Preference-based Reinforcement Learning beyond Pairwise Comparisons: Benefits of Multiple Options

Joongkyu Lee
Seoul National University
jklee0717@snu.ac.kr

Seouh-won Yi
Seoul National University
uniqueseouh@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

We study online preference-based reinforcement learning (PbRL) with the goal of improving sample efficiency. While a growing body of theoretical work has emerged—motivated by PbRL's recent empirical success, particularly in aligning large language models (LLMs)—most existing studies focus only on pairwise comparisons. A few recent works [96, 51, 79] have explored using multiple comparisons and ranking feedback, but their performance guarantees fail to improve—and can even deteriorate—as the feedback length increases, despite the richer information available. To address this gap, we adopt the Plackett–Luce (PL) model for ranking feedback over action subsets and propose M-AUPO, an algorithm that selects multiple actions by maximizing the average uncertainty within the offered subset. We prove that M-AUPO achieves a suboptimality gap of $\tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}}\right)$, where T is the total number of rounds, d is the feature dimension, and $|S_t|$ is the size of the subset at round t. This result shows that larger subsets directly lead to improved performance and, notably, the bound avoids the exponential dependence on the unknown parameter's norm, which was a fundamental limitation in most previous works. Moreover, we establish a near-matching lower bound of $\Omega\left(\frac{d}{K\sqrt{T}}\right)$, where K is the maximum subset size. To the best of our knowledge, this is the first theoretical result in PbRL with ranking feedback that explicitly shows improved sample efficiency as a function of the subset size.

1 Introduction

The framework of *Preference-based Reinforcement Learning* (PbRL) [12, 83, 84, 72] was introduced to address the difficulty of designing effective reward functions, which often demands substantial and complex engineering effort [82, 84]. PbRL has been successfully applied in diverse domains, including robot training, stock prediction, recommender systems, and clinical trials [30, 67, 18, 38, 54]. Notably, PbRL also serves as a foundational framework for Reinforcement Learning from Human Feedback (RLHF) when feedback is provided in the form of preferences rather than explicit scalar rewards. This preference-based approach has proven highly effective in aligning Large Language Models (LLMs) with human values and preferences [18, 59, 64].

Given its practical success, the field has also seen significant theoretical advances [16, 49, 72, 96, 89, 94, 93, 86, 74, 53, 13, 66, 22, 19, 51, 77, 73, 79, 88, 14, 39]. However, despite this progress, most existing models remain limited to handling only *pairwise* comparison feedback. A few works [96, 51, 79] explore the more general setting of *multiple* comparisons, offering a strict extension beyond the pairwise case. Zhu et al. [96] study the offline setting, where a dataset of questions (or contexts) along with corresponding ranking feedback over K answers (or actions), labeled by human annotators, is available. Mukherjee et al. [51] investigate the online learning-to-rank problem [63], where a dataset of questions with K candidate answers is provided, but no feedback is initially available.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: Comparisons of settings and theoretical guarantees in related works on PbRL with ranking feedback. Here, T denotes the number of rounds (or the number of data points in the offline setting), K is the (maximum) size of the offered action set (i.e., *assortment*), and d is the feature dimension, and $1/\kappa = \mathcal{O}(e^B)$. ρ represents the unknown context distribution. Here, $\tilde{\mathcal{O}}$ hides logarithmic factors and polynomial dependencies on B. "Sq. Pred. Error" refers to the squared prediction error.

	Setting	Context	Assortment	Measure	Result
Zhu et al. [96]	Offline	Accessible $\mathcal{X}$	Given	Suboptimality	$\tilde{\mathcal{O}}\left(\frac{K^2}{\kappa}\sqrt{\frac{d}{T}}\right)$
Mukherjee et al. [51]	Online	Accessible $\mathcal{X}$	Given	Pred. Error	$\tilde{\mathcal{O}}\left(\frac{K^3 d}{\kappa\sqrt{T}}\right)$
Thekumparampil et al. [79]	Online	No context	Select K	Pred. Error	$\tilde{\mathcal{O}}\left(\frac{K^3 d}{\kappa\sqrt{T}}\right)$
This work (Theorem 1, 2)	Online	Sampled $x \sim \rho$	Select $\leqslant K$	Suboptimality	$\tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{\lvert S_t\rvert}}\right)$
This work (Theorem 3)	Lower Bound	Sampled $x \sim \rho$	Select $\leqslant K$	Suboptimality	$\Omega\left(\frac{d}{K\sqrt{T}}\right)$

Thekumparampil et al. [79] consider a context-free setting (i.e., a singleton context), and the goal is to learn the ranking of $N \geqslant K$ answers based on ranking feedback obtained from subsets of size K. However, all of their theoretical performance guarantees fail to show that using multiple comparisons provides any advantage over the pairwise setting (see Table 1). This is counterintuitive, as ranking feedback is inherently more informative than pairwise feedback. Specifically, since a ranking over K actions provides $\binom{K}{2}$ pairwise comparisons, it should, in principle, enable faster learning and lead to stronger performance guarantees. Thus, the following fundamental question remains open:

Can we design an algorithm that achieves a strictly better theoretical guarantee under multiple-option feedback compared to the pairwise comparisons in the online PbRL setting?

In this paper, we assume that the ranking feedback follows the Plackett-Luce (PL) model [62, 47], where, in each round, the learner receives ranking feedback over a subset of up to K actions (with $K \leqslant N$) selected from a universe of N actions. This problem setup is closely related to that of Thekumparampil et al. [79]; however, unlike their work, which focuses solely on a context-free setting (or equivalently, a fixed singleton context), we study a more general setting where contexts are diverse and drawn from an unknown distribution.

Under this problem setup, we provide an affirmative answer to the above question by introducing a novel algorithm, *Maximizing Average Uncertainty for Preference Optimization* (`M-AUPO`), which explicitly exploits the richer information available from ranking feedback under the Plackett–Luce (PL) model. `M-AUPO` selects action subsets by maximizing *average uncertainty* and achieves a suboptimality gap that strictly improves upon what is attainable with pairwise comparisons. In particular, we show that its suboptimality gap decreases with longer ranking feedback.

Furthermore, our suboptimality gap eliminates the *exponential* dependence on the parameter norm bound, $\mathcal{O}(e^B)$, in the leading term, by employing novel matrix concentration inequalities for the Hessian matrix H_t (see the proof sketch in Section 5.1 for details). This represents a significant improvement over most prior works, where performance guarantees typically depend on $\mathcal{O}(e^B)$ [68, 72, 96, 89, 94, 19, 88, 79, 39]. Very recently, a few works [14, 20] have successfully avoided the $\mathcal{O}(e^B)$ dependency by relying on auxiliary techniques or additional information—such as specialized sampling schemes [14] or prior knowledge of κ [20]—which, however, are often impractical. Moreover, their methods are limited to pairwise comparison settings. In contrast, our approach eliminates the $\mathcal{O}(e^B)$ dependency without using any auxiliary techniques and considers more general ranking feedback beyond pairwise comparisons. Our main contributions are summarized as follows:

- **Improved sample efficiency via larger subsets:** We propose `M-AUPO`, a novel algorithm for online PbRL (or RLHF) with PL ranking feedback, which achieves a suboptimality gap of $\tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}}\right)$, where $|S_t|$ is the size of the action subset offered at round t. This result provides the first rigorous theoretical guarantee that larger subsets directly improve sample efficiency. To the best of our knowledge, this is the first theoretical work in PbRL that explicitly demonstrates performance improvements as a function of the subset size $|S_t|$.

- **Free of $\mathcal{O}(e^B)$ dependency:** Our result eliminates the exponential dependence on the parameter norm bound, $\mathcal{O}(e^B)$, in the leading term, without relying on any auxiliary techniques. This demonstrates that the $\mathcal{O}(e^B)$ dependence commonly observed in PbRL (or RLHF) and dueling bandit analyses is not fundamentally necessary, but rather an artifact of loose analysis. We believe that our key lemmas (Lemmas 1 and E.1) can be directly applied to existing PbRL or dueling bandit analyses—including regret-minimization settings—whenever elliptical potential lemmas are used, without requiring any modification to the original algorithm. To the best of our knowledge, this is the first PbRL work with ranking feedback involving more than two options that avoids $\mathcal{O}(e^B)$ dependence.
- **Lower bound:** We establish a near-matching lower bound of $\Omega\left(\frac{d}{K\sqrt{T}}\right)$ under the Plackett–Luce (PL) model with ranking feedback, matching our upper bound up to a factor of K. This result demonstrates that incorporating richer ranking information (i.e., larger K) provably enhances sample efficiency.
- **Experiment:** We empirically evaluate `M-AUPO` on both synthetic and real-world datasets, showing its improved performance for larger K and its superiority over existing baselines.

2 Related Works

Fueled by the remarkable success of LLMs [18, 59, 64], the theoretical study of PbRL has rapidly emerged as a central focus within the research community. Early work in this area traces back to the dueling bandits literature [91, 98, 70, 8].

Dueling bandits. The dueling bandit framework, introduced by Yue et al. [91], departs from the classical multi-armed bandit setting by requiring the learner to select two arms and observe only their pairwise preference. For general preferences, a single best arm that is globally dominant may not exist. To address this, various alternative winners have been proposed, including the Condorcet winner [97, 36], Copeland winner [98, 85, 37], Borda winner [31, 25, 28, 71, 87], and von Neumann winner [65, 24, 7], each with its own corresponding performance metric.

To address scalability and contextual information, Saha [68] proposed a structured contextual dueling bandit setting in which preferences are modeled using a Bradley–Terry–Luce (BTL) model [11] based on the unknown intrinsic rewards of each arm. In a similar setting, Bengs et al. [9] studied a contextual linear stochastic transitivity model, and Di et al. [21] proposed a layered algorithm that achieves variance-aware regret bounds. However, most prior dueling bandit works suffer from an exponential dependency of $\mathcal{O}(e^B)$. In recent work, only a few studies [20, 14] have succeeded in eliminating the $\mathcal{O}(e^B)$ dependency by incorporating additional complex subroutines.

Preference-based reinforcement learning (PbRL). Building upon this line of work, subsequent research has extended the dueling bandit framework to the RL framework, considering both online [90, 54, 16, 72, 86] and offline settings [96, 94, 46]. More recently, under the active learning framework—where the full set of contexts $\mathcal{X}$ is accessible—many studies aim to improve sample efficiency by selecting prompts either based on the differences in estimated rewards for their responses [52] or through D-optimal design methods [49, 73, 19, 51, 79, 39]. However, most of these works focus exclusively on pairwise preference feedback and cannot be extended to more general ranking feedback cases. Mukherjee et al. [51] study the online learning-to-rank problem when prompts are given along with K candidate answers, while Thekumparampil et al. [79] investigate learning to rank $N \geqslant K$ answers from partial rankings over K answers, but under a context-free setting. In this paper, we consider a stochastic contextual setting (more general than Thekumparampil et al. [79]), where contexts are sampled from an unknown but fixed distribution, and aim to minimize the suboptimality gap using ranking feedback of up to length K.

For further related work, see Appendix A.

3 Problem Setting and Preliminaries

Notations. Given a set $\mathcal{X}$, we use $|\mathcal{X}|$ to denote its cardinality. For a positive integer n, we denote $[n] := \{1, 2, \dots, n\}$. For a real-valued matrix A, we let $\|A\|_2 := \sup_{x:\|x\|_2=1} \|Ax\|_2$ which is the maximum singular value of A. We write $A \geq A'$ if $A - A'$ is positive semidefinite. For a univariate function f, we denote $\dot{f}$ as its derivative.

We have a set of contexts (or prompts), denoted by $\mathcal{X}$, and a set of possible actions (or answers), denoted by $\mathcal{A} := \{a_1, \ldots, a_N\}$.[1]. We consider preference feedback in the form of partial rankings over subsets of $\mathcal{A}$, and model this feedback using the Plackett-Luce (PL) distribution:

Definition 1 (PL model)**.** *Let $\mathcal{S} := \{S \subseteq \mathcal{A} \mid 2 \leqslant |S| \leqslant K\}$ be the collection of all action subsets whose sizes range from 2 to K. For any $S \in \mathcal{S}$, let σ denote the labeler's ranking feedback—that is, a permutation of the elements in S. We write σ_j for the j-th most preferred action under σ. We model the distribution of such rankings using the Plackett-Luce (PL) model [62, 47], defined as:*

$$\mathbb{P}(\sigma|x, S; \boldsymbol{\theta}^\star) = \prod_{j=1}^{|S|} \frac{\exp\left(r_{\boldsymbol{\theta}^\star}(x, \sigma_j)\right)}{\sum_{k=j}^{|S|} \exp\left(r_{\boldsymbol{\theta}^\star}(x, \sigma_k)\right)}, \quad \textit{where } (x, S) \in \mathcal{X} \times \mathcal{S}. \tag{1}$$

Here, $r_{\boldsymbol{\theta}^\star}$ represents a reward model parameterized by the unknown parameter $\theta^\star$.

When $K = 2$, this reduces to the pairwise comparison framework considered in the Bradley-Terry-Luce (BTL) model [11]. The probability that a is preferred to a' given x can be expressed as:

$$\mathbb{P}(a \succ a'|x; \boldsymbol{\theta}^\star) = \frac{\exp\left(r_{\boldsymbol{\theta}^\star}(x, a)\right)}{\exp\left(r_{\boldsymbol{\theta}^\star}(x, a)\right) + \exp\left(r_{\boldsymbol{\theta}^\star}(x, a')\right)} = \mu\left(r_{\boldsymbol{\theta}^\star}(x, a) - r_{\boldsymbol{\theta}^\star}(x, a')\right), \tag{2}$$

where $\mu(w) = \frac{1}{1+e^{-w}}$ is the sigmoid function. In this work, we assume a linear reward model:

Assumption 1. *Let $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature map satisfying $\max_{x,a} \|\phi(x, a)\|_2 \leqslant 1$, and let $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ denote the true but unknown parameter. The reward is assumed to follow a linear structure given by $r_{\boldsymbol{\theta}^\star}(x, a) = \phi(x, a)^\top \boldsymbol{\theta}^\star$. To ensure identifiability of $\boldsymbol{\theta}^\star$, we assume that $\boldsymbol{\theta}^\star \in \Theta$, where $\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 \leqslant B\}$. Without loss of generality, we assume $B \geqslant 1$.*

At each round $t \in [T]$, a context $x_t \in \mathcal{X}$ is drawn from a fixed but unknown distribution ρ. Given the context x_t, the learning agent selects a subset of actions $S_t \in \mathcal{S}$—referred to as an *assortment* throughout the paper—and receives a ranking over S_t as feedback, generated according to the PL model. After T rounds of interaction with the labeler, the goal is to output a policy $\widehat{\pi}_T : \mathcal{X} \to \mathcal{A}$ that minimizes the *suboptimality gap*, defined as:

$$\textbf{SubOpt}(T) := \mathbb{E}_{x \sim \rho}\left[r_{\boldsymbol{\theta}^\star}\left(x, \pi^\star(x)\right) - r_{\boldsymbol{\theta}^\star}\left(x, \widehat{\pi}_T(x)\right)\right],$$

where $\pi^\star(x) = \operatorname{argmax}_a r_{\boldsymbol{\theta}^\star}(x, a)$ is the optimal policy under the true reward $r_{\boldsymbol{\theta}^\star}$.

3.1 Loss Functions and Rank Breaking

In this paper, we consider two different losses for estimating the parameter: one directly induced by the PL model, and the other obtained by splitting the ranking feedback into pairwise comparisons.

Plackett-Luce (PL) loss. The PL loss function for round t is defined as follows:

$$\ell_t(\boldsymbol{\theta}) := \sum_{j=1}^{|S_t|} \ell_t^{(j)}(\boldsymbol{\theta}), \quad \text{where } \ell_t^{(j)}(\boldsymbol{\theta}) := -\log\left(\frac{\exp\left(\phi(x_t, \sigma_{tj})^\top \boldsymbol{\theta}\right)}{\sum_{k=j}^{|S_t|} \exp\left(\phi(x_t, \sigma_{tk})^\top \boldsymbol{\theta}\right)}\right). \tag{3}$$

Here, $\ell_t^{(j)}(\boldsymbol{\theta})$ denotes the negative log-likelihood loss under the Multinomial Logit (MNL) model [48], conditioned on the assortment being the remaining actions in S_t after removing the previously selected actions $\sigma_{t1}, \ldots, \sigma_{t(j-1)}$—that is, over the set $S_t \backslash \{\sigma_{t1}, \ldots, \sigma_{t(j-1)}\}$.

Rank-Breaking (RB) loss. In addition to this standard approach, one can replace the full $|S_t|$-action ranking with its $\binom{|S_t|}{2}$ pairwise comparisons. This technique, referred to as *rank breaking* (RB), decomposes (partial) ranking data into individual pairwise comparisons, treating each comparison as independent [6, 34, 32, 69]. Thus, the RB loss is defined as:

$$\ell_t(\boldsymbol{\theta}) := \sum_{j=1}^{|S_t|-1} \sum_{k=j+1}^{|S_t|} \ell_t^{(j,k)}(\boldsymbol{\theta}), \quad \text{where } \ell_t^{(j,k)}(\boldsymbol{\theta}) := -\log\left(\frac{\exp\left(\phi(x_t, \sigma_{tj})^\top \boldsymbol{\theta}\right)}{\sum_{m \in \{j,k\}} \exp\left(\phi(x_t, \sigma_{tm})^\top \boldsymbol{\theta}\right)}\right). \tag{4}$$

This approach is applied in the current RLHF for LLM (e.g., Ouyang et al. [59]) and is also studied in the theoretical RLHF paper [96] under the offline setting.

[1]For simplicity, we assume a stationary action space $\mathcal{A}$, though it may depend on the context $x \in \mathcal{X}$.

Procedure 1 `OMD-PL`, OMD for PL Loss

Input: $\widehat{\boldsymbol{\theta}}_t^{(1)}, S_t, H_t$
for $j = 1$ to $|S_t|$ **do**
  Update $\tilde{H}_t^{(j)}$, $\widehat{\boldsymbol{\theta}}_t^{(j+1)}$ via (5)
end for
return $\widehat{\boldsymbol{\theta}}_t^{(|S_t|+1)}$

Procedure 2 `OMD-RB`, OMD for RB Loss

Input: $\widehat{\boldsymbol{\theta}}_t^{(1,2)}, S_t, H_t$
for $j = 1$ to $|S_t| - 1$ **do**
  for $k = 2$ to $|S_t|$ **do**
    Update $\tilde{H}_t^{(j,k)}$, $\widehat{\boldsymbol{\theta}}_t^{(j,k+1)}$ via (7)
  end for
end for
return $\widehat{\boldsymbol{\theta}}_t^{(|S_t|-1,|S_t|+1)}$

3.2 Online Parameter Estimation

Motivated by recent advances in Multinomial Logit (MNL) bandits [95, 41, 43], we adopt an online mirror descent (OMD) algorithm to estimate the underlying parameter $\boldsymbol{\theta}^\star$, instead of relying on maximum likelihood estimation (MLE). This enables a constant per-round computational cost, in contrast to the MLE-based approach, whose cost grows linearly with the number of rounds t.

OMD update for PL loss. For the the PL loss (3), we estimate the true parameter $\boldsymbol{\theta}^\star$ as follows:

$$\widehat{\boldsymbol{\theta}}_t^{(j+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \langle \nabla \ell_t^{(j)}(\widehat{\boldsymbol{\theta}}_t^{(j)}), \boldsymbol{\theta} \rangle + \frac{1}{2\eta} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^{(j)}\|_{\tilde{H}_t^{(j)}}^2, \quad j = 1, \dots, |S_t|, \tag{5}$$

where we write $\widehat{\boldsymbol{\theta}}_t^{(|S_t|+1)} = \widehat{\boldsymbol{\theta}}_{t+1}^{(1)}$, and η is the step-size parameter to be specified later. The matrix $\tilde{H}_t^{(j)}$ is given by $\tilde{H}_t^{(j)} := H_t + \eta \sum_{j'=1}^{j} \nabla^2 \ell_t^{(j')}(\widehat{\boldsymbol{\theta}}_t^{(j')})$, where

$$H_t := \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d, \quad \lambda > 0. \tag{6}$$

The optimization problem (5) can be solved using a single projected gradient step [57], which enjoys a computational cost of only $\mathcal{O}(Kd^3)$—independent of t [50], unlike MLE—and requires only $\mathcal{O}(d^2)$ storage, thanks to the incremental updates of $\tilde{H}_t^{(j)}$ and H_t.

OMD update for RB loss. Similarly, for the RB loss (4), we estimate the underlying parameter as:

$$\widehat{\boldsymbol{\theta}}_t^{(j,k+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \langle \nabla \ell_t^{(j,k)}(\widehat{\boldsymbol{\theta}}_t^{(j,k)}), \boldsymbol{\theta} \rangle + \frac{1}{2\eta} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^{(j,k)}\|_{\tilde{H}_t^{(j,k)}}^2, \quad 1 \leqslant j < k \leqslant |S_t|, \tag{7}$$

where we set $\widehat{\boldsymbol{\theta}}_t^{(j,|S_t|+1)} = \widehat{\boldsymbol{\theta}}_t^{(j+1,j+2)}$ for all $j < |S_t| - 1$ and for the final pair, let $\widehat{\boldsymbol{\theta}}_t^{(|S_t|-1,|S_t|+1)} = \widehat{\boldsymbol{\theta}}_{t+1}^{(1,2)}$. Also, the matrix $\tilde{H}_t^{(j,k)}$ is defined as $\tilde{H}_t^{(j,k)} := H_t + \eta \sum_{(j',k') \leqslant (j,k)} \nabla^2 \ell_t^{(j',k')}(\widehat{\boldsymbol{\theta}}_t^{(j',k')})$ [2], where

$$H_t := \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d, \quad \lambda > 0. \tag{8}$$

Remark 1 (Computational cost of OMD)**.** *The per-round computational cost of the PL parameter update is $\mathcal{O}(K^2 d^3)$, since the parameter is updated $|S_t| \leqslant K$ times per round. Similarly, the cost for the RB parameter update is $\mathcal{O}(K^3 d^3)$, as the parameter is updated $\binom{|S_t|}{2}$ times per round.*

4 `M-AUPO`: Maximizing Average Uncertainty

In this section, we propose a new algorithm, `M-AUPO`, designed to select an assortment that maximizes *average uncertainty* of S_t, thereby leveraging the potential benefits of a large K. At each round t, a context x_t is drawn from a fixed but unknown distribution ρ. The algorithm then selects a *reference*

[2] We write $(j', k') \leqslant (j, k)$ to indicate lexicographic order, i.e., $j' < j$ or $j' = j$ and $k' \leqslant k$.

Algorithm 3 M-AUPO: **M**aximizing **A**verage **U**ncertainty for **P**reference **O**ptimization

```
1: Inputs: maximum assortment size K, regularization parameter λ, step size η
2: Initialize: H_1 = λI_d, θ̂_1 ∈ Θ
3: for round t = 1 to T do
4:     Observe x_t and select (ā_t, S_t) via (9)
5:     Observe ranking feedback σ_t for S_t
6:     θ̂_{t+1} ← OMD-PL(θ̂_t, S_t, H_t) (Proc. 1)        ▷ or OMD-RB(θ̂_t, S_t, H_t) (Proc. 2) if RB loss
7:     Update H_{t+1} ← H_t + Σ_{j=1}^{|S_t|} ∇²ℓ_t^{(j)}(θ̂_t^{(j+1)}) via (6)        ▷ or via (8) if RB loss
8: end for
9: Return: π̂_T(x) ← argmax_{a∈A} φ(x, a)^⊤ θ̂_{T+1}
```

action-assortment pair $(\bar{a}_t, S_t)$ by maximizing the average feature uncertainty—measured in the H_t^{-1}-norm—relative to a candidate reference action $\bar{a}$ (Line 4):

$$(\bar{a}_t, S_t) = \operatorname*{argmax}_{\bar{a} \in \mathcal{A}} \operatorname*{argmax}_{\substack{S \in \mathcal{S} \\ \bar{a} \in S}} \frac{1}{|S|} \sum_{a \in S} \|\phi(x_t, a) - \phi(x_t, \bar{a})\|_{H_t^{-1}} . \tag{9}$$

By construction, the reference action $\bar{a}_t$ is always included in the selected assortment S_t. This selection strategy plays a key role in our algorithm, as it promotes rapid reduction in reward uncertainty—particularly when the assortment size $|S_t|$ is large—by encouraging informative comparisons centered around the reference action. Importantly, the assortment selection rule in Equation (9) can be computed efficiently, without enumerating all $\binom{N}{K}$ possible subsets.

