Optimal, Efficient and Practical Algorithms for Assortment Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the problem of active online assortment optimization problem with preference feedback, which is a framework for modeling user choices and subsetwise utility maximization. The framework is useful in various real-world applications including ad placement, online retail, recommender systems, and fine-tuning language models, amongst many others. The problem, although has been studied in the past, lacks an intuitive and practical solution approach with simultaneously efficient algorithm and optimal regret guarantee. E.g., popularly used assortment selection algorithms often require the presence of a 'strong reference' which is always included in the choice sets, further they are also designed to offer the same assortments repeatedly until the reference item gets selected—all such requirements are quite unrealistic for practical applications. In this paper, we designed efficient algorithms for the problem of regret minimization in assortment selection with *Plackett Luce* (PL) based user choices. We designed a novel concentration guarantee for estimating the score parameters of the PL model using '*Pairwise Rank-Breaking*', which builds the foundation of our proposed algorithms. Moreover, our methods are practical, provably optimal, and devoid of the aforementioned limitations of the existing methods. Empirical evaluations corroborate our findings and outperform the existing baselines.

1 Introduction

Studies have shown that it is often easier, faster and less expensive to collect feedback on a relative scale rather than asking ratings on an absolute scale. E.g., to understand the liking for a given pair of items, say (A,B), it is easier for the users to answer preference-based queries like: "Do you prefer Item A over B?", rather than their absolute counterparts: "How much do you score items A and B in a scale of [0-10]?". Due to the widespread applicability and ease of data collection with relative feedback, learning from preferences has gained much popularity in the machine-learning community, especially the active learning literature which has applications in Medical surveys, AI tutoring systems, Multi-player sports/games, or any real-world systems that have ways to collect feedback in terms of preferences. The problem is famously studied as the *Dueling-Bandit* (DB) problem in the active learning community [41, 3, 45, 46, 44], which is an online learning framework for identifying a set of 'good' items from a fixed decision-space (set of items) by querying preference feedback of actively chosen item-pairs. Consequently, the generalization of Dueling-Bandits, with *subset-wise* preferences has also been developed into an active field of research. For instance, applications like Web search (e.g. Google, Bing, or even in some versions of ChatGPT), online shopping (Amazon, App stores, Google Flights), recommender systems (e.g. Youtube, Netflix, Google News/Maps, Spotify) typically involve users expressing preferences by choosing one result (or a handful of results) from a subset of offered items and often the objective of the system is to

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

identify the ‘most-profitable’ subset to offer to their users. The problem, popularly termed as ‘Assortment Optimization’ is studied in many interdisciplinary literature, e.g. Online learning and bandits [10], Operations research [40, 2], Game theory [15], RLHF [20, 30], to name a few.

Problem (Informal): Active Optimal Assortment (AOA) Active Assortment Optimization (a.k.a. Utility Maximization with Subset Choices) [13, 2, 23, 22] is an active learning framework for finding the ‘optimal’ profit-maximizing subset. Formally, assume we have a decision set of $[K] := \{1, 2, \dots K\}$ of K items, with each item being associated with the score (or utility) parameters $\boldsymbol{\theta} := (\theta_1, \theta_2, \dots, \theta_K)$ (without loss of generality assume $\theta_1 \geq \theta_2 \geq \dots \geq \theta_K \geq 0$). At each round $t = 1, 2, \dots$, the learner or the algorithm gets to query an assortment (typically subsets containing up to m-items) $S_t \subseteq [K]$, upon which it gets to see some (noisy) relative preferences across the items in S_t, typically generated according to an underlying Plackett-Luce (PL) choice model with parameters $\boldsymbol{\theta}$ (1). Further, to allow the event where no items are selected, we also model a No-Choice (NC) item, indexed by item-0, with PL parameter $\theta_0 \in \mathbb{R}_+$.

(Objective 1.) Top-m: identify the top-m item-set: $\{\theta_1, \dots, \theta_m\}$, for some $m \in [1, K]$.

(Objective 2.) Wtd-Top-m: A more general objective could also consider a weight (or price) $r_i \in \mathbb{R}_+$ associated with the item $i \in [K]$, and the goal could be to identify the assortment (subset) with maximum weighted utility [1], as detailed in Sec. 2.

Related Works and Limitations: As stated above, the problem of AOA is fundamental in many practical scenarios, and thus widely studied in multiple research areas, including Online ML/learning theory and operations research.

- In the Online ML literature, the problem is well-studied as *Multi-Dueling Bandits* [39, 14], or Battling Bandits [35, 34, 11], which is an extension of the famous *Dueling Bandit* problem [46, 45]. The main limitation of this line of work is the lack of practical objectives, which either aim to identify the ‘best-item’ $1 (= \arg\max_{i \in [K]} \theta_i)$ within a PAC (probably approximately correct) framework [36, 16, 17, 31] or quantifying regret against the best items [35, 12]. Note the latter actually leads to the optimal subset choice of repeatedly selecting the optimal item, $\arg\max_i \theta_i$, m times, i.e. $(1, 1, \dots 1)$, which is unrealistic from the viewpoint of real-world system design. Selecting an assortment of distinct top-m items (Top-m-AOA) or maximum expected utility (Wtd-Top-m-AOA) makes more sense.

- On the other hand, a similar line of the problem has been studied in operations research and dynamic assortment selection literature, where the goal is to offer a subset of items to the customers in order to maximize expected revenue. The problem has been studied under different user choice models, e.g. PL or Multinomial-Logit models [2], Mallows and mixture of Mallows [22], Markov chain-based choice models [23], single transition model [27] etc. While these works indeed consider a more practical objective of finding the best assortment (subset) with the highest expected utility for a regret minimization objective, (1) a major drawback in their approach lies in the algorithm design which *requires to keep on querying the same set multiple times*, e.g. [2, 29, 18, 1]. Such design techniques could be impractical to be deployed in real systems where users could easily get annoyed if the same items are shown again and again. For example, in ad-placement, music/movies/news/tweets/reels recommendations, offering the same assortment could increase user dissatisfaction and disengagement.

(2) The second major drawback of this line of work lies in the *structural assumption of their underlying choice models which requires the existence of a reference/default item, that needs to be part of every assortment S_t*. This leads to assuming a No-Choice item, typically denoted as item-0, which is a default choice of any assortment S_t. Further a stronger and more unrealistic assumption lies in the fact that they require to assume that the above pivot is stronger than the rest of the K items, i.e. $\theta_0 \geq \max_{i \in [K]} \theta_i$, i.e. the No-Choice (NC) action is the most likely outcome of any assortment S_t. This is often unrealistic, e.g., during user interactions with language models, or online shopping, or Route recommendation in GPS navigation, a NC action is highly improbable. Consequently, such assumption limits the use in real-systems. In the existing literature [2, 28, 1, 24], such assumptions are primarily

[1] This is equivalent to finding the set with maximum expected revenue when r_is represents the price of item i [2]

adapted solely for theoretical needs, precisely for maintaining concentration bounds of the PL parameters $\boldsymbol{\theta}$, and hence not well justified from a practical viewpoint. Some recent developments also generalized the AOA problem to linear MNL scores to incorporate large actions embedded in d-dimension [43, 42, 28], however, their approaches are either limited to the above restrictions or suffer sub-optimal regret guarantees without those assumptions (e.g. the regret bound of [28] is $O(d^{3/2}\sqrt{T})$ which is suboptimal by a d-factor). Considering the above limitations of the AOA literature, we set to answer two questions:

(1) Can we consider a general AOA model where the default item, like the NC item defined above, is not necessarily the strongest one, i.e. $\theta_0 \geq \max_{i\in[K]} \theta_i$?

(2) Can we design a practical and regret optimal algorithm for the AOA framework, without needing to play the same repetitive actions and yet converge to the optimal assortment?

Contributions We answer these questions in the affirmative and present best of all scenarios. We design practical algorithms on practical AOA framework with practical objectives–Unlike the existing approaches of the AOA, literature [2, 18], we do not have to keep playing the same assortment multiple times, neither require a strongest default item (like NC satisfying $\theta_0 \geq \max_{i\in[K]} \theta_i$). Moreover, our objectives do not require us to converge to a multiset of replicated arms like $(1, 1, \ldots 1)$, but converge to the utility-maximizing set of distinct items. We list our contributions below:

1. **A General AOA Setup:** We work with a general problem of AOA for PL model, which requires no additional structural assumption of the $\boldsymbol{\theta}$ parameters such as $\theta_0 \geq \max_i \theta_i$, unlike the existing works. We designed algorithms for two separate objectives Top-m and Wtd-Top-m as discussed above (Sec. 2).

2. **Practical, Efficient and Optimal Algorithm:** In Sec. 3, we give a practical, efficient and optimal algorithm for MNL Assortment (up to log factors and the magnitude of $\theta_{\max}$). The regret bound of our algorithm AOA-RB$_{\text{PL}}$ (Alg. 1) yields $\tilde{O}(\sqrt{KT})$ regret for both Top-m and Wtd-Top-m objective. Our algorithms use a novel parameter estimation technique for discrete choice models based on the concept of *Rank-Breaking* (RB) which is one of our key contributions towards designing the efficient and optimal algorithm. This enables our algorithm to perform optimally without requiring the No-Choice item to be the strongest. Appendix A details the key concept of our parameter estimation technique exploiting the concept of RB. Our resulting algorithm plays optimistically based on the UCB estimates of PL parameters and does not require repeating the same subset multiple times, justifying our title.

3. **Improvement with Adaptive Pivots:** In Sec. 4, we refine the performance of our algorithm by employing the novel idea of 'adaptive pivots' (a reference item) and proposed AOA-RB$_{\text{PL}}$-Adaptive. Performance-wise this removes the asymptotic dependence on $\theta_{\max} = \max_i \theta_i/\theta_0$ in the regret analysis. This enables the algorithm to work effectively in scenarios where the No-Choice item is less likely to be selected, i.e., $\theta_{\max} \gg 1$. This leads to a huge improvement in our experiments, especially in the range of low θ_0, where AOA-RB$_{\text{PL}}$-Adaptive drastically outperforms over the existing baseline. Comparison of our regret bound with existing work is detailed in Table 1.

4. **Emperical Analysis.** Finally, we corroborate our theoretical results with empirical evaluations (Sec. 5), which certify our superior performance in the general AOA setups.

Work	Framework	Assume $\theta_0 = \theta_{\max} = 1$	Regret
Our (Alg. 1)	MNL model (Obj. 2)	No	$\sqrt{\min\{\theta_{\max}, K\}KT}\log T$
[2] (Thm 1)	MNL model (Obj. 2)	Yes	$\sqrt{KT}\log T$
[2] (Thm 4)	MNL model (Obj. 2)	No	$\sqrt{\theta_{\max}KT}\log T$
[1]	MNL model (Obj. 2)	Yes	$\sqrt{KT\log(mT)} + K\log^2(mT)$
[24]	MNL model with constraints (Obj. 2)	No	$\sqrt{\frac{KT}{\min_i r_i}}\log T$

Table 1: Our Contribution vs the Existing Results in the K-armed MNL-Assortment literature

It is also worth mentioning that our proposed algorithm and their respective regret analysis could be extended to any general random utility (RUM) based preference models [38, 37], as explained in Rem. 1. However, to keep the focus on the AOA problem and ease the presentation, we stick to the special case of MNL choice model based preferences.

