Learning from Ranking Feedback: Improved Regret Bounds via Independence Preserving Rank Breaking

Nigel Q. D. Strachan
Delft University of Technology
n.q.d.strachan@tudelft.nl

Sattar Vakili
University College London
s.vakili@ucl.ac.uk

Matthijs T. J. Spaan
Delft University of Technology
m.t.j.spaan@tudelft.nl

Julia Olkhovskaya
Delft University of Technology
julia.olkhovskaya@gmail.com

Abstract

We study a sequential decision-making problem where at each time the learner selects a bounded amount of items, which we call an assortment, and receives a ranking of the selected items. Ranking feedback arises naturally when human evaluators or automated LLM judges compare multiple options simultaneously. It is often easier and more reliable than assigning absolute scores, which are typically poorly calibrated across evaluators, while yielding substantially more information than a single pairwise comparison. We focus on rankings that are constructed according to the random utility model with linear utilities, which includes the popular Plackett-Luce model as a special case. We show that improved regret minimisation is possible under rank breaking with a refined analysis based on Baranyai's theorem that fully exploits independence in the pairwise comparisons. This gives a tighter confidence interval around the true utility parameter which is crucial in our analysis for regret that improves with maximum assortment size. Furthermore, it matches the regret lower bound for our model in case the learner is permitted to play multi-sets of the maximum bounded size. Our work gives a principled approach to learning from ranking feedback that shows consistent improvement with the maximum assortment size.

1 Introduction

In many real-world learning problems, obtaining scalar, well-calibrated feedback is expensive, time-consuming, or even infeasible [Wirth et al., 2017]. Evaluating a candidate molecule's binding affinity, a new recommendation's favourability, or a generated image's perceptual quality often requires costly experiments or expert assessment. In contrast, both *human evaluators*, such as crowd-sourced annotators, and, more recently, *large language models (LLMs) acting as judges* [Gu et al., 2026], are far more consistent and efficient when asked to *rank* or *compare* multiple options rather than to assign absolute numerical scores.

The prevalence of ranking feedback has motivated a growing body of work on learning from preference feedback [Wirth et al., 2017], where only relative judgments between items are observed. In many settings, however, feedback consists of a comparison of multiple items [Tang and Feng, 2025]. This naturally leads to considering models that process multiple comparisons simultaneously.

Problem setting We study a sequential decision-making problem of learning from ranking feedback. In this model, at each time t, the learner offers an assortment, which is a set of items of size at most q from an item set $\mathcal{A}$, and observes a noisy ranking based on an unobserved latent utility function

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

$f : \mathcal{A} \to \mathbb{R}$. The goal is to minimise regret, to be defined as the expected value of the cumulative sum of the difference in utility of the best item and the average utility of items offered in the assortment. This notion of regret has been considered by Saha [2021] for offering assortments, but only for receiving winner feedback, that is observing only the preferred item from the assortment, under Plackett-Luce model [Plackett, 1975]. The Plackett-Luce model can be interpreted as sequentially choosing a winner without replacement where the probability weights of each item are the exponential of their utility.

The case when $q = 2$, which is also called duelling bandits, has by now been well-studied [Yue and Joachims, 2009, Ailon et al., 2014, Saha, 2021, Pásztor et al., 2024, Xu et al., 2024, Kayal et al., 2025]. An extensive survey on duelling bandits is provided by Bengs et al. [2021]. In the special case where the utility function is linear, that is $f(x) = \langle \theta^*, x \rangle$ for some $\theta^* \in \mathbb{R}^d$, results which are comparable to the classical linear bandit have been derived [Saha, 2021]. It turns out that no accelerated regret learning is possible under just *winner feedback*, for playing sets instead of pairs where comparisons are constructed according to the Plackett-Luce model [Saha, 2021].

Even though accelerated learning is impossible under receiving just winner feedback in the linear utility setting [Saha, 2021], this leaves open the possibility that accelerated learning is possible when observing the *full ranking* feedback of the items offered, that is receiving an ordering according to preference which includes all items in the assortment. This line of research has mostly focused on the Plackett-Luce model [Zhu et al., 2023, Thekumparampil et al., 2025, Mukherjee et al., 2024, Lee et al., 2025].

Existing theoretical results Somewhat counter-intuitively, some recent literature on learning from ranking feedback with linear utilities have some theoretical guarantees that worsen when q is large [Zhu et al., 2023, Mukherjee et al., 2024, Thekumparampil et al., 2025], albeit not for regret minimisation as we consider. The first of these works has results on sub-optimality gaps, while the latter two provide guarantees for a prediction error type. The main reason that these works do not show improved performance in q is that the confidence bounds for the estimator of θ^* used, as provided by Zhu et al. [2023], are not strong enough to show improved performance with q. Therefore, the confidence intervals need to be further tightened. In this paper, we prove a tightened confidence interval for the rank breaking estimator (see Theorem 3.1) for the more general random utility model. Lee et al. [2025] aim to address the scaling in assortment size for sub-optimality gaps, but do not show consistent improvement in q, rather only stating that when the assortments are large that accelerated regret learning is possible and no formal lower bound on the size of the assortment is given.

Approach In this paper, we view the Plackett-Luce model from a more probabilistic viewpoint contrary to the mostly analytical approaches taken in these recent previous works [Zhu et al., 2023, Mukherjee et al., 2024, Thekumparampil et al., 2025] in the sense that we use the Gumbel top-k trick [Kool et al., 2019] for *rankings*. This enables us to maximally exploit independence in the pairwise comparisons in the following sense: we partition all pairwise comparisons into groups such that all the pairwise comparisons in the group are independent.

The key idea is to use that we can have *few* groups, so as to be able to exploit independence as much as possible under the pairwise comparisons. Using this perspective in the construction of confidence intervals yields the improved performance sought. The idea to split the sources of independence was introduced by Negahban et al. [2018]. We demonstrate in our work that we can also utilise this idea in the linear utility ranking setup for regret minimisation: the best possible scaling in regret is of the optimal order $\frac{1}{\sqrt{q}}$ compared to standard duelling bandits and it is achieved by a greedy phased elimination algorithm with our improved analysis on the rank breaking confidence interval.

Contributions Our work makes three main contributions. Firstly, we sharpen the confidence interval around θ^* under rank breaking for the Plackett-Luce model which holds for any d and q, and which also holds for the more general random utility model, by Zhu et al. [2023]. We do this by removing a factor of $\frac{1}{q}$ in the design matrix. Secondly, we utilise the improved confidence interval under rank breaking to design an algorithm that has a regret guarantee that scales favourably in q. We note that without the improved confidence interval strict improvement is not possible. We obtain a regret bound of $\tilde{O}\left(\frac{d}{\sqrt{q}}\sqrt{T}\right)$, where $\tilde{O}$ suppresses logarithmic factors in T under $q \leq \sqrt{d}$ (a

discussion of this assumption is given after Assumption 2.2). Finally, we show that the regret bound we derive is tight for the random utility model under Gaussian noise for an algorithm that is allowed to play only *multi-sets* of size q (a framework considered by Sui et al. [2017]). We thereby provide a principled approach to learning from ranking feedback.

Notation We write the Euclidean ball in dimension d with radius R as $B_2^d(0;R)$ where dimension may be omitted if it is understood. In matrix additions I is the identity matrix. For a positive semi-definite matrix we write $\|x\|_A := \sqrt{x^T A x}$. We write $A \succeq B$ for positive semi-definite A and B if $x^T(A-B)x \geq 0$ for all $x \in \mathbb{R}^d$ and call this relation Loewner order. For a finite set $\mathcal{X}$, including $\mathcal{A}$, it can be assumed that there is an ordering $\mathcal{X} = \{x_1, x_2, \cdots, x_{|\mathcal{X}|}\}$. Then formally $\binom{\mathcal{X}}{2} = \{\{x_i, x_j\} : i < j\}$. Using this viewpoint, it is unambiguous to write $\sum_{\{x,y\} \in \binom{\mathcal{X}}{2}}$ to be the sum over all pairs. We view vectors in $\mathbb{R}^d$ as functions $x : [d] \to \mathbb{R}$. Therefore, $x(i) = x_i$ and for $I \subset [d]$ we define $x(I) := (x_i \mathbb{I}(i \in I))_{i=1}^d$, i.e. the restriction of x to I. The subgaussian norm of a scalar random variable X is denoted by $\|X\|_{\psi_2}$. The subgaussian norm of a random vector X is defined by $\|X\|_{\psi_2} = \sup_{z \in B_2^d(0;1)} \|z^T X\|_{\psi_2}$ (see the textbook by Vershynin [2018] for an overview of the properties of subgaussian random variables/vectors). We write $\mathbb{I}(A)$ for the indicator of event A. Finally, we write $x \succ y$ for the event that x is preferred to y in the ranking.

2 Learning Framework

We describe the learning framework in detail along with all the necessary assumptions on the theory.

2.1 Ranking Feedback Description

At each time t in the sequential decision-making process, the learner chooses to rank an assortment $A_t = \{x_{t,i}\}_{i=1}^{|A_t|}$, which is a subset of size at most q from a finite but potentially large set of possible items $\mathcal{A}$, which we assume to be contained in the Euclidean unit ball $B_2^d(0;1)$. To lighten notation, we sometimes write q_t for $|A_t|$. The learner then receives feedback according to the *random utility model* (see the work by Hajek et al. [2014] and references therein). In this model, the learner receives a ranking based on the sorted values in decreasing order of $f(x_{t,i}) + \zeta_{t,i}$ for $i = 1, \cdots |A_t|$, with $\zeta_{t,i}$ being i.i.d. noise. In this paper we assume that f takes a linear form, that is $f(x) = \langle \theta^*, x \rangle$ for $\theta^* \in \mathbb{R}^d$. We assume that $\|\theta^*\|_2 \leq S$ with S being a known upper bound.

We make some assumptions on the noise distribution. The assumption on the noise is mild, as it mostly excludes discrete and categorial distributions.

Assumption 2.1. For all values of $x, y \in \mathbb{R}$, the noise satisfies $\mathbb{P}(x + \zeta = y + \zeta') = 0$ where ζ, ζ' are two independent identically distributed copies of the noise. This assumption is needed to ensure that the sorting procedure is well-defined. Moreover, assume that for each $x \in [-2S, 2S]$ that $\mu(x) := \mathbb{P}_{\zeta,\zeta'}(x \geq \zeta - \zeta')$ is differentiable in x and satisfies $\dot{\mu}_{\min} := \min\left\{1, \min_{x \in [-2S,2S]} \dot{\mu}(x)\right\} > 0$.

To put this learning context in perspective, we now turn to some examples which follow the described model.

Example 2.1 (Gaussian noise)**.** *Consider the case that $\zeta_{t,i}$ follows the standard Gaussian distribution. This case was also considered as a model for ranking by Plackett [1975]. We note that in this case, we have by the two stability of Gaussian random variables that $\zeta - \zeta'$ is also mean zero Gaussian with $\sigma = \sqrt{2}$. It follows that $\mu(x) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{x} \exp(-\frac{t^2}{4}) \mathrm{d}t$. Thus, our assumption is satisfied with $\dot{\mu}_{\min} := \frac{1}{2\sqrt{\pi}} \exp(-S^2)$.*

A popular model for ranking feedback is the Plackett-Luce model [Plackett, 1975]. The Plackett-Luce model is indeed a special case of our more general learning framework.

Example 2.2 (Plackett-Luce – Gumbel noise)**.** *The following fact is known as the Gumbel soft-max trick in machine learning, which gives an explicit probability that an item ends up as the winner of the ranking under the assumption that ζ_i is are independent and distributed according to standard Gumbel noise:*

$$\mathbb{P}\left(j = \arg\max_{i \in I} \langle x_i, \theta^* \rangle + \zeta_i\right) = \frac{\exp(\langle \theta^*, x_j \rangle)}{\sum_{i \in I} \exp(\langle \theta^*, x_i \rangle)}.$$

Beyond the winner feedback setting, the Plackett-Luce model for the general full ranking is constructed as follows: conditioned on the chosen set A_t and parameter $\theta^\star$, the probability of observing permutation π_t is

$$\Pr(\pi_t \mid A_t, \theta^\star) = \prod_{r=1}^{q_t} \frac{\exp(\langle \theta^*, x_{t,\pi_t(r)}^\top \rangle)}{\sum_{s=r}^{q_t} \exp(\langle \theta^\star, x_{t,\pi_t(s)} \rangle)}.$$

The Gumbel top-k trick [Kool et al., 2019] shows that, under standard Gumbel noise, the Plackett-Luce model coincides with the random utility model studied in this paper. We provide a proof in Appendix F for the sake of self-containedness and some small additional details. Direct computation shows that $\dot{\mu}_{\min} = \exp(-O(S))$.

We now see that our model naturally incorporates popular frameworks for modelling ranking feedback. Next, we explain different ways of incorporating ranking feedback and how it is richer than just pairwise feedback.

2.2 Incorporating Ranking Feedback

In duelling bandits, one can model the feedback as a Bernoulli random variable indicating whether item A is preferred over item B. In the case of ranking feedback, no such natural numerical correspondence exists and there are several ways of processing ranking feedback. For the Plackett-Luce model, one can perform maximum likelihood estimation (MLE) on the *full ranking*, as the ranking probabilities are explicit (as done for example by Zhu et al. [2023]). Another approach is to extract all pairwise comparisons, i.e. rank breaking [Soufiani et al., 2014, Hajek et al., 2014, Khetan and Oh, 2016, Negahban et al., 2018, Zhu et al., 2023]. To the best of our knowledge, the Plackett-Luce model is one of the few models that allows for a sufficiently nice closed form analytical expression of the ranking probabilities. On the other hand, computing pairwise probabilities is often much easier, such as in the Gaussian case. This makes rank breaking an appealing estimator for the random utility model.