Remark 2 (Computational cost of S_t-selection)**.** *The optimization in Equation* (9) *can be efficiently solved with a computational cost of $\tilde{\mathcal{O}}(N^2 K)$ (see Appendix B for details). Furthermore, in Appendix H.1, we will show that the reference action $\bar{a}_t$ can be chosen arbitrarily, which further reduces the computational cost to $\tilde{\mathcal{O}}(NK)$.*

Then, we observe the ranking feedback σ_t from a labeler and update the parameter according to Procedure 1 if using the PL loss, or Procedure 2 if using the RB loss (Line 6). After T rounds, the algorithm returns the final policy $\widehat{\pi}_T$, which selects actions by maximizing the estimated reward under the final parameter estimate, i.e., $\widehat{\pi}_T(x) := \operatorname{argmax}_a \phi(x, a)^\top \widehat{\boldsymbol{\theta}}_{T+1}$ (Line 7).

5 Main Results

In this section, we present our main theoretical contributions. In Section 5.1, we show that M-AUPO achieves a suboptimality gap that decreases with the size of the presented assortment $|S_t|$, implying improved performance when larger action subsets are offered for ranking feedback. In Section 5.2, we establish the near-matching lower bound.

5.1 Suboptimality Gap of M-AUPO

We begin by presenting the online confidence bound for the PL loss, derived by extending the results of Lee and Oh [43], who analyzed the MNL model [48]. Since the PL model constructs ranking probabilities as a product of MNL probabilities, their confidence bound can be directly applied to our setting by replacing the round t with the cumulative number of updates $\sum_{s=1}^{t} |S_s|$.

Corollary 1 (Online confidence bound for PL loss)**.** *Let $\delta \in (0, 1]$. We set $\eta = (1 + 3\sqrt{2}B)/2$ and $\lambda = \max\{12\sqrt{2}B\eta, 144\eta d, 2\}$. Then, under Assumption 1, with probability at least $1 - \delta$, we have*

$$\|\widehat{\boldsymbol{\theta}}_t^{(j)} - \boldsymbol{\theta}^\star\|_{H_t^{(j)}} \leqslant \beta_t(\delta) = \mathcal{O}\left(B\sqrt{d \log(tK/\delta)} + B\sqrt{\lambda}\right), \quad \forall t \geqslant 1, j \leqslant |S_t|,$$

where $H_t^{(j)} := H_t + \sum_{j'=1}^{j-1} \nabla^2 \ell_s^{(j')}(\widehat{\boldsymbol{\theta}}_s^{(j'+1)}) + \lambda \mathbf{I}_d$.

This confidence bound is free of any polynomial dependency on K, which is primarily made possible by the improved self-concordant-like properties proposed by Lee and Oh [43]. Moreover, for the RB loss, we can derive a confidence bound of the same order (see Corollary E.1). Based on this confidence bound, we derive the suboptimality gap for M-AUPO, with the proof deferred to Appendix D.

Theorem 1. *Let $\delta \in (0,1]$. We set $\lambda = \Omega\big(d\log(KT/\delta) + \eta(B+d)\big)$ and $\eta = \frac{1}{2}(1+3\sqrt{2}B)$. Define $\kappa := e^{-4B}$. If Assumption 1 holds, then, with probability at least $1-\delta$,* `M-AUPO` *(Algorithm 3) achieves the following suboptimality gap:*

$$\textbf{SubOpt}(T) = \tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}} + \frac{d^2K^2}{\kappa T}\right).$$

Discussion of Theorem 1. For sufficiently large T, the second (non-leading) term becomes negligible, and Theorem 1 shows that the suboptimality gap of `M-AUPO` decreases as the assortment size $|S_t|$ increases. This establishes a strict advantage of receiving ranking feedback over larger assortments. Moreover, our result does not involve any $\mathcal{O}(e^B)$ dependency in the leading term, a harmful dependency that commonly appears in prior works [68, 72, 96, 89, 94, 19, 79, 39]. Although very recent studies [20, 14] also achieve $\mathcal{O}(e^B)$-free performance in the leading term, they rely on auxiliary techniques and are restricted to pairwise preference feedback. To the best of our knowledge, this is the first theoretical study that simultaneously establishes (i) the performance benefits of utilizing richer ranking feedback over larger assortments, and (ii) the elimination of the $\mathcal{O}(e^B)$ dependence in the leading term of the PbRL framework when accommodating multiple (i.e., more than two) options.

Proof sketch of Theorem 1. We provide a proof sketch of Theorem 1. For simplicity, the main paper assumes that the term $\left\|\phi(x_t,a) - \phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}$ remains sufficiently small for all a during all rounds. This is justified because the regret incurred in the rounds where this condition fails is bounded by lower-order terms and thus has a negligible impact (see Lemma D.4 and D.5).

1) Regret decomposition and assortment selection. The proof begins by decomposing the suboptimality gap into two components: the realized regrets and a martingale difference sequence (MDS). Since the MDS term can be readily bounded using the Azuma–Hoeffding inequality, the analysis focuses on bounding the realized regrets.

$$\begin{aligned}\textbf{SubOpt}(T) &= \frac{1}{T}\sum_{t=1}^{T}\underbrace{\big(\phi(x_t,\pi^\star(x_t)) - \phi(x_t,\widehat{\pi}_T(x_t))\big)^\top\boldsymbol{\theta}^\star}_{\text{realized regret of } \widehat{\pi}_T \text{ at round } t} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\mathrm{MDS}_t}_{=\tilde{\mathcal{O}}(1/\sqrt{T})} \\ &\lesssim \frac{1}{T}\sum_{t=1}^{T}\left\|\phi(x_t,\pi^\star(x_t)) - \phi(x_t,\widehat{\pi}_T(x_t))\right\|_{H_t^{-1}}\underbrace{\left\|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_T}}_{=\tilde{\mathcal{O}}(\sqrt{d})} + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right).\end{aligned}$$

In the inequality, we first use the fact that $\phi(x_t,\widehat{\pi}_T(x_t))^\top\big(\widehat{\boldsymbol{\theta}}_{T+1} - \boldsymbol{\theta}^\star\big) \geqslant 0$, which follows from definition of $\widehat{\pi}_T$, and then apply Hölder's inequality together with the inequality $H_{T+1} \geq H_t$. We now apply Corollary 1 to upper bound $\left\|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_T}$ by $\tilde{\mathcal{O}}(\sqrt{d})$. Next, using our assortment selection rule (9), we can bound $\left\|\phi(x_t,\pi^\star(x_t)) - \phi(x_t,\widehat{\pi}_T(x_t))\right\|_{H_t^{-1}}$ as follows:

$$\frac{1}{2}\left\|\phi(x_t,\pi^\star(x_t)) - \phi(x_t,\widehat{\pi}_T(x_t))\right\|_{H_t^{-1}} \leqslant \frac{1}{|S_t|}\sum_{a\in S_t}\left\|\phi(x_t,a) - \phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}. \tag{10}$$

Hence, the performance is expected to improve as the subset size $|S_t|$ increases.

2) Avoiding $\mathcal{O}(e^B)$ by matrix concentration. To further bound the right-hand side of Equation (10), we first express H_t as follows:

$$H_t = \frac{1}{2}\sum_{s=1}^{t-1}\sum_{j=1}^{|S_s|}\mathbb{E}_{(a,a')\sim P_s^{(j)}\times P_s^{(j)}}\left[\big(\phi(x_s,a) - \phi(x_s,a')\big)\big(\phi(x_s,a) - \phi(x_s,a')\big)^\top\right] + \lambda\mathbf{I}_d,$$

where $P_s^{(j)}$ denote the (true) MNL distribution over the remaining actions in S_s after removing the first $j-1$ selected actions, i.e., $P_s^{(j)}(a) := \frac{\exp\left(\phi(x_s,a)^\top\boldsymbol{\theta}^\star\right)}{\sum_{a'\in S_s^{(j)}}\exp(\phi(x_s,a')^\top\boldsymbol{\theta}^\star)}$, where $a \in S_s^{(j)} := \{\sigma_{sj},\ldots,\sigma_{s|S_s|}\}$.

Furthermore, we define the regularized sample covariance matrix of feature differences, Λ_t, which, unlike H_t, does not incorporate local information:

$$\Lambda_t := \sum_{s=1}^{t-1} \sum_{a \in S_s} \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top + \lambda \mathbf{I}_d.$$

Aside from the very recent works [20, 14], which avoid the $\mathcal{O}(e^B)$ dependency but only in pairwise comparison settings, most previous works on linear contextual dueling bandits and PbRL [68, 72, 96, 89, 94, 19, 79, 39] exhibit performance (either in terms of cumulative regret or suboptimality gap) that depends on $\mathcal{O}(Ke^B)$ (or $\mathcal{O}(e^B)$ in the case of pairwise comparisons). This dependency arises because these works apply a crude lower bound on H_t by using the inequality $P_s^{(j)}(a)P_s^{(j)}(a') \gtrsim \frac{1}{K^2 e^{2B}}$. As a result, they derive $H_t \gtrsim \frac{1}{K^2 e^{2B}} \Lambda_t$, which further implies $\left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{H_t^{-1}} \lesssim Ke^B \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{\Lambda_t^{-1}}$. This leads directly to performance bounds that scale with $\mathcal{O}(Ke^B)$.

To tackle this problem, we leverage the concentration lemma for covariance matrices (Corollary F.1) and for PSD matrices (Lemma F.4). The following lemma shows that, even without introducing additional algorithmic complexities specifically designed to avoid the $\mathcal{O}(e^B)$ dependency—as done in Di et al. [20] and Chen et al. [14]—the standard analysis techniques are sufficient to eliminate the $\mathcal{O}(Ke^B)$ dependency. In particular, we establish that H_t approximates Λ_t up to a constant factor.

Lemma 1. *Let $\lambda = \Omega(d \log(KT/\delta))$. Then, with probability at least $1 - \delta$, we have*

$$H_t \succeq \frac{1}{50} \Lambda_t.$$

Remark 3 (Applicability of Lemma 1)**.** *Lemma 1 is expected to readily apply to most existing PbRL (or RLHF) and dueling bandit algorithms without requiring any modification to their original formulations, thereby eliminating the $\mathcal{O}(e^B)$ dependency in the leading term.*

By applying Lemma 1 and Cauchy-Schwartz inequality, we obtain:

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{|S_t|} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{H_t^{-1}} \lesssim \frac{1}{T} \sqrt{\sum_{t=1}^{T} \frac{|S_t|}{|S_t|^2}} \underbrace{\sqrt{\sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{\Lambda_t^{-1}}^2}}_{= \tilde{\mathcal{O}}(\sqrt{d})}.$$

Finally, applying the elliptical potential lemma (Lemma D.3), we concludes the proof.

Furthermore, we establish a similar suboptimality gap when using the RB loss (4) in place of the PL loss (3). The proof is provided in Appendix E.

Theorem 2. *Under the same setting as Theorem 1, let $\kappa := \frac{e^{-4B}}{4}$. Then, with probability at least $1 - \delta$, `M-AUPO` (Algorithm 3) achieves the following suboptimality gap:*

$$\textbf{SubOpt}(T) = \tilde{\mathcal{O}}\left(\frac{d}{T} \sqrt{\sum_{t=1}^{T} \frac{1}{|S_t|}} + \frac{d^2}{\kappa T} \right).$$

Discussion of Theorem 2. For sufficiently large T, the suboptimality gap in Theorem 2 matches the leading-order term of Theorem 1, while its second (non-leading) term is tighter by a factor of $\mathcal{O}(K^2)$. However, the per-round computational cost of the RB parameter update is K times higher than that of the PL parameter update (see Remark 1). Despite this, the result is particularly notable as it offers a rigorous theoretical explanation for the empirical success of RLHF in LLMs (e.g., Ouyang et al. [59]), where ranking feedback is decomposed into pairwise comparisons for parameter estimation.

5.2 Lower Bound

In this subsection, we derive a lower bound for our setting: PbRL with linear rewards under ranking feedback generated by a Plackett-Luce (PL) model. The proof is deferred to Appendix G.

Theorem 3 (Lower bound)**.** *Suppose $T \geqslant d^2/(8K^2)$. Define the feature space as $\Phi := \mathcal{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$, and let the parameter space be $\Theta = \{-\mu, \mu\}^d$, where $\mu = \sqrt{d/(8K^2T)}$. Then, for any policy $\widehat{\pi}_T \in \triangle_\Phi$ returned after collecting T samples (using any sampling policy), the expected suboptimality gap is lower bounded as:*

$$\textbf{SubOpt}(T) = \Omega\left(\frac{d}{K\sqrt{T}}\right).$$

Discussion of Theorem 3. Theorem 3 provides theoretical support for our upper bounds, particularly with respect to the dependency on K. Compared to the upper bounds in Theorems 1 and 2, the remaining gap is only a factor of $\frac{1}{\sqrt{K}}$. Closing this gap remains an open problem for future work. To the best of our knowledge, this is the first lower bound on the suboptimality gap that incorporates PL ranking feedback in PbRL and formally shows that the suboptimality gap can diminish as K grows, highlighting the advantage of utilizing ranking feedback over simple pairwise comparisons.

6 Numerical Experiments

We conduct two sets of experiments to empirically validate our theoretical findings: (i) one using synthetic data (Subsection 6.1), and (ii) another using two real-world datasets (Subsection 6.2). We compare our proposed algorithm, `M-AUPO`, against three baselines: (i) `DopeWolfe` [79], which selects K actions in a non-contextual setting; (ii) `Uniform`, which uniformly samples assortments of size K at random; and (iii) `Best&Ref` constructs an action pair ($|S_t| = 2$) by combining the action that maximizes the current reward estimate with another sampled from a reference policy (e.g., uniform random or SFT), following the setup in Online GSHF [89] and XPO [88]. In our experiments, the reference policy for `Best&Ref` is set to the uniform random policy.

6.1 Synthetic Data

In the synthetic data experiment, for each instance, we sample the underlying parameter $\boldsymbol{\theta}^\star \sim \mathcal{N}(0, I_d)$ and normalize it to ensure that $\|\boldsymbol{\theta}^\star\|_2 \leqslant 1$. At every round t, a context $x \in \mathcal{X}$ is drawn uniformly at random, and its feature vector $\phi(x, \cdot)$ lies within the unit ball. We set $d = 5$, $|\mathcal{A}| = N = 100$, and $|\mathcal{X}| = 100$. We measure the suboptimality gap every 25 rounds and report the mean over 20 independent runs, together with one standard error.

The first two plots in Figure 1 show the suboptimality gap of `M-AUPO` under both the PL loss (3) and RB loss (4) as the maximum assortment size K varies. The results clearly show that performance improves as K increases, supporting our theoretical findings. In the third plot of Figure 1, we compare the performance of `M-AUPO` with other three baselines under the PL loss with $K = 5$ at the final round, demonstrating that our algorithm outperforms other baselines significantly. While `DopeWolfe` also considers the selection of K actions from N actions, it treats each context x independently and is specifically designed for the context-free setting (i.e., a singleton context). As a result, `DopeWolfe` cannot leverage information sharing across varying contexts and performs poorly in our setting. Furthermore, `M-AUPO` outperforms naive assortment selection strategies such as `Uniform` and `Best&Ref`, as it explicitly chooses assortments that maximize the expected uncertainty, thereby achieving more efficient exploration. See Appendix I.1 for additional experimental details and results.

6.2 Real-World Dataset

We also conduct experiments using real-world datasets from TREC Deep Learning (TREC-DL)[3] and NECTAR[4]. The TREC-DL dataset provides 100 candidate answers for each question, while the NECTAR dataset offers 7 candidate answers per question. We sample $|\mathcal{X}| = 5000$ prompts from each dataset, with the corresponding set of actions (100 or 7 actions, respectively).

We use the gemma-2b[5] [78] LLM to construct the feature $\phi(x, a)$. Specifically, $\phi(x, a)$ is obtained by extracting the embedding of the concatenated prompt and response from the last hidden layer of

[3]https://microsoft.github.io/msmarco/TREC-Deep-Learning

[4]https://huggingface.co/datasets/berkeley-nest/Nectar

[5]https://huggingface.co/google/gemma-2b-it

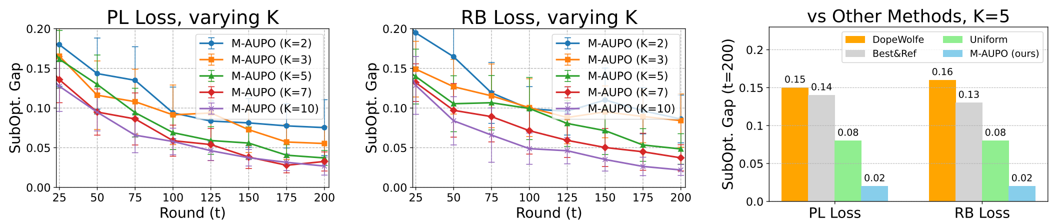

Figure 1: Synthetic data experiment: suboptimality gap of `M-AUPO` under varying K, evaluated with PL loss (left) and RB loss (middle), along with comparison against `DopeWolfe` (right).

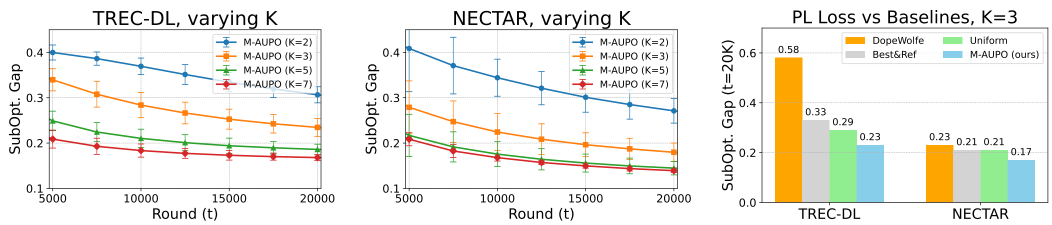

Figure 2: Real-world dataset experiment: suboptimality gap of `M-AUPO` under varying K on the TREC-DL dataset (left) and the NECTAR dataset (middle), along with comparison against `DopeWolfe` (right). The results are rescaled to align the performances between the two datasets.

the LLM, with size $d = 2048$. Additionally, we use the Mistral-7B [33] reward model[6] as the true reward model $r_{\boldsymbol{\theta}^\star}$ to generate ranking feedback and compute the suboptimality gap accordingly. We measure the suboptimality gap every 2,500 rounds and report the average over 10 independent runs, along with the standard error. In these experiments, we present only the results under the PL loss, as the performance difference between the PL and RB losses is negligible, as shown in Figure 1.

The first two plots in Figure 2 show the suboptimality gap of `M-AUPO` under the PL loss on two real-world datasets as the maximum assortment size K varies. Consistent with our theoretical findings, the performance improves as K increases. In the third plot of Figure 2, we compare the performance of `M-AUPO` with other baselines under the PL loss with $K = 3$ at the final round, showing that `M-AUPO` outperforms baselines by a large margin, consistent with the results from the synthetic data experiment. See Appendix I.2 for additional experimental details and results.

7 Conclusion

To the best of our knowledge, this work presents the first theoretical result in online PbRL showing that the suboptimality gap decreases as more options are revealed to the labeler for ranking feedback. By demonstrating its statistical efficiency, our results provide a solid theoretical foundation for moving beyond the prevalent reliance on pairwise comparisons. We hope this finding will encourage future research to explore richer feedback formats beyond pairwise comparisons.

Moreover, our analysis eliminates the $\mathcal{O}(e^B)$ dependency in the leading term without introducing any additional algorithm. This result implies that all existing PbRL and dueling bandit algorithms can likewise avoid this harmful dependency without modification—indicating that the limitation lies in their analyses rather than in the optimality of the algorithms themselves. The key takeaway is that in PbRL, the $\mathcal{O}(e^B)$ dependency is theoretically avoidable and thus no longer poses a limitation.

We believe that these two implications are both conceptually significant and provide meaningful contributions toward a deeper theoretical understanding of PbRL.

[6] https://huggingface.co/Ray2333/reward-model-Mistral-7B-instruct-Unified-Feedback

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023-00222663, and RS-2025-25420849), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263754), and by AI-Bio Research Grant through Seoul National University.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

[3] Priyank Agrawal, Theja Tulabandhula, and Vashist Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 310(2):737–750, 2023.

[4] 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.

[5] 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.

[6] Hossein Azari Soufiani, William Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. *Advances in Neural Information Processing Systems*, 26, 2013.

[7] Akshay Balsubramani, Zohar Karnin, Robert E Schapire, and Masrour Zoghi. Instance-dependent regret bounds for dueling bandits. In *Conference on Learning Theory*, pages 336–360. PMLR, 2016.

[8] 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*, 22(7):1–108, 2021.

[9] 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.

[10] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[11] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[12] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, 97:327–351, 2014.

[13] Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=SQnitDuow6`.

[14] Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. Avoiding $\exp(r_{\max})$ dependency in preference-based reinforcement learning. *arXiv preprint arXiv:2502.00666*, 2025.

[15] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *The Journal of Machine Learning Research*, 21(1):8918–8961, 2020.

[16] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

[17] Wooseong Cho, Taehyun Hwang, Joongkyu Lee, and Min-hwan Oh. Randomized exploration for reinforcement learning with multinomial logistic function approximation. *Advances in Neural Information Processing Systems*, 37:76643–76720, 2024.

[18] 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.

[19] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 96–112. Springer, 2025.

[20] Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. In *Forty-second International Conference on Machine Learning*, 2024.

[21] Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *Transactions on Machine Learning Research*, 2024.

[23] Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR, 2019.

[24] Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.

[25] Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. *Advances in Neural Information Processing Systems*, 30, 2017.

[26] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

[27] Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

[28] Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin Wainwright. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 1057–1066. PMLR, 2018.

[29] Taehyun Hwang and Min-hwan Oh. Model-based reinforcement learning with multinomial logistic function approximation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7971–7979, 2023.

[30] Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

[31] Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pages 416–424. PMLR, 2015.

[32] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. *Advances in Neural Information Processing Systems*, 30, 2017.

[33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

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

[35] Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072, 2022.

[36] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR, 2015.

[37] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, pages 1235–1244. PMLR, 2016.

[38] Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. In *Robotics Research: Volume 1*, pages 161–176. Springer, 2017.

[39] Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

[40] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[41] Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[42] Joongkyu Lee and Min-hwan Oh. Combinatorial reinforcement learning with preference feedback. In *Forty-second International Conference on Machine Learning*, 2025.

[43] Joongkyu Lee and Min-hwan Oh. Improved online confidence bounds for multinomial logistic bandits. In *Forty-second International Conference on Machine Learning*, 2025.

[44] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4482. PMLR, 2024.

[45] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *Advances in Neural Information Processing Systems*, 37:124640–124685, 2024.

[46] Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=2cQ3lPhkeO`.

[47] R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

[48] Daniel McFadden. Modelling the choice of residential location. 1977.

[49] Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*, 2023.

[50] Zakaria Mhammedi, Wouter M Koolen, and Tim Van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory*, pages 2490–2511. PMLR, 2019.

[51] Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Anand Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

[52] William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

[53] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2023.

[54] Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 1029–1038. PMLR, 2020.

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

[56] Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9205–9213, 2021.

[57] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[58] Mingdong Ou, Nan Li, Shenghuo Zhu, and Rong Jin. Multinomial logit bandit with linear utility functions. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2602–2608. International Joint Conferences on Artificial Intelligence Organization, 2018.

[59] 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.

[60] Jaehyun Park, Junyeop Kwon, and Dabeen Lee. Infinite-horizon reinforcement learning with multinomial logit function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 361–369. PMLR, 2025.

[61] Noemie Perivier and Vineet Goyal. Dynamic pricing and assortment under a contextual mnl demand. *Advances in Neural Information Processing Systems*, 35:3461–3474, 2022.

[62] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

[63] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791, 2008.

[64] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[65] Siddartha Y Ramamohan, Arun Rajkumar, and Shivani Agarwal. Dueling bandits: Beyond condorcet winners to general tournament solutions. *Advances in Neural Information Processing Systems*, 29, 2016.

[66] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42258–42274. PMLR, 21–27 Jul 2024.

[67] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. *Active preference-based learning of reward functions*. 2017.

[68] Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

[69] Aadirupa Saha and Pierre Gaillard. Stop relying on no-choice and do not repeat the moves: Optimal, efficient and practical algorithms for assortment optimization. *arXiv preprint arXiv:2402.18917*, 2024.

[70] Aadirupa Saha and Aditya Gopalan. Battle of bandits. In *UAI*, pages 805–814, 2018.

[71] Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244. PMLR, 2021.

[72] Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International conference on artificial intelligence and statistics*, pages 6263–6289. PMLR, 2023.

[73] Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*, 2024.

[74] Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, 36:11261–11295, 2023.

[75] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on learning theory*, pages 3–24. PMLR, 2013.

[76] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[77] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 47345–47377, 2024.