2 Problem Setup

We write $[n] = \{1, 2, ..., n\}$ and $\mathbb{1}\{\cdot\}$ denotes the indicator function. The symbol $\lesssim$, employed in the proof sketches, represents a coarse inequality.

We consider the sequential decision-making problem of Active Optimal Assortment (AOA), with preference/choice feedback. Formally, the learner is given $[K]$, a finite set of K items ($K > 2$). At each decision round $t = 1, 2, \ldots$, the learner selects a subset $S_t \subseteq [K]$ of up to m items, and receives some (stochastic) feedback about the item preferences of S_t, drawn according to some unknown underlying Plackett-Luce (PL) choice model (1) with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K) \in \mathbb{R}_+^K$. We assume $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_K$ without loss of generality. An interested reader may check App. A.1 for a detailed discussion on PL models. Given any assortment S_t we also consider the possibility of 'no-selection' of any items given an S_t. Following the literature of [2], we model this mathematically as a No-Choice (NC) item, indexed by item-0, and its corresponding PL utility parameter θ_0. Unlike most existing literature on assortment selection, we are not assuming $\theta_0 \not\geq \max_{i\in[K]} \theta_i$. Further, since the PL model is scale independent, we set $\theta_0 = 1$ and scale the rest of the PL parameters.

Feedback model The feedback model formulates the information received (from the 'environment') once the learner plays a subset $S_t \subseteq [K]$ of at most m items. Given S_t we consider the algorithm receives a winner feedback (or index of an item) $i_t \in S_t \cup \{0\}$, drawn according to the underlying PL choice model as:

$$\mathbb{P}(i_t = i|S_t) = \theta_i / \big(\theta_0 + \textstyle\sum_{j\in S_t} \theta_j\big), \quad \forall i \in S_t. \tag{1}$$

We consider the following two objectives for the learner:

1. Top-m-Ojective. One simple objective could be to identify the top-m item-set: $\{\theta_1, \ldots, \theta_m\}$, for some $m \in [1, K]$. The performance of the learner can be captured by minimizing the following regret:

$$Reg_T^{\texttt{top}} := \sum_{t=1}^{T} \frac{\Theta_{S^*} - \Theta_{S_t}}{m}, \quad \text{where} \quad S^* := \underset{S\subseteq[K]:|S|=m}{\text{argmax}} \left\{\Theta_S := \sum_{i\in S} \theta_i\right\}.$$

2. Wtd-Top-m-Objective. Here, each item-i is associated with a weight (for example price) $r_i \in \mathbb{R}_+$, and the goal is to identify the set of size at most m with maximum weighted utility. One could measure the regret of the learner as:

$$Reg_T^{\texttt{wtd}} := \sum_{t=1}^{T} (\mathcal{R}(S^*, \boldsymbol{\theta}) - \mathcal{R}(S_t, \boldsymbol{\theta})), \text{ where } \mathcal{R}(S, \boldsymbol{\theta}) := \sum_{i\in S} \frac{r_i \theta_i}{\theta_0 + \sum_{j\in S} \theta_j}, \ \forall S \subseteq [K], \tag{2}$$

denotes $S^* := \text{argmax}_{S\subseteq[K]||S|\leq m} \mathcal{R}(S, \boldsymbol{\theta})$ is the optimal utility-maximizing subset. This objective corresponds to the standard objective in the MNL litterature [2].

3 A Practical and Efficient Algorithm for AOA with PL

In this section, we introduce our first algorithm, which works for both objectives.

3.1 Algorithm Design

At each time t, our algorithm (Alg. 1) maintains a pairwise preference matrix $\widehat{\mathbf{P}}_t \in [0, 1]^{n\times n}$, whose (i, j)-th entry $\widehat{p}_{ij,t}$ records the empirical probability of i having beaten j in a pairwise duel, and a corresponding upper confidence bound $p^{\text{ucb}}_{ij,t}$. Let $[\tilde{K}] := [K] \cup \{0\}$. We define for each pair $(i,j) \in [\tilde{K}] \times [\tilde{K}]$,

$$p^{\text{ucb}}_{ij,t} := \widehat{p}_{ij,t} + \sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{3x}{n_{ij,t}}, \qquad \text{where} \quad \widehat{p}_{ij,t} := \frac{w_{ij,t}}{n_{ij,t}}\,, \tag{3}$$

where $w_{ij,t} = \sum_{s=1}^{t-1} \mathbb{1}\{i_s = i, j \in S_s\}$ denotes the number of pairwise wins of item-i over j and $n_{ij,t} = w_{ij,t} + w_{ji,t}$ being the number of times (i,j) has been compared. The above UCB estimates $p^{\text{ucb}}_{ij,t}$ are further used to design UCB estimates of the PL parameters θ_i as follows

$$\theta^{\text{ucb}}_{i,t} = p^{\text{ucb}}_{i0,t}/(1-p^{\text{ucb}}_{i0,t})_+.$$

The estimates $\theta^{\text{ucb}}_{i,t}$s are then used to select the set S_t, that maximizes the underlying objective. This optimization problem transforms into a static assortment optimization problem with upper confidence bounds $\theta^{\texttt{ucb}}_{i,t}$ as the parameters, and efficient solution methods for this case are available (see e.g., [7, 21, 32]).

Algorithm 1 AOA for PL model with RB (AOA-RB$_{\mathbf{PL}}$)

```
input: x > 0
init: K̃ ← K + 1, [K̃] = [K] ∪ {0}, W_1 ← [0]_{K̃×K̃}
for t = 1, 2, 3, ..., T do
    Set N_t = W_t + W_t^⊤, and P̂_t = W_t / N_t. Denote N_t = [n_{ij,t}]_{K̃×K̃} and P̂_t = [p̂_{ij,t}]_{K̃×K̃}.
    Define for all i, p^ucb_{ii,t} = 1/2 and for all i, j ∈ [K̃], i ≠ j
        p^ucb_{ij,t} = p̂_{ij,t} + ( 2p̂_{ij,t}(1 − p̂_{ij,t})x / n_{ij,t} )^{1/2} + 3x / n_{ij,t}
    θ^ucb_{i,t} := p^ucb_{i0,t} / (1 − p^ucb_{i0,t})_+
    S_t ← { Top-m items from argsort({θ^ucb_{1,t}, ..., θ^ucb_{K,t}}),
                for Top-m objective
            argmax_{S⊆[K]||S|≤m} R(S, θ^ucb_t),
                for Wtd-Top-m objective
    Play S_t
    Receive the winner i_t ∈ [K̃] (drawn as per (1))
    Update: W_{t+1} = [w_{ij,t+1}]_{K̃×K̃} s.t. w_{i_t j,t+1} ← w_{i_t j,t} + 1  ∀j ∈ S_t ∪ {0}
end for
```

3.2 Analysis: Concentration Lemmas

We start the analysis by providing two technical lemmas, whose proofs are deferred to the appendix and that provide confidence bounds for the θ_i.

Lemma 1. *Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - 3KTe^{-x}$, for all $t \in [T]$ and $i \in [K]$: $\theta_i \leq \theta^{ucb}_{i,t}$ atleast one of the following two inequalities is satisfied*

$$n_{i0,t} < 69x(\theta_0 + \theta_i) \quad \text{ or } \quad \theta^{ucb}_{i,t} \leq \theta_i + 4(\theta_0+\theta_i)\sqrt{\frac{2\theta_0\theta_i x}{n_{i0,t}}} + \frac{22x(\theta_0+\theta_i)^2}{n_{i0,t}}\,.$$

The above lemma depends on $n_{i0,t}$ the number of times items i have been compared with item 0 up to round t. The latter is controlled using the following lemma:

Lemma 2. *Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - KTe^{-x}$: simultaneously for all $t \in [T]$ and $i \in [K]$*

$$\tau_{i,t} < 2x(\theta_0 + \Theta_{S^*})^2 \quad \text{or} \quad n_{i0,t} \geq \frac{(\theta_0+\theta_i)\tau_{i,t}}{2(\theta_0+\Theta_{S^*})}\,, \tag{4}$$

where $\tau_{i,t} = \sum_{s=1}^{t-1} \mathbb{1}\{i \in S_s\}$ denotes the number of rounds item i got selected before round t.

3.3 Analysis: Top-m Objective:

We are now ready to provide the regret upper bound for Algorithm 1 with Top-m objective.

Theorem 3 (Top-m Objective)**.** *Let $\theta_{\max} \geq 1$. Consider any instance of PL model on K items with parameters $\theta \in [0, \theta_{\max}]^K$, $\theta_0 = 1$. The regret of Alg. 1 with parameter $x = 2\log T$ is bounded as*

$$Reg_T^{top} = O\big(\theta_{\max}^{3/2}\sqrt{KT\log T}\big) \quad \text{when } T \to \infty\,.$$

The above rate of $\tilde{O}(KT)$ is optimal (up to log-factors), as a lower bound can be derived from standard multi-armed bandits [5, 6]. We only state here a sketch of the proof of Theorem 3. The detailed proof is deferred to the App. B.

Proof Sketch of Theorem 3. Let us define for any $S \subseteq [K]$,

$$\Theta_S = \sum_{i\in S}\theta_i, \quad \text{and} \quad \Theta_S^{\mathtt{ucb}} := \sum_{i\in S}\theta_i^{\mathtt{ucb}}.$$

Let $\mathcal{E}$ be the high-probability event such that both Lemma 1 and 2 holds true. Then, $\mathbb{P}(\mathcal{E}) \geq 1 - 4TKe^{-x}$. Let us first assume that $\mathcal{E}$ holds true. Then, by Lemma 1, $\Theta_{S^*} \leq \Theta_{S^*}^{\mathtt{ucb}} \leq \Theta_{S_t}^{\mathtt{ucb}}$, which yields

$$Reg_T^{\mathtt{top}} = \frac{1}{m}\sum_{t=1}^T \Theta_{S^*} - \Theta_{S_t} \leq \frac{1}{m}\sum_{t=1}^T \Theta_{S_t}^{\mathtt{ucb}} - \Theta_{S_t} \lesssim \tau_0 + \frac{1}{m}\sum_{t=1}^T\sum_{i\in S_t}(\theta_{i,t}^{\mathtt{ucb}} - \theta_i)\mathbb{1}\big\{\tau_{i,t} \geq \tau_0\big\}\,,$$

where $\tau_0 = 138x(m+1)^2\theta_{\max}^2$ corresponds to an exploration phase needed for the confidence upper bounds of Lem 1 and 2 to be satisfied. Then, noting that if $\mathcal{E}$ holds true, we can show by Lemma 2, that $\mathbb{1}\{\tau_{i,t} \geq \tau_0\} \leq \mathbb{1}\{n_{i0,t} \geq 69x(\theta_0 + \theta_i)\}$. Therefore, we can apply Lemma 1 that entails,

$$\begin{aligned}
\frac{1}{m}\sum_{t=1}^T\sum_{i\in S_t}(\theta_{i,t}^{\mathtt{ucb}} - \theta_i)\mathbb{1}\big\{\tau_{i,t} \geq \bar{n}_{i0}\big\} &\lesssim \frac{1}{m}\sum_{t=1}^T\sum_{i\in S_t}\left((\theta_0+\theta_i)\sqrt{\frac{\theta_0\theta_i x}{n_{i0,t}}}\mathbb{1}\big\{\tau_{i,t} \geq \tau_0\big\}\right) \\
&\overset{\text{Lem. 2}}{\lesssim} \frac{1}{m}\sum_{t=1}^T\sum_{i\in S_t}\theta_{\max}^{3/2}\sqrt{\frac{mx}{\tau_{i,t}}} \lesssim \frac{1}{m}\sum_{i=1}^K\theta_{\max}^{3/2}\sqrt{mx\tau_{i,t}} \lesssim \theta_{\max}^{3/2}\sqrt{xKT}\,.
\end{aligned}$$

where we used $\sum_{i=1}^n 1/\sqrt{i} \leq 2\sqrt{n}$ and $\sum_i \tau_{i,t} = mT$ together with Jensen's inequality in the last inequality. We thus have under the event $\mathcal{E}$ that $Reg_T^{\mathtt{top}} \leq O(\theta_{\max}^{3/2}\sqrt{xKT})$ and the proof is concluded by taking the expectation with $x = 2\log T$ to control $\mathbb{P}(\mathcal{E}^c)$. □

3.4 Analysis: Wtd-Top-m Objective

We turn now to the analysis of the Wtd-Top-m objective (2). We start by stating a lemma from [2] that shows that the expected utility $\mathcal{R}(S^*, \theta)$ that corresponds to the optimal assortment $S^* = \text{argmax}_{S\subset[K],|S|\leq m}\mathcal{R}(S,\theta)$ is non-decreasing in the parameters θ.