2.3 Regret Definition and Framework

We now introduce the notion of regret and interaction sequence for our learning from ranking feedback framework. At each time t, the learner selects an assortment A_t, which is a subset of $\mathcal{A}$ with size at most q . The learner observes a permutation according to the random utility model of the assortment A_t. We adopt the notion of regret by Saha [2021] for the case of offering assortments. Let $x^* := \arg\max_{x \in \mathcal{A}} \langle \theta^*, x \rangle$ be the optimal item. We define the random regret as

$$\tilde{R}_T := \sum_{t=1}^{T} \left\langle \theta^*, x^* - \frac{1}{|A_t|} \sum_{x \in A_t} x \right\rangle.$$

This is a direct generalisation of the regret as in the duelling bandit case [Saha, 2021]. Distinctly from the work by Lee et al. [2025], this regret notion differs as we do not value the assortment as the item of maximum utility but their average, which makes this regret notion stronger. We aim to show a regret bound that scales favourably in q. To this end, we need the following assumption.

Assumption 2.2. We assume that q satisfies $q \leq \sqrt{d}$.

This assumption should be considered also in the context with the multi-armed bandit case of learning from ranking feedback, i.e. when $\mathcal{A}$ is a standard basis. In this case, the dimension dictates the number of items, and the assumption reads that we have formal guarantees whenever the number of items being compared does not exceed to the square root of the total number of items, which should be considered reasonable. Thus, this assumption can be seen as a linear function approximation version of such an assumption.

Moreover, we comment that this does not ensure that the selection strategy by [Lee et al., 2025] chooses large sets. For example, in one of their experiments for the contextual setting they report a problem instance with $d = 2048$ and 7 arms, yet the algorithm does not choose 7 arms, even if it is allowed to.

Under Assumption 2.2, we derive an algorithm that achieves $R_T = \tilde{O}(\frac{1}{\dot{\mu}_{\min}} \frac{d}{\sqrt{q}} \sqrt{T})$. Here, $\tilde{O}$ suppresses logarithmic factors. Along the way, we also show that N, satisfying that $N \geq$

$C\frac{1}{\dot{\mu}_{\min}^2}\frac{S^2d^2\log(\frac{d}{\delta})}{q\varepsilon^2}\log(N)$, ranking samples are sufficient for establishing a suboptimality gap within an additive factor of ε with probability at least $1-\delta$, where $C>0$ is an absolute constant (see Remark C.8).

3 Algorithm via Independence-Preserving Rank-Breaking and Analysis

In this section, we prove the main contributions of our paper. We start out by presenting the improved confidence interval based on rank breaking for the random utility model. Next, we utilise the improved confidence interval to show accelerated learning for learning from ranking feedback, and show that a greedy phased elimination algorithm can be used to learn from ranking feedback. Finally, we discuss a lower bound for an adjacent multi-set variant of the learning from ranking feedback problem.

3.1 Confidence Interval for Rank Breaking

We now prove our confidence interval for the rank breaking estimator around the true utility parameter θ^*. Our analysis leverages techniques used for confidence intervals suitable for generalised linear bandits as provided by Filippi et al. [2010]. We note that this confidence interval does *not* need Assumption 2.2. In this section we only provide a sketch, and the full proof with all the details can be found in Appendix B. Formally, we prove the following result.

Theorem 3.1. *Let $\mathcal{X}:=\{A_t=\{x_{t,i}\}_{i=1}^{q_t}\}_{t=1}^T$, with $q_t\leq q$ for all time points, be derived from an offline sampling strategy.*

We define the gradient space function as

$$g(\theta):=\lambda(q-1)\theta+\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}\mu(\langle\theta,x_{t,i}-x_{t,j}\rangle)(x_{t,i}-x_{t,j})$$

Let θ_{rb} be the rank breaking approximation of the true parameter vector in the case of rank breaking loss, which we define as a minimiser over the Euclidean ball with radius S of

$$\|g(\theta)-\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}\mathbb{I}(x_{t,i}\succ x_{t,j})(x_{t,i}-x_{t,j})\|_{((q-1)^2V_T)^{-1}},$$

where we define the design matrix as $V_T:=\lambda I+\frac{1}{q-1}\sum_{t=1}^T\sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T$. Then, we have for all $x\in\mathbb{R}^d$ uniformly the following inequality

$$|\langle\theta_{rb}-\theta^*,x\rangle|\leq\frac{C}{\dot{\mu}_{\min}}\left(\sqrt{d\log\left(\frac{d}{\delta}\right)+\lambda S^2}\right)\|x\|_{V_T^{-1}},$$

where $C>0$ is an absolute constant, with probability at least $1-\delta$. Here $\dot{\mu}_{\min}$ denotes the curvature of the rank breaking loss.

Remark 3.2. For the Plackett-Luce model, one can also compute the convex and coercive MLE program

$$\hat{\theta}=\arg\min_{\theta\in\mathbb{R}^d}-\frac{1}{q-1}\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}\log(\mathbb{P}(1\{x_{t,i}\succ x_{t,j}\}|\theta)+\frac{\lambda}{2}\|\theta\|_2^2.$$

Here $\mathbb{P}(1\{x_{t,i}\succ x_{t,j}\}|\theta)$ is the probability of observing the outcome $1\{x_{t,i}\succ x_{t,j}\}$ if θ were the true parameter. If this estimator is contained in the Euclidean S-ball, then it is not necessary to compute the projection estimator as given in the theorem. This follows by the first order optimality condition, as the term inside the weighted norm is then zero and thus a minimiser of the program. This is generally preferred as this is only a root finding problem which is in general much easier. Filippi et al. [2010] discuss the need of projection.

This result should be compared to the confidence interval derived by Zhu et al. [2023]. We note that their design matrix has a scaling factor of $\frac{1}{q(q-1)}$ in front of the part of the design matrix that stems

from rank broken MLE. The improvement to $\frac{1}{q-1}$ makes the notion of uncertainty smaller, which we crucially utilise to show improved sampling complexity, and hence on regret.

This confidence interval is possible by improving a subgaussian estimate of the random vector $(\mathbb{I}(x_{t,i} \succ x_{t,j}) - \mu(\langle \theta^*, x_{t,i} - x_{t,j} \rangle))_{i,j \in \binom{[q_t]}{2}}$. A trivial upper bound would give a subgaussian norm of $O(\sqrt{q(q-1)})$ by using that each component is bounded by 1 and using that to upper bound the Euclidean norm. This is, however, suboptimal, and can be improved to $O(\sqrt{q-1})$.

We prove this by using the random utility model interpretation. Observe that for i, j, i', j' distinct the random variables $\mathbb{I}(x_{t,i} \succ x_{t,j})$ and $\mathbb{I}(x_{t,i'} \succ x_{t,j'})$ are independent, as per the random utility model the first comparison only depends on the noises $\zeta_{t,i}, \zeta_{t,j}$ and the second one on $\zeta_{t,i'}\zeta_{t,j'}$, which are independent. This of course works more generally with q-wise comparisons, where we can extract $q/2$ independent comparisons from a given assortment. We will use combinatorial design theory to maximally utilise these independence relations in estimating the subgaussian norm, which we call independence preserving rank breaking. The use of combinatorial design theory to separate independence was first introduced by Negahban et al. [2018], and we use it to improve a confidence interval by Zhu et al. [2023]. We need the following definition that formalises the notion that we want to separate as much independence in the pairs as possible.

Definition 3.3. Let $\mathcal{X}$ be a finite set. An *independence chain* I is a subset of $P = \binom{\mathcal{X}}{2}$ such that for each $x, y \in I$ with $x \neq y$ we have $x \cap y = \varnothing$. An *independence grid* G is a collection of independence chains such that $I \cap J = \varnothing$ for all $I, J \in G$ with $I \neq J$ and such that the union of all of them is precisely P, that is $\cup_{I \in G} I = P$. The cardinality of G is the *size* of the independence grid.

The following example is illuminating for the technical definition above.

Example 3.1. *Let $q = 6$, then the following grid provides an independence grid for $[q]$. Here, we use a bar to separate the independence chains*

$$(1,2)(3,4)(5,6) \mid (1,3)(2,5)(4,6) \mid (1,4)(2,6)(3,5) \mid (1,5)(2,4)(3,6) \mid (1,6)(2,3)(4,5)$$

We see that the size of the independence grid is 5.

In Appendix A, we discuss the example in more detail. It is crucial to have as few independence chains in an independence grid for our improved estimate of the subgaussian norm (see also Lemma 3.6 with its proof sketch and full proof in Appendix A). Clearly, an independence grid needs at least $|\mathcal{X}| - 1$ independence chains, due to the pairs (x, y) for $y \neq x$. It turns out that, for even $|\mathcal{X}|$, this lower bound is tight. This is a consequence of a special case of Baranyai's theorem [Chapter 38, van Lint and Wilson, 2001]). For a proof of the special case we refer to Theorem 9.1 by Harary [1969]. Baranyai's theorem in full generality is given as Theorem 38.1 in the textbook [van Lint and Wilson, 2001].

Theorem 3.4 (Special case of Baranyai's theorem [van Lint and Wilson, 2001])**.** *Let $|\mathcal{X}|$ be even. Then, there exists an independence grid of $\mathcal{X}$ of size at most $|\mathcal{X}| - 1$.*

Remark 3.5. For odd $|\mathcal{X}|$, the bound of $2(|\mathcal{X}| - 1)$ can trivially be obtained by using an independence grid of $\mathcal{X} \setminus \{x\}$ for any $x \in \mathcal{X}$, and appending the missing pairs on separate independence chains.

We sketch the following bound on the subgaussian norm, whose proof can be found in more detail in Appendix A.

Lemma 3.6. *The random vector $X = (\mathbb{I}(x_{t,i} \succ x_{t,j}) - \mu(\langle \theta^*, x_{t,i} - x_{t,j} \rangle)_{i,j \in \binom{[q_t]}{2}}$ with $q_t \leq q$ is $O(\sqrt{q-1})$ subgaussian.*

Proof. Let $z \in B_2^{\binom{q_t}{2}}(0; 1)$ and let G be an independence grid of size $O(q-1)$ as $q_t \leq q$. Then, by the triangle inequality and Cauchy-Schwarz we have

$$\|z^T X\|_{\psi_2} \leq \sum_{I \in G} \|z^T(I) X(I)\|_{\psi_2} \leq \sqrt{|G|} \sqrt{\sum_{I \in G} \|z^T(I) X(I)\|_{\psi_2}^2}$$

As each component of $X(I)$ is bounded by one uniformly and are independent, we have $\|z^T(I) X(I)\|_{\psi_2}^2 \leq \|z(I)\|_2^2$ [Vershynin, 2018, Proposition 2.6.1]. Therefore, the subgaussian norm is bounded by

$$C\sqrt{q-1}\sqrt{\sum_{I\in G}\|z(I)\|_2^2} = C\sqrt{q-1}\|z\|_2 = C\sqrt{q-1},$$

for $C > 0$ an absolute constant. We used that each component appears only once, that the union of the independence grid is all pairs, and finally that $\|z\|_2 = 1$. □

Therefore, we see that the subgaussian norm is improved. We refer the reader of the exact details of the proof in Appendix B. We are now in a position to give a proof sketch of the main confidence interval.

Proof sketch of Theorem 3.1. From standard analysis by Filippi et al. [2010] we have

$$|\langle\theta_{rb} - \theta^*, x\rangle| = |\langle G_T^{-1}(g(\theta_{rb}) - g(\theta^*)), x\rangle| \leq \|g(\theta_{rb}) - g(\theta^*)\|_{G_T^{-1}}\|x\|_{G_T^{-1}}, \tag{1}$$

for some matrix G_T satisfying up to a factor depending on only on $\dot{\mu}_{\min}$ the Loewner order bound $G_T \succeq \lambda(q-1)I + \sum_{t=1}^T \sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$. We multiply (1) by $\frac{\sqrt{q-1}}{\sqrt{q-1}}$ and distribute it under each of the weighted norms. The second factor gives that $\sqrt{q-1}\|x\|_{G_T^{-1}} \leq \|x\|_{(\lambda I + \frac{1}{q-1}\sum_{t=1}^T \sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T)^{-1}}$ by expanding definitions and using the Loewner order lower bound. For the first factor, we perform a triangle inequality with $\sum_{t=1}^T \sum_{\{i,j\}\in\binom{[q_t]}{2}} \mathbb{I}(x_{t,i} \succ x_{t,j})(x_{t,i} - x_{t,j})$. We then use the defining property of θ_{rb} to upper bound the first term by $2\|g(\theta^*) - \sum_{t=1}^T \sum_{\{i,j\}\in\binom{[q_t]}{2}} \mathbb{I}(x_{t,i} \succ x_{t,j})(x_{t,i} - x_{t,j})\|_{((q-1)V_T)^{-1}}$ as $\|\theta^*\|_2 \leq S$. We then split the first factor into two parts, namely $\|\lambda(q-1)\|_{((q-1)V_T)^{-1}}$ and $\|\sum_{t=1}^T \sum_{\{i,j\}\in\binom{[q_t]}{2}}(\mu(\langle\theta^*, x_{t,i} - x_{t,j}\rangle) - \mathbb{I}(x_{t,i} \succ x_{t,j}))(x_{t,i} - x_{t,j})\|_{((q-1)V_T)^{-1}}$. The first can be trivially bounded by the Loewner order lower bound of the identity matrix. For the second part we use Hoeffding's inequality which can be applied due to non-adaptive sampling assumption. When adjusted with the factor $\frac{1}{\sqrt{q-1}}$, these bounds result in a $O(\sqrt{d\log(d/\delta) + \lambda S^2})$ bound. Details of these steps are in Appendix B. □

We stress that independence grids are only used during analysis and are not needed for the actual implementation of our algorithm. We now design a multi-round algorithm based on the previous confidence interval for the rank breaking estimator.

3.2 Algorithm for learning with ranking feedback

Our algorithm, Algorithm 1, is based on phased elimination similar to Chapter 22 of Lattimore and Szepesvári [2020]. The key step in the analysis is to show that, under the correctness of the used confidence intervals, that the optimal item never gets eliminated and that all items with a sub-optimality gap greater than $\varepsilon_\ell := 2S(2^{-\ell})$ get eliminated at the end of round ℓ.

