Nearly Minimax Optimal Regret for Multinomial Logistic Bandit

Joongkyu Lee
Seoul National University
Seoul, South Korea
jklee0717@snu.ac.kr

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

Abstract

In this paper, we study the contextual multinomial logistic (MNL) bandit problem in which a learning agent sequentially selects an assortment based on contextual information, and user feedback follows an MNL choice model. There has been a significant discrepancy between lower and upper regret bounds, particularly regarding the maximum assortment size K. Additionally, the variation in reward structures between these bounds complicates the quest for optimality. Under uniform rewards, where all items have the same expected reward, we establish a regret lower bound of $\Omega(d\sqrt{T/K})$ and propose a constant-time algorithm, `OFU-MNL+`, that achieves a matching upper bound of $\tilde{\mathcal{O}}(d\sqrt{T/K})$. We also provide instance-dependent minimax regret bounds under uniform rewards. Under non-uniform rewards, we prove a lower bound of $\Omega(d\sqrt{T})$ and an upper bound of $\tilde{\mathcal{O}}(d\sqrt{T})$, also achievable by `OFU-MNL+`. Our empirical studies support these theoretical findings. To the best of our knowledge, this is the first work in the contextual MNL bandit literature to prove minimax optimality — for either uniform or non-uniform reward setting — and to propose a computationally efficient algorithm that achieves this optimality up to logarithmic factors.

1 Introduction

The multinomial logistic (MNL) bandit framework [47, 48, 7, 8, 40, 41, 44, 5, 53] describes sequential assortment selection problems in which an agent offer a sequence of assortments of at most K item from a set of N possible items and receives feedback *only* for the chosen decisions. The choice probability of each outcome is characterized by an MNL model [37]. This framework allows modeling of various real-world situations such as recommender systems and online retails, where selections of assortments are evaluated based on the user-choice feedback among offered multiple options.

In this paper, we study the *contextual* MNL bandit problem [8, 7, 43, 16, 40, 41, 44, 5], where the features of items and possibly contextual information about a user at each round are available. Despite many recent advances, [16, 40, 41, 44, 5], however, no previous studies have proven the minimax optimality of contextual MNL bandits. Chen et al. [16] proposed a regret lower bound of $\Omega(d\sqrt{T}/K)$, where d is the number of features, T is the total number of rounds, and K is the maximum size of assortments, assuming the uniform rewards, i.e., rewards are all same for each of the total N items. Furthermore, Chen and Wang [15] established a regret lower bound of $\Omega(\sqrt{NT})$ in the non-contextual setting (hence, dependence on N appears instead of d), which is tighter in terms of K. It is important to note the difference in the assumptions for the *attraction parameter for the outside option* v_0. Chen and Wang [15] assumed for the attraction parameter for the outside option to be $v_0 = K$, whereas Chen et al. [16] assumed $v_0 = 1$. Therefore, it remains an *open question whether and how the value of* v_0 *affects both lower and upper bounds of regret*.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Comparisons of lower and upper regret bounds in related works on MNL bandits with T rounds, N items, the maximum size of assortments K, d-dimensional feature vectors, and problem-dependent constants $1/\kappa = \mathcal{O}(K^2)$ and $\kappa' = \mathcal{O}(1/K)$. $\tilde{\mathcal{O}}$ represents big-$\mathcal{O}$ notation up to logarithmic factors. For the computational cost (abbreviated as "Comput."), we consider only the dependence on the number of rounds t. "Intractable" means a non-polynomial runtime. The notation "$-$" denotes *not applicable*. The starred (*) papers only consider the non-contextual setting.

		Regret	Contexts	Rewards	v_0	Comput. per Round
Lower Bound	Chen et al. [16]	$\Omega(d\sqrt{T}/K)$	–	Uniform	$\Theta(1)$	–
	Agrawal et al. [8]*	$\Omega(\sqrt{NT/K})$	–	Uniform	$\Theta(K)$	–
	Chen and Wang [15]*	$\Omega(\sqrt{NT})$	–	Uniform	$\Theta(K)$	–
	This work (Theorem 1)	$\Omega(\frac{\sqrt{v_0 K}}{v_0+K} d\sqrt{T})$	–	Uniform	Any value	–
	This work (Theorem 3)	$\Omega(d\sqrt{T})$	–	Non-uniform	$\Theta(1)$	–
Upper Bound	Chen et al. [16]	$\tilde{\mathcal{O}}(d\sqrt{T})$	Stochastic	Non-uniform	$\Theta(1)$	Intractable
	Oh and Iyengar [41]	$\tilde{\mathcal{O}}(d\sqrt{T}/\kappa)$	Stochastic	Non-uniform	$\Theta(1)$	$\mathcal{O}(t)$
	Oh and Iyengar [40]	$\tilde{\mathcal{O}}(d^{3/2}\sqrt{T}/\kappa)$	Adversarial	Non-uniform	$\Theta(1)$	$\mathcal{O}(t)$
	Perivier and Goyal [44]	$\tilde{\mathcal{O}}(dK\sqrt{\kappa' T})$	Adversarial	Uniform	$\Theta(1)$	Intractable
	This work (Theorem 2)	$\tilde{\mathcal{O}}\left(\frac{\sqrt{v_0 K}}{v_0+K} d\sqrt{T}\right)$	Adversarial	Uniform	Any value	$\mathcal{O}(1)$
	This work (Theorem 4)	$\tilde{\mathcal{O}}(d\sqrt{T})$	Adversarial	Non-uniform	$\Theta(1)$	$\mathcal{O}(1)$

Regarding regret upper bounds, Chen et al. [16] proposed an exponential runtime algorithm that achieves a regret of $\tilde{\mathcal{O}}(d\sqrt{T})$ in the setting with *stochastic* contexts and the *non-uniform* rewards. Under the same setting, Oh and Iyengar [41] and Oh and Iyengar [40] introduced polynomial-time algorithms that attain regrets of $\tilde{\mathcal{O}}(d\sqrt{T}/\kappa)$ and $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T}/\kappa)$ respectively, where $1/\kappa = \mathcal{O}(K^2)$ is a problem-dependent constant. Recently, Perivier and Goyal [44] improved the dependency on κ in the *adversarial* context setting, achieving a regret of $\tilde{\mathcal{O}}(dK\sqrt{\kappa' T})$, where $\kappa' = \mathcal{O}(1/K)$. However, their approach focuses solely on the setting with *uniform* rewards, which is a special case of non-uniform rewards, and currently, there is no tractable method to implement the algorithm.

As summarized in Table 1, there has been a gap between the upper and lower bounds in the existing works of contextual MNL bandits. No previous studies have confirmed whether lower or upper bounds are tight, obscuring what the optimal regret should be. This ambiguity is further exacerbated because many studies introduce their methods under varying conditions such as different reward structures and values of v_0, without explicitly explaining how these factors impact regret. Additionally, there is currently no computationally efficient algorithm whose regret does not scale with $1/\kappa = \mathcal{O}(K^2)$ or directly with K. Intuitively, increasing K provides more information at least in the uniform reward setting, potentially leading to a more statistically efficient learning process. However, no previous results have reflected such intuition. Hence, the following research questions arise:

- *What is the optimal regret lower bound in contextual MNL bandits?*
- *Can we design a computationally efficient, nearly minimax optimal algorithm under the adversarial context setting?*

In this paper, we affirmatively answer the questions by first tackling the contextual MNL bandit problem separately based on the structure of rewards—uniform and non-uniform—and the value of the outside option v_0. In the setting of uniform rewards, we establish the tightest regret lower bound, explicitly demonstrating the dependence of regret on v_0. Specifically, we prove a regret lower bound of $\Omega(d\sqrt{T/K})$ when $v_0 = \Theta(1)$, a common assumption in contextual settings [6, 18, 43, 8, 40, 41, 9, 44, 5, 53, 33] (see Appendix C.1 for more details), and a lower bound of $\Omega(d\sqrt{T})$ when $v_0 = \Theta(K)$. Furthermore, in the adversarial context setting, we introduce a computationally efficient and provably optimal (up to logarithmic factors) algorithm, `OFU-MNL+`. We prove that our proposed algorithm achieves a regret of $\tilde{\mathcal{O}}(d\sqrt{T/K})$ when $v_0 = \Theta(1)$ and $\tilde{\mathcal{O}}(d\sqrt{T})$ when $v_0 = \Theta(K)$, each of which matches the respective lower bounds that we establish up to logarithmic factors. Furthermore, in the non-uniform reward setting, we provide the optimal lower bound of $\Omega(d\sqrt{T})$ assuming $v_0 = \Theta(1)$. In the same setting, our proposed algorithm also attains a matching upper bound of $\tilde{\mathcal{O}}(d\sqrt{T})$ up to logarithmic factors. Our main contributions are summarized as follows:

- Under uniform rewards, we establish a regret lower bound of $\Omega(\sqrt{v_0 K}/(v_0+K)d\sqrt{T})$ (Theorem 1), which is the tightest known lower bound in contextual MNL bandits. We propose, for the first time, a computationally efficient and provably optimal algorithm, `OFU-MNL+`, achieving a matching upper bound of $\mathcal{O}(\sqrt{v_0 K}/(v_0+K)d\sqrt{T})$ (Theorem 2) up to logarithmic factors, while requiring only a constant computation cost per round. The results indicate that the regret improves as the assortment size K increases, unless $v_0 = \Theta(K)$. To the best of our knowledge, this is the first study to demonstrate the dependence of regret on the attraction parameter for the outside option v_0 and to highlight the advantages of a larger assortment size K which aligns with intuition. That is, this is the first work to show that a regret upper bound (in either contextual or non-contextual setting) decreases as K increases. Additionally, we provide instance-dependent minimax regret bounds (Proposition 1 and 2), up to logarithmic factors.
- Under non-uniform rewards, with setting $v_0 = \Theta(1)$ following the convention in contextual MNL bandits [6, 18, 43, 8, 40, 41, 9, 44, 5, 53, 33], we establish a regret lower bound of $\Omega(d\sqrt{T})$ (Theorem 3). To the best of our knowledge, this is the first and tightest lower bound established under non-uniform rewards. Moreover, `OFU-MNL+` also achieves a matching upper bound (up to logarithmic factors) of $\tilde{\mathcal{O}}(d\sqrt{T})$ (Theorem 4) in this setting.
- We also conduct numerical experiments and show that our algorithm consistently outperforms the existing MNL bandit algorithms while maintaining a constant computation cost per round. Furthermore, the empirical results corroborate our theoretical findings regarding the dependence of regret on the reward structure, v_0 and K.

Overall, our paper addresses the long-standing open problem of closing the gap between upper and lower bounds for contextual MNL bandits. Our proposed algorithm is the first to achieve both provably optimality (up to logarithmic factors) and practicality with improved computation.

2 Related Work

Lower bounds of MNL bandits. In contextual MNL bandits, to the best of our knowledge, only Chen et al. [16] proved a lower bound of $\Omega(d\sqrt{T}/K)$ with the attraction parameter for the outside option set at $v_0 = 1$. However, in the non-contextual setting, there exist improved lower bounds in terms of K. Agrawal et al. [8] demonstrated a lower bound of $\Omega(\sqrt{NT/K})$, and Chen and Wang [15] established a lower bound of $\Omega(\sqrt{NT})$. By setting $d = N$, one can derive equivalent lower bounds for the contextual setting, specifically $\Omega(\sqrt{dT/K})$ and $\Omega(\sqrt{dT})$, respectively. However, Agrawal et al. [8] and Chen and Wang [15] assumed $v_0 = K$ when establishing their lower bounds, which differs from the setting used by Chen et al. [16], where $v_0 = 1$. Moreover, to the best of our knowledge, all existing works Chen et al. [16], Agrawal et al. [8], Chen and Wang [15] have established the lower bounds under uniform rewards. Consequently, it remains unclear what the optimal regret is, depending on the value of v_0 and the reward structure.

Upper bounds of contextual MNL bandits. Ou et al. [43] formulated a linear utility model and achieved $\tilde{\mathcal{O}}(dK\sqrt{T})$ regret; however, they assumed that utilities are fixed over time. Chen et al. [16] considered contextual MNL bandits with changing and stochastic contexts, establishing a regret of $\tilde{\mathcal{O}}(d\sqrt{T} + d^2K^2)$. However, they encountered computational issues due to the need to enumerate all possible (N choose K) assortments. To address this, Oh and Iyengar [41] proposed a polynomial-time assortment optimization algorithm, which maintains the confidence bounds in the parameter space and then calculates the upper confidence bounds of attraction parameter for each item, achieving a regret of $\tilde{\mathcal{O}}(d\sqrt{T}/\kappa)$, where $1/\kappa = \mathcal{O}(K^2)$ is a problem-dependent constant. Perivier and Goyal [44] considered the adversarial context and uniform reward setting and improved the dependency on κ to $\tilde{\mathcal{O}}(dK\sqrt{\kappa' T} + d^2K^4/\kappa)$, where $\kappa' = \mathcal{O}(1/K)$. However, their algorithm is intractable. Agrawal et al. [5] considered a uniform rewards setting (with $v_0 = 1$) and achieved a regret of $\tilde{\mathcal{O}}(d\sqrt{T})$. However, due to significant technical errors in their paper (refer Appendix L), we do not include a comparison with their results in this work.

Recently, Zhang and Sugiyama [53] utilized an online parameter update to construct a constant time algorithm. However, they consider a *multiple-parameter* choice model in which the learner estimates K parameters and shares the contextual information x_t across the items in the assortment. This model

differs from ours; we use a *single-parameter* choice model with varying the context for each item in the assortment. Additionally, they make a stronger assumption regarding the reward than we do (see Assumption 1). Moreover, while they fix the assortment size at K, we allow it to be smaller than or equal to K. On the other hand, Zhang and Luo [52] considered a general function approximation, achieving a regret bound of $\tilde{\mathcal{O}}(K^{2.5}\sqrt{dNT})$. However, this bound scales with K and N, and the proposed algorithm is not tractable. To the best of our knowledge, all existing methods fail to show that the regret upper bound can improve as the assortment size K increases.

3 Problem Setting

Notations. For a positive integer, n, we denote $[n] := \{1, 2, \dots, n\}$. For a real-valued matrix A, we denote $\|A\|_2 := \sup_{x:\|x\|_2=1} \|Ax\|_2$ as the maximum singular value of A. For two symmetric matrices, V and W of the same dimensions, $V \succeq W$ means that $V - W$ is positive semi-definite. Finally, we define $\mathcal{S}$ to be the set of candidate assortment with size constraint at most K, i.e., $\mathcal{S} = \{S \subseteq [N] : |S| \leqslant K\}$. While, for simplicity, we consider both $\mathcal{S}$ and the set of items $[N]$ to be stationary in this paper, it is important to note that both $\mathcal{S}$ and $[N]$ can vary over time.

Contextual MNL bandits. We consider a sequential assortment selection problem which is defined as follows. At each round t, the agent observes feature vectors $x_{ti} \in \mathbb{R}^d$ for every item $i \in [N]$. Based on this contextual information, the agent presents an assortment $S_t = \{i_1, \dots, i_l\} \in \mathcal{S}$, where $l \leqslant K$, and then observes the user purchase decision $c_t \in S_t \cup \{0\}$, where $\{0\}$ represents the "outside option" which indicates that the user did not select any of the items in S_t. The distribution of these selections follows a multinomial logistic (MNL) choice model [37], where the probability of choosing any item $i_k \in S_t$ (or the outside option) is defined as:

$$p_t(i_k|S_t, \mathbf{w}^\star) := \frac{\exp(x_{ti_k}^\top \mathbf{w}^\star)}{v_0 + \sum_{j \in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}, \quad p_t(0|S_t, \mathbf{w}^\star) := \frac{v_0}{v_0 + \sum_{j \in S_t} \exp(x_{tj}^\top \mathbf{w}^\star)}, \tag{1}$$

where v_0 is a *known* attraction parameter for the outside option and $\mathbf{w}^\star \in \mathbb{R}^d$ is an *unknown* parameter.

Remark 1. *In the existing literature on MNL bandits, it is commonly assumed that $v_0 = 1$ [40, 41, 44, 5, 53]. On the other hand, Chen and Wang [15], Agrawal et al. [8] assume that $v_0 = K$ [1] to induce a tighter lower bound in terms of K. Later, we will explore how these differing assumptions create fundamentally different problems, leading to different regret lower bounds (Subsection 5.1).*

The choice response for each item $i \in S_t \cup \{0\}$ is defined as $y_{ti} := \mathbb{1}(c_t = i) \in \{0, 1\}$. Hence, the choice feedback variable $\mathbf{y}_t := (y_{t0}, y_{ti_1}, \dots y_{ti_l})$ is sampled from the following multinomial (MNL) distribution: $\mathbf{y}_t \sim \mathrm{MNL}\{1, (p_t(0|S_t, \mathbf{w}^\star), \dots, p_t(i_l|S_t, \mathbf{w}^\star))\}$, where the parameter 1 indicates that $\mathbf{y}_t$ is a single-trial sample, i.e., $y_{t0} + \sum_{k=1}^{l} y_{ti_k} = 1$. For each $i \in S_t \cup \{0\}$, we define the noise $\epsilon_{ti} := y_{ti} - p_t(i|S_t, \mathbf{w}^\star)$. Since each ϵ_{ti} is a bounded random variable in $[0, 1]$, ϵ_{ti} is $1/4$-sub-Gaussian. At every round t, the reward r_{ti} for each item i is also given. Then, we define the expected revenue of the assortment S as

$$R_t(S, \mathbf{w}^\star) := \sum_{i \in S} p_t(i|S, \mathbf{w}^\star) r_{ti} = \frac{\sum_{i \in S} \exp(x_{ti}^\top \mathbf{w}^\star) r_{ti}}{v_0 + \sum_{j \in S} \exp(x_{tj}^\top \mathbf{w}^\star)}$$

and define $S_t^\star$ as the offline optimal assortment at time t when $\mathbf{w}^\star$ is known a prior, i.e., $S_t^\star = \mathrm{argmax}_{S \in \mathcal{S}} R_t(S, \mathbf{w}^\star)$. Our objective is to minimize the cumulative regret over the T periods:

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \sum_{t=1}^{T} R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star).$$

When $K = 1$, $r_{t1} = 1$, and $v_0 = 1$, the MNL bandit recovers the binary logistic bandit with $R_t(S = \{x\}, \mathbf{w}^\star) = \sigma\left(x^\top \mathbf{w}^\star\right) = 1/(1 + \exp(-x^\top \mathbf{w}^\star))$, where $\sigma(\cdot)$ is the sigmoid function.

Consistent with previous works on MNL bandits [41, 44, 5, 53], we make the following assumptions:

Assumption 1 (Bounded assumption)**.** *We assume that $\|\mathbf{w}^\star\|_2 \leqslant 1$, and for all $t \geqslant 1$, $i \in [N]$, $\|x_{ti}\|_2 \leqslant 1$ and $r_{ti} \in [0, 1]$.*

[1] Chen and Wang [15] indeed set $v_0 = 1$ and $v_1, \dots, v_N = \Theta(1/K)$. However, this is equivalent to the setting with $v_0 = K$ and $v_1, \dots, v_N = \Theta(1)$.

Assumption 2 (Problem-dependent constant)**.** *There exist $\kappa > 0$ such that for every item $i \in S$ and any $S \in \mathcal{S}$, and all round t, $\min_{\mathbf{w} \in \mathcal{W}} p_t(i|S, \mathbf{w}) p_t(0|S, \mathbf{w}) \geqslant \kappa$, where $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|_2 \leqslant 1\}$.*

In Assumption 1, we assume that the reward for each item i is bounded by a constant, allowing the norm of the reward vector to depend on K, e.g., $\|\boldsymbol{\rho}_t\|_2 \leqslant \sqrt{K}$. In contrast, Zhang and Sugiyama [53] assume that the norm of the reward vector $\boldsymbol{\rho}_t = [r_{t1}, \dots r_{t|S_t|}]^\top \in \mathbb{R}^{|S_t|}$ is bounded by a constant, independent of K, e.g., $\|\boldsymbol{\rho}_t\|_2 \leqslant 1$. Thus, our assumption regarding rewards is weaker than theirs.

Assumption 2 is common in contextual MNL bandits [16, 41, 44, 53]. Note that $1/\kappa$ depends on the maximum size of the assortment K, i.e., $1/\kappa = \mathcal{O}(K^2)$. One of the primary goals of this paper is to show that as the assortment size K increases, we can achieve an improved (or at least not worsened) regret bound. To this end, we design a dynamic assortment policy that enjoys improved dependence on κ. Note that our algorithm does not need to know κ a priori, whereas Oh and Iyengar [40, 41] do.

4 Existing Gap between Upper and Lower Bounds in MNL Bandits

The primary objective of this paper is to establish minimax regrets in contextual MNL bandits. To explore the optimality of regret, we analyze how it depends on the attraction parameter for the outside option v_0, the maximum assortment size K, and the structure of rewards.

Dependence on v_0. Currently, the established lower bounds are $\Omega(d\sqrt{T}/K)$ by Chen et al. [16], $\Omega(\sqrt{dT/K})$ by the contextual version of Agrawal et al. [8], and $\Omega(\sqrt{dT})$, which is the tightest in terms of K, by the contextual version of Chen and Wang [15]. These results can be misleading, as many subsequent studies [41, 39, 17, 52] have claimed that a K-independent regret is achievable, without clearly addressing the influence of the value of v_0. In fact, the improved regret bounds (in terms of K) obtained by Agrawal et al. [8] and Chen and Wang [15] were possible when $v_0 = K$. However, in the contextual setting, it is more common to set $v_0 = \Theta(1)$. This is because, given the context for the outside option x_{t0}, it is straightforward to construct an equivalent choice model where $v_0 = \Theta(1)$ (refer Appendix C.1). In this paper, we rigorously show the regret dependency on the value of v_0. In Theorem 1, we establish a regret lower bound of $\Omega\left(\sqrt{v_0 K}/(v_0 + K)d\sqrt{T}\right)$, which implies that the value of v_0, indeed, affects the regret. Then, in Theorem 2, we show that our proposed computationally efficient algorithm, `OFU-MNL+` achieves a regret of $\tilde{\mathcal{O}}\left(\sqrt{v_0 K}/(v_0 + K)d\sqrt{T}\right)$, which is minimax optimal up to logarithmic factors in terms of all d, T, K and even v_0.

Dependence on K & Uniform/Non-uniform rewards. To the best of our knowledge, the regret bound in all existing works in contextual MNL bandits does not decrease as the assortment size K increases [16, 40, 41, 44]. However, intuitively, as the assortment size increases, we can gain more information because we receive more feedback. Therefore, it makes sense that regret could be reduced as K increases, at least in the uniform reward setting. Under uniform rewards, the expected revenue (to be specified later) increases as more items are added in the assortment. Consequently, both the optimistically chosen assortment and the optimal assortment always have a size of K. Thus, the agent obtain information about exactly K items in each round. This phenomenon is also demonstrated empirically in Figure 1. In the uniform reward setting, as K increases, the cumulative regrets of not only our proposed algorithm but also other baseline algorithms decrease. This indicates that the existing regret bounds are not tight enough in terms of K. Conversely, in the non-uniform reward setting, the sizes of both the optimistically chosen assortment and the optimal assortment can be less than K, so performance improvement is not guaranteed. In this paper, we show that the regret dependence on K varies by case: uniform and non-uniform rewards. When $v_0 = \Theta(1)$, we obtain a regret lower bound of $\Omega(d\sqrt{T/K})$ (Theorem 1) and a regret upper bound of $\tilde{\mathcal{O}}(d\sqrt{T/K})$ (Theorem 2) under uniform rewards. Additionally, we achieve a regret lower bound of (Theorem 3) and a regret upper bound of $\tilde{\mathcal{O}}(d\sqrt{T})$ (Theorem 4) under non-uniform rewards.

5 Algorithms and Main Results

In this section, we begin by proving the tightest regret lower bound under uniform rewards (Subsection 5.1), explicitly showing the dependence on the attraction parameter for the outside option v_0. We then introduce `OFU-MNL+`, an algorithm that achieves minimax optimality, up to logarithmic factors, under *uniform rewards* (Subsection 5.2). Notably, `OFU-MNL+` is designed for efficiency, requiring

only constant computation and storage costs. Finally, we establish the tightest regret lower bound and a matching minimax optimal regret upper bound under *non-uniform rewards* (Subsection 5.3).

