, \nu)],$$

for every $\sigma \in \mathcal{S}_k$.

We denote by $\mathbb{P}_u$ and $\mathbb{P}_v$ the probability measures on $(\Omega, \mathcal{F})$ induced by $\text{RUM}(\mathbf{u}, \nu)$ and $\text{RUM}(\mathbf{v}, \nu)$, respectively. The next result provides an upper bound on the KL divergence between the two ranking distributions.

Lemma 10. *Let P and Q be the ranking distributions introduced above. Then, the KL divergence is upper-bounded as*

$$D_{\text{KL}}(P \parallel Q) \leq \frac{\pi^2 \varepsilon^2}{V}.$$

Proof. In what follows, we abuse notation and use $P[\cdot]$ and $\mathbb{P}_u[\cdot]$ to denote the same probability measure.

Consider any sample from $\text{RUM}(\mathbf{u}, \nu)$. Let $\sigma \in \mathcal{S}_k$ and denote by i_j the index of the j -th order statistic. Then, we can factorize the probability of observing σ as

$$P(\sigma) = \mathbb{P}_u[i_1 = \sigma^{-1}(1)] \cdot \prod_{t=2}^{k-1} \mathbb{P}_u[i_t = \sigma^{-1}(t) \mid i_j = \sigma^{-1}(j) \text{ for } j \leq t-1]. \quad (32)$$

A similar factorization holds for Q , and these allow us to think of the sampling process as a sequential one: first, the top item i_1 is chosen among all items $1, \dots, k$; then, item i_2 is picked among all the remaining items; and so on, until only one item is left and is chosen as the last element of the ranking.

According to this interpretation, if $S_t \subseteq [k]$ is the set of items left after the first $t-1$ items have been chosen, we define

$$P[i \mid S_t] := \frac{e^{u_i/\beta}}{\sum_{\ell \in S_t} e^{u_\ell/\beta}} \quad \text{for all } i \in S_t.$$

Now, using (32), we can write

$$D_{\text{KL}}(P \parallel Q) = \sum_{t=1}^{k-1} \mathbb{E}_{\nu, \mathbf{u}}[D_{\text{KL}}(P[\cdot \mid S_t] \parallel Q[\cdot \mid S_t])], \quad (33)$$

where the expectation is taken with respect to $\mathbb{P}_u$ over all pairs (i, S_t) where necessarily $i \in S_t$.

For simplicity, we introduce the parameter $\tilde{\varepsilon} := \varepsilon/\beta$, which allows us to rewrite

$$u_1/\beta = \tilde{\varepsilon} \text{ and } u_\ell/\beta = 0 \text{ for all } \ell > 1,$$

as well as

$$v_2/\beta = \tilde{\varepsilon} \text{ and } v_\ell/\beta = 0 \text{ for all } \ell \neq 2.$$

Recall that $\tilde{\varepsilon} \in (0, 1/2)$, which is exactly the small-gap condition $\varepsilon < \sqrt{6V}/(2\pi)$ assumed above. To estimate the KL divergence between P and Q , we consider three different cases. In the following analysis, we fix any $t \in [k-1]$.

Case 1: $\{1, 2\} \cap S_t = \emptyset$.

In this case, 1 and 2 have already been picked as the t -th decision round starts, so the conditional winning probabilities given S_t induced by $\mathbf{u}$ and $\mathbf{v}$ are the same:

$$D_{\text{KL}}(P[\cdot \mid S_t] \parallel Q[\cdot \mid S_t]) = 0.$$

Case 2: $1, 2 \in S_t$.

In this case, we have

$$\begin{aligned} D_{\text{KL}}(P[\cdot \mid S_t] \parallel Q[\cdot \mid S_t]) &= \sum_{i_t \in S_t} P[i_t \mid S_t] \log \frac{P[i_t \mid S_t]}{Q[i_t \mid S_t]} \\ &= \sum_{i_t \in S_t \setminus \{1, 2\}} P[i_t \mid S_t] \log \frac{P[i_t \mid S_t]}{Q[i_t \mid S_t]} + \sum_{\ell=1, 2} P[\ell \mid S_t] \log \frac{P[\ell \mid S_t]}{Q[\ell \mid S_t]} \\ &\stackrel{(i)}{=} \frac{e^{\tilde{\varepsilon}}}{|S_t| - 1 + e^{\tilde{\varepsilon}}} \log(e^{\tilde{\varepsilon}}) + \frac{1}{|S_t| - 1 + e^{\tilde{\varepsilon}}} \log(e^{-\tilde{\varepsilon}}) \\ &= \frac{(e^{\tilde{\varepsilon}} - 1)\tilde{\varepsilon}}{|S_t| - 1 + e^{\tilde{\varepsilon}}} \\ &\stackrel{(ii)}{\leq} \frac{\tilde{\varepsilon}(\tilde{\varepsilon} + \tilde{\varepsilon}^2)}{|S_t|}, \end{aligned}$$

where (i) follows from the fact that the first sum is equal to zero and (ii) is due to $e^x - 1 \leq x + x^2$ for $x \in (0, 1/2)$.

Case 3: $|S_t \cap \{1, 2\}| = 1$.

In this case, we consider two further subcases corresponding to: (a) $1 \in S_t$ and $2 \notin S_t$, and (b) $2 \in S_t$ and $1 \notin S_t$.

Consider (a) first. Setting $W := \sum_{j \in S_t} e^{u_j/\beta}$ and $W' := \sum_{j \in S_t} e^{v_j/\beta}$, we obtain

$$\begin{aligned} D_{\text{KL}}(P[\cdot | S_t] || Q[\cdot | S_t]) &= \sum_{j \in S_t} P[i_t = j | S_t] \log \frac{P[i_t = j | S_t]}{Q[i_t = j | S_t]} \\ &= \frac{e^{\tilde{\varepsilon}}}{W} \log \frac{W' \cdot e^{\tilde{\varepsilon}}}{W} + \sum_{j \in S_t \setminus \{1\}} \frac{1}{W} \log \frac{W'}{W} \\ &= \frac{e^{\tilde{\varepsilon}}}{W} \left(\tilde{\varepsilon} + \log \frac{W'}{W} \right) + \sum_{j \in S_t \setminus \{1\}} \frac{1}{W} \log \frac{W'}{W} \\ &= \frac{e^{\tilde{\varepsilon}} \tilde{\varepsilon}}{W} + \log \frac{W'}{W}. \end{aligned}$$