We start by explaining why we see improvement in q. For this, we introduce a notion of uncertainty for an assortment $A = \{x_i\}_{i=1}^{|A|}$, which we call $\sigma_t(A)$ defined as

$$\sigma_t(A) := \sum_{\{i,j\}\in\binom{|A|}{2}} \|x_i - x_j\|^2_{V_{t-1}^{-1}}. \tag{2}$$

Recall that V_t has the form $\lambda I + \frac{1}{q-1}\sum_{s=1}^t \sum_{\{i,j\}\in\binom{[q_s]}{2}}(x_{s,i} - x_{s,j})(x_{s,i} - x_{s,j})^T$. We can use a version of the elliptical potential lemma of Oh and Iyengar [2021] to bound the cumulative sum of the uncertainties $\sum_{s=1}^t \sigma_{s-1}(A_s)$. The improved scaling in q in V_t gives an upper bound which scales *linearly* in q, while each time we sum a quadratic number of pairwise weighted norms. It is by using this insight that we can show that the upper confidence bound is realised to a sufficient degree q times as fast for all pairwise differences of items in $\mathcal{A}_{\ell-1}$. This leads to a decreased sampling complexity and thus to lower regret.

We achieve this goal in two phases. The first is called the exploration phase, which is of duration $E_\ell = \lceil c\frac{(d\log(T)+dS^2)}{q\varepsilon_\ell^2} d\log(T)\rceil$ for an absolute constant $c > 0$. During this phase, we ensure that

every item has a 'projection' into A_{E_ℓ} which maximizes (2) up to the accuracy of $\frac{1}{12}\varepsilon_\ell$. Formally, let $\sqrt{\beta}$ be the confidence width as described in Algorithm 1. Using the design matrix V_{E_ℓ} from the samples collected in the exploration phase, for each x not in the final assortment A_{E_ℓ} there exists a $\pi(x) \in A_{E_\ell}$ such that $\sqrt{\beta}\|x - \pi(x)\|_{V_{E_\ell}^{-1}} \leq \frac{1}{12}\varepsilon_\ell$. Let F_ℓ be the design matrix of all the samples contained in round ℓ. We will control the final uncertainty between pairs as follows

$$\sqrt{\beta}\|x - y\|_{F_\ell^{-1}} \leq \sqrt{\beta}(\|x - \pi(x)\|_{F_\ell^{-1}} + \|\pi(x) - \pi(y)\|_{F_\ell^{-1}} + \|y - \pi(y)\|_{F_\ell^{-1}}). \tag{3}$$

Thus, if we can show that for all $z, w \in A_{E_\ell}$ we have $\|z - w\|_{F_\ell^{-1}} \leq \frac{1}{12}\varepsilon_\ell$, we see that for all $x, y \in A_\ell$ we have that (3) is bounded by $\frac{1}{4}\varepsilon_\ell$. Thus, we only need to 'push down' the worst case uncertainty among the pairs in the final assortment. We show that P_ℓ ranking samples with $P_\ell = \lceil c'\varepsilon_\ell^{-2}(d\log(T) + dS^2)(q-1)\rceil \leq \lceil c'\varepsilon_\ell^{-2}(d\log(T) + dS^2)\frac{d}{q}\rceil$ by Assumption 2.2, for an absolute constant $c' > 0$, suffices by a direct linear algebraic argument and foregoing the Elliptical potential lemma altogether. In case we have only fewer than q items remaining by elimination, the push-down phase by itself will ensure the desired accuracy.

Formally, we show the following result, under all the previously stated assumptions. The full proof can be found in Appendix C. The condition on λ is technical and needed to make the Elliptical potential lemma work, for details see Lemma C.1 and text above that Lemma.

Theorem 3.7. *Assume that Assumption 2.2 holds. Choose the regulariser to be $\lambda \in [2\sqrt{d}, 2d]$ and let $\sqrt{\beta} := \frac{C}{\dot{\mu}_{\min}}\left(\sqrt{d\log(dT^2) + \lambda S^2}\right)$, which is the width parameter with $\delta = \frac{1}{T^2}$ from Theorem 3.1. Then, Algorithm 1 incurs regret $R_T = \tilde{O}\left(\frac{1}{\dot{\mu}_{\min}}\frac{d}{\sqrt{q}}\sqrt{T}\right)$.*

Proof (sketch). We estimate sub-optimality gaps for all remaining items in a round. Using correctness of the confidence interval at the end of the round and that combining the exploration and push down phase gives a confidence up to an additive factor of $\frac{1}{4}\varepsilon_\ell$ for all pairs in $\mathcal{A}_{\ell-1}$, we can show that an that the optimal item will not be removed and all items with sub-optimality gap larger than ε_ℓ get removed by domination of the optimal item. The suboptimality gap in round ℓ is then at most $2\varepsilon_\ell$ We note that we may use our confidence interval as our selection strategy in a round is non-adaptive in round ℓ. Let R_ℓ be the time points in phase ℓ and denote by $|R_\ell|$ the length of the ℓ-th phase. Then, by the definition of the exploration length we can estimate the regret in the following way, using that an upper bound on each individual sub-optimality gap is an upper bound on the average, we estimate

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{q_t}\sum_{x\in A_t}\langle\theta^*, x^* - x\rangle &\leq 2\sqrt{T}\sqrt{\sum_{\ell=1}^{L}\sum_{t\in R_\ell}\varepsilon_\ell^2|R_\ell|} \\
&\leq C\sqrt{T}\sqrt{\sum_{\ell=1}^{L}\varepsilon_\ell^2\frac{1}{\dot{\mu}_{\min}^2}(d\log(dT) + dS^2)\left(\frac{d\log(T)}{q\varepsilon_\ell^2} + \frac{q}{\varepsilon_\ell^2}\right)} \\
&= O\left(\frac{1}{\dot{\mu}_{\min}}\frac{d}{\sqrt{q}}\sqrt{T}\log^{3/2}(dT)\right).
\end{aligned}$$

where in the first inequality we used Cauchy-Schwarz and in the O-estimate we used that $q \leq \sqrt{d}$, so that $q \leq \frac{d}{q}$ and $L \leq \log T$ and $C > 0$ is an absolute constant. □

Remark 3.8. While the primary focus of this work is theoretical, showing that improved statistical performance for regret minimisation under rank breaking is possible, we note that a potential issue is in its scalability due to the action space being of order $O(N^q)$. Due to elimination the primary computation costs are expected at the beginning of the learning process. A heuristic for improved computational efficiency is to perform maximum likelihood estimation on a single independence chain by using non-adaptive strategies inspired by duelling bandits. A more extensive discussion is given in Appendix E.

Numerical experiments highlighting different values of q are given in Appendix G. By the established upper bound on the regret of our algorithm, it is natural to ask how tight the regret bound is.

Algorithm 1 Multi-round algorithm for learning under ranking feedback

Input: Upper bound S of $\|\theta^*\|_2$, Arm set $\mathcal{A}$, bound on assortment size q, $\mathcal{A}_0 := \mathcal{A}$, horizon T and confidence width parameter $\sqrt{\beta} = \frac{C}{\dot{\mu}_{\min}}\sqrt{d\log(dT^2) + \lambda S^2}$, L is the number of rounds the algorithm will perform with choices E_ℓ and P_ℓ. Algorithm will stop when it reaches horizon.
for $\ell = 1$ **to** L **do**
  for $t = 1$ **to** E_ℓ **do**
    Play assortment $A_t := \arg\max_{A \in \binom{\mathcal{A}_{\ell-1}}{q}} \sigma_{t-1}(A)$ and receive ranking of A_t
  end for
  for $t = E_\ell + 1$ **to** $P_\ell + E_\ell$ **do**
    Play assortment A_{E_ℓ} and receive ranking feedback
  end for
  Compute estimator θ_ℓ based on feedback in round ℓ, and put $F_\ell = \lambda I + \frac{1}{q-1}\sum_{t=1}^{E_\ell+P_\ell}\sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$. Eliminate under performing arms from $\mathcal{A}_{\ell-1}$, that is arms $y \in \mathcal{A}_{\ell-1}$ such that for some remaining arm $x \in \mathcal{A}_{\ell-1}$

$$\langle \theta_\ell, y - x\rangle + \sqrt{\beta}\|x - y\|_{F_\ell^{-1}} < 0.$$

  Call the new set $\mathcal{A}_\ell$
end for

3.3 Lower bound

We show, for the multi-set variant of the problem under Gaussian noise, that the regret bound is not improvable. Note that we only play assortments smaller than q we have fewer than q items remaining. The deterministic push-down phase will still work to reduce the final uncertainty of all pairs. Our result is based on individual suboptimality gaps, so the analysis will still hold if we append extra items from the set $\mathcal{A}_{\ell-1}$. This makes it a reasonable comparison. We build on top of the lower bound by Saha [2021] for the duelling bandit problem and the lower bound for linear bandits under Gaussian noise in Chapter 24 by Lattimore and Szepesvári [2020]. The principal idea is that a ranking algorithm should, at timepoints a multiple of q, select q actions and then commit to playing them. The ranking can then be reconstructed from the noisy observations and thereby it includes any strategy that works solely on the sorted values. The lower bound for the unit ball scales with $d\sqrt{T}$ for linear bandit [Lattimore and Szepesvári, 2020]. We note that we essentially stretch time by a factor of q, and our regret is normalised by a factor of q, yielding a lower bound of order $\Omega(\frac{d}{\sqrt{q}}\sqrt{T})$.

Theorem 3.9. *Consider the random utility model with Gaussian noise. Suppose that the action set is $\mathcal{A} := B_2(0;1)$. For T sufficiently large depending only on d, we have that any algorithm that at time t is allowed to play multi-sets of size q incurs regret at least $\Omega(\frac{d}{\sqrt{q}}\sqrt{T})$ for an instance with $\|\theta^*\|_2 \leq 1$.*

We provide the formal proof in Appendix D. We note that this multi-set lower bound does not depend on Assumption 2.2, so that it is not possible to improve the bound in this regime. While in theory we have only provided an algorithm for finite item sets, a discretisation argument and the lower bound presented shows that, in general, our algorithm guarantee cannot be further tightened. This is because our confidence interval is *independent of the number of arms*.

4 Related Literature

In this section we review the literature most closely related to our work. We go over previous results on rank breaking, duelling bandits in the linear and kernelised setting, as well as the recent advances on learning from ranking feedback in case of linear utility.

Rank breaking Rank breaking, i.e. the principle of extracting all pairwise comparisons from a ranking, is a popular approach for learning from ranking feedback, both in theory and practice [Azari Soufiani et al., 2013, Soufiani et al., 2014, Hajek et al., 2014, Shah et al., 2015, Khetan and Oh, 2016, Negahban et al., 2018]. A rank breaking estimator is, for example, used for the alignment of

large language models [Ouyang et al., 2022]. In the case of linear utilities, Zhu et al. [2023] provide a self-normalising bound for rank breaking and MLE with the Plackett-Luce ranking probabilities.

Dueling Bandits Learning from preference feedback has been extensively studied for pairwise comparisons, i.e. duelling bandits [Yue and Joachims, 2009, 2011, Yue et al., 2012]. More closely related to our work is the work by Saha [2021] that learns from preferential feedback in the contextual setting with linear utilities. Saha [2021] achieves $\tilde{O}(d\sqrt{T})$ regret for the general arms case and $\tilde{O}(\sqrt{dT\log|\mathcal{A}|})$ for the finite arm case, with $|\mathcal{A}|$ an upper bound on the actions in a round. The linear utility setting for duelling bandits has been generalised to the kernelised setting [Xu et al., 2024, Pásztor et al., 2024, Kayal et al., 2025]. Here, the complexity scales with maximum information gain to the corresponding reproducing kernel Hilbert space utility function, instead of the dimension. Let Γ_T be the maximum information gain over a horizon of T. For the general arm set case a regret bound of $\tilde{O}(\Gamma_T\sqrt{T})$ was derived by Pásztor et al. [2024]. For finite action set, regret that scales as $\tilde{O}(\sqrt{\Gamma_T\log(|\mathcal{A}|)T})$ is derived by Kayal et al. [2025].

Ranking Feedback Chen et al. [2022] aim to learn from preferential feedback with general function approximation in the reinforcement learning setting. They also provide a bound from n-wise comparisons, but they model the feedback as pairwise feedback with Bernoulli random variables that are independent, but we do not assume that in our ranking model as we adhere to the random utility model. For the Plackett-Luce model in the multi-armed bandit version, a regret bound that decreases with top-m feedback is provided by Saha and Gopalan [2019a]. Saha and Gopalan [2019b] study PAC-learning from ranking feedback by maintaining pairwise preferential estimates. The paper by Sui et al. [2017] studies a related multi-dueling bandit problem, where only some pairwise comparisons may be seen, but no theoretical guarantees are given for the kernel case they consider.

There are also recent works that focus on learning from ranking feedback with linear utilities. This type of research has mostly focussed on the Plackett-Luce model [Mukherjee et al., 2024, Thekumparampil et al., 2025, Lee et al., 2025]. The works [Mukherjee et al., 2024, Thekumparampil et al., 2025] study optimal designs for ranking feedback. The work by Lee et al. [2025] is more closely related to ours. They aim to estimate the sub-optimality gaps in the contextual setting with stochastic context. They achieve a sub-optimality bound of $\tilde{O}(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}})$ with $|S_t|$ the size of the assortment offered at time t. However, their $|S_t|$ is also stochastic, which makes it unclear how good the regret bound is.