[78] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[79] Kiran Koshy Thekumparampil, Gaurush Hiranandani, Kousha Kalantari, Shoham Sabach, and Branislav Kveton. Comparing few to rank many: Active human preference learning using randomized frank-wolfe method. In *Forty-second International Conference on Machine Learning*, 2024.

[80] Joel A Tropp. User-friendly tail bounds for matrix martingales. *ACM Report*, 1, 2011.

[81] Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

[82] Christian Wirth and Johannes Fürnkranz. Preference-based reinforcement learning: A preliminary survey. In *Proceedings of the ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*, 2013.

[83] Christian Wirth, Johannes Fürnkranz, and Gerhard Neumann. Model-free preference-based reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[84] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017.

[85] Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

[86] Runzhe Wu and Wen Sun. Making rl with preference-based feedback efficient via randomization. In *The Twelfth International Conference on Learning Representations*, 2023.

[87] Yue Wu, Tao Jin, Qiwei Di, Hao Lou, Farzad Farnoud, and Quanquan Gu. Borda regret minimization for generalized linear dueling bandits. In *International Conference on Machine Learning*, pages 53571–53596. PMLR, 2024.

[88] Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. In *The Thirteenth International Conference on Learning Representations*, 2024.

[89] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

[90] Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

[91] 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.

[92] Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

[93] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[94] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Provable reward-agnostic preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[96] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

[97] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*, pages 10–18. PMLR, 2014.

[98] Masrour Zoghi, Zohar S Karnin, Shimon Whiteson, and Maarten De Rijke. Copeland dueling bandits. *Advances in neural information processing systems*, 28, 2015.

Appendix

Table of Contents

A Further Related Work

In this section, we provide additional related work that complements Section 2.

Logistic and MNL bandits. Our work is also closely related to logistic bandits and multinomial logit (MNL) bandits. The logistic bandit problem [23, 26, 2, 27, 44, 45] is a special case of the MNL bandit model in which the agent offers only a single item (i.e., $K = 1$) at each round and receives binary feedback indicating whether the item was selected (1) or not (0). Faury et al. [26] examined how the regret in logistic bandits depends on the non-linearity parameter κ of the logistic link function and proposed the first algorithm whose regret bound eliminates explicit dependence on $1/\kappa = \mathcal{O}(e^B)$. Abeille et al. [2] further improved the theoretical dependency on $1/\kappa$ and established a matching, problem-dependent lower bound. Building on this, Faury et al. [27] developed a computationally efficient algorithm whose regret still matches the lower bound established by Abeille et al. [2].

Multinomial logit (MNL) bandits tackle a more sophisticated problem than logistic bandits. Instead of offering a single item and observing binary feedback, the learner chooses a subset of items—underscoring the combinatorial nature of the task—and receives non-uniform rewards driven by an MNL choice model [5, 4, 58, 15, 55, 56, 61, 3, 41, 43]. A recent breakthrough by Lee and Oh [41] closed a long-standing gap by providing a computationally efficient algorithm that attains the minimax-optimal regret for this setting. Building on this result, Lee and Oh [43] further reduced the regret bound by a factor polynomial in B and logarithmic in K, and established the first variance-dependent regret bounds for MNL bandits.

Procedure A.1 Greedy Selection of Reference Action and Assortment

Input: $x_t, H_t^{-1}, \mathcal{A}, K$
Initialize: $(\bar{a}_t^\star, S_t^\star, \text{max\_avg}) \leftarrow (\text{None}, \text{None}, -\infty)$
for all $\bar{a} \in \mathcal{A}$ **do**
  Initialize $S \leftarrow \{\bar{a}\}$, prev_avg $\leftarrow 0$
  while $|S| < K$ **do**
    Find
$$a^\star \leftarrow \operatorname*{argmax}_{a \in \mathcal{A} \setminus S} \|\phi(x_t, a) - \phi(x_t, \bar{a})\|_{H_t^{-1}}$$
    Tentatively update $S' \leftarrow S \cup \{a^\star\}$
    Compute
$$\text{cur\_avg} \leftarrow \frac{1}{|S'|} \sum_{a \in S'} \|\phi(x_t, a) - \phi(x_t, \bar{a})\|_{H_t^{-1}}$$
    if cur_avg $<$ prev_avg **then**
      break
    else
      $S \leftarrow S'$
      prev_avg $\leftarrow$ cur_avg
    end if
  end while
  if prev_avg $>$ max_avg **then**
    $(\bar{a}_t^\star, S_t^\star, \text{max\_avg}) \leftarrow (\bar{a}, S, \text{prev\_avg})$
  end if
end for
return $(\bar{a}_t^\star, S_t^\star)$

Our work extends the online confidence bound analysis of Lee and Oh [43] to the Plackett–Luce (PL) model. This extension is natural because the PL probability distribution decomposes into a sequence of MNL probabilities over successive choices. Crucially, we leverage their key insight—that the MNL loss exhibits an ℓ_∞-self-concordant property—to eliminate the harmful $\mathcal{O}(e^B)$ dependence. This is one of the main contributions of our work (see Lemma D.2).

RL with MNL models. Recent work has extended the Multinomial Logit (MNL) framework beyond bandit formulations to reinforcement learning. Lee and Oh [42] introduced *combinatorial RL with preference feedback*, a framework in which an agent learns to select subsets of items so as to maximize long-term cumulative rewards.

Another line of research incorporates MNL models directly into the transition dynamics. Hwang and Oh [29] proposed MNL-MDPs, a class of Markov decision processes whose transition probabilities follow an MNL parameterization. Building upon this formulation, Cho et al. [17] improved the regret bounds by improving the exponential dependence on B, and Park et al. [60] extended the analysis to the infinite-horizon setting.

B Efficient Assortment Selection

In this section, we describe how the assortment selection rule in Equation (9) can be solved efficiently.

Given x_t, the reference action–assortment pair $(\bar{a}_t, S_t)$ is selected by evaluating each candidate reference action $\bar{a} \in \mathcal{A}$. For each $\bar{a}$, we construct an assortment S beginning with the singleton $\bar{a}$, and iteratively add actions $a \in \mathcal{A} \setminus \{\bar{a}\}$ in decreasing order of their uncertainty relative to $\bar{a}$, measured by

$$\|\phi(x_t, a) - \phi(x_t, \bar{a})\|_{H_t^{-1}} .$$

Let $a_{tk}(\bar{a})$ denote the action with the k-th highest uncertainty with respect to $\bar{a}$ at round t. For example, $a_{t1}(\bar{a}) = \operatorname{argmax}_{a \in \mathcal{A} \setminus \{\bar{a}\}} \|\phi(x_t, a) - \phi(x_t, \bar{a})\|_{H_t^{-1}}$. We add actions greedily to the set S, as long as the average uncertainty continues to increase:

$$\frac{1}{|S|}\sum_{a\in S}\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}\right)\|_{H_t^{-1}}, \quad \text{where } \bar{a}\in S.$$

Among all candidates $\bar{a}\in\mathcal{A}$, we select the pair $(\bar{a}_t, S_t)$ that achieves the highest average uncertainty. The pseudocode is given in Procedure A.1.

For each candidate reference action, the algorithm incrementally constructs a subset of actions by greedily adding those with the highest uncertainty relative to the reference—stopping once the average uncertainty no longer increases. This greedy strategy guarantees that, for each reference, the selected subset maximizes the average uncertainty. By applying this procedure across all possible reference actions and selecting the pair that achieves the highest score, the algorithm obtains the global optimum over all reference–assortment combinations.

As for the computational cost, each greedy addition step involves searching over $\mathcal{O}(N)$ candidate actions, resulting in a total of $\tilde{\mathcal{O}}(NK)$ operations per each reference action $\bar{a}$. Repeating this process for all N candidate references yields a total cost of $\tilde{\mathcal{O}}(N^2K)$.

C Notation

Let T denote the total number of rounds, with $t\in[T]$ representing the current round. We use N for the total number of items, K for the maximum assortment size, d for the feature vector dimension, and B as an upper bound on the norm of the unknown parameter. For notational convenience, we provide Table C.1.

For clarity, we derive the first- and second-order derivatives (i.e., gradients and Hessians) of the loss functions. For the PL loss at round t for the j'th ranking, let $y_{ti}^{(j)}=1$ if $i=j$, and $y_{ti}^{(j)}=0$ for otherwise. Then, we have

$$\begin{aligned}
\ell_t^{(j)}(\boldsymbol{\theta}) &= -\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top\boldsymbol{\theta}\right)}{\sum_{k=j}^{|S_t|}\exp\left(\phi(x_t,\sigma_{tk})^\top\boldsymbol{\theta}\right)}\right) = -\sum_{i=j}^{|S_t|} y_{ti}^{(j)}\log\Bigg(\underbrace{\frac{\exp\left(\phi(x_t,\sigma_{ti})^\top\boldsymbol{\theta}\right)}{\sum_{k=j}^{|S_t|}\exp\left(\phi(x_t,\sigma_{tk})^\top\boldsymbol{\theta}\right)}}_{=:P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti})}\Bigg) \\
&= -\sum_{i=j}^{|S_t|} y_{ti}^{(j)}\log P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti}), \\
\nabla\ell_t^{(j)}(\boldsymbol{\theta}) &= \sum_{i=j}^{|S_t|}\left(P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti}) - y_{ti}^{(j)}\right)\phi(x_t,\sigma_{ti}), \\
\nabla^2\ell_t^{(j)}(\boldsymbol{\theta}) &= \sum_{i=j}^{|S_t|} P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti})\phi(x_t,\sigma_{ti})\phi(x_t,\sigma_{ti})^\top - \sum_{i=j}^{|S_t|}\sum_{k=j}^{|S_t|} P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti})P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{tk})\phi(x_t,\sigma_{ti})\phi(x_t,\sigma_{tk})^\top \\
&= \frac{1}{2}\sum_{i=j}^{|S_t|}\sum_{k=j}^{|S_t|} P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{ti})P_{t,\boldsymbol{\theta}}^{(j)}(\sigma_{tk})\big(\phi(x_t,\sigma_{ti})-\phi(x_t,\sigma_{tk})\big)\big(\phi(x_t,\sigma_{ti})-\phi(x_t,\sigma_{tk})\big)^\top.
\end{aligned}$$

For the RB loss at round t for the pairwise comparison between σ_{tj} and σ_{tk}, let $y_{ti}^{(j,k)}=1$ if $i=j$, and $y_{ti}^{(j,k)}=0$ for otherwise (i.e., when $i=k$). Then, we have

$$\begin{aligned}
\ell_t^{(j,k)}(\boldsymbol{\theta}) &= -\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top\boldsymbol{\theta}\right)}{\exp\left(\phi(x_t,\sigma_{tj})^\top\boldsymbol{\theta}\right)+\exp\left(\phi(x_t,\sigma_{tk})^\top\boldsymbol{\theta}\right)}\right) \\
&= -\log\mu\Big(\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big)^\top\boldsymbol{\theta}\Big), \quad \text{where } \mu(w)=\frac{1}{1+e^{-w}}, \\
\nabla\ell_t^{(j,k)}(\boldsymbol{\theta}) &= \Big(\mu\Big(\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big)^\top\boldsymbol{\theta}\Big)-1\Big)\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big), \\
\nabla^2\ell_t^{(j,k)}(\boldsymbol{\theta}) &= \dot{\mu}\Big(\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big)^\top\boldsymbol{\theta}\Big)\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big)\big(\phi(x_t,\sigma_{tj})-\phi(x_t,\sigma_{tk})\big)^\top.
\end{aligned}$$

Table C.1: Symbols

$\mathcal{X}, \mathcal{A}, \mathcal{S}$	context (prompt) space, action (answer) space, assortment space
$\phi(x,a) \in \mathbb{R}^d$	feature representation of context-action pair (x,a)
z_{tjk}	$:= \phi(x_t, \sigma_{tj}) - \phi(x_t, \sigma_{tk})$, feature difference between σ_{tj} and σ_{tk} under context x_t
S_t	assortment chosen by an algorithm at round t
$\ell_t^{(j)}(\boldsymbol{\theta})$	$:= -\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top \boldsymbol{\theta}\right)}{\sum_{k=j}^{\lvert S_t\rvert} \exp(\phi(x_t,\sigma_{tk})^\top \boldsymbol{\theta}))}\right)$, PL loss at round t for j'th ranking
$\ell_t^{(j,k)}(\boldsymbol{\theta})$	$:= -\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top \boldsymbol{\theta}\right)}{\sum_{m\in\{j,k\}} \exp(\phi(x_t,\sigma_{tm})^\top \boldsymbol{\theta})}\right)$, RB loss at round t for comparison σ_{tj} vs σ_{tk}
$\nabla^2 \ell_t^{(j)}(\boldsymbol{\theta})$	$= \sum_{k=j}^{\lvert S_t\rvert} \sum_{k'=j}^{\lvert S_t\rvert} \frac{\exp\left((\phi(x_t,\sigma_{tk})+\phi(x_t,\sigma_{tk'}))^\top \boldsymbol{\theta}\right)}{2\left(\sum_{k'=j}^{\lvert S_t\rvert} \exp(\phi(x_t,\sigma_{tk'})^\top \boldsymbol{\theta})\right)^2} \cdot z_{tkk'} z_{tkk'}^\top$
$\nabla^2 \ell_t^{(j,k)}(\boldsymbol{\theta})$	$= \dot{\mu}\left(z_{tjk}^\top \boldsymbol{\theta}\right) z_{tjk} z_{tjk}^\top$, where $\mu(w) = \frac{1}{1+e^{-w}}$ is sigmoid function
$\widehat{\boldsymbol{\theta}}_t^{(j+1)}$	online parameter estimate using PL loss at round t, after j'th update
$\widehat{\boldsymbol{\theta}}_t^{(j,k+1)}$	online parameter estimate using RB loss at round t, after (j,k)'th comparison update
η	$:= \frac{1}{2}(1+3\sqrt{2}B)$, step-size parameter
λ	$:= \Omega\big(d\log(KT/\delta) + \eta(B+d)\big)$, regularization parameter
H_t	$:= \sum_{s=1}^{t-1} \sum_{j=1}^{\lvert S_s\rvert} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d$ (or $\sum_{s=1}^{t-1} \sum_{j=1}^{\lvert S_s\rvert-1} \sum_{k=j+1}^{\lvert S_s\rvert} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d$)
$\tilde{H}_t^{(j)}$	$:= H_t + \eta \sum_{j'=1}^{j} \nabla^2 \ell_t^{(j')}(\widehat{\boldsymbol{\theta}}_t^{(j')})$ (for PL loss)
$\tilde{H}_t^{(j,k)}$	$:= H_t + \eta \sum_{(j',k')\leqslant(j,k)} \nabla^2 \ell_t^{(j',k')}(\widehat{\boldsymbol{\theta}}_t^{(j',k')})$ (for RB loss)
$\beta_t(\delta)$	$:= \mathcal{O}\left(B\sqrt{d\log(tK/\delta)} + B\sqrt{\lambda}\right)$, confidence radius for $\boldsymbol{\theta}_t$ at round t
$\mathcal{T}^w$	$:= \left\{t \in [T] : \max_{a\in\mathcal{A}} \lVert \phi(x_t,a) - \phi(x_t,\bar{a}_t)\rVert_{H_t^{-1}} \geqslant \frac{1}{3\sqrt{2}\beta_{T+1}(\delta)}\right\}$, warm-up rounds
Λ_t	$:= \sum_{s\in[t-1]\setminus\mathcal{T}^w} \sum_{a\in S_s} \big(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\big)\big(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\big)^\top + \lambda \mathbf{I}_d$
$\mathcal{T}_0$	$:= \left\{t \in [T] : \sum_{a\in S_t} \lVert \phi(x_t,a) - \phi(x_t,\bar{a}_t)\rVert_{\Lambda_t^{-1}} \geqslant 1.\right\}$, large EP rounds

D Proof of Theorem 1

In this section, we present the proof of Theorem 1.

D.1 Main Proof of Theorem 1

PL loss and OMD. We begin by recalling the loss function and the parameter update rule. Specifically, we use the PL loss defined in Equation (3) and update the parameter according to Equation (5).

$$\ell_t(\boldsymbol{\theta}) := \sum_{j=1}^{|S_t|} \underbrace{-\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top \boldsymbol{\theta}\right)}{\sum_{k=j}^{|S_t|} \exp\left(\phi(x_t,\sigma_{tk})^\top \boldsymbol{\theta}\right))}\right)}_{=:\ell_t^{(j)}(\boldsymbol{\theta})} = \sum_{j=1}^{|S_t|} \ell_t^{(j)}(\boldsymbol{\theta}).$$

and

$$\widehat{\boldsymbol{\theta}}_t^{(j+1)} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \langle \nabla \ell_t^{(j)}(\widehat{\boldsymbol{\theta}}_t^{(j)}), \boldsymbol{\theta}\rangle + \frac{1}{2\eta}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^{(j)}\|_{\tilde{H}_t^{(j)}}^2, \quad j = 1, \dots, |S_t|,$$

where $\widehat{\boldsymbol{\theta}}_t^{(|S_t|+1)} = \widehat{\boldsymbol{\theta}}_{t+1}^{(1)}$, and $\eta := \frac{1}{2}(1+3\sqrt{2}B)$ is the step-size parameter. The matrix $\tilde{H}_t^{(j)}$ is given by $\tilde{H}_t^{(j)} := H_t + \eta \sum_{j'=1}^{j} \nabla^2 \ell_t^{(j')}(\widehat{\boldsymbol{\theta}}_t^{(j')})$, where

$$H_t := \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d, \quad \lambda > 0.$$

Online confidence bound for PL loss. Now, we present the confidence bound for online parameter estimation in MNL models, as recently proposed by Lee and Oh [43].

Lemma D.1 (Online confidence bound, Theorem 4.2 of Lee and Oh 43)**.** *Let* $\delta \in (0, 1]$. *We set* $\eta = (1 + 3\sqrt{2}B)/2$ *and* $\lambda = \max\{12\sqrt{2}B\eta, 144\eta d, 2\}$. *Then, under Assumption 1, with probability at least* $1 - \delta$*, we have*

$$\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\|_{H_t} \leqslant \beta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(t/\delta)} + B\sqrt{\lambda}\right), \quad \forall t \geqslant 1.$$

We now extend this result to our setting. Since the total number of updates up to round t is $\sum_{s=1}^{t} |S_s|$, the corresponding confidence bound can be expressed as follows:

Corollary D.1 (Restatement of Corollary 1, Online confidence bound for PL loss)**.** *Let* $\delta \in (0, 1]$. *We set* $\eta = (1 + 3\sqrt{2}B)/2$ *and* $\lambda = \max\{12\sqrt{2}B\eta, 144\eta d, 2\}$. *Then, under Assumption 1, with probability at least* $1 - \delta$*, we have*

$$\|\widehat{\boldsymbol{\theta}}_t^{(j)} - \boldsymbol{\theta}^\star\|_{H_t^{(j)}} \leqslant \beta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(tK/\delta)} + B\sqrt{\lambda}\right), \quad \forall t \geqslant 1, j \leqslant |S_t|,$$

where $H_t^{(j)} := H_t + \sum_{j'=1}^{j-1} \nabla^2 \ell_s^{(j')}(\widehat{\boldsymbol{\theta}}_s^{(j'+1)}) + \lambda \mathbf{I}_d$ *and* $\widehat{\boldsymbol{\theta}}_t^{(1)} = \widehat{\boldsymbol{\theta}}_t$.

Useful definitions. We define the set of *warm-up rounds*, denoted by $\mathcal{T}^w$, which consists of rounds with large uncertainty, as follows:

$$\mathcal{T}^w := \left\{ t \in [T] : \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}} \geqslant \frac{1}{3\sqrt{2}\beta_{T+1}(\delta)} \right\}, \tag{D.1}$$

where $\beta_{T+1}(\delta)$ denotes the confidence radius as defined in Corollary D.1. Furthermore, we define the regularized sample covariance matrix of feature differences (with respect to $\phi(x_s, \bar{a}_s)$) over the non-warm-up rounds as:

$$\Lambda_t := \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{a \in S_s} \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top + \lambda \mathbf{I}_d. \tag{D.2}$$

To control the elliptical potentials, we also define the set of *large elliptical potential (EP) rounds*, denoted by $\mathcal{T}_0$, as follows:

$$\mathcal{T}_0 := \left\{ t \in [T] : \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{\Lambda_t^{-1}} \geqslant 1. \right\}, \tag{D.3}$$

Key lemmas. We now present key lemmas needed to prove Theorem 1. The following lemma, one of our main contributions, is crucial for avoiding the $1/\kappa = \mathcal{O}(e^B)$ dependency in the leading term.

Lemma D.2 (Restatement of Lemma 1)**.** *Let* Λ_t *be defined as in Equation* (D.2). *Set* $\lambda = \Omega(d\log(KT/\delta))$. *Then, for all* $t \in [T]$*, with probability at least* $1 - \delta$*, we have*

$$H_t \succeq \frac{1}{50}\Lambda_t.$$

The proof is deferred to Appendix D.2.1.

The following lemma is a variant of the elliptical potential lemma [1], adapted specifically to the assortment offering setting. For completeness, we provide the proof tailored to our setting.

Lemma D.3 (Elliptical potential lemma for S_t)**.** *Let* $\{z_{ta}\}_{t \geqslant 1, a \in S_t}$ *be a bounded sequence in* $\mathbb{R}^d$ *satisfying* $\max_{t \geqslant 1} \|z_{ta}\|_2 \leqslant X$. *For any* $t \geqslant 1$*, we define* $\Lambda_t := \sum_{s=1}^{t-1} \sum_{a \in S_s} z_{sa} z_{sa}^\top + \lambda \mathbf{I}_d$ *with* $\lambda > 0$. *Then, we have*

$$\sum_{t=1}^{T} \min\left\{1, \sum_{a \in S_t} \|z_{ta}\|_{\Lambda_t^{-1}}^2\right\} \leqslant 2d\log\left(1 + \frac{X^2 KT}{d\lambda}\right).$$

The proof is deferred to Appendix D.2.2.

The cardinality of the set $\mathcal{T}_0$ can be bounded by a variant of the elliptical potential counting lemma [40, 35].

Lemma D.4 (Elliptical potential count lemma for S_t)**.** *Let $\{z_{ta}\}_{t\geqslant 1, a\in S_t}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\max_{t\geqslant 1}\|z_{ta}\|_2 \leqslant X$. For any $t \geqslant 1$, we define $\Lambda_t := \sum_{s=1}^{t-1}\sum_{a\in S_s} z_{sa}z_{sa}^\top + \lambda \mathbf{I}_d$ with $\lambda > 0$. Let $\mathcal{T}_0 \subseteq [T]$ be the set of indices where $\sum_{a\in S_t}\|z_{ta}\|^2_{\Lambda_t^{-1}} \geqslant L$. Then,*

$$|\mathcal{T}_0| \leqslant \frac{2d}{\log(1+L)}\log\left(1+\frac{X^2K}{\log(1+L)\lambda}\right).$$

The proof is deferred to Appendix D.2.3.

The size of the set $\mathcal{T}^w \cap (\mathcal{T}_0)^c$ is bounded as described in the following lemma:

Lemma D.5. *Let $\mathcal{T}_0 := \{t \in [T] : \sum_{a\in S_t}\|\phi(x_t,a)-\phi(x_t,\bar{a}_t)\|_{\Lambda_t^{-1}} \geqslant 1.\}$ and $\mathcal{T}^w = \{t\in[T] : \max_{a\in\mathcal{A}}\|\phi(x_t,a)-\phi(x_t,\bar{a}_t)\|_{H_t^{-1}} \geqslant \frac{1}{3\sqrt{2}\beta_{T+1}(\delta)}\}$. Define $\kappa := e^{-4B}$. Then, the size of the set $\mathcal{T}^w \cap (\mathcal{T}_0)^c$ is bounded as follows:*

$$|\mathcal{T}^w \cap (\mathcal{T}_0)^c| \leqslant \frac{12\sqrt{2}K^2}{\kappa}\beta_{T+1}(\delta)^2 d\log\left(1+\frac{2KT}{d\lambda}\right).$$

The proof is deferred to Appendix D.2.4.

We are now ready to provide the proof of Theorem 1.