To bound the term with the logarithm, observe that $\log(1/y) \leq 1/y - 1$ for every $y > 0$. Thus, we have

$$\log \frac{W'}{W} = \log \frac{|S_t|}{W} \leq \frac{|S_t|}{W} - 1 = \frac{1 - e^{\tilde{\varepsilon}}}{W},$$

where we used the fact that $W' = |S_t|$ and $W = |S_t| - 1 + e^{\tilde{\varepsilon}}$. It follows that

$$\begin{aligned} D_{\text{KL}}(P[\cdot | S_t] || Q[\cdot | S_t]) &\leq \frac{e^{\tilde{\varepsilon}} \tilde{\varepsilon}}{W} + \frac{1 - e^{\tilde{\varepsilon}}}{W} \\ &= \frac{e^{\tilde{\varepsilon}} \tilde{\varepsilon} + 1 - e^{\tilde{\varepsilon}}}{W} \\ &\leq \frac{\tilde{\varepsilon}^2}{|S_t|}, \end{aligned}$$

using the inequality $W \geq |S_t|$ and observing that

$$e^x x + 1 - e^x = e^x(x - 1) + 1 \leq x^2 \text{ for } x \in (0, 1/2).$$

The calculations for scenario (b) are similar and yield

$$\begin{aligned} D_{\text{KL}}(P[\cdot | S_t] || Q[\cdot | S_t]) &= \log \frac{W'}{W} - \frac{\tilde{\varepsilon}}{W} \\ &= \log \left(1 + \frac{e^{\tilde{\varepsilon}} - 1}{|S_t|} \right) - \frac{\tilde{\varepsilon}}{|S_t|} \\ &\stackrel{(i)}{\leq} \frac{e^{\tilde{\varepsilon}} - 1}{|S_t|} - \frac{\tilde{\varepsilon}}{|S_t|} \\ &= \frac{e^{\tilde{\varepsilon}} - 1 - \tilde{\varepsilon}}{|S_t|} \\ &\stackrel{(ii)}{\leq} \frac{\tilde{\varepsilon}^2 + \tilde{\varepsilon} - \tilde{\varepsilon}}{|S_t|} \\ &= \frac{\tilde{\varepsilon}^2}{|S_t|}, \end{aligned} \tag{34}$$

where (i) is due to $\log(1 + x) \leq x$ for $x > 0$, whereas for (ii), we relied again on the fact that $e^{\tilde{\varepsilon}} - 1 \leq \tilde{\varepsilon} + \tilde{\varepsilon}^2$.

Now, observe that $\mathbb{E}_{\nu, \mathbf{u}} [\mathbf{D}_{\text{KL}} (P[\cdot | S_t] || Q[\cdot | S_t])]$ can be computed by averaging the KL divergence inside the expectation over all choices of S_t , where weights are induced by P . Using the bounds derived in the three cases above and the fact that $\tilde{\varepsilon}^3 \leq \tilde{\varepsilon}^2$, we obtain

$$\begin{aligned} \mathbb{E}_{\nu, \mathbf{u}} [\mathbf{D}_{\text{KL}} (P[\cdot | S_t] || Q[\cdot | S_t])] &\leq P[1, 2 \in S_t] \frac{\tilde{\varepsilon}^2 + \tilde{\varepsilon}^3}{|S_t|} + P[|\{1, 2\} \cap S_t| = 1] \frac{\tilde{\varepsilon}^2}{|S_t|} \\ &\leq (1 - P[1, 2 \notin S_t]) \frac{2\tilde{\varepsilon}^2}{|S_t|}, \end{aligned} \quad (35)$$

where, with abuse of notation, $P(E)$ denotes the probability of event E under the PL model associated with $\mathbf{u}$.

Next, let us observe that

$$P[1 \in S_t] = \prod_{\ell=1}^{t-1} \frac{k - \ell}{k - \ell + e^{\tilde{\varepsilon}}},$$

where $k - \ell + e^{\tilde{\varepsilon}} \geq k - \ell + 1$ for every $\ell \in [t-1]$. This implies

$$P[1 \in S_t] \leq \prod_{\ell=1}^{t-1} \frac{k - \ell}{k - \ell + 1} = \frac{k - t + 1}{k} = \frac{|S_t|}{k}.$$

On the other hand, one has

$$P[2 \in S_t] \leq \prod_{\ell=1}^{t-1} \frac{k - (\ell + 1) + e^{\tilde{\varepsilon}}}{k - \ell + e^{\tilde{\varepsilon}}} = \frac{|S_t| + e^{\tilde{\varepsilon}} - 1}{k + e^{\tilde{\varepsilon}} - 1}.$$

By using a union bound, we have

$$(1 - P[1, 2 \notin S_t]) = P[1 \in S_t \text{ or } 2 \in S_t] \leq \frac{|S_t|}{k} + \frac{|S_t| + e^{\tilde{\varepsilon}} - 1}{k + e^{\tilde{\varepsilon}} - 1}.$$

Since $\tilde{\varepsilon} > 0$, $k + e^{\tilde{\varepsilon}} - 1 > k$; moreover, $|S_t| \geq 1$. Therefore, we can bound

$$\frac{|S_t| + e^{\tilde{\varepsilon}} - 1}{k + e^{\tilde{\varepsilon}} - 1} \leq \frac{|S_t|(1 + e^{\tilde{\varepsilon}} - 1)}{k} = \frac{e^{\tilde{\varepsilon}} |S_t|}{k}.$$

From this, we obtain

$$P[1 \in S_t \text{ or } 2 \in S_t] \leq \frac{|S_t|(1 + e^{\tilde{\varepsilon}})}{k} \leq \frac{3|S_t|}{k},$$

where we used the fact that $e^x + 1 \leq 3$ for every $x \in (0, 1/2)$. Together with (35), this allows us to derive our final bound

$$\mathbb{E}_{\nu, \mathbf{u}} [\mathbf{D}_{\text{KL}} (P[\cdot | S_t] || Q[\cdot | S_t])] \leq \frac{6\tilde{\varepsilon}^2}{k}.$$

Combining this with expression (33) yields

$$\begin{aligned} \mathbf{D}_{\text{KL}} (P || Q) &\leq \sum_{t=1}^{k-1} \frac{6\tilde{\varepsilon}^2}{k} \\ &= \frac{6(k-1)\tilde{\varepsilon}^2}{k} \\ &\leq 6\tilde{\varepsilon}^2. \end{aligned}$$