5 Conclusion and Discussion

We considered the regret minimization problem where the regret is the (expected) cumulative sum of the difference in the best utility and the average of the items offered at each time. We extend the winner feedback setting of Saha [2021] and show that, unlike for winner feedback, ranking feedback shows improvement with the maximum assortment size. Specifically, we show that it is possible to achieve improved regret minimisation of the order $\tilde{O}\left(\frac{d}{\sqrt{q}}\sqrt{T}\right)$ with the additional assumption $q \leq \sqrt{d}$, and thereby show that improved regret minimisation is possible in stark contrast to receiving winner feedback. We achieved this by using independence preserving rank breaking in the construction of confidence intervals. For a related multi-set variant of the problem we showed that, in certain cases, $\Omega(\frac{d}{\sqrt{q}}\sqrt{T})$ is unavoidable. We see a few extensions of this work. Firstly, it is natural to extend the analysis to the kernelised setting or a different form of function approximation. Secondly, relaxing $q \leq \sqrt{d}$ to $q \leq d$ (or removing the assumption) is also of interest. We think the former might be done by improving our analysis of pushing down the final uncertainty with a refined analysis rather than our worst case analysis, as the rest of the proof would go unaltered by the assumption $q \leq d$. Finally, it would also be of interest to show whether the push down phase can be eliminated by extending the exploration phase by similar order.

Broader societal impact There may be societal consequences of our work, but none we feel must be highlighted here.

Acknowledgements JO is supported by VENI grant VI.Veni.232.068 issued by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). NS thanks all anonymous reviewers of this work for their suggestions which helped improve the current paper.

References

M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 856–864, Bejing, China, 22–24 Jun 2014. PMLR.

H. Azari Soufiani, W. Chen, D. C. Parkes, and L. Xia. Generalized method-of-moments for rank aggregation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

F. Bach. The gumbel trick. `https://francisbach.com/the-gumbel-trick/`, 2019. Accessed: 2026-01-27.

V. Bengs, R. Busa-Fekete, A. E. Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3773–3793. PMLR, 17–23 Jul 2022.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, and J. Guo. A survey on llm-as-a-judge. *The Innovation*, page 101253, 2026. ISSN 2666-6758. doi: https://doi.org/10.1016/j.xinn.2025.101253.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

F. Harary. Graph theory. 1969.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

A. Kayal, S. Vakili, L. Toni, D. Shiu, and A. Bernacchia. Bayesian optimization from human feedback: Near-optimal regret bounds. In *Forty-second International Conference on Machine Learning*, 2025.

A. Khetan and S. Oh. Data-driven rank breaking for efficient rank aggregation. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 89–98, New York, New York, USA, 20–22 Jun 2016. PMLR.

W. Kool, H. Van Hoof, and M. Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3499–3508. PMLR, 09–15 Jun 2019.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

J. Lee, S. Yi, and M. hwan Oh. Preference-based reinforcement learning beyond pairwise comparisons: Benefits of multiple options. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

S. Mukherjee, A. Lalitha, K. Kalantari, A. Deshmukh, G. Liu, Y. Ma, and B. Kveton. Optimal design for human preference elicitation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 90132–90159. Curran Associates, Inc., 2024. doi: 10.52202/079017-2861.

S. Negahban, S. Oh, K. K. Thekumparampil, and J. Xu. Learning from comparisons and choices. *Journal of Machine Learning Research*, 19(40):1–95, 2018.

M. Oh and G. Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9205–9213, 2021.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

B. Pásztor, P. Kassraie, and A. Krause. Bandits with preference feedback: A stackelberg game perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. ISSN 00359254, 14679876.

A. Saha. Optimal algorithms for stochastic contextual preference bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30050–30062. Curran Associates, Inc., 2021.

A. Saha and A. Gopalan. Combinatorial bandits with relative feedback. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

A. Saha and A. Gopalan. Pac battling bandits in the plackett-luce model. In A. Garivier and S. Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 700–737. PMLR, 22–24 Mar 2019b.

N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 856–865, San Diego, California, USA, 09–12 May 2015. PMLR.

H. A. Soufiani, D. Parkes, and L. Xia. Computing parametric ranking models via rank-breaking. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 360–368, Bejing, China, 22–24 Jun 2014. PMLR.

Y. Sui, V. Zhuang, J. W. Burdick, and Y. Yue. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

Y. Tang and Y. Feng. Beyond pairwise: Empowering llm alignment with ranked choice modeling. *arXiv preprint arXiv:2510.23631*, 2025.

K. K. Thekumparampil, G. Hiranandani, K. Kalantari, S. Sabach, and B. Kveton. Comparing few to rank many: Active human preference learning using randomized frank-Wolfe method. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 59355–59376. PMLR, 13–19 Jul 2025.

J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2 edition, 2001.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

W. Xu, W. Wang, Y. Jiang, B. Svetozarevic, and C. N. Jones. Principled preferential bayesian optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1201–1208, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553527.

Y. Yue and T. Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 241–248, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.028. JCSS Special Issue: Cloud Computing 2011.

B. Zhu, M. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023.

A Details on Independence Grids and subgaussian estimates

In this section, we provide an example of an independence grid and explicitly show how an independence grid can be used to obtain a better subgaussian estimate. We first give an example of an independence grid.

Example A.1. *Let $q = 6$, then the following grid provides an independence grid for $[q]$, where each row is an independence chain.*

$$(1,2)(3,4)(5,6)$$
$$(1,3)(2,5)(4,6)$$
$$(1,4)(2,6)(3,5)$$
$$(1,5)(2,4)(3,6)$$
$$(1,6)(2,3)(4,5)$$

Let us now provide a bound on the subgaussian norm of a stochastic random vector that appears when we prove confidence intervals. Recall that a mean zero random variable is called K-*subgaussian* if its moment generating function satisfies $\mathbb{E}(\exp(\lambda X)) \leq \exp(\lambda^2 K^2)$ for all $\lambda \in \mathbb{R}$. A mean zero random vector $X \in \mathbb{R}^d$ is called K-subgaussian if for every $x \in \mathbb{S}^{d-1}$ the one-dimensional projection of X onto x is K-subgaussian. The number K in this definition is called the *subgaussian* norm, which we will denote $\|X\|_{\psi_2}$.

Theorem A.1. *Let q_t be even and X be the random vector*

$$\left(1\{x_{t,i} \succ x_{t,j}\} - \mu(\langle \theta^*, x_{t,i} - x_{t,j}\rangle)\right)_{(\{i,j\} \in \binom{[q_t]}{2})}.$$

Then, this random vector is $O(\sqrt{q_t - 1})$ subgaussian.

Proof. We note first that each component is bounded by 1, which makes them constant order subgaussian. Then, let $\{R_i\}_{i=1}^{q_t - 1}$ be an independence grid associated to $[q_t]$. Let z be a unit vector. By the triangle inequality, we can estimate the subgaussian norm as

$$\|z^T X\|_{\psi_2} = \sum_{i=1}^{q-1} \|z^T X(R_i)\|_{\psi_2}. \tag{4}$$

Next, we note that, as each component is clearly mean zero, that the one-dimensional projection must also be mean zero. Furthermore, by using that $\{R_i\}$ is an independence grid, we have that each partial random vector has independent coordinates. Thus, by means of rotational invariance, Proposition 2.6.1 [Vershynin, 2018], we obtain

$$\|z^T X(R_i)\|_{\psi_2}^2 \leq C \sum_{p \in R_i} z_p^2 \|X_p\|_{\psi_2}^2 \leq C' \sum_{p \in R_i} z_p^2$$

where the first inequality follows from Proposition 2.6.1 [Vershynin, 2018] and the second inequality follows from the boundedness by one of the components, which makes each X_p constant order subgaussian. Therefore, we have that

$$\|z^T X(R_i)\|_{\psi_2} \leq C' \|z(R_i)\|_2.$$

By plugging this bound into (4) and using the Cauchy-Schwarz inequality, we obtain the desired result

$$\|z^T X\|_{\psi_2} \leq C' \sqrt{q_t - 1} \|z\|_2 \leq C' \sqrt{q_t - 1}$$

as $\|z\|_2 = 1$. This concludes the proof. □

The case of odd q_t still has the same order due to Remark 3.5.

B Formal proof of confidence interval

In this section, we formally prove our confidence interval. The proof is mainly inspired by that by Filippi et al. [2010]. For convenience, we provide the full statement again and afterwards give its proof.

Theorem B.1. *Let $\mathcal{X} := \{A_t = \{x_{t,i}\}_{i=1}^{q_t}\}_{t=1}^{T}$, with $q_t \leq q$ for all time points, be an offline sampling strategy.*

We define the gradient space function as

$$g(\theta) := \lambda(q-1)\theta + \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \mu(\langle \theta, x_{t,i} - x_{t,j} \rangle)(x_{t,i} - x_{t,j}).$$

Let θ_{rb} be the rank breaking approximation of the true parameter vector in the case of rank breaking loss, which we define as a minimiser of the expression

$$\|g(\theta) - \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \mathbb{I}(x_{t,i} > x_{t,j})(x_{t,i} - x_{t,j})\|_{((q-1)^2 V_T)^{-1}}$$

over the ball $B_2^d(0; S)$ where we define $V_T := \lambda I + \frac{1}{q-1} \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$. Then, we have for all $x \in \mathbb{R}^d$ the following inequality

$$|\langle \theta_{rb} - \theta^*, x \rangle| \leq C \frac{1}{\dot{\mu}_{\min}} \left(\sqrt{d \log\left(\frac{d}{\delta}\right)} + \sqrt{\lambda} S \right) \|x\|_{V_T^{-1}}$$

with probability at least $1 - \delta$ for an absolute $C > 0$.

Proof. We aim to estimate the quantity $|\langle \theta_{\mathrm{rb}} - \theta^*, x \rangle|$ uniformly over all x. We start by noting that by the mean value theorem we have the following expression

$$\begin{aligned} &g(\theta^*) - g(\theta_{\mathrm{rb}}) = \lambda(q-1)(\theta^* - \theta_{\mathrm{rb}}) \\ &+ \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \int_0^1 \dot{\mu}(\langle r\theta^* + (1-r)\theta_{rb}, x_{t,i} - x_{t,j} \rangle) \mathrm{d}r \langle \theta^* - \theta_{rb}, x_{t,i} - x_{t,j} \rangle (x_{t,i} - x_{t,j}). \end{aligned}$$

Let us now define the matrix

$$G_t := \lambda(q-1)I + \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \int_0^1 \dot{\mu}(\langle r\theta^* + (1-r)\theta_{rb}, x_{t,i} - x_{t,j} \rangle) \mathrm{d}r (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T.$$

As we have assumed a lower bound on the derivative of the link function and *both* θ^* and θ_{rb} are S-bounded in Euclidean norm (θ^* satisfies the bound by assumption and θ_{rb} satisfies the bound by construction) and the set $\mathcal{A}$ is contained in the unit ball of radius one (so that the difference of the two vectors is contained in the ball of radius two), we see that for all $r \in [0, 1]$, $t = 1, \ldots, T$ and i, j we have by Cauchy-Schwarz

$$|\langle r\theta^* + (1-r)\theta_{\mathrm{rb}}, x_{t,i} - x_{t,j} \rangle| \leq 2S.$$

By the assumption that $\dot{\mu}_{\min} = \min\{1, \min_{x \in [-2S, 2S]} \dot{\mu}(x)\} > 0$, we see that this matrix is positive definite, as each of the integral terms are lower bounded by $\dot{\mu}_{\min}$.

Then, we can express the difference in θ^* and θ_{rb} as follows:

$$\theta^* - \theta_{\mathrm{rb}} = G_T^{-1}(g(\theta^*) - g(\theta_{\mathrm{rb}})).$$

It follows that for any $x \in \mathbb{R}^d$, we can express the inner product as

$$|\langle \theta^* - \theta_{\mathrm{rb}}, x \rangle| = \left| G_T^{-1}(g(\theta^*) - g(\theta_{\mathrm{rb}})), x \rangle \right|.$$

By the Cauchy-Schwarz inequality and multiplying by $\frac{\sqrt{q-1}}{\sqrt{q-1}}$, we can upper bound this quantity by

$$\|g(\theta^*) - g(\theta_{\mathrm{rb}})\|_{((q-1)G_T)^{-1}} \|x\|_{(\frac{1}{q-1}G_T)^{-1}}.$$

In order to have a single good event, we lower bound the matrix G_T from below in Loewner order. As we assume that the link function satisfies $\min_{x\in[-2S,2S]} \dot{\mu}(x) \geq \dot{\mu}_{\min}$ we have that the matrix G_t can be lower bounded by the deterministic matrix

$$D_T := \lambda(q-1)I + \dot{\mu}_{\min} \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$$

Because both matrices are positive definite and $G_T \succeq D_T$ with D_T invertible, we can bound the weighted matrix norms appearing in the upper bound thus far by D_t. Doing this result in the following bound:

$$\|g(\theta^*) - g(\theta_{\mathrm{rb}})\|_{((q-1)D_T)^{-1}} \|x\|_{(\frac{1}{q-1}D_T)^{-1}}. \tag{5}$$

We note that we may rewrite D_T as the following matrix

$$\begin{aligned} D_T &= \dot{\mu}_{\min}(\lambda(q-1)\frac{1}{\dot{\mu}_{\min}}I + \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T \\ &\succeq \dot{\mu}_{\min}(\lambda(q-1)I + \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T, \end{aligned} \tag{6}$$

where the Loewner order lower bound follows by $\dot{\mu}_{\min} \leq 1$. Substituting the lower bound obtained in (6) into (5) and recalling that we defined $V_T := \lambda I + \frac{1}{q-1}\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$, we see that we must upper bound

$$|\langle \theta^* - \theta_{\mathrm{rb}}, x\rangle| \leq \frac{1}{\dot{\mu}_{\min}} \|x\|_{V_T^{-1}} \|g(\theta^*) - g(\theta_{\mathrm{rb}})\|_{((q-1)^2V_T)^{-1}}. \tag{7}$$

We use the triangle inequality and the defining property of θ_{rb} to eliminate the $g(\theta_{\mathrm{rb}})$ term in our bound. We note that by the triangle inequality we may estimate

$$\begin{aligned} \|g(\theta^*) - g(\theta_{\mathrm{rb}})\|_{((q-1)^2V_T)^{-1}} &\leq \|g(\theta^*) - \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} \mathbb{I}(x_{t,i} \succ x_{t,j})(x_{t,i} - x_{t,j})\|_{((q-1)^2V_T)^{-1}} \\ &\quad + \|\sum_{s=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} \mathbb{I}(x_{t,i} \succ x_{t,j})(x_{t,i} - x_{t,j}) - g(\theta_{\mathrm{rb}})\|_{((q-1)^2V_T)^{-1}} \\ &\leq 2\|g(\theta^*) - \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} \mathbb{I}(x_{t,i} \succ x_{t,j})(x_{t,i} - x_{t,j})\|_{((q-1)^2V_T)^{-1}}, \end{aligned} \tag{8}$$

where the last inequality follows by the minimising property of θ_{rb} and the fact that $\|\theta^*\|_2 \leq S$.