Proof of Theorem 1. To begin, we define a martingale difference sequence (MDS) ζ_t as follows:

$$\zeta_t := \mathbb{E}_{x\sim\rho}\left[\left(\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right)^\top\boldsymbol{\theta}^\star\right]-\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\boldsymbol{\theta}^\star,$$

which satisfies $|\zeta_t| \leqslant 2B$. Then, by the definition of the suboptimality gap, we have

$$\begin{aligned}
\textbf{SubOpt}(T) &= \mathbb{E}_{x\sim\rho}\left[\left(\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right)^\top\boldsymbol{\theta}^\star\right]\\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\boldsymbol{\theta}^\star + \frac{1}{T}\sum_{t=1}^{T}\zeta_t && \text{(Def. of } \zeta_t)\\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right) + \frac{1}{T}\sum_{t=1}^{T}\zeta_t\\
& && (\widehat{\pi}_T(x_t) = \operatorname{argmax}_{a\in\mathcal{A}}\phi(x_t,a)^\top\widehat{\boldsymbol{\theta}}_{T+1})\\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right), && \text{(D.4)}
\end{aligned}$$

where the last inequality follows from the Azuma–Hoeffding inequality. Specifically, for any $T \geqslant 1$, with probability at least $1-\delta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\zeta_t \leqslant \frac{1}{T}\sqrt{8B^2T\log(1/\delta)} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right).$$

To complete the proof, it remains to bound the first term in Equation (D.4).

Recall the definitions of the set of *large elliptical potential (EP) rounds* (Equation (D.3)), denoted by $\mathcal{T}_0$, and the set of *warm-up rounds* (Equation (D.1)), denoted by $\mathcal{T}^w$:

$$\begin{aligned}
\mathcal{T}_0 &:= \left\{t\in[T] : \sum_{a\in S_t}\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\|_{\Lambda_t^{-1}} \geqslant 1.\right\}, && \text{(large EP rounds)}\\
\mathcal{T}^w &:= \left\{t\in[T] : \max_{a\in\mathcal{A}}\|\phi(x_t,a)-\phi(x_t,\bar{a}_t)\|_{H_t^{-1}} \geqslant \frac{1}{3\sqrt{2}\beta_{T+1}(\delta)}\right\}, && \text{(warm-up rounds)}
\end{aligned}$$

where Λ_t is defined in Equation (D.2). Then, by applying the elliptical potential count lemma (Lemma D.4) and the bound on the cardinality of the set $|\mathcal{T}^w \cap (\mathcal{T}_0)^c|$ lemma (Lemma D.5), we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\quad=\frac{1}{T}\sum_{t\in\mathcal{T}_0}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\quad+\frac{1}{T}\sum_{t\in\mathcal{T}^w\cap(\mathcal{T}_0)^c}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\quad+\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\quad\leqslant\frac{4B}{T}|\mathcal{T}_0|+\frac{4B}{T}|\mathcal{T}^w\cap(\mathcal{T}_0)^c|+\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\hspace{30em}\text{(Assumption 1)}\\
&\quad\leqslant\frac{8B}{\log(2)T}d\log\left(1+\frac{2K}{\log(2)\lambda}\right)+\frac{48\sqrt{2}BK^2}{\kappa T}\beta_{T+1}(\delta)^2 d\log\left(1+\frac{2KT}{d\lambda}\right)\\
&\hspace{30em}\text{(Lemma D.4 and D.5)}\\
&\quad+\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right).
\end{aligned}
\tag{D.5}
$$

To further bound the last term of Equation (D.5), we get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\\
&\qquad\leqslant\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_{T+1}^{-1}}\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_{T+1}}\\
&\hspace{30em}\text{(Hölder's ineq.)}\\
&\qquad\leqslant\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}}\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_{T+1}}\\
&\hspace{30em}(H_{T+1}\succeq H_t)\\
&\qquad\leqslant\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}}.\\
&\hspace{24em}\text{(Corollary D.1, with prob. } 1-\delta)
\end{aligned}
$$

We denote $S_t^\star=\{\pi^\star(x_t),\widehat{\pi}_T(x_t)\}$. Then, we have

$$
\begin{aligned}
&\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}}\\
&\qquad=\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\sum_{a\in S_t^\star}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}}\\
&\qquad=\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{|S_t^\star|}{|S_t^\star|}\sum_{a\in S_t^\star}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}}\\
&\qquad=\frac{2\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{1}{|S_t^\star|}\sum_{a\in S_t^\star}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\right\|_{H_t^{-1}},
\end{aligned}
\tag{D.6}
$$

where the last equality holds due to the fact that $|S_t^\star| = 2$. To proceed, by our efficient assortment selection rule in Equation (9), we obtain

$$
\begin{aligned}
&\frac{2\beta_{T+1}(\delta)}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \frac{1}{|S_t^\star|} \sum_{a \in S_t^\star} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \widehat{\pi}_T(x_t)\right) \right\|_{H_t^{-1}} \\
&\quad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \frac{1}{|S_t|} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi(x_t, \bar{a}_t) \right\|_{H_t^{-1}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (S_t \text{ selection rule, Eqn. (9)}) \\
&\quad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\frac{1}{|S_t|}\right)^2 |S_t|} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi(x_t, \bar{a}_t) \right\|_{H_t^{-1}}^2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Cauchy-Schwartz ineq.}) \\
&\quad = \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\frac{1}{|S_t|}\right)^2 |S_t|} \sqrt{50 \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi(x_t, \bar{a}_t) \right\|_{\Lambda_t^{-1}}^2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Lemma D.2, with prob. } 1-\delta) \\
&\quad \leqslant \frac{15\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^T \frac{1}{|S_t|}} \sqrt{\sum_{t=1}^T \min\left\{1, \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi(x_t, \bar{a}_t) \right\|_{\Lambda_t^{-1}}^2 \right\}} \\
&\quad \leqslant \frac{15\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^T \frac{1}{|S_t|}} \sqrt{2d \log\left(1 + \frac{2KT}{d\lambda}\right)} \qquad\qquad (\text{Lemma D.3}) \\
&\quad = \mathcal{O}\left(\frac{\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^T \frac{1}{|S_t|}} \cdot \sqrt{d \log\left(KT\right)} \right). \qquad\qquad (\text{D.7})
\end{aligned}
$$

By combining Equations (D.4), (D.5), and (D.7), and setting $\beta_{T+1}(\delta) = \mathcal{O}\big(B\sqrt{d\log(KT)} + B\sqrt{\lambda}\big)$, we derive that, with probability at least $1-3\delta$ (omitting logarithmic terms and polynomial dependencies on B for brevity),

$$
\textbf{SubOpt}(T) = \tilde{\mathcal{O}}\left(\frac{d}{T} \sqrt{\sum_{t=1}^T \frac{1}{|S_t|}} + \frac{d^2K^2}{\kappa T} \right).
$$

Substituting $\delta \leftarrow \frac{\delta}{3}$, we conclude the proof of Theorem 1.

□

D.2 Proofs of Lemmas for Theorem 1

D.2.1 Proof of Lemma D.2

Proof of Lemma D.2. Recall the definition of H_t.

$$
H_t = \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d \succeq \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d
$$

Here, we can equivalently express the MNL loss at step j and round s, denoted by $\nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)})$, as follows:

$$
\begin{aligned}
\ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) &= -\log\left(\frac{\exp\left(\phi(x_s, \sigma_{sj})^\top \widehat{\boldsymbol{\theta}}_s^{(j+1)}\right)}{\sum_{k=j}^{|S_s|} \exp\left(\phi(x_s, \sigma_{sk})^\top \widehat{\boldsymbol{\theta}}_s^{(j+1)})\right)} \right) = -\log\left(\frac{\exp\left(a_{sj}\right)}{\sum_{k=j}^{|S_s|} \exp\left(a_{sk}\right)} \right) \\
&=: \bar{\ell}_s^{(j)}(\mathbf{a}_s^{(j)}), \qquad\qquad (\text{D.8})
\end{aligned}
$$

where $a_{sj} = \phi(x_s, \sigma_{sj})^\top \widehat{\boldsymbol{\theta}}_s^{(j+1)}$, $\mathbf{a}_s^{(j)} = (a_{sk})_{k=j}^{|S_s|} \in \mathbb{R}^{|S_s|-j+1}$. Define the matrix

$$\mathbf{\Phi}_s^{(j)} = \begin{pmatrix} \phi(x_s, \sigma_{sj})^\top \\ \vdots \\ \phi(x_s, \sigma_{s|S_s|})^\top \end{pmatrix} \in \mathbb{R}^{(|S_s|-j+1)\times d},$$

where each row corresponds to the feature vector of an action ranked from position j to $|S_s|$ in the ranking σ_s. Moreover, we define $a^\star_{sj} = \phi(x_s, \sigma_{sj})^\top \boldsymbol{\theta}^\star$, $\mathbf{a}_s^{\star,(j)} = (a^\star_{sk})_{k=j}^{|S_s|} \in \mathbb{R}^{|S_s|-j+1}$.

Then, using the ℓ_∞-norm self-concordant property of the MNL loss [43], for any $s \in [t-1]\backslash\mathcal{T}^w$, we obtain

$$\begin{aligned}
\nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) &= \left(\mathbf{\Phi}_s^{(j)}\right)^\top \nabla^2_{\mathbf{a}} \bar{\ell}_s^{(j)}(\mathbf{a}_s^{(j)})\, \mathbf{\Phi}_s^{(j)} && \text{(Eqn. (D.8))} \\
&\succeq e^{-3\sqrt{2}\|\mathbf{a}_s^{(j)} - \mathbf{a}_s^{\star,(j)}\|_\infty} \left(\mathbf{\Phi}_s^{(j)}\right)^\top \nabla^2_{\mathbf{a}} \bar{\ell}_s^{(j)}(\mathbf{a}_s^{\star,(j)})\, \mathbf{\Phi}_s^{(j)} && \text{(Lemma F.1)} \\
&\succeq \frac{1}{e} \left(\mathbf{\Phi}_s^{(j)}\right)^\top \nabla^2_{\mathbf{a}} \bar{\ell}_s^{(j)}(\mathbf{a}_s^{\star,(j)})\, \mathbf{\Phi}_s^{(j)} && (\|\mathbf{a}_s^{(j)} - \mathbf{a}_s^{\star,(j)}\|_\infty \leqslant \tfrac{1}{3\sqrt{2}}) \\
&= \frac{1}{e} \nabla^2 \ell_s^{(j)}(\boldsymbol{\theta}^\star), && \text{(Eqn. (D.8))}
\end{aligned}$$

where the last inequality holds because, for any $s \in [t-1]\backslash\mathcal{T}^w$ and $j \leqslant |S_s|$, the following holds:

$$\begin{aligned}
\|\mathbf{a}_s^{(j)} - \mathbf{a}_s^{\star,(j)}\|_\infty &= \max_{k=j,\dots,|S_s|} \left| \phi(x_k, \sigma_{sk})^\top \left(\widehat{\boldsymbol{\theta}}_s^{(k+1)} - \boldsymbol{\theta}^\star\right) \right| \\
&\leqslant \max_{k=j,\dots,|S_s|} \|\phi(x_k, \sigma_{sk})\|_{H_s^{-1}} \left\| \widehat{\boldsymbol{\theta}}_s^{(k+1)} - \boldsymbol{\theta}^\star \right\|_{H_s} && \text{(Hölder's inequality)} \\
&\leqslant \frac{1}{3\sqrt{2}\beta_{T+1}(\delta)} \max_{k=j,\dots,|S_s|} \left\| \widehat{\boldsymbol{\theta}}_s^{(k+1)} - \boldsymbol{\theta}^\star \right\|_{H_s^{(k+1)}} && (s \notin \mathcal{T}^w,\ H_s \preceq H_s^{(k+1)}) \\
&\leqslant \frac{\beta_{T+1}(\delta)}{3\sqrt{2}\beta_{T+1}(\delta)} && \text{(Corollary D.1, } \beta_t(\delta) \text{ is non-decreasing)} \\
&= \frac{1}{3\sqrt{2}}.
\end{aligned}$$

Therefore, we get

$$H_t \succeq \sum_{s\in[t-1]\backslash\mathcal{T}^w} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d \succeq \frac{1}{e} \sum_{s\in[t-1]\backslash\mathcal{T}^w} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\boldsymbol{\theta}^\star) + \lambda \mathbf{I}_d. \tag{D.9}$$

Now, for better presentation, we define the Multinomial Logit (MNL) choice probability [48] for a given assortment S at round s as follows:

$$P_s(a|S;\boldsymbol{\theta}) := \frac{\exp\left(\phi(x_s, a)^\top \boldsymbol{\theta}\right)}{\sum_{a'\in S} \exp\left(\phi(x_s, a')^\top \boldsymbol{\theta}\right)}, \quad \forall a \in S.$$

Let $S_s = \{\sigma_{s1}, \dots, \sigma_{s|S_s|}\}$. Thus, the PL model in Equation (1) can be rewritten as follows:

$$\begin{aligned}
\mathbb{P}(\sigma_s|x_s, S_s;\boldsymbol{\theta}) &= P_s(\sigma_{s1}|S_s;\boldsymbol{\theta}) \cdot P_s(\sigma_{s2}|S_s\backslash\{\sigma_{s1}\};\boldsymbol{\theta}) \cdot\ \dots\ \cdot P_s(\sigma_{s|S_s|}|\{\sigma_{s|S_s|}\};\boldsymbol{\theta}) \\
&= \prod_{j=1}^{|S_s|} P_s\left(\sigma_{sj}|\{\sigma_{sj}, \dots, \sigma_{s|S_s|}\};\boldsymbol{\theta}\right).
\end{aligned}$$

For simplicity, we define $S_s^{(j)} := \{\sigma_{sj}, \dots, \sigma_{s|S_s|}\}$, and let $P_s^{(j)}$ denote the (true) MNL distribution over the remaining actions in S_s after removing the first $j-1$ selected actions, i.e., $P_s^{(j)} = P_s(\cdot|S_s^{(j)};\boldsymbol{\theta}^\star)$. Then, to further lower bound the right-hand side of Equation (D.9), we proceed as follows:

$$\begin{aligned}
\sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\boldsymbol{\theta}^\star) &= \sum_{j=1}^{|S_s|}\sum_{k=j}^{|S_s|}\sum_{k'=j}^{|S_s|} \frac{\exp\left(\left(\phi(x_s,\sigma_{sk}) + \phi(x_s,\sigma_{sk'})\right)^\top \boldsymbol{\theta}^\star\right)}{2\left(\sum_{k'=j}^{|S_s|}\exp\left(\phi(x_s,\sigma_{sk'})^\top \boldsymbol{\theta}^\star\right)\right)^2} \cdot z_{skk'} z_{skk'}^\top \\
&= \frac{1}{2}\sum_{j=1}^{|S_s|}\sum_{k=j}^{|S_s|}\sum_{k'=j}^{|S_s|} P_s(\sigma_{sk}|S_s^{(j)};\boldsymbol{\theta}^\star) P_s(\sigma_{sk'}|S_s^{(j)};\boldsymbol{\theta}^\star) z_{skk'} z_{skk'}^\top \\
&= \frac{1}{2}\sum_{j=1}^{|S_s|} \mathbb{E}_{(a,a')\sim P_s^{(j)}\times P_s^{(j)}}\left[\left(\phi(x_s,a)-\phi(x_s,a')\right)\left(\phi(x_s,a)-\phi(x_s,a')\right)^\top\right],
\end{aligned} \tag{D.10}$$

where $z_{skk'} = \phi(x_s,\sigma_{sk}) - \phi(x_s,\sigma_{sk'})$. Let the action $\bar{a}_s \in S_s$ be ranked at position $\bar{k}_s$ in the ranking σ_s. That is,

$$\sigma_s = (\underbrace{\sigma_{s1},\dots\sigma_{s\bar{k}_s-1}}_{\bar{k}_s-1 \text{ actions}}, \bar{a}_s, \sigma_{s\bar{k}_s+1},\dots\sigma_{s|S_s|-1}).$$

Note that $\bar{a}_s \in S_s^{(j)}$ for $j \leqslant \bar{k}_s$. We also note that $P_s^{(j)}$ is measurable with respect to the filtration $\mathcal{F}'_{s-1,j-1} = \boldsymbol{\sigma}\left(S_1,\sigma_{11},\sigma_{12},\dots,S_s,\sigma_{s1},\dots\sigma_{sj-1}\right)$. Then, by plugging Equation (D.10) into Equation (D.9) and applying the covariance matrix concentration result (Corollary F.1), since $\lambda = \Omega(d\log(KT/\delta))$, we have, with probability at least $1-\delta$,

$$\begin{aligned}
H_t &\succeq \frac{1}{2e}\sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{j=1}^{|S_s|} \mathbb{E}_{(a,a')\sim P_s^{(j)}\times P_s^{(j)}}\left[\left(\phi(x_s,a)-\phi(x_s,a')\right)\left(\phi(x_s,a)-\phi(x_s,a')\right)^\top\right] + \lambda \mathbf{I}_d \\
&\succeq \frac{1}{2e}\sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{j=1}^{\bar{k}_s} \mathbb{E}_{(a,a')\sim P_s^{(j)}\times P_s^{(j)}}\left[\left(\phi(x_s,a)-\phi(x_s,a')\right)\left(\phi(x_s,a)-\phi(x_s,a')\right)^\top\right] + \lambda \mathbf{I}_d \\
&\qquad\qquad (\bar{k}_s \leqslant |S_s|) \\
&\succeq \frac{3}{10e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{j=1}^{\bar{k}_s}\left(\phi(x_s,\sigma_{sj})-\phi(x_s,\bar{a}_s)\right)\left(\phi(x_s,\sigma_{sj})-\phi(x_s,\bar{a}_s)\right)^\top + \lambda\mathbf{I}_d\right) \\
&\qquad\qquad (\text{Corollary F.1, } \bar{a}_s \in S_s^{(j)} \text{ for } j \leqslant \bar{k}_s) \\
&= \frac{3K}{10e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w}\underbrace{\frac{1}{K}\sum_{j=1}^{\bar{k}_s}\left(\phi(x_s,\sigma_{sj})-\phi(x_s,\bar{a}_s)\right)\left(\phi(x_s,\sigma_{sj})-\phi(x_s,\bar{a}_s)\right)^\top}_{=:X(\sigma_s)} + \lambda\mathbf{I}_d\right).
\end{aligned}$$

Here, $\{X(\sigma_s)\}_{s\in[t-1]\setminus\mathcal{T}^w}$ is a sequence of positive semi-definite (PSD) random matrices, where each matrix $X(\sigma_s)$ depends on the sampled ranking σ_s, and satisfies $\lambda_{\max}\left(X(\sigma_s)\right) \leqslant 1$.

Note that the ranking σ_s is drawn from the PL distribution $\mathbb{P}(\cdot \mid x_s, S_s;\boldsymbol{\theta}^\star)$, which is measurable with respect to the filtration $\mathcal{F}_{s-1} = \boldsymbol{\sigma}(S_1,\sigma_1,\dots,S_s)$. Furthermore, $X(\sigma_s)$ is measurable with respect to $\boldsymbol{\sigma}\left(\mathcal{F}_{s-1},\sigma_s\right)$. Then, by applying the concentration lemma for PSD matrices (Lemma F.4) two times, with probability at least $1-2\delta$, we get

$$\begin{aligned}
H_t &\succeq \frac{3K}{10e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w} X(\sigma_s) + \lambda\mathbf{I}_d\right) \\
&\succeq \frac{K}{10e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w} \mathbb{E}_{\sigma\sim\mathbb{P}(\cdot|x_s,S_s;\boldsymbol{\theta}^\star)}\left[X(\sigma)\right] + \lambda\mathbf{I}_d\right) && \text{(Lemma F.4)} \\
&\succeq \frac{3K}{50e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w} X(\tilde{\sigma}_s) + \lambda\mathbf{I}_d\right), && \text{(Lemma F.4)}
\end{aligned}$$

where $\tilde{\sigma}_s$ denotes an arbitrary ranking in which $\bar{a}_s$ is placed last. For example, $\tilde{\sigma}_s = (\sigma_{s1}, \dots, \sigma_{s\bar{k}_s - 1}, \sigma_{s\bar{k}_s+1}, \sigma_{s|S_s|-1}, \bar{a}_s)$. Note that $\tilde{\sigma}_s$ is a possible *virtual* ranking feedback for the assortment S_s, whereas σ_s denotes the *actual* ranking feedback observed at round s. Hence, since $\bar{a}_s$ occupies the final position in the virtual sequence $\tilde{\sigma}_s$, it follows that:

$$
\begin{aligned}
H_t &\succeq \frac{3K}{50e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w} X(\tilde{\sigma}_s) + \lambda \mathbf{I}_d\right) \\
&= \frac{3}{50e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{j=1}^{|S_s|}\big(\phi(x_s,\tilde{\sigma}_{sj}) - \phi(x_s,\bar{a}_s)\big)\big(\phi(x_s,\tilde{\sigma}_{sj}) - \phi(x_s,\bar{a}_s)\big)^\top + \lambda \mathbf{I}_d\right) \\
&= \frac{3}{50e}\left(\sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{a\in S_s}\big(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\big)\big(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\big)^\top + \lambda \mathbf{I}_d\right) \\
&= \frac{3}{50e}\Lambda_t \succeq \frac{1}{50}\Lambda_t. && \text{(Def. of } \Lambda_t\text{, Eqn. (D.2))}
\end{aligned}
$$

By substituting $\delta \leftarrow \frac{\delta}{3}$, we conclude the proof of Lemma D.2. □

D.2.2 Proof of Lemma D.3

Proof of Lemma D.3. By the definition of Λ_t, we have

$$
\begin{aligned}
\det(\Lambda_{t+1}) &= \det\left(\Lambda_t + \sum_{a\in S_t} z_{ta} z_{ta}^\top\right) \\
&\geqslant \det(\Lambda_t)\left(1 + \sum_{a\in S_t}\|z_{ta}\|^2_{\Lambda_t^{-1}}\right) \\
&\geqslant \det(\lambda \mathbf{I}_d)\prod_{s=1}^t\left(1 + \sum_{a\in S_s}\|z_{sa}\|^2_{\Lambda_s^{-1}}\right) \\
&\geqslant \det(\lambda \mathbf{I}_d)\prod_{s=1}^t\left(1 + \min\left\{1, \sum_{a\in S_s}\|z_{sa}\|^2_{\Lambda_s^{-1}}\right\}\right). && \text{(D.11)}
\end{aligned}
$$

Then, using the fact that $a \leqslant 2\log(1+a)$ for any $a \in [0,1]$, we get

$$
\begin{aligned}
\sum_{t=1}^T \min\left\{1, \sum_{a\in S_t}\|z_{ta}\|^2_{\Lambda_t^{-1}}\right\} &\leqslant 2\sum_{t=1}^T \log\left(1 + \min\left\{1, \sum_{a\in S_t}\|z_{ta}\|^2_{\Lambda_t^{-1}}\right\}\right) \\
&\leqslant 2\log\left(\frac{\det(\Lambda_{T+1})}{\det(\lambda \mathbf{I}_d)}\right) && \text{(Eqn. (D.11))} \\
&\leqslant 2d\log\left(1 + \frac{X^2 KT}{d\lambda}\right),
\end{aligned}
$$

where the last inequality holds because

$$
\begin{aligned}
\det(\Lambda_{T+1}) &\leqslant \left(\frac{\lambda_1 + \cdots + \lambda_d}{d}\right)^d && (\lambda_1, \cdots, \lambda_d \text{ are eigenvalues of } \Lambda_{T+1}\text{, AM-GM ineq.)} \\
&= \left(\frac{\operatorname{trace}(\Lambda_{T+1})}{d}\right)^d \\
&= \left(\frac{\lambda d + \sum_{t=1}^T\sum_{a\in S_t}\|z_{ta}\|_2^2}{d}\right)^d \leqslant \left(\lambda + \frac{X^2 KT}{d}\right)^d.
\end{aligned}
$$

This concludes the proof of Lemma D.3. □

D.2.3 Proof of Lemma D.4

Proof of Lemma D.4. Let $W_t := \lambda \mathbf{I}_d + \sum_{s \in \mathcal{T}_0, s<t} \sum_{a \in S_s} z_{sa} z_{sa}^\top + \lambda \mathbf{I}_d$. Then, we have

$$\begin{aligned}
\left(\lambda + \frac{X^2 |\mathcal{T}_0| K}{d}\right)^d &\geqslant \left(\frac{\lambda d + \sum_{t \in \mathcal{T}_0} \sum_{a \in S_t} \|z_{ta}\|_2^2}{d}\right)^d \\
&= \left(\frac{\operatorname{trace}(W_{T+1})}{d}\right)^d \\
&\geqslant \det(W_{T+1}) && \text{(AM-GM ineq.)} \\
&= \det(\lambda \mathbf{I}_d) \prod_{t \in \mathcal{T}_0} \left(1 + \sum_{a \in S_t} \|z_{ta}\|^2_{W_t^{-1}}\right) && \text{(update equality for det.)} \\
&\geqslant \det(\lambda \mathbf{I}_d) \prod_{t \in \mathcal{T}_0} \left(1 + \sum_{a \in S_t} \|z_{ta}\|^2_{\Lambda_t^{-1}}\right) && (W_t \preceq \Lambda_t) \\
&\geqslant \lambda^d (1+L)^{|\mathcal{T}_0|}. && \left(\textstyle\sum_{a \in S_t} \|z_{ta}\|^2_{\Lambda_t^{-1}} \geqslant L \text{ for } t \in \mathcal{T}_0\right)
\end{aligned}$$

Hence, we get

$$\begin{aligned}
|\mathcal{T}_0| &\leqslant \frac{d}{\log(1+L)} \log\left(1 + \frac{X^2 |\mathcal{T}_0| K}{d\lambda}\right) && \text{(D.12)} \\
&= \frac{d}{\log(1+L)} \left(\log\left(\frac{|\mathcal{T}_0|}{2d/\log(1+L)}\right) + \log\left(\frac{2d}{\log(1+L)}\left(\frac{1}{|\mathcal{T}_0|} + \frac{X^2 K}{d\lambda}\right)\right)\right) \\
&\leqslant \frac{|\mathcal{T}_0|}{2} + \frac{d}{\log(1+L)} \log\left(\frac{2d}{e\log(1+L)}\left(\frac{1}{|\mathcal{T}_0|} + \frac{X^2 K}{d\lambda}\right)\right),
\end{aligned}$$

which implies that

$$|\mathcal{T}_0| \leqslant \frac{2d}{\log(1+L)} \log\left(\frac{2d}{e\log(1+L)}\left(\frac{1}{|\mathcal{T}_0|} + \frac{X^2 K}{d\lambda}\right)\right). \tag{D.13}$$

Now, we fix $c > 0$ and consider two cases:

- Case 1: $|\mathcal{T}_0| < cd$
  In this case, from Equation (D.12), we have $|\mathcal{T}_0| \leqslant \frac{d}{\log(1+L)} \log\left(1 + \frac{X^2 c K}{\lambda}\right)$.
- Case 2: $|\mathcal{T}_0| \geqslant cd$
  In this case, from Equation (D.13), we have $|\mathcal{T}_0| \leqslant \frac{2d}{\log(1+L)} \log\left(\frac{2}{e\log(1+L)}\left(\frac{1}{c} + \frac{X^2 K}{\lambda}\right)\right)$.