Lemma 4 (Lemma A.3 of [2])**.** *Assume $\theta_i^{ucb} \geq \theta_i$ for all $i \in [K]$, then $\mathcal{R}(S^*,\theta) \leq \mathcal{R}(S^*,\theta^{ucb})$.*

Theorem 5 (Wtd-Top-m Objective)**.** *Let $\theta_{\max} \geq 1$. Then, for any $\theta \in [0,\theta_{\max}]^K$ and weights $\mathbf{r} \in [0,1]^K$, the weighted regret of AOA-RB$_{PL}$ (Alg. 1) with $x = 2\log T$*

$$Reg_T^{wtd} = O(\sqrt{\theta_{\max}KT}\log T) \qquad \text{when} \quad T \to \infty\,.$$

The complete proof is postponed to App. B. The rate $\Omega(\sqrt{KT})$ is optimal as proved by the lower bound in [19] for MNL bandit problems for $\theta_{\max} = 1$. Our result recovers (up to a factor $\sqrt{\log T}$) the one of [2] when $\theta_{\max} = 1$. However, their algorithm relies on more sophisticated estimators that necessitate epochs repeating the same assortment until the No-Choice item is selected. Note for our problem setting, where it is possible to have $\theta_{\max} \gg \theta_0 = 1$, the length of these epochs could be of $O(K\theta_{\max})$, which could be potentially very large when $\theta_{\max} \gg 1$. This reduces the number of effective epochs, leading to poor estimation of the PL parameters. We see this tradeoff in our experiments (Sec. 5) where the MNL-UCB algorithm of [2] yields linear $O(T)$ regret for such choice of the problem parameters.

Remark 1 (Beyond MNL Models)**.** *Although, in this paper, we primarily focused on MNL based choice models, it is worth mentioning that our proposed algorithms can be generalized to more general random utility based models (RUMs) [9, 33] pursuing the ideas from [36] that extends the RB based parameter estimation technique to any RUM($\boldsymbol{\theta}$) choice models. Our algorithms and analyses thus apply to any general RUM($\boldsymbol{\theta}$) based choice models; we stick to the special case of MNL models in this paper for brevity and keep the main focus on the AOA problem and the related algorithmic novelties.*

Proof sketch of Thm. 5. Let $\mathcal{E}$ be the high-probability event such that both Lemma 1 and 2 are satisfied. Then,

$$\begin{aligned}
Reg_T^{\texttt{wtd}} &= \sum_{t=1}^{T} \mathbb{E}\big[\mathcal{R}(S^*,\theta) - \mathcal{R}(S_t,\theta)\big] \lesssim \sum_{t=1}^{T} \mathbb{E}\big[(\mathcal{R}(S^*,\theta) - \mathcal{R}(S_t,\theta))\mathbb{1}\{\mathcal{E}\}\big] + T\mathbb{P}(\mathcal{E}^c) \\
&\lesssim \sum_{t=1}^{T} \mathbb{E}\big[(\mathcal{R}(S_t,\theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta))\mathbb{1}\{\mathcal{E}\}\big] + T\mathbb{P}(\mathcal{E}^c) \qquad (5)
\end{aligned}$$

because $\mathcal{R}(S_t,\theta_t^{\texttt{ucb}}) \geq \mathcal{R}(S^*,\theta_t^{\texttt{ucb}}) \geq \mathcal{R}(S^*,\theta)$ under the event $\mathcal{E}$ by Lemma 4. We now upper-bound the first term of the right-hand-side

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big)\Big)\mathbb{1}\{\mathcal{E}\}\Big] &= \sum_{t=1}^{T}\mathbb{E}\Bigg[\bigg(\sum_{i\in S_t}\frac{r_i\theta_{i,t}^{\texttt{ucb}}}{\theta_0+\Theta_{S_t,t}^{\texttt{ucb}}} - \frac{r_i\theta_i}{\theta_0+\Theta_{S_t}}\bigg)\mathbb{1}\{\mathcal{E}\}\Bigg] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\Bigg[\bigg(\sum_{i\in S_t}\frac{r_i(\theta_{i,t}^{\texttt{ucb}}-\theta_i)}{\theta_0+\Theta_{S_t}}\bigg)\mathbb{1}\{\mathcal{E}\}
\end{aligned}$$

Because $\Theta_{S_t,t}^{\texttt{ucb}} \geq \Theta_{S_t}$ under the event $\mathcal{E}$ by Lemma 1. Then, using $r_i \leq 1$, we further upper-bound using an exploration parameter $\tau_0 = O(\log(T))$ so that the upper-confidence-bounds in Lemmas 1 and 2 are satisfied

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big)\Big)\mathbb{1}\{\mathcal{E}\}\Big] \leq \sum_{i=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\bigg(\frac{|\theta_{i,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\bigg)\mathbb{1}\{i\in S_t,\mathcal{E}\}\right] \\
&\lesssim O(\tau_0) + \sum_{i=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\frac{|\theta_{i,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\mathbb{1}\{i\in S_t,\tau_{i,t}\geq\tau_0,\mathcal{E}\}\right] \\
&\lesssim O(\tau_0) + \sum_{i=1}^{K}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\frac{\theta_i\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}\right]} \times \underbrace{\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\bigg(\frac{\theta_{i,t}^{\texttt{ucb}}-\theta_i}{\theta_0+\Theta_{S_t}}\bigg)^2\frac{\theta_0+\Theta_{S_t}}{\theta_i}\mathbb{1}\{i\in S_t,\tau_{i,t}\geq\tau_0,\mathcal{E}\}\right]}}_{=:A_T(i)} \qquad (6)
\end{aligned}$$

where the last inequality is by Cauchy-Schwarz inequality. Now, the term $A_T(i)$ above may be upper-bounded using Lemmas 1 and 2,

$$\begin{aligned}
A_T(i) &= \mathbb{E}\left[\frac{(\theta_{i,t}^{\texttt{ucb}}-\theta_i)^2}{\theta_i(\theta_0+\Theta_{S_t})}\mathbb{1}\{i\in S_t,\tau_{i,t}\geq\tau_0,\mathcal{E}\}\right] \lesssim \sum_{t=1}^{T}\mathbb{E}\left[\frac{(\theta_0+\theta_i)^2 x}{n_{i0,t}(\theta_0+\Theta_{S_t})}\mathbb{1}\{i\in S_t\}\right] \\
&\lesssim \theta_{\max}x\sum_{t=1}^{T}\mathbb{E}\left[\frac{(\theta_0+\theta_i)\mathbb{1}\{i\in S_t\}}{(\theta_0+\Theta_{S_t})n_{i0,t}}\right] = \theta_{\max}x\mathbb{E}\left[\sum_{t=1}^{T}\frac{\mathbb{1}\{i_t\in\{i,0\},i\in S_t\}}{n_{i0,t}}\right] \lesssim \theta_{\max}x\log T
\end{aligned}$$

where in the last inequality we used that $\sum_{n=1}^{T} n^{-1} \leq 1+\log T$. Substituting into (6), Jensen's inequality entails,

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big(\mathcal{R}(S_t,\theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big)\mathbb{1}\{\mathcal{E}\}\Big] \lesssim O(\tau_0) + \mathbb{E}\left[\sqrt{\theta_{\max}x\log T}\sum_{i=1}^{K}\sqrt{\sum_{t=1}^{T}\frac{\theta_i\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}}\right]. \qquad (7)$$

The proof is finally concluded by applying Cauchy-Schwarz inequality which yields:

$$\sum_{i=1}^{K}\sqrt{\sum_{t=1}^{T}\frac{\theta_i\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}} \le \sqrt{K\sum_{t=1}^{T}\frac{\sum_{i=1}^{K}\theta_i\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}} \le \sqrt{KT}\,.$$

Finally, combining the above result with (5) and (7) concludes the proof

$$Reg_T^{\mathtt{wtd}} \lesssim TP(\mathcal{E}^c) + O(\tau_0) + \sqrt{\theta_{\max} x K T\log T}\,.$$

Choosing $x = 2\log T$ ensures $TP(\mathcal{E}^c) \le O(1)$ and $\tau_0 \le O(\log T)$. □

4 Improved dependance on $\theta_{\max}$ with Adaptive Pivot Selection

A problem with Algorithm 1 stems from estimating all θ_i based on pairwise comparisons with item 0. When $\theta_{\max} \gg \theta_0 = 1$, item 0 may not be sampled enough as the winner, leading to poor estimators. This deficiency contributes to the suboptimal dependence on $\theta_{\max}$ observed in Theorems 3 and 5 and in prior work, such as [2]. We propose the following fix to optimize the pivot. For all $i, j \in [K]\cup\{0\}$ we define $\gamma_{ij} = \frac{\theta_i}{\theta_j}$, and the estimators:

$$\gamma_{ij,t}^{\mathtt{ucb}} = p_{ij,t}^{\mathtt{ucb}}/(1-p_{ij,t}^{\mathtt{ucb}})_+ \qquad \text{and} \qquad \gamma_{ii,t}^{\mathtt{ucb}} = 1\,,$$

where $p_{ij,t}^{\mathtt{ucb}}$ are defined in (3). For all rounds t, the algorithm AOA-RB$_{\mathrm{PL}}$-Adaptive selects

$$S_t = \underset{|S|\le m}{\operatorname{argmax}}\,\mathcal{R}(S,\widehat{\theta}_t^{\mathtt{ucb}}) \qquad \text{where} \qquad \widehat{\theta}_{i,t}^{\mathtt{ucb}} := \min_{j\in[K]\cup\{0\}} \gamma_{ij,t}^{\mathtt{ucb}}\gamma_{j0,t}^{\mathtt{ucb}}\,.$$

We offer below a regret bound that underscores the value of optimizing the pivot when $\theta_{\max} \gg K$. Note that while the algorithm and analysis are presented for the weighted objective with winner feedback only, it can be adapted to other objectives by replacing $\mathcal{R}(S,\theta)$ with the new objective in the analysis, as long as Lemma 4 remains valid.