Recalling the definition of $g(\theta)$, we will use one more triangle inequality to further bound this expression. We note that by the triangle inequality we have that

$$\begin{aligned} &\|\lambda(q-1)\theta^*\|_{((q-1)^2V_T)^{-1}} \\ &+ \left\| \sum_{t=1}^{T} \sum_{\{i,j\}\in\binom{[q_t]}{2}} (\mu(\langle \theta^*, x_{t,i} - x_{t,j}\rangle) - \mathbb{I}(x_{t,i} \succ x_{t,j}))(x_{t,i} - x_{t,j}) \right\|_{((q-1)^2V_T)^{-1}}. \end{aligned} \tag{9}$$

We can easily bound the first term by $\sqrt{\lambda}S$. Indeed, we note that due to the additional factor of $(q-1)$ over the standard Hessian matrix, we can completely eliminate the factor $q-1$ in the first term, we see that we can upper bound this quantity by $\sqrt{\lambda}S$ by the Loewner order lower bound $(q-1)^2 V_T \succeq (q-1)^2 \lambda I$.

We now define

$$m_t := \sum_{\{i,j\} \in \binom{[q_t]}{2}} (\mu(\langle \theta^*, x_{t,i} - x_{t,j} \rangle) - \mathbb{I}(x_{t,i} \succ x_{t,j}))(x_{t,i} - x_{t,j}).$$

We can now rewrite the weighted matrix norm of the of the m_t terms as follows:

$$\begin{aligned}
\left\| \sum_{t=1}^T m_t \right\|^2_{((q-1)^2 V_T)^{-1}} &= \left\| ((q-1)^2 V_T)^{-1/2} \left(\sum_{s=1}^T m_t \right) \right\|_2^2 \\
&= \sum_{i=1}^d \left| e_i^T ((q-1)^2 V_T)^{-1/2} \left(\sum_{s=1}^T m_t \right) \right|^2 \\
&= \sum_{i=1}^d \langle ((q-1)^2 V_T)^{-1/2} e_i, \sum_{t=1}^T m_t \rangle^2
\end{aligned}$$

We therefore analyse the tail behaviour of the random variable $\langle z, \sum_{t=1}^T m_t \rangle = \sum_{s=1}^T \langle z, m_t \rangle$. Note that as m_t is mean zero, then so is the one dimensional projection $\langle z, m_t \rangle$. We observe that by Theorem A.1 that

$$\| \langle z, m_t \rangle \|_{\psi_2} \leq C_0 \sqrt{q-1} \| z \|_{\sum_{\{i,j\} \in \binom{[q_t]}{2}} (x_{s,i} - x_{s,j})(x_{s,i} - x_{s,j})^T} \tag{10}$$

subgaussian with $C_0 > 0$ an absolute constant. Introduce the matrix $X_T := \sum_{t=1}^T \sum_{\{i,j\} \in \binom{[q_t]}{2}} (x_{t,i} - x_{t,j})(x_{t,i} - x_{t,j})^T$. By the offline sampling strategy we now deal with independent mean zero random variables, so that again by Proposition 2.6.1 in [Vershynin, 2018] and (10) the random variable $\langle z, \sum_{s=1}^T m_s \rangle$ is

$$C\sqrt{q-1} \| z \|_{X_T} \leq C\sqrt{q-1} \| z \|_{(q-1)V_T}$$

subgaussian for some absolute constant $C > 0$, where the inequality follows by the definition of V_T and $(q-1)V_T \succeq X_T$ as $X_T + (q-1)\lambda I = (q-1)V_T$.

Next, we observe that for the choice of $z_i = ((q-1)^2 V_T)^{-1/2} e_i$, where (e_i) is the standard basis, we see that

$$C\sqrt{q-1} \| z_i \|_{(q-1)V_T} \leq C\sqrt{q-1} \| ((q-1)V_T)^{1/2} \frac{1}{\sqrt{q-1}} ((q-1)V_T)^{-1/2} e_i \|_2 \leq C,$$

where $C > 0$ is an absolute constant.

Thus, by the definition of subgaussian random variables, we have the following tail inequality

$$\mathbb{P}\left(\left| \left\langle z_i, \sum_{t=1}^T m_s \right\rangle \right| \geq r \right) \leq 2 \exp(-cr^2),$$

for $c > 0$ and absolute constant. Thus, we have for $r = c' \sqrt{\log d/\delta}$ with c' an absolute constant that the probability of this event is bounded above by $\frac{\delta}{d}$.

Now, by a standard decoupling argument, we can estimate now

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{s=1}^{T} m_s\right\|^2_{((q-1)^2V_T)^{-1}} \geq d(c')^2\log(d/\delta)\right) &= \mathbb{P}\left(\sum_{i=1}^{d}\langle z_i, \sum_{s=1}^{T} m_s\rangle^2 \geq d(c')^2\log(d/\delta)\right)\\ &\leq \sum_{i=1}^{d}\mathbb{P}\left(|\langle z_i, \sum_{s=1}^{T} m_s\rangle|^2 \geq (c')^2\log(d/\delta)\rangle\right)\\ &\leq \sum_{i=1}^{d}\frac{\delta}{d}\\ &\leq \delta.\end{aligned}$$

Thus, we now see, by combining (7), (8) and (9) and the complement of the event above, which happens with probability at least $1-\delta$, that for all $x \in \mathbb{R}^d$ (as the event is x independent) that

$$|\langle \theta^* - \theta_{\mathrm{rb}}, x\rangle| \leq \frac{1}{\dot{\mu}_{\min}}\left(\sqrt{\lambda}S + c'\sqrt{d\log(d/\delta)}\right)\|x\|_{V_T^{-1}}.$$

Therefore, the theorem is proven.

□

Remark B.2. We note that the estimate $\int_0^1 \dot{\mu}(\langle r\theta^* + (1-r)\theta_{rb}, x_{t,i} - x_{t,j}\rangle)\mathrm{d}r \geq \dot{\mu}_{\min}$ is worst-case. Under assumptions on the shape of μ, such as self-concordance, this might be improvable. In particular, in each phase ℓ we have $|\langle\theta^*, x-y\rangle| = |\langle\theta^*, x-x^*\rangle| + |\langle\theta^*, x^*-y\rangle| \leq 4\varepsilon_\ell$ if all the previous confidence intervals hold, for x, y in the current round. This can potentially be used with results in Section F from the paper byAbeille et al. [2021] for the case of the Plackett-Luce model. We note that Abeille et al. [2021] do not study regret in terms of utilities, but rather regret is defined through the logistic link function.

C Formal regret analysis

In this section, we study the formal regret analysis of our multi-round algorithm. Before we can delve into the details of our algorithm, we first give a version of the elliptical potential lemma as first derived by Oh and Iyengar [2021]. We provide proof for the sake of self-containedness as small details need to be changed for our design matrix, but the proof mostly follows that by Oh and Iyengar [2021].

C.1 Elliptical potential for rank breaking

Here, we recall the following form of the Elliptical potential lemma we need in order to finish our form of the proof. Note that $\lambda \geq 2\sqrt{d}$ ensures $\lambda \geq 2q$ by Assumption 2.2. This version of Elliptical potential is by Oh and Iyengar [2021].

Lemma C.1. *(Adapted from 7 [Oh and Iyengar, 2021, Lemma 7]) Let $\lambda \geq 2q$ and $V_T := \lambda I + \sum_{t=1}^{T}\frac{1}{q-1}\sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T$. Assume that $x_{t,i}$ lies in the Euclidean unit ball for all t, i and $q_t \leq q$ for all t. Then, we have that*

$$\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}\|x_{t,i}-x_{t,j}\|^2_{V_{t-1}^{-1}} \leq 2d(q-1)\log\left(1+\frac{4Tq^2}{\lambda(q-1)d}\right). \tag{11}$$

Proof. We start by putting

$$W_T = (q-1)V_T = \lambda(q-1) + \sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T.$$

We then observe that we can rewrite the elliptical potential sum in the following way:

$$\sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{V_{t-1}^{-1}} = \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{(\frac{q-1}{q-1} V_{t-1})^{-1}}$$
$$= (q-1) \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}}. \tag{12}$$

It is therefore sufficient to control the elliptical potential sum on the right hand side. We start by observing that as our regularizer is at least $2q$, that we can bound each of the terms $\sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}}$ by 1. Indeed, we compute

$$\begin{aligned}
\sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} &\leq \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{(\lambda(q-1))^{-1} I} \\
&= \frac{1}{\lambda(q-1)} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|_2^2 \\
&\leq \frac{4}{\lambda(q-1)} \frac{(q_t - 1) q_t}{2} \\
&\leq \frac{4}{2q(q-1)} \frac{(q_t - 1) q_t}{2} \\
&\leq \frac{(q_t - 1) q_t}{q(q-1)} \\
&\leq 1.
\end{aligned}$$

In the first inequality we used the Loewner order lower bound $\lambda(q-1)I \preceq W_{t-1}$. In the third line we used that $\|x_{t,i}\|_2 \leq 1$, and in the final line we use that $q_t \leq q$ for all q.

We conclude that under the current hypothesis we have

$$(q-1) \sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} = (q-1) \sum_{t=1}^{T} 1 \wedge \left(\sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} \right).$$

We now proceed to bound the summation on the right. From this point, we mostly follow the proof of Lemma 7 in [Oh and Iyengar, 2021].

From the standard inequality $x \leq 2 \log(1 + x)$ valid for any $x \in [0, 1]$, we see that

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} &\leq 2 \sum_{t=1}^{T} \log \left(1 + \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} \right) \\
&= 2 \log \left(\Pi_{t=1}^{T} \left(1 + \sum_{\{i,j\} \in \binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} \right) \right). \qquad (13)
\end{aligned}$$

We will show that $\frac{\det(W_T)}{\det(\lambda(q-1)I)}$ is an upper bound for the term inside the logarithm. We proceed by the following inductive argument. Start by observing the following equalities due to the determinant

being multiplicative.

$$\begin{aligned}\det(W_T) &= \det\left(W_{T-1} + \sum_{\{i,j\}\in\binom{[q_T]}{2}} (x_{T,i}-x_{T,j})(x_{T,i}-x_{T,j})^T\right)\\ &= \det(W_{T-1})\det\left(I + \sum_{\{i,j\}\in\binom{[q_T]}{2}} W_{T-1}^{-1/2}(x_{T,i}-x_{T,j})(x_{T,i}-x_{T,j})^T W_{T-1}^{-1/2}\right)\\ &= \det(\lambda(q-1)I)\Pi_{t=1}^{T}\det\left(I + \sum_{\{i,j\}\in\binom{[q_t]}{2}} W_{t-1}^{-1/2}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}\right)\end{aligned}$$

where the last step follows by recursively applying the idea behind the first two equalities. It follows that

$$\frac{\det(W_T)}{\det(\lambda(q-1)I)} = \Pi_{t=1}^{T}\det\left(I + W_{t-1}^{-1/2}(x_{t,i}-x_{T,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}\right).$$

Therefore, all that needs to be shown is that $1+\sum_{\{i,j\}\in\binom{[q_t]}{2}} \|x_{t,i}-x_{t,j}\|^2_{W_{t-1}^{-1}}$ can be upper bounded by

$$\det\left(I + \sum_{\{i,j\}\in\binom{[q_t]}{2}} W_{t-1}^{-1/2}(x_{t,i}-x_{T,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}\right).$$

We now proceed with its proof. Observe that the matrix $\sum_{\{i,j\}\in\binom{[q_t]}{2}} W_{t-1}^{-1/2}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}$ is positive semi-definite with eigenvalues λ_i with $i=1,\dots,d$. Moreover, we use that it is simultaneously diagonisable with the identity matrix (as any positive semi-definite matrix is) with eigenvalues $1+\lambda_i$. Therefore, we see that

$$\begin{aligned}&\det\left(I + \sum_{\{i,j\}\in\binom{[q_t]}{2}} W_{t-1}^{-1/2}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}\right)\\ &= \Pi_{i=1}^{d}(1+\lambda_i)\\ &\geq 1+\sum_{i=1}^{d}\lambda_i\\ &= 1-d+\sum_{i=1}^{d}(1+\lambda_i)\\ &= 1-d\\ &+\operatorname{trace}\left(I + \sum_{\{i,j\}\in\binom{[q_t]}{2}} W_{t-1}^{-1/2}(x_{t,i}-x_{t,j})(x_{t,i}-x_{t,j})^T W_{t-1}^{-1/2}\right)\\ &= 1+\sum_{\{i,j\}\in\binom{[q_t]}{2}} \|x_{t,i}-x_{t,j}\|^2_{W_{t-1}^{-1}}. \qquad (14)\end{aligned}$$

In this last step we used that $W_{t-1}^{-1/2}$ is a symmetric matrix and the cyclic property of the trace. This proves the claim. We now turn to the final step of the proof, which is to give a suitable upper bound on $\log\left(\frac{\det(W_T)}{\det(\lambda(q-1)I)}\right)$. Note that evidently $\det(\lambda(q-1)I)) = (\lambda(q-1))^d$. To give an upper bound

on $\det(W_T)$, we start by noting that by the arithmetic-geometric inequality we have

$$\begin{aligned}
\det(W_T) &\leq \left(\frac{\operatorname{trace}(W_T)}{d}\right)^d \\
&\leq \left(\frac{\lambda(q-1)d + \sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|_2^2}{d}\right)^d \\
&\leq \left(\frac{\operatorname{trace}(W_T)}{d}\right)^d \\
&\leq \left(\frac{\lambda(q-1)d + \sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|_2^2}{d}\right)^d \\
&\leq \left(\frac{\lambda(q-1)d + \frac{q^2}{2}T\cdot 4}{d}\right)^d
\end{aligned}$$