By setting $c = \frac{2}{e\log(1+L)}$, we obtain

$$|\mathcal{T}_0| \leqslant \frac{2d}{\log(1+L)} \log\left(1 + \frac{X^2 K}{\log(1+L)\lambda}\right),$$

which concludes the proof of Lemma D.4. □

D.2.4 Proof of Lemma D.5

Proof of Lemma D.5. For simplicity, let $\widetilde{\mathcal{T}}_t^w = \{s \in [t-1] \mid s \in \mathcal{T}^w \cap (\mathcal{T}_0)^c\}$. Clearly, $\widetilde{\mathcal{T}}_{T+1}^w = \mathcal{T}^w \cap (\mathcal{T}_0)^c$. Recall that by the definition of H_t, we have

$$
\begin{aligned}
H_t &= \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|} \nabla^2 \ell_s^{(j)}(\widehat{\boldsymbol{\theta}}_s^{(j+1)}) + \lambda \mathbf{I}_d \\
&= \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|} \sum_{k=j}^{|S_s|} \sum_{k'=j}^{|S_s|} \frac{\exp\left(\left(\phi(x_s, \sigma_{sk}) + \phi(x_s, \sigma_{sk'})\right)^\top \widehat{\boldsymbol{\theta}}_s^{(j+1)}\right)}{2\left(\sum_{k'=j}^{|S_s|} \exp\left(\phi(x_s, \sigma_{sk'})^\top \widehat{\boldsymbol{\theta}}_s^{(j+1)}\right)\right)^2} \cdot z_{skk'} z_{skk'}^\top + \lambda \mathbf{I}_d \\
&\succeq \frac{\kappa}{2K^2} \sum_{s=1}^{t-1} \sum_{k=j}^{|S_s|} \sum_{k'=j}^{|S_s|} z_{skk'} z_{skk'}^\top + \lambda \mathbf{I}_d && (\kappa = e^{-4B}) \\
&\succeq \frac{\kappa}{2K^2} \sum_{s=1}^{t-1} \sum_{a \in S_s} \left(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\right)\left(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\right)^\top + \lambda \mathbf{I}_d && (\bar{a}_s \in S_s \text{ by Eqn. (9)}) \\
&\succeq \frac{\kappa}{2K^2} \underbrace{\left(\sum_{s \in \widetilde{\mathcal{T}}_t^w} \sum_{a \in S_s} \left(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\right)\left(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\right)^\top + \lambda \mathbf{I}_d\right)}_{=:\Lambda_t^w}, && \text{(D.14)}
\end{aligned}
$$

where $z_{skk'} = \phi(x_s, \sigma_{sk}) - \phi(x_s, \sigma_{sk'})$.

Let $\tilde{a}_t = \operatorname{argmax}_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}$. Then, we get

$$
\begin{aligned}
&\sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \\
&\qquad \leqslant \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \|\phi(x_t, \tilde{a}_t) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \\
&\qquad \leqslant \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 && (\tilde{a}_t, \bar{a}_t \in S_t \text{ by Eqn. (9)}) \\
&\qquad \leqslant \frac{2K^2}{\kappa} \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{\left(\Lambda_t^w\right)^{-1}}^2 && \text{(Eqn. (D.14))} \\
&\qquad \leqslant \frac{2K^2}{\kappa} \sum_{t=1}^{T} \min\left\{1, \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{\left(\Lambda_t^w\right)^{-1}}^2\right\} && (t \notin \mathcal{T}_0 \text{ and } \widetilde{\mathcal{T}}_{T+1}^w \subseteq [T]) \\
&\qquad \leqslant \frac{4K^2}{\kappa} d \log\left(1 + \frac{2KT}{d\lambda}\right). && \text{(Lemma D.3)}
\end{aligned}
$$

On the other hand, for $t \in \widetilde{\mathcal{T}}_{T+1}^w = \mathcal{T}^w \cap (\mathcal{T}_0)^c$, we know that

$$
\sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \geqslant \frac{|\widetilde{\mathcal{T}}_{T+1}^w|}{3\sqrt{2}\beta_{T+1}(\delta)^2}.
$$

By combining the two results above, we obtain

$$
|\widetilde{\mathcal{T}}_{T+1}^w| = |\mathcal{T}^w \cap (\mathcal{T}_0)^c| \leqslant \frac{12\sqrt{2}K^2}{\kappa} \beta_{T+1}(\delta)^2 d \log\left(1 + \frac{2KT}{d\lambda}\right),
$$

which concludes the proof. □

E Proof of Theorem 2

E.1 Main Proof of Theorem 2

In this section, we present the proof of Theorem 2, which is obtained by using the RB loss (4) instead of the PL loss (3). Note that this approach is based on the concept of *rank breaking* (RB), which decomposes (partial) ranking data into individual pairwise comparisons, treats each comparison as independent, and has been extensively studied in previous works [6, 34, 32, 69]. Moreover, this RB approach is applied in the current RLHF for LLM (e.g., Ouyang et al. [59]) and is also studied theoretically in Zhu et al. [96] under the offline setting.

RB loss and OMD. We begin by recalling the loss function and the parameter update rule. Specifically, we use the PL loss defined in Equation (4) and update the parameter according to Equation (7).

$$\ell_t(\boldsymbol{\theta}) := \sum_{j=1}^{|S_t|-1} \sum_{k=j+1}^{|S_t|} \underbrace{-\log\left(\frac{\exp\left(\phi(x_t,\sigma_{tj})^\top \boldsymbol{\theta}\right)}{\exp\left(\phi(x_t,\sigma_{tj})^\top \boldsymbol{\theta}\right) + \exp\left(\phi(x_t,\sigma_{tk})^\top \boldsymbol{\theta}\right)}\right)}_{=:\ell_t^{(j,k)}(\boldsymbol{\theta})} = \sum_{j=1}^{|S_t|-1} \sum_{k=j+1}^{|S_t|} \ell_t^{(j,k)}(\boldsymbol{\theta}).$$

and

$$\widehat{\boldsymbol{\theta}}_t^{(j,k+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \langle \nabla \ell_t^{(j,k)}(\widehat{\boldsymbol{\theta}}_t^{(j,k)}), \boldsymbol{\theta}\rangle + \frac{1}{2\eta}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^{(j,k)}\|^2_{\tilde{H}_t^{(j,k)}}, \quad 1 \leqslant j < k \leqslant |S_t|,$$

where if $k = |S_t|$, we set $\widehat{\boldsymbol{\theta}}_t^{(j,k+1)} = \widehat{\boldsymbol{\theta}}_t^{(j+1,j+2)}$, and for the final pair, let $\widehat{\boldsymbol{\theta}}_t^{(|S_t|-1,|S_t|+1)} = \widehat{\boldsymbol{\theta}}_{t+1}^{(1,2)}$. Also, the matrix $\tilde{H}_t^{(j,k)}$ is defined as $\tilde{H}_t^{(j,k)} := H_t + \eta \sum_{(j',k')\leqslant(j,k)} \nabla^2 \ell_t^{(j',k')}(\widehat{\boldsymbol{\theta}}_t^{(j',k')})$ [7], where

$$H_t := \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d, \quad \lambda > 0.$$

Online confidence bound for RB loss. Now, we introduce the online confidence bound for RB loss. Since the total number of updates up to round t is $\sum_{s=1}^{t}\binom{|S_s|}{2}$, a modification of Lemma D.1 yields the following result:

Corollary E.1 (Online confidence bound for RB loss)**.** *Let $\delta \in (0,1]$. We set $\eta = (1+3\sqrt{2}B)/2$ and $\lambda = \max\{12\sqrt{2}B\eta, 144\eta d, 2\}$. Then, under Assumption 1, with probability at least $1-\delta$, we have*

$$\|\widehat{\boldsymbol{\theta}}_t^{(j,k)} - \boldsymbol{\theta}^\star\|_{H_t^{(j,k)}} \leqslant \beta_t(\delta) = \mathcal{O}\left(B\sqrt{d\log(tK/\delta)} + B\sqrt{\lambda}\right), \quad \forall t \geqslant 1,\ 1 \leqslant j < k \leqslant |S_t|,$$

where $H_t^{(j,k)} := H_t + \sum_{(j',k')<(j,k)} \nabla^2 \ell_t^{(j',k')}(\widehat{\boldsymbol{\theta}}_t^{(j',k'+1)}) + \lambda \mathbf{I}_d$ and $\widehat{\boldsymbol{\theta}}_t^{(1,2)} = \widehat{\boldsymbol{\theta}}_t$.

Useful definitions. We use the same or similar definitions for the set of *warm-up rounds* $\mathcal{T}^w$ (given in Equation (D.1)), the set of *large elliptical potential (EP) rounds* $\mathcal{T}_0$ (given in Equation (D.3)), and the regularized covariance matrix Λ_t (given in Equation (D.2)).

$$\mathcal{T}^w := \left\{ t \in [T] : \max_{a\in\mathcal{A}} \|\phi(x_t,a) - \phi(x_t,\bar{a}_t)\|_{H_t^{-1}} \geqslant \frac{1}{\beta_{T+1}(\delta)} \right\}, \qquad \text{(warm-up rounds)}$$

$$\mathcal{T}_0 := \left\{ t \in [T] : \sum_{a\in S_t} \|\phi(x_t,a) - \phi(x_t,\bar{a}_t)\|_{\Lambda_t^{-1}} \geqslant 1. \right\}, \qquad \text{(large EP rounds)}$$

$$\Lambda_t := \sum_{s\in[t-1]\setminus\mathcal{T}^w} \sum_{a\in S_s} \left(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\right)\left(\phi(x_s,a) - \phi(x_s,\bar{a}_s)\right)^\top + \lambda \mathbf{I}_d.$$

Key Lemmas. We can avoid the $1/\kappa = \mathcal{O}(e^B)$ dependency in the leading term, thanks to the following lemma.

[7] We write $(j',k') \leqslant (j,k)$ to indicate lexicographic order, i.e., $j' < j$ or $j' = j$ and $k' \leqslant k$.

Lemma E.1. *Let Λ_t be defined as in Equation* (D.2)*. Set $\lambda = \Omega(d \log(KT/\delta))$. Then, for all $t \in [T]$, with probability at least $1 - \delta$, we have*

$$H_t \succeq \frac{1}{10} \Lambda_t.$$

The proof is deferred to Appendix E.2.1.

Lemma E.2. *Let $\mathcal{T}_0 := \{t \in [T] : \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{\Lambda_t^{-1}} \geqslant 1.\}$ and $\mathcal{T}^w = \{t \in [T] : \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}} \geqslant \frac{1}{\beta_{T+1}(\delta)}\}$. Define $\kappa := \frac{e^{-4B}}{4}$. Then, the size of the set $\mathcal{T}^w \cap (\mathcal{T}_0)^c$ is bounded as follows:*

$$|\mathcal{T}^w \cap (\mathcal{T}_0)^c| \leqslant \frac{2}{\kappa} \beta_{T+1}(\delta)^2 d \log\left(1 + \frac{2KT}{d\lambda}\right).$$

The proof is deferred to Appendix E.2.2.

We are now ready to provide the proof of Theorem 2.

Proof of Theorem 2. The overall proof structure is similar to that of Theorem 1. We begin with Equation (D.5), but apply Lemma E.2 instead of Lemma D.5. With probability at least $1 - \delta$, we have

$$\begin{aligned}
\textbf{SubOpt}(T) &= \mathbb{E}_{x \sim \rho}\left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x))\right)^\top \boldsymbol{\theta}^\star\right] \\
&\leqslant \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right) + \frac{8B}{\log(2)T} d \log\left(1 + \frac{2K}{\log(2)\lambda}\right) + \frac{8B}{\kappa T} \beta_{T+1}(\delta)^2 d \log\left(1 + \frac{2KT}{d\lambda}\right) \\
&\hspace{30em} \text{(Lemma E.2)} \\
&\quad + \frac{1}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\phi(x_t, \pi^\star(x_t)) - \phi(x_t, \widehat{\pi}_T(x_t))\right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\right). \qquad \text{(E.1)}
\end{aligned}$$

To further bound the last term of Equation (E.1), by following the same logic from Equation (D.5) to Equation (D.6), with probability at least $1 - \delta$, we obtain

$$\begin{aligned}
&\frac{1}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\phi(x_t, \pi^\star(x_t)) - \phi(x_t, \widehat{\pi}_T(x_t))\right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1}\right) \\
&\qquad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \frac{1}{|S_t^\star|} \sum_{a \in S_t^\star} \|\phi(x_t, a) - \phi(x_t, \widehat{\pi}_T(x_t))\|_{H_t^{-1}} \\
&\hspace{30em} (S_t^\star := \{\pi^\star(x_t), \widehat{\pi}_T(x_t)\}) \\
&\qquad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \frac{1}{|S_t|} \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}. \\
&\hspace{30em} (S_t \text{ selection rule, Eqn. (9)})
\end{aligned}$$

To further bound the right-hand side, by applying the Cauchy-Schwartz inequality, we get

$$
\begin{aligned}
&\frac{2\beta_{T+1}(\delta)}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \frac{1}{|S_t|} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{H_t^{-1}} \\
&\quad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\frac{1}{|S_t|}\right)^2 |S_t|} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{H_t^{-1}}^2} \\
&\hspace{22em} \text{(Cauchy-Schwartz ineq.)} \\
&\quad = \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \left(\frac{1}{|S_t|}\right)^2 |S_t|} \sqrt{10 \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{\Lambda_t^{-1}}^2} \\
&\hspace{16em} \text{(Lemma E.1, with probability at least } 1-\delta) \\
&\quad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^{T} \frac{1}{|S_t|}} \sqrt{10 \sum_{t \notin \mathcal{T}^w} \min\left\{1, \sum_{a \in S_t} \left\| \phi\left(x_t, a\right) - \phi\left(x_t, \bar{a}_t\right) \right\|_{\Lambda_t^{-1}}^2\right\}} \\
&\hspace{20em} (t \notin \mathcal{T}_0 \text{ and } \mathcal{T}_0 \cup \mathcal{T}^w \subseteq [T]) \\
&\quad \leqslant \frac{2\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^{T} \frac{1}{|S_t|}} \sqrt{20 d \log\left(1 + \frac{2K(T - |\mathcal{T}^w|)}{d\lambda}\right)} \qquad \text{(Lemma D.3)} \\
&\quad = \mathcal{O}\left(\frac{\beta_{T+1}(\delta)}{T} \sqrt{\sum_{t=1}^{T} \frac{1}{|S_t|}} \cdot \sqrt{d \log\left(KT\right)}\right). \qquad \text{(E.2)}
\end{aligned}
$$

By plugging Equation (E.2) into Equation (E.1) and setting $\beta_{T+1}(\delta) = \mathcal{O}\big(B\sqrt{d\log(KT)} + B\sqrt{\lambda}\big)$, then with probability at least $1 - 3\delta$, we derive that

$$
\textbf{SubOpt}(T) = \tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T} \frac{1}{|S_t|}} + \frac{d^2}{\kappa T}\right).
$$

Substituting $\delta \leftarrow \frac{\delta}{3}$, we conclude the proof of Theorem 2. $\square$

E.2 Proofs of Lemmas for Theorem 2

E.2.1 Proof of Lemma E.1

Proof of Lemma E.1. Recall that, under the Bradley–Terry-Luce (BTL) model defined in Equation (2), the probability that action a is preferred over action a' is given by:

$$
\mathbb{P}\left(a \succ a' | x_t, ; \boldsymbol{\theta}\right) = \frac{\exp\left(\phi(x_t, a)^\top \boldsymbol{\theta}\right)}{\exp\left(\phi(x_t, a)^\top \boldsymbol{\theta}\right) + \exp\left(\phi(x_t, a')^\top \boldsymbol{\theta}\right)} = \mu\left(\left(\phi(x_t, a) - \phi(x_t, a')\right)^\top \boldsymbol{\theta}\right).
$$

Then, we can derive a lower bound on the matrix H_t as follows:

$$
\begin{aligned}
H_t &= \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d \\
&\geq \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d \\
&= \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \dot{\mu}\left(z_{sjk}^\top \widehat{\boldsymbol{\theta}}_s^{(j,k+1)}\right) z_{sjk} z_{sjk}^\top + \lambda \mathbf{I}_d \\
&\geq \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \dot{\mu}\left(z_{sjk}^\top \boldsymbol{\theta}^\star\right) e^{-\left|z_{sjk}^\top\left(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)} - \boldsymbol{\theta}^\star\right)\right|} z_{sjk} z_{sjk}^\top + \lambda \mathbf{I}_d \qquad \text{(Lemma F.2)} \\
&\geq e^{-1} \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \dot{\mu}\left(z_{sjk}^\top \boldsymbol{\theta}^\star\right) z_{sjk} z_{sjk}^\top + \lambda \mathbf{I}_d, \qquad \text{(E.3)}
\end{aligned}
$$

where the last inequality holds because, for any $s \notin \mathcal{T}^w$, the following property is satisfied:

$$
\begin{aligned}
\left|z_{sjk}^\top\big(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)} - \boldsymbol{\theta}^\star\big)\right| &= \left|\big(\phi(x_s, \sigma_{sj}) - \phi(x_s, \sigma_{sk})\big)^\top \left(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)} - \boldsymbol{\theta}^\star\right)\right| \\
&\leqslant \big\|\phi(x_s, \sigma_{sj}) - \phi(x_s, \sigma_{sk})\big\|_{H_s^{-1}} \big\|\widehat{\boldsymbol{\theta}}_s^{(j,k+1)} - \boldsymbol{\theta}^\star\big\|_{H_s} \qquad \text{(Hölder's inequality)} \\
&\leqslant \frac{1}{\beta_{T+1}(\delta)} \big\|\widehat{\boldsymbol{\theta}}_s^{(j,k+1)} - \boldsymbol{\theta}^\star\big\|_{H_s^{(j,k+1)}} \qquad (s \neq \mathcal{T}^w, H_s \preceq H_s^{(j,k+1)}) \\
&\leqslant \frac{\beta_t(\delta)}{\beta_{T+1}(\delta)} \qquad \text{(Corollary E.1)} \\
&\leqslant 1. \qquad (\beta_t(\delta) \text{ is non-decreasing})
\end{aligned}
$$

For simplicity, we write $\mathbb{P}_s(a \succ a') = \mathbb{P}(a \succ a' | x_s; \boldsymbol{\theta}^\star)$. Let $P_{s,\{a,a'\}}$ denote the Bernoulli distribution over the support $\{a, a'\}$, where a occurs with probability $\mu\big(\big(\phi(x_s, a) - \phi(x_s, a')\big)^\top \boldsymbol{\theta}^\star\big)$. Then, to further lower bound the right-hand side of Equation (E.3), we proceed as follows:

$$
\begin{aligned}
&\sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \dot{\mu}\left(z_{sjk}^\top \boldsymbol{\theta}^\star\right) z_{sjk} z_{sjk}^\top \\
&= \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \mu\left(z_{sjk}^\top \boldsymbol{\theta}^\star\right) \mu\left(z_{skj}^\top \boldsymbol{\theta}^\star\right) z_{sjk} z_{sjk}^\top \\
&= \frac{1}{2} \sum_{a \in S_s} \sum_{a' \in S_s} \mathbb{P}_s(a \succ a') \mathbb{P}_s(a' \succ a) \big(\phi(x_s, a) - \phi(x_s, a')\big)\big(\phi(x_s, a) - \phi(x_s, a')\big)^\top \\
&\geq \frac{1}{2} \sum_{a \in S_s} 2\mathbb{P}_s(a \succ \bar{a}_s) \mathbb{P}_s(\bar{a}_s \succ a) \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top \\
&\qquad (\bar{a}_s \in S_s \text{ by Eqn. (9)}) \\
&= \frac{1}{2} \sum_{a \in S_s} \mathbb{E}_{(a', a'') \sim P_{s,\{a,\bar{a}_s\}}^{\otimes 2}} \left[\big(\phi(x_s, a') - \phi(x_s, a'')\big)\big(\phi(x_s, a') - \phi(x_s, a'')\big)^\top\right], \qquad \text{(E.4)}
\end{aligned}
$$

where $P_{s,\{a,\bar{a}_s\}}^{\otimes 2} = P_{s,\{a,\bar{a}_s\}} \times P_{s,\{a,\bar{a}_s\}}$ denotes the the product distribution over two independent samples from $P_{s,\{a,\bar{a}_s\}}$. Note that the Bernoulli distribution $P_{s,\{a,\bar{a}_s\}}$, where $a \in S_s$, is measurable with respect to the filtration $\mathcal{F}_{s-1} = \boldsymbol{\sigma}\left(S_1, \sigma_1, \ldots, S_{s-1}, \sigma_{s-1}, S_s\right)$. Then, plugging Equation (E.4) into Equation (E.3), we get

$$
\begin{aligned}
&H_t \\
&\succeq \frac{e^{-1}}{2} \sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{a \in S_s} \mathbb{E}_{(a', a'') \sim P^{\otimes 2}_{s, \{a, \bar{a}_s\}}} \left[\big(\phi(x_s, a') - \phi(x_s, a'')\big)\big(\phi(x_s, a') - \phi(x_s, a'')\big)^\top \right] + \lambda \mathbf{I}_d \\
&\succeq \frac{3e^{-1}}{10} \underbrace{\left(\sum_{s \in [t-1] \setminus \mathcal{T}^w} \sum_{a \in S_s} \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top + \lambda \mathbf{I}_d \right)}_{=\Lambda_t} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(covariance matrix concentration lemma (Corollary F.1))} \\
&\succeq \frac{1}{10} \Lambda_t,
\end{aligned}
$$

which conclude the proof of Lemma E.1. □

E.2.2 Proof of Lemma E.2

Proof of Lemma E.2. For simplicity, let $\widetilde{\mathcal{T}}_t^w = \{s \in [t-1] \mid s \in \mathcal{T}^w \cap (\mathcal{T}_0)^c\}$. Clearly, $\widetilde{\mathcal{T}}_{T+1}^w = \mathcal{T}^w \cap (\mathcal{T}_0)^c$. Recall that by the definition of H_t, we have

$$
\begin{aligned}
H_t &= \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} \nabla^2 \ell_s^{(j,k)}(\widehat{\boldsymbol{\theta}}_s^{(j,k+1)}) + \lambda \mathbf{I}_d \\
&\succeq \kappa \sum_{s=1}^{t-1} \sum_{j=1}^{|S_s|-1} \sum_{k=j+1}^{|S_s|} z_{sjk} z_{sjk}^\top + \lambda \mathbf{I}_d && (\kappa = e^{-4B}/4) \\
&\succeq \kappa \sum_{s=1}^{t-1} \sum_{a \in S_s} \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top + \lambda \mathbf{I}_d && (\bar{a}_s \in S_s) \\
&\succeq \kappa \underbrace{\left(\sum_{s \in \widetilde{\mathcal{T}}_t^w} \sum_{a \in S_s} \big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)\big(\phi(x_s, a) - \phi(x_s, \bar{a}_s)\big)^\top + \lambda \mathbf{I}_d \right)}_{=:\Lambda_t^w}. && \text{(E.5)}
\end{aligned}
$$