Theorem 6. *Let $\theta_{\max} \ge 1$. For any $\theta \in [0,\theta_{\max}]^K$ and weights $\mathbf{r} \in [0,1]^K$, the weighted regret of AOA-RB$_{PL}$-Adaptive is upper-bounded as*

$$Reg_T^{\mathtt{wtd}} = O\big(\sqrt{\min\{\theta_{\max},K\}KT}\log T\big)$$

as $T\to\infty$ for the choice $x = 2\log T$ (when definining $p_{ij,t}^{\mathtt{ucb}}$).

Asymptotically, when $\theta_{\max}$ is constant, the regret is $O(K\sqrt{T}\log T)$, eliminating any dependence on $\theta_{\max}$. This allows for handling scenarios where the No-Choice item is highly unlikely, which is not achievable in previous works such as [2, 1]. [2] did attempt in their Thm. 4 to relax the assumption of $\theta_{\max} = \theta_0$ and shows a bound of order $O\big(\max\{\theta_{\max}/\theta_0, 1\}^{1/2}\sqrt{KT}\big)$, which unfortunately blows to ∞ as $\theta_0 \to 0$ or equivalently $\theta_{\max}\to\infty$, leading to a vacuous bound. Here, lies the stark improvement and one of the key contributions, as also corroborated in our experimental evaluation Sec. 5 (Fig. 2).

The proof is deferred to the App. B, with a key step relying on selecting the pivot $j_t = \operatorname{argmax}_{j\in S_t\cup\{0\}}\theta_j$. The use of $|\widehat{\theta}_{i,t}^{\mathtt{ucb}} - \theta_i| \le |\gamma_{ij_t,t}^{\mathtt{ucb}} - \theta_i|$ provides confidence upper-bounds with an improved dependence on $\theta_{\max}$, leveraging the fact that $\theta_{j_t} \ge \theta_i$. Due to the varying pivot over time, a telescoping argument introduces an additive factor $\sqrt{K}$.

5 Experiments

We provide here a synthetic experiments. All results are averaged across 100 runs. We evaluate the performance of our main algorithm AOA-RB$_{\mathrm{PL}}$-Adaptive (Sec. 4), referred as "Our Alg-1 (Adaptive Pivot)", with the following two algorithms: AOA-RB$_{\mathrm{PL}}$ (Sec. 3) referred as "Our Alg-2 (No-Choice Pivot)", and MNL-UCB, the state-of-the-art algorithm for AOA ([2], Alg. 1).

Different PL ($\boldsymbol{\theta}$) Environments. We report our experiment results on two datasets with $K = 50$ items: (1) Arith50 with PL parameters $\theta_i = 1-(i-1)0.2,\ \forall i\in[50]$. (2) Bad50 with PL parameters $\theta_i = 0.6,\ \forall i \in [50] \setminus \{25\}$ and $\theta_{25} = 0.8$. For simplicity of computing the assortment choices S_t, we assume $r_i = 1,\ \forall i \in [K]$.

(1). Averaged Regret with weak NC ($\theta_{\max}/\theta_0 \gg 1$) **(Fig. 1):** In our first experiment, we set set $m = 5$ and $\theta_0/\theta_{\max} = 0.01$ and report the average regret of the above three algorithms for our two objectives.

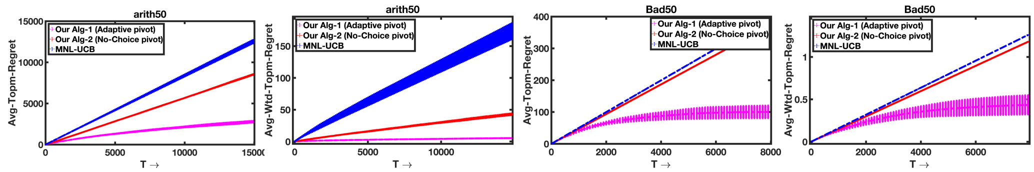

Figure 1: Averaged Regret for $m = 5$, $\theta_0 = 0.01$

Fig. 1 shows that our algorithm AOA-RB$_{\text{PL}}$-Adaptive (with adaptive pivot) significantly outperforms the other two algorithms, while our algorithm AOA-RB$_{\text{PL}}$ with no-choice (NC) pivot still outperforms MNL-UCB.

(2). Averaged Regret vs No-Choice PL Parameter ($\theta_{\max}/\theta_0$) **(Fig. 2):** In this experiment, we evaluate the regret performance of our algorithm AOA-RB$_{\text{PL}}$-Adaptive. We report the experiment on Artith50 PL dataset and set the subsetsize $m = 5$, $\theta_{\max}/\theta_0 = \{1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$. Fig. 2 shows the increase in the performance gap between our algorithm AOA-RB$_{\text{PL}}$-Adaptive (with adaptive pivot) with decreasing $\theta_0/\theta_{\max}$.

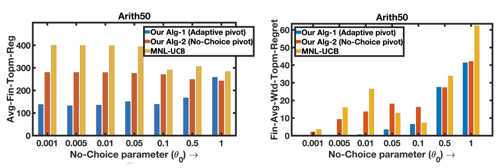

Figure 2: Comparative performance for varying $\theta_0/\theta_{\max}$, $m = 5$

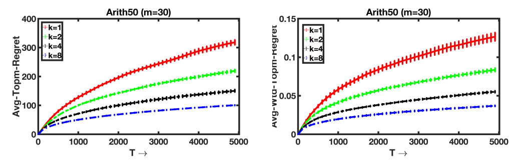

Figure 3: Tradofff: Averaged Regret vs length of the k rank-ordered feedback

(3). Averaged Regret vs Length of the rank-ordered feedback (k) **(Fig. 3):** We also run a thought experiment to understand the tradeoff between learning rate with k-length rank-ordered feedback, where given any assortment $S_t \subseteq [K]$ of size m, the learner gets to see the top-k draws ($k \le m$) from the PL model without replacement. This is a stronger feedback than the winner (i.e. top-1 for $k = 1$) feedback and, as expected, we see in Fig. 3 an improved regret (for both notions) when increasing k. The experiment are run on the Artith50 dataset with $m = 30$ and $k \in \{1, 2, 4, 8\}$.

6 Conclusion

We address the Active Optimal Assortment Selection problem with PL choice models, introducing a versatile framework (*AOA*) that eliminates the need for a strong default item, typically assumed as the No-Choice (NC) item in the existing literature. Our proposed algorithms employ a novel 'Rank-Breaking' technique to establish tight concentration guarantees for estimating the score parameters of the PL model. Our approach stands out for its practicality and avoids the suboptimal practice of repeatedly selecting the same set of items until the default item prevails. This is beneficial when the default item's quality (θ_0) is significantly lower than the quality of the best item ($\theta_{\max}$). Our algorithms are computationally efficient, optimal (up to log factors), and free from restrictive assumptions on the default item.

Future Works. Among many interesting questions to address in the future, it will be interesting to understand the role of the No-Choice (NC) item in the algorithm design, precisely, can we design efficient algorithms without the existence of NC items with a regret rate still linear in $\theta_{\max}$? Further, it will be interesting to extend our results to more general choice models beyond the PL model [18, 22, 23]. What is the tradeoff between the subsetsize m and the regret for such general choice models? Extending our results to large (potentially infinite) decision spaces and contextual settings would also be a very useful and practical contribution to the literature of assortment optimization.

References

[1] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pages 76–78. PMLR, 2017.

[2] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

[3] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

[4] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[5] Peter Auer. Using upper confidence bounds for online learning. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 270–279. IEEE, 2000.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[7] Vashist Avadhanula, Jalaj Bhandari, Vineet Goyal, and Assaf Zeevi. On the tightness of an lp relaxation for rational optimization and its applications. *Operations Research Letters*, 44(5):612–617, 2016.

[8] Hossein Azari, David Parkes, and Lirong Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

[9] Hossein Azari, David Parks, and Lirong Xia. Random utility theory for social choice. *Advances in Neural Information Processing Systems*, 25, 2012.

[10] Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021.

[11] Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *J. Mach. Learn. Res.*, 22:7–1, 2021.

[12] Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pages 1764–1786. PMLR, 2022.

[13] Gerardo Berbeglia and Gwenaël Joret. Assortment optimisation under a general discrete choice model: A tight analysis of revenue-ordered assortments. *arXiv preprint arXiv:1606.01371*, 2016.

[14] Brian Brost, Yevgeny Seldin, Ingemar J. Cox, and Christina Lioma. Multi-dueling bandits and their application to online ranker evaluation. *CoRR*, abs/1608.06253, 2016.

[15] Niladri S Chatterji, Aldo Pacchiano, Peter L Bartlett, and Michael I Jordan. On the theory of reinforcement learning with once-per-episode feedback. *arXiv preprint arXiv:2105.14363*, 2021.

[16] Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1245–1264. SIAM, 2017.

[17] Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2504–2522. SIAM, 2018.

[18] Xi Chen, Chao Shi, Yining Wang, and Yuan Zhou. Dynamic assortment planning under nested logit models. *Production and Operations Management*, 30(1):85–102, 2021.

[19] Xi Chen and Yining Wang. A note on a tight lower bound for mnl-bandit assortment selection models. *arXiv preprint arXiv:1709.06109*, 2017.

[20] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[21] James Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Work in Progress*, 2013.

[22] Antoine Désir, Vineet Goyal, Srikanth Jagabathula, and Danny Segev. Assortment optimization under the mallows model. In *Advances in Neural Information Processing Systems*, pages 4700–4708, 2016.

[23] Antoine Désir, Vineet Goyal, Danny Segev, and Chun Ye. Capacity constrained assortment optimization under the markov chain based choice model. *Operations Research*, 2016.

[24] James A Grant and David S Leslie. Learning to rank under multinomial logit choice. *Journal of Machine Learning Research*, 24(260):1–49, 2023.

[25] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems*, pages 1685–1695, 2017.

[26] Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.

[27] Kameng Nip, Zhenbo Wang, and Zizhuo Wang. Assortment optimization under a single transition model. 2017.

[28] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Mingdong Ou, Nan Li, Shenghuo Zhu, and Rong Jin. Multinomial logit bandit with linear utility functions. *arXiv preprint arXiv:1805.02971*, 2018.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[31] Wenbo Ren, Jia Liu, and Ness B Shroff. PAC ranking from pairwise and listwise queries: Lower bounds and upper bounds. *arXiv preprint arXiv:1806.02970*, 2018.

[32] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

[33] Aadirupa Saha and Suprovat Ghoshal. Exploiting correlation to achieve faster learning rates in low-rank preference bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 456–482. PMLR, 2022.

[34] Aadirupa Saha and Aditya Gopalan. Active ranking with subset-wise preferences. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[35] Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019.

[36] Aadirupa Saha and Aditya Gopalan. PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, pages 700–737, 2019.

[37] Aadirupa Saha and Aditya Gopalan. Best-item learning in random utility models with subset choices. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4291. PMLR, 2020.

[38] Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *ICML*, pages 360–368, 2014.

[39] Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Multi-dueling bandits with dependent arms. In *Conference on Uncertainty in Artificial Intelligence*, UAI'17, 2017.