Thus, we can now upper bound $\frac{\det(W_T)}{\det(\lambda(q-1)I)}$ as follows:

$$\begin{aligned}
\frac{\det(W_T)}{\det(\lambda(q-1)I)} &\leq \frac{\left(\frac{\lambda(q-1)d + \frac{q^2}{2}T\cdot 4}{d}\right)^d}{(\lambda(q-1))^d} \\
&= \left(\frac{\lambda(q-1)d + \frac{q^2}{2}T\cdot 4}{d(q-1)\lambda}\right)^d \\
&\leq \left(1 + \frac{q^2}{2d(q-1)\lambda}T\cdot 4\right)^d \qquad (15)
\end{aligned}$$

Therefore, by combining (13), (14), and (15) and we conclude that

$$\sum_{t=1}^{T}\sum_{\{i,j\}\in\binom{[q_t]}{2}} \|x_{t,i} - x_{t,j}\|^2_{W_{t-1}^{-1}} \leq d\log\left(1 + \frac{2q^2}{d(q-1)\lambda}T\right). \qquad (16)$$

The bound from the theorem statement now follows follows by (12) and (16). □

C.2 Sufficient condition for sampling complexity

Let us now recall the upper confidence index for our confidence interval, which was of the form $\sqrt{\beta}\|x\|_{V_t^{-1}}$. Recall that $\sqrt{\beta} = C\sqrt{\lambda S^2 + d\log(T^2 d)}$ if we set $\delta = \frac{1}{T^2}$ in our confidence bound (this will be the level that we require for the proof). Thus, the sampling complexity to get the upper bound bound below a given threshold is then given by how quickly we can decrease the weighted norm $\|x\|_{V_t^{-1}}$. The key insight is that the Elliptical potential lemma at each time step (according to our algorithm definition) sums $\frac{q(q-1)}{2}$ pairwise uncertainties, but the upper bound only scales linearly in $q-1$. We will exploit this to show that search length to get below a desired threshold is accelerated by $\frac{1}{q}$. Below we give the proof of this result. We start out with an elementary lemma.

Lemma C.2. *Suppose that $\sum_{i=1}^{q} x_i \leq q\varepsilon$ for some $\varepsilon > 0$. Then there exists an index $i \in [q]$ such that $x_i \leq \varepsilon$*

Proof. Suppose that for all $x_i > \varepsilon$. Then $\sum_{i=1}^{q} x_i > q\varepsilon$, which contradicts the upper bound. □

We now want to bound the number of samples we need to get to get an upper bound on the desired threshold. We will do this in two steps. First, we will get an estimate on the number of samples that

is needed to get two times the uncertainty of all sets of size q in the current round, that is $\sigma_{t-1}(A)$ below $\frac{1}{12^2}q(q-1)\varepsilon_\ell^2$. Subsequently, we prove that by pushing down the final assortment played, we can get the uncertainty of all pairwise difference $x-y$ for $x,y \in A_{\ell-1}$ below $\frac{1}{4}\varepsilon_\ell$. The projection works when we have at least q items remaining. If this is not the case, we can use our final pushing down phase on the entire collection to obtain the desired uncertainty estimates for the difference of two vectors, with the push down length bounded by Assumption 2.2. Let us now start off with the sampling complexity to get the uncertainties below the desired threshold of $\frac{1}{12^2}q(q-1)\varepsilon^2$.

Proposition C.3. *Let $\lambda \geq 2q$. Let $\mathcal{A}_{\ell-1}$ be the active set at round ℓ, and let $A \subset \mathcal{A}_{\ell-1}$ be of size q Then, the number of samples according to our greedy maximisation rule that is needed achieve*

$$\beta \sum_{x\in A} \sum_{\substack{y\in A \\ x\neq y}} \|x-y\|^2_{V_t^{-1}} \leq q(q-1)\varepsilon^2$$

is satisfied upon the following inequality

$$N \geq \frac{2\beta d \log(1+N)}{q\varepsilon^2}.$$

Proof. We start by noting that $2\sigma_{t-1}(A_t) = \sum_{x\in A_t} \sum_{\substack{y\in A_t \\ y\neq x}} \|x-y\|^2_{V_{t-1}^{-1}}$. By Lemma C.1 we have that for any time that our sampling strategy gives

$$\beta \sum_{s=1}^{t} 2\sigma_{s-1}(A_s) \leq \beta 2d(q-1)\log(1+t)$$

As the uncertainties are monotonically decreasing over time, we can also lower bound the quantity on the left by $t\beta\sigma_{t-1}(A_t)$. By combining these two inequalities, we see that our uncertainty multiplied by our confidence width parameter β by sampling t point for our sampling strategy is upper bounded by

$$\frac{2\beta d(q-1)\log(1+t)}{t}$$

To get this bound below $q(q-1)\varepsilon^2$, we solve for t. We then obtain $t \geq \frac{2\beta d \log(1+t)}{q\varepsilon^2}$. As the final assortment was the maximiser, we conclude that the bound holds uniformly over all sets of size q. □

The main reason we aim to provide such a bound is due to the following lemma.

Lemma C.4. *Let A be a set of size q. Suppose we have that $\beta \sum_{x\in A} \sum_{\substack{y\in A \\ y\neq x}} \|x-y\|^2_{V_{t-1}^{-1}} \leq q(q-1)\varepsilon^2$, then there exists an $x \in A$ such that $\beta \sum_{\substack{y\in A \\ y\neq x}} \|x-y\|^2_{V_{t-1}^{-1}} \leq (q-1)\varepsilon^2$.*

Proof. This is direct application of Lemma C.2. Indeed, write $S_x := \beta \sum_{\substack{y\in A \\ y\neq x}} \|x-y\|^2_{V_{t-1}^{-1}}$. By hypothesis, we have $\sum_{x\in A} S_x \leq q((q-1)\varepsilon^2)$. Thus, by applying Lemma C.2 we find that there exists an x such that $S_x \leq (q-1)\varepsilon^2$, which is precisely what had to be proven. □

Now, we observe that if A is the set that maximises the final expression, we obtain a basic inequality in the uncertainties by directly comparing them. Indeed, suppose we have $x' \notin A$. Then, we define A' as A with x replaced by x'. Then, by the maximising property of A, we have that

$$\sum_{y\in A'} \sum_{\substack{z\in A' \\ z\neq y}} \|z-y\|^2_{V_{t-1}^{-1}} \leq \sum_{y\in A} \sum_{\substack{z\in A \\ z\neq y}} \|z-y\|^2_{V_{t-1}^{-1}}.$$

By rearranging the terms such that any two items which are not already present in both, we find that

$$\beta \sum_{\substack{y\in A \\ y\neq x}} \|x'-y\|^2_{V_{t-1}^{-1}} \leq \beta \sum_{\substack{y\in A \\ y\neq x}} \|x-y\|^2_{V_{t-1}^{-1}}.$$

Recall that by Lemma C.4 we have that the latter sum is upper bounded by $(q-1)\varepsilon^2$. By applying Lemma C.2 we see that there exists a point $y \in A$ such that $\beta\|x'-y\|^2_{V_{t-1}^{-1}} \leq \varepsilon^2$. We conclude that

any x' not present in the final set has a projected point $\pi(x) \in A$ in the final set such that the squared upper confidence bound parameter is at most ε, that is $\sqrt{\beta}\|\pi(x) - x\|_{V_t^{-1}} \leq \varepsilon$.

The final ingredient we is to prove that, by sampling at most $\beta q \varepsilon^{-2}$ actions, we can also push down the uncertainty of the pairs in the final assortment. Its proof is provided in the following lemma.

Lemma C.5. *By sampling the final action A_{E_ℓ} for $P_\ell = \lceil \beta q \varepsilon^{-2} \rceil$ times, we have for any pair $x, y \in A_{E_\ell}$ that*

$$\beta \|x - y\|^2_{(V_{E_\ell} + P)^{-1}} \leq \varepsilon^2,$$

with V_{E_ℓ} the matrix of the original samples and P the matrix coming from playing the final assortment P_ℓ times.

Proof. Let $A_{E_\ell} = \{x_i\}_{i=1}^{|A_{E_\ell}|}$ Recall that the design matrix grows by $\frac{1}{q-1}\sum_{\{i,j\} \in \binom{[|A_{E_\ell}|]}{2}} (x_i - x_j)(x_i - x_j)^T$ by playing A_{E_ℓ}. Thus, by the condition on P_ℓ, we obtain that the weighted matrix is lower bounded in Loewner order by $\lambda I + \frac{\beta}{\varepsilon^2}\sum_{\{i,j\} \in \binom{[|A_{E_\ell}|]}{2}} (x_i - x_j)(x_i - x_j)^T$, which in turn is bounded below in Loewner order by $X_{i,j} := \lambda I + \frac{\beta}{\varepsilon^2}(x_i - x_j)(x_i - x_j)^T$ for any pair i, j. Note that for any pair, we have that $x_i - x_j$ is an eigenvector with eigenvalue $\lambda + \frac{\beta}{\varepsilon^2}\|x_i - x_j\|_2^2$ of $X_{i,j}$. It follows that $X_{i,j}^{-1}(x_i - x_j) = \frac{1}{\lambda + \frac{\beta}{\varepsilon^2}\|x_i - x_j\|_2^2}(x_i - x_j)$. Therefore, it follows that

$$\|x_i - x_j\|^2_{X_{i,j}^{-1}} \leq \frac{\|x_i - x_j\|_2^2}{\lambda + \frac{\beta}{\varepsilon^2}\|x_i - x_j\|_2^2} \leq \frac{\varepsilon^2}{\beta}$$

Multiplying both sides by β and using that $X_{i,j}$ was a lower bound in Loewner order concludes the proof. □

Lemma C.6. *By combining both phases, we have that the uncertainty in all pairs from $\mathcal{A}_\ell$ is at most 3ε.*

Proof. We can distinguish three cases. As defined in the main text, let F_ℓ be the final matrix from the data collected at round ℓ. We note that $F_\ell \succeq V_{E_\ell - 1}, P + \lambda I$, which means that all the upper bounds in ε also hold when we weigh with F_ℓ.

Case 1, both items are in the final assortment played In this case, the final pushing down phase ensures that the width with respect to the data collected in that phase is at most ε.

Case 2, one item is not in the final assortment and the other item is inside Let x be the item that is outside the final assortment and y be the item that is inside the final assortment. Then, there exists a point $\pi(x)$ such that $\|x - \pi(x)\|_{(F_\ell)^{-1}} \leq \varepsilon$. It follows by the triangle inequality that

$$\begin{aligned}
\sqrt{\beta}\|x - y\|_{(F_\ell)^{-1}} &\leq \sqrt{\beta}\|x - \pi(x) + \pi(x) - y\|_{(F_\ell)^{-1}} \\
&\leq \sqrt{\beta}\|x - \pi(x)\|_{(F_\ell)^{-1}} + \sqrt{\beta}\|\pi(x) - y\|_{(F_\ell)^{-1}} \\
&\leq 2\varepsilon.
\end{aligned}$$

Case 3, both items are not in the final assortment: Similarly, both points have a projection point in the final assortment played. Thus, we can estimate using the triangle inequality and case 1.

$$\begin{aligned}
\sqrt{\beta}\|x - y\|_{(F_\ell)^{-1}} &\leq \sqrt{\beta}\|x - \pi(x) + \pi(x) - \pi(y) + \pi(y) - y\|_{(F_\ell)^{-1}} \\
&\leq \sqrt{\beta}\left(\|x - \pi(x)\|_{(F_\ell)^{-1}} + \|\pi(x) - \pi(y)\|_{(F_\ell)^{-1}} + \|\pi(y) - y\|_{(F_\ell)^{-1}}\right) \\
&\leq 3\varepsilon.
\end{aligned}$$

By combining all the cases, we conclude this lemma. □

We have now established sufficient conditions to get the confidence width below 3ε. By adjusting $\varepsilon \leftarrow \varepsilon/3$ (which increases the exploration length by a constant order), we get the bounds on both phases in round ℓ as presented in the main text. We will use the above with $\varepsilon = \frac{1}{4}\varepsilon_\ell$, which will we then shift to $\frac{1}{12}\varepsilon_\ell$ by the readjusting step to get the final bound stated in the main text.

Now that we have all the ingredients in place, we are ready to prove the regret bound.

Theorem C.7. *Under Assumption 2.1 and Assumption 2.2, we have that Algorithm 1 incurs regret at most* $\tilde{O}\left(\frac{d}{\sqrt{q}}\sqrt{T}\right)$.

Proof. We begin by noticing that there are at most $L \leq \log(T)$ phases in our algorithm. We prove now that if with probability at least $1 - \frac{1}{T^2}$ our confidence bounds hold, then with the probability at least $1 - \frac{1}{T}$ we can ensure that both the optimal item never gets eliminated and that the sub-optimality gap in round ℓ is at most $\varepsilon_\ell := (2S)2^{-\ell}$ after elimination, which means that in each round ℓ the suboptimality gap is at most $2\varepsilon_\ell$, which for $\ell = 1$ is trivial. We prove that if the upper confidence index is at most $\frac{1}{4}\varepsilon_\ell$, then any item with sub-optimality gap at least ε_ℓ will get eliminated.

First we prove that if all confidence intervals succeed, then the optimal item does not get eliminated. Indeed, consider the first round ℓ that the optimal item gets eliminated. Then, we have by concentration inequality the following guarantee

$$|\langle \theta^* - \theta_{\mathrm{rb}}, x - y\rangle| \leq \sqrt{\beta}\|x - y\|_{(F_\ell)^{-1}}.$$

By rearranging the terms, we see that for x^* and $y \in \mathcal{A}_\ell$ that

$$0 \leq \langle \theta^*, x^* - y\rangle \leq \langle \theta_{\mathrm{rb}}, x^* - y\rangle + \sqrt{\beta}\|x^* - y\|_{(F_\ell)^{-1}}.$$

Thus, the optimal item will not get eliminated.