Let $\tilde{a}_t = \operatorname{argmax}_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}$. Then, we get

$$
\begin{aligned}
&\sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \\
&\qquad \leqslant \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \|\phi(x_t, \tilde{a}_t) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \\
&\qquad \leqslant \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 && (\tilde{a}_t, \bar{a}_t \in S_t \text{ by Eqn. (9)}) \\
&\qquad \leqslant \frac{1}{\kappa} \sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{(\Lambda_t^w)^{-1}}^2 && \text{(Eqn. (E.5))} \\
&\qquad \leqslant \frac{1}{\kappa} \sum_{t=1}^{T} \min\left\{1, \sum_{a \in S_t} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{(\Lambda_t^w)^{-1}}^2 \right\} && (t \neq \mathcal{T}_0 \text{ and } \widetilde{\mathcal{T}}_{T+1}^w \subseteq [T]) \\
&\qquad \leqslant \frac{2}{\kappa} d \log\left(1 + \frac{2KT}{d\lambda}\right). && \text{(Lemma D.3)}
\end{aligned}
$$

On the other hand, for $t \in \widetilde{\mathcal{T}}_{T+1}^w = \mathcal{T}^w \cap (\mathcal{T}_0)^c$, we know that

$$
\sum_{t \in \widetilde{\mathcal{T}}_{T+1}^w} \max_{a \in \mathcal{A}} \|\phi(x_t, a) - \phi(x_t, \bar{a}_t)\|_{H_t^{-1}}^2 \geqslant \frac{|\widetilde{\mathcal{T}}_{T+1}^w|}{\beta_{T+1}(\delta)^2}.
$$

By combining the two results above, we get

$$|\widetilde{\mathcal{T}}_{T+1}^{w}| \leqslant \frac{2}{\kappa}\beta_{T+1}(\delta)^2 d\log\left(1+\frac{2KT}{d\lambda}\right),$$

which concludes the proof. □

F Technical Lemmas

Lemma F.1 (Proposition B.5 of Lee and Oh 43)**.** *The Hessian of the multinomial logistic loss $\bar{\ell}:\mathbb{R}^M\to\mathbb{R}$ satisfies that, for any $\mathbf{a}_1,\mathbf{a}_2\in\mathbb{R}^M$, we have:*

$$e^{-3\sqrt{2}\|\mathbf{a}_1-\mathbf{a}_2\|_\infty}\nabla^2\bar{\ell}(\mathbf{a}_1)\preceq\nabla^2\bar{\ell}(\mathbf{a}_2)\preceq e^{3\sqrt{2}\|\mathbf{a}_1-\mathbf{a}_2\|_\infty}\nabla^2\bar{\ell}(\mathbf{a}_1).$$

Lemma F.2 (Lemma 9 of Abeille et al. 2)**.** *Let f be a strictly increasing function such that $|\ddot{f}|\leqslant\dot{f}$, and let $\mathcal{Z}$ be any bounded interval of $\mathbb{R}$. Then, for all $z_1,z_2\in\mathcal{Z}$, we have*

$$\dot{f}(z_2)\exp\left(-|z_2-z_1|\right)\leqslant\dot{f}(z_1)\leqslant\dot{f}(z_2)\exp\left(|z_2-z_1|\right).$$

Lemma F.3 (Concentration of covariances, Lemma 39 of Zanette et al. 92)**.** *Let μ_i be the conditional distribution of $\phi\in\mathbb{R}^d$ given the sampled $\phi_1,\ldots,\phi_{i-1}$. Assume $\|\phi\|_2\leqslant 1$. Define $\Sigma=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\phi\sim\mu_i}\phi\phi^\top$. If $\lambda=\Omega(d\log(n/\delta))$, then, with probability at least $1-\delta$, for any $n\geqslant 1$, we have*

$$\frac{1}{3}\left(n\Sigma+\lambda\mathbf{I}_d\right)\preceq\sum_{i=1}^n\phi_i\phi_i^\top+\lambda\mathbf{I}_d\preceq\frac{5}{3}\left(n\Sigma+\lambda\mathbf{I}_d\right).1$$

We now extend the previous sequence $\phi_1,\ldots,\phi_{i-1}$ to a general setting where conditioning is performed with respect to an arbitrary filtration $\mathcal{F}_{i-1}$. For instance, $\mathcal{F}_{i-1}$ may be the σ-algebra generated by previous samples $\tilde{\phi}_1,\ldots,\tilde{\phi}_{i-1}$, where each $\tilde{\phi}_j$ is drawn from a potentially different distribution $\tilde{\mu}_j$. This generalization is valid because a martingale difference sequence can be defined with respect to any filtration provided that the σ-algebras satisfy the usual properties (e.g., nestedness and making $\{\phi_i\}$ adapted).

Corollary F.1 (Generalized version of covariance concentration)**.** *Let μ_i denote the conditional distribution of $\phi\in\mathbb{R}^d$ conditioned on the filtration $\mathcal{F}_{i-1}$. Assume $\|\phi\|_2\leqslant 1$. Define $\Sigma=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\phi\sim\mu_i}\phi\phi^\top$. If $\lambda=\Omega(d\log(n/\delta))$, then with probability at least $1-\delta$, for any $n\geqslant 1$, we have*

$$\frac{1}{3}\left(n\Sigma+\lambda\mathbf{I}_d\right)\preceq\sum_{i=1}^n\phi_i\phi_i^\top+\lambda\mathbf{I}_d\preceq\frac{5}{3}\left(n\Sigma+\lambda\mathbf{I}_d\right).$$

We also provide a concentration lemma for the more general case of positive semi-definite (PSD) random matrices.

Lemma F.4 (Concentration of PSD matrices)**.** *Let μ_i denote the conditional distribution of a positive semi-definite $M\in\mathbb{R}^{d\times d}$ conditioned on the filtration $\mathcal{F}_{i-1}$. Assume $\lambda_{\max}(M)\leqslant 1$. Define $\overline{M}:=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{M\sim\mu_i}M$. If $\lambda=\Omega(d\log(n/\delta))$, then with probability at least $1-\delta$, for any $n\geqslant 1$, we have*

$$\frac{1}{3}\left(n\overline{M}+\lambda\mathbf{I}_d\right)\preceq\sum_{i=1}^n M_i+\lambda\mathbf{I}_d\preceq\frac{5}{3}\left(n\overline{M}+\lambda\mathbf{I}_d\right).$$

Proof of Lemma F.4. The overall structure of the proof closely follows that of Lemma 39 in Zanette et al. 92. For completeness, we provide the full proof below.

Fix $x\in\mathbb{R}^d$ such that $\|x\|_2=1$. Let $\overline{M}_i=\mathbb{E}_{M\sim\mu_i}M$ and $\overline{M}=\frac{1}{n}\sum_{i=1}^n\overline{M}_i$. Then, we have

$$\mathbb{E}_{M\sim\mu_i}x^\top Mx=x^\top\mathbb{E}_{M\sim\mu_i}Mx=x^\top\overline{M}_ix.$$

Since M is a positive semi-definite matrix, the random variable $x^\top M x$ is non-negative with the maximum value $x^\top M x \leqslant \lambda_{\max}(M)\|x\|_2^2 \leqslant 1$. Thus, the conditional variance is at most $x^\top \overline{M}_i x$ because

$$\mathrm{Var}_{M\sim\mu_i}(x^\top M x) \leqslant \mathbb{E}_{M\sim\mu_i}(x^\top M x)^2 \leqslant \mathbb{E}_{M\sim\mu_i} x^\top M x = x^\top \overline{M}_i x.$$

Then, by Lemma F.5, with probability at least $1-\delta$, for some constant c, we have

$$\left|\frac{1}{n}\sum_{i=1}^n \left(x^\top M_i x - x^\top \overline{M}_i x\right)\right| = \left|\frac{1}{n}\sum_{i=1}^n x^\top M_i x - x^\top \overline{M} x\right| \leqslant c\left(\sqrt{2\frac{x^\top \overline{M} x}{n}\log(2/\delta)} + \frac{\log(2/\delta)}{3n}\right).$$

Now, we will show that if $\lambda = \Omega(\log(1/\delta))$, we can derive

$$c\left(\sqrt{2\frac{x^\top \overline{M} x}{n}\log(2/\delta)} + \frac{\log(2/\delta)}{3n}\right) \leqslant \frac{1}{2}\left(x^\top \overline{M} x + \frac{\lambda}{n}\right). \tag{F.1}$$

Case 1. $x^\top \overline{M} x \leqslant \frac{\lambda}{n}$.
In this case, it is sufficient to satisfy for some constants c', c''

$$\begin{aligned}
\sqrt{2\frac{\log(2/\delta)}{n}} \leqslant c'\sqrt{\frac{\lambda}{n}} &\quad\longleftrightarrow\quad \Omega(\log(1/\delta)) \leqslant \lambda \\
\frac{\log(2/\delta)}{3n} \leqslant c''\left(\frac{\lambda}{n}\right) &\quad\longleftrightarrow\quad \Omega(\log(1/\delta)) \leqslant \lambda.
\end{aligned}$$

Case 2. $x^\top \overline{M} x > \frac{\lambda}{n}$.
In this case, it is sufficient to satisfy for some constants c''', c'''

$$\begin{aligned}
\sqrt{2\frac{x^\top \overline{M} x}{n}\log(2/\delta)} \leqslant c'''\left(\frac{\lambda}{n}\right) &\quad\longleftrightarrow\quad \Omega(\log(1/\delta)) \leqslant \lambda \\
\frac{\log(2/\delta)}{3n} \leqslant c''''\left(\frac{\lambda}{n}\right) &\quad\longleftrightarrow\quad \Omega(\log(1/\delta)) \leqslant \lambda.
\end{aligned}$$

Therefore, Equation (F.1) is satisfied. Since $\|x\|_2 \leqslant 1$, this implies

$$\left|x^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)x\right| \leqslant \frac{1}{2}x^\top\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right)x. \tag{F.2}$$

We denote the boundary of the unit ball by $\partial\mathcal{B} = \{\|x\|_2 = 1\}$. Then, for any $x \in \partial\mathcal{B}$, we know there exists a x' in the ϵ-covering such that $\|x - x'\|_2 \leqslant \epsilon$. Let $\mathcal{N}_\epsilon$ be the ϵ-covering number of $\partial\mathcal{B}$. Then, by the covering number of Euclidean ball lemma (Lemma F.6), we get

$$\mathcal{N}_\epsilon \leqslant \left(\frac{3}{\epsilon}\right)^d. \tag{F.3}$$

Taking a union bound over x' and the number of samples n, with probability at least $1 - n\mathcal{N}_\epsilon\delta$, we obtain

$$\begin{aligned}
\left|x^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)x\right| &\leqslant \left|(x')^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)x'\right| + \left|(x-x')^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)x'\right| \\
&\quad + \left|(x')^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)(x-x')\right| \\
&\leqslant \left|(x')^\top\left(\frac{1}{n}\sum_{i=1}^n M_i - \overline{M}\right)x'\right| + 4\epsilon. \\
&\qquad\qquad (\|x-x'\|_2 \leqslant \epsilon \text{ and } M_i, \|\overline{M}\|_2 \leqslant 1) \\
&\leqslant \frac{1}{2}(x')^\top\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right)x' + 4\epsilon \qquad \text{(Eqn. (F.2))} \\
&\leqslant \frac{1}{2}x^\top\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right)x + \frac{9}{2}\epsilon \qquad (\|x-x'\|_2 \leqslant \epsilon \text{ and } \|\overline{M}\|_2 \leqslant 1) \\
&\leqslant \frac{2}{3}x^\top\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right)x, \qquad (\text{set } \epsilon = \mathcal{O}(\tfrac{1}{n}))
\end{aligned}$$

where $\lambda = \Omega\left(\log\left(\frac{2n\mathcal{N}_\epsilon}{\delta}\right)\right)$. By substituting $\delta \leftarrow \delta/(n\mathcal{N}_\epsilon+1)$ and combining this with Equation (F.3), we obtain:

$$\frac{1}{3}\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right) \preceq \frac{1}{n}\sum_{i=1}^{n} M_i + \frac{\lambda}{n}\mathbf{I}_d \preceq \frac{5}{3}\left(\overline{M} + \frac{\lambda}{n}\mathbf{I}_d\right),$$

which concludes the proof. $\square$

Lemma F.5 (Bernstein for martingales, Theorem 1 of Beygelzimer et al. 10 and Lemma 45 of Zanette et al. 92)**.** *Consider the stochastic process $\{X_n\}$ adapted to the filtration $\{\mathcal{F}_n\}$. Assume $\mathbb{E}X_n = 0$ and $cX_n \leqslant 1$ for every n; then for every constant $z \neq 0$ it holds that*

$$\Pr\left(\sum_{n=1}^{N} X_n \leqslant z\sum_{n=1}^{N}\mathbb{E}(X_n^2 \mid \mathcal{F}_n) + \frac{1}{z}\log\frac{1}{\delta}\right) \geqslant 1-\delta.$$

By optimizing the bound as a function of z, we also have

$$\Pr\left(\sum_{n=1}^{N} X_n \leqslant c\sqrt{\sum_{n=1}^{N}\mathbb{E}(X_n^2 \mid \mathcal{F}_n)\log\frac{1}{\delta}} + \log\frac{1}{\delta}\right) \geqslant 1-\delta.$$

Lemma F.6 (Covering number of Euclidean ball)**.** *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^2$.*

G Proof of Theorem 3

G.1 Main Proof of Theorem 3

Throughout the proof, we consider the setting where the context space is a singleton, i.e., $\mathcal{X} = \{x\}$. As a result, the problem reduces to a context-free setting, and we focus solely on the action space $\mathcal{A}$. Note that this is equivalent to assuming that ρ is a Dirac distribution.

We first present the following theorem, which serves as the foundation for our analysis.

Theorem G.1 (Lower bound on adaptive PL model parameter estimation)**.** *Let $\Phi = \mathcal{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, and let $\Theta = \{-\mu, \mu\}^d$ for some $\mu \in (0, 1/\sqrt{d}]$. We consider a query model where, at each round $t = 1, \ldots, T$, the learner selects a subset $S_t \subseteq \Phi$ of feature vectors, with cardinality satisfying $2 \leqslant |S_t| \leqslant K$, and then receives a ranking feedback σ_t drawn from the Plackett–Luce (PL) model defined as:*

$$\mathbb{P}(\sigma_t | S_t; \boldsymbol{\theta}) = \prod_{j=1}^{|S_t|}\frac{\exp\left(\phi_{\sigma_{tj}}^\top\boldsymbol{\theta}\right)}{\sum_{k=j}^{|S_t|}\exp\left(\phi_{\sigma_{tk}}^\top\boldsymbol{\theta}\right)},$$

where $\sigma_t = (\sigma_{t1}, \ldots, \sigma_{t|S_t|})$ is a permutation of the actions in S_t, $\phi_a \in \Phi$ denotes the feature vector associated with action $a \in \mathcal{A}$ in the selected subset at round t, and $\boldsymbol{\theta} \in \Theta$. Then, we have

$$\inf_{\widehat{\boldsymbol{\theta}},\pi}\max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right] \geqslant \frac{d\mu^2}{2}\left(1 - \sqrt{\frac{2K^2T\mu^2}{d}}\right),$$

where the infimum is over all measurable estimators $\widehat{\boldsymbol{\theta}}$ and measurable (but possibly adaptive) query rules π, and $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ denotes the expectation over the randomness in the observations and decision rules if $\boldsymbol{\theta}$ is the true instance. In particular, if $T \geqslant \frac{d^2}{8K^2}$, by choosing $\mu = \sqrt{d/(8K^2T)}$, we obtain

$$\inf_{\widehat{\boldsymbol{\theta}},\pi}\max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right] \geqslant \frac{d^2}{32K^2T}.$$

Proof of Theorem G.1. The analysis of this result closely follows the proof of Theorem 3 in Shamir [75]. The key distinction lies in the input structure: our setting involves a set of feature vectors, while theirs is restricted to a single feature vector.

To begin with, since the worst-case expected regret with respect to $\boldsymbol{\theta}$ can be lower bounded by the average regret under the uniform prior over Θ, we have:

$$
\begin{aligned}
\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right\|_2^2\right] &\geqslant \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right\|_2^2\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d \left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)^2\right] \\
&\geqslant \mu^2 \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d \mathbb{I}\left\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i < 0\right\}\right]. \qquad \text{(G.1)}
\end{aligned}
$$

As in Shamir [75], we assume that the query strategy is deterministic conditioned on the past: that is, S_t is a deterministic function of the previous queries and observations, i.e., $S_1, \sigma_1, \ldots, S_{t-1}, \sigma_{t-1}$. This assumption is made without loss of generality, since any randomized querying strategy can be viewed as a distribution over deterministic strategies. Therefore, a lower bound that holds uniformly for all deterministic strategies also applies to any randomized strategy. Then, we use the following lemma.

Lemma G.1 (Lemma 4 of Shamir 75)**.** *Let $\boldsymbol{\theta}$ be a random vector, none of whose coordinates is supported on 0, and let $y_1, y_2, \ldots, y_T$ be a sequence of queries obtained by a deterministic strategy returning a point $\widehat{\boldsymbol{\theta}}$ (that is, $\boldsymbol{\psi}_t$ is a deterministic function of $\boldsymbol{\psi}_1, y_1, \ldots, \boldsymbol{\psi}_{t-1}, y_{t-1}$, and $\widehat{\boldsymbol{\theta}}$ is a deterministic function of $y_1, \ldots, y_T$). Then, we have*

$$
\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d \mathbb{I}\left\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i < 0\right\}\right] \geqslant \frac{d}{2}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^d\sum_{t=1}^T U_{ti}}\right),
$$

where

$$
U_{ti} := \sup_{\boldsymbol{\theta}_j, j\neq i} D_{KL}\big(P\big(y_t|\boldsymbol{\theta}_i>0, \{\boldsymbol{\theta}_j\}_{j\neq i}, \{y_s\}_{s=1}^{t-1}\big) \,\|\, P\big(y_t|\boldsymbol{\theta}_i<0, \{\boldsymbol{\theta}_j\}_{j\neq i}, \{y_s\}_{s=1}^{t-1}\big)\big).
$$

In our setting, we interpret $y_t = \sigma_t$, and $\boldsymbol{\psi}_t = \{\phi_a\}_{a\in S_t} \subseteq \Phi$. Then, we can write U_{ti} as follows:

$$
U_{ti} = \sup_{\boldsymbol{\theta}_j, j\neq i} D_{\mathrm{KL}}(\mathbb{P}(\sigma_t|S_t;\boldsymbol{\theta}_i>0,\{\boldsymbol{\theta}_j\}_{j\neq i},) \,\|\, \mathbb{P}(\sigma_t|S_t;\boldsymbol{\theta}_i<0,\{\boldsymbol{\theta}_j\}_{j\neq i},)).
$$

For simplicity, let $\mathbb{P}_{\boldsymbol{\theta}}(\sigma|S) = \mathbb{P}(\sigma|S;\boldsymbol{\theta})$. Then, we can upper bound U_{ti} using the following lemma.

Lemma G.2. *For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, let $\mathbb{P}_{\boldsymbol{\theta}}(\cdot \mid S)$ denote the PL distribution over rankings induced by the action set S and parameter vector $\boldsymbol{\theta}$. Then, we have*

$$
D_{KL}\big(\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)\|\mathbb{P}_{\boldsymbol{\theta}'}(\cdot|S)\big) \leqslant \frac{K}{2}\sum_{a\in S}\left(\phi_a^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2.
$$

The proof is deferred to Appendix G.2.1.

By applying Lemma G.2, we have

$$
\begin{aligned}
\sum_{i=1}^d U_{ti} &\leqslant \frac{K}{2}\sum_{i=1}^d\sum_{a\in S_t}(2\mu\cdot[\phi_a]_i)^2 = 2K\mu^2\sum_{a\in S_t}\underbrace{\sum_{i=1}^d([\phi_a]_i)^2}_{=1} \\
&= 2K\mu^2\cdot|S_t| && (\phi_a\in\mathcal{S}^{d-1}) \\
&\leqslant 2K^2\mu^2. && (|S_t|\leqslant K)
\end{aligned}
$$

Hence, by Lemma G.1, we get

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{Unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^d \mathbb{I}\left\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i < 0\right\}\right] &\geqslant \frac{d}{2}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^d\sum_{t=1}^T U_{ti}}\right) \\
&\geqslant \frac{d}{2}\left(1-\sqrt{\frac{2K^2T\mu^2}{d}}\right). \qquad \text{(G.2)}
\end{aligned}
$$

Combining Equation (G.1) and (G.2), we prove the first inequality of Theorem G.1. The second inequality directly follows by choosing $\mu = \sqrt{d/(8K^2T)}$. $\square$

We are now ready to present the proof of Theorem 3.

Proof of Theorem 3. The structure of our proof is similar to that of Theorem 2 in Wagenmaker et al. [81]. However, while they consider the linear bandit setting, we focus on the Plackett–Luce (PL) bandit setting.