By recalling that $\tilde{\varepsilon} = \varepsilon/\beta$, we finally get

$$\mathbf{D}_{\text{KL}} (P || Q) \leq \frac{6\varepsilon^2}{\beta^2} = \frac{\pi^2 \varepsilon^2}{V}.$$

□

Let $\mathcal{A}^* = (T^*, \hat{\sigma}^*)$ be an (ε, δ) -PAC algorithm, and define the $\mathcal{F}_{T^*}$ -measurable event

$$E = \{\hat{\sigma}^*(1) < \hat{\sigma}^*(2)\}.$$

Then

$$\mathbb{P}_u[E] \geq 1 - \delta \quad \text{and} \quad \mathbb{P}_v[E] \leq \delta.$$

Let us assume that $\delta \in (0, 1/4]$, as the confidence parameter is typically chosen to be very close to zero. For $x \geq 1 - \delta$ and $y \leq \delta$, the binary relative entropy satisfies

$$d(x, y) \geq d(1 - \delta, \delta) \geq \log \left(\frac{1}{3\delta} \right).$$

Using this, a direct application of Lemma 16 yields

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] \mathbf{D}_{\text{KL}}(P \| Q) \geq \log \left(\frac{1}{3\delta} \right).$$

Hence, by Lemma 10, we obtain

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] \geq \frac{V}{\pi^2 \varepsilon^2} \log \left(\frac{1}{3\delta} \right).$$

In conclusion, since $\log(1/(3\delta)) \geq \frac{1}{5} \log(1/\delta)$ for every $\delta \in (0, 1/4]$, we have

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] = \Omega \left(\frac{V}{\varepsilon^2} \log \left(\frac{1}{\delta} \right) \right).$$

D Technical Results and Proofs for Section 5

D.1 Example Distributions

In this section, we briefly present some log-concave distributions that satisfy Assumptions (WO1) and (WO2). For each of them, we give an explicit condition on the scale parameter so that Assumption (V) is met, and formally demonstrate the monotonicity of the RHR ratio. In particular, the support of each of the distributions below is the real line $\mathbb{R}$, meaning that (WO1) is verified for any choice of $a < b$. A summary of the properties presented here can be found in Table 1.

Gumbel. We let s be the scale parameter and set the location parameter μ so that ν is centered. The variance is given by $\pi^2 s^2/6$, so we require

$$s \leq \frac{\sqrt{6V}}{\pi}$$

to ensure that (V) is fulfilled. Next, observe that the function in (WO2) can be rewritten as $r(x + t)/r(x)$, where r is the reversed hazard rate. Explicitly, we have

$$f(x) = \frac{1}{s} \exp\left(-\frac{x - \mu}{s} - e^{-(x - \mu)/s}\right) \quad \text{and} \quad F(x) = \exp(-e^{-(x - \mu)/s}).$$

A direct calculation yields

$$\frac{r(x + t)}{r(x)} = e^{-t/s},$$

which is constant for every fixed $t > 0$. □

Normal. In order to fulfill (V), we let $\sigma \leq \sqrt{V}$. Following standard conventions, we denote by ϕ and Φ the density and CDF of a standard normal random variable. Thus, if X is distributed according to $\nu = \text{Normal}(0, \sigma^2)$, we will have

$$f(x) = \frac{1}{\sigma} \phi\left(\frac{x}{\sigma}\right) \quad \text{and} \quad F(x) = \Phi\left(\frac{x}{\sigma}\right).$$

Therefore, the reversed hazard rate is given by

$$r(x) = \frac{1}{\sigma} \cdot \frac{\phi(x/\sigma)}{\Phi(x/\sigma)},$$

whose first derivative is

$$r'(x) = -\frac{r(x)}{\sigma^2} [x + \sigma^2 \cdot r(x)].$$

Now, for a fixed $t > 0$, consider the function

$$h(x) := 1 - \frac{r(x+t)}{r(x)}, \quad x \in \mathbb{R}.$$

Proving that (WO2) holds is equivalent to showing that h is non-decreasing for every $t > 0$. To this end, we compute the first derivative

$$h'(x) = \frac{1}{\sigma^2} \cdot \frac{r(x+t)}{r(x)} [t + \sigma^2(r(x+t) - r(x))], \quad (36)$$

where

$$r(x+t) - r(x) = \int_x^{x+t} r'(y) dy. \quad (37)$$

It is known from the properties of truncated normal random variables that the conditional variance of X given $X < x$ can be computed as

$$\text{Var}[X | X < x] = \sigma^2 (1 + \sigma^2 r'(x)).$$

This quantity is strictly greater than zero for every $x \in \mathbb{R}$, so $r'(x) > -1/\sigma^2$. Using (37), we obtain

$$r(x+t) - r(x) > -\frac{t}{\sigma^2}.$$

Plugging this back into (36) yields

$$h'(x) > \frac{1}{\sigma^2} \cdot \frac{r(x+t)}{r(x)} \left[t - \sigma^2 \frac{t}{\sigma^2} \right] = 0.$$

In particular, h is increasing, and so (WO2) is satisfied. $\square$

Logistic. We set the location parameter equal to zero, while the scale parameter s is chosen so that $s \leq \sqrt{3V}/\pi$. This ensures that the variance bound given by (V) is satisfied. The PDF and CDF of ν are given respectively by

$$f(x) = \frac{e^{-x/s}}{s(1 + e^{-x/s})^2} \quad \text{and} \quad F(x) = \frac{1}{1 + e^{-x/s}}.$$

Thus, for every $t > 0$, by direct calculations, we obtain

$$\frac{d}{dx} \left(\frac{r(x+t)}{r(x)} \right) = \frac{e^{x/s} - e^{(x+t)/s}}{s(e^{(x+t)/s} + 1)^2}.$$

The latter is negative due to $x+t > x$, so the function

$$x \mapsto 1 - \frac{r(x+t)}{r(x)}$$

is non-decreasing for every $t > 0$. This is equivalent to saying that (WO2) holds. $\square$

D.2 Concentration Bounds

The following result provides anytime-concentration bounds for the unbiased estimators introduced in Section 5. Our inequalities are based on [18]. The proof of the lemma below is analogous to the one of Lemma 6.