We can now prove that items with too large a sub-optimality gap will be eliminated. By again the concentration inequality, we obtain

$$\langle \theta_{rb}, x - x^*\rangle + \sqrt{\beta}\|x - x^*\|_{F_\ell^{-1}} \leq \langle \theta^*, x - x^*\rangle + 2\sqrt{\beta}\|x - x^*\|_{F_\ell^{-1}} \leq -\varepsilon_\ell + \frac{2}{4}\varepsilon_\ell \leq -\frac{\varepsilon_\ell}{2} < 0.$$

Therefore, we have that at the beginning of round ℓ the sub-optimality gap is at most $2\varepsilon_\ell$.

Each of the confidence intervals hold with probability at least $1 - \frac{1}{T^2}$ which follows by Theorem 3.1 by noticing that within round ℓ the sequencing of the actions is deterministic. Thus, we find that all confidence intervals holds with probability at least $1 - \frac{L}{T^2} \geq 1 - \frac{1}{T}$. We can now bound the cumulative regret. We note that as the sub-optimality gaps are bounded by $2\varepsilon_\ell$ at every round, then the instantaneous regret at a time t in round ℓ are also bounded by $2\varepsilon_\ell$, as the average of quantities uniformly bounded by ε_ℓ is again bounded by ε_ℓ. We shall call R_ℓ the set of time points in phase ℓ. It follows by previous considerations, the length of the exploration and push down phase in round ℓ, and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{q_t} \sum_{x \in A_t} \langle \theta^*, x^* - x\rangle &\leq \sqrt{T} 2 \sqrt{\sum_{\ell=1}^{L} \sum_{t \in R_\ell} \varepsilon_\ell^2 |R_\ell|} \\
&\leq C\sqrt{T}\sqrt{\sum_{\ell=1}^{L} \varepsilon_\ell^2 \frac{1}{\dot{\mu}_{\min}^2}(d\log(dT) + dS^2)\left(\frac{d\log(1+T)}{q\varepsilon_\ell^2} + \frac{q}{\varepsilon_\ell^2}\right)} \\
&\leq \frac{C}{\dot{\mu}_{\min}}\sqrt{T}\sqrt{\sum_{\ell=1}^{L} \varepsilon_\ell^2 (d\log(dT) + dS^2)\left(\frac{d\log(1+T)}{q\varepsilon_{\ell^2}} + \frac{d}{q\varepsilon_\ell^2}\right)} \\
&\leq \frac{C}{\dot{\mu}_{\min}}\sqrt{T}\sqrt{\frac{1}{q}\log(T)\,(d\log(dT) + dS^2)\,(d\log(1+T) + d)} \\
&= O\left(\frac{1}{\dot{\mu}_{\min}}\frac{d}{\sqrt{q}}\sqrt{T}\log^{3/2}(dT)\right).
\end{aligned}$$

In the first step we used the Cauchy-Schwarz inequality along with our bound on the suboptimality gaps on the event previously described. The second inequality uses the bound on the length of all phases. The third inequality used $q \leq \sqrt{d}$. The fourth inequality multiplies the ε_ℓ terms, and the final inequality collects the order of the inequalities proven. On the complement of the event under which the above estimation hold we trivially bound it, results in an $O(S)$ additive term. This completes the proof. □

Remark C.8. Using essentially the same confidence interval and elimination we can achieve the sub-optimality gap sampling complexity claimed in the main text. One should now set the confidence level to $1-\delta$ and apply the sampling scheme and pushing down phase, and then perform the same elimination step. Any of the remaining items has then the desired sub-optimality gap with probability $1-\delta$ under sampling complexity $N \geq C\frac{d^2\log(\frac{1}{\delta})}{q\varepsilon^2}\log N$ for absolute constant $C>0$. For example, the greedy maximiser with respect to our current estimate will not get eliminated and will thus be an item within this sub-optimality gap.

D Lower bound

In this section we prove the lower bound presented in the main text.

Theorem D.1. *Suppose that the item set is $\mathcal{A} := B_2(0;1)$. For T sufficiently large depending only on d, we have that an algorithm that at time t is allowed to play multi-sets of size q incurs regret at least $\Omega(\frac{1}{\sqrt{q}}d\sqrt{T})$. In the instance we have $\|\theta^*\|_2 \leq 1$.*

Proof. The assumption that T be sufficiently large is to ensure that the parameter used to construct the lower bound is contained in the Euclidean ball and also required for the lower bound of linear bandits, see Theorem 24.2 [Lattimore and Szepesvári, 2020] as the length of the parameter in the counter example is $\|\theta^*\|_2 := \frac{d^2}{48T}$, thus for $T \geq T(d)$ this parameter is contained in the unit ball. The proof is a straightforward generalisation of the lower bound for duelling bandits provided in [Saha, 2021]. We note that Theorem 24.2 in [Lattimore and Szepesvári, 2020] asserts that the linear bandit problem incurs regret at least $\Omega(d\sqrt{T})$ under Gaussian noise. One can solve a linear bandit with ranking feedback in the following way. By playing a *multi-set*, the learner effectively chooses at each time point that is a multiple of q an assortment of size q and observes the *stochastic values*, and can therefore also sample from the random utility model by not using the stochastic values but instead only the items in sorted order. Thus, by playing the items recommended by the ranking algorithm we can play the linear bandit. Thus, by the lower bound, the non-normalised version of our regret incurs regret at least $\Omega(d\sqrt{qT})$. Since we normalise by q, we obtain a lower bound for our regret minimisation problem of the order $\Omega(\frac{d}{\sqrt{q}}\sqrt{T})$. □

E A note on computational efficiency

A potential issue of the proposed greedy algorithm is its scalability, as the number of possible q subsets is of order $O(N^q)$, with N being the total number of items. An idea to improve the computational efficiency is to consider performing maximum likelihood estimation for just a single independence chain, that is by only extracting $q/2$ valuable pairwise comparisons which do not overlap.

Consider the design matrix obtained by incorporating independent pairwise comparisons. That is, we consider the matrix

$$D_t := \lambda I + \sum_{s=1}^{t}(x_{s,1}-x_{s,2})(x_{s,1}-x_{s,2})^T.$$

Then, choosing the next most informative pairs (x_{t+1}, y_{t+1}) can be done by solving

$$(x_{t+1}, y_{t+1}) = \arg\max_{x,y\in\mathcal{A}_{\ell-1}} \|x-y\|_{V_t^{-1}},$$

which is *non-adaptive*. Thus, one can now compute ahead the $q/2$ most informative pairs in time $O(qN^2)$ per time step, ignoring matrix multiplication costs. One can then perform maximum likelihood estimation along the single independence chain or perform full rank breaking MLE on this set. A small technical caveat is that one might obtain the same point twice during this process, which is why we present it as heuristic rather than an sampling strategy with formal guarantees. We do not expect it to happen in practice that often, as the matrix also pushes down along the directions of previously chosen points.

Finally, another line of research focusses on optimal designs for ranking feedback [Mukherjee et al., 2024, Thekumparampil et al., 2025]. Their optimisation problem is not affected by the sub-optimal factors in q obtained in their analysis, so that they could potentially be incorporated into our algorithm.

The main benefit is that the computational costs due to the large action space are incurred once at the solving of the optimisation problem and then once more for a final pushing down phase. We leave such refinements for future work.

F Gumbel softmax trick

We provide a computation of the Gumbel top-k trick, see also [Kool et al., 2019] for the computation of the trick. This proof is based on the proof in the blog post [Bach, 2019] and the proof of [Kool et al., 2019]. We will show that the max and argmax are independent and we will derive their distribution. Finally, we will prove the Gumbel top-k trick for rankings. We start by recalling that the probability density function of a standard Gumbel random variable at x_i is $\exp(-(x-x_i) - \exp(-(x-x_i))$ and the cumulative distribution function is given by $\exp(-\exp(-(x-x_i)))$. We formally prove the following Lemma.

Lemma F.1. *Let $x_i+\zeta_i$ be standard Gumbel centered at x_i. Then $\arg\max_{i\in I} x_i+\zeta_i$ and $\max_{i\in I} x_i+\zeta_i$ are independent. Moreover, $\mathbb{P}(j = \arg\max_{i\in I} x_i + \zeta_i) = \frac{\exp(x_j)}{\sum_{i\in I}\exp(x_i)}$ and $\max_{i\in I} x_i + \zeta_i$ is standard Gumbel centered at $\log(\sum_{i\in I}\exp(x_i))$. Finally, we have that*

$$\mathbb{P}(x_1+\zeta_1 > x_2+\zeta_2 > \cdots > x_{|I|}+\zeta_i) = \Pi_{i=1}^{|I|}\frac{\exp(x_i)}{\sum_{j=i}^{|I|}\exp(x_i)}.$$

Proof. We start by the following elementary analysis. We write Z_i to be $x_i+\zeta_i$ that is the standard Gumbel centered at x_i and f_i to be its PDF.

$$\begin{aligned}
&\mathbb{P}\left(j = \arg\max_{i\in I} x_i+\zeta_i, \max_{i\in I} x_i+\zeta_i \le u\right) = \mathbb{P}\left(x_j+\zeta_j \le u, \forall i \neq j\ x_j+\zeta_j > x_i+\zeta_i\right)\\
&= \int_{-\infty}^{u} \Pi_{i\neq j}\mathbb{P}(Z_i < x) f_j(x)dx\\
&= \int_{-\infty}^{u} \Pi_{i\neq j}\exp(-\exp(-(x-x_i)))f_j(x)dx\\
&= \int_{-\infty}^{u} \Pi_{i\neq j}\exp(-\exp(-(x-x_i)))(\exp(-(x-x_j)) - \exp(-(x-x_j)))dx\\
&= \int_{-\infty}^{u} \Pi_{i\in I}\exp(-\exp(-(x-x_i)))\exp(-(x-x_j))dx\\
&= \int_{-\infty}^{u} \Pi_{i\in I}\exp(-\exp(x_i-x))\exp(-(x-x_j))dx\\
&= \int_{-\infty}^{u} \exp(-\exp(-x)\sum_{i\in I}\exp(x_i))\exp(-(x-x_j))dx
\end{aligned}$$

From here, we write $\sum_{i\in I}\exp(x_i) = \exp(\log(\sum_{i\in I}\exp(x_i))$. This allows us to rewrite the last integral as follows:

$$\begin{aligned}
&\int_{-\infty}^{u} \exp\left(-\exp\left(-\left(x-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\right)\right)\exp(-(x-x_j))\\
&= \int_{-\infty}^{u} \exp\left(-\exp\left(-\left(x-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\right)\right)\exp\left(-\left(x-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\right)dx\\
&\cdot \exp\left(x_j-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\\
&= \int_{-\infty}^{u} \exp\left(-\exp\left(-\left(x-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\right)\right)\exp\left(-\left(x-\log\left(\sum_{i\in I}\exp(x_i)\right)\right)\right)dx
\end{aligned}$$

$$\cdot \frac{\exp(x_j)}{\sum_{i\in I}\exp(x_i)}$$

where in the second line we added and subtracted the quantity $\log\left(\sum_{i\in I}\exp(x_i)\right)$ in the exponential. By letting $u \to \infty$ we get that the probability is the maximum is given by $\frac{\exp(x_j)}{\sum_{i\in I}\exp(x_i)}$. By integrating out the argmax, we obtain that the maximum is distributed according to a standard Gumbel centered at $\log(\sum_{i\in I}\exp(x_i))$. The above formula also shows independence.

Finally, the Gumbel top-k trick in case of our ranking model is then given by the following string of algebra.

$$\begin{aligned}
\mathbb{P}\left(Z_1 > Z_2 \cdots > Z_q\right) &= \mathbb{P}\left(Z_1 > \cdots > Z_{q-2} > \max_{i\in\{q-1,q\}} Z_i \wedge q-1 = \arg\max_{i\in\{q-1,q\}} Z_i\right) \\
&= \mathbb{P}\left(Z_1 > Z_2 > Z_3 > \cdots > Z_{q-2} > \max_{i\in\{q-1,q\}} Z_i\right) \\
&\cdot \mathbb{P}\left(q-1 = \arg\max_{i\in\{q-1,q\}} Z_i\right) \\
&= \mathbb{P}\left(Z_1 > Z_2 > Z_3 > \cdots > Z_{q-2} > \max_{i\in\{q-1,q\}} Z_i\right) \\
&\cdot \frac{\exp(x_{q-1})}{\exp(x_{q-1})+\exp(x_q)}
\end{aligned}$$

The first equality stems from independence and the second from the Gumbel softmax trick. The trick is now to keep popping off the final one, absorbing it into the max, and then use the Gumbel trick recursively. Thus, the Gumbel softmax trick as promised. □

G Numerical experiments

In this section we present numerical results for different values of q. The code used to run the experiments is available at [1]. The experiments were computed on a laptop with an Intel(R) Core(TM) Ultra 7 155U (1.70 GHz) processor with 16 GB of RAM. The experiments were run on CPU resources. To produce the plots we use the scientific packages Numpy, Scipy, and Matplotlib to generate the plots [Harris et al., 2020, Virtanen et al., 2020, Hunter, 2007]. A rough estimation of the compute required for the project is estimated to be of the order of a few days.