We adopt the same instance construction as in Theorem G.1, where $\Phi = \mathcal{S}^{d-1}$ and $\Theta = \{-\mu, \mu\}^d$. Define $\phi^\star(\boldsymbol{\theta}) = \operatorname{argmax}_{a \in \mathcal{A}} \phi_a^\top \boldsymbol{\theta}$. Then, since $\phi^\star(\boldsymbol{\theta}) \in \Phi$ and $\boldsymbol{\theta} \in \Phi$, it is clear that

$$\phi^\star(\boldsymbol{\theta}) = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 = \boldsymbol{\theta}/(\sqrt{d}\mu), \quad \text{and } \phi^\star(\boldsymbol{\theta})^\top \boldsymbol{\theta} = \sqrt{d}\mu. \tag{G.3}$$

Fix the suboptimality gap $\epsilon > 0$. By definition, a policy $\pi \in \triangle_\Phi$ is said to be ϵ-optimal if it satisfies

$$\mathbb{E}_{\phi\sim\pi}\left[\phi^\top \boldsymbol{\theta}\right] = \big(\underbrace{\mathbb{E}_{\phi\sim\pi}\left[\phi\right]}_{=:\phi_\pi}\big)^\top \boldsymbol{\theta} \geqslant \phi^\star(\boldsymbol{\theta})^\top \boldsymbol{\theta} - \epsilon = \sqrt{d}\mu - \epsilon. \tag{G.4}$$

Moreover, by Jensen's inequality, we have

$$\|\phi_\pi\|_2^2 \leqslant \mathbb{E}_{\phi\sim\pi}\left[\|\phi\|_2^2\right] = 1.$$

Let $\Delta = \phi_\pi - \phi^\star(\boldsymbol{\theta})$. Then, we get

$$\begin{aligned} & 1 \geqslant \|\phi_\pi\|_2^2 = \|\phi^\star(\boldsymbol{\theta}) + \Delta\|_2^2 = 1 + \|\Delta\|_2^2 + 2\phi^\star(\boldsymbol{\theta})^\top \Delta \\ \iff & \phi^\star(\boldsymbol{\theta})^\top \Delta \leqslant -\frac{1}{2}\|\Delta\|_2^2 \\ \iff & \boldsymbol{\theta}^\star \Delta \leqslant -\frac{\sqrt{d}\mu}{2}\|\Delta\|_2^2. && \text{(Eqn. (G.3))} \end{aligned}$$

Hence, if a policy π is ϵ-optimal for a parameter $\boldsymbol{\theta}$, then the following bound holds:

$$\begin{aligned} & -\epsilon \leqslant -\frac{\sqrt{d}\mu}{2}\|\Delta\|_2^2. && \text{(Eqn. (G.4))} \\ \iff & \|\Delta\|_2^2 \leqslant \frac{2\epsilon}{\sqrt{d}\mu}, \quad \text{where } \boldsymbol{\theta} = \sqrt{d}\mu(\phi_\pi - \Delta). \end{aligned}$$

We now assume that we are given an ϵ-optimal policy $\widehat{\pi}$. Define $\widehat{\phi} := \phi_{\widehat{\pi}}$ and the following estimator

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \boldsymbol{\theta}' & \text{if } \exists \boldsymbol{\theta}' \in \Theta \text{ with } \boldsymbol{\theta}' = \sqrt{d}\mu(\widehat{\phi} - \Delta') \text{ for some } \Delta' \in \mathbb{R}^d, \|\Delta'\|_2^2 \leqslant \frac{2\epsilon}{\sqrt{d}\mu}; \\ \text{any } \boldsymbol{\theta}' \in \Theta & \text{otherwise.} \end{cases}$$

If $\widehat{\pi}$ is indeed ϵ-optimal for some $\boldsymbol{\theta} \in \Theta$, then the first condition is satisfied, and we have:

$$\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2 = \big\|\sqrt{d}\mu(\widehat{\phi} - \Delta') - \sqrt{d}\mu(\widehat{\phi} - \Delta)\big\|_2 \leqslant 2\sqrt{d}\mu\sqrt{\frac{2\epsilon}{\sqrt{d}\mu}} = \sqrt{8\sqrt{d}\mu\epsilon}. \tag{G.5}$$

We denote $\mathcal{E}$ as the event that $\widehat{\pi}$ is ϵ-optimal for $\boldsymbol{\theta} \in \Theta$. Then, we get

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2 \cdot \mathbb{I}\{\mathcal{E}\} + \big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2 \cdot \mathbb{I}\{\mathcal{E}^c\}\right] \\ &\leqslant 8\sqrt{d}\mu\epsilon + \mathbb{E}_{\boldsymbol{\theta}}\left[\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2 \cdot \mathbb{I}\{\mathcal{E}^c\}\right] && \text{(Eqn. (G.5))} \\ &\leqslant 8\sqrt{d}\mu\epsilon + 2d\mu^2 \cdot P_{\boldsymbol{\theta}}[\mathcal{E}^c]. && (\max\{\|\widehat{\boldsymbol{\theta}}\|_2^2, \|\boldsymbol{\theta}\|_2^2\} \leqslant d\mu^2) \end{aligned}$$

On the other hand, by Theorem G.1, there exists a parameter $\boldsymbol{\theta} \in \Theta$ such that, if we collect T samples and set $\mu = \sqrt{d/(8K^2T)}$, then the following lower bound holds:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2\right] \geqslant \frac{d^2}{32K^2T}.$$

To satisfy both inequalities, we require:

$$\begin{aligned} & \frac{2\sqrt{2}d\epsilon}{\sqrt{K^2T}} + \frac{d^2}{4K^2T} \cdot P_{\boldsymbol{\theta}}[\mathcal{E}^c] \geqslant \frac{d^2}{32K^2T} \\ \iff & P_{\boldsymbol{\theta}}[\mathcal{E}^c] \geqslant \frac{1}{8} - \frac{4\sqrt{2}K\sqrt{T}\epsilon}{d}. \end{aligned}$$

It follows that if

$$\frac{1}{8} - \frac{4\sqrt{2}K\sqrt{T}\epsilon}{d} \geqslant 0.1 \iff \frac{0.025^2}{32} \cdot \frac{d^2}{K^2\epsilon^2} \geqslant T,$$

then we have that $P_{\boldsymbol{\theta}}[\mathcal{E}^c] \geqslant 0.1$. In words, this means that with constant probability, any algorithm must either collect more than $c \cdot \frac{d^2}{K^2\epsilon^2}$ samples, or output a policy that is not ϵ-optimal. This implies that $T = \Omega(\frac{d^2}{K^2\epsilon^2})$ samples are necessary to guarantee an ϵ-optimal policy. Equivalently, after T rounds, the suboptimality gap ϵ is lower bounded as

$$\textbf{SubOpt}(T) = \Omega\left(\frac{d}{K\sqrt{T}}\right).$$

This concludes the proof of Theorem 3. □

G.2 Proof of Lemmas for Theorem 3

G.2.1 Proof of Lemma G.2

Proof of Lemma G.2. By the definition of KL divergence, we have

$$D_{\mathrm{KL}}\big(\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)\|\mathbb{P}_{\boldsymbol{\theta}'}(\cdot|S)\big) = \mathbb{E}_{\sigma\sim\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)}\left[\sum_{j=1}^{|S|}\left(\phi_{\sigma_j}^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right) - \log\frac{\sum_{k=j}^{|S|} e^{\phi_{\sigma_k}^\top\boldsymbol{\theta}}}{\sum_{k=j}^{|S|} e^{\phi_{\sigma_k}^\top\boldsymbol{\theta}'}}\right)\right]. \tag{G.6}$$

Fix a stage j and a ranking σ. We define

$$p_{k'}(\boldsymbol{\theta}) := \frac{\exp\left(\phi_{\sigma_{k'}}^\top\boldsymbol{\theta}\right)}{\sum_{k=j}^{|S|}\exp\left(\phi_{\sigma_k}^\top\boldsymbol{\theta}\right)}, \quad \text{where } k' \in \{j, \dots, |S|\},$$

which corresponds to the Multinomial Logit (MNL) probability of selecting action $\sigma_{k'}$ at position j, given the parameter $\boldsymbol{\theta}$ and the choice set S. Moreover, we define

$$f(\boldsymbol{\theta}) := \log\left(\sum_{k=j}^{|S|} e^{\phi_{\sigma_k}^\top\boldsymbol{\theta}}\right).$$

Then, by applying the mean value form of Taylor's theorem, there exists $\bar{\boldsymbol{\theta}} = (1-c)\boldsymbol{\theta} + c\boldsymbol{\theta}'$ for some $c \in (0,1)$ such that

$$\begin{aligned}
-\log\frac{\sum_{k=j}^{|S|} e^{\phi_{\sigma_k}^\top\boldsymbol{\theta}}}{\sum_{k=j}^{|S|} e^{\phi_{\sigma_k}^\top\boldsymbol{\theta}'}} &= f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})^\top\left(\boldsymbol{\theta}'-\boldsymbol{\theta}\right) + \frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\theta}\right)^\top\nabla^2_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\theta}})\left(\boldsymbol{\theta}'-\boldsymbol{\theta}\right) \quad \text{(Taylor's theorem)} \\
&\leqslant \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})^\top\left(\boldsymbol{\theta}'-\boldsymbol{\theta}\right) + \frac{1}{2}\sum_{k=j}^{|S|} p_k(\bar{\boldsymbol{\theta}})\left(\phi_{\sigma_k}^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2 \\
&\leqslant \sum_{k=j}^{|S|} p_k(\boldsymbol{\theta})\phi_{\sigma_k}^\top\left(\boldsymbol{\theta}'-\boldsymbol{\theta}\right) + \frac{1}{2}\sum_{a\in S}\left(\phi_a^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2,
\end{aligned} \tag{G.7}$$

where the first inequality holds because

$$\nabla^2_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\theta}}) = \sum_{k=j}^{|S|} p_k(\bar{\boldsymbol{\theta}})\phi_{\sigma_k}\phi_{\sigma_k}^\top - \left(\sum_{k=j}^{|S|} p_k(\bar{\boldsymbol{\theta}})\phi_{\sigma_k}\right)\left(\sum_{k=j}^{|S|} p_k(\bar{\boldsymbol{\theta}})\phi_{\sigma_k}\right)^\top \preceq \sum_{k=j}^{|S|} p_k(\bar{\boldsymbol{\theta}})\phi_{\sigma_k}\phi_{\sigma_k}^\top.$$

Plugging Equation (G.7) into Equation (G.6), we get

$$
\begin{aligned}
&D_{\mathrm{KL}}\big(\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)\|\mathbb{P}_{\boldsymbol{\theta}'}(\cdot|S)\big)\\
&\quad\leqslant \mathbb{E}_{\sigma\sim\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)}\left[\sum_{j=1}^{|S|}\left(\phi_{\sigma_j}^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right)-\sum_{k=j}^{|S|}p_k(\boldsymbol{\theta})\phi_{\sigma_k}^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right)+\frac{1}{2}\sum_{a\in S}\left(\phi_a^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2\right)\right]\\
&\quad= \mathbb{E}_{\sigma\sim\mathbb{P}_{\boldsymbol{\theta}}(\cdot|S)}\left[\sum_{j=1}^{|S|}\underbrace{\mathbb{E}_{\sigma_j}\left[\phi_{\sigma_j}^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right)-\sum_{k=j}^{|S|}p_k(\boldsymbol{\theta})\phi_{\sigma_k}^\top\left(\boldsymbol{\theta}-\boldsymbol{\theta}'\right)\,\Big|\,\sigma_1,\dots,\sigma_{j-1}\right]}_{=0}\right] && \text{(Tower rule)}\\
&\qquad+\frac{|S|}{2}\sum_{a\in S}\left(\phi_a^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2\\
&\quad\leqslant \frac{K}{2}\sum_{a\in S}\left(\phi_a^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\right)^2, && (|S|\leqslant K)
\end{aligned}
$$

which concludes the proof. □

H Additional Discussions

In this section, we provide additional discussion of our approach. In Subsection H.1, we propose a more efficient assortment selection rule than Equation (9), by using an arbitrary reference action $\bar{a}_t \in \mathcal{A}$ instead of selecting the one that maximizes average uncertainty. In Subsection H.2, we show that under a sufficient feature diversity condition, selecting S_t uniformly at random can still achieve a comparable suboptimality gap. Finally, in Subsection H.3, we extend our approach to the active learning setting, as studied in [19].

H.1 Arbitrary Reference Action for More Efficient Assortment Selection

As described in the main paper, the reference action $\bar{a}_t$ is selected to maximize the average uncertainty across the subset S_t, according to Equation (9). This selection incurs a computational cost of $\tilde{\mathcal{O}}(N^2K)$ (see Remark 2).

However, in this subsection, we show that $\bar{a}_t$ can, in fact, be selected arbitrarily—i.e., any $\bar{a}_t \in \mathcal{A}$ is valid. Specifically, we modify our assortment selection rule as follows:

$$
S_t = \operatorname*{argmax}_{\substack{S\in\mathcal{S}\\ \bar{a}_t\in S}} \frac{1}{|S|}\sum_{a\in S}\left\|\phi\left(x_t,a\right)-\phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}, \quad \text{for any } \bar{a}_t\in\mathcal{A}. \tag{H.1}
$$

This results in only a constant-factor increase (specifically, by a factor of 2) in the suboptimality gap, while reducing the computational cost to $\tilde{\mathcal{O}}(NK)$, as it removes the need to enumerate over all possible reference actions.

To show this explicitly, we return to Equation (D.6). Let $\bar{a}_t$ be an arbitrary action in $\mathcal{A}$ (e.g., selected uniformly at random). Then, we have

$$
\begin{aligned}
&\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\widehat{\pi}_T(x_t)\right)\pm\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}\\
&\leqslant \frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\left\|\phi\left(x_t,\pi^\star(x_t)\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}+\left\|\phi\left(x_t,\widehat{\pi}_T(x_t)\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}\right)\\
&= \frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\sum_{a\in S_t^\star}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}+\sum_{a'\in\widehat{S}_t}\left\|\phi\left(x_t,a'\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}\right)\\
&&& \text{(Let } S_t^\star:=\{\pi^\star(x_t),\bar{a}_t\} \text{ and } \widehat{S}_t:=\{\widehat{\pi}_T(x_t),\bar{a}_t\})\\
&\leqslant \frac{4\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{1}{|S_t|}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}. && (S_t \text{ selection rule, Eqn. (H.1)})
\end{aligned}
$$

The remaining steps of the proof follow exactly as in the proof of Theorem 1 (or Theorem 2).

H.2 Suboptimality Gap Under Sufficient Diversity Condition

So far, we have considered the general case where the feature vectors ϕ are not required to be diverse, and as a result, the induced matrix $H_t - \lambda \mathbf{I}_d$ (or $\Lambda_t - \lambda \mathbf{I}_d$) may be singular. In this subsection, we discuss the case where the following diversity assumption holds:

Assumption H.1 (Diverse features)**.** *For any $S \in \mathcal{S}$ and $a' \in S$, there exists a constant $\lambda_0 > 0$ such that $\lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\frac{1}{|S|}\sum_{a\in S}(\phi(x,a)-\phi(x,a'))(\phi(x,a)-\phi(x,a'))^\top\right]\right) \geqslant \lambda_0$.*

Under this condition, it is sufficient to randomly select *exactly* K actions, rather than solving the optimization problem in Equation (9) to construct the assortment. Specifically, we can select S_t as:

$$S_t \sim \text{Unif}\left(\{S \subseteq \mathcal{A} : |S| = K\}\right), \quad \forall t \in [T]. \tag{H.2}$$

Theorem H.1 (Suboptimality Gap of Random Assortment Selection Under Diversity)**.** *Let $\mathcal{T}^w := \left\{t \in [T] : \max_{a\in\mathcal{A}} \|\phi(x_t,a) - \phi(x_t,a'_t)\|_{H_t^{-1}} \geqslant \frac{1}{\beta_{T+1}(\delta)}\right\}$, where $a'_t \in S_t$ is an arbitrary action selected from the assortment S_t. Suppose $T = \Omega\big(\log(dT)/\lambda_0\big)$ and $T > |\mathcal{T}^w|$. Then, under the same setting as Theorem 1 and Assumption H.1, if S_t is randomly selected according to Equation* (H.2)*, then with probability at least $1-\delta$, we have:*

$$\textbf{SubOpt}(T) = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{\lambda_0(T - |\mathcal{T}^w|)K}}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{\lambda_0\big(T - \min\{(dK)^2/\kappa, T-1\}\big)K}}\right).$$

Discussion of Theorem H.1. Theorem H.1 shows that for sufficiently large T (i.e., $T = \Omega((dK)^2/\kappa + \log(dT)/\lambda_0)$), the suboptimality gap under the uniform random assortment selection strategy achieves $\widetilde{\mathcal{O}}\big(\sqrt{\frac{d}{\lambda_0 TK}}\big)$. This result suggests that when the feature space is sufficiently diverse, uniform random selection is effective for learning. It also provides a theoretical explanation for the empirical success of many RLHF implementations [76, 59], where the feature space is sufficiently diverse and prompt-action (sub)set pairs are often selected uniformly at random.

Note that the lower bound we establish in Theorem 3 does not rely on the diversity assumption (Assumption H.1). As a result, deriving a lower bound under the diversity assumption remains an open question, which we leave for future work.

Proof of Theorem H.1. To provide the proof of Theorem H.1, we first introduce useful concentration inequalities.

Lemma H.1 (Matrix Chernoff, Adapted Sequence from Tropp 80)**.** *Consider a finite adapted sequence $\{X_k\}$ with filtratio $\{\mathcal{F}_t\}_{t\geqslant 0}$ of positive-semi definite matrices with dimension d, and suppose that $\lambda_{\max}(X_k) \leqslant R$ almost surely. Define the finite series*

$$Y := \sum_k X_k, \quad \textit{and} \quad W := \sum_k \mathbb{E}_{k-1} X_k.$$

Then, for all $\mu \geqslant 0$, we have

$$\mathbb{P}\left\{\lambda_{\min}(Y) \leqslant (1-\delta)\mu \ \textit{and}\ \lambda_{\min}(W) \geqslant \mu\right\} \leqslant d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu/R}, \quad \delta \in [0,1).$$

By setting $\delta = \frac{3}{4}$, $\mu = \lambda_0 t$, $R = x_{\max}$ in Lemma H.1, we obtain the following result.

Corollary H.1 (Eigenvalue Growth of Adaptive Gram Matrix)**.** *If $\|X_i\|_2 \leqslant x_{\max}$ and $\lambda_{\min}\left(\mathbb{E}\left[X_i X_i^\top \mid \mathcal{F}_{i-1}\right]\right) \geqslant \lambda_0$, then, with probability at least $1 - d\exp\left(-c_1 \frac{\lambda_0 t}{x_{\max}}\right)$,*

$$\lambda_{\min}\left(\sum_{i=1}^t X_i X_i^\top\right) \geqslant \frac{\lambda_0}{4} t$$

holds for some absolute constant c_1.

Now we are ready to provide the proof of Theorem H.1. For simplicity, we present only the case of the PL loss, since the extension to the RB loss directly follows from similar arguments in the proof of Theorem 2. By the definition of the suboptimality gap, we have

$$
\begin{aligned}
\textbf{SubOpt}(T) &= \mathbb{E}_{x\sim\rho}\left[\left(\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right)^\top \boldsymbol{\theta}^\star\right] \\
&\leqslant \mathbb{E}_{x\sim\rho}\left[\left(\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right)^\top \left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\right] \\
&\qquad\qquad (\widehat{\pi}_T(x)=\operatorname{argmax}_{a\in\mathcal{A}}\phi(x,a)^\top\widehat{\boldsymbol{\theta}}_{T+1}) \\
&\leqslant \left\|\mathbb{E}_{x\sim\rho}\left[\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right]\right\|_{H_{T+1}^{-1}}\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_{T+1}} \qquad \text{(Hölder's ineq.)} \\
&\leqslant \beta_{T+1}(\delta)\left\|\mathbb{E}_{x\sim\rho}\left[\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right]\right\|_{H_{T+1}^{-1}}. \\
&\qquad\qquad \text{(Corollary D.1, with prob. } 1-\delta)
\end{aligned}
$$

To proceed, we slightly modify the definition of Λ_t, as we no longer compute the reference action explicitly. Let a_s' be an arbitrary action selected from S_s, which can simply be chosen by sampling uniformly from S_s. Additionally, the regularization term λ is no longer required. Then, we redefine Λ_t as follows:

$$
\Lambda_t := \sum_{s\in[t-1]\setminus\mathcal{T}^w}\sum_{a\in S_s}\left(\phi(x_s,a)-\phi(x_s,a_s')\right)\left(\phi(x_s,a)-\phi(x_s,a_s')\right)^\top, \quad a_s'\in S
$$

where

$$
\mathcal{T}^w := \left\{t\in[T]:\max_{a\in\mathcal{A}}\|\phi(x_t,a)-\phi(x_t,a_t')\|_{H_t^{-1}}\geqslant\frac{1}{\beta_{T+1}(\delta)}\right\}.
$$

Then, by Lemma D.2, we obtain

$$
\begin{aligned}
\left\|\mathbb{E}_{x\sim\rho}\left[\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right]\right\|_{H_{T+1}^{-1}} &\leqslant \sqrt{50}\left\|\mathbb{E}_{x\sim\rho}\left[\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right]\right\|_{\Lambda_{T+1}^{-1}} \\
&\qquad\qquad \text{(Lemma D.2, with prob. } 1-\delta) \\
&\leqslant \frac{10\sqrt{2}}{\lambda_{\min}\left(\Lambda_{T+1}\right)} \qquad (\|\phi(x,a)\|_2\leqslant 1)
\end{aligned}
$$

Under the diversity assumption (Assumption H.1), for $T\geqslant\frac{c}{\lambda_0}\log\frac{d}{\delta}$ with some constant $c>0$, by Corollary H.1, we have, with probability at least $1-\delta$,

$$
\lambda_{\min}\left(\Lambda_{T+1}\right)\geqslant\frac{\lambda_0}{4}(T-|\mathcal{T}^w|)K.
$$

Suppose $T-|\mathcal{T}^w|>0$. Then, combining the above results, we get

$$
\begin{aligned}
\textbf{SubOpt}(T) \leqslant \beta_{T+1}(\delta)\frac{20\sqrt{2}}{\sqrt{\lambda_0(T-|\mathcal{T}^w|)K}} &= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{\lambda_0(T-|\mathcal{T}^w|)K}}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{\lambda_0\left(T-\min\{(dK)^2/\kappa,T-1\}\right)K}}\right),
\end{aligned}
$$

where in the last inequality we use the fact that it holds that $|\mathcal{T}^w|\leqslant|\mathcal{T}^w\cap(\mathcal{T}_0)^c|+|\mathcal{T}_0|=\tilde{\mathcal{O}}\left(\frac{d^2K^2}{\kappa}\right)$, which follows from Lemmas D.4 and D.5. This concludes the proof of Theorem H.1. □

H.3 Extension to Active Learning Setting

In this subsection, we consider a different setting—referred to as the *active learning setting*—where the learner has access to the entire context set $\mathcal{X}$, and the objective is to minimize the following *worst-case suboptimality gap*, defined as:

$$
\textbf{WorstSubOpt}(T) := \max_{x\in\mathcal{X}}\left[r_{\boldsymbol{\theta}^\star}\left(x,\pi^\star(x)\right)-r_{\boldsymbol{\theta}^\star}\left(x,\widehat{\pi}(x)\right)\right].
$$

This setting has received increasing attention in recent work [52, 49, 73, 19, 51, 79, 39]. However, most existing approaches focus exclusively on pairwise preference feedback. Mukherjee et al. [51]

study an online learning-to-rank problem where, for each context, a fixed set of K actions is provided, and the goal is to recover the true ranking based on feedback over these K actions. In contrast, we consider a more general setting in which, for each context, a set of N actions is available. The learner selects at most K actions from this set and receives ranking feedback over the selected subset. Thekumparampil et al. [79] investigate the problem of ranking $N \geqslant K$ items using partial rankings over K candidates, but under a context-free setting. In contrast, we study a stochastic contextual setting, where contexts are drawn from an unknown (and fixed) distribution.

In the active learning setting, the algorithm jointly selects the context x_t—which is no longer given but actively chosen—and the assortment S_t by maximizing the average uncertainty objective. For computational efficiency, we employ the arbitrary reference action strategy described in Equation (H.1). (Note that one may alternatively use the reference action selection method from Equation (9), which selects $\bar{a}_t$ to maximize uncertainty.)

$$(x_t, S_t) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \underset{\substack{S \in \mathcal{S} \\ \bar{a}_t \in S}}{\operatorname{argmax}} \frac{1}{|S|} \sum_{a \in S} \|\phi(x, a) - \phi(x, \bar{a}_t)\|_{H_t^{-1}}, \quad \text{for any } \bar{a}_t \in \mathcal{A}. \tag{H.3}$$

The rest of the algorithm proceeds in the same manner as Algorithm 3. With the above context-assortment selection strategy, `M-AUPO` achieves the following bound on the worst-case suboptimality gap, matching the order established in Theorem 1 (and in Theorem 2):

Theorem H.2. *Under the same setting as Theorem 1 and 2, with probability at least* $1 - \delta$, `M-AUPO` *achieves the following worst-case suboptimality gap:*

$$\textbf{WorstSubOpt}(T) = \begin{cases} \tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}} + \frac{d^2K^2}{\kappa T}\right), & \textit{(PL loss)} \\ \tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^{T}\frac{1}{|S_t|}} + \frac{d^2}{\kappa T}\right). & \textit{(RB loss)} \end{cases}$$

Proof of Theorem H.2. We present only the proof using the PL loss (3), as extending it to the RB loss case (4) follows similarly to the extension from Theorem 1 to Theorem 2.

By the definition of the worst-cacse suboptimality gap, we have

$$\begin{aligned} \textbf{WorstSubOpt}(T) &= \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \boldsymbol{\theta}^\star \right] \\ &\leqslant \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right] \\ &\qquad\qquad (\widehat{\pi}_T(x) = \operatorname{argmax}_{a \in \mathcal{A}} \phi(x,a)^\top \widehat{\boldsymbol{\theta}}_{T+1}) \\ &= \frac{1}{T} \sum_{t=1}^{T} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right]. \end{aligned}$$

We adopt the same definitions for $\mathcal{T}_0$ (Equation (D.3)), $\mathcal{T}^w$ (Equation (D.1)), and Λ_t (Equation (D.2)) as in Theorem 1. Then, we have

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^{T} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right] \\ &\quad = \frac{1}{T} \sum_{t \in \mathcal{T}_0} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right] \\ &\quad + \frac{1}{T} \sum_{t \in \mathcal{T}^w \cap (\mathcal{T}_0)^c} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right] \\ &\quad + \frac{1}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right] \\ &\quad \leqslant \frac{8B}{\log(2)T} d \log\left(1 + \frac{2K}{\log(2)\lambda}\right) + \frac{48\sqrt{2}BK^2}{\kappa T} \beta_{T+1}(\delta)^2 d \log\left(1 + \frac{2KT}{d\lambda}\right) \\ &\qquad\qquad \text{(Lemma D.4 and D.5)} \\ &\quad + \frac{1}{T} \sum_{t \notin \mathcal{T}_0 \cup \mathcal{T}^w} \max_{x \in \mathcal{X}} \left[\left(\phi(x, \pi^\star(x)) - \phi(x, \widehat{\pi}_T(x)) \right)^\top \left(\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}_{T+1} \right) \right]. \end{aligned} \tag{H.4}$$

To further bound the last term of Equation (H.4), we get

$$\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\max_{x\in\mathcal{X}}\left[\left(\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right)^\top\left(\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right)\right]$$

$$\leqslant\frac{1}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\max_{x\in\mathcal{X}}\left[\left\|\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right\|_{H_{T+1}^{-1}}\left\|\boldsymbol{\theta}^\star-\widehat{\boldsymbol{\theta}}_{T+1}\right\|_{H_{T+1}}\right]$$

(Hölder's ineq.)

$$\leqslant\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\max_{x\in\mathcal{X}}\left[\left\|\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right\|_{H_t^{-1}}\right].$$

($H_{T+1}\geq H_t$ and Corollary D.1, with prob. $1-\delta$)

Then, for any arbitrary $\bar{a}_t\in\mathcal{A}$, we have

$$\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\max_{x\in\mathcal{X}}\left[\left\|\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\widehat{\pi}_T(x)\right)\right\|_{H_t^{-1}}\right]$$

$$\leqslant\frac{\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\max_{x\in\mathcal{X}}\left[\left\|\phi\left(x,\pi^\star(x)\right)-\phi\left(x,\bar{a}_t\right)\right\|_{H_t^{-1}}+\left\|\phi\left(x,\widehat{\pi}_T(x)\right)-\phi\left(x,\bar{a}_t\right)\right\|_{H_t^{-1}}\right]$$

$$\leqslant\frac{4\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{1}{|S_t|}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}.$$

((x_t,S_t) selection rule, Eqn. (H.3))

Hence, we further obtain

$$\frac{4\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{1}{|S_t|}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi(x_t,\bar{a}_t)\right\|_{H_t^{-1}}$$

$$\leqslant\frac{4\beta_{T+1}(\delta)}{T}\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\frac{1}{|S_t|}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}} \qquad \text{(Eqn. (H.3))}$$

$$\leqslant\frac{4\beta_{T+1}(\delta)}{T}\sqrt{\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\frac{1}{|S_t|}\right)^2|S_t|}\sqrt{\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{H_t^{-1}}^2}$$

(Cauchy-Schwartz ineq.)

$$=\frac{4\beta_{T+1}(\delta)}{T}\sqrt{\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\left(\frac{1}{|S_t|}\right)^2|S_t|}\sqrt{50\sum_{t\notin\mathcal{T}_0\cup\mathcal{T}^w}\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{\Lambda_t^{-1}}^2}$$

(Lemma D.2, with prob. $1-\delta$)

$$\leqslant\frac{30\beta_{T+1}(\delta)}{T}\sqrt{\sum_{t=1}^T\frac{1}{|S_t|}}\sqrt{\sum_{t=1}^T\min\left\{1,\sum_{a\in S_t}\left\|\phi\left(x_t,a\right)-\phi\left(x_t,\bar{a}_t\right)\right\|_{\Lambda_t^{-1}}^2\right\}}$$

$$\leqslant\frac{30\beta_{T+1}(\delta)}{T}\sqrt{\sum_{t=1}^T\frac{1}{|S_t|}}\sqrt{2d\log\left(1+\frac{2KT}{d\lambda}\right)} \qquad \text{(Lemma D.3)}$$

$$=\mathcal{O}\left(\frac{\beta_{T+1}(\delta)}{T}\sqrt{\sum_{t=1}^T\frac{1}{|S_t|}}\cdot\sqrt{d\log\left(KT\right)}\right). \tag{H.5}$$

Plugging Equation (H.5) into Equation (H.4), and setting $\beta_{T+1}(\delta)=\mathcal{O}\big(B\sqrt{d\log(KT)}+B\sqrt{\lambda}\big)$, with probability at least $1-3\delta$, we have

$$\textbf{SubOpt}(T)=\tilde{\mathcal{O}}\left(\frac{d}{T}\sqrt{\sum_{t=1}^T\frac{1}{|S_t|}}+\frac{d^2K^2}{\kappa T}\right).$$

Substituting $\delta\leftarrow\frac{\delta}{3}$, we conclude the proof of Theorem H.2. □

I Experimental Details and Additional Results

I.1 Synthetic Data

Setup. In the synthetic data experiment, we sample the true but unknown parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ from a d-dimensional standard normal distribution, i.e., $\boldsymbol{\theta}^\star \sim \mathcal{N}(0, I_d)$, and then normalize it to ensure $\|\boldsymbol{\theta}^\star\|_2 \leqslant 1$. We consider four different types of context sets $\mathcal{X}$:

1. **Instance 1 (Stochastic contexts)**: For each $x \in \mathcal{X}$, the feature vectors $\phi(x, \cdot)$ are sampled from a standard normal distribution and then normalized to satisfy $\|\phi(x, \cdot)\|_2 \leqslant 1$. Here, $|\mathcal{X}| = 100$.

2. **Instance 2 (Non-contextual)**: A single shared context is used for all rounds, i.e., $\mathcal{X} = \{x_1\}$ and $|\mathcal{X}| = 1$. The corresponding feature vectors $\phi(x_1, \cdot)$ are sampled from a standard normal distribution and then normalized to satisfy $\|\phi(x_1, \cdot)\|_2 \leqslant 1$.

3. **Instance 3 (Hard-to-learn contexts)**: For each $x \in \mathcal{X}$, the feature vectors $\phi(x, \cdot)$ are constructed such that most of them are approximately orthogonal to the true parameter $\boldsymbol{\theta}^\star$. Here, $|\mathcal{X}| = 100$.

4. **Instance 4 (Skewed stochastic contexts)**: For each $x \in \mathcal{X}$, the feature vectors $\phi(x, \cdot)$ are sampled in a skewed or biased manner and then normalized to satisfy $\|\phi(x, \cdot)\|_2 \leqslant 1$. Here, $|\mathcal{X}| = 100$. This is our main experimental setup in Section 6.1.

Additionally, we set the feature dimension to $d = 5$ and the number of available actions to $|\mathcal{A}| = N = 100$. The suboptimality gap is measured every 25 rounds. All results are averaged over 20 independent runs with different random seeds, and standard errors are reported to indicate variability. The experiments are run on a Xeon(R) Gold 6226R CPU @ 2.90GHz (16 cores).

Baselines. We evaluate our proposed algorithm, `M-AUPO`, against three baselines: (i) `DopeWolfe` [79], a method designed for non-contextual K-subset selection; (ii) `Uniform`, which selects assortments of size K uniformly at random; and (iii) `Best&Ref`, which forms a pair of actions ($|S_t| = 2$) by combining one action from the current policy with another from a reference policy (e.g., uniform random or SFT), following the setup in Online GSHF [89] and XPO [88].

Thekumparampil et al. [79] propose a D-optimal design approach for the Plackett-Luce objective to efficiently select informative subsets of items for comparison. Recognizing the computational complexity inherent in this method, they introduce a randomized Frank-Wolfe algorithm, named `DopeWolfe`, which approximates the optimal design by solving linear maximization sub-problems on randomly chosen variables. This approach reduces computational overhead while maintaining effective learning performance. However, their approach is specifically tailored to the single-context setting (e.g., **Instance 2**) and may not generalize well to the multiple-context scenarios (e.g., **Instances 1, 3,** and **4**). While their original implementation updates the model parameters using a maximum likelihood estimation (MLE) procedure, we instead adopt an online update strategy (as described in Procedures 1 and 2) to ensure a fair comparison across all methods. For sampling size parameter R, we set $R = \min\{\binom{N}{K}, 100,000\}$.

The uniform random assortment selection strategy, `Uniform`, selects K actions uniformly at random from the available action set $\mathcal{A}$ at each round, without utilizing any uncertainty or reward-based information. This approach can be effective when the feature representations are sufficiently diverse (e.g., **Instances 1, 2,** and **4**), but may perform poorly when the diversity parameter λ_0 in Assumption H.1 is very small (e.g., **Instance 3**).

`Best&Ref` constructs an action pair ($|S_t| = 2$) by combining two distinct sources of actions. The first action is chosen to maximize the current reward estimate, while the second is sampled from a reference policy—such as a uniform random policy or a supervised fine-tuned (SFT) model. This pairing mechanism follows the framework introduced in Online GSHF [89] and XPO [88]. In our experiments, we use the uniform random policy as the reference.

Performance measure. Since computing the exact suboptimality gap is challenging under a general distribution ρ, we instead evaluate the *average realized regret*, which serves as a slightly relaxed

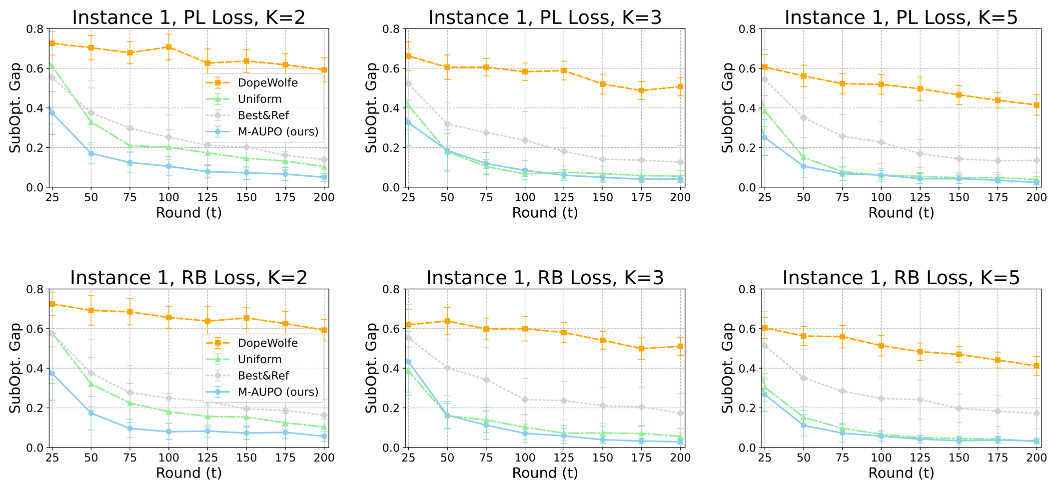

Figure I.1: Performance comparisons for Instance 1 (Stochastic contexts) with $K = 2$, 3, and 5, evaluated under the PL loss (first row) and RB loss (second row).

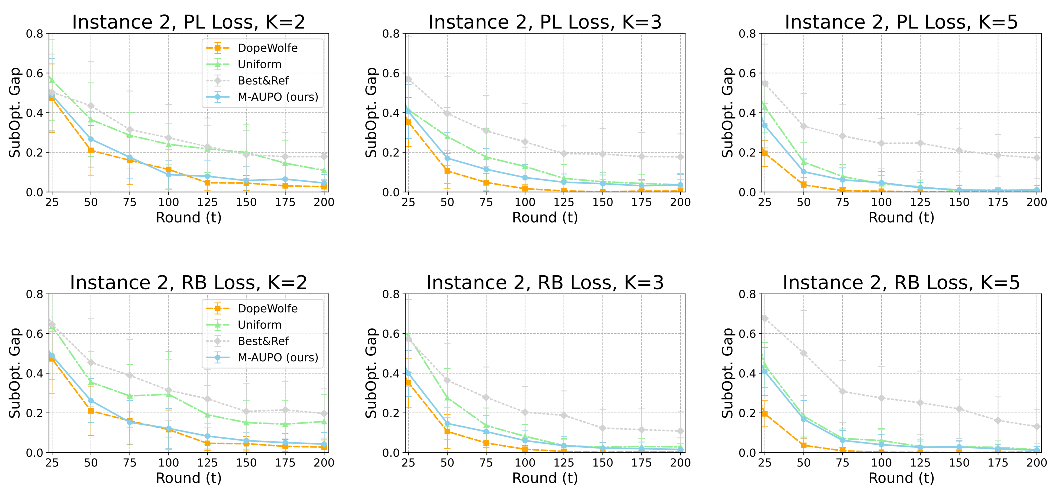

Figure I.2: Performance comparisons for Instance 2 (Non-contextual) with $K = 2$, 3, and 5, evaluated under the PL loss (first row) and RB loss (second row).

proxy for the suboptimality gap.

$$\begin{aligned}\textbf{SubOpt}(T) &\lesssim \frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t, \pi^\star(x_t)\right) - \phi\left(x_t, \widehat{\pi}_T(x_t)\right)\right)^\top \boldsymbol{\theta}^\star + \underbrace{\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)}_{\text{incurred by MDS terms}} \\ &\leqslant \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\phi\left(x_t, \pi^\star(x_t)\right) - \phi\left(x_t, \pi_t(x_t)\right)\right)^\top \boldsymbol{\theta}^\star}_{=:\textit{average realized regret}} + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right),\end{aligned}$$

where we define $\pi_t(x) := \operatorname{argmax}_a \phi(x, a)^\top \widehat{\boldsymbol{\theta}}_t$, and let $\widehat{\pi}_T$ denote the best policy among $\{\pi_t\}_{t=1}^T$, possibly selected using a validation set.

Results. We present performance comparisons in Figures I.1 through I.4, corresponding to Instances 1 through 4, respectively. Overall, our algorithm, `M-AUPO`, consistently outperforms other baseline

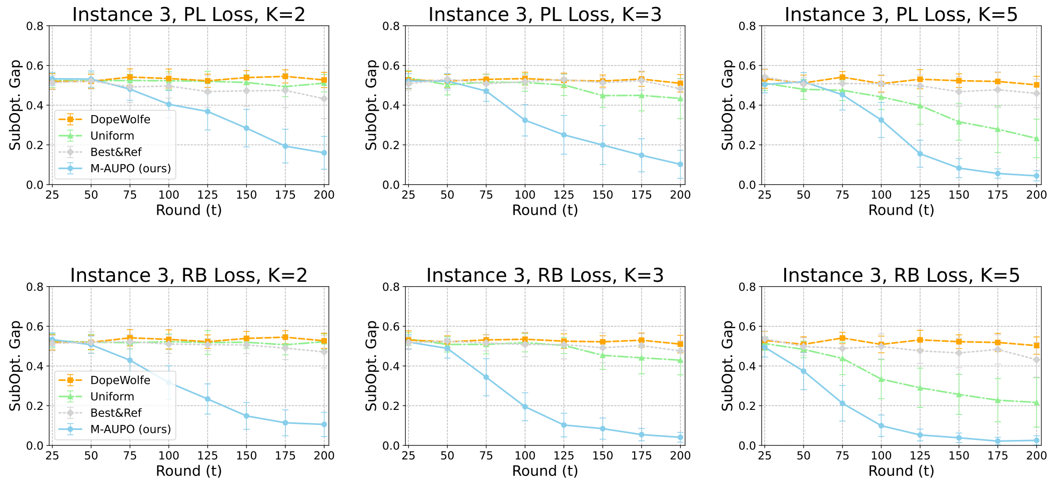

Figure I.3: Performance comparisons for Instance 3 (Hard-to-learn contexts) with $K = 2, 3$, and 5, evaluated under the PL loss (first row) and RB loss (second row).

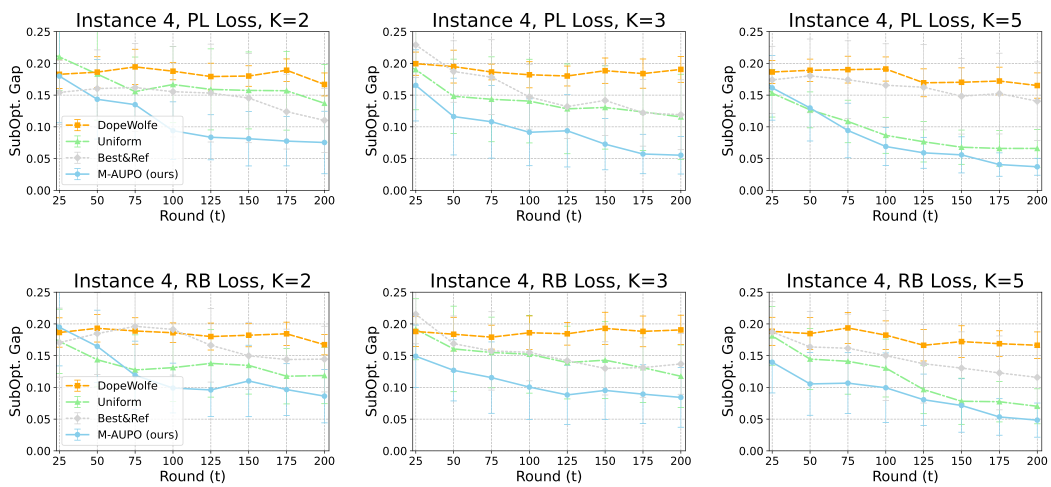

Figure I.4: Performance comparisons for Instance 4 (Skewed stochastic contexts) with $K = 2, 3$, and 5, evaluated under the PL loss (first row) and RB loss (second row).

methods. The only exception is in Instance 2 (Figure I.2), a special case of the non-contextual setting, where `M-AUPO` performs slightly worse than `DopeWolfe`. This is an expected outcome, as `DopeWolfe` leverages a D-optimal design strategy, which is known to be highly effective in the single-context setting. However, it is important to note that `DopeWolfe` completely fails in more general contextual scenarios (Figures I.1, I.3, and I.4), and its computational cost is significantly higher than that of our approach (see Table I.1).

The uniform random assortment selection strategy, `Uniform`, demonstrates competitive performance—though still worse than `M-AUPO`—in Instances 1, 2, and 4, as illustrated in Figures I.1, I.2, and I.4, respectively. However, in Instance 3 (Figure I.3), where the diversity parameter λ_0 is very small due to most feature vectors lying within a hyperplane, `Uniform` performs significantly worse, as discussed in Appendix H.2.

The `Best&Ref` algorithm performs consistently worse than our algorithm and does not benefit from larger K, since it always selects only a pair of actions.

K	DopeWolfe	Uniform	Best&Ref	M-AUPO (ours)
2	7.28 s	0.10 s	0.10s	1.94 s
3	99.6 s	0.18 s	0.10s	2.37 s
5	150.5 s	0.35 s	0.10s	2.94 s
7	218.8 s	0.58 s	0.10s	4.17 s
10	331.1 s	0.99 s	0.10s	4.50 s

Table I.1: Runtime comparison over 200 rounds (seconds)

Moreover, the suboptimality gap consistently decreases with larger K across the three algorithms—M-AUPO, Uniform, and DopeWolfe—while Best&Ref shows no such improvement. For both M-AUPO and Uniform, this trend is consistent with our theoretical results (Theorems 1, 2, and H.1). In contrast, the improvement observed for DopeWolfe suggests that its current theoretical guarantees may be loose, as their bound actually deteriorates with increasing K (recall that their theoretical guarantee worsens for larger K). This indicates that tighter bounds might be achievable by incorporating some of the techniques introduced in our work.

Table I.2 presents the average assortment size $|S_t|$ of M-AUPO for various values of the maximum assortment size K. In most cases, the algorithm selects the full K actions, i.e., $|S_t| = K$. An exception occurs when K is large (e.g., 30 or more), which may be impractical in real-world applications due to the increased annotation burden on human labelers.

K	2	3	5	7	10	30	50
PL loss, $\lvert S_t\rvert$	2.00	3.00	5.00	7.00	10.00	18.31	18.69
RB loss, $\lvert S_t\rvert$	2.00	3.00	5.00	7.00	10.00	18.39	18.40

Table I.2: Assortment size $|S_t|$ of M-AUPO with varying maximum size K in the synthetic data experiment

I.2 Real-World Dataset

Setup. In our real-world dataset experiments, we evaluate performance on two widely used benchmark datasets: TREC Deep Learning (TREC-DL) and NECTAR. The TREC-DL dataset provides 100 candidate answers for each query, offering a rich and diverse set of responses suitable for learning from listwise feedback. In contrast, the NECTAR dataset presents a more concise setup, with only 7 candidate answers per question. From each dataset, we randomly sample $|\mathcal{X}| = 5000$ prompts (i.e., questions), each paired with its corresponding set of candidate actions—100 for TREC-DL and 7 for NECTAR.

We use the Gemma-2B language model [78] to construct the feature representation $\phi(x, a)$. To obtain $\phi(x, a)$, we first concatenate the input prompt x and the candidate response a into a single sequence, which is then fed into Gemma-2B. The resulting feature vector is extracted from the last hidden layer of the model and has a dimensionality of $d = 2048$. We then apply ℓ_1 normalization to enhance numerical stability and ensure consistent scaling. For each round t, we sample the context index from an exponential distribution with rate $\lambda = 0.1$, which assigns higher probability to smaller indices and thus biases the selection toward earlier contexts. To generate ranking feedback and evaluate the suboptimality gap, we use the Mistral-7B reward model [33] as the ground-truth reward function, denoted by $r_{\boldsymbol{\theta}^\star}$.

We measure the suboptimality gap every 2,500 rounds throughout the training process and report the average performance over 10 independent runs, each with a different random seed. Along with the average, we also include the standard error to indicate variability across runs. In these experiments, we report results under the PL loss only, since the performance difference between PL and RB losses is minimal, as demonstrated in the synthetic data experiments. The experiments are conducted on a Xeon(R) Gold 6226R CPU @ 2.90GHz (16 cores) and a single GeForce RTX 3090 GPU.

Baselines. We use the same set of baselines as in the synthetic data experiments. For DopeWolfe [79], we set the sampling size parameter R as $R = \min\{\binom{N}{K}, 1000\}$. Although a small value of $R \leqslant 1000$

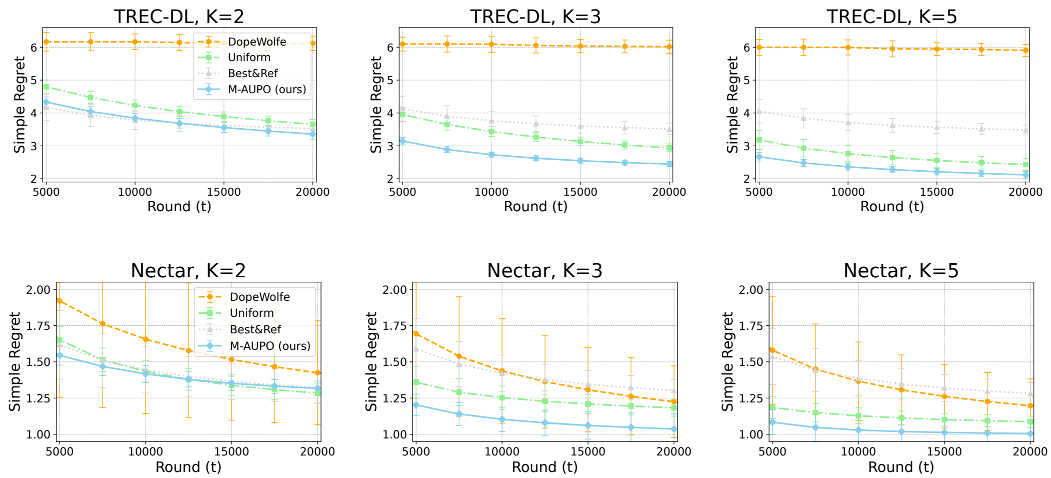

Figure I.5: Performance comparisons on the TREC-DL dataset (top row) and the NECTAR dataset (bottom row) for varying values of $K = 2, 3$, and 5.

may introduce significant approximation error—since the theoretically minimal-error choice is $R = \mathcal{O}\big(\binom{N}{K}\big)$—we adopt this smaller value in our experiment to reduce computational overhead.

Performance measure. We measure the average realized regret as in the synthetic experiment (Appendix I.1).

Results. We present performance comparisons in Figure I.5. Our algorithm, `M-AUPO`, consistently outperforms all baselines by a significant margin. As in the synthetic data experiments, the suboptimality gap for all methods decreases as K increases. Notably, `DopeWolfe` performs particularly poorly on the TREC-DL dataset. This may be attributed to the use of a small sampling size R, which is insufficient compared to the full subset space of size $\binom{N}{K} = \mathcal{O}(N^K) \gg 1000 \geqslant R$. This result highlights an important practical limitation of `DopeWolfe`: despite its use of approximate optimization to reduce runtime, the method still depends on combinatorial sampling to perform well, which becomes computationally infeasible in large-scale settings. In contrast, our algorithm, `M-AUPO`, maintains strong performance while requiring only $\tilde{\mathcal{O}}(NK)$ computational cost, making it significantly more scalable and practical for real-world applications.

Table I.3 reports the actual assortment size $|S_t|$ selected by `M-AUPO` on both datasets. In the TREC-DL experiment, $|S_t|$ is nearly equal to K for all values of K, as the number of available actions is large ($N = 100$). In contrast, in the NECTAR experiment, where the number of available actions is much smaller ($N = 7$), the actual assortment size $|S_t|$ is often smaller than K, especially when $K = N$. This reduction occurs because the limited action space constrains the potential informativeness of larger assortments—for example, it becomes difficult to achieve high average uncertainty when there are too few actions to choose from.

K	2	3	5	7
TREC-DL dataset, $\lvert S_t\rvert$	2.00	3.00	4.99	6.95
NECTAR dataset, $\lvert S_t\rvert$	2.00	2.99	4.31	4.74

Table I.3: Assortment size $|S_t|$ of `M-AUPO` with varying maximum size K in the real-world dataset experiment