Lemma 11. Fix $\delta \in (0, 1)$ and let δ_n be as defined by (14). Then, with probability at least $1 - \delta$, the following hold simultaneously for all items $i \in [k]$ and every $n \in \mathbb{N} \setminus \{1\}$:

$$|P_i - P_i^n| \leq \sqrt{\frac{2V_i^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)}, \quad (38)$$

$$\left| \sqrt{V_i^n} - \sqrt{P_i(1 - P_i)} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}}. \quad (39)$$

Proof. For fixed $i \in [k]$ and $n \geq 2$, we define the events

$$A_i^n := \left\{ |P_i - P_i^n| \leq \sqrt{\frac{2V_i^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} \right\}$$

and

$$B_i^n := \left\{ \left| \sqrt{V_i^n} - \sqrt{P_i(1-P_i)} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}} \right\}.$$

By Theorem 4 and Theorem 10 in [18], we have $\mathbb{P}[A_i^n] \geq 1 - \delta_n$ and $\mathbb{P}[B_i^n] \geq 1 - \delta_n$. By taking a union bound, we obtain

$$\begin{aligned} \mathbb{P} \left[\bigcap_{n \in \mathbb{N}} \left(\bigcap_{i \in [k]} (A_i^n \cap B_i^n) \right) \right] &= 1 - \mathbb{P} \left[\bigcup_{n \in \mathbb{N}} \left(\bigcup_{i \in [k]} (A_i^n)^c \cup (B_i^n)^c \right) \right] \\ &\geq 1 - k \sum_{n=1}^{\infty} 2\delta_n \\ &= 1 - \frac{6\delta}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \\ &\stackrel{(i)}{=} 1 - \delta, \end{aligned}$$

where (i) follows from the well-known fact that $\sum_{n \in \mathbb{N}} 1/n^2 = \pi^2/6$. $\square$

D.3 Monotonicity of the Winning Probabilities

In the FR case, we showed that the pairwise preference probability P_{ij} is larger than $1/2$ if and only if $u_i > u_j$. In the WO case, we rely on an analogous result for the winning probabilities.

Lemma 12. *Let $(a, b) \subseteq \text{Supp}(f)$, where $-\infty \leq a < b \leq \infty$, and let $\mathbf{u} \in \mathbb{R}^k$ be a utility vector such that $|u_l - u_m| < b - a$ for every $l, m \in [k]$. Then, for any two items $i, j \in [k]$ with $u_i \geq u_j$, their winning probabilities satisfy:*

$$P_i > P_j \iff u_i > u_j.$$

Proof. The direct implication can be proven by contrapositive. Indeed, if $u_i = u_j$, then the winning probabilities P_i and P_j takes on the same expression, so $P_i = P_j$.

For the converse implication, let us assume $u_i > u_j$. Explicitly, the winning probability of item i can be expressed as

$$\begin{aligned} P_i &= \int_{\mathbb{R}} \mathbb{P}[U_i > U_l \forall l \in [k] \setminus \{i\} \mid U_i = x] f(x - u_i) dx = \\ &= \int_{\mathbb{R}} \prod_{l \neq i} F(x - u_l) f(x - u_i) dx = \\ &= \int_{\mathbb{R}} \prod_{l \neq i} F(x + u_i - u_l) f(x) dx. \end{aligned}$$

For every $x \in \mathbb{R}$, since $u_i > u_j$, we have

$$F(x + u_i - u_l) \geq F(x + u_j - u_l) \quad \forall l \in [k] \setminus \{i\}.$$

In particular, because $|u_l - u_m| < b - a$ for any two indices $l, m \in [n]$, there exists a subinterval $(c, d) \subseteq \text{Supp}(f)$ such that $f(x)F(x + u_i - u_l) > 0$ for all $x \in (c, d)$ and $l \neq i$, thereby implying

$$\prod_{l \in [k], l \neq i} F(x + u_i - u_l) > \prod_{l \in [k], l \neq i} F(x + u_j - u_l).$$

In the product on the right-hand side, we can further lower bound the term corresponding to $l = j$:

$$F(x + u_j - u_j) \geq F(x + u_i - u_j).$$

Hence, the winning probability of item i satisfies

$$\begin{aligned} P_i &= \int_{\mathbb{R}} \prod_{l \neq i} F(x + u_i - u_l) f(x) dx > \\ &> \int_{\mathbb{R}} \prod_{l \neq j} F(x + u_j - u_l) f(x) dx = P_j. \end{aligned}$$

This completes the proof. $\square$

D.4 Utility-to-probability Relation

To derive the utility-to-probability relation given by (13) in Section 5, we used the monotonicity of ξ . Together with the monotonicity of ζ , this is the content of the following lemma.

Lemma 13. *The functions ζ and ξ defined by (12) are non-decreasing on $(0, b - a)$.*

Proof. By definition, the function ζ is given by

$$\zeta(t) = 1 - \int_{\text{Supp}(f)} \frac{r(x+t)}{r(x)} f(x) dx = 1 - \int_{\text{Supp}(f)} r(x+t) F(x) dx,$$

where we used the definition of the reversed hazard rate $r = f/F$. In particular, observe that

$$r(x) = \frac{d}{dx} (\log(F(x))), \quad \forall x \in \text{Supp}(f).$$

Since f is log-concave, its CDF F is also log-concave. This is equivalent to requiring that $\log(F)$ is concave, thus implying that $(\log(F))' = r$ is non-increasing. Therefore, if $0 < t < s < b - a$, we have

$$r(x+t)F(x) \geq r(x+s)F(x) \quad \forall x \in \text{Supp}(f),$$

where we used the fact that F is non-negative. Using the above expression for ζ , by integrating over x , we obtain

$$\zeta(t) \leq \zeta(s).$$

Finally, $\xi = 1/(1 - \zeta)$, which implies that ξ is also non-decreasing. $\square$

D.5 Uniform Lower Bound on the Resolution Function

Owing to (V), (WO1), and (WO2), we can derive a uniform lower bound on the resolution function, emulating what we did in the FR case.

Lemma 14. *Let ν be any noise distribution in the class $\mathcal{N}_V(a, b)$. Then*

$$\eta(t) \geq \eta^*(t) \quad \forall t \in (0, b - a),$$

where η and $\eta^*(\varepsilon)$ are given by (15) and (16), respectively.