[40] Kalyan Talluri and Garrett Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

[41] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[42] Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2827–2836, 2019.

[44] Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pages 307–315, 2015.

[45] Masrour Zoghi, Shimon Whiteson, Remi Munos, Maarten de Rijke, et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pages 10–18. JMLR, 2014.

[46] Masrour Zoghi, Shimon A Whiteson, Maarten De Rijke, and Remi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 73–82. ACM, 2014.

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: In the abstract, we list the main claims of this paper in a general fashion. Then, in the introduction we state them in more detail. They accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer NA 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 NA 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 discuss the limitations and assumptions of our work throughout the paper. Additional limitations are highlighted in the discussion.

   Guidelines:

   - The answer NA 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: The assumptions can be found in the Problem Setup section and in the paragraphs before the theorems and remarks. We provide proof sketches in the main text and complete proofs in the appendix.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: experimental details including algorithms and setups are clearly provided. In addition, the main contribution of the paper is theoretical and synthetic experiments are mostly provided as an illustration.

Guidelines:

- The answer NA 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: [No]

Justification: experimental setups are synthetic and can easily be reproduced.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All details are provided to reproduce the experiments.

Guidelines:

- The answer NA 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: [Yes]

Guidelines:

- The answer NA 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: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA 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: We do not see any potential negative social impact of this work and it follows the NeurIPS code of ethics.

   Guidelines:

   - The answer NA 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: [NA]

Justification: This work addresses the problem of designing efficient and optimal algorithms for different assortment selection problems with MNL models. Our work is purely theoretical and studies a fundamental mathematical optimization framework that is unrelated to societal considerations

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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.

Supplementary: Optimal, Efficient and Practical Algorithms for Assortment Optimization

A Preliminaries: Some Useful Concepts for PL choice models

A.1 Plackett-Luce (PL): A Discrete Choice Model

A discrete choice model specifies the relative preferences of two or more discrete alternatives in a given set. A widely studied class of discrete choice models is the class of *Random Utility Models* (RUMs), which assume a ground-truth utility score $\theta_i \in \mathbb{R}$ for each alternative $i \in [n]$, and assign a conditional distribution $\mathcal{D}_i(\cdot|\theta_i)$ for scoring item i. To model a winning alternative given any set $S \subseteq [n]$, one first draws a random utility score $X_i \sim \mathcal{D}_i(\cdot|\theta_i)$ for each alternative in S, and selects an item with the highest random score.

One widely used RUM is the *Multinomial-Logit (MNL)* or *Plackett-Luce model (PL)*, where the $\mathcal{D}_i$s are taken to be independent Gumbel distributions with parameters θ'_i [8], i.e., with probability densities

$$\mathcal{D}_i(x_i|\theta'_i) = e^{-(x_j-\theta'_j)}e^{-e^{-(x_j-\theta'_j)}}, \qquad \theta'_i \in R,\ \forall i \in [n]\,.$$

Moreover assuming $\theta'_i = \ln\theta_i$, $\theta_i > 0\ \forall i \in [n]$, it can be shown in this case the probability that an alternative i emerges as the winner in the set $S \ni i$ becomes: $\mathbb{P}(i|S) = \frac{\theta_i}{\sum_{j\in S}\theta_j}$.

Other families of discrete choice models can be obtained by imposing different probability distributions over the utility scores X_i, e.g. if $(X_1,\ldots X_n) \sim \mathcal{N}(\boldsymbol{\theta},\boldsymbol{\Lambda})$ are jointly normal with mean $\boldsymbol{\theta} = (\theta_1,\ldots\theta_n)$ and covariance $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$, then the corresponding RUM-based choice model reduces to the *Multinomial Probit (MNP)*.

A.2 Rank Breaking

Rank breaking (RB) is a well-understood idea involving the extraction of pairwise comparisons from (partial) ranking data, and then building pairwise estimators on the obtained pairs by treating each comparison independently [26, 25], e.g., a winner a sampled from among a, b, c is rank-broken into the pairwise preferences $a \succ b$, $a \succ c$. We use this idea to devise estimators for the pairwise win probabilities $p_{ij} = \mathbb{P}(i|\{i,j\}) = \theta_i/(\theta_i+\theta_j)$ for our problem setting. We used the idea of RB in both our algorithms (AOA-RB$_{\text{PL}}$ and AOA-RB$_{\text{PL}}$-Adaptive) to update the pairwise win-count estimates $w_{i,j,t}$ for all the item pairs $(i,j) \in [K]\times[K]$, which is further used for deriving the empirical pairwise preference estimates $\widehat{p}_{ij,t}$, at any time t.

A.3 Parameter Estimation with PL based preference data

Lemma 7 (Pairwise win-probability estimates for the PL model [34])**.** *Consider a Plackett-Luce choice model with parameters $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots,\theta_n)$, and fix two items $i,j \in [n]$. Let $S_1,\ldots,S_T$ be a sequence of (possibly random) subsets of $[n]$ of size at least 2, where T is a positive integer, and $i_1,\ldots,i_T$ a sequence of random items with each $i_t \in S_t$, $1 \le t \le T$, such that for each $1 \le t \le T$, (a) S_t depends only on $S_1,\ldots,S_{t-1}$, and (b) i_t is distributed as the Plackett-Luce winner of the subset S_t, given $S_1, i_1,\ldots,S_{t-1}, i_{t-1}$ and S_t, and (c) $\forall t : \{i,j\} \subseteq S_t$ with probability 1. Let $n_i(T) = \sum_{t=1}^T \mathbb{P}(i_t = i)$ and $n_{ij}(T) = \sum_{t=1}^T \mathbb{P}(\{i_t \in \{i,j\}\})$. Then, for any positive integer v, and $\eta \in (0,1)$,*

$$\mathbb{P}\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i+\theta_j} \ge \eta,\ n_{ij}(T) \ge v\right) \le e^{-2v\eta^2},$$
$$\mathbb{P}\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i+\theta_j} \le -\eta,\ n_{ij}(T) \ge v\right) \le e^{-2v\eta^2}.$$

B Omitted Proofs from Sec. 3 and Sec. 4

B.1 A concentration bounds for the $p_{ij,t}$

We first prove below a concentration inequality based on Bernstein's inequality for the estimators $p_{ij,t}$.

Lemma 8. *Let $(i,j) \in [K] \times [K]$. Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - 3Te^{-x}$,*

$$p_{ij} \leq p_{ij,t}^{\mathtt{ucb}} \leq p_{ij} + 2\sqrt{\frac{2p_{ij}(1-p_{ij})x}{n_{ij,t}}} + \frac{11x}{n_{ij,t}}\,, \tag{8}$$

simultaneously for all $t \in [T]$.

Proof of Lemma 8. Let $T \geq 1$, $x > 0$ and $i,j \in [K]$. Applying Thm. 1 of [4], with probability at least $1 - \beta(x,T)$, we get simultaneously for all $t \in [T]$,

$$\left|\widehat{p}_{ij,t} - p_{ij}\right| \leq \sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{3x}{n_{ij,t}}\,, \tag{9}$$

where $\beta(x,T) = 3\inf_{1<\alpha\leq 3}\min\left\{\frac{\log T}{\log\alpha}, T\right\}e^{-x/\alpha} \leq 3Te^{-x}$. Note that the inequality holds true although $n_{ij,t}$ is a random variable. This, shows the first inequality

$$p_{ij} \leq p_{ij,t}^{\mathtt{ucb}}\,.$$

For the second inequality, (9) implies

$$\begin{aligned} p_{ij,t}^{\mathtt{ucb}} &= \widehat{p}_{ij,t} + \sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{3x}{n_{ij,t}} \\ &\leq p_{ij} + 2\sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{6x}{n_{ij,t}}\,. \end{aligned} \tag{10}$$

Furthermore, because $x \mapsto x(1-x)$ is 1-Lipschitz on $[0,1]$, we have

$$\begin{aligned} \left|\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t}) - p_{ij}(1-p_{ij})\right| &\leq \left|\widehat{p}_{ij,t} - p_{ij}\right| \\ &\overset{(9)}{\leq} \sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{3x}{n_{ij,t}}\,. \end{aligned}$$

Therefore,

$$\begin{aligned} \widehat{p}_{ij,t}(1-\widehat{p}_{ij,t}) &\leq p_{ij}(1-p_{ij}) + \sqrt{\frac{2\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})x}{n_{ij,t}}} + \frac{3x}{n_{ij,t}} \\ &\leq \left(\sqrt{p_{ij}(1-p_{ij})} + \sqrt{\frac{3x}{n_{ij,t}}}\right)^2, \end{aligned}$$

which yields

$$\sqrt{\widehat{p}_{ij,t}(1-\widehat{p}_{ij,t})} \leq \sqrt{p_{ij}(1-p_{ij})} + \sqrt{\frac{3x}{n_{ij,t}}}\,. \tag{11}$$

Plugging back into (10), we get

$$p_{ij,t}^{\mathtt{ucb}} \leq 2\sqrt{\frac{2p_{ij}(1-p_{ij})x}{n_{ij,t}}} + \frac{11x}{n_{ij,t}}\,.$$

□

B.2 Proof of Lemma 1

Proof. Let $i \in [K]$ and $x > 0$. Then, by a union bound on Lemma 8 and 2, with probability at least $1 - 4Te^{-x}$, (8) and (4) hold true for all $t \in [T]$. We consider this high-probability event in the rest of the proof. Define the function $f : x \mapsto x/(1-x)_+$ on $[0,1]$ (with the convention $f(1) = +\infty$), so that $\theta_{i,t}^{\texttt{ucb}} = f(p_{i0,t}^{\texttt{ucb}})$ and $\theta_i = f(p_{i0})$. Because f is non-decreasing, and $p_{i0,t}^{\texttt{ucb}} \ge p_{i0}$ by (8), we have

$$\theta_{i,t}^{\texttt{ucb}} \ge \theta_i \,. \tag{12}$$

Furthermore, denote

$$\Delta_{i,t} := 2\sqrt{\frac{2p_{ij}(1-p_{ij})x}{n_{i0,t}}} + \frac{11x}{n_{i0,t}} = 2\sqrt{\frac{2\theta_0\theta_i x}{(\theta_0+\theta_i)^2 n_{i0,t}}} + \frac{11x}{n_{i0,t}} \,. \tag{13}$$

In the rest of the proof we assume, $n_{i0,t} \ge 69x(\theta_0 + \theta_i)$. Then, using that $\theta_0\theta_i \le \theta_0 + \theta_i$ since $\theta_0 = 1$, it implies

$$(\theta_0+\theta_i)\Delta_{i,t} \le 2\sqrt{\frac{2\theta_0\theta_i x}{n_{i0,t}}} + \frac{11x(\theta_0+\theta_i)}{n_{i0,t}} \le \frac{1}{2} \,,$$

and

$$p_{i0} + \Delta_{i,t} = \frac{\theta_i}{\theta_0+\theta_i} + \Delta_{i,t} \le \frac{\theta_i + 1/2}{\theta_i + 1} < 1.$$