We now describe the set up for Figure 1 and Figure 2. The run time typically takes between 20 to 40 minutes per plot for the plots in Figure 1 and Figure 2. In the experiments for Figure 1 and Figure 2 we compute the projected estimator. In our code there is also an option to compute the first order solution only, which is typically faster as that corresponds to a root finding problem. We will use root finding for the plots Figure 3 and Figure 4. In the experiments we use standard Gumbel noise. We defined the exploration length to be the ceiling of $\frac{d^2}{q2^{-2\ell}}\log(T)$ in round ℓ and our push down length to be the ceiling of $\frac{dq}{2^{-2\ell}}\log(T)$. We use $\lambda = 1$. In our experiments, we use $\beta = d$ where d is the dimension. The setup is as follows: we draw a random θ^* uniformly from the sphere. Then, either we use an arm set where each arms is drawn uniformly from the sphere of size 16 and dimension 20 or use the standard basis in dimension 16. In case of uniform arms case, the horizon is 10000 and in case of the standard basis the horizon is 12000. We then run 10 trials for each plot using the fixed drawn θ^* and arm set so as to average over only the randomness in the Plackett-Luce rankings. We do this for each plot, so that different plots show different behaviour on θ^* and possibly different arm sets. We compute the standard deviation at time t as $\sqrt{\frac{1}{s-1}\sum_{i=1}^{s}(R_t[i] - \frac{1}{s}\sum_{j=1}^{s}R_t[j])^2}$ and plot 1 standard deviation bars, where $R_t[i]$ is the regret at time t for trials i, where we have s trials.

The results for the uniform arms case are found in Figure 1. While the average of the cumulative regret does seem to be somewhat lower for larger q, we note that the shaded areas overlap meaning we cannot draw definitive conclusions from these experiments in terms of performance in q. The results for the standard orthonormal basis, shown in Figure 2, are comparable. While on the whole

[1] `https://github.com/NigelTUD/SupplementaryMaterialRankingIndependence`

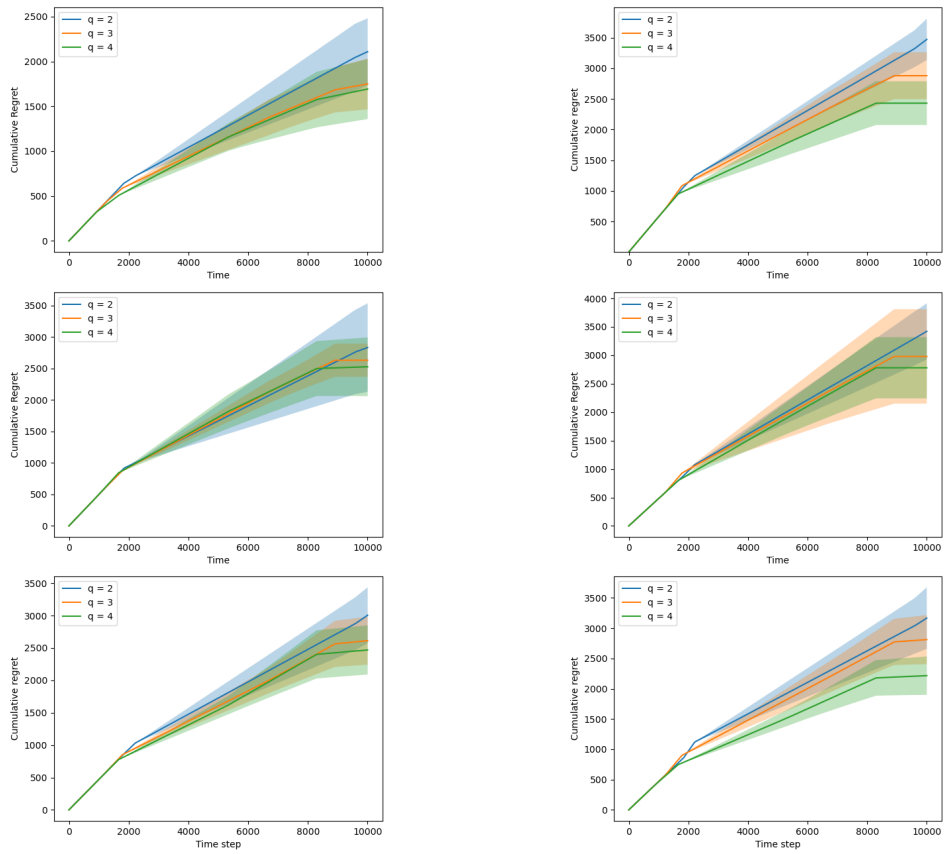


Figure 1: Cumulative regret plots with different values of q for the uniform arms case of size 16 in dimension 20.

larger values of q seems have smaller average cumulative regret, that is not always the case. This is also indicated by the shaded areas. Thus, for the orthonormal arms case the results are somewhat weaker than for the uniform arms case.

We see that the variability in Figures 1 and 2 is quite high. Therefore, we study what happens if we take a less aggressive algorithm by making the exploration phase longer and the confidence parameter wider. We now describe the set up for Figure 3 and Figure 4. The noise is again standard Gumbel. We choose now $\sqrt{\beta} = \sqrt{d \log T + \sqrt{d} \log T}$ and the exploration length is set to $\frac{d^2}{q 2^{-2\ell}} \log^2(T)$ and the push down phase to the length of $\frac{dq}{2^{-2\ell}} \log T$. Here, the regularizer λ is the square root of the dimension. From here on, we do not use the projection step and instead use root finding to get an estimate of θ^*. As before, we draw θ^* uniformly from the sphere once, as well as the arm set in case of uniform arms set. We do this procedure for every plot. We average solely over the rankings obtained. The horizon is 30000 here. Each plot with 10 trials can be computed in approximately 40 minutes. The experiments with the uniform arms case tend to take the longest.

In Figures 3 and 4 we see that in general the larger assortments do tend to perform better than their pairwise counterparts. The standard deviation is also slightly lower. In these two figures we look at two cases for arm set, namely uniform sphere (Figure 3) and standard orthonormal basis (Figure 4). The standard deviation in Figure 4a is larger, but the different values of q do find the optimal action in the time horizon. In Figure 3 we see that the larger assortment typically tend to perform slightly better. However, this is not enough to make a definitive statement regarding the performance of different values of q.

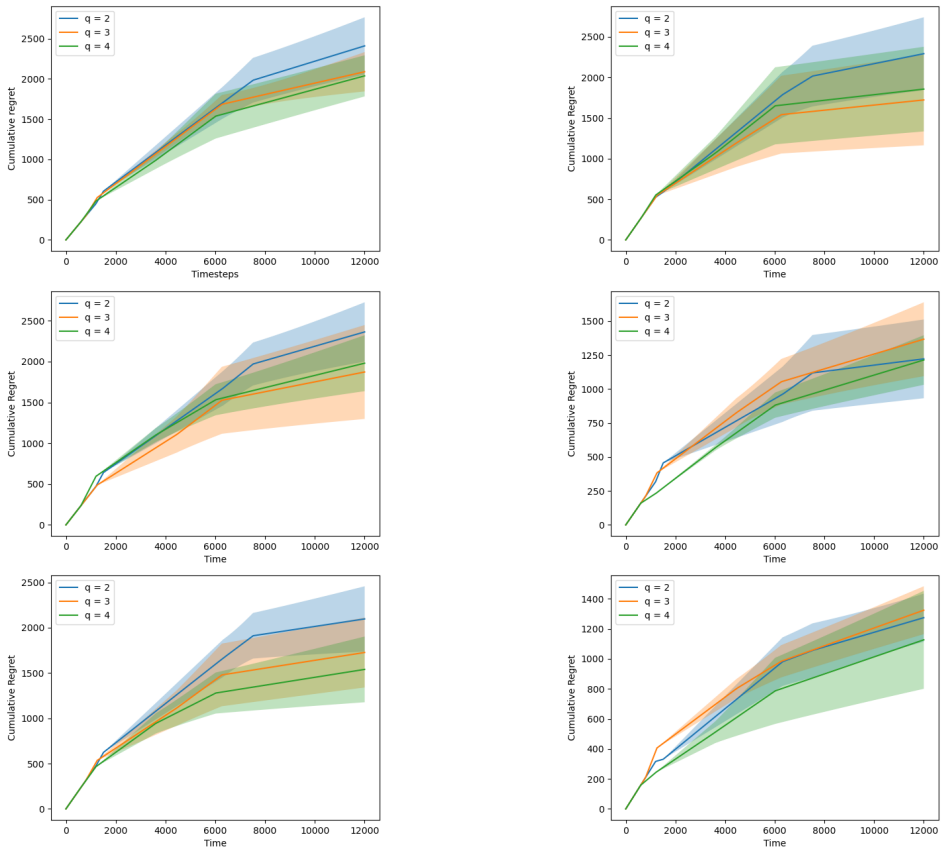


Figure 2: Regret plots for different values of q and the standard orthonormal basis in dimension $d = 16$.

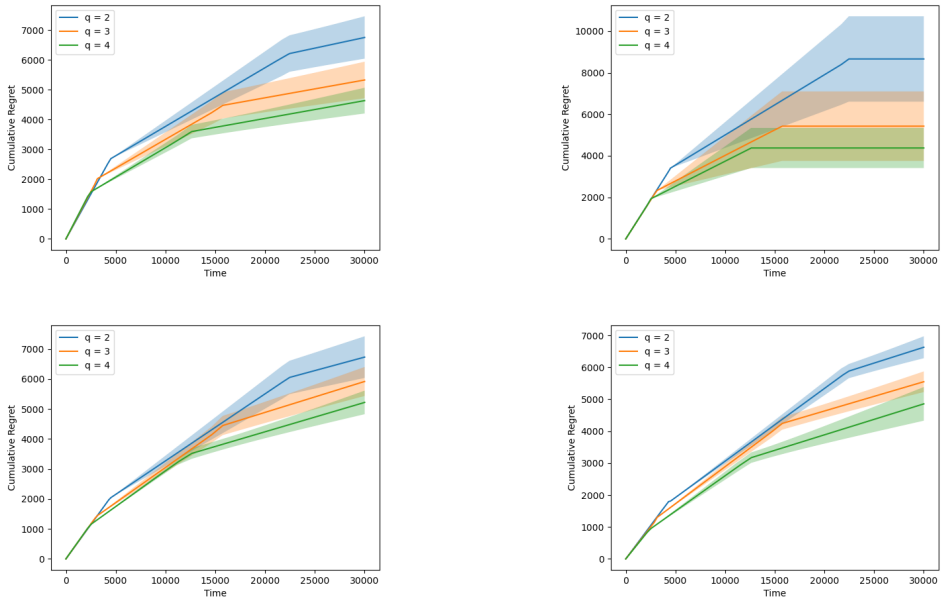
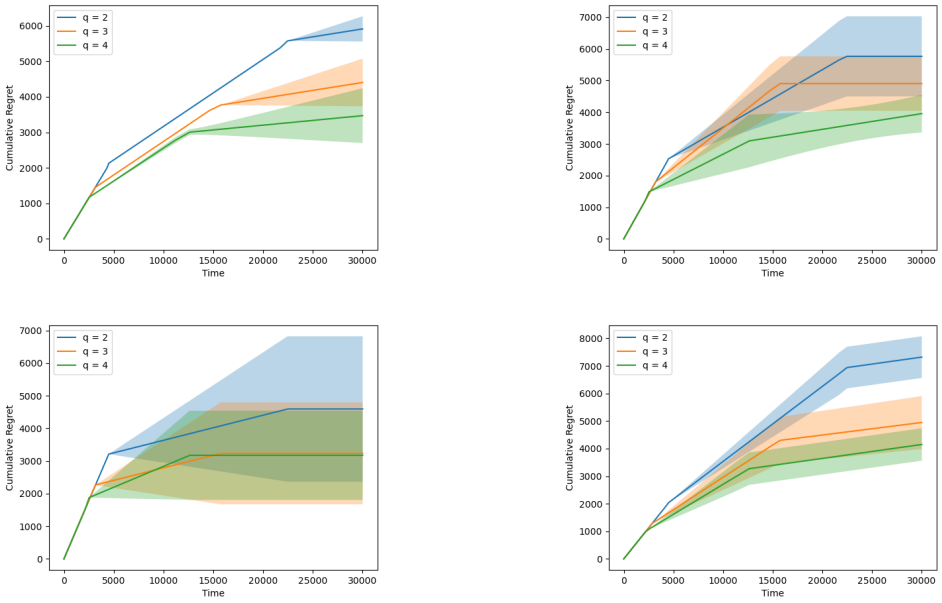


Figure 3: Regret plots for different values of q. We take the uniform sphere case in this figure with 16 items. The dimension is $d = 9$.



(a) Higher variability plot

Figure 4: Regret plots for different values of q. We take standard orthonormal basis in dimension $d = 9$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly state the main contributions that the paper makes and provide further details in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide a clear explanation of the assumptions along with a discussion of how they might be limiting and could potentially be improved. We point to the possibility to relax assumptions, such as the $q \leq \sqrt{d}$ assumption made in our paper needed for the full regret bound claim. We also discuss when assumptions are not needed, such as confidence interval. We also point out a limitation in computational efficiency, which we discuss in more depth in the Appendix due to space constraints.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Full proofs are given in the Appendix due to space constraints. The main text points towards the specific section in the Appendix where the proof may be found. The assumptions are all clearly stated in the main text, so that no additional assumption is presented in the Appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We give a description of the process by which we generate the experiments.

   Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: We include the code used to generate the plots as a GitHub link in the Appendix.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We list all the required details to be able to reproduce the experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We compute standard deviation in regret plots and show the average with an error strip which shows one standard deviation error in both directions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We state our resources on which we ran the experiments and the approximate time to run the code with our specified hardware in the Appendix. We also mention CPU resources.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our paper does not involve any of the risks mentioned in the Code of Ethics, and does not violate the Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: The goal of this paper is to show the information theoretic rate for regret minimization with ranking feedback. Due to its primarily theoretical nature, we do not think our work has immediate societal impact. We include a statement regarding this statement at the end of the paper.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: We do not use any data or models that have high risk of misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We cite the papers that the scientific packages (Numpy, Matplotlib, Scipy) that we use to generate the plots in the Appendix want to be cited, thereby we properly credit these packages. We include the license statements of these packages in the code distribution.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We have provided a readme file containing a reference of what should be installed and how to run the code.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our paper does not deal with human subjects.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: Our paper does not deal with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were not used any stage to produce this paper.

Guidelines:

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