Proof. Consider any noise distribution $\nu \in \mathcal{N}_V(a, b)$, with density f and CDF F . The support of f is an interval, say (c, d) , and by assumption, we have $(a, b) \subseteq (c, d)$. Let us extend the definition of the reversed-hazard rate $r : \mathbb{R} \rightarrow [0, \infty)$ to the whole real line by setting

$$r(x) := \begin{cases} \frac{f(x)}{F(x)} & \text{if } x > c \\ 0 & \text{otherwise} \end{cases}.$$

Both r and F are non-negative, so the integral of $r \cdot F$ is well-defined in $[0, \infty]$. In particular, we have

$$\int_{\mathbb{R}} r(x) F(x) dx = \int_c^d \frac{f(x)}{F(x)} F(x) dx = \int_c^d f(x) dx = 1. \quad (40)$$

Since we want to bound the function η , it may be convenient to rewrite it as

$$\eta(t) = \frac{\zeta(t)}{2 - \zeta(t)},$$

where ζ is defined by (12). The function ζ takes values in $(0, 1)$, so we can bound

$$\eta(t) \geq \frac{\zeta(t)}{2}.$$

Next, we conveniently rewrite

$$\begin{aligned} \zeta(t) &= 1 - \int_c^d \frac{f(x+t)}{F(x+t)} F(x) dx \\ &= 1 - \int_c^d r(x+t) F(x) dx \\ &\stackrel{(i)}{=} \int_c^d r(x) F(x) dx - \int_c^d r(x+t) F(x) dx \\ &\stackrel{(ii)}{=} \int_c^d r(x) F(x) dx - \int_{c+t}^{d+t} r(y) F(y-t) dy \\ &\stackrel{(iii)}{=} \int_c^d r(x) F(x) dx - \int_c^d r(y) F(y-t) dy \\ &= \int_c^d r(x) [F(x) - F(x-t)] dx. \end{aligned}$$

In (i), we used (40); (ii) follows by setting $y = x+t$ in the second integral; (iii) is due to $F(y-t) = 0$ for $y \in (c, c+t)$ and $r(y) = 0$ for $y > d$. Rewriting the CDF difference in the square brackets as an integral involving the density f , the last expression is equal to

$$\begin{aligned} &\int_c^d r(x) \int_{x-t}^x f(y) dy dx \\ &\stackrel{(iv)}{=} \int_{c-t}^d f(y) \int_y^{y+t} r(x) dx dy \\ &= \int_c^d f(y) \int_y^{\min(y+t, d)} r(x) dx dy, \end{aligned}$$

where (iv) follows by applying Fubini's theorem (for positive functions). Next, observe that $r(x) \geq f(x)$ for every $x \in (c, d)$, thus implying

$$\int_y^{\min(y+t, d)} r(x) dx \geq \int_y^{\min(y+t, d)} f(x) dx = F(\min(y+t, d)) - F(y).$$

Now, let X and Y be two i.i.d. random variables drawn from ν , and let $Z := X - Y$. Then,

$$\begin{aligned} \int_c^d f(y) [F(\min(y+t, d)) - F(y)] dy &= \int_c^d f(y) \mathbb{P}[\{y < X < y+t\} \cap \{X < d\}] dy \\ &\stackrel{(v)}{=} \int_c^d f(y) \mathbb{P}[y < X < y+t] dy \\ &= \mathbb{P}[Y < X < Y+t] \\ &= \mathbb{P}[0 < X - Y < t] \\ &= \mathbb{P}[0 < Z < t], \end{aligned}$$

where (v) is due to $X < d$ $\mathbb{P}$ -almost surely. In summary, we have derived:

$$\zeta(t) \geq \mathbb{P}[0 < Z < t]. \quad (41)$$

Being the difference between two independent log-concave random variables, Z is log-concave and symmetric, with $\text{Var}[Z] \leq 2V$. If v denotes the true variance of the noise distribution, then $\text{Var}[Z] = 2v$, so applying Corollary 20 with $\sigma^2 = 2v$ yields

$$\mathbb{P}[Z \in (0, t)] \geq \frac{t}{4e^3 \sqrt{2v}} \quad \forall t \in (0, \sqrt{2v}).$$

Using the fact that $v \leq V$ and Equation (41), we obtain:

$$\zeta(t) \geq \frac{t}{4e^3\sqrt{2V}} \quad \forall t \in (0, \sqrt{2v}].$$

On the other hand, for $t \geq \sqrt{2v}$, we have

$$\mathbb{P}[0 < Z < t] \geq \mathbb{P}[0 < Z < \sqrt{2v}] \geq \frac{1}{4e^3}.$$

In conclusion, using $\eta \geq \zeta/2$, we obtain the desired inequality, that is:

$$\eta(\varepsilon) \geq \min\left(\frac{1}{8e^3}, \frac{\varepsilon}{8e^3\sqrt{2V}}\right) \quad \forall \varepsilon \in (0, b-a).$$

□

D.6 Sample Complexity Upper Bound

In this section, we derive the sample complexity upper bound given by Theorem 3 for the WO-VUR algorithm.

To do so, we let δ_n be defined as in Equation (14) and define $\mathcal{E}$ as the event consisting of those outcomes for which the following hold simultaneously for all $i \in [k]$ and every $n \in \mathbb{N} \setminus \{1\}$:

$$|P_i - P_i^n| \leq \sqrt{\frac{2V_i^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)}, \quad (42)$$

$$\left| \sqrt{V_i^n} - \sqrt{P_i(1-P_i)} \right| \leq \sqrt{\frac{2 \log(2/\delta_n)}{n-1}}. \quad (43)$$

Arguing exactly as in Appendix C.5, condition (43) allows us to replace the sample variance V_i^n by the true one, at the price of an extra additive term:

$$\sqrt{\frac{2V_i^n \log(4/\delta_n)}{n}} + \frac{7 \log(4/\delta_n)}{3(n-1)} \leq \sqrt{\frac{2P_i(1-P_i) \log(4/\delta_n)}{n}} + \frac{13 \log(4/\delta_n)}{3(n-1)}. \quad (44)$$

The left-hand side is the quantity appearing both in (42) and in the stopping rule (17), so that, whenever $\mathcal{E}$ occurs, both can be controlled through the right-hand side.