Thus, because f is non-decreasing

$$\begin{aligned}
\theta_{i,t}^{\texttt{ucb}} - \theta_i &= f(p_{i0,t}^{\texttt{ucb}}) - f(p_{i0}) \\
&\overset{(8)}{\le} f\big(p_{i0} + \Delta_{i,t}\big) - f(p_{i0}) \\
&= \frac{p_{i0} + \Delta_{i,t}}{1 - p_{i0} - \Delta_{i,t}} - \frac{p_{i0}}{1-p_{i0}} \\
&= \frac{\Delta_{i,t}}{(1-p_{i0})(1-p_{i0}-\Delta_{i,t})} \\
&= \frac{(\theta_0+\theta_i)^2\Delta_{i,t}}{1-(\theta_0+\theta_i)\Delta_{i,t}} \\
&\le 2(\theta_0+\theta_i)^2\Delta_{i,t} \\
&\overset{(13)}{\le} 4(\theta_0+\theta_i)\sqrt{\frac{2\theta_0\theta_i x}{n_{i0,t}}} + \frac{22x(\theta_0+\theta_i)^2}{n_{i0,t}} \,,
\end{aligned}$$

which concludes the proof. □

B.3 Proof of Lemma 2

Proof. Let $T \ge 1$ and $i \in [K]$. Recall that $\tau_{i,t} = \sum_{s=1}^{t-1} \mathbb{1}\{i \in S_s\}$ is the number of times i was played at the start of round t and $n_{i0,t} = \sum_{s=1}^{t-1} \mathbb{1}\{i_t \in \{i,0\}, i \in S_t\}$ is the number of times i or 0 won up to round t when played together. When i is played the probability of 0 or i to win is

$$\mathbb{P}(i_t \in \{i,0\}|S_t) = \frac{\theta_0+\theta_i}{\theta_0+\Theta_{S_t}} \ge \frac{\theta_0+\theta_i}{\theta_0+\Theta_{S^*}} \,.$$

Therefore, applying Chernoff-Hoeffding inequality together with a union bound (to deal with the fact that $\tau_{i,t}$ is random), we have with probability at least $1 - Te^{-x}$

$$n_{i0,t} \ge \frac{\theta_0+\theta_i}{\theta_0+\Theta_{S^*}}\tau_{i,t} - \sqrt{\frac{\tau_{i,t}x}{2}}$$

simultaneously for all $t \in [T]$. Noting that

$$\frac{\theta_0+\theta_i}{\theta_0+\Theta_{S^*}}\tau_{i,t} - \sqrt{\frac{\tau_{i,t}x}{2}} \ge \frac{\theta_0+\theta_i}{2(\theta_0+\Theta_{S^*})}\tau_{i,t}$$

if $\tau_{i,t} \ge 2x(\theta_0+\Theta_{S^*})^2 \ge \frac{2x(\theta_0+\Theta_{S^*})^2}{(\theta_0+\theta_i)^2}$ concludes the proof. □

B.4 Proof of Theorem 3

Proof. Let us define for any $S \subseteq [K]$,

$$\Theta_S = \sum_{i \in S} \theta_i, \quad \text{and} \quad \Theta_S^{\texttt{ucb}} := \sum_{i \in S} \theta_i^{\texttt{ucb}}.$$

Let $\mathcal{E}$ be the high-probabality event such that both Lemma 1 and 2 holds true. Then, $\mathbb{P}(\mathcal{E}) \geq 1 - 4TKe^{-x}$. Let us first assume that $\mathcal{E}$ holds true. Then, by Lemma 1,

$$\begin{aligned} Reg_T^{\texttt{top}} &= \frac{1}{m} \sum_{t=1}^T \Theta_{S^*} - \Theta_{S_t} \\ &\leq \frac{1}{m} \sum_{t=1}^T \min\Big\{\Theta_{S^*}, \Theta_{S_t}^{\texttt{ucb}} - \Theta_{S_t}\Big\} \quad \leftarrow \text{ because } \Theta_{S^*} \leq \Theta_{S^*}^{\texttt{ucb}} \leq \Theta_{S_t}^{\texttt{ucb}} \text{ under the event } \mathcal{E} \\ &= \frac{1}{m} \sum_{t=1}^T \min\Big\{\Theta_{S^*}, \sum_{i \in S_t} \theta_{i,t}^{\texttt{ucb}} - \theta_i\Big\} \\ &\leq \frac{1}{m} \Theta_{S^*} \sum_{i=1}^K \bar{\tau}_{i0} + \frac{1}{m} \sum_{t=1}^T \sum_{i \in S_t} (\theta_{i,t}^{\texttt{ucb}} - \theta_i) \mathbb{1}\big\{\tau_{i,t} \geq \bar{\tau}_{i0}\big\} \end{aligned}$$

where $\bar{\tau}_{i0} = 2x(\theta_0 + \Theta_{S^*}) \max\{\theta_0 + \Theta_{S^*}, 69\} \leq 138x(m+1)^2\theta_{\max}^2$, where $\theta_{\max} := \max_i \theta_i$. Then, noting that if $\mathcal{E}$ holds true, by Lemma 2, we also have $n_{i0,t} \geq \frac{1}{2(\theta_0 + \Theta_{S^*})}(\theta_0 + \theta_i)\tau_{i,t}$, which yields

$$\mathbb{1}\{\tau_{i,t} \geq \bar{\tau}_{i0}\} \leq \mathbb{1}\{n_{i0,t} \geq 69x(\theta_0 + \theta_i)\}.$$

Therefore, we can apply Lemma 1 that entails,

$$\begin{aligned} &\frac{1}{m} \sum_{t=1}^T \sum_{i \in S_t} (\theta_{i,t}^{\texttt{ucb}} - \theta_i) \mathbb{1}\big\{\tau_{i,t} \geq \bar{\tau}_{i0}\big\} \\ &\quad \stackrel{\text{Lem. 1}}{\leq} \frac{1}{m} \sum_{t=1}^T \sum_{i \in S_t} \Big(4(\theta_0 + \theta_i)\sqrt{\frac{2\theta_0\theta_i x}{n_{i0,t}}} + \frac{22x(\theta_0 + \theta_i)^2}{n_{i0,t}}\Big) \mathbb{1}\big\{n_{i0,t} \geq 69x(\theta_0 + \theta_i)\big\} \\ &\quad \stackrel{\text{Lem 2}}{\leq} \frac{1}{m} \sum_{t=1}^T \sum_{i \in S_t} \left(8\sqrt{\frac{(\theta_0 + \Theta_{S^*})(\theta_0 + \theta_i)\theta_0\theta_i x}{\tau_{i,t}}} + \frac{44x(\theta_0 + \Theta_{S^*})(\theta_0 + \theta_i)}{\tau_{i,t}}\right) \\ &\quad \leq \frac{1}{m} \sum_{i=1}^K 16\sqrt{(\theta_0 + \Theta_{S^*})(\theta_0 + \theta_i)\theta_0\theta_i x \tau_{i,T}} + 44x(\theta_0 + \Theta_{S^*}) \sum_{i=1}^K (\theta_0 + \theta_i)(1 + \log(\tau_{i,T}))\,, \end{aligned}$$

where we used $\sum_{i=1}^n 1/\sqrt{i} \leq 2\sqrt{n}$ and $\sum_{i=1}^n i^{-1} \leq 1 + \log n$. We thus have

$$\begin{aligned} Reg_T^{\texttt{top}} &\leq 138x(m+1)^2 K\theta_{\max}^3 + \frac{1}{m} \sum_{i=1}^K 16\theta_{\max}^{3/2}\sqrt{(m+1)x\tau_{i,T}} \\ &\qquad\qquad + 44x(m+1)(1+\theta_{\max})^2 \sum_{i=1}^K (1 + \log(\tau_{i,T})) \\ &\leq 138x(m+1)^2 K\theta_{\max}^3 + 16\theta_{\max}^{3/2}\sqrt{2xKT} + 88x(m+1)K\theta_{\max}^2\Big(1 + \log\Big(\frac{mT}{K}\Big)\Big)\,. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[Reg_T^{\texttt{top}}] \leq 12\sqrt{2}xmK\theta_{\max}^3 + 16\theta_{\max}^{3/2}\sqrt{2xKT} + 88xmK\theta_{\max}^2\Big(1 + \log\Big(\frac{mT}{K}\Big)\Big) \\ + 4mKT^2e^{-x}\theta_{\max}\,. \end{aligned}$$

Choosing $x = 2\log T$ concludes the proof. $\square$

B.5 Proof of Theorem 5

Proof. Let $\mathcal{E}$ be the high-probabality event such that Lemma 1 and 2 are satisfied, so that $\mathbb{P}(\mathcal{E}) \geq 1 - 4KTe^{-x}$. Then, denoting $x \wedge y := \min\{x, y\}$,

$$\begin{aligned} Reg_T^{\texttt{wtd}} &= \sum_{t=1}^{T} \mathbb{E}\big[\mathcal{R}(S^*, \theta) - \mathcal{R}(S_t, \theta)\big] \\ &= \sum_{t=1}^{T} \mathbb{E}\big[(\mathcal{R}(S^*, \theta) - \mathcal{R}(S_t, \theta))\mathbb{1}\{\mathcal{E}\} + (\mathcal{R}(S^*, \theta) - \mathcal{R}(S_t, \theta))\mathbb{1}\{\mathcal{E}^c\}\big] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\big((\mathcal{R}(S_t, \theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t, \theta)) \wedge \mathcal{R}(S^*, \theta)\big)\mathbb{1}\{\mathcal{E}\} + \mathcal{R}(S^*, \theta)\mathbb{1}\{\mathcal{E}^c\}\Big] \end{aligned} \tag{14}$$

because $\mathcal{R}(S_t, \theta_t^{\texttt{ucb}}) \geq \mathcal{R}(S^*, \theta_t^{\texttt{ucb}}) \geq \mathcal{R}(S^*, \theta)$ under the event $\mathcal{E}$ by Lemma 4. Then, using $\mathcal{R}(S^*, \theta) \leq \max_i r_i \leq 1$, we get

$$\begin{aligned} Reg_T^{\texttt{wtd}} &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\big((\mathcal{R}(S_t, \theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t, \theta)) \wedge 1\big)\mathbb{1}\{\mathcal{E}\} + \mathbb{1}\{\mathcal{E}^c\}\Big] \\ &\leq 4T^2Ke^{-x} + \sum_{t=1}^{T} \mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t, \theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t, \theta)\big) \wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big] . \end{aligned}$$