5.1 Regret Lower Bound under Uniform Rewards

In this subsection, we present a lower bound for the worst-case expected regret in the uniform reward setting ($r_{ti} = 1$). This covers all applications where the objective is to maximize the appropriate "click-through rate" by offering the assortment.

Theorem 1 (Regret lower bound, Uniform rewards)**.** *Let d be divisible by 4 and let Assumption 1 hold true. Suppose $T \geqslant C \cdot d^4(v_0 + K)/K$ for some constant $C > 0$. Then, in the uniform reward setting, for any policy π, there exists a worst-case problem instance with $N = \Theta(K \cdot 2^d)$ items such that the worst-case expected regret of π is lower bounded as follows:*

$$\sup_{\mathbf{w}} \mathbb{E}^{\pi}_{\mathbf{w}} \left[\mathbf{Reg}_T(\mathbf{w})\right] = \Omega\left(\frac{\sqrt{v_0 K}}{v_0 + K} \cdot d\sqrt{T}\right).$$

Discussion of Theorem 1. If $v_0 = \Theta(1)$, Theorem 1 demonstrates a regret lower bound of $\Omega(d\sqrt{T/K})$. This indicates that, under uniform rewards, increasing the assortment size K leads to an improvement in regret. Compared to the lower bound $\Omega(d\sqrt{T}/K)$ proposed by Chen et al. [16], our lower bound is improved by a factor of $\sqrt{K}$. This improvement is mainly due to the establishment of a tighter upper bound for the KL divergence (Lemma D.2). Notably, Chen et al. [16] also considered uniform rewards with $v_0 = \Theta(1)$. On the other hand, Chen and Wang [15] and Agrawal et al. [8] established regret lower bounds of $\Omega(\sqrt{NT})$ and $\Omega(\sqrt{NT/K})$, respectively, in non-contextual MNL bandits with uniform rewards, by setting $v_0 = K$ to achieve these regrets. Theorem 1 shows that if $v_0 = \Theta(K)$, we can obtain a regret lower bound of $\Omega(d\sqrt{T})$, which is consistent with the K-independent regret in Chen and Wang [15]. To the best of our knowledge, this result is the first to explicitly show the dependency of regret on the attraction parameter for the outside option v_0. The proof is deferred to Appendix D.

5.2 Minimax Optimal Regret Upper Bound under Uniform Rewards

In this subsection, we propose a new algorithm `OFU-MNL+`, which enjoys minimax optimal regret up to logarithmic factors in the case of uniform rewards. Note that, since the revenue is an increasing function when rewards are uniform, maximizing the expected revenue $R_t(S, \mathbf{w})$ over all $S \in \mathcal{S}$ always yields exactly K items, i.e., $|S_t| = |S_t^\star| = K$.

Our first step involves constructing the confidence set for the online parameter.

Online parameter estimation. Instead of performing MLE as in previous works [16, 41, 44], inspired by Zhang and Sugiyama [53], we use the online mirror descent algorithm to estimate parameter. We first define the multinomial logistic loss function at round t as:

$$\ell_t(\mathbf{w}) := -\sum_{i \in S_t} y_{ti} \log p_t(i|S_t, \mathbf{w}). \tag{2}$$

In Proposition C.1, we will show that the loss function has the constant parameter self-concordant-like property. We estimate the true parameter $\mathbf{w}^\star$ as follows:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \langle \nabla \ell_t(\mathbf{w}_t), \mathbf{w} \rangle + \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_t\|^2_{\tilde{H}_t}, \quad \forall t \geqslant 1, \tag{3}$$

where $\eta > 0$ is the step-size parameter to be specified later, and $\tilde{H}_t := H_t + \eta \mathcal{G}_t(\mathbf{w}_t)$, where

$$\mathcal{G}_t(\mathbf{w}) := \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_t(i|S_t, \mathbf{w}) p_t(j|S_t, \mathbf{w}) x_{ti} x_{tj}^\top,$$

and $H_t := \lambda \mathbf{I}_d + \sum_{s=1}^{t-1} \mathcal{G}_s(\mathbf{w}_{s+1})$. Note that $\mathcal{G}_t(\mathbf{w}) = \nabla^2 \ell_t(\mathbf{w})$. This online estimator is efficient in terms of both computation and storage. By a standard online mirror descent formulation [42], the optimization problem in (3) can be solved using a single projected gradient step through the following equivalent formula:

$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta \tilde{H}_t^{-1} \nabla \ell_t(\mathbf{w}_t), \quad \text{and} \quad \mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w} - \mathbf{w}'_{t+1}\|_{\tilde{H}_t}, \tag{4}$$

Algorithm 1 OFU-MNL+

Inputs: regularization parameter λ, probability δ, confidence radius $\beta_t(\delta)$, step size η.
Initialize: $H_1 = \lambda \mathbf{I}_d$ and $\mathbf{w}_1$ as any point in $\mathcal{W}$.
for round $t = 1, 2, \cdots, T$ **do**
    Compute $\alpha_{ti} = x_{ti}^\top \mathbf{w}_t + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$ for all $i \in [N]$.
    Offer $S_t = \operatorname{argmax}_{S \in \mathcal{S}} \tilde{R}_t(S)$ and observe $\mathbf{y}_t$.
    Update $\tilde{H}_t = H_t + \eta \mathcal{G}_t(\mathbf{w}_t)$, and update the estimator $\mathbf{w}_{t+1}$ by (3).
    Update $H_{t+1} = H_t + \mathcal{G}_t(\mathbf{w}_{t+1})$.
end for

which enjoys a computational cost of only $\mathcal{O}(Kd^3)$, completely independent of t [38, 53]. Regarding storage costs, the estimator does not need to store all historical data because both $\tilde{H}_t$ and H_t can be updated incrementally, requiring only $\mathcal{O}(d^2)$ storage.

Furthermore, the estimator allows for a κ-independent confidence set, leading to an improved regret.

Lemma 1 (Online parameter confidence set)**.** *Let $\delta \in (0, 1]$. Under Assumption 1, with $\eta = \frac{1}{2}\log(K+1) + 2$ and $\lambda = 84\sqrt{2}d\eta$, we define the following confidence set*

$$\mathcal{C}_t(\delta) := \{\mathbf{w} \in \mathcal{W} \mid \|\mathbf{w}_t - \mathbf{w}\|_{H_t} \leqslant \beta_t(\delta)\},$$

where $\beta_t(\delta) = \mathcal{O}\left(\sqrt{d}\log t \log K\right)$. Then, we have $\Pr[\forall t \geqslant 1, \mathbf{w}^\star \in \mathcal{C}_t(\delta)] \geqslant 1 - \delta$.

Armed with the online estimator, we construct the computationally efficient optimistic revenue.

Computationally efficient optimistic expected revenue. To balance the exploration and exploitation trade-off, we use the upper confidence bounds (UCB) technique, which have been widely studied in many bandit problems, including K-arm bandits [11, 32] and linear bandits [1, 20].

At each time t, given the confidence set in Lemma 1, we first calculate the optimistic utility α_{ti} as:

$$\alpha_{ti} := x_{ti}^\top \mathbf{w}_t + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}, \quad \forall i \in [N]. \tag{5}$$

The optimistic utility α_{ti} is composed of two parts: the mean utility estimate $x_{ti}^\top \mathbf{w}_t$ and the standard deviation $\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$. In the proof of the regret upper bound, we show that α_{ti} serves as an upper bound for $x_{ti}^\top \mathbf{w}^\star$, assuming that the true parameter $\mathbf{w}^\star$ falls within the confidence set $\mathcal{C}_t(\delta)$. Based on α_{ti}, we construct the optimistic expected revenue for the assortment S, defined as follows:

$$\tilde{R}_t(S) := \frac{\sum_{i \in S} \exp(\alpha_{ti}) r_{ti}}{v_0 + \sum_{j \in S} \exp(\alpha_{tj})}, \tag{6}$$

where $r_{ti} = 1$. Then, we offer the set S_t that maximizes the optimistic expected revenue, $S_t = \operatorname{argmax}_{S \in \mathcal{S}} \tilde{R}_t(S)$. Given our assumption that all rewards are of unit value, the optimization problem is equivalent to selecting the K items with the highest optimistic utility α_{ti}. Consequently, solving the optimization problem incurs a constant computational cost of $\mathcal{O}(N)$.

Remark 2 (Comparison to Zhang and Sugiyama [53])**.** *In Zhang and Sugiyama [53], the MNL choice model is outlined with a shared context x_t and distinct parameters $\mathbf{w}_1^\star, \ldots, \mathbf{w}_K^\star$ for each choice. Conversely, our model employs a single parameter $\mathbf{w}^\star$ across all choices and has varying contexts for each item in the assortment S, $x_{t1}, \ldots x_{ti_{|S|}}$. Due to this discrepancy in the choice model, directly applying Proposition 1 from Zhang and Sugiyama [53], which constructs the optimistic revenue by adding bonus terms to the estimated revenue, incurs an exponential computational cost in our problem setting. This complexity arises because the optimistic revenue must be calculated for every possible assortment $S \in \mathcal{S}$; therefore, it is necessary to enumerate all potential assortments (N choose K) to identify the one that maximizes the optimistic revenue As a result, extending the approach in Zhang and Sugiyama [53] to our setting is non-trivial, requiring a different analysis.*

We now present the regret upper bound of OFU-MNL+ in the uniform reward setting.

Theorem 2 (Regret upper bound of OFU-MNL+, Uniform rewards)**.** *Let $\delta \in (0, 1]$ and Assumptions 1 and 2 hold. In the uniform reward setting, by setting $\eta = \frac{1}{2}\log(K+1) + 2$ and $\lambda = 84\sqrt{2}d\eta$, with*

probability at least $1-\delta$*, the cumulative regret of* OFU-MNL+ *is upper-bounded by*

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(\frac{\sqrt{v_0 K}}{v_0 + K} \cdot d\sqrt{T} + \frac{1}{\kappa} d^2\right).$$

Discussion of Theorem 2. If $T \geqslant \mathcal{O}(d^2(v_0+K)^2/(\kappa^2 v_0 K))$, Theorem 2 shows that our algorithm OFU-MNL+ achieves minimax optimal regret (up to logarithmic factor) in terms of all d, T, K, and even v_0. To the best of our knowledge, ignoring logarithmic factors, our proposed algorithm is the first computationally efficient, minimax optimal algorithm in (adversarial) contextual MNL bandits. When $v_0 = \Theta(1)$, which is the convention in existing MNL bandit literature [40, 41, 44, 5, 53], OFU-MNL+ obtains $\tilde{\mathcal{O}}(d\sqrt{T/K})$ regret. This represents an improvement over the previous upper bound of Perivier and Goyal [44] [2], which is $\tilde{\mathcal{O}}(dK\sqrt{\kappa' T} + d^2 K^4/\kappa)$, where $\kappa' = \mathcal{O}(1/K)$, by a factor of K. This improvement can be attributed to two key factors: an improved, constant, self-concordant-like property of the loss function (Proposition C.1) and a K-free elliptical potential lemma (Lemma E.2). Furthermore, by employing an improved bound for the second derivative of the revenue (Lemma E.3), we achieve an enhancement in the regret for the second term, d^2/κ, by a factor of K^4, in comparison to Perivier and Goyal [44]. Unless $v_0 = \Theta(K)$, Theorem 2 indicates that the regret decreases as the assortment size K increases. To the best of our knowledge, this is the first algorithm in MNL bandits to show that increasing K results in a reduction in regret. Moreover, when reduced to the logistic bandit, i.e., $K = 1$, $r_{t1} = 1$, and $v_0 = 1$, our algorithm can also achieve a regret of $\tilde{\mathcal{O}}(d\sqrt{\kappa T})$ by Corollary 1 in Zhang and Sugiyama [53], which is consistent with the results in Abeille et al. [4], Faury et al. [23]. The proof is deferred to Appendix E.

Remark 3 (Efficiency of OFU-MNL+)**.** *The proposed algorithm is computationally efficient in both parameter updates and assortment selections. Since we employ online parameter estimation, akin to Zhang and Sugiyama [53], our algorithm demonstrates computational efficiency in parameter estimation, incurring only incurring* $\mathcal{O}(Kd^3)$ *computation cost and* $\mathcal{O}(d^2)$ *storage cost, which are completely independent of* t*. Furthermore, a naive approach to selecting the optimistic assortment requires enumerating all possible (*N *choose* K*) assortments, resulting in exponential computational cost [16]. However, by constructing the optimistic expected revenue according to* (6) *(inspired by Oh and Iyengar [41]), our algorithm needs only* $\mathcal{O}(N)$ *computational cost.*

5.3 Regret Upper & Lower Bounds under Non-Uniform Rewards

In this subsection, we propose regret upper and lower bounds in the non-uniform reward setting. In this scenario, the sizes of both the chosen assortment S_t, and the optimal assortment, $S_t^\star$ are not fixed at K. Therefore, we cannot guarantee an improvement in regret even as K increases.

We first prove the regret lower bound in the non-uniform reward setting.

Theorem 3 (Regret lower bound, Non-uniform rewards)**.** *Under the same conditions as Theorem 1, let the rewards be non-uniform and* $v_0 = \Theta(1)$*. Then, for any policy* π*, there exists a worst-case problem instance such that the worst-case expected regret of* π *is lower bounded as follows:*

$$\sup_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}^{\pi}\left[\mathbf{Reg}_T(\mathbf{w})\right] = \Omega\left(d\sqrt{T}\right).$$

Discussion of Theorem 3. In contrast to Theorem 1, which considers uniform rewards, the regret lower bound is independent of the assortment size K. Note that Theorem 3 does not claim that non-uniform rewards inherently make the problem more difficult. Rather, it implies that there exists an instance with *adversarial* non-uniform rewards, where regret does not improve even with an increase in K. Moreover, the assumption that $v_0 = \Theta(1)$ is common in the existing literature on contextual MNL bandits [40, 41, 44, 5, 53] (refer Appendix C.1). To the best of our knowledge, this is the first established lower bound for non-uniform rewards in MNL bandits, even in the non-contextual setting. The proof is deferred to Appendix G.

We also prove a matching upper bound up to logarithmic factors. The algorithm OFU-MNL+ is also applicable in the case of non-uniform rewards. However, because the optimistic expected revenue $\tilde{R}_t(S)$ is no longer an increasing function of α_{ti}, optimizing for $S_t = \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$ no longer equates to simply selecting the top K items with the highest optimistic utility. Instead, we employ

[2] Perivier and Goyal [44] also consider the uniform rewards ($r_{ti} = 1$) with $v_0 = 1$.

assortment optimization methods introduced in Rusmevichientong et al. [47], Davis et al. [21], which are efficient polynomial-time (independent of t) [3] algorithms available for solving this optimization problem. Therefore, our algorithm is also computationally efficient under non-uniform rewards.

Theorem 4 (Regret upper bound of OFU-MNL+, Non-uniform rewards)**.** *Under the same assumptions and parameter settings as Theorem 2, if the rewards are non-uniform and $v_0 = \Theta(1)$, then with a probability at least $1-\delta$, the cumulative regret of* OFU-MNL+ *is upper-bounded by*

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\kappa}d^2\right).$$

Discussion of Theorem 4. If $T \geqslant \mathcal{O}(d^2/\kappa^2)$, our algorithm achieves a regret of $\tilde{\mathcal{O}}(d\sqrt{T})$ when the reward for each item is non-uniform, demonstrating that OFU-MNL+ is minimax optimal up to a logarithmic factor. Recall that we relax the bounded assumption on the reward compared to Zhang and Sugiyama [53] (refer Assumption 1); thus, we allow the sum of the squared rewards in the assortment to scale with K. Consequently, we need a novel approach to achieve the regret that does not scale with K. To this end, we *centralize* the features, i.e., $\tilde{x}_{ti} = x_{ti} - \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}_{t+1})}[x_{tj}]$, and propose a novel elliptical potential lemma for them, as detailed in Lemma H.3. Note that our algorithm is capable of achieving $1/\kappa$-free regret (in the leading term) under both uniform and non-uniform rewards. In contrast, the algorithm in Perivier and Goyal [44] is limited to achieving this only in the uniform reward setting. Furthermore, compared to the regret bound in Chen et al. [16], which is $\tilde{\mathcal{O}}(d\sqrt{T})$, our regret bounds has the same order of regret with theirs. However, their algorithm is computationally intractable as it requires enumerating all possible assortments, whereas our algorithm incurs only a constant computational cost per round. The proof is deferred to Appendix H.

6 Instance-Dependent Bounds

In this section, we show that instance-dependent upper and lower bounds are also achievable under uniform rewards. We define the degree of non-linearity for the optimal assortment $S_t^\star$ at round t under the true parameter $\mathbf{w}^\star$ as $\kappa_t^\star := \sum_{i\in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star)p_t(0|S_t^\star, \mathbf{w}^\star)$. We first establish the instance-dependent lower bound under uniform rewards.

Proposition 1 (Instance-dependent regret lower bound, Uniform rewards)**.** *Under the same conditions as Theorem 1, for any policy π and for $T \geqslant d^2/\kappa$, there exists a worst-case problem instance such that the worst-case expected regret of π is lower bounded as follows:*

$$\sup_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}^{\pi}\left[\mathbf{Reg}_T(\mathbf{w})\right] = \Omega\left(d\sqrt{\sum_{t=1}^{T}\kappa_t^\star}\right).$$

The proof is deferred to Appendix I. We also provide the matching upper bound.

Proposition 2 (Instance-dependent regret upper bound of OFU-MNL+, Uniform rewards)**.** *Under the same assumptions, parameter settings, and reward structure as Theorem 2, with a probability at least $1-\delta$, the cumulative regret of* OFU-MNL+ *is upper-bounded by*

$$\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\kappa_t^\star} + \frac{1}{\kappa}d^2\right).$$

The proof is deferred to Appendix J. For sufficiently large T, the regret upper bound (Proposition 2) matches the regret lower bound (Proposition 1), up to logarithmic factor. To the best of our knowledge, these are the first minimax instance-dependent regret bounds under uniform rewards. Note that, in the worst case, $\kappa_t^\star = \tilde{\mathcal{O}}(\sqrt{v_0 K}/(v_0+K))$, which indicates that these results provide a strict improvement over the worst-case bounds given in Theorems 1 and 2.

Some readers may expect instance-dependent regret bounds for non-uniform rewards as well. Unfortunately, we were unable to establish these. Recall that $\kappa_t^\star$ represents the degree of non-linearity for

[3] An interior point method would generally solve the problem with a computational complexity of $\mathcal{O}(N^{3.5})$.

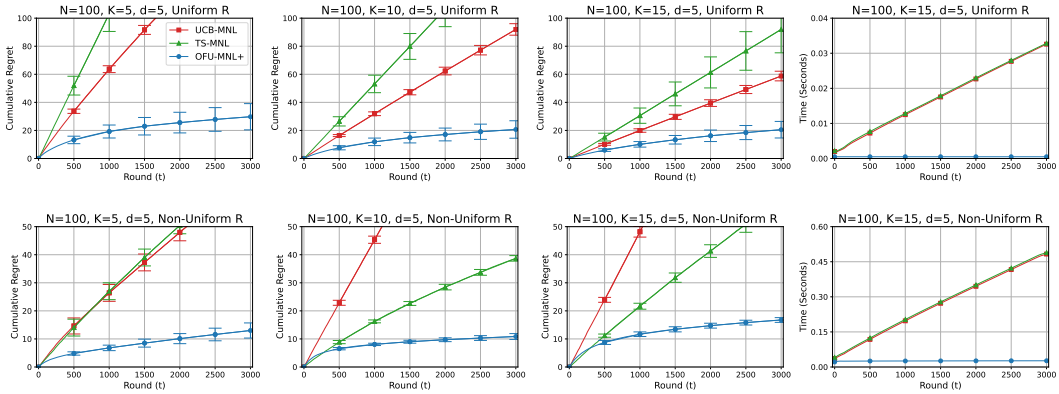

Figure 1: Cumulative regret (left three, $K = 5, 10, 15$) and runtime per round (rightmost one, $K = 15$) under uniform rewards (first row) and non-uniform rewards (second row) with $v_0 = 1$.

the optimal assortment $S_t^\star$. However, in the proofs for regret bounds, we encounter terms associated with the chosen assortment S_t, such as $\sum_{i \in S_t^\star} p_t(i|S_t, \mathbf{w}^\star) p_t(0|S_t, \mathbf{w}^\star)$. To address this, we use the mean value theorem-based analysis (Lemma I.5) to replace this quantity with $\kappa_t^\star$. Under non-uniform rewards, however, the mean value theorem does not apply because the sizes and rewards of $S_t^\star$ and S_t may differ. Addressing this problem would be an interesting direction for future research.

7 Numerical Experiments

In this section, we empirically evaluate the performance of our algorithm `OFU-MNL+`. We measure cumulative regret over $T = 3000$ rounds. For each experimental setup, we run the algorithms across 20 independent instances and report the average performance. In each instance, the underlying parameter $\mathbf{w}^\star$ is randomly sampled from a d-dimensional uniform distribution, where each element of $\mathbf{w}^\star$ lies within the range $[-1/\sqrt{d}, 1/\sqrt{d}]$ and is not known to the algorithms. Additionally, the context features x_{ti} are drawn from a d-dimensional multivariate Gaussian distribution, with each element of x_{ti} clipped to the range $[-1/\sqrt{d}, 1/\sqrt{d}]$. This setup ensures compliance with Assumption 1. In the uniform reward setting (first row of Figure 1), the combinatorial optimization step to choose the assortment reduces to sorting items by their utility estimate. In the non-uniform reward setting (second row of Figure 1), the rewards are sampled from a uniform distribution in each round, i.e., $r_{ti} \sim \mathrm{Unif}(0, 1)$. Refer Appendix K for more details.

We compare the performance of `OFU-MNL+` with those of the practical and state-of-the-art algorithms: the Upper Confidence Bound-based algorithm, `UCB-MNL` [40], and the Thompson Sampling-based algorithm, `TS-MNL` [40]. Figure 1 demonstrates that our algorithm significantly outperforms other baseline algorithms. In the uniform reward setting, as K increases, the cumulative regrets of all algorithms tend to decrease. In contrast, this trend is not observed in the non-uniform reward setting. Furthermore, the results also show that our algorithm maintains a constant computation cost per round, while the other algorithms exhibit a linear dependence on t. In Appendix K, we present the additional runtime curves (Figure K.1) as well as the regret curves of the other configuration where $v_0 = \Theta(K)$ (Figure K.2). All of these empirical results align with our theoretical results.

8 Conclusion

In this paper, we propose minimax optimal lower and upper bounds for both uniform and non-uniform reward settings. We propose a computationally efficient algorithm, `OFU-MNL+`, that achieves a regret of $\tilde{\mathcal{O}}(d\sqrt{T/K})$ under uniform rewards and $\tilde{\mathcal{O}}(d\sqrt{T})$ under non-uniform rewards. We also prove matching lower bounds of $\Omega(d\sqrt{T/K})$ and $\Omega(d\sqrt{T})$ for each setting, respectively. Moreover, our empirical results support our theoretical findings, demonstrating that `OFU-MNL+` is not only provably but also experimentally efficient.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1006859, 2022R1A4A1030579, and RS-2023-00222663) 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] Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

[3] Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017.

[4] 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.

[5] 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.

[6] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[7] 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.

[8] 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.

[9] Sanae Amani and Christos Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34:2913–2924, 2021.

[10] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

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

[12] Nicolo Campolongo and Francesco Orabona. Temporal variability in implicit online learning. *Advances in neural information processing systems*, 33:12377–12387, 2020.

[13] Junyu Cao and Wei Sun. Tiered assortment: Optimization and online learning. *Management Science*, 70(8):5481–5501, 2024.

[14] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

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

[16] 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.

[17] Xi Chen, Yining Wang, and Yuan Zhou. Optimal policy for dynamic assortment planning under multinomial logit models. *Mathematics of Operations Research*, 46(4):1639–1657, 2021.

[18] Wang Chi Cheung and David Simchi-Levi. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658*, 2017.

[19] Hyun-jun Choi, Rajan Udwani, and Min-hwan Oh. Cascading contextual assortment bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[21] James M Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

[22] 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.

[23] 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.

[24] Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 586–594, Red Hook, NY, USA, 2010. Curran Associates Inc.

[25] Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on learning theory*, pages 167–208. PMLR, 2018.