First, we determine a positive integer n_0 such that, for every $n \geq n_0$, the following holds for every $i \in [k]$:

$$\sqrt{\frac{2P_i(1-P_i) \log(4/\delta_n)}{n}} + \frac{13 \log(4/\delta_n)}{3(n-1)} \leq \frac{P_i}{2}. \quad (45)$$

By (44), this guarantees $|P_i - P_i^n| \leq P_i/2$ on $\mathcal{E}$. The inequality is satisfied if the following hold simultaneously:

$$\sqrt{\frac{2P_i(1-P_i) \log(4/\delta_n)}{n}} < \frac{P_i}{4} \quad \text{and} \quad \frac{13 \log(4/\delta_n)}{3(n-1)} < \frac{P_i}{4}.$$

For these conditions to be satisfied, we can enforce the stronger requirement:

$$n > \frac{32 \log(4/\delta_n)}{P_i}.$$

Since we want this to hold for every $i \in [k]$, let us define $P_{\min} := \min_{i \in [k]} P_i$, and let n_0 be the minimum positive integer n such that

$$n \geq \max\left\{2, \frac{32 \log(4/\delta_n)}{P_{\min}}\right\}.$$

It follows that

$$n_0 \leq \frac{32 \log(4/\delta_{n_0})}{P_{\min}} + 1 \leq \frac{128}{P_{\min}} \log\left(2\pi n_0 \sqrt{\frac{k}{3\delta}}\right).$$

By Lemma 17, there exists a universal constant $C_1 > 0$ such that

$$n_0 \leq \frac{C_1}{P_{\min}} \log \left(\frac{C_1}{P_{\min}} \cdot \frac{k}{\delta} \right).$$

Now assume that $n > n_0$. If $\mathcal{E}$ occurs, then the condition given by (45) ensures $P_i < 2P_i^n$ for every $i \in [k]$, so that $\frac{\eta^*(\varepsilon)}{2} P_i^n > \frac{\eta^*(\varepsilon)}{4} P_i$. In this regime, using (44) once more, the stopping condition of $\mathcal{A}$ is fulfilled if the following hold for every $i \in [k]$:

$$\sqrt{\frac{2P_i(1-P_i) \log(4/\delta_n)}{n}} < \frac{P_i}{8} \eta^*(\varepsilon) \quad \text{and} \quad \frac{13 \log(4/\delta_n)}{3(n-1)} < \frac{P_i}{8} \eta^*(\varepsilon). \quad (46)$$

In particular, these hold simultaneously for every $i \in [k]$ if the following stronger requirement is met:

$$n \geq \frac{256}{P_{\min} \eta^*(\varepsilon)^2} \log \left(\frac{4}{\delta_n} \right).$$

If we define n_1 as the minimum positive integer $n > n_0$ such that

$$n \geq \left\lceil \frac{256 \log(4/\delta_n)}{P_{\min} \eta^*(\varepsilon)^2} \right\rceil,$$

which exists since the right-hand side grows logarithmically in n , then either $n_1 = n_0 + 1$ or the inequality above fails at $n_1 - 1$, in which case $n_1 \leq 256 \log(4/\delta_{n_1}) / (P_{\min} \eta^*(\varepsilon)^2) + 2$. In both cases, we can apply the same reasoning as above to find

$$n_1 \leq \frac{C_2}{P_{\min} \eta^*(\varepsilon)^2} \log \left(\frac{C_2}{P_{\min} \eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right),$$

for some universal constant $C_2 > 0$.

Up to re-tuning C_1 and C_2 , our derivations show that, under event $\mathcal{E}$, the stopping condition is met within a number of steps equal to

$$T_0 := \left\lceil \frac{C_2}{P_{\min} \eta^*(\varepsilon)^2} \log \left(\frac{C_2}{P_{\min} \eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right) \right\rceil.$$

Therefore, we have

$$\mathbb{E}[T] = \mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}}] + \mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}^c}] \leq T_0 \cdot \mathbb{P}[\mathcal{E}] + \mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}^c}].$$

Using the tail-sum formula for the expectation, we can rewrite

$$\mathbb{E}[T \cdot \mathbb{1}_{\mathcal{E}^c}] = \sum_{n=1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n] \leq T_0 \cdot \mathbb{P}[\mathcal{E}^c] + \sum_{n=T_0+1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n].$$

Now, observe that:

$$T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n \iff T \geq n \text{ and } \mathcal{E}^c \text{ occurs.}$$

However, we showed above that if $n \geq T_0 + 1$ and $T \geq n$, then necessarily $\mathcal{E}^c$ must have occurred, as otherwise we would have found $T \leq T_0$. Thus, for every $n \geq T_0 + 1$, we have

$$\mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n] = \mathbb{P}[T \geq n].$$

If $T \geq n$, then the stopping rule was not satisfied at time $n - 1$. This implies that at least one of the two conditions given by (42) and (43) was violated for at least one item $i \in [k]$. By a union bound over the k items and the two conditions, this occurs with probability at most $2k\delta_{n-1}$. Hence, we have

$$\begin{aligned} \sum_{n=T_0+1}^{\infty} \mathbb{P}[T \cdot \mathbb{1}_{\mathcal{E}^c} \geq n] &\leq 2k \sum_{n=T_0+1}^{\infty} \delta_{n-1} \\ &= \delta \cdot \sum_{n=T_0}^{\infty} \frac{6}{\pi^2 n^2} \\ &\leq C_3, \end{aligned}$$

where $C_3 > 0$ is a universal constant. This implies $\mathbb{E}[T] \leq T_0 + C_3$.

In conclusion, there exists a universal constant $C > 0$ such that

$$\mathbb{E}[T] \leq \frac{C}{P_{\min} \eta^*(\varepsilon)^2} \log \left(\frac{C}{P_{\min} \eta^*(\varepsilon)^2} \cdot \frac{k}{\delta} \right).$$

Using the definition of $\eta^*(\varepsilon)$, when $\sqrt{2V} \geq \varepsilon$, the inequality above can be rewritten as

$$\mathbb{E}[T] \leq \frac{C \cdot V}{P_{\min} \varepsilon^2} \log \left(\frac{C \cdot V}{P_{\min} \varepsilon^2} \cdot \frac{k}{\delta} \right).$$

□

D.7 Information-theoretic Lower Bound

Finally, we prove Theorem 4, thus deriving an information-theoretic lower bound for the ranking problem under WO feedback.

For every $\delta \in (0, 1)$, algorithm $\mathcal{A}^*$ recovers an accurate representation of the utility ranking with probability at least $1 - \delta$, provided that Assumptions (WO1), (WO2), and (V) are satisfied. In other words, if $\mathbf{u}$ and $\mathbf{v}$ are chosen so that these conditions are met, then:

$$\mathbb{P}[\hat{\sigma}^* \text{ is } \varepsilon\text{-accurate for } \sigma_u \mid \text{RUM}(\mathbf{u}, \mathbf{v})] \geq 1 - \delta.$$

We now construct two specific instances of the problem, and we use them to provide a lower bound on the sample complexity of $\mathcal{A}^*$. Throughout, we assume $k \geq 3$. To do so, choose the *minimum-probability level* $p \in (0, 1/k)$ and set an accuracy level $\varepsilon \in (0, b - a)$. As we shall see, the latter must be chosen “sufficiently small”, which, in this case, means $\varepsilon \in (0, \beta)$. However, this restriction does not invalidate our sample complexity result due to its asymptotic nature.