Let us upper-bound the second term of the right-hand-side

$$\sum_{t=1}^{T} \mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t, \theta_t^{\texttt{ucb}}) - \mathcal{R}(S_t, \theta)\big) \wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big] \tag{15}$$

$$\begin{aligned} &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\left(\sum_{i \in S_t} \frac{r_i\theta_{i,t}^{\texttt{ucb}}}{\theta_0 + \Theta_{S_t,t}^{\texttt{ucb}}} - \frac{r_i\theta_i}{\theta_0 + \Theta_{S_t}}\right) \wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[\left(\left(\sum_{i \in S_t} \frac{r_i(\theta_{i,t}^{\texttt{ucb}} - \theta_i)}{\theta_0 + \Theta_{S_t}}\right) \wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right] \qquad \text{because } \Theta_{S_t,t}^{\texttt{ucb}} \geq \Theta_{S_t} \text{ under } \mathcal{E} \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[\left(\left(\sum_{i \in S_t} \frac{|\theta_{i,t}^{\texttt{ucb}} - \theta_i|}{\theta_0 + \Theta_{S_t}}\right) \wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right] \qquad \text{because } r_i \leq 1 \\ &\leq \sum_{i=1}^{K} \mathbb{E}\left[\sum_{t=1}^{T} \left(\frac{|\theta_{i,t}^{\texttt{ucb}} - \theta_i|}{\theta_0 + \Theta_{S_t}} \wedge 1\right)\mathbb{1}\{i \in S_t\}\mathbb{1}\{\mathcal{E}\}\right] \\ &\leq 138xm^2K\theta_{\max}^2 + \sum_{i=1}^{K} \mathbb{E}\left[\sum_{t=1}^{T} \frac{|\theta_{i,t}^{\texttt{ucb}} - \theta_i|}{\theta_0 + \Theta_{S_t}}\mathbb{1}\{i \in S_t, \tau_{i,t} \geq 138x(m+1)^2\theta_{\max}^2\}\mathbb{1}\{\mathcal{E}\}\right] \\ &\leq 138xm^2K\theta_{\max}^2 + \sum_{i=1}^{K} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\frac{(\frac{\theta_0}{m} + \theta_i)\mathbb{1}\{i \in S_t\}}{\theta_0 + \Theta_{S_t}}\right]} \\ &\qquad \times \underbrace{\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\left(\frac{|\theta_{i,t}^{\texttt{ucb}} - \theta_i|}{\theta_0 + \Theta_{S_t}}\right)^2 \frac{\theta_0 + \Theta_{S_t}}{\frac{\theta_0}{m} + \theta_i}\mathbb{1}\{i \in S_t, \tau_{i,t} \geq 138x(m+1)^2\theta_{\max}^2\}\mathbb{1}\{\mathcal{E}\}\right]}}_{=:A_T(i)} \end{aligned} \tag{16}$$

where the last inequality is by Cauchy-Schwarz inequality. Now, the term $A_T(i)$ above may be upper-bounded as follows

$$A_T(i) := \sum_{t=1}^{T} \mathbb{E}\left[\left(\frac{|\theta_{i,t}^{\texttt{ucb}} - \theta_i|}{\theta_0 + \Theta_{S_t}}\right)^2 \frac{\theta_0 + \Theta_{S_t}}{\frac{\theta_0}{m} + \theta_i}\mathbb{1}\{i \in S_t, \tau_{i,t} \geq 138x(m+1)^2\theta_{\max}^2\}\mathbb{1}\{\mathcal{E}\}\right]$$

$$= \mathbb{E}\left[\frac{(\theta^{\texttt{ucb}}_{i,t}-\theta_i)^2}{\big(\frac{\theta_0}{m}+\theta_i\big)\theta_0+\Theta_{S_t}}\mathbb{1}\{i\in S_t, \tau_{i,t}\geq 138x(m+1)^2\theta_{\max}^2\}\mathbb{1}\{\mathcal{E}\}\right].$$

Now, since under the event $\mathcal{E}$ by Lemma 2, $\tau_{i,t}\geq 138x(m+1)^2\theta_{\max}^2$ implies

$$n_{i0,t}\geq 69x(\theta_0+\theta_i)(m+1)\theta_{\max}\geq 69x(\theta_0+\theta_i).$$

Therefore, we can apply Lemma 1, which further upper-bounds

$$\begin{aligned}
A_T(i) &\leq \sum_{t=1}^T \mathbb{E}\Bigg[\bigg(\frac{2^6(\theta_0+\theta_i)^2x}{n_{i0,t}}+\frac{2(22x)^2(\theta_0+\theta_i)^4}{n_{i0,t}^2(\frac{\theta_0}{m}+\theta_i)}\bigg) \\
&\qquad\qquad\qquad \times \frac{\mathbb{1}\{i\in S_t,\tau_{i,t}\geq 138x(m+1)^2\theta^2_{\max}\}}{\theta_0+\Theta_{S_t}}\mathbb{1}\{\mathcal{E}\}\Bigg] \\
&\leq \sum_{t=1}^T \mathbb{E}\Bigg[\bigg(\frac{2^6(\theta_0+\theta_i)^2x}{n_{i0,t}}+\frac{15x(\theta_0+\theta_i)^3}{n_{i0,t}\theta_{\max}(\theta_0+m\theta_i)}\bigg)\times\frac{\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}\mathbb{1}\{\mathcal{E}\}\Bigg]
\end{aligned}$$

where we used $n_{i0,t}\geq 69x(\theta_0+\theta_i)m\theta_{\max}$ in the last inequality. Then, we get

$$\begin{aligned}
A_T(i) &\leq \sum_{t=1}^T \mathbb{E}\Bigg[\bigg(\frac{(\theta_0+\theta_i)^2x}{n_{i0,t}}+\frac{30x(\theta_0+\theta_i)}{n_{i0,t}}\bigg)\times\frac{\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}\mathbb{1}\{\mathcal{E}\}\Bigg] \\
&\leq (94+64\theta_i)x\sum_{t=1}^T\mathbb{E}\left[\frac{(\theta_0+\theta_i)\mathbb{1}\{i\in S_t\}}{(\theta_0+\Theta_{S_t})n_{i0,t}}\right] \\
&= (94+64\theta_i)x\mathbb{E}\left[\sum_{t=1}^T\frac{\mathbb{1}\{i_t\in\{i,0\}, i\in S_t\}}{n_{i0,t}}\right] \\
&= (94+64\theta_i)x\mathbb{E}\Big[1+\log\big(n_{i0}(T)\big)\Big] \\
&\leq 158\theta_{\max}x(1+\log T).
\end{aligned}$$

Substituting into (16), we then obtain using Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sum_{t=1}^T\mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\theta^{\texttt{ucb}}_t)-\mathcal{R}(S_t,\theta)\big)\wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big] \\
&\quad\leq 138xm^2K\theta^2_{\max}+13\sqrt{\theta_{\max}x(1+\log T)}\sum_{i=1}^K\sqrt{\sum_{t=1}^T\mathbb{E}\left[\frac{\big(\frac{\theta_0}{m}+\theta_i\big)\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}\right]} \\
&\quad\leq 138xm^2K\theta^2_{\max}+13\sqrt{\theta_{\max}x(1+\log T)}\sqrt{\mathbb{E}\left[K\sum_{t=1}^T\frac{\sum_{i=1}^K\big(\frac{\theta_0}{m}+\theta_i\big)\mathbb{1}\{i\in S_t\}}{\theta_0+\Theta_{S_t}}\right]} \\
&\quad= 138xm^2K\theta^2_{\max}+13\sqrt{\theta_{\max}x(1+\log T)KT}.
\end{aligned}$$

Finally, replacing into Inequality (15) yields

$$Reg^{\texttt{wtd}}_T\leq 4T^2Ke^{-x}+138xm^2K\theta^2_{\max}+13\sqrt{\theta_{\max}x(1+\log T)KT}.$$

Choosing $x=2\log T$ concludes the proof. □

B.6 Proof of Theorem 6

The proof follows the one of Theorem 5, except that the concentration lemmas should be generalized to any pairs (i,j) instead of only with respect to item 0, whose proofs are left to the reader and closely follows the one of Lemma 1 and 2. For simplicity, this proof is performed up to universal multiplicative constants, using the rough inequality $\lesssim$.

Lemma 9. *Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - 3K(K+1)Te^{-x}$, simultaneously for all $t \in [T]$ and $i \neq j$ in $[\tilde{K}]$: $\gamma_{ij} := \frac{\theta_i}{\theta_j} \leq \gamma_{ij,t}^{\mathtt{ucb}}$ and one of the following two inequalities is satisfied*

$$n_{ij,t} < 69x(1+\gamma_{ij}) \qquad or \qquad \gamma_{ij,t}^{\mathtt{ucb}} \leq \gamma_{ij} + 4(\gamma_{ij}+1)\sqrt{\frac{2\gamma_{ij}x}{n_{ij,t}}} + \frac{22x(\gamma_{ij}+1)^2}{n_{ij,t}}\,.$$

Lemma 10. *Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - 3K(K+1)Te^{-x}$, simultaneously for all $t \in [T]$ and $i \in [K]$: $\widehat{\theta}_{i,t}^{\mathtt{ucb}} := \min_j \gamma_{ij,t}^{\mathtt{ucb}}\gamma_{j0,t}^{\mathtt{ucb}} \geq \theta_i$ and for all j one of the following two inequalities is satisfied*

$$n_{ij,t} \lesssim x(1+\gamma_{ij}) \qquad or \qquad n_{j0,t} \lesssim x(1+\theta_j)^2\theta_j^{-1}$$

or

$$\gamma_{ij,t}^{\mathtt{ucb}}\gamma_{j0,t}^{\mathtt{ucb}} - \theta_i \lesssim \sqrt{(\gamma_{ij}+1)\theta_i x}\left(\sqrt{\frac{(\theta_i+\theta_j)}{n_{ij,t}}} + \sqrt{\frac{(1+\theta_j)}{n_{j0,t}}}\right) + (\gamma_{ij}+1)\frac{(\theta_i+\theta_j)x}{n_{ij,t}} + \frac{\gamma_{ij}(1+\theta_j)^2 x}{n_{j0,t}}\,.$$

Proof of Lemma 10. The proof follows from Lemma 9. If $n_{ij,t} > Cx(1+\gamma_{ij})$ and $n_{j0,t} > Cx(1+\theta_j)$ for some large enough constant C, we have

$$\gamma_{ij,t}^{\mathtt{ucb}} \leq \gamma_{ij} + 4(\gamma_{ij}+1)\sqrt{\frac{2\gamma_{ij}x}{n_{ij,t}}} + \frac{22x(\gamma_{ij}+1)^2}{n_{ij,t}}$$

and

$$\gamma_{j0,t}^{\mathtt{ucb}} \leq \gamma_{j0} + 4(\gamma_{j0}+1)\sqrt{\frac{2\gamma_{j0}x}{n_{j0,t}}} + \frac{22x(\gamma_{j0}+1)^2}{n_{j0,t}} \leq 2\gamma_{j0}\,.$$

This implies,

$$\begin{aligned}
\gamma_{ij,t}^{\mathtt{ucb}}\gamma_{j0,t}^{\mathtt{ucb}} - \theta_i &= \gamma_{ij,t}^{\mathtt{ucb}}\gamma_{j0,t}^{\mathtt{ucb}} - \gamma_{ij}\gamma_{j0} = (\gamma_{ij,t}^{\mathtt{ucb}} - \gamma_{ij})\gamma_{j0,t}^{\mathtt{ucb}} + \gamma_{ij}(\gamma_{j0,t}^{\mathtt{ucb}} - \gamma_{j0}) \\
&\leq 2(\gamma_{ij,t}^{\mathtt{ucb}} - \gamma_{ij})\gamma_{j0} + \gamma_{ij}(\gamma_{j0,t}^{\mathtt{ucb}} - \gamma_{j0}) \\
&\leq 8\gamma_{j0}(\gamma_{ij}+1)\sqrt{\frac{2\gamma_{ij}x}{n_{ij,t}}} + \frac{44x\gamma_{j0}(\gamma_{ij}+1)^2}{n_{ij,t}} \\
&\qquad\qquad + 4\gamma_{ij}(\gamma_{j0}+1)\sqrt{\frac{2\gamma_{j0}x}{n_{j0,t}}} + \frac{22x\gamma_{ij}(\gamma_{j0}+1)^2}{n_{j0,t}}\,.
\end{aligned}$$

Replacing $\gamma_{ij} = \theta_i/\theta_j$ and $\gamma_{j0} = \theta_j$ concludes the proof. □

Lemma 11. *Let $T \geq 1$ and $x > 0$. Then, with probability at least $1 - K(K+1)Te^{-x}$*

$$\tau_{ij,t} < 2x\frac{(\theta_0 + \Theta_{S^*})^2}{\theta_i + \theta_j} \quad or \quad n_{ij,t} \geq \frac{(\theta_i+\theta_j)\tau_{ij,t}}{2(\theta_0+\Theta_{S^*})}\,, \tag{17}$$

where $\tau_{ij,t} := \sum_{s=1}^{t-1}\mathbb{1}\{\{i,j\} \subseteq S_s\}$ simultaneously for all $t \in [T]$ and $i \neq j \in [K]$.