[26] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[27] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

[28] Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

[29] Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8300–8307, 2023.

[30] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

[31] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

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

[33] 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.

[34] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024.

[35] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[36] Xutong Liu, Jinhang Zuo, Siwei Wang, John CS Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*, pages 22559–22593. PMLR, 2023.

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

[38] 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.

[39] Sentao Miao and Xiuli Chao. Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management*, 23(2):525–545, 2021.

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

[41] 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.

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

[43] 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.

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

[45] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

[46] Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 752–776. PMLR, 08 Feb–11 Feb 2020.

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

[48] Denis Sauré and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[49] Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences: Proceedings of the 3rd International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2015-Part I*, pages 155–168. Springer, 2015.

[50] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008. doi: https://doi.org/10.1007/b13794.

[51] Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pages 392–401. PMLR, 2016.

[52] Mengxiao Zhang and Haipeng Luo. Contextual multinomial logit bandits with general value functions. *arXiv preprint arXiv:2402.08126*, 2024.

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

[54] Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton. Cascading bandits for large-scale recommendation problems. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, page 835–844. AUAI Press, 2016.

Appendix

Table of Contents

A Further Related Work

In this section, we discuss additional related works that complement Section 2. For simplicity, we consider only the dependence on the number of rounds t for a computation cost in big-$\mathcal{O}$ notation.

Logistic Bandits. The logistic bandit model [24, 22, 4, 23] focuses on environments with *binary* rewards and explores the impact of non-linearity on the exploration-exploitation trade-off for parametrized bandits. The main research interest has been the algorithms' dependence on the degree of non-linearity κ, which can grow exponentially in terms of the diameter of the decision domain $\mathcal{W}$. Zhang et al. [51] introduced the first efficient algorithm for binary logistic bandits with a $\mathcal{O}(1)$ computation cost, achieving a regret of $\tilde{\mathcal{O}}(d\sqrt{T/\kappa})$. Faury et al. [22] enhanced the regret to $\tilde{\mathcal{O}}(d\sqrt{T/\kappa})$ with a $\mathcal{O}(t)$ computation cost. However, their regret bounds still suffered from a harmful dependence on $1/\kappa$. Abeille et al. [4] addressed this by achieving the tightest regret upper bound of $\tilde{\mathcal{O}}(d\sqrt{\kappa T})$ with a $\mathcal{O}(t)$ computation cost, while Faury et al. [23] achieved the same regret with an improved computation cost of $\mathcal{O}(\log t)$. More recently, Zhang and Sugiyama [53] proposed a jointly efficient algorithm that achieves the optimal regret with a constant $\mathcal{O}(1)$ computation cost. Note that the logistic bandit is a special case of the multinomial logistic (MNL) bandit. When the maximum assortment size is one ($K = 1$), rewards are uniform ($r_{t1} = 1$), and the attraction parameter for the outside option is one $v_0 = 1$, the MNL bandit reduces to the logistic bandit. In this logistic bandit setting, our proposed algorithm, `OFU-MNL+`, can achieve a regret upper bound of $\tilde{\mathcal{O}}(d\sqrt{\kappa T})$ with a constant $\mathcal{O}(1)$ computation cost, consistent with the result in Zhang and Sugiyama [53].

Multinomial Logistic (MNL) Bandits. There are two main approaches to multinomial logistic (MNL) bandits: the *multiple-parameter* choice model and the *single-parameter* choice model. In the multiple-parameter choice model, the learner estimates parameters for each choice in the assortment $(\mathbf{w}_1^\star, \dots, \mathbf{w}_K^\star)$ with a shared context x_t. In this setting, Amani and Thrampoulidis [9] proposed a feasible algorithm that achieves a regret upper bound of $\tilde{\mathcal{O}}(dK\sqrt{\kappa T})$ with a $\mathcal{O}(t)$ computation cost. They also proposed an intractable algorithm that achieves an improved regret of $\tilde{\mathcal{O}}(dK^{3/2}\sqrt{T})$. Zhang and Sugiyama [53] introduced a computationally and statistically efficient algorithm that obtains a regret of $\tilde{\mathcal{O}}(dK\sqrt{T})$. Recently, Lee et al. [33] further improved the regret by a factor of $\sqrt{K}$, achieving $\tilde{\mathcal{O}}(d\sqrt{KT})$ regret. In the multiple-parameter case, the regret's dependence on K is unavoidable since the number of unknown parameters depends on K.

On the other hand, the single-parameter choice model, closely related to ours, shares the parameter $\mathbf{w}^\star$ cross the choices, with varying contexts for each choice. The learner offers a set of items S_t, with $|S_t| \leqslant K$ at each round. This setting involves a combinatorial optimization to choose the assortment S_t, making it more challenging to devise a tractable algorithm. As extensively discussed in Section 2, no previous studies have definitively confirmed whether the existing lower or upper bounds are tight. As shown in Table 1, many studies have presented their results in inconsistent settings with varying reward structures and values of v_0, adding to the ambiguity about the bounds' optimality. In this paper, we address these issues by bridging the gap between the lower and upper bounds of regret through a careful categorization of the settings. We propose an algorithm that is both provably optimal, up to logarithmic factors, and computationally efficient, significantly enhancing the theoretical and practical understanding of MNL bandits.

Generalized Linear Bandits. In generalized linear bandits [24, 27, 35, 3, 31, 28, 29, 34], the expected rewards are modeled using a generalized linear model. These problems generalize logistic bandits by incorporating a general exponential family link function instead of the logistic link function. The algorithms proposed for generalized linear bandits also exhibit a dependence on the nonlinear term κ. However, our problem setting (single-parameter MNL bandits) considers a more complex state space where multiple arms are pulled simultaneously.

Combinatorial Bandits. Another related stream of literature is combinatorial bandits [14, 45, 30, 54, 46, 36], particularly top-k combinatorial bandits [46]. In top-k combinatorial bandits, the decision set includes all subsets of size k out of n arms, and the reward for each action is the sum of the rewards of the k selected arms. In this framework, the rewards are assumed to be independent of the entire set of arms played in round t. In contrast, in our setting, the reward for each individual arm depends on the whole set of arms played.

Recently, Choi et al. [19], Cao and Sun [13] have considered the cascading assortment bandit problem, which encompasses the MNL bandit problem as a special case where the cascading length

is 1. However, these works do not strictly encompass our results. Choi et al. [19] only consider uniform rewards, and achieve a regret upper bound of $\tilde{\mathcal{O}}(d\sqrt{T})$, which avoids dependence on both the cascading length and κ. When the cascading length is 1, our result (Theorem 2) improves upon theirs by a factor of $\sqrt{K}$. Moreover, their computation cost per round is $\mathcal{O}(t)$ since they employ MLE to estimate the parameter. Cao and Sun [13] consider non-uniform rewards, and achieve a regret upper bound of $\tilde{\mathcal{O}}(h^2 d\sqrt{MT})$, where M is the cascading length and $h \geqslant \frac{p(i|S,\mathbf{w})}{p(i|S,\mathbf{w}')}$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $S \in \mathcal{S}$, and $i \in S \cup \{0\}$. However, their regret bound still suffers from a harmful dependence on h^2, which can be exponentially large.

B Notation

We denote T as the total number of rounds and $t \in [T]$ as the current round. We denote N as the total number of items, K as the maximum size of assortments, and d as the dimension of feature vectors.

For notational simplicity, we define the loss function in two different forms throughout the proof:

$$\ell_t(\mathbf{w}) = -\sum_{i \in S_t} y_{ti} \log p_t(i|S_t, \mathbf{w}) = -\sum_{i \in S_t} y_{ti} \log \left(\frac{\exp(x_{ti}^\top \mathbf{w})}{v_0 + \sum_{j \in S_t} \exp(x_{tj}^\top \mathbf{w})} \right),$$

$$\ell(\mathbf{z}_t, \mathbf{y}_t) = -\sum_{i \in S_t} y_{ti} \log \left(\frac{\exp(z_{ti})}{v_0 + \sum_{j \in S_t} \exp(z_{tj})} \right),$$

where $z_{ti} = x_{ti}^\top \mathbf{w}$, $\mathbf{z}_t = (z_{ti})_{i \in S_t} \in \mathbb{R}^{|S_t|}$, and $\mathbf{y}_t = (y_{ti})_{i \in S_t} \in \mathbb{R}^{|S_t|}$. Thus, $\ell_t(\mathbf{w}) = \ell(\mathbf{z}_t, \mathbf{y}_t)$.

We offer a Table B.1 for convenient reference.

Table B.1: Symbols

x_{ti}	feature vector for item i given at round t
r_{ti}	reward for item i given at round t
S_t	assortment chosen by an algorithm at round t
0	outside option
y_{ti}	choice response for each item $i \in S_t \cup \{0\}$ at round t
$R_t(S, \mathbf{w}^\star)$	$:= \sum_{i \in S} p_t(i\|S, \mathbf{w}^\star) r_{ti}$, expected revenue of the assortment S at round t
$\ell_t(\mathbf{w})$	$:= -\sum_{i \in S_t} y_{ti} \log \left(\frac{\exp(x_{ti}^\top \mathbf{w})}{v_0 + \sum_{j \in S_t} \exp(x_{tj}^\top \mathbf{w})} \right)$, loss function at round t
$\ell(\mathbf{z}_t, \mathbf{y}_t)$	$:= -\sum_{i \in S_t} y_{ti} \log \left(\frac{\exp(z_{ti})}{v_0 + \sum_{j \in S_t} \exp(z_{tj})} \right)$, loss function at round t, $z_{ti} = x_{ti}^\top \mathbf{w}$
λ	regularization parameter
$\mathcal{G}_t(\mathbf{w})$	$:= \sum_{i \in S_t} p_t(i\|S_t, \mathbf{w}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_t(i\|S_t, \mathbf{w}) p_t(j\|S_t, \mathbf{w}) x_{ti} x_{tj}^\top$
H_t	$:= \lambda \mathbf{I}_d + \sum_{s=1}^{t-1} \mathcal{G}_s(\mathbf{w}_{s+1})$
$\tilde{H}_t$	$:= H_t + \eta \mathcal{G}_t(\mathbf{w}_t)$
α_{ti}	$:= x_{ti}^\top \mathbf{w}_t + \beta_t(\delta) \|x_{ti}\|_{H_t^{-1}}$, optimistic utility for item i at round t
$\beta_t(\delta)$	$:= \mathcal{O}\left(\sqrt{d} \log t \log K\right)$, confidence radius at round t
$\tilde{R}_t(S)$	$:= \frac{\sum_{i \in S} \exp(\alpha_{ti}) r_{ti}}{v_0 + \sum_{j \in S} \exp(\alpha_{tj})}$, optimistic expected revenue for the assortment S at round t

C Properties of MNL function

In this section, we present key properties of the MNL function and its associated loss, which are used throughout the paper.

C.1 Attraction parameter for Outside Option: $v_0 = \Theta(1)$ is Common in Contextual MNL Bandits

In this subsection, we explain why the assumption that $v_0 = \Theta(1)$ is made without loss of generality. Let the original feature vectors be $x'_{ti} \in \mathbb{R}^d$ for every item $i \in [N]$. Suppose that a context for the outside option x'_{t0} is given and the probability of choosing any item $i \in S_t \cup \{0\}$ is defined as

$$p_t(i|S_t, \mathbf{w}^\star) = \frac{\exp((x'_{ti})^\top \mathbf{w}^\star)}{\sum_{j \in S_t \cup \{0\}} \exp((x'_{tj})^\top \mathbf{w}^\star)}.$$

Then, by dividing the denominator and numerator by $\exp\left((x'_{t0})^\top \mathbf{w}^\star\right)$, and defining $x_{ti} := x'_{ti} - x'_{t0}$, we obtain the MNL probability in the form presented in (1) with $v_0 = \exp(0) = 1$. Note that this division does not change the probability. Therefore, $v_0 = \Theta(1)$ is natural and common in contextual MNL bandit literature.

C.2 Self-concordant-like Function

Definition C.1 (Self-concordant-like function, Tran-Dinh et al. 49)**.** *A convex function $f \in \mathcal{C}^3(\mathbb{R}^m)$ is M-self-concordant-like function with constant M if:*

$$|\phi'''(s)| \leqslant M\|\mathbf{b}\|_2 \phi''(s).$$

for $s \in \mathbb{R}$ and $M > 0$, where $\phi(s) := f(\mathbf{a} + s\mathbf{b})$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$.

Then, the MNL loss defined in (2) is $3\sqrt{2}$-self-concordant-like function.

Proposition C.1. *For any $t \in [T]$, the multinomial logistic loss $\ell_t(\mathbf{w})$, defined in* (2)*, is $3\sqrt{2}$-self-concordant-like.*

Proof. Consider the function $\phi(s) := \log\left(\sum_{i=0}^n e^{a_i s + b_i}\right)$, where $\mathbf{a} = [a_0, \ldots, a_n]^\top \in \mathbb{R}^{n+1}$ and $\mathbf{b} = [b_0, \ldots, b_n]^\top \in \mathbb{R}^{n+1}$. Then, by simple calculus, we have

$$\phi''(s) = \frac{\sum_{i<j}(a_i - a_j)^2 e^{a_i s + b_i} e^{a_j s + b_j}}{\left(\sum_{i=0}^n e^{a_i s + b_i}\right)^2} \geqslant 0,$$

and

$$\phi'''(s) = \frac{\sum_{i<j}(a_i - a_j)^2 e^{a_i s + b_i} e^{a_j s + b_j} \left[\sum_{k=0}^n (a_i + a_j - 2a_k) e^{a_k s + b_k}\right]}{\left(\sum_{i=0}^n e^{a_i s + b_i}\right)^3}. \tag{C.1}$$

Note that for all $i, j, k = 0, \ldots, n$,

$$|a_i + a_j - 2a_k| \leqslant \sqrt{6}\sqrt{a_i^2 + a_j^2 + a_k^2} \leqslant 3\sqrt{2} \max_{i=0,\ldots,n} |a_i|. \tag{C.2}$$

Therefore, we have

$$\left|\sum_{k=0}^n (a_i + a_j - 2a_k) e^{a_k s + b_k}\right| \leqslant \sum_{k=0}^n |a_i + a_j - 2a_k|\, e^{a_k s + b_k} \leqslant 3\sqrt{2} \max_{i=0,\ldots,n} |a_i| \sum_{k=0}^n e^{a_k s + b_k}. \tag{C.3}$$

Plugging in (C.3) into (C.1), we obtain

$$\phi'''(s) \leqslant 3\sqrt{2} \max_{i=0,\ldots,n} |a_i| \phi''(s). \tag{C.4}$$

Now, we are ready to prove the proposition. For any $t \in [T]$, let $n = |S_t|$ and $c_1 = x_{ti_1}, c_2 = x_{ti_2}, \ldots, c_n = x_{ti_n}$. Define a function $f \in \mathcal{C}^3 : \mathbb{R}^d \to \mathbb{R}$ as $f(\boldsymbol{\theta}) := \log\left(v_0 + \sum_{i=1}^n e^{c_i^\top \boldsymbol{\theta}}\right)$. Let $\boldsymbol{\delta} \in \mathbb{R}^d$ and let $f(\boldsymbol{\theta} + s\boldsymbol{\delta}) = \log\left(v_0 + \sum_{i=1}^n e^{c_i^\top \boldsymbol{\theta} + s c_i^\top \boldsymbol{\delta}}\right) = \log\left(\sum_{i=0}^n e^{a_i s + b_i}\right) = \phi(s)$, where $a_i = c_i^\top \boldsymbol{\delta}$, $b_i = c_i^\top \boldsymbol{\theta}$ for $i = 1, \ldots, n$, and $a_i = 0$ and $b_i = \log v_0$ for $i = 0$. Then, by (C.4), we get

$$\begin{aligned}
|\phi'''(s)| &\leqslant 3\sqrt{2} \max_{i=0,\ldots,n} |a_i| \phi''(s) = 3\sqrt{2} \max_{i=1,\ldots,n} |c_i^\top \boldsymbol{\delta}| \phi''(s) \\
&\leqslant 3\sqrt{2} \max_{i=1,\ldots,n} \|c_i\|_2 \|\boldsymbol{\delta}\|_2 \phi''(s) \leqslant 3\sqrt{2} \|\boldsymbol{\delta}\|_2 \phi''(s),
\end{aligned}$$

where the last inequality holds due to Assumption 1 that $\|c_i\|_2 = \|x_{tj_i}\|_2 \leqslant 1$. Then, by Definition C.1, f is $3\sqrt{2}$-self-concordant-like. Since ℓ_t is the sum of f and a linear operator, which has third derivatives equal to zero, it follows that ℓ_t is also $3\sqrt{2}$-self-concordant-like function. □

Remark C.1. *Contrary to the findings of Perivier and Goyal [44], which suggest that the MNL loss function $\sqrt{6K}$-self-concordant-like, our loss function is $3\sqrt{2}$-self-concordant-like. This yields an improved regret bound on the order of $\mathcal{O}(\sqrt{K})$. The improvement arises due to a K-independent self-concordant-like property of ℓ_t, as shown in Proposition C.1. In Perivier and Goyal [44], Lemma 4 from Tran-Dinh et al. [49] is used, which describes a $\sqrt{6}\|\mathbf{a}\|_2$ self-concordant-like property. However, in the analysis of C.2, we show that their analysis is not tight because they bound the term $\sqrt{a_i^2 + a_j^2 + a_k^2}$ by $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=0}^{n} a_i^2}$, thus making its upper bound dependent on K, i.e., $n = |S_t| \leqslant K$. In contrast, we bound the same term by a constant, $\max_{i=1,\dots,n} \|a_i\|_2$, which allows our loss function to exhibit a constant $3\sqrt{2}$-self-concordant-like property. This key difference accounts for the $\sqrt{K}$-improved regret.*

Lemma C.1 (Theorem 3 of Tran-Dinh et al. 49)**.** *A convex function $\ell \in \mathcal{C}^3 : \mathbb{R}^d \to \mathbb{R}$ is M-self-concordant-like if and only if for any $\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in \mathbb{R}^d$, we have*

$$|\langle D^3\ell(\mathbf{v})[\mathbf{u}_1]\mathbf{u}_2, \mathbf{u}_3\rangle| \leqslant M\|\mathbf{u}_1\|_2\|\mathbf{u}_2\|_{\nabla^2\ell(\mathbf{v})}\|\mathbf{u}_3\|_{\nabla^2\ell(\mathbf{v})}.$$

D Proof of Theorem 1

In this section, we provide the proof of Theorem 1. The proof structure is similar to the one presented in Chen et al. [16]. However, unlike their approach, we explicitly derive a bound that includes v_0. Furthermore, by establishing a tighter upper bound for the KL divergence (Lemma D.2), we derive a bound that is tighter than the one provided by Chen et al. [16].

D.1 Adversarial Construction and Bayes Risk

Let $\epsilon \in (0, 1/d\sqrt{d})$ be a small positive parameter to be specified later. For every subset $V \subseteq [d]$, we define the corresponding parameter $\mathbf{w}_V \in \mathbb{R}^d$ as $[\mathbf{w}_V]_j = \epsilon$ for all $j \in V$, and $[\mathbf{w}_V]_j = 0$ for all $j \notin V$. Then, we consider the following parameter set

$$\mathbf{w} \in \mathcal{W} := \{\mathbf{w}_V : V \in \mathcal{V}_{d/4}\} := \{\mathbf{w}_V : V \subseteq [d], |V| = d/4\},$$

where $\mathcal{V}_k$ denotes the class of all subsets of $[d]$ whose size is k. Moreover, note that $d/4$ is a positive integer, as d is divisible by 4 by construction.

The context vectors $\{x_{ti}\}$ are constructed to be invariant across rounds t. For each t and $U \in \mathcal{V}_{d/4}$, K identical context vectors [4] x_U are constructed as follows:

$$[x_U]_j = 1/\sqrt{d} \quad \text{for } j \in U; \quad [x_U]_j = 0 \quad \text{for } j \notin U.$$

For all $V, U \in \mathcal{V}_{d/4}$, it can be verified that $\mathbf{w}_V$ and x_U satisfy the requirements of a bounded assumption 1 as follows:

$$\|\mathbf{w}_V\|_2 \leqslant \sqrt{d\epsilon^2} \leqslant 1, \quad \|x_U\|_2 \leqslant \sqrt{d \cdot 1/d} = 1.$$

Therefore, the worst-case expected regret of any policy π can be lower bounded by the worst-case expected regret of parameters belonging to $\mathcal{W}$, which can be further lower bounded by the "average"

[4] Recall that K is the maximum allowed assortment capacity.

regret over a uniform prior over $\mathcal{W}$ as follows:

$$\begin{aligned}
\sup_{\mathbf{w}} \mathbb{E}^{\pi}_{\mathbf{w}} \left[\mathbf{Reg}_T(\mathbf{w})\right] &= \sup_{\mathbf{w}} \mathbb{E}^{\pi}_{\mathbf{w}} \sum_{t=1}^{T} R(S^\star, \mathbf{w}) - R(S_t, \mathbf{w}) \\
&\geqslant \max_{\mathbf{w}_V} \mathbb{E}^{\pi}_{\mathbf{w}_V} \sum_{t=1}^{T} R(S^\star, \mathbf{w}_V) - R(S_t, \mathbf{w}_V) \\
&\geqslant \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}^{\pi}_{\mathbf{w}_V} \sum_{t=1}^{T} R(S^\star, \mathbf{w}_V) - R(S_t, \mathbf{w}_V) \\
&= \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}^{\pi}_{\mathbf{w}_V} \sum_{t=1}^{T} \left[\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in S_t} p(i|S_t, \mathbf{w}_V) \right]. \qquad \text{(D.1)}
\end{aligned}$$

This reduces the task of lower bounding the worst-case regret of any policy to the task of lower bounding the *Bayes risk* of the constructed parameter set.

D.2 Main Proof of Theorem 1

Proof of Theorem 1. For any sequence of assortments $\{S_t\}_{t=1}^T$ produced by policy π, we denote an alternative sequence $\{\tilde{S}_t\}_{t=1}^T$ that provably enjoys less regret under parameterization $\mathbf{w}_V$.

Let $x_{U_1}, \ldots, x_{U_L}$ be the distinct feature vectors contained in assortments S_t (if $S_t = \varnothing$, then one may choose an arbitrary feature x_U) with $U_1, \ldots, U_L \in \mathcal{V}_{d/4}$. Let $U^\star$ be the subset among $U_1, \ldots, U_L$ that maximizes $x_U^\top \mathbf{w}_V$, i.e., $U^\star \in \operatorname{argmax}_{U \in \{U_1, \ldots, U_L\}} x_U^\top \mathbf{w}_V$, where $\mathbf{w}_V$ is the underlying parameter. Then, we define $\tilde{S}_t$ as the assortment consisting of all K items corresponding to feature $x_{U^\star}$, i.e., $\tilde{S}_t = \{\underbrace{x_{U^\star}, \ldots, x_{U^\star}}_{K}\}$.

Since the expected revenue is an increasing function, we have the following observation:

Proposition D.1 (Proposition 1 in Chen et al. 16)**.**

$$\sum_{i \in S_t} p(i|S_t, \mathbf{w}_V) \leqslant \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V).$$

Proposition D.1 implies that $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in S_t} p(i|S_t, \mathbf{w}_V) \geqslant \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V)$. Hence, it is sufficient to bound $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V)$ instead of $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in S_t} p(i|S_t, \mathbf{w}_V)$.

To simplify notation, we denote $\tilde{U}_t$ as the unique $U^\star \in \mathcal{V}_{d/4}$ in $\tilde{S}_t$. We also use $\mathbb{E}_V$ and $\mathbb{P}_V$ to denote the expected value and probability, respectively, as governed by the law parameterized by $\mathbf{w}_V$ and under policy π. Then, we can establish a lower bound for $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V)$ as follows:

Lemma D.1. *Suppose $\epsilon \in (0, 1/d\sqrt{d})$ and define $\delta := d/4 - |\tilde{U}_t \cap V|$. Then, we have*

$$\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \geqslant \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\delta \epsilon}{2\sqrt{d}}.$$

For any $j \in V$, define random variables $\tilde{M}_j := \sum_{t=1}^T \mathbf{1}\{j \in \tilde{U}_t\}$. Then, by Lemma D.1, for all $V \in \mathcal{V}_{d/4}$, we have

$$\mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \geqslant \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{4} - \sum_{j \in V} \mathbb{E}_V[\tilde{M}_j] \right). \qquad \text{(D.2)}$$

Furthermore, we define $\mathcal{V}_{d/4}^{(j)} := \{V \in \mathcal{V}_{d/4} : j \in V\}$ and $\mathcal{V}_{d/4-1} := \{V \subseteq [d] : |V| = d/4 - 1\}$. By taking the average of both sides of Equation (D.2) with respect to all $V \in \mathcal{V}_{d/4}$, we obtain

$$
\begin{aligned}
&\frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \\
&\geqslant \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \cdot \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \left(\frac{dT}{4} - \sum_{j \in V} \mathbb{E}_V[\tilde{M}_j] \right) \\
&= \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{4} - \frac{1}{|\mathcal{V}_{d/4}|} \sum_{j=1}^{d} \sum_{V \in \mathcal{V}_{d/4}^{(j)}} \mathbb{E}_V[\tilde{M}_j] \right) \\
&= \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{4} - \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4-1}} \sum_{j \notin V} \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] \right) \\
&\geqslant \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{4} - \frac{|\mathcal{V}_{d/4-1}|}{|\mathcal{V}_{d/4}|} \max_{V \in \mathcal{V}_{d/4-1}} \sum_{j \notin V} \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] \right) \\
&= \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{4} - \frac{|\mathcal{V}_{d/4-1}|}{|\mathcal{V}_{d/4}|} \max_{V \in \mathcal{V}_{d/4-1}} \sum_{j \notin V} \mathbb{E}_V[\tilde{M}_j] + \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] - \mathbb{E}_V[\tilde{M}_j] \right).
\end{aligned}
$$

For any fixed V, we get $\sum_{j \notin V} \mathbb{E}_V[\tilde{M}_j] \leqslant \sum_{j=1}^{d} \mathbb{E}_V[\tilde{M}_j] \leqslant dT/4$. Also, we have $\frac{|\mathcal{V}_{d/4-1}|}{|\mathcal{V}_{d/4}|} = \binom{d}{d/4-1} / \binom{d}{d/4} = \frac{d/4}{3d/4+1} \leqslant \frac{1}{3}$. Consequently, we derive that

$$
\begin{aligned}
&\frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \\
&\qquad \geqslant \frac{v_0 K}{(v_0 + Ke)^2} \cdot \frac{\epsilon}{2\sqrt{d}} \left(\frac{dT}{6} - \max_{V \in \mathcal{V}_{d/4-1}} \sum_{j \notin V} \left| \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] - \mathbb{E}_V[\tilde{M}_j] \right| \right). \qquad \text{(D.3)}
\end{aligned}
$$

Now we bound the term $\left| \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] - \mathbb{E}_V[\tilde{M}_j] \right|$ in (D.3) for any $V \in \mathcal{V}_{d/4-1}$. For simplicity, let $P = \mathbb{P}_V$ and $Q = \mathbb{P}_{V \cup \{j\}}$ denote the laws under $\mathbf{w}_V$ and $\mathbf{w}_{V \cup j}$, respectively. Then, we have

$$
\begin{aligned}
\left| \mathbb{E}_P[\tilde{M}_j] - \mathbb{E}_Q[\tilde{M}_j] \right| &\leqslant \sum_{t=0}^{T} t \cdot \left| P[\tilde{M}_j = t] - Q[\tilde{M}_j = t] \right| \\
&\leqslant T \cdot \sum_{t=0}^{T} \left| P[\tilde{M}_j = t] - Q[\tilde{M}_j = t] \right| \\
&\leqslant T \cdot \|P - Q\|_{\mathrm{TV}} \leqslant T \cdot \sqrt{\frac{1}{2} \mathrm{KL}(P \| Q)}, \qquad \text{(D.4)}
\end{aligned}
$$

where $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$ l is the total variation distance between P and Q, $\mathrm{KL}(P\|Q) = \int (\log \mathrm{d}P/\mathrm{d}Q) \mathrm{d}P$ is s the Kullback-Leibler (KL) divergence between P and Q, and the last inequality holds by Pinsker's inequality. Now, we bound the KL divergence term using the following Lemma.

Lemma D.2. *For any $V \in \mathcal{V}_{d/4-1}$ and $j \in [d]$, there exists a positive constant $C_{\mathrm{KL}} > 0$ such that*

$$
\mathrm{KL}(P_V \| Q_{V \cup \{j\}}) \leqslant C_{\mathrm{KL}} \cdot \frac{v_0 K}{(v_0 + K)^2} \cdot \frac{\mathbb{E}_V[\tilde{M}_j] \epsilon^2}{d}.
$$

Therefore, combining (D.3), (D.4), and Lemma D.2, we have

$$\begin{aligned}
&\frac{1}{\left|\mathcal{V}_{d/4}\right|}\sum_{V\in\mathcal{V}_{d/4}}\mathbb{E}_V\sum_{i\in S^\star}p(i|S^\star,\mathbf{w}_V)-\sum_{i\in\tilde{S}_t}p(i|\tilde{S}_t,\mathbf{w}_V)\\
&\quad\geqslant\frac{v_0K}{(v_0+Ke)^2}\cdot\frac{\epsilon}{2\sqrt{d}}\left(\frac{dT}{6}-T\sum_{j=1}^{d}\sqrt{C_{\mathrm{KL}}\cdot\frac{v_0K}{(v_0+K)^2}\cdot\frac{\mathbb{E}_V[\tilde{M}_j]\epsilon^2}{d}}\right)\\
&\quad\geqslant\frac{v_0K}{(v_0+Ke)^2}\cdot\frac{\epsilon}{2\sqrt{d}}\left(\frac{dT}{6}-T\sqrt{d}\cdot\sqrt{\sum_{j=1}^{d}C_{\mathrm{KL}}\cdot\frac{v_0K}{(v_0+K)^2}\cdot\frac{\mathbb{E}_V[\tilde{M}_j]\epsilon^2}{d}}\right)\\
&\quad\geqslant\frac{v_0K}{(v_0+Ke)^2}\cdot\frac{\epsilon}{2\sqrt{d}}\left(\frac{dT}{6}-T\sqrt{d}\cdot\sqrt{\frac{C_{\mathrm{KL}}}{4}\cdot\frac{v_0K}{(v_0+K)^2}\cdot T\epsilon^2}\right),
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwartz inequality and the last inequality holds because $\sum_{j=1}^{d}\mathbb{E}_V[\tilde{M}_j]\leqslant dT/4$. Let $C'_{\mathrm{KL}}=C_{\mathrm{KL}}/4$.

By setting $\epsilon=\sqrt{\frac{d}{144C'_{\mathrm{KL}}T}\cdot\frac{(v_0+K)^2}{v_0K}}$, we have

$$\begin{aligned}
\sup_{\mathbf{w}}\mathbb{E}^{\pi}_{\mathbf{w}}\left[\mathbf{Reg}_T(\mathbf{w})\right]&\geqslant\frac{v_0K}{(v_0+Ke)^2}\cdot\frac{\epsilon}{2\sqrt{d}}\left(\frac{dT}{6}-\sqrt{C'_{\mathrm{KL}}\cdot\frac{v_0K}{(v_0+K)^2}dT\epsilon^2}\right)\\
&=\frac{v_0K}{(v_0+Ke)^2}\cdot\sqrt{\frac{(v_0+K)^2}{v_0K}}\cdot\frac{1}{288\sqrt{C'_{\mathrm{KL}}}}d\sqrt{T}\\
&=\Omega\left(\frac{\sqrt{v_0K}}{v_0+K}\cdot d\sqrt{T}\right).
\end{aligned}$$

This concludes the proof of Theorem 1. □

D.3 Proofs of Lemmas for Theorem 1

D.3.1 Proof of Lemma D.1

Proof of Lemma D.1. Let $x=x_V$ and $\hat{x}=x_{\tilde{U}_t}$ be the corresponding context vectors. Then, we have

$$\begin{aligned}
\sum_{i\in S^\star}p(i|S^\star,\mathbf{w}_V)-\sum_{i\in\tilde{S}_t}p(i|\tilde{S}_t,\mathbf{w}_V)&=\frac{K\exp\left(x^\top\mathbf{w}_V\right)}{v_0+K\exp\left(x^\top\mathbf{w}_V\right)}-\frac{K\exp\left(\hat{x}^\top\mathbf{w}_V\right)}{v_0+K\exp\left(\hat{x}^\top\mathbf{w}_V\right)}\\
&=\frac{v_0K\left(\exp\left(x^\top\mathbf{w}_V\right)-\exp\left(\hat{x}^\top\mathbf{w}_V\right)\right)}{\left(v_0+K\exp\left(x^\top\mathbf{w}_V\right)\right)\left(v_0+K\exp\left(\hat{x}^\top\mathbf{w}_V\right)\right)}\\
&\geqslant\frac{v_0K\left(\exp\left(x^\top\mathbf{w}_V\right)-\exp\left(\hat{x}^\top\mathbf{w}_V\right)\right)}{(v_0+Ke)^2},
\end{aligned}\tag{D.5}$$

where the inequality holds since $\max\left\{\exp\left(x^\top\mathbf{w}_V\right),\exp\left(\hat{x}^\top\mathbf{w}_V\right)\right\}\leqslant e$. To further bound the right-hand side of (D.5), we use the fact that $1+a\leqslant e^a\leqslant 1+a+a^2/2$ for all $a\in[0,1]$, which can be easily shown by Taylor expansion. Thus, we get

$$\begin{aligned}
\sum_{i\in S^\star}p(i|S^\star,\mathbf{w}_V)-\sum_{i\in\tilde{S}_t}p(i|\tilde{S}_t,\mathbf{w}_V)&\geqslant\frac{v_0K\left((x-\hat{x})^\top\mathbf{w}_V-(\hat{x}^\top\mathbf{w}_V)^2/2\right)}{(v_0+Ke)^2}\\
&\geqslant\frac{v_0K\left(\delta\epsilon/\sqrt{d}-(\sqrt{d}\epsilon)^2/2\right)}{(v_0+Ke)^2}\\
&\geqslant\frac{v_0K\delta\epsilon}{2\sqrt{d}(v_0+Ke)^2},
\end{aligned}$$

where the last inequality holds because $(\sqrt{d}\epsilon)^2\leqslant\delta\epsilon/\sqrt{d}$ when $\epsilon\in(0,1/d\sqrt{d})$. This concludes the proof. □

D.3.2 Proof of Lemma D.2

Proof of Lemma D.2. Fix a round t, an assortment $\tilde{S}_t$, and $\tilde{U}_t$. Let $U = \tilde{U}_t$. Define $m_j(\tilde{S}_t) := \sum_{x_U \in \tilde{S}_t} \mathbf{1}\{j \in U\}/K$. Let $\{p_i\}_{i \in \tilde{S}_t \cup \{0\}}$ and $\{q_i\}_{i \in \tilde{S}_t \cup \{0\}}$ be the probabilities of choosing item i under parameterizations $\mathbf{w}_V$ and $\mathbf{w}_{V \cup \{j\}}$, respectively. Then, we have

$$\mathrm{KL}\left(\mathbb{P}_V(\cdot|\tilde{S}_t) \| \mathbb{P}_{V \cup \{j\}}(\cdot|\tilde{S}_t)\right) = \sum_{i \in \tilde{S}_t \cup \{0\}} p_i \log \frac{p_i}{q_i} \leqslant \sum_{i \in \tilde{S}_t \cup \{0\}} p_i \frac{p_i - q_i}{q_i} \leqslant \sum_{i \in \tilde{S}_t \cup \{0\}} \frac{(p_i - q_i)^2}{q_i},$$

where the first inequality holds because $\log(1+x) \leqslant x$ for all $x > -1$.

Let $\hat{x} = x_U$. Now, we separately upper bound $(p_i - q_i)^2/q_i$, by analyzing the following three cases:

Case 1. The outside option, $i = 0$.

For $i = 0$, $q_i \geqslant \frac{v_0}{v_0 + Ke}$. Thus, we have

$$\begin{aligned}
|p_i - q_i| &= \left| \frac{v_0}{v_0 + \sum_{i \in \tilde{S}_t} \exp\left(x_i^\top \mathbf{w}_V\right)} - \frac{v_0}{v_0 + \sum_{i \in \tilde{S}_t} \exp\left(x_i^\top \mathbf{w}_{V \cup \{j\}}\right)} \right| \\
&= \left| \frac{v_0}{v_0 + K \exp\left(\hat{x}^\top \mathbf{w}_V\right)} - \frac{v_0}{v_0 + K \exp\left(\hat{x}^\top \mathbf{w}_{V \cup \{j\}}\right)} \right| \\
&\leqslant \frac{v_0 K}{(v_0 + K/e)^2} \left| \exp\left(\hat{x}^\top \mathbf{w}_V\right) - \exp\left(\hat{x}^\top \mathbf{w}_{V \cup \{j\}}\right) \right| \\
&= \frac{v_0 K}{(v_0 + K/e)^2} \left| e^{\bar{c}_1} (\hat{x}^\top \mathbf{w}_V - \hat{x}^\top \mathbf{w}_{V \cup \{j\}}) \right| \\
&\leqslant \frac{v_0 K e}{(v_0 + K/e)^2} \left| \hat{x}^\top \left(\mathbf{w}_V - \mathbf{w}_{V \cup \{j\}}\right) \right| \leqslant \frac{v_0 K e}{(v_0 + K/e)^2} \cdot \frac{m_j(\tilde{S}_t)\epsilon}{\sqrt{d}},
\end{aligned}$$

where the third equality holds by applying the mean value theorem for the exponential function, with $\bar{c}_1 := (1-u)(\hat{x}^\top \mathbf{w}_V) + u(\hat{x}^\top \mathbf{w}_{V \cup \{j\}})$ for some $u \in (0,1)$. Then, there exist an absolute constant C_0 such that

$$\begin{aligned}
\frac{(p_0 - q_0)^2}{q_0} &\leqslant \frac{v_0^2 K^2 e^2}{(v_0 + K/e)^4} \cdot \frac{\left(m_j(\tilde{S}_t)\right)^2 \epsilon^2}{d} \cdot \frac{v_0 + Ke}{v_0} \\
&\leqslant C_0 \cdot \frac{v_0 K^2}{(v_0 + K)^3} \cdot \frac{m_j(\tilde{S}_t)\epsilon^2}{d},
\end{aligned} \tag{D.6}$$

where the last inequality holds since $m_j(\tilde{S}_t) \leqslant 1$.

Case 2. $i \in \tilde{S}_t$ and $j \notin U$.

Then, for any $i \in \tilde{S}_t$ corresponding to $x_i = \hat{x}$ and $j \notin U$, we have

$$|p_i - q_i| = \left| \frac{\exp\left(\hat{x}^\top \mathbf{w}_V\right)}{v_0 + K \exp\left(\hat{x}^\top \mathbf{w}_V\right)} - \frac{\exp\left(\hat{x}^\top \mathbf{w}_{V \cup \{j\}}\right)}{v_0 + K \exp\left(\hat{x}^\top \mathbf{w}_{V \cup \{j\}}\right)} \right| = 0,$$

where the last equality holds because $\exp\left(\hat{x}^\top \mathbf{w}_V\right) = \exp\left(\hat{x}^\top \mathbf{w}_{V \cup \{j\}}\right)$, given that $j \notin U$. Thus, we get

$$\sum_{i \in \tilde{S}_t, j \notin U} \frac{(p_i - q_i)^2}{q_i} = 0, \tag{D.7}$$

Case 3. $i \in \tilde{S}_t$ and $j \in U$.

Recall that for any $i \in \tilde{S}_t$, $q_i \geqslant \frac{e^{-1}}{v_0+Ke}$. Then, for any $i \in \tilde{S}_t$ corresponding to $x_i = \hat{x}$ and $j \in U$, we have

$$\begin{aligned}
|p_i - q_i| &= \left| \frac{\exp\left(\hat{x}^\top \mathbf{w}_V\right)}{v_0 + K\exp\left(\hat{x}^\top \mathbf{w}_V\right)} - \frac{\exp\left(\hat{x}^\top \mathbf{w}_{V\cup\{j\}}\right)}{v_0 + K\exp\left(\hat{x}^\top \mathbf{w}_{V\cup\{j\}}\right)} \right| \\
&= \left| \frac{\exp\left(\bar{c}_2\right)}{v_0 + K\exp\left(\bar{c}_2\right)} \cdot \hat{x}^\top \left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right) - \frac{K\exp\left(2\bar{c}_2\right)}{\left(v_0 + K\exp\left(\bar{c}_2\right)\right)^2} \cdot \hat{x}^\top \left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right) \right| \\
&= \frac{\exp\left(\bar{c}_2\right) v_0}{\left(v_0 + K\exp\left(\bar{c}_2\right)\right)^2} \left| \hat{x}^\top \left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right) \right| \\
&\leqslant \frac{v_0 e}{(v_0 + K/e)^2} \left| \hat{x}^\top \left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right) \right| \leqslant \frac{v_0 e}{(v_0 + K/e)^2} \cdot \frac{m_j(\tilde{S}_t)\epsilon}{\sqrt{d}},
\end{aligned}$$

the second equality holds by applying the mean value theorem, with $\bar{c}_2 := (1-u)(\hat{x}^\top \mathbf{w}_V) + u(\hat{x}^\top \mathbf{w}_{V\cup\{j\}})$ for some $u \in (0,1)$. Then, there exist an absolute constant C_1 such that

$$\begin{aligned}
\sum_{i\in\tilde{S}_t, j\in U} \frac{(p_i - q_i)^2}{q_i} &\leqslant K m_j(\tilde{S}_t) \cdot \frac{v_0^2 e^2}{(v_0 + K/e)^4} \cdot \frac{\left(m_j(\tilde{S}_t)\right)^2 \epsilon^2}{d} \cdot \frac{v_0 + Ke}{e^{-1}} \\
&\leqslant C_1 \cdot \frac{v_0^2 K}{(v_0+K)^3} \cdot \frac{m_j(\tilde{S}_t)\epsilon^2}{d}, \qquad \text{(D.8)}
\end{aligned}$$

where the last inequality holds since $m_j(\tilde{S}_t) \leqslant 1$.

Combining (D.6), (D.7), and (D.8), we derive that

$$\begin{aligned}
\sum_{i\in\tilde{S}_t\cup\{0\}} \frac{(p_i - q_i)^2}{q_i} &\leqslant \left(C_0 \cdot \frac{v_0 K^2}{(v_0+K)^3} + C_1 \cdot \frac{v_0^2 K}{(v_0+K)^3} \right) \cdot \frac{m_j(\tilde{S}_t)\epsilon^2}{d} \\
&\leqslant \max\{C_0, C_1\} \cdot \frac{v_0 K}{(v_0+K)^2} \cdot \frac{m_j(\tilde{S}_t)\epsilon^2}{d} \\
&= C_{\mathrm{KL}} \cdot \frac{v_0 K}{(v_0+K)^2} \cdot \frac{m_j(\tilde{S}_t)\epsilon^2}{d},
\end{aligned}$$

where $C_{\mathrm{KL}} = \max\{C_0, C_1\}$. Since $\tilde{M}_j = \sum_{t=1}^T m_j(\tilde{S}_t)$ by definition, and subsequently summing over all $t = 1$ to T, we have

$$\begin{aligned}
\mathrm{KL}\left(P_V \| Q_{V\cup\{j\}}\right) &= \sum_{t=1}^T \mathbb{E}_V \left[\mathrm{KL}\left(\mathbb{P}_V(\cdot|\tilde{S}_t) \| \mathbb{P}_{V\cup\{j\}}(\cdot|\tilde{S}_t) \right) \right] \\
&\leqslant C_{\mathrm{KL}} \cdot \frac{v_0 K}{(v_0+K)^2} \cdot \frac{\mathbb{E}_V[\tilde{M}_j]\epsilon^2}{d},
\end{aligned}$$

where the equality holds by the chain rule of relative entropy (cf. Exercise 14.11 of Lattimore and Szepesvári [32]). This concludes the proof. □

E Proof of Theorem 2

In this section, we present the proof of Theorem 2. Note that when the rewards are uniform, the revenue increases as a function of the assortment size. Therefore, maximizing the expected revenue $R_t(S, \mathbf{w})$ across all possible assortments $S \in \mathcal{S}$ always contains exactly K items. In other words, the size of the chosen assortment S_t and the size of the optimal assortment $S_t^\star$ both equal to K.

E.1 Main Proof of Theorem 2

Before presenting the proof, we introduce useful lemmas, whose proof can be found in Appendix E.2. Lemma E.1 shows the optimistic utility for the context vectors.

Lemma E.1. *Let* $\alpha_{ti} = x_{ti}^\top \mathbf{w}_t + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$. *If* $\mathbf{w}^\star \in \mathcal{C}_t(\delta)$, *then we have*

$$0 \leqslant \alpha_{ti} - x_{ti}^\top \mathbf{w}^\star \leqslant 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}.$$

Lemma E.3 is a K-free elliptical potential lemma that improves upon the one presented in Lemma 10 of Perivier and Goyal [44] in terms of K. Lemma 10 of Perivier and Goyal [44] states: $\sum_{s=1}^t \sum_{i\in S_s} p_s(i|S_s,\mathbf{w}^\star)p_s(0|S_s,\mathbf{w}^\star)\|x_{si}\|^2_{H_s(\mathbf{w}^\star)^{-1}} \leqslant 2dK\log\left(\lambda_{t+1}+\frac{2tK}{d}\right)$ and $\sum_{s=1}^t \max_{i\in S_s}\|x_{si}\|^2_{H_s(\mathbf{w}^\star)^{-1}} \leqslant 2d\left(K+\frac{1}{\kappa}\right)\log\left(\lambda_{t+1}+\frac{2tK}{d}\right)$, where $H_t(\mathbf{w}) = \sum_{s=1}^{t-1}\mathcal{G}_s(\mathbf{w}) + \lambda_t\mathbf{I}_d$.

Lemma E.2. *Let* $H_t = \lambda\mathbf{I}_d + \sum_{s=1}^{t-1}\mathcal{G}_s(\mathbf{w}_{s+1})$, *where* $\mathcal{G}_s(\mathbf{w}) = \sum_{i\in S_s} p_s(i|S_s,\mathbf{w})x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s,\mathbf{w})p_s(j|S_s,\mathbf{w})x_{si}x_{sj}^\top$. *Suppose* $\lambda \geqslant 1$. *Then the following statements hold true:*

(1) $\sum_{s=1}^t \sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})p_s(0|S_s,\mathbf{w}_{s+1})\|x_{si}\|^2_{H_s^{-1}} \leqslant 2d\log\left(1+\frac{t}{d\lambda}\right)$,

(2) $\sum_{s=1}^t \max_{i\in S_s}\|x_{si}\|^2_{H_s^{-1}} \leqslant \frac{2}{\kappa}d\log\left(1+\frac{t}{d\lambda}\right)$.

Moreover, we provide a tighter bound for the second derivative of the expected revenue than that presented in Lemma 12 of Perivier and Goyal [44]. Lemma 12 of Perivier and Goyal [44] states: $\left|\frac{\partial^2 Q}{\partial i\partial j}\right| \leqslant 5$.

Lemma E.3. *Define* $Q : \mathbb{R}^K \to \mathbb{R}$, *such that for any* $\mathbf{u} = (u_1,\ldots,u_K) \in \mathbb{R}^K$, $Q(\mathbf{u}) = \sum_{i=1}^K \frac{\exp(u_i)}{v_0+\sum_{k=1}^K\exp(u_k)}$. *Let* $p_i(\mathbf{u}) = \frac{\exp(u_i)}{v_0+\sum_{k=1}^K\exp(u_k)}$. *Then, for all* $i \in [K]$, *we have*

$$\left|\frac{\partial^2 Q}{\partial i\partial j}\right| \leqslant \begin{cases} 3p_i(\mathbf{u}) & \text{if } i=j, \\ 2p_i(\mathbf{u})p_j(\mathbf{u}) & \text{if } i\neq j. \end{cases}$$

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. First, we bound the regret as follows:

$$\begin{aligned}
\sum_{t=1}^T R_t(S_t^\star,\mathbf{w}^\star) - R_t(S_t,\mathbf{w}^\star) &= \sum_{t=1}^T\left[\sum_{i\in S_t^\star} p_t(i|S_t^\star,\mathbf{w}^\star) - \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\right] \\
&= \sum_{t=1}^T\left[\frac{\sum_{i\in S_t^\star}\exp(x_{ti}^\top\mathbf{w}^\star)}{v_0+\sum_{j\in S_t^\star}\exp(x_{tj}^\top\mathbf{w}^\star)} - \frac{\sum_{i\in S_t}\exp(x_{ti}^\top\mathbf{w}^\star)}{v_0+\sum_{j\in S_t}\exp(x_{tj}^\top\mathbf{w}^\star)}\right] \\
&\leqslant \sum_{t=1}^T\left[\frac{\sum_{i\in S_t^\star}\exp(\alpha_{ti})}{v_0+\sum_{j\in S_t^\star}\exp(\alpha_{tj})} - \frac{\sum_{i\in S_t}\exp(x_{ti}^\top\mathbf{w}^\star)}{v_0+\sum_{j\in S_t}\exp(x_{tj}^\top\mathbf{w}^\star)}\right] \\
&\leqslant \sum_{t=1}^T\left[\frac{\sum_{i\in S_t}\exp(\alpha_{ti})}{v_0+\sum_{j\in S_t}\exp(\alpha_{tj})} - \frac{\sum_{i\in S_t}\exp(x_{ti}^\top\mathbf{w}^\star)}{v_0+\sum_{j\in S_t}\exp(x_{tj}^\top\mathbf{w}^\star)}\right] = \sum_{t=1}^T \tilde{R}_t(S_t) - R_t(S_t,\mathbf{w}^\star),
\end{aligned}$$

where the first inequality holds by Lemma E.1, and the last inequality holds by the assortment selection of Algorithm 1.

Now, we define $Q : \mathbb{R}^K \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1,\ldots,u_K)^\top \in \mathbb{R}^K$, $Q(\mathbf{u}) = \sum_{i=1}^K \frac{\exp(u_i)}{v_0+\sum_{j=1}^K\exp(u_j)}$. Noting that S_t always contains K elements since the expected revenue is an increasing function in the uniform reward setting, we can write $S_t = \{i_1,\ldots,i_K\}$. Moreover, for all $t \geqslant 1$, let $\mathbf{u}_t = (u_{ti_1},\ldots u_{ti_K})^\top = (\alpha_{ti_1},\ldots,\alpha_{ti_K})^\top$ and $\mathbf{u}_t^\star = (u^\star_{ti_1},\ldots u^\star_{ti_K})^\top =$

$(x_{ti_1}^\top \mathbf{w}^\star, \ldots, x_{ti_K}^\top \mathbf{w}^\star)^\top$. Then, by a second order Taylor expansion, we have

$$\begin{aligned}\sum_{t=1}^T \tilde{R}_t(S_t) - R_t(S_t, \mathbf{w}^\star) &= \sum_{t=1}^T Q(\mathbf{u}_t) - Q(\mathbf{u}_t^\star) \\ &= \underbrace{\sum_{t=1}^T \nabla Q(\mathbf{u}_t^\star)^\top (\mathbf{u}_t - \mathbf{u}_t^\star)}_{\texttt{(A)}} + \underbrace{\frac{1}{2}\sum_{t=1}^T (\mathbf{u}_t - \mathbf{u}_t^\star)^\top \nabla^2 Q(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star)}_{\texttt{(B)}},\end{aligned} \tag{E.1}$$

where $\bar{\mathbf{u}}_t = (\bar{u}_{ti_1}, \ldots, \bar{u}_{ti_K})^\top \in \mathbb{R}^K$ is the convex combination of $\mathbf{u}_t$ and $\mathbf{u}_t^\star$.

First, we bound the term `(A)`.

$$\begin{aligned}&\sum_{t=1}^T \nabla Q(\mathbf{u}_t^\star)^\top (\mathbf{u}_t - \mathbf{u}_t^\star) \\ &= \sum_{t=1}^T \sum_{i \in S_t} \frac{\exp(x_{ti}^\top \mathbf{w}^\star)}{v_0 + \sum_{k \in S_t} \exp(x_{tk}^\top \mathbf{w}^\star)}(u_{ti} - u_{ti}^\star) - \sum_{i \in S_t}\sum_{j \in S_t} \frac{\exp(x_{ti}^\top \mathbf{w}^\star)\exp(x_{tj}^\top \mathbf{w}^\star)}{(v_0 + \sum_{k \in S_t} \exp(x_{tk}^\top \mathbf{w}^\star))^2}(u_{ti} - u_{ti}^\star) \\ &= \sum_{t=1}^T \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star)(u_{ti} - u_{ti}^\star) - \sum_{i \in S_t}\sum_{j \in S_t} p_t(i|S_t, \mathbf{w}^\star) p_t(j|S_t, \mathbf{w}^\star)(u_{ti} - u_{ti}^\star) \\ &= \sum_{t=1}^T \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star)\left(1 - \sum_{j \in S_t} p_t(j|S_t, \mathbf{w}^\star)\right)(u_{ti} - u_{ti}^\star) \\ &= \sum_{t=1}^T \sum_{t=1}^T \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) p_t(0|S_t, \mathbf{w}^\star)(u_{ti} - u_{ti}^\star) \\ &\leqslant \sum_{t=1}^T \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) p_t(0|S_t, \mathbf{w}^\star) 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} \\ &\leqslant 2\beta_T(\delta) \sum_{t=1}^T \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) p_t(0|S_t, \mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}},\end{aligned} \tag{E.2}$$

where the first inequality holds by Lemma E.1, and the last inequality holds because $\beta_t(\delta)$ is increasing for $t \in [T]$.

Now we bound the term `(B)`. Let $p_i(\bar{\mathbf{u}}_t) = \frac{\exp(\bar{u}_{ti})}{v_0 + \sum_{k=1}^K \exp(\bar{u}_{tk})}$. Then, we have

$$\begin{aligned}&\frac{1}{2}\sum_{t=1}^T (\mathbf{u}_t - \mathbf{u}_t^\star)^\top \nabla^2 Q(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star) \\ &= \frac{1}{2}\sum_{t=1}^T \sum_{i \in S_t}\sum_{j \in S_t} (u_{ti} - u_{ti}^\star)\frac{\partial^2 Q}{\partial i \partial j}(u_{tj} - u_{tj}^\star) \\ &= \frac{1}{2}\sum_{t=1}^T \sum_{i \in S_t}\sum_{j \in S_t, j \neq i} (u_{ti} - u_{ti}^\star)\frac{\partial^2 Q}{\partial i \partial j}(u_{tj} - u_{tj}^\star) + \frac{1}{2}\sum_{t=1}^T \sum_{i \in S_t} (u_{ti} - u_{ti}^\star)\frac{\partial^2 Q}{\partial i \partial i}(u_{ti} - u_{ti}^\star) \\ &\leqslant \sum_{t=1}^T \sum_{i \in S_t}\sum_{j \in S_t, j \neq i} |u_{ti} - u_{ti}^\star| p_i(\bar{\mathbf{u}}_t) p_j(\bar{\mathbf{u}}_t) |u_{tj} - u_{tj}^\star| + \frac{3}{2}\sum_{t=1}^T \sum_{i \in S_t} (u_{ti} - u_{ti}^\star)^2 p_i(\bar{\mathbf{u}}_t),\end{aligned} \tag{E.3}$$

where the inequality is by Lemma E.3. To bound the first term in (E.3), by applying the AM-GM inequality, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t, j\neq i}|u_{ti}-u_{ti}^\star|p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)|u_{tj}-u_{tj}^\star| \\
&\leqslant \sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}|u_{ti}-u_{ti}^\star|p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)|u_{tj}-u_{tj}^\star| \\
&\leqslant \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}(u_{ti}-u_{ti}^\star)^2 p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t)+\frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t}(u_{tj}-u_{tj}^\star)^2 p_i(\bar{\mathbf{u}}_t)p_j(\bar{\mathbf{u}}_t) \\
&\leqslant \sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2 p_i(\bar{\mathbf{u}}_t). \qquad \text{(E.4)}
\end{aligned}$$

By plugging (E.4) into (E.3), we have

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}(\mathbf{u}_t-\mathbf{u}_t^\star)^\top\nabla^2 Q(\bar{\mathbf{u}}_t)(\mathbf{u}_t-\mathbf{u}_t^\star) &\leqslant \frac{5}{2}\sum_{t=1}^{T}\sum_{i\in S_t}(u_{ti}-u_{ti}^\star)^2 p_i(\bar{\mathbf{u}}_t) \\
&\leqslant 10\sum_{t=1}^{T}\sum_{i\in S_t}p_i(\bar{\mathbf{u}}_t)\beta_t(\delta)^2\|x_{ti}\|^2_{H_t^{-1}} \\
&\leqslant 10\sum_{t=1}^{T}\max_{i\in S_t}\beta_t(\delta)^2\|x_{ti}\|^2_{H_t^{-1}} \\
&\leqslant 10\beta_T(\delta)^2\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}}, \qquad \text{(E.5)}
\end{aligned}$$

where the second inequality holds by Lemma E.1. Combining the upper bound for the terms `(A)` and `(B)`, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\tilde{R}_t(S_t)-R_t(S_t,\mathbf{w}^\star) \leqslant{}& 2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}} \\
&+10\beta_T(\delta)^2\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}}. \qquad \text{(E.6)}
\end{aligned}$$

Now, we bound each term of (E.6) respectively. For the first term, we decompose it as follows:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}} \\
&\quad=\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}_{t+1})p_t(0|S_t,\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}} \\
&\quad+\sum_{t=1}^{T}\sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}^\star)-p_t(i|S_t,\mathbf{w}_{t+1})\right)p_t(0|S_t,\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}} \\
&\quad+\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\left(p_t(0|S_t,\mathbf{w}^\star)-p_t(0|S_t,\mathbf{w}_{t+1})\right)\|x_{ti}\|_{H_t^{-1}}. \qquad \text{(E.7)}
\end{aligned}$$

To bound the first term on the right-hand side of (E.7), we apply the Cauchy-Schwarz inequality.

$$
\begin{aligned}
&\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}} \\
&\leqslant \sqrt{\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1})} \sqrt{\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}}^2} \\
&\leqslant \frac{\sqrt{v_0 K}}{(v_0 + K e^{-1})} \sqrt{T} \sqrt{\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}}^2} \\
&\leqslant \frac{\sqrt{v_0 K}}{(v_0 + K e^{-1})} \sqrt{T \cdot 2d \log\left(1 + \frac{T}{d\lambda}\right)},
\end{aligned} \tag{E.8}
$$

where the last inequality holds by Lemma E.2.

Now, we bound the second term on the right-hand side of (E.7). Let the *virtual* context for the outside option be $x_{t0} = \mathbf{0}$. Then, by the mean value theorem, there exists $\boldsymbol{\xi}_t = (1-c)\mathbf{w}^\star + c\mathbf{w}_{t+1}$ for some $c \in (0,1)$ such that

$$
\begin{aligned}
&\sum_{i \in S_t} \left(p_t(i|S_t, \mathbf{w}^\star) - p_t(i|S_t, \mathbf{w}_{t+1})\right) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}} \\
&= \sum_{i \in S_t} \nabla p_t(i|S_t, \boldsymbol{\xi}_t)^\top (\mathbf{w}^\star - \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}} \\
&= \sum_{i \in S_t} \left(p_t(i|S_t, \boldsymbol{\xi}_t) x_{ti} - p_t(i|S_t, \boldsymbol{\xi}_t) \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) x_{tj}\right)^\top (\mathbf{w}^\star - \mathbf{w}_{t+1}) p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}} \\
&\leqslant \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \left|x_{ti}^\top (\mathbf{w}^\star - \mathbf{w}_{t+1})\right| p_t(0|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|_{H_t^{-1}} \\
&+ \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}} \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \left|x_{tj}^\top (\mathbf{w}^\star - \mathbf{w}_{t+1})\right| p_t(0|S_t, \mathbf{w}_{t+1}) \\
&\leqslant \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} + \left(\sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}\right)^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t}.
\end{aligned}
$$

Then, since $x_{t0} = \mathbf{0}$, we can further bound the right-hand side as:

$$
\begin{aligned}
&\sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} + \left(\sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}\right)^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} \\
&= \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} + \left(\sum_{i \in S_t \cup \{0\}} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}\right)^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} \\
&\leqslant \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} + \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} \\
&\leqslant 2 \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \|\mathbf{w}^\star - \mathbf{w}_{t+1}\|_{H_t} \\
&\leqslant 2\beta_t(\delta) \sum_{i \in S_t} p_t(i|S_t, \boldsymbol{\xi}_t) \|x_{ti}\|_{H_t^{-1}}^2 \leqslant 2\beta_t(\delta) \max_{i \in S_t} \|x_{ti}\|_{H_t^{-1}}^2,
\end{aligned}
$$

where the first inequality holds due to Jensen's inequality and the second-to-last inequality holds by Lemma 1. Hence, we get

$$\sum_{t=1}^{T}\sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}^\star)-p_t(i|S_t,\mathbf{w}_{t+1})\right)p_t(0|S_t,\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}} \leqslant 2\beta_T(\delta)\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} \leqslant \frac{4d}{\kappa}\beta_T(\delta)\log\left(1+\frac{T}{d\lambda}\right) \tag{E.9}$$

where the last inequality holds by Lemma E.2.

Finally, we bound the third term on the right-hand side of (E.7). By the mean value theorem, there exists $\boldsymbol{\xi}_t' = (1-c')\mathbf{w}^\star + c'\mathbf{w}_{t+1}$ for some $c'\in(0,1)$ such that

$$\begin{aligned}
&\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\left(p_t(0|S_t,\mathbf{w}^\star)-p_t(0|S_t,\mathbf{w}_{t+1})\right)\|x_{ti}\|_{H_t^{-1}}\\
&=\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\nabla p_t(0|S_t,\boldsymbol{\xi}_t')^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}}\\
&=-\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\boldsymbol{\xi}_t')\sum_{j\in S_t} p_t(j|S_t,\boldsymbol{\xi}_t')x_{tj}^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}}\\
&\leqslant \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}p_t(0|S_t,\boldsymbol{\xi}_t')\sum_{j\in S_t} p_t(j|S_t,\boldsymbol{\xi}_t')\|x_{tj}\|_{H_t^{-1}}\|\mathbf{w}^\star-\mathbf{w}_{t+1}\|_{H_t}\\
&\leqslant \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}\sum_{j\in S_t} p_t(j|S_t,\boldsymbol{\xi}_t')\|x_{tj}\|_{H_t^{-1}}\|\mathbf{w}^\star-\mathbf{w}_{t+1}\|_{H_t}\\
&\leqslant \beta_t(\delta)\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}\sum_{j\in S_t} p_t(j|S_t,\boldsymbol{\xi}_t')\|x_{tj}\|_{H_t^{-1}}\\
&\leqslant \beta_t(\delta)\left(\max_{i\in S_t}\|x_{ti}\|_{H_t^{-1}}\right)^2=\beta_t(\delta)\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}},
\end{aligned}$$

where the third inequality holds by Lemma 1, and the last inequality holds since $(\max_i a_i)^2 = \max_i a_i^2$ for any $a_i\geqslant 0$. Therefore, we have

$$\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)\left(p_t(0|S_t,\mathbf{w}^\star)-p_t(0|S_t,\mathbf{w}_{t+1})\right)\|x_{ti}\|_{H_t^{-1}} \leqslant \beta_T(\delta)\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} \leqslant \frac{2d}{\kappa}\beta_T(\delta)\log\left(1+\frac{T}{d\lambda}\right), \tag{E.10}$$

where the last inequality holds by Lemma E.2. By plugging (E.8), (E.9), and (E.10) into (E.7) and multiplying $2\beta_T(\delta)$, we get

$$\begin{aligned}
&2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}\\
&\leqslant 2\sqrt{2}\frac{\sqrt{v_0K}}{(v_0+Ke^{-1})}\beta_T(\delta)\sqrt{dT}\sqrt{\log\left(1+\frac{T}{d\lambda}\right)}+\frac{12d}{\kappa}\beta_T(\delta)^2\log\left(1+\frac{T}{d\lambda}\right).
\end{aligned} \tag{E.11}$$

Moreover, by applying Lemma E.2, we can directly bound the second term of (E.6).

$$10\beta_T(\delta)^2\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} \leqslant 10\beta_T(\delta)^2\cdot\frac{2}{\kappa}d\log\left(1+\frac{T}{d\lambda}\right). \tag{E.12}$$

Finally, plugging (E.11) and (E.12) into (E.6), we obtain

$$\begin{aligned}
\mathbf{Reg}_T(\mathbf{w}^\star) &\leqslant 2\sqrt{2}\frac{\sqrt{v_0K}}{(v_0+Ke^{-1})}\beta_T(\delta)\sqrt{dT}\sqrt{\log\left(1+\frac{T}{d\lambda}\right)}+\frac{32d}{\kappa}\beta_T(\delta)^2\log\left(1+\frac{T}{d\lambda}\right)\\
&=\tilde{\mathcal{O}}\left(\frac{\sqrt{v_0K}}{v_0+K}d\sqrt{T}+\frac{1}{\kappa}d^2\right),
\end{aligned}$$

where $\beta_T(\delta) = \mathcal{O}\left(\sqrt{d}\log T \log K\right)$. This concludes the proof of Theorem 2. □

Remark E.1. *If the boundedness assumption on the parameter is relaxed to $\|\mathbf{w}\|_2 \leqslant B$, since $\beta_t(\delta) = \mathcal{O}\left(B\sqrt{d}\log t \log K + B^{3/2}\sqrt{d\log K}\right)$ (refer Corollary F.1), we have $\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(B^{3/2}e^B\frac{\sqrt{v_0 K}}{v_0+K}d\sqrt{T} + \frac{1}{\kappa}d^2\right)$. It's important to note that one of our main goals is to explicitly demonstrate the regret depends on K and v_0. In deriving such a result, the dependence on e^B is unavoidable to our best knowledge. Note that for non-uniform rewards, the regret bound does not depend on e^B (refer Remark H.1).*

E.2 Proofs of Lemmas for Theorem 2

E.2.1 Proof of Lemma E.1

Proof of Lemma E.1. Under the condition $\mathbf{w}^\star \in \mathcal{C}_t(\delta)$, we have

$$\left|x_{ti}^\top \mathbf{w}_t - x_{ti}^\top \mathbf{w}^\star\right| \leqslant \|x_{ti}\|_{H_t^{-1}}\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} \leqslant \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}},$$

where the first inequality is by the Hölder's inequality, and the last inequality holds by Lemma 1. Hence, it follows that

$$\alpha_{ti} - x_{ti}^\top \mathbf{w}^\star = x_{ti}^\top \mathbf{w}_t - x_{ti}^\top \mathbf{w}^\star + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} \leqslant 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}.$$

Moreover, from $x_{ti}^\top \mathbf{w}_t - x_{ti}^\top \mathbf{w}^\star \geqslant -\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$, we also have

$$\alpha_{ti} - x_{ti}^\top \mathbf{w}^\star = x_{ti}^\top \mathbf{w}_t - x_{ti}^\top \mathbf{w}^\star + \beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} \geqslant 0.$$

This concludes the proof. □

E.2.2 Proof of Lemma E.2

Proof of Lemma E.2. Since $xx^\top + yy^\top \succeq xy^\top + yx^\top$ for any $x, y \in \mathbb{R}^d$, it follows that

$$\begin{aligned}
&\mathcal{G}_s(\mathbf{w}_{s+1}) \\
&= \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})p_s(j|S_s, \mathbf{w}_{s+1}) x_{si}x_{sj}^\top \\
&= \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top - \frac{1}{2}\sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})p_s(j|S_s, \mathbf{w}_{s+1}) (x_{si}x_{sj}^\top + x_{sj}x_{si}^\top) \\
&\geq \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top - \frac{1}{2}\sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})p_s(j|S_s, \mathbf{w}_{s+1}) (x_{si}x_{si}^\top + x_{sj}x_{sj}^\top) \\
&= \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})p_s(j|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathcal{G}_s(\mathbf{w}_{s+1}) &\geq \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})\left(1 - \sum_{j\in S_s} p_s(j|S_s, \mathbf{w}_{s+1})\right) x_{si}x_{si}^\top \\
&= \sum_{i\in S_s} p_s(i|S_s, \mathbf{w}_{s+1})p_s(0|S_s, \mathbf{w}_{s+1}) x_{si}x_{si}^\top, \qquad \text{(E.13)}
\end{aligned}$$

which implies that

$$H_{t+1} \geq H_t + \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}_{t+1})p_t(0|S_t, \mathbf{w}_{t+1}) x_{ti}x_{ti}^\top.$$

Then, we get

$$\det\left(H_{t+1}\right) \geqslant \det\left(H_t\right)\left(1 + \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}_{t+1})p_0(i|S_t, \mathbf{w}_{t+1})\|x_{ti}\|^2_{H_t^{-1}}\right).$$

Since $\lambda \geqslant 1$, for all $t \geqslant 1$, we have $\sum_{i \in S_t} p_t(i|S_t, \mathbf{w}_{t+1}) p_0(i|S_t, \mathbf{w}_{t+1}) \|x_{ti}\|^2_{H_t^{-1}} \leqslant 1$. Then, using the fact that $z \leqslant 2\log(1+z)$ for any $z \in [0,1]$, we get

$$\begin{aligned}
&\sum_{s=1}^{t} \sum_{i \in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) p_s(0|S_s, \mathbf{w}_{s+1}) \|x_{si}\|^2_{H_s^{-1}} \\
&\qquad \leqslant 2 \sum_{s=1}^{t} \log\left(1 + p_s(i|S_s, \mathbf{w}_{s+1}) p_s(0|S_s, \mathbf{w}_{s+1}) \|x_{si}\|^2_{H_s^{-1}}\right) \\
&\qquad \leqslant 2 \sum_{s=1}^{t} \log\left(\frac{\det(H_{s+1})}{\det(H_s)}\right) \\
&\qquad \leqslant 2d \log\left(\frac{\operatorname{tr}(H_{t+1})}{d\lambda}\right) \leqslant 2d \log\left(1 + \frac{t}{d\lambda}\right).
\end{aligned}$$

This proves the first inequality.

To establish the proof for the second inequality, we return to (E.13):

$$\mathcal{G}_s(\mathbf{w}_{s+1}) \geq \sum_{i \in S_s} p_s(i|S_s, \mathbf{w}_{s+1}) p_s(0|S_s, \mathbf{w}_{s+1}) x_{si} x_{si}^\top \geq \kappa \sum_{i \in S_s} x_{si} x_{si}^\top,$$

which implies that

$$H_{t+1} = H_t + \mathcal{G}_t(\mathbf{w}_{t+1}) \geq H_t + \kappa \sum_{i \in S_t} x_{ti} x_{ti}^\top.$$

Since $\lambda \geqslant 1$, for all $t \geqslant 1$, we have $\kappa \max_{i \in S_t} \|x_{ti}\|^2_{H_t^{-1}} \leqslant \kappa$. We then conclude on the same way:

$$\begin{aligned}
\sum_{s=1}^{t} \max_{i \in S_s} \|x_{si}\|^2_{H_s^{-1}} &\leqslant \frac{2}{\kappa} \sum_{s=1}^{t} \log\left(1 + \kappa \max_{i \in S_s} \|x_{si}\|^2_{H_s^{-1}}\right) \\
&\leqslant \frac{2}{\kappa} \sum_{s=1}^{t} \log\left(\frac{\det(H_{s+1})}{\det(H_s)}\right) \leqslant \frac{2}{\kappa} d \log\left(1 + \frac{t}{d\lambda}\right),
\end{aligned}$$

which proves the second inequality. □

E.2.3 Proof of Lemma E.3

Proof of Lemma E.3. Let $i, j \in [K]$. We first have

$$\frac{\partial Q}{\partial i} = \frac{e^{u_i}}{v_0 + \sum_{k=1}^K e^{u_k}} - \frac{e^{u_i}\left(\sum_{k=1}^K e^{u_k}\right)}{(v_0 + \sum_{k=1}^K e^{u_k})^2}$$

Then, we get

$$\begin{aligned}
&\frac{\partial^2 Q}{\partial i \partial j} \\
&= \frac{\mathbb{1}_{i=j} e^{u_i}}{v_0 + \sum_{k=1}^K e^{u_k}} - \frac{e^{u_i} e^{u_j}}{(v_0 + \sum_{k=1}^K e^{u_k})^2} - \frac{\mathbb{1}_{i=j} e^{u_i}\left(\sum_{k=1}^K e^{u_k}\right) + e^{u_i} e^{u_j}}{(v_0 + \sum_{k=1}^K e^{u_k})^2} \\
&+ \frac{e^{u_i}\left(\sum_{k=1}^K e^{u_k}\right) 2e^{u_j}\left(v_0 + \sum_{k=1}^K e^{u_k}\right)}{(v_0 + \sum_{k=1}^K e^{u_k})^4} \\
&= \frac{\mathbb{1}_{i=j} e^{u_i}}{v_0 + \sum_{k=1}^K e^{u_k}} - \frac{e^{u_i} e^{u_j}}{(v_0 + \sum_{k=1}^K e^{u_k})^2} - \frac{\mathbb{1}_{i=j} e^{u_i}\left(\sum_{k=1}^K e^{u_k}\right) + e^{u_i} e^{u_j}}{(v_0 + \sum_{k=1}^K e^{u_k})^2} + \frac{e^{u_i}\left(\sum_{k=1}^K e^{u_k}\right) 2e^{u_j}}{(v_0 + \sum_{k=1}^K e^{u_k})^3}.
\end{aligned}$$

Let $p_i(\mathbf{u}) = \frac{e^{u_i}}{v_0+\sum_{k=1}^K e^{u_k}}$ and $p_0(\mathbf{u}) = \frac{v_0}{v_0+\sum_{k=1}^K e^{u_k}}$. For $i = j$, we have

$$\begin{aligned}
\left|\frac{\partial^2 Q}{\partial i \partial j}\right| &= \left| p_i(\mathbf{u}) - p_i(\mathbf{u})p_j(\mathbf{u}) - p_i(\mathbf{u})\frac{\sum_{k=1}^K e^{u_k}}{v_0+\sum_{k=1}^K e^{u_k}} - p_i(\mathbf{u})p_j(\mathbf{u}) \right. \\
&\quad \left. + 2p_i(\mathbf{u})p_j(\mathbf{u})\frac{\sum_{k=1}^K e^{u_k}}{v_0+\sum_{k=1}^K e^{u_k}} \right| \\
&= \left| p_i(\mathbf{u})p_0(\mathbf{u}) - 2p_i(\mathbf{u})p_j(\mathbf{u}) + 2p_i(\mathbf{u})p_j(\mathbf{u})\frac{\sum_{k=1}^K e^{u_k}}{v_0+\sum_{k=1}^K e^{u_k}} \right| \\
&= \left| p_i(\mathbf{u})p_0(\mathbf{u}) - 2p_i(\mathbf{u})p_j(\mathbf{u})p_0(\mathbf{u}) \right| \\
&\leqslant 3p_i(\mathbf{u})
\end{aligned}$$

For $i \neq j$, we have

$$\begin{aligned}
\left|\frac{\partial^2 Q}{\partial i \partial j}\right| &= \left| - p_i(\mathbf{u})p_j(\mathbf{u}) - p_i(\mathbf{u})p_j(\mathbf{u}) + 2p_i(\mathbf{u})p_j(\mathbf{u})\frac{\sum_{k=1}^K e^{u_k}}{v_0+\sum_{k=1}^K e^{u_k}} \right| \\
&= \left| - 2p_i(\mathbf{u})p_j(\mathbf{u})p_0(\mathbf{u}) \right| \\
&\leqslant 2p_i(\mathbf{u})p_j(\mathbf{u}).
\end{aligned}$$

This concludes the proof. □

F Proof of Lemma 1

In this section, we provide the proof of Lemma 1. First, we present the main proof of Lemma 1, followed by the proof of the technical lemma utilized within the main proof.

F.1 Main Proof of Lemma 1

Proof of Lemma 1. The proof is similar to the analysis presented in Zhang and Sugiyama [53]. However, their MNL choice model is constructed using a shared context x_t and varying parameters across the choices $\mathbf{w}_1^\star, \ldots, \mathbf{w}_K^\star$, whereas our approach considers an MNL choice model that shares the parameter $\mathbf{w}^\star$ across the choices and has varying contexts for each item in the assortment S, $x_{t1}, \ldots x_{ti_{|S|}}$. Moreover, Zhang and Sugiyama [53] only consider a fixed assortment size, whereas we consider a more general setting where the assortment size can vary in each round t. We denote $K_t = |S_t|$ in the proof of Lemma 1. Note that $K_t \leqslant K$ for all $t \geqslant 1$.

Lemma F.1. *Let the update rule be*

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \tilde{\ell}_t(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|_{H_t}^2,$$

where $\tilde{\ell}_t(\mathbf{w}) = \ell_t(\mathbf{w}_t) + \langle \mathbf{w}-\mathbf{w}_t, \nabla\ell_t(\mathbf{w}_t)\rangle + \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_{\nabla^2\ell_t(\mathbf{w}_t)}^2$ *and* $H_t = \lambda\mathbf{I}_d + \sum_{s=1}^{t-1}\mathcal{G}_s(\mathbf{w}_{s+1})$. *Let* $\eta = \frac{1}{2}\log(K+1)+2$ *and* $\lambda > 0$. *Then, we have*

$$\begin{aligned}
\|\mathbf{w}_{t+1}-\mathbf{w}^\star\|_{H_{t+1}}^2 &\leqslant 2\eta\left(\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1})\right) + 4\lambda + 12\sqrt{2}\eta\sum_{s=1}^t \|\mathbf{w}_{s+1}-\mathbf{w}_s\|_2^2 \\
&\quad - \sum_{s=1}^t \|\mathbf{w}_{s+1}-\mathbf{w}_s\|_{H_s}^2 . \qquad \text{(F.1)}
\end{aligned}$$

We first bound the first term in (F.1). For simplicity, we define the softmax function at round t $\boldsymbol{\sigma}_t(\mathbf{z}) : \mathbb{R}^{K_t} \to \mathbb{R}^{K_t}$ as follows:

$$[\boldsymbol{\sigma}_t(\mathbf{z})]_i = \frac{\exp([\mathbf{z}]_i)}{v_0+\sum_{k=1}^{K_t}\exp([\mathbf{z}]_k)}, \quad \forall i \in [K_t], \qquad \text{(F.2)}$$

where $[\cdot]_i$ denotes i'th element of the input vector. We denote the probability of choosing the outside option as $[\boldsymbol{\sigma}_t(\mathbf{z})]_0 = \frac{v_0}{v_0+\sum_{k=1}^{K_t}\exp([\mathbf{z}]_k)}$. Although $[\boldsymbol{\sigma}_t(\mathbf{z})]_0$ is not the output of the softmax function $\boldsymbol{\sigma}_t(\mathbf{z})$, we represent it in a form similar to that in (F.2) for simplicity. Then, the user choice model in (1) can be equivalently expressed as $p_t(i|S_t,\mathbf{w}) = \left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]_i$ for all $i \in [K_t]$ and $p_t(0|S_t,\mathbf{w}) = \left[\boldsymbol{\sigma}_t\left((x_{tj}^\top \mathbf{w})_{j\in S_t}\right)\right]_0$. Furthermore, the loss function (2) can also be written as $\ell(\mathbf{z}_t,\mathbf{y}_t) = \sum_{k=0}^{K_t}\mathbf{1}\{y_{ti}=1\}\cdot\log\left(\frac{1}{[\boldsymbol{\sigma}_t(\mathbf{z}_t)]_k}\right)$.

Define a pseudo-inverse function of $\boldsymbol{\sigma}_t(\cdot)$ as $\boldsymbol{\sigma}_t^+ : \mathbb{R}^{K_t} \to \mathbb{R}^{K_t}$, where $[\boldsymbol{\sigma}_t^+(\mathbf{q})]_i = \log\left(q_i/(1-\|\mathbf{q}\|_1)\right)$ for any $\mathbf{q} \in \{\mathbf{p}\in[0,1]^{K_t} \mid \|\mathbf{p}\|_1 < 1\}$. Then, inspired by the previous studies on binary logistic bandit [23], we decompose the regret into two terms by introducing an intermediate term.

$$\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) = \underbrace{\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s,\mathbf{y}_s)}_{(a)} + \underbrace{\sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s,\mathbf{y}_s) - \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1})}_{(b)}, \quad \text{(F.3)}$$

where $\tilde{\mathbf{z}}_s := \boldsymbol{\sigma}_s^+\left(\mathbb{E}_{\mathbf{w}\sim P_s}\left[\boldsymbol{\sigma}_s\left((x_{sj}^\top\mathbf{w})_{j\in S_s}\right)\right]\right)$, and $P_s := \mathcal{N}(\mathbf{w}_s, cH_s^{-1})$ is the Gaussian distribution with mean $\mathbf{w}_s$ and covariance matrix cH_s^{-1}, where $c>0$ is a positive constant to be specified later. We first show that the term (a) is bounded by $\mathcal{O}\left(\log K(\log t)^2\right)$ with high probability.

Lemma F.2. *Let $\delta \in (0,1]$. Under Assumptions 1, for all $t\in[T]$, with probability at least $1-\delta$, we have*

$$\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s,\mathbf{y}_s) \leqslant 11\cdot(3\log(1+(K+1)t)+3)\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)+2.$$

Furthermore, we can bound the term (b) by the following lemma.

Lemma F.3. *For any $c>0$, let $\lambda \geqslant \max\{2, 72cd\}$. Then, under Assumption 1, for all $t\geqslant 1$, we have*

$$\sum_{s=1}^t \left(\ell(\tilde{\mathbf{z}}_s,\mathbf{y}_s) - \ell_s(\mathbf{w}_{s+1})\right) \leqslant \frac{1}{2c}\sum_{s=1}^t \|\mathbf{w}_s - \mathbf{w}_{s+1}\|_{H_s}^2 + \sqrt{6}cd\log\left(1+\frac{t+1}{2\lambda}\right).$$

Now, we are ready to prove the Lemma 1. By combining Lemma F.1, Lemma F.2, and Lemma F.3, we derive that

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}^\star\|_{H_{t+1}}^2 \\
&\leqslant 2\eta\left[11\cdot(3\log(1+(K+1)t)+3)\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)+2+\sqrt{6}cd\log\left(1+\frac{t+1}{2\lambda}\right)\right] \\
&+4\lambda+12\sqrt{2}\eta\sum_{s=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_2^2+\left(\frac{\eta}{c}-1\right)\sum_{i=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_{H_s}^2 \\
&\leqslant 2\eta\left[11\cdot(3\log(1+(K+1)t)+3)\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)+2+\sqrt{6}cd\log\left(1+\frac{t+1}{2\lambda}\right)\right] \\
&+4\lambda =: \beta'_{t+1}(\delta)^2 = \mathcal{O}\left(d(\log t\log K)^2\right), \qquad \text{(F.4)}
\end{aligned}$$

where the second inequality holds because by setting $c=7\eta/6$ and $\lambda \geqslant \max\{84\sqrt{2}\eta, 84d\eta\}$, we obtain:

$$\begin{aligned}
&12\sqrt{2}\eta\sum_{s=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_2^2+\left(\frac{\eta}{c}-1\right)\sum_{i=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_{H_s}^2 \\
&\quad= 12\sqrt{2}\eta\sum_{s=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_2^2-\frac{1}{7}\sum_{i=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_{H_s}^2 \\
&\quad\leqslant\left(12\sqrt{2}\eta-\frac{\lambda}{7}\right)\sum_{s=1}^t\|\mathbf{w}_{s+1}-\mathbf{w}_s\|_2^2\leqslant 0,
\end{aligned}$$

where the first inequality holds since $H_s \succeq \lambda \mathbf{I}_d$.

By setting $\eta = \frac{1}{2}\log(K+1) + 2$ and $\lambda = 84\sqrt{2}d\eta$, we derive that

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}^\star\|_{H_t} &\leqslant \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{H_t} + \|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_t} \\
&\leqslant \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{H_t} + \|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_{t+1}} \\
&\leqslant \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_{H_t} + \beta'_{t+1}(\delta) \\
&\leqslant \frac{2\eta}{\lambda} + \beta'_{t+1}(\delta) := \beta_t(\delta) = \mathcal{O}\left(\sqrt{d}\log t \log K\right).
\end{aligned}$$

where the third inequality follows from the definition of$\beta'_{t+1}(\delta)$ in Equation (F.4), and the last inequality holds by Lemma F.9.

This concludes the proof of Lemma 1. □

Corollary F.1. *If the boundedness assumption on the parameter is relaxed to* $\|\mathbf{w}\|_2 \leqslant B$*, then* $\beta_t(\delta) = \mathcal{O}\left(B\sqrt{d}\log t \log K + B^{3/2}\sqrt{d\log K}\right)$.

Proof of Corollary F.1. When $\|\mathbf{w}\|_2 \leqslant B$, following the same analysis as in Lemma 1, we can set $\eta = \frac{1}{2}\log(K+1) + B + 1$, $c = 7\eta/6$, and $\lambda = \mathcal{O}(dB\log K)$. Under these settings, we have

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|^2_{H_{t+1}} \\
&\leqslant 2\eta\left[11\cdot(3\log(1+(K+1)t)+3)\log\left(\frac{2\sqrt{1+2t}}{\delta}\right) + 2 + \sqrt{6}cd\log\left(1+\frac{t+1}{2\lambda}\right)\right] + 4B^2\lambda \\
&= \mathcal{O}\left(B\sqrt{d}\log t\log K + B^{3/2}\sqrt{d\log K}\right).
\end{aligned}$$

which concludes the proof. □

F.2 Proofs of Lemmas for Lemma 1

F.2.1 Proof of Lemma F.1

Proof of Lemma F.1. Let $\tilde{\ell}_s(\mathbf{w}) = \ell_s(\mathbf{w}_s) + \langle\nabla\ell_s(\mathbf{w}_s), \mathbf{w}-\mathbf{w}_s\rangle + \frac{1}{2}\|\mathbf{w}-\mathbf{w}_s\|^2_{\nabla^2\ell_s(\mathbf{w}_s)}$ be a second-order approximation of the original function $\ell_s(\mathbf{w})$ at the point $\mathbf{w}_s$. The update rule (3) can also be expressed as

$$\mathbf{w}_{s+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \tilde{\ell}_s(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_s\|^2_{H_s}.$$

Then, by Lemma F.4, we have

$$\langle\nabla\tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \leqslant \frac{1}{2\eta}\left(\|\mathbf{w}_s - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{H_s}\right). \quad \text{(F.5)}$$

To utilize Lemma F.6, we can rewrite the loss function as $\ell\left((x_{si}^\top\mathbf{w})_{i\in S_s}, \mathbf{y}_s\right) = \ell_s(\mathbf{w})$. Consequently, according to Lemma F.6, it follows that

$$\ell_s(\mathbf{w}_{s+1}) - \ell_s(\mathbf{w}^\star) \leqslant \langle\nabla\ell_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle - \frac{1}{\zeta}\|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}, \quad \text{(F.6)}$$

where $\zeta = \log(K+1) + 4$. Then, by combining (F.5) and (F.6), we have

$$\begin{aligned}
\ell_s(\mathbf{w}_{s+1}) - \ell_s(\mathbf{w}^\star) \leqslant{}& \langle\nabla\ell_s(\mathbf{w}_{s+1}) - \nabla\tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star\rangle \\
&+ \frac{1}{\zeta}\left(\|\mathbf{w}_s - \mathbf{w}^\star\|^2_{H_s} - \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|^2_{H_{s+1}} - \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{H_s}\right).
\end{aligned}$$

In above, we can further bound the first term of the right-hand side as:

$$\begin{aligned}
\langle \nabla \ell_s(\mathbf{w}_{s+1}) - &\nabla \tilde{\ell}_s(\mathbf{w}_{s+1}), \mathbf{w}_{s+1} - \mathbf{w}^\star \rangle \\
&= \langle \nabla \ell_s(\mathbf{w}_{s+1}) - \nabla \ell_s(\mathbf{w}_s) - \nabla^2 \ell_s(\mathbf{w}_s)(\mathbf{w}_{s+1} - \mathbf{w}_s), \mathbf{w}_{s+1} - \mathbf{w}^\star \rangle \\
&= \langle D^3 \ell_s(\boldsymbol{\xi}_{s+1})[\mathbf{w}_{s+1} - \mathbf{w}_s](\mathbf{w}_{s+1} - \mathbf{w}_s), \mathbf{w}_{s+1} - \mathbf{w}^\star \rangle \\
&\leqslant 3\sqrt{2} \|\mathbf{w}_{s+1} - \mathbf{w}^\star\|_2 \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{\nabla^2 \ell_s(\boldsymbol{\xi}_{s+1})} \\
&\leqslant 6\sqrt{2} \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{\nabla^2 \ell_s(\boldsymbol{\xi}_{s+1})} \\
&\leqslant 6\sqrt{2} \|\mathbf{w}_{s+1} - \mathbf{w}_s\|_2^2
\end{aligned}$$

where in the second equality, we apply the Taylor expansion by introducing $\boldsymbol{\xi}_{s+1}$, a convex combination of $\mathbf{w}_{s+1}$ and $\mathbf{w}_s$. The first inequality follows from Lemma C.1 and Proposition C.1, the second inequality holds by Assumption 1, and the last inequality holds because

$$\begin{aligned}
\nabla^2 \ell_s(\boldsymbol{\xi}_{s+1}) &= \mathcal{G}_s(\boldsymbol{\xi}_{s+1}) \\
&= \sum_{i \in S_s} p_s(i|S_s, \boldsymbol{\xi}_{s+1}) x_{si} x_{si}^\top - \sum_{i \in S_s} \sum_{j \in S_s} p_s(i|S_s, \boldsymbol{\xi}_{s+1}) p_s(j|S_s, \boldsymbol{\xi}_{s+1}) x_{si} x_{sj}^\top \\
&= \sum_{i \in S_s \cup \{0\}} p_s(i|S_s, \boldsymbol{\xi}_{s+1}) x_{si} x_{si}^\top - \sum_{i \in S_s \cup \{0\}} \sum_{j \in S_s \cup \{0\}} p_s(i|S_s, \boldsymbol{\xi}_{s+1}) p_s(j|S_s, \boldsymbol{\xi}_{s+1}) x_{si} x_{sj}^\top \\
&= \mathbb{E}_{i \sim p_s(\cdot|S, \boldsymbol{\xi}_{s+1})} \left[x_{si} x_{si}^\top \right] - \mathbb{E}_{i \sim p_s(\cdot|S_s, \boldsymbol{\xi}_{s+1})} \left[x_{si} \right] \left(\mathbb{E}_{i \sim p_s(\cdot|S_s, \boldsymbol{\xi}_{s+1})} \left[x_{si} \right] \right)^\top \\
&\preceq \mathbb{E}_{i \sim p_s(\cdot|S_s, \boldsymbol{\xi}_{s+1})} \left[x_{si} x_{si}^\top \right] \preceq \mathbf{I}_d,
\end{aligned}$$

where the third equality holds by setting $x_{s0} = \mathbf{0}$ for all $s \geqslant 1$.

Now, by taking the summation over s and rearranging the terms, we obtain

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|^2_{H_{t+1}} \\
&\leqslant \zeta \left(\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) \right) + \|\mathbf{w}_1 - \mathbf{w}^\star\|^2_{H_1} + 6\sqrt{2}\zeta \sum_{s=1}^t \|\mathbf{w}_{s+1} - \mathbf{w}_s\|_2^2 \\
&- \sum_{s=1}^t \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{H_s} \\
&\leqslant \zeta \left(\sum_{s=1}^t \ell_s(\mathbf{w}^\star) - \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) \right) + 4\lambda + 6\sqrt{2}\zeta \sum_{s=1}^t \|\mathbf{w}_{s+1} - \mathbf{w}_s\|_2^2 - \sum_{s=1}^t \|\mathbf{w}_{s+1} - \mathbf{w}_s\|^2_{H_s},
\end{aligned}$$

where the last inequality holds since $\|\mathbf{w}_1 - \mathbf{w}^\star\|^2_{H_1} \leqslant \lambda \|\mathbf{w}_1 - \mathbf{w}^\star\|_2^2 \leqslant 4\lambda$. Plugging in $\zeta = 2\eta$, we conclude the proof. □

F.2.2 Proof of Lemma F.2

Proof of Lemma F.2. Since the norm of $\tilde{\mathbf{z}}_s = \boldsymbol{\sigma}_s^+ \left(\mathbb{E}_{\mathbf{w} \sim P_s} \left[\boldsymbol{\sigma}_s \left((x_{sj}^\top \mathbf{w})_{j \in S_s} \right) \right] \right)$ is unbounded in general, as suggested by Foster et al. [25], we use the smoothed version $\tilde{\mathbf{z}}_s^\mu = \boldsymbol{\sigma}_s^+ \left(\mathrm{smooth}_s^\mu \, \mathbb{E}_{\mathbf{w} \sim P_s} \left[\boldsymbol{\sigma}_s \left((x_{sj}^\top \mathbf{w})_{j \in S_s} \right) \right] \right)$ as an intermediate-term, where the smooth function is defined by $\mathrm{smooth}_s^\mu(\mathbf{q}) = (1 - \mu)\mathbf{q} + \mu \mathbf{1}/(K_s + 1)$, where $\mathbf{1} \in \mathbb{R}^{K_s}$ is an all one vector.

Note that $\tilde{\mathbf{z}}_s^\mu = \boldsymbol{\sigma}_s^+(\mathrm{smooth}_s^\mu(\boldsymbol{\sigma}_s(\tilde{\mathbf{z}}_s)))$ by the definition of the pseudo inverse function $\boldsymbol{\sigma}_s^+$ such that $\boldsymbol{\sigma}_s^+(\boldsymbol{\sigma}_s(\mathbf{q})) = \mathbf{q}$ for any $\mathbf{q} \in \{\mathbf{p} \in [0,1]^{K_s} \mid \|\mathbf{p}\|_1 < 1\}$. Then, by Lemma F.7, we have

$$\sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s^\mu, \mathbf{y}_s) - \sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s, \mathbf{y}_s) \leqslant 2\mu t, \quad \text{and} \quad \|\tilde{\mathbf{z}}_s^\mu\|_\infty \leqslant \log(1 + (K+1)/\mu). \tag{F.7}$$

Hence, to prove the lemma, we need only to bound the gap between the loss of $\mathbf{w}^\star$ and $\tilde{\mathbf{z}}_s^\mu$. To enhance clarity in our presentation, let $\ell(\mathbf{z}_s^\star, \mathbf{y}_s) = \ell_s(\mathbf{w}^\star)$, where $\mathbf{z}_s^\star = \left(x_{sj}^\top \mathbf{w}^\star \right)_{j \in S_s} \in \mathbb{R}^{K_s}$. Then,

we have

$$\begin{aligned}\sum_{s=1}^{t}\ell_s(\mathbf{w}^\star)-\sum_{s=1}^{t}\ell(\tilde{\mathbf{z}}_s^\mu,\mathbf{y}_s)&=\sum_{s=1}^{t}\ell(\mathbf{z}_s^\star,\mathbf{y}_s)-\sum_{s=1}^{t}\ell(\tilde{\mathbf{z}}_s^\mu,\mathbf{y}_s)\\&\leqslant\sum_{s=1}^{t}\langle\nabla_z\ell(\mathbf{z}_s^\star,\mathbf{y}_s),\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\rangle-\sum_{s=1}^{t}\frac{1}{c_\mu}\|\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\|^2_{\nabla_z^2\ell(\mathbf{z}_s^\star,\mathbf{y}_s)}\\&=\sum_{s=1}^{t}\langle\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s,\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\rangle-\sum_{s=1}^{t}\frac{1}{c_\mu}\|\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)},\end{aligned}\tag{F.8}$$

where $c_\mu=\log(K+1)+2\log(1+(K+1)/\mu)+2$, the inequality holds by Lemma F.6, and the last equality holds by a direct calculation of the first order and Hessian of the logistic loss as follows:

$$\nabla_z\ell(\mathbf{z}_s,\mathbf{y}_s)=\boldsymbol{\sigma}_s(\mathbf{z}_s)-\mathbf{y}_s,\quad\nabla_z^2\ell(\mathbf{z}_s,\mathbf{y}_s)=\operatorname{diag}(\boldsymbol{\sigma}_s(\mathbf{z}_s))-\boldsymbol{\sigma}_s(\mathbf{z}_s)\boldsymbol{\sigma}_s(\mathbf{z}_s)^\top.$$

We first bound the first term of the right-hand side. Define $\mathbf{d}_s=(\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu)/(c_\mu+1)$. Let $\mathbf{d}_s'$ be $\mathbf{d}_s$ extended with zero padding. Specifically, we define $\mathbf{d}_s'=[\mathbf{d}_s^\top,0,\ldots,0]^\top\in\mathbb{R}^K$, where the zeros are appended to increase the dimension of $\mathbf{d}_s$ to K. Similarly, we also extend $\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s$ with zero padding and define $\boldsymbol{\varepsilon}_s=[(\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s)^\top,0,\ldots,0]^\top\in\mathbb{R}^K$.

Then, one can easily verify that $\|\mathbf{d}_s'\|_\infty\leqslant 1$ since $\|\mathbf{z}_s^\star\|_\infty\leqslant\max_{i\in S_s}\|x_{si}\|_2\|\mathbf{w}^\star\|_2\leqslant 1$ and $\|\tilde{\mathbf{z}}_s^\mu\|_\infty\leqslant\log(1+(K+1)/\mu)$. On the other hand, $\mathbf{d}_s'$ is $\mathcal{F}_s$-measurable since $\mathbf{z}_s^\star$ and $\tilde{\mathbf{z}}_s^\mu$ are independent of $\mathbf{y}_s$. Moreover, we have $\|\mathbf{d}_s'\|^2_{\mathbb{E}[\boldsymbol{\varepsilon}_s\boldsymbol{\varepsilon}_s^\top|\mathcal{F}_s]}=\|\mathbf{d}_s\|^2_{\mathbb{E}[(\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s)(\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s)^\top|\mathcal{F}_s]}=\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}$ and $\|\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s\|_1\leqslant 2$. Thus, by Lemma F.5, with probability at least $1-\delta$, for any $t\geqslant 1$, we have

$$\begin{aligned}&\sum_{s=1}^{t}\langle\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s,\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\rangle=(c_\mu+1)\sum_{s=1}^{t}\langle\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)-\mathbf{y}_s,\mathbf{d}_s\rangle\\&=(c_\mu+1)\sum_{s=1}^{t}\langle\boldsymbol{\varepsilon}_s,\mathbf{d}_s'\rangle\\&\leqslant(c_\mu+1)\sqrt{\tilde{\lambda}+\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}}\left(\frac{\sqrt{\tilde{\lambda}}}{4}+\frac{4}{\sqrt{\tilde{\lambda}}}\log\cdot\left(\frac{2\sqrt{1+\frac{1}{\tilde{\lambda}}\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}}}{\delta}\right)\right)\\&\leqslant(c_\mu+1)\sqrt{\tilde{\lambda}+\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}}\cdot\left(\frac{\sqrt{\tilde{\lambda}}}{4}+\frac{4}{\sqrt{\tilde{\lambda}}}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right),\end{aligned}\tag{F.9}$$

where in the second inequality, we set $\tilde{\lambda}>1$, and the last inequality holds because $\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}=\mathbf{d}_s^\top\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)\mathbf{d}_s\leqslant 2$. Then, combining (F.8) and (F.9), we obtain

$$\begin{aligned}&\sum_{s=1}^{t}\ell_s(\mathbf{w}^\star)-\sum_{s=1}^{t}\ell(\tilde{\mathbf{z}}_s^\mu,\mathbf{y}_s)\\&\leqslant(c_\mu+1)\sqrt{\tilde{\lambda}+\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}}\cdot\left(\frac{\sqrt{\tilde{\lambda}}}{4}+\frac{4}{\sqrt{\tilde{\lambda}}}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)-\sum_{s=1}^{t}\frac{1}{c_\mu}\|\mathbf{z}_s^\star-\tilde{\mathbf{z}}_s^\mu\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}\\&\leqslant(c_\mu+1)\sqrt{\tilde{\lambda}+\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}}\cdot\left(\frac{\sqrt{\tilde{\lambda}}}{4}+\frac{4}{\sqrt{\tilde{\lambda}}}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)-(c_\mu+1)\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}\\&\leqslant(c_\mu+1)\left(\tilde{\lambda}+\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}\right)+(c_\mu+1)\left(\frac{\sqrt{\tilde{\lambda}}}{4}+\frac{4}{\sqrt{\tilde{\lambda}}}\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)^2\\&\quad-(c_\mu+1)\sum_{s=1}^{t}\|\mathbf{d}_s\|^2_{\nabla\boldsymbol{\sigma}_s(\mathbf{z}_s^\star)}\\&=(c_\mu+1)\left(\frac{17}{16}\tilde{\lambda}+2\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)+\frac{16}{\tilde{\lambda}}\left(\log\left(\frac{2\sqrt{1+2t}}{\delta}\right)\right)^2\right),\end{aligned}\tag{F.10}$$

where the third inequality holds due to the AM-GM inequality. Finally, combining (F.7) and (F.10), by setting $\mu = 1/t$ and $\tilde{\lambda} = \frac{16}{\sqrt{17}} \log\left(\frac{2\sqrt{1+2t}}{\delta}\right)$, we have

$$\begin{aligned}\sum_{s=1}^{t} \left(\ell_s(\mathbf{w}^\star) - \ell(\tilde{\mathbf{z}}_s, \mathbf{y}_s)\right) &\leqslant (c_\mu + 1)(2\sqrt{17} + 2) \log\left(\frac{2\sqrt{1+2t}}{\delta}\right) + 2\mu t \\ &\leqslant 11 \cdot (3\log(1 + (K+1)t) + 3) \log\left(\frac{2\sqrt{1+2t}}{\delta}\right) + 2\end{aligned}$$

where the last inequality holds by the definition of $c_\mu = \log(K+1) + 2\log(1 + (K+1)/\mu) + 2$. This concludes the proof. □

F.2.3 Proof of Lemma F.3

Proof of Lemma F.3. The proof with an observation from Proposition 2 in Foster et al. [25], which notes that $\tilde{\mathbf{z}}_s$ is an aggregation forecaster for the logistic function. Hence, it satisfies

$$\ell(\tilde{\mathbf{z}}_s, \mathbf{y}_s) \leqslant -\log\left(\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right]\right) = -\log\left(\frac{1}{Z_s}\int_{\mathbb{R}^d} e^{-L_s(\mathbf{w})}\mathrm{d}\mathbf{w}\right), \tag{F.11}$$

where $L_s(\mathbf{w}) := \ell_s(\mathbf{w}) + \frac{1}{2c}\|\mathbf{w} - \mathbf{w}_s\|^2_{H_s}$ and $Z_s := \sqrt{(2\pi)^d c|H_s^{-1}|}$.

Then, by the quadratic approximation, we get

$$\tilde{L}_s(\mathbf{w}) = L_s(\mathbf{w}_{s+1}) + \langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w} - \mathbf{w}_{s+1}\rangle + \frac{1}{2c}\|\mathbf{w} - \mathbf{w}_{s+1}\|^2_{H_s}. \tag{F.12}$$

Applying Lemma F.8 and considering the fact that ℓ_s is $3\sqrt{2}$-self-concordant-like function by Proposition C.1, we have

$$L_s(\mathbf{w}) \leqslant \tilde{L}_s(\mathbf{w}) + e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w} - \mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}. \tag{F.13}$$

We define the function $\tilde{f}_{s+1} : \mathcal{W} \to \mathbb{R}$ as

$$\tilde{f}_{s+1}(\mathbf{w}) = \exp\left(-\frac{1}{2c}\|\mathbf{w} - \mathbf{w}_{s+1}\|^2_{H_s} - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w} - \mathbf{w}_{s+1}\|^2_{\nabla^2\ell_s(\mathbf{w}_{s+1})}\right).$$

Then, we can establish a lower bound for the expectation in (F.11) as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right] &= \frac{1}{Z_s}\int_{\mathbb{R}^d} \exp(-L_s(\mathbf{w}))\mathrm{d}\mathbf{w} \\ &\geqslant \frac{1}{Z_s}\int_{\mathbb{R}^d} \exp(-\tilde{L}_s(\mathbf{w}) - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w} - \mathbf{w}_{s+1}\|^2_{\nabla\ell_s(\mathbf{w}_{s+1})})\mathrm{d}\mathbf{w} \\ &= \frac{\exp(-L_s(\mathbf{w}_{s+1}))}{Z_s}\int_{\mathbb{R}^d} \tilde{f}_{s+1}(\mathbf{w}) \cdot \exp(-\langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w} - \mathbf{w}_{s+1}\rangle)\mathrm{d}\mathbf{w},\end{aligned} \tag{F.14}$$

where the first inequality holds by (F.13) and the last equality holds by (F.12). We define $\tilde{Z}_{s+1} = \int_{\mathbb{R}^d} \tilde{f}_{s+1}(\mathbf{w})\mathrm{d}\mathbf{w} \leqslant +\infty$. Moreover, we denote the distribution whose density function is $\tilde{f}_{s+1}(\mathbf{w})/\tilde{Z}_{s+1}$ as $\tilde{P}_{s+1}$. Then, we can rewrite Equation (F.14) as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{w}\sim P_s}\left[e^{-\ell_s(\mathbf{w})}\right] &\geqslant \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s}\mathbb{E}_{\mathbf{w}\sim\tilde{P}_{s+1}}\left[\exp(-\langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w} - \mathbf{w}_{s+1}\rangle)\right] \\ &\geqslant \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s}\exp\left(-\mathbb{E}_{\mathbf{w}\sim\tilde{P}_{s+1}}\left[\langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w} - \mathbf{w}_{s+1}\rangle\right]\right) \\ &= \frac{\exp(-L_s(\mathbf{w}_{s+1}))\tilde{Z}_{s+1}}{Z_s},\end{aligned} \tag{F.15}$$

where the second inequality follows from Jensen's inequality, and the equality holds because $\tilde{P}_{s+1}$ is symmetric around $\mathbf{w}_{s+1}$, thus $\mathbb{E}_{\mathbf{w}\sim\tilde{P}_{s+1}}\left[\langle \nabla L_s(\mathbf{w}_{s+1}), \mathbf{w} - \mathbf{w}_{s+1}\rangle\right] = 0$.

By plugging (F.15) into (F.11), we have

$$\ell(\tilde{\mathbf{z}}_s, \mathbf{y}_s) \leqslant L_s(\mathbf{w}_{s+1}) + \log Z_s - \log \tilde{Z}_{s+1}. \tag{F.16}$$

In the above, we can bound the last term, $-\log \tilde{Z}_{s+1}$, by

$$\begin{aligned}
-\log \tilde{Z}_{s+1} &= -\log\left(\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{H_s}^2 - e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right)\mathrm{d}\mathbf{w}\right) \\
&= -\log\left(\widehat{Z}_{s+1}\cdot\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\exp\left(-e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right)\right]\right) \\
&\leqslant -\log\widehat{Z}_{s+1} + \mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right] \\
&= -\log Z_s + \mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right],
\end{aligned} \tag{F.17}$$

where $\widehat{P}_{s+1} = \mathcal{N}(\mathbf{w}_{s+1}, cH_s^{-1})$ and $\widehat{Z}_{s+1} = \int_{\mathbb{R}^d}\exp\left(-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{H_s}^2\right)\mathrm{d}\mathbf{w}$. In (F.17), the inequality holds due to Jensen's inequality, and the last inequality is by the fact that $\widehat{Z}_{s+1} = \int_{\mathbb{R}^d}\exp\left(-\frac{1}{2c}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{H_s}^2\right)\mathrm{d}\mathbf{w} = \sqrt{(2\pi)^d c|H_s^{-1}|} = Z_s$.

By applying the Cauchy-Schwarz inequality, we can further bound the second term on the right-hand side of (F.17) by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right] \\
&\qquad\leqslant \underbrace{\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]}}_{\texttt{(a)-1}}\underbrace{\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^4\right]}}_{\texttt{(a)-2}}.
\end{aligned} \tag{F.18}$$

Note that, since $\widehat{P}_{s+1} = \mathcal{N}(\mathbf{w}_{s+1}, cH_s^{-1})$, there exist orthogonal bases $\mathbf{e}_1, \ldots, \mathbf{e}_d \in \mathbb{R}^d$ such that $\mathbf{w}-\mathbf{w}_{s+1}$ follows the same distribution as

$$\sum_{j=1}^d \sqrt{c\lambda_j\left(H_s^{-1}\right)}X_j\mathbf{e}_j, \quad \text{where } X_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \forall j\in[d], \tag{F.19}$$

and $\lambda_j\left(H_s^{-1}\right)$ denotes the j-th largest eigenvalue of H_s^{-1}. Then, we can bound the term `(a)-1` in (F.18) as follows:

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]} &= \sqrt{\mathbb{E}_{X_j}\left[\prod_{j=1}^d e^{36c\lambda_j\left(H_s^{-1}\right)X_j^2}\right]} \leqslant \sqrt{\prod_{j=1}^d \mathbb{E}_{X_j}\left[e^{\frac{36c}{\lambda}X_j^2}\right]} \\
&= \left(\mathbb{E}_{X\sim\chi^2}\left[e^{\frac{36c}{\lambda}X}\right]\right)^{\frac{d}{2}} \leqslant \mathbb{E}_{X\sim\chi^2}\left[e^{\frac{18cd}{\lambda}X}\right],
\end{aligned}$$

where the first inequality holds since $\lambda_j\left(H_s^{-1}\right) \leqslant \frac{1}{\lambda}$. In the second equality, χ^2 denotes the chi-square distribution, and the last inequality is due to Jensen's inequality. By setting $\lambda \geqslant 72cd$, we get

$$\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{36\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\right]} \leqslant \mathbb{E}_{X\sim\chi^2}\left[e^{\frac{X}{4}}\right] \leqslant \sqrt{2}, \tag{F.20}$$

where the last inequality holds due to the fact that the moment-generating function for χ^2-distribution is bounded by $\mathbb{E}_{X\sim\chi^2}[e^{tX}] \leqslant 1/\sqrt{1-2t}$ for all $t \leqslant 1/2$.

Now, we bound the term `(a)-2` in (F.18).

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\|\mathbf{w}-\mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^4\right]} &= \sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,cH_s^{-1})}\left[\|\mathbf{w}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^4\right]} \\
&= \sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,c\bar{H}_s^{-1})}\left[\|\mathbf{w}\|_2^4\right]},
\end{aligned}$$

where $\bar{H}_s = (\nabla^2 \ell_s(\mathbf{w}_{s+1}))^{-1/2} H_s (\nabla^2 \ell_s(\mathbf{w}_{s+1}))^{-1/2}$. Let $\bar{\lambda}_j = \lambda_j \left(c\bar{H}_s^{-1}\right)$ be the j-th largest eigenvalue of the matrix $c\bar{H}_s^{-1}$. Then, conducting an analysis similar to that in equation (F.19) yields that

$$\begin{aligned}
\sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,c\bar{H}_s^{-1})}\left[\|\mathbf{w}\|_2^4\right]} &= \sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left\|\sum_{j=1}^d \sqrt{\bar{\lambda}_j} X_j \mathbf{e}_j\right\|_2^4\right]} \\
&= \sqrt{\mathbb{E}_{X_j\sim\mathcal{N}(0,1)}\left[\left(\sum_{j=1}^d \bar{\lambda}_j X_j^2\right)^2\right]} \\
&= \sqrt{\sum_{j=1}^d \sum_{j'=1}^d \bar{\lambda}_j \bar{\lambda}_{j'} \mathbb{E}_{X_j, X_{j'}\sim\mathcal{N}(0,1)}\left[X_j^2 X_{j'}^2\right]} \\
&\leqslant \sqrt{3\sum_{j=1}^d \sum_{j'=1}^d \bar{\lambda}_j \bar{\lambda}_{j'}} = \sqrt{3}c \operatorname{Tr}\left(\bar{H}_s^{-1}\right),
\end{aligned}$$

where the inequality holds due to $\mathbb{E}_{X_j, X_{j'}\sim\mathcal{N}(0,1)}[X_j^2 X_{j'}^2] \leqslant 3$ for all $j, j' \in [d]$, and the last equality holds because $\sum_{j=1}^d \bar{\lambda}_j = \operatorname{Tr}\left(c\bar{H}_s^{-1}\right)$. Here, $\operatorname{Tr}(A)$ denotes the trace of the matrix A.

We define the matrix $M_{s+1} := \lambda \mathbf{I}_d/2 + \sum_{\tau=1}^s \nabla^2 \ell_\tau(\mathbf{w}_{\tau+1})$. Under the condition $\lambda \geqslant 2$, for any $s \in [T]$ and $\mathbf{w} \in \mathcal{W}$, we have $\nabla^2 \ell_s(\mathbf{w}) \preceq \mathbf{I}_d \leqslant \frac{\lambda}{2}\mathbf{I}_d$. Thus, we have $H_s \succeq M_{s+1}$. Then, we can bound the trace as follows:

$$\begin{aligned}
\operatorname{Tr}\left(\bar{H}_s^{-1}\right) &= \operatorname{Tr}\left(H_s^{-1}\nabla^2 \ell_s(\mathbf{w}_{s+1})\right) \leqslant \operatorname{Tr}\left(M_{s+1}^{-1}\nabla^2 \ell_s(\mathbf{w}_{s+1})\right) \\
&= \operatorname{Tr}\left(M_{s+1}^{-1}(M_{s+1} - M_s)\right) \leqslant \log\frac{\det(M_{s+1})}{\det(M_s)},
\end{aligned}$$

where the last inequality holds by Lemma 4.5 of Hazan et al. [26]. Therefore we can bound the term `(a)-2` as

$$\sqrt{\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[\|\mathbf{w} - \mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^4\right]} \leqslant \sqrt{3}c \log\frac{\det(M_{s+1})}{\det(M_s)}. \tag{F.21}$$

By plugging (F.20) and (F.21) into (F.18), we have

$$\mathbb{E}_{\mathbf{w}\sim\widehat{P}_{s+1}}\left[e^{18\|\mathbf{w}-\mathbf{w}_{s+1}\|_2^2}\|\mathbf{w} - \mathbf{w}_{s+1}\|_{\nabla^2\ell_s(\mathbf{w}_{s+1})}^2\right] \leqslant \sqrt{6}c \log\frac{\det(M_{s+1})}{\det(M_s)}. \tag{F.22}$$

Combining (F.16), (F.17), and (F.22), and taking summation over s, we derive that

$$\begin{aligned}
\sum_{s=1}^t \ell(\tilde{\mathbf{z}}_s, \mathbf{y}_s) &\leqslant \sum_{s=1}^t L_s(\mathbf{w}_{s+1}) + \sqrt{6}c\sum_{s=1}^t \log\frac{\det(M_{s+1})}{\det(M_s)} \\
&= \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) + \frac{1}{2c}\sum_{s=1}^t \|\mathbf{w}_s - \mathbf{w}_{s+1}\|_{H_s}^2 + \sqrt{6}c\sum_{s=1}^t \log\frac{\det(M_{s+1})}{\det(M_s)} \\
&= \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) + \frac{1}{2c}\sum_{s=1}^t \|\mathbf{w}_s - \mathbf{w}_{s+1}\|_{H_s}^2 + \sqrt{6}c\log\left(\frac{\det(M_{s+1})}{\det\left(\frac{\lambda}{2}\mathbf{I}_d\right)}\right) \\
&\leqslant \sum_{s=1}^t \ell_s(\mathbf{w}_{s+1}) + \frac{1}{2c}\sum_{s=1}^t \|\mathbf{w}_s - \mathbf{w}_{s+1}\|_{H_s}^2 + \sqrt{6}c\cdot d\log\left(1 + \frac{t+1}{2\lambda}\right),
\end{aligned}$$

By rearranging the terms, we conclude the proof. □

F.3 Technical Lemmas for Lemma 1

Lemma F.4 (Proposition 4.1 of Campolongo and Orabona 12)**.** *Let the* $\mathbf{w}_{t+1}$ *be the solution of the update rule*

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{V}} \eta_t \ell_t(\mathbf{w}) + D_\psi(\mathbf{w}, \mathbf{w}_t),$$

where $\mathcal{V} \subseteq \mathcal{W} \subseteq \mathbb{R}^d$ is a non-empty convex set and $D_\psi(\mathbf{w}_1, \mathbf{w}_2) = \psi(\mathbf{w}_1) - \psi(\mathbf{w}_2) - \langle \nabla\psi(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle$ is the Bregman Divergence w.r.t. a strictly convex and continuously differentiable function $\psi : \mathcal{W} \to \mathbb{R}$. Further supposing $\psi(\mathbf{w})$ is 1-strongly convex w.r.t. a certain norm $\|\cdot\|$ in $\mathcal{W}$, then there exists a $\mathbf{g}'_t \in \partial\ell_t(\mathbf{w}_{t+1})$ such that

$$\langle \eta_t \mathbf{g}'_t, \mathbf{w}_{t+1} - \mathbf{u}\rangle \leqslant \langle \nabla\psi(\mathbf{w}_t) - \nabla\psi(\mathbf{w}_{t+1}), \mathbf{w}_{t+1} - \mathbf{u}\rangle$$

for any $\mathbf{u} \in \mathcal{W}$.

Lemma F.5 (Lemma 15 of Zhang and Sugiyama 53)**.** *Let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration. Let $\{\mathbf{z}_t\}_{t=1}^\infty$ be a stochastic process in $\mathcal{B}_2(K) = \{\mathbf{z} \in \mathbb{R}^K \mid \|\mathbf{z}\|_\infty \leqslant 1\}$ such that $\mathbf{z}_t$ is $\mathcal{F}_t$ measurable. Let $\{\boldsymbol{\varepsilon}_t\}_{t=1}^\infty$ be a martingale difference sequence such that $\boldsymbol{\varepsilon}_t \in \mathbb{R}^K$ is $\mathcal{F}_{t+1}$ measurable. Furthermore, assume that, conditional on $\mathcal{F}_t$, we have $\|\boldsymbol{\varepsilon}_t\|_1 \leqslant 2$ almost surely. Let $\Sigma_t = \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top | \mathcal{F}_t]$. and $\lambda > 0$. Then, for any $t \geqslant 1$ define*

$$U_t = \sum_{s=1}^{t-1} \langle \boldsymbol{\varepsilon}_s, \mathbf{z}_s\rangle \quad \text{and} \quad H_t = \lambda + \sum_{s=1}^{t-1} \|\mathbf{z}_s\|_{\Sigma_s}^2,$$

Then, for any $\delta \in (0, 1]$, we have

$$\Pr\left[\exists t \geqslant 1, U_t \geqslant \sqrt{H_t}\left(\frac{\sqrt{\lambda}}{4} + \frac{4}{\sqrt{\lambda}}\log\left(\sqrt{\frac{H_t}{\lambda}}\right) + \frac{4}{\sqrt{\lambda}}\log\left(\frac{2}{\delta}\right)\right)\right] \leqslant \delta.$$

Lemma F.6 (Lemma 1 of Zhang and Sugiyama 53)**.** *Let $C > 0$, $\mathbf{a} \in [-C, C]^K$, $\mathbf{y} \in \mathbb{R}^{K+1}$ be a one-hot vector and $\mathbf{b} \in \mathbb{R}^K$. Then, we have*

$$\ell(\mathbf{a}, \mathbf{y}) \geqslant \ell(\mathbf{b}, \mathbf{y}) + \nabla\ell(\mathbf{b}, \mathbf{y})^\top(\mathbf{a} - \mathbf{b}) + \frac{1}{\log(K+1) + 2(C+1)}(\mathbf{a} - \mathbf{b})^\top \nabla^2 \ell(\mathbf{b}, \mathbf{y})(\mathbf{a} - \mathbf{b}).$$

Lemma F.7 (Lemma 17 of Zhang and Sugiyama 53)**.** *Let $\mathbf{z} \in \mathbb{R}^K$ be a K-dimensional vector. Let $\ell(\mathbf{z}, \mathbf{y}) = \sum_{k=0}^K \mathbf{1}\{y_i = 1\} \cdot \log\left(\frac{1}{[\boldsymbol{\sigma}(\mathbf{z})]_k}\right)$, where $\mathbf{y} = [y_0, \ldots, y_K]^\top \in \mathbb{R}^{K+1}$, and the softmax function $\boldsymbol{\sigma}(\mathbf{z}) : \mathbb{R}^K \to \mathbb{R}^K$ is defined as $[\boldsymbol{\sigma}(\mathbf{z})]_i = \frac{\exp([\mathbf{z}]_i)}{v_0 + \sum_{k=1}^K \exp([\mathbf{z}]_k)}$ for all $i \in [K]$, and $[\boldsymbol{\sigma}(\mathbf{z})]_0 = \frac{v_0}{v_0 + \sum_{k=1}^K \exp([\mathbf{z}]_k)}$. Define $\mathbf{z}^\mu := \boldsymbol{\sigma}^+(\mathrm{smooth}_\mu(\boldsymbol{\sigma}(\mathbf{z})))$, where $\mathrm{smooth}_\mu(\mathbf{q}) = (1-\mu)\mathbf{q} + \mu\mathbf{1}/(K+1)$. Then, for $\mu \in [0, 1/2]$, we have*

$$\ell(\mathbf{z}^\mu, \mathbf{y}) - \ell(\mathbf{z}, \mathbf{y}) \leqslant 2\mu$$

We also have $\|\mathbf{z}^\mu\|_\infty \leqslant \log(1 + (K+1)/\mu)$.

Lemma F.8 (Lemma 18 of Zhang and Sugiyama 53)**.** *Let $L_t(\mathbf{w}) = \ell_t(\mathbf{w}) + \frac{1}{2c}\|\mathbf{w} - \mathbf{w}_t\|_{H_t}^2$. Assume that ℓ_t is a M-self-concordant-like function. Then, for any $\mathbf{w}, \mathbf{w}_t \in \mathcal{W}$, the quadratic approximation $\tilde{L}_t(\mathbf{w}) = L_t(\mathbf{w}_{t+1}) + \langle \nabla^2 L_t(\mathbf{w}_{t+1}), \mathbf{w} - \mathbf{w}_{t+1}\rangle + \frac{1}{2c}\|\mathbf{w} - \mathbf{w}_{t+1}\|_{H_t}^2$ satisfies*

$$L_t(\mathbf{w}) \leqslant \tilde{L}_t(\mathbf{w}) + e^{M^2\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2}\|\mathbf{w} - \mathbf{w}_{t+1}\|_{\nabla\ell_t(\mathbf{w}_{t+1})}^2.$$

Lemma F.9. *Let $\mathbf{w}_{t+1} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\langle \nabla\ell_t(\mathbf{w}_t), \mathbf{w}\rangle + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|_{H_t}^2$. Then, we have*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_{H_t} \leqslant \eta\|\nabla\ell_t(\mathbf{w}_t)\|_{H_t^{-1}} \leqslant \frac{\eta}{\lambda}\|\nabla\ell_t(\mathbf{w}_t)\|_2 \leqslant \frac{2\eta}{\lambda}.$$

Proof of Lemma F.9. We make a slight improvement to Lemma 20 of Zhang and Sugiyama [53].

Lemma F.10 (Lemma 20 of Zhang and Sugiyama 53)**.** *Under the same conditions as Lemma F.9, the following holds:*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_{H_t} \leqslant \eta\|\nabla\ell_t(\mathbf{w}_t)\|_{H_t^{-1}} \leqslant \frac{\eta}{\lambda}\|\nabla\ell_t(\mathbf{w}_t)\|_2 \leqslant \frac{\eta\sqrt{K}}{\lambda}.$$

In the proof of Lemma 20 by Zhang and Sugiyama [53], the bound for $\|\nabla\ell_t(\mathbf{w}_t)\|_2$ was given as $\|\nabla\ell_t(\mathbf{w}_t)\|_2 \leqslant 2\sqrt{K}$. However, a tighter bound can be established: $\|\nabla\ell_t(\mathbf{w}_t)\|_2 \leqslant 2$. Using this tighter bound, we can refine and derive the improved result. □

G Proofs of Theorem 3

In this section, we provide the proof of Theorem 3. In addition to the adversarial construction presented in Section D.1, we construct the adversarial non-uniform rewards.

G.1 Adversarial Rewards Construction

Under the adversarial construction in Section D.1, we observe that there are K identical context vectors, invariant across rounds t. Therefore, in total, there are $N = K \cdot \binom{d}{d/4}$ items. Let the rewards be also time-invariant. Given $\mathbf{w}_V$, we define a unique item $i^\star \in [N]$ as an item that maximizes $x_i^\top \mathbf{w}_V$, i.e., $x_{i^\star} = x_V$, and has a reward of 1, i.e., $r_{i^\star} = 1$. Then, we construct the non-uniform rewards as follows:

$$r_i = \begin{cases} 1, & \text{for } i = i^\star \\ \gamma, & \text{for } i \neq i^\star, \end{cases} \tag{G.1}$$

where we define γ as

$$\gamma = \min_{S \in \mathcal{S}} \frac{\min_{i \in S} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \min_{i \in S} \exp(x_i^\top \mathbf{w}_V)} = \frac{1}{v_0 + 1}.$$

Note that $\gamma < 1$.

G.2 Main Proof of Theorem 3

Proof of Theorem 3. Given the rewards construction as (G.1), any reward in the optimal assortment $S_t^\star$ is larger than the expected revenues.

Lemma G.1. *Let* $R(S^\star, \mathbf{w}_V) = \frac{\sum_{i \in S^\star} \exp(x_i^\top \mathbf{w}_V) r_i}{v_0 + \sum_{j \in S^\star} \exp(x_i^\top \mathbf{w}_V)}$. *Then, we have*

$$r_i \geqslant R(S^\star, \mathbf{w}_V), \quad \forall i \in S^\star.$$

Lemma G.1 implies that $S^\star$ contains only one item $i^\star$. This is because if $S^\star = \{x_{i^\star}\}$, adding any item $i \neq i^\star$ to the assortment results in lower expected revenue, since $r_i = \gamma \leqslant R(S^\star = \{x_{i^\star}\}, \mathbf{w}_v)$.

Furthermore, we can bound the expected revenue for any assortment as follows:

Lemma G.2. *Under the same parameters and context vectors as those in Section D, if the rewards are constructed according to Equation* (G.1)*, for any* $S \in \mathcal{S}$*, we have*

$$R(S, \mathbf{w}_V) \leqslant \frac{\max_{i \in S} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \max_{i \in S} \exp(x_i^\top \mathbf{w}_V)}.$$

Let $x_{U_1}, \ldots, x_{U_L}$ be the distinct feature vectors contained in assortments S_t with $U_1, \ldots, U_L \in \mathcal{V}_{d/4}$. Let $U^\star$ be the subset among $U_1, \ldots, U_L$ that maximizes $x_U^\top \mathbf{w}_V$, i.e., $U^\star \in \operatorname{argmax}_{U \in \{U_1, \ldots, U_L\}} x_U^\top \mathbf{w}_V$, where $\mathbf{w}_V$ is the underlying parameter. For simplicity, we denote $\tilde{U}_t$ as the unique $U^\star \in \mathcal{V}_{d/4}$ in S_t. Then, we have

$$\begin{aligned}
\sum_{t=1}^{T} R(S^\star, \mathbf{w}_V) - R(S_t, \mathbf{w}_V) &= \sum_{t=1}^{T} \frac{\exp(x_V^\top \mathbf{w}_V)}{v_0 + \exp(x_V^\top \mathbf{w}_V)} - R(S_t, \mathbf{w}_V) \\
&\geqslant \sum_{t=1}^{T} \frac{\exp(x_V^\top \mathbf{w}_V)}{v_0 + \exp(x_V^\top \mathbf{w}_V)} - \frac{\max_{i \in S_t} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \max_{i \in S_t} \exp(x_i^\top \mathbf{w}_V)} \\
&= \sum_{t=1}^{T} \frac{\exp(x_V^\top \mathbf{w}_V)}{v_0 + \exp(x_V^\top \mathbf{w}_V)} - \frac{\exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)}{v_0 + \exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)},
\end{aligned}$$

where the first equality holds becauset $S^\star$ contains only one item $i^\star$ by Lemma G.1 (and recall that $x_{i^\star} = x_V$), and the inequality holds by Lemma G.2. Hence, the problem is not easier than solving

the MNL bandit problems with the assortment size 1, i.e., $K = 1$. By putting $K = 1$ and $v_0 = \Theta(1)$ in Theorem 1, we derive that

$$\begin{aligned}
\sup_{\mathbf{w}} \mathbb{E}^{\pi}_{\mathbf{w}} \left[\mathbf{Reg}_T(\mathbf{w})\right] &\geqslant \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}^{\pi}_{\mathbf{w}_V} \sum_{t=1}^{T} R(S^\star, \mathbf{w}_V) - R(S_t, \mathbf{w}_V) \\
&\geqslant \frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}^{\pi}_{\mathbf{w}_V} \sum_{t=1}^{T} \frac{\exp(x_V^\top \mathbf{w}_V)}{v_0 + \exp(x_V^\top \mathbf{w}_V)} - \frac{\exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)}{v_0 + \exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)} \\
&= \Omega\left(d\sqrt{T}\right).
\end{aligned}$$

This concludes the proof of Theorem 3. □

G.3 Proofs of Lemmas for Theorem 3

G.3.1 Proof of Lemma G.1

Proof of Lemma G.1. We prove by contradiction. Assume that there exists $i \in S^\star$ such that $r_i < R(S^\star, \mathbf{w}_V)$. Then, removing the item i from the assortment $S^\star$ yields higher expected revenue. This contradicts the optimality of $S^\star$. Thus, we have

$$r_i \geqslant R(S^\star, \mathbf{w}_V), \quad \forall i \in S^\star.$$

This concludes the proof. □

G.3.2 Proof of Lemma G.2

Proof of Lemma G.2. We provide a proof by considering the following cases:

Case 1. $i^\star \in S$.

Recall that, by the construction of rewards, we have

$$\gamma = \min_{S \in \mathcal{S}} \frac{\min_{i \in S} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \min_{i \in S} \exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\min_{i \in S} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \min_{i \in S} \exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\exp(x_{i^\star}^\top \mathbf{w}_V)}{v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V)}. \tag{G.2}$$

This implies that

$$\begin{aligned}
&\left\{\sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\right\} \gamma \left(v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V)\right) \leqslant \left\{\sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\right\} \exp(x_{i^\star}^\top \mathbf{w}_V) \\
&\Leftrightarrow \left(\exp(x_{i^\star}^\top \mathbf{w}_V) + \sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\gamma\right)\left(v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V)\right) \\
&\leqslant \exp(x_{i^\star}^\top \mathbf{w}_V)\left(v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V) + \sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\right) \\
&\Leftrightarrow \frac{\exp(x_{i^\star}^\top \mathbf{w}_V) + \sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\gamma}{v_0 + \sum_{i \in S} \exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\exp(x_{i^\star}^\top \mathbf{w}_V)}{v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V)}.
\end{aligned} \tag{G.3}$$

Therefore, for all $S \in \mathcal{S}$, we have

$$\begin{aligned}
R(S, \mathbf{w}_V) = \frac{\sum_{i \in S} \exp(x_i^\top \mathbf{w}_V) r_i}{v_0 + \sum_{i \in S} \exp(x_i^\top \mathbf{w}_V)} &= \frac{\exp(x_{i^\star}^\top \mathbf{w}_V) + \sum_{i \in S \setminus \{i^\star\}} \exp(x_i^\top \mathbf{w}_V)\gamma}{v_0 + \sum_{i \in S} \exp(x_i^\top \mathbf{w}_V)} \\
&\leqslant \frac{\exp(x_{i^\star}^\top \mathbf{w}_V)}{v_0 + \exp(x_{i^\star}^\top \mathbf{w}_V)} \leqslant \frac{\max_{i \in S} \exp(x_i^\top \mathbf{w}_V)}{v_0 + \max_{i \in S} \exp(x_i^\top \mathbf{w}_V)},
\end{aligned}$$

where the first inequality holds by (G.3), and the last inequality holds since $f(x) = \frac{x}{v_0 + x}$ is an increasing function.

Case 2. $i^\star \notin S$.

Let us return to (G.2). Since $\frac{v_0+\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \geqslant 1$ for any $S \in \mathcal{S}$, we have

$$\gamma \leqslant \frac{\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{v_0+\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{v_0+\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \cdot \frac{v_0+\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)},$$

which is equivalent to

$$\frac{\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)\gamma}{v_0+\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{v_0+\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)}.$$

Hence, for all $S \in \mathcal{S}$, we get

$$R(S,\mathbf{w}_V) = \frac{\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)\gamma}{v_0+\sum_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{v_0+\min_{i\in S}\exp(x_i^\top \mathbf{w}_V)} \leqslant \frac{\max_{i\in S}\exp(x_i^\top \mathbf{w}_V)}{v_0+\max_{i\in S}\exp(x_i^\top \mathbf{w}_V)}.$$

This concludes the proof. □

H Proofs of Theorem 4

In this section, we provide the proof of Theorem 4. Since we now consider the case of non-uniform rewards, the sizes of both the chosen assortment S_t, and the optimal assortment, $S_t^\star$ are no longer fixed at K.

We begin the proof by introducing additional useful lemmas. Lemma H.1 shows that $\tilde{R}_t(S_t)$, defined in (6), is an upper bound of the true expected revenue of the optimal assortment, $R_t(S_t^\star, \mathbf{w}^\star)$.

Lemma H.1 (Lemma 4 in Oh and Iyengar 41)**.** *Let* $\tilde{R}_t(S) = \frac{\sum_{i\in S}\exp(\alpha_{ti})r_{ti}}{v_0+\sum_{j\in S}\exp(\alpha_{tj})}$. *And suppose* $S_t = \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$. *If for every item* $i \in S_t^\star$, $\alpha_{ti} \geqslant x_{ti}^\top \mathbf{w}^\star$, *then for all* $t \geqslant 1$, *the following inequalities hold:*

$$R_t(S_t^\star, \mathbf{w}^\star) \leqslant \tilde{R}_t(S_t^\star) \leqslant \tilde{R}_t(S_t).$$

Note that Lemma H.1 does not claim that the expected revenue is a monotone function in general. Instead, it specifically states that the value of the expected revenue, when associated with the optimal assortment, increases with an increase in the MNL parameters [8, 41].

Lemma H.2 shows that $\tilde{R}_t(S_t)$ increases as the utilities of items in S_t increase.

Lemma H.2. *Let* $\tilde{R}_t(S) = \frac{\sum_{i\in S}\exp(\alpha_{ti})r_{ti}}{v_0+\sum_{j\in S}\exp(\alpha_{tj})}$ *and* $S_t = \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$. *Assume* $\alpha'_{ti} \geqslant \alpha_{ti} \geqslant 0$ *for all* $i \in [N]$. *Then, we have*

$$\tilde{R}_t(S_t) \leqslant \frac{\sum_{i\in S_t}\exp(\alpha'_{ti})r_{ti}}{v_0+\sum_{j\in S_t}\exp(\alpha'_{tj})}.$$

Furthermore, we provide a novel elliptical potential Lemma H.3 for the *centralized* context vectors $\tilde{x}_{ti}$.

Lemma H.3. *Let* $H_t = \lambda \mathbf{I}_d + \sum_{s=1}^{t-1} \mathcal{G}_s(\mathbf{w}_{s+1})$, *where* $\mathcal{G}_s(\mathbf{w}) = \sum_{i\in S_s} p_s(i|S_s,\mathbf{w}) x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s,\mathbf{w})p_s(j|S_s,\mathbf{w})x_{si}x_{sj}^\top$. *Define* $\tilde{x}_{si} = x_{si} - \mathbb{E}_{j\sim p_s(\cdot|S_s,\mathbf{w}_{s+1})}[x_{sj}]$. *Suppose* $\lambda \geqslant 2$. *Then the following statements hold true:*

(1) $\sum_{s=1}^{t}\sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})\|\tilde{x}_{si}\|^2_{H_s^{-1}} \leqslant 2d\log\left(1+\frac{t}{d\lambda}\right)$,

(2) $\sum_{s=1}^{t}\max_{i\in S_s}\|\tilde{x}_{si}\|^2_{H_s^{-1}} \leqslant \frac{2}{\kappa}d\log\left(1+\frac{t}{d\lambda}\right)$.

Now, we prove the Theorem 4.

H.1 Proof of Theorem 4

Proof of Theorem 4. Let $\alpha'_{ti} = x_{ti}^\top \mathbf{w}^\star + 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}}$. If $\mathbf{w}^\star \in \mathcal{C}_t(\delta)$, then, by Lemma E.1, we have

$$\alpha_{ti} \leqslant x_{ti}^\top \mathbf{w}^\star + 2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} = \alpha'_{ti}.$$

We denote $\tilde{\bar{R}}_t(S_t) = \frac{\sum_{i\in S_t}\exp(\alpha'_{ti})r_{ti}}{v_0+\sum_{j\in S_t}\exp(\alpha'_{tj})}$. Then, we can bound the regret as follows:

$$\sum_{t=1}^T R_t(S_t^\star, \mathbf{w}^\star) - R_t(S_t, \mathbf{w}^\star) \leqslant \sum_{t=1}^T \tilde{R}_t(S_t) - R_t(S_t, \mathbf{w}^\star) \leqslant \sum_{t=1}^T \tilde{\bar{R}}_t(S_t) - R_t(S_t, \mathbf{w}^\star),$$

where the first inequality holds by Lemma H.1 and the last inequality holds by Lemma H.2.

Now, we define $\tilde{Q} : \mathbb{R}^{|S_t|} \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_{|S_t|})^\top \in \mathbb{R}^{|S_t|}$, $\tilde{Q}(\mathbf{u}) = \sum_{k=1}^{|S_t|} \frac{\exp(u_k)r_{ti_k}}{v_0+\sum_{j=1}^{|S_t|}\exp(u_j)}$. Let $S_t = \{i_1, \ldots, i_{|S_t|}\}$. Moreover, for all $t \geqslant 1$, let $\mathbf{u}_t = (u_{ti_1}, \ldots u_{ti_{|S_t|}})^\top = (\alpha'_{ti_1}, \ldots, \alpha'_{ti_{|S_t|}})^\top$ and $\mathbf{u}_t^\star = (u^\star_{ti_1}, \ldots u^\star_{ti_{|S_t|}})^\top = (x_{ti_1}^\top \mathbf{w}^\star, \ldots, x_{ti_{|S_t|}}^\top \mathbf{w}^\star)^\top$. Then, by applying a second order Taylor expansion, we obtain

$$\begin{aligned}\sum_{t=1}^T \tilde{\bar{R}}_t(S_t) - R_t(S_t, \mathbf{w}^\star) &= \sum_{t=1}^T \tilde{Q}(\mathbf{u}_t) - \tilde{Q}(\mathbf{u}_t^\star) \\ &= \underbrace{\sum_{t=1}^T \nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t - \mathbf{u}_t^\star)}_{\texttt{(C)}} + \underbrace{\frac{1}{2}\sum_{t=1}^T (\mathbf{u}_t - \mathbf{u}_t^\star)^\top \nabla^2\tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star)}_{\texttt{(D)}},\end{aligned}$$

where $\bar{\mathbf{u}}_t = (\bar{u}_{ti_1}, \ldots, \bar{u}_{ti_{|S_t|}})^\top \in \mathbb{R}^{|S_t|}$ is the convex combination of $\mathbf{u}_t$ and $\mathbf{u}_t^\star$.

We first bound the term `(C)`.

$$\begin{aligned}&\sum_{t=1}^T \nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t - \mathbf{u}_t^\star) \\ &= \sum_{t=1}^T \sum_{i\in S_t} \frac{\exp(x_{ti}^\top\mathbf{w}^\star)r_{ti}}{v_0 + \sum_{k\in S_t}\exp(x_{tk}^\top\mathbf{w}^\star)}(u_{ti} - u^\star_{ti}) - \sum_{j\in S_t} \frac{\exp(x_{tj}^\top\mathbf{w}^\star)r_{tj}\sum_{i\in S_t}\exp(x_{ti}^\top\mathbf{w}^\star)}{(v_0 + \sum_{k\in S_t}\exp(x_{tk}^\top\mathbf{w}^\star))^2}(u_{ti} - u^\star_{ti}) \\ &= \sum_{t=1}^T \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)r_{ti}(u_{ti} - u^\star_{ti}) - \sum_{i\in S_t}\sum_{j\in S_t} p_t(i|S_t, \mathbf{w}^\star)r_{ti}p_t(j|S_t, \mathbf{w}^\star)(u_{tj} - u^\star_{tj}) \\ &= \sum_{t=1}^T \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)r_{ti}\left((u_{ti} - u^\star_{ti}) - \sum_{j\in S_t} p_t(j|S_t, \mathbf{w}^\star)(u_{tj} - u^\star_{tj})\right) \\ &= \sum_{t=1}^T \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)r_{ti}\left(2\beta_t(\delta)\|x_{ti}\|_{H_t^{-1}} - \sum_{j\in S_t} p_t(j|S_t, \mathbf{w}^\star)2\beta_t(\delta)\|x_{tj}\|_{H_t^{-1}}\right) \\ &= 2\sum_{t=1}^T \beta_t(\delta)\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}} - \sum_{j\in S_t} p_t(j|S_t, \mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right).\end{aligned}$$

Let $x_{t0} = \mathbf{0}$. Then, we can further bound the right-hand side as follows:

$$
\begin{aligned}
&2\sum_{t=1}^{T}\beta_t(\delta)\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\sum_{j\in S_t}p_t(j|S_t,\mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right)\\
&\quad=2\sum_{t=1}^{T}\beta_t(\delta)\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\sum_{j\in S_t\cup\{0\}}p_t(j|S_t,\mathbf{w}^\star)\|x_{tj}\|_{H_t^{-1}}\right)\\
&\quad=2\sum_{t=1}^{T}\beta_t(\delta)\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)r_{ti}\left(\|x_{ti}\|_{H_t^{-1}}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right)\\
&\quad\leqslant 2\sum_{t=1}^{T}\beta_t(\delta)\sum_{i\in S_t^+}p_t(i|S_t,\mathbf{w}^\star)\left(\|x_{ti}\|_{H_t^{-1}}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right)\\
&\quad\leqslant 2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t^+}p_t(i|S_t,\mathbf{w}^\star)\left(\|x_{ti}\|_{H_t^{-1}}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[\|x_{tj}\|_{H_t^{-1}}\right]\right)\\
&\quad\leqslant 2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t^+}p_t(i|S_t,\mathbf{w}^\star)\left(\|x_{ti}\|_{H_t^{-1}}-\left\|\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[x_{tj}\right]\right\|_{H_t^{-1}}\right)\\
&\quad\leqslant 2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t^+}p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[x_{tj}\right]\|_{H_t^{-1}}\\
&\quad\leqslant 2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\left\|x_{ti}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[x_{tj}\right]\right\|_{H_t^{-1}},
\end{aligned}
$$

where, in the first inequality, we define $S_t^+ \subseteq S_t$ as the subset of items in S_t such that the term $\|x_{ti}\|_{H_t^{-1}} - \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[\|x_{tj}\|_{H_t^{-1}}\right] \geqslant 0$ and $r_{ti} \in [0,1]$, the second inequality holds because $\beta_1(\delta) \leqslant \cdots \leqslant \beta_T(\delta)$, the third inequality holds due to Jensen's inequality, and the second-to-last inequality holds due to the fact that $\|\mathbf{a}\| = \|\mathbf{a}-\mathbf{b}+\mathbf{b}\| \leqslant \|\mathbf{a}-\mathbf{b}\| + \|\mathbf{b}\|$ for any vectors $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$. For simplicity, we denote $\mathbb{E}_{\mathbf{w}}[x_{tj}] = \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w})}[x_{tj}]$. Let $\bar{x}_{ti} = x_{ti} - \mathbb{E}_{\mathbf{w}^\star}[x_{tj}]$ and $\tilde{x}_{ti} = x_{ti} - \mathbb{E}_{\mathbf{w}_{t+1}}[x_{tj}]$. Then, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[x_{tj}\right]\|_{H_t^{-1}}=\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|\bar{x}_{ti}\|_{H_t^{-1}}\\
&\leqslant\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|\bar{x}_{ti}-\tilde{x}_{ti}\|_{H_t^{-1}}+\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&=\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|\bar{x}_{ti}-\tilde{x}_{ti}\|_{H_t^{-1}}+\sum_{t=1}^{T}\sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}^\star)-p_t(i|S_t,\mathbf{w}_{t+1})\right)\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&+\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|_{H_t^{-1}},
\end{aligned}
\tag{H.1}
$$

where the inequality holds by the triangle inequality. Now, we bound the terms on the right-hand side of (H.1) individually. For the first term, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \|\bar{x}_{ti} - \tilde{x}