Then, let ν be the centered Gumbel distribution with scale and location parameters given by

$$\beta = \frac{\sqrt{6V}}{\pi} \quad \text{and} \quad \mu = -\beta\gamma,$$

where γ denotes the Euler–Mascheroni constant. In particular, this ensures that ν is centered and that its variance is equal to V . Since the location parameter shifts all sampled utilities by the same constant, the implied winning probabilities do not depend on μ .

Now, consider the utility vectors $\mathbf{u}$ and $\mathbf{v}$ given by:

- $u_1 = \varepsilon/2, u_2 = -\varepsilon/2, u_3 = \dots = u_{k-1} = 0$, and $u_k = L$.
- $v_1 = -\varepsilon/2, v_2 = \varepsilon/2, v_3 = \dots = v_{k-1} = 0$, and $v_k = L$.

Here $k \geq 3$ guarantees that item k , which carries the utility L , is distinct from items 1 and 2; for $k = 3$ the intermediate block is empty, and the two instances reduce to $(\varepsilon/2, -\varepsilon/2, L)$ and $(-\varepsilon/2, \varepsilon/2, L)$. Up to choosing ε sufficiently small, the positive constant L is chosen so that the minimum winning probability of both instances is equal to p . We let P and Q be the *winning distributions* corresponding to $\mathbf{u}$ and $\mathbf{v}$, that is,

$$P_i := \frac{e^{u_i/\beta}}{\sum_{\ell=1}^k e^{u_\ell/\beta}} \quad \text{and} \quad Q_i := \frac{e^{v_i/\beta}}{\sum_{\ell=1}^k e^{v_\ell/\beta}} \quad \forall i \in [k].$$

Lemma 15. *Let P and Q be the winning distributions introduced above. Then, the KL divergence between P and Q is upper bounded as*

$$D_{\text{KL}}(P \parallel Q) \leq \frac{\varepsilon^2 \pi^2}{6V} \cdot C \cdot p,$$

where $C > 0$ is a universal constant.

Proof. By direct calculations, we obtain

$$D_{\text{KL}}(P \parallel Q) = \frac{\varepsilon}{\beta} (P_1 - P_2).$$

Up to choosing $\varepsilon \in (0, \beta)$, there exists a universal constant $C > 0$ such that

$$P_1 - P_2 \leq \frac{\varepsilon}{\beta} \cdot C \cdot P_2.$$

Due to our choice of L , $P_2 = P_{\min} = p$, thus implying

$$D_{\text{KL}}(P \parallel Q) \leq \left(\frac{\varepsilon}{\beta}\right)^2 \cdot C \cdot p = \frac{\varepsilon^2 \pi^2}{6V} \cdot C \cdot p.$$

□

In order to employ Lemma 16, we think of our setting as a trivial multi-armed bandit problem, in which the number of arms is $n = 1$. At time t , the only available arm is played. The observed reward is the feedback signal, which consists of the winner of the current round. The two bandit models for the problem are $\text{RUM}(\mathbf{u}, \nu)$ and $\text{RUM}(\mathbf{v}, \nu)$, where $\mathbf{u}$ and $\mathbf{v}$ are the ones introduced above. Since the noise distribution is of the Gumbel type, all winning probabilities are non-zero. Thus, the two reward distributions are mutually absolutely continuous, as needed in order to apply Lemma 16.

Now, let $\mathcal{E}$ be the event that item 1 is ranked before item 2. Since the output ranking is an $\mathcal{F}_T$ -measurable function, $\mathcal{E} \in \mathcal{F}_T$. If we define $\mathbb{P}_u$ and $\mathbb{P}_v$ as the probability measures induced by $\text{RUM}(\mathbf{u}, \nu)$ and $\text{RUM}(\mathbf{v}, \nu)$, respectively, a direct application of Lemma 16 yields

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] D_{\text{KL}}(P \parallel Q) \geq d(\mathbb{P}_u(\mathcal{E}), \mathbb{P}_v(\mathcal{E})).$$

Let us assume that $\delta \in (0, 1/4]$. Algorithm $\mathcal{A}^*$ is (ε, δ) -PAC, so

$$\mathbb{P}_u(\mathcal{E}) \geq 1 - \delta \quad \text{and} \quad \mathbb{P}_v(\mathcal{E}) \leq \delta.$$

Due to the properties of the binary relative entropy, we have

$$d(\mathbb{P}_u(\mathcal{E}), \mathbb{P}_v(\mathcal{E})) \geq d(1 - \delta, \delta) \geq \log\left(\frac{1}{3\delta}\right).$$

It follows that

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] \geq \frac{1}{D_{\text{KL}}(P \parallel Q)} \log\left(\frac{1}{3\delta}\right).$$

Owing to Lemma 15, we conclude that

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] \geq \frac{6V}{C p \pi^2 \varepsilon^2} \log\left(\frac{1}{3\delta}\right).$$

In conclusion, since $\log(1/(3\delta)) \geq \frac{1}{5} \log(1/\delta)$ for every $\delta \in (0, 1/4]$, we have shown that

$$\mathbb{E}_{\nu, \mathbf{u}}[T^*] = \Omega\left(\frac{V}{p \varepsilon^2} \log\left(\frac{1}{\delta}\right)\right).$$

E Auxiliary

E.1 Change-of-measure Lemma for Multi-armed Bandit Problems

The next result restates Lemma 1 of [19]. As this is typically used in the context of multi-armed bandit problems, we need to introduce some preliminary notation. The number of arms is any positive integer n , and we identify the arm set with $[n]$. At time t , let A_t and Z_t be the (possibly random) arm being chosen and the corresponding reward, respectively. Let $\mathcal{F}_t = \sigma(A_1, Z_1, \dots, A_t, Z_t)$ be the sigma-algebra generated by the trajectory followed up to time t . For every $i \in [n]$, we define $N_i(t)$ as the number of times that arm i was played up to time t . We use the subscripts $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ to refer to the probability measure induced by a given bandit model ν and the implied expectation functional.