Proof of Theorem 6. Let $\mathcal{E}$ be the high-probabality event of Lemmas 10 and 11 are satisfied, so that $\mathbb{P}(\mathcal{E}) \geq 1 - 4K^2Te^{-x}$. First, note that since we have under the event $\mathcal{E}$, $\widehat{\theta}_t^{\mathtt{ucb}} \leq \theta_t^{\mathtt{ucb}}$, our procedure also satisfies the regret upper-bound

$$Reg_T^{\mathtt{wtd}} \leq O(\sqrt{\theta_{\max}KT}\log T)$$

of Theorem 5. Indeed, all upper-bounds of the proof of Theorem 5 remain valid upper-bounds except the probability of the event $\mathcal{E}^c$ which is $O(T^{-1})$ for $x = 2\log T$.

Let us now prove that we also have $R_T \leq O(K\sqrt{T}\log T)$ with no asymptotic dependence on $\theta_{\max}$ when $T \to \infty$.

Then,

$$
\begin{aligned}
Reg_T^{\texttt{wtd}} &= \sum_{t=1}^{T} \mathbb{E}\big[\mathcal{R}(S^*,\theta) - \mathcal{R}(S_t,\theta)\big] \qquad (18)\\
&= \sum_{t=1}^{T} \mathbb{E}\big[(\mathcal{R}(S^*,\theta) - \mathcal{R}(S_t,\theta))\mathbb{1}\{\mathcal{E}\} + (\mathcal{R}(S^*,\theta) - \mathcal{R}(S_t,\theta))\mathbb{1}\{\mathcal{E}^c\}\big]\\
&\le \sum_{t=1}^{T} \mathbb{E}\Big[\big((\mathcal{R}(S_t,\widehat{\theta}_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)) \wedge \mathcal{R}(S^*,\theta)\big)\mathbb{1}\{\mathcal{E}\} + \mathcal{R}(S^*,\theta)\mathbb{1}\{\mathcal{E}^c\}\Big]\,.
\end{aligned}
$$

Then, using $\mathcal{R}(S^*,\theta) \le \max_i r_i \le 1$, we get

$$
\begin{aligned}
Reg_T^{\texttt{wtd}} &\le \sum_{t=1}^{T} \mathbb{E}\Big[\big((\mathcal{R}(S_t,\widehat{\theta}_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)) \wedge 1\big)\mathbb{1}\{\mathcal{E}\} + \mathbb{1}\{\mathcal{E}^c\}\Big]\\
&\le 4T^2K(K+1)^2e^{-x} + \sum_{t=1}^{T} \mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\widehat{\theta}_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big) \wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big]\,. \qquad (19)
\end{aligned}
$$

Follow the proof of Theorem 5, we upper-bound the second term of the right-hand-side of (19):

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\widehat{\theta}_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big) \wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big] \qquad (20)\\
&\quad= \sum_{t=1}^{T}\mathbb{E}\left[\left(\left(\min_{j\in[K]}\sum_{i\in S_t}\frac{r_i\widehat{\theta}_{i,t}^{\texttt{ucb}}}{1+\sum_{j\in S_t}\widehat{\theta}_{j,t}^{\texttt{ucb}}} - \frac{r_i\theta_i}{1+\sum_{j\in S_t}\theta_j}\right)\wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right]\\
&\quad\le \sum_{t=1}^{T}\mathbb{E}\left[\left(\left(\sum_{i\in S_t}\frac{r_i(\widehat{\theta}_{i,t}^{\texttt{ucb}}-\theta_i)}{\theta_0+\Theta_{S_t}}\right)\wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right] \qquad \text{because } \sum_{i\in S_t}\widehat{\theta}_{i,t}^{\texttt{ucb}} \ge \Theta_{S_t} \text{ under } \mathcal{E}\\
&\quad\le \sum_{t=1}^{T}\mathbb{E}\left[\left(\left(\sum_{i\in S_t}\frac{|\widehat{\theta}_{i,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\right)\wedge 1\right)\mathbb{1}\{\mathcal{E}\}\right] \qquad \text{because } r_i \le 1\\
&\quad\le \sum_{i=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{|\widehat{\theta}_{i,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\wedge 1\right)\mathbb{1}\{i\in S_t\}\mathbb{1}\{\mathcal{E}\}\right]\\
&\quad\le \sum_{i=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{|\gamma_{ij_t,t}^{\texttt{ucb}}\gamma_{j_t0,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\wedge 1\right)\mathbb{1}\{i\in S_t\}\mathbb{1}\{\mathcal{E}\}\right]
\end{aligned}
$$

where $j_t = \operatorname{argmax}_{j\in S_t\cup\{0\}}\theta_j$, where the last inequality is by definition of $\widehat{\theta}_{i,t}^{\texttt{ucb}}$. Now, from Lemma 10, paying an additive exploration cost to ensure that $n_{ij,t} \gtrsim x(1+\gamma_{ij})$ and $n_{j0,t} \gtrsim x(1+\theta_j)^2\theta_j$ for all $j\in S_t$ such that $\theta_j \ge \theta_0$. From Lemma 11, this is satisfied if for some constant $C > 0$

$$
\tau_{ij,t} > Cm^2\theta_{\max}^2 x\,.
$$

Such a condidtion can be wrong for a couple $(i,j)\in S_t^2$ at most during $CK^2m^2\theta_{\max}^2 x = O(\log T)$ rounds (since $\tau_{ij,t}$ increases then). Thus, for C large enough,

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t,\widehat{\theta}_t^{\texttt{ucb}}) - \mathcal{R}(S_t,\theta)\big) \wedge 1\Big)\mathbb{1}\{\mathcal{E}\}\Big]\\
&\qquad\le O(\log T) + \sum_{i=1}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\frac{|\gamma_{ij_t,t}^{\texttt{ucb}}\gamma_{j_t0,t}^{\texttt{ucb}}-\theta_i|}{\theta_0+\Theta_{S_t}}\mathbb{1}\{i\in S_t, \tau_{ij_t,t}\wedge\tau_{j_t,t} \ge Cxm^2\theta_{\max}^2\}\mathbb{1}\{\mathcal{E}\}\right]
\end{aligned}
$$

$$\lesssim O(\log T) + \sum_{i=1}^{K} \mathbb{E}\Bigg[\sum_{t=1}^{T} \Bigg(\sqrt{(\gamma_{ij_t}+1)\theta_i x} \Bigg(\sqrt{\frac{(\theta_i+\theta_{j_t})}{n_{ij_t,t}}} + \sqrt{\frac{(1+\theta_j)}{n_{j_t 0,t}}} \Bigg) + (\gamma_{ij_t}+1)\frac{(\theta_i+\theta_{j_t})x}{n_{ij_t,t}} + \frac{\gamma_{ij_t}(1+\theta_{j_t})^2 x}{n_{j_t 0,t}} \Bigg) \frac{\mathbb{1}\{i \in S_t\}}{\theta_0+\Theta_{S_t}} \Bigg]$$

$$\le O(\log T) + \sum_{i=1}^{K} \mathbb{E}\Bigg[\sum_{t=1}^{T} \sqrt{(\gamma_{ij_t}+1)\theta_i x} \Bigg(\sqrt{\frac{(\theta_i+\theta_{j_t})}{n_{ij_t,t}}} + \sqrt{\frac{(1+\theta_{j_t})}{n_{j_t 0,t}}} \Bigg) \frac{\mathbb{1}\{i \in S_t\}}{\theta_0+\Theta_{S_t}} \Bigg]$$

where the last inequality is because using that $\{i, j_t, 0\} \subseteq S_t$, we have

$$\mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{1+\theta_{j_t}}{(1+\Theta_{S_t}) n_{j_t 0,t}} \Bigg] = \mathbb{E}\Bigg[\sum_{t=1}^{T} \sum_{j=1}^{K} \frac{\mathbb{1}\{i_t \in \{j,0\}\}}{n_{j0,t}} \mathbb{1}\{j = j_t\} \Bigg] \le K(1+\log T).$$

and

$$\mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{\theta_i+\theta_{j_t}}{(1+\Theta_{S_t}) n_{ij_t,t}} \Bigg] = \mathbb{E}\Bigg[\sum_{t=1}^{T} \sum_{j=1}^{K} \frac{\mathbb{1}\{i_t \in \{j,i\}\}}{n_{j0,t}} \mathbb{1}\{j = j_t\} \Bigg] \le K(1+\log T).$$

Then, by Cauchy-Schwarz inequality we further get

$$\sum_{t=1}^{T} \mathbb{E}\Big[\Big(\big(\mathcal{R}(S_t, \widehat{\theta}_t^{\mathsf{ucb}}) - \mathcal{R}(S_t, \theta)\big) \wedge 1 \Big) \mathbb{1}\{\mathcal{E}\} \Big]$$

$$\lesssim O(\log T) + \sum_{i=1}^{K} \sqrt{ \mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{(\gamma_{ij_t}+1)\theta_i \mathbb{1}\{i \in S_t\} x}{\theta_0+\Theta_{S_t}} \Bigg] } \times \sqrt{ \mathbb{E}\Bigg[\sum_{t=1}^{T} \Bigg(\frac{(\theta_i+\theta_{j_t})}{n_{ij_t,t}} + \frac{(1+\theta_{j_t})}{n_{j_t 0,t}} \Bigg) \frac{\mathbb{1}\{i \in S_t\}}{\theta_0+\Theta_{S_t}} \Bigg] } \tag{21}$$

$$\lesssim O(\log T) + \sum_{i=1}^{K} \sqrt{ \mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{(\gamma_{ij_t}+1)\theta_i \mathbb{1}\{i \in S_t\} x}{\theta_0+\Theta_{S_t}} \Bigg] } \sqrt{K \log T}$$

$$\lesssim O(\log T) + \sum_{i=1}^{K} \sqrt{ \mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{\theta_i \mathbb{1}\{i \in S_t\} x}{\theta_0+\Theta_{S_t}} \Bigg] } \sqrt{K \log T} \text{ (because } \gamma_{ij_t} \le 1 \text{ by definition of } j_t)$$

$$\le O(K\sqrt{Tx \log T}) = O(K\sqrt{T} \log T)\,, \tag{22}$$

where the last inequality is by Jensen's inequality and the equality by setting $x = 2\log T$ to control the probability that $\mathcal{E}^c$ occurs. This concludes the proof. $\square$