Lemma 16. *Let $\nu = (\nu_i)_{i=1}^n$ and $\nu' = (\nu'_i)_{i=1}^n$ be two bandit models, where ν_i (resp. ν'_i) is the reward distribution of arm $i \in [n]$. Assume that, for every $i \in [n]$, the probability distributions ν_i and ν'_i are equivalent, i.e., they are mutually absolutely continuous. For any almost-surely finite stopping time τ with respect to the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$, the following holds:*

$$\sum_{i=1}^n \mathbb{E}_\nu[N_i(\tau)] D_{\text{KL}}(\nu_i \parallel \nu'_i) \geq \sup_{E \in \mathcal{F}_\tau} d(\mathbb{P}_\nu[E], \mathbb{P}_{\nu'}[E]),$$

where $d(x, y) := x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is the binary relative entropy, with the convention that $d(0, 0) = d(1, 1) = 0$.

As we shall motivate later, the setting considered in this lemma encompasses ours by choosing the number of arms equal to one.

E.2 Technical Result for Bounding the Sample Complexity

The next result is a technical lemma used to derive the sample complexity upper bounds given by Theorem 1 and Theorem 3. A more general version can be found as Lemma 12 in [32]. The slightly modified version presented below is sufficient for our purposes.

Lemma 17. *Let $n \in \mathbb{N}$ and $a, b > 0$. Suppose that $n \leq a \log(bn)$. Then*

$$n \leq 2a \log(ab).$$

Proof. For every $x > 0$, $\log(x) \leq \sqrt{x}$. Therefore, we have

$$n \leq a \log(bn) \leq a\sqrt{bn} = a\sqrt{b}\sqrt{n},$$

from which we obtain

$$\sqrt{n} \leq a\sqrt{b}.$$

Equivalently, $n \leq a^2 b$. Since the logarithm is increasing and $b > 0$, we have

$$n \leq a \log(bn) \leq a \log(b(a^2 b)) = 2a \log(ab).$$

□

E.3 Deviation Bounds for Unimodal and Log-concave Distributions

The following is a standard result providing useful deviation bounds for symmetric unimodal distributions. The version for standardized random variables can be found as Theorem 7.1 in [33].

Lemma 18. *Let Z be a symmetric unimodal random variable with $\text{Var}[Z] = \sigma^2 > 0$. Then*

$$\mathbb{P}[Z \geq t] \leq \begin{cases} \frac{1}{2}(1 - \frac{t}{\sqrt{3}\sigma}) & \text{if } 0 \leq t < 2\sigma/\sqrt{3} \\ \frac{2\sigma^2}{9t^2} & \text{if } t \geq 2\sigma/\sqrt{3} \end{cases}. \quad (47)$$

Next, we derive a useful small-deviation bound for centered log-concave distributions with a normalized variance. The proof is due to Pinelis [34].

Lemma 19. *Let f be a log-concave density with $\int_{\mathbb{R}} x f(x) dx = 0$ and $\int_{\mathbb{R}} x^2 f(x) dx = 1$. Then, for every $t \in (0, 1)$, we have*

$$\int_{-t}^t f(x) dx \geq \frac{t}{2e^3}.$$

Proof. Since f is log-concave, there exists a concave function $g : \mathbb{R} \rightarrow [-\infty, \infty)$ such that $f = e^g$. A general result on log-concave densities [14, Lemma 5.5] ensures that $g(0) \geq -\log(8)$.

Suppose $g(1/2) \leq -3$. Then, using concavity, we find

$$g(x) \leq -3 - (x - 1/2) \frac{3 - \log(8)}{1/2} =: h(x), \quad \text{for all } x > 1/2.$$

It follows that

$$\int_{1/2}^{\infty} x^2 f(x) dx \leq \int_{1/2}^{\infty} x^2 e^{h(x)} dx < 0.04.$$

Similarly, if $g(-1/2) \leq -3$, by concavity,

$$g(x) \leq -3 + (x + 1/2) \frac{3 - \log(8)}{1/2} =: k(x), \quad \text{for all } x < -1/2.$$

It follows that

$$\int_{-\infty}^{-1/2} x^2 f(x) dx \leq \int_{-\infty}^{-1/2} x^2 e^{k(x)} dx < 0.04.$$

Hence, if $g(1/2) \leq -3$ and $g(-1/2) \leq -3$, we have

$$\begin{aligned} \int_{\mathbb{R}} x^2 f(x) dx &= \int_{-\infty}^{-1/2} x^2 f(x) dx + \int_{1/2}^{\infty} x^2 f(x) dx + \int_{-1/2}^{1/2} x^2 f(x) dx \\ &< 0.08 + \int_{-1/2}^{1/2} x^2 f(x) dx \\ &\leq 0.08 + \frac{1}{4} < 1. \end{aligned}$$

This contradicts our assumption that f has variance equal to one. Therefore, $g(1/2) \geq -3$ or $g(-1/2) \geq -3$. Without loss of generality, we can assume that $g(1/2) \geq -3$. By concavity, this implies that $g(x) \geq -3$ for every $x \in (0, 1/2)$. Thus, for $t \in [0, 1/2]$, we have

$$\int_{-t}^t f(x) dx \geq \int_0^t e^{-3} dx = e^{-3}t.$$

On the other hand, if $t \in [1/2, 1]$, we obtain

$$\int_{-t}^t f(x) dx \geq \int_{-1/2}^{1/2} f(x) dx \geq \frac{e^{-3}}{2} \geq \frac{t}{2e^3},$$

where the last inequality follows from $t/2 \leq 1/2$. □

As a useful corollary, we derive a small-deviation bound for symmetric log-concave distributions with arbitrary variance.

Corollary 20. *Let X be a symmetric log-concave random variable with density f and $\text{Var}[X] = \sigma^2 > 0$. Then, for every $t \in (0, \sigma]$, we have*

$$\mathbb{P}[X \in (0, t)] \geq \frac{t}{4e^3\sigma}.$$

Proof. Let $t \in (0, \sigma]$, and observe that

$$X \in (-t, t) \iff \frac{X}{\sigma} \in \left(-\frac{t}{\sigma}, \frac{t}{\sigma}\right).$$

The variance of the random variable X/σ is equal to one, so by Lemma 19, we have

$$\begin{aligned} \mathbb{P}[X \in (0, t)] &= \frac{1}{2} \mathbb{P}[X \in (-t, t)] \\ &\geq \frac{t}{4e^3\sigma}. \end{aligned}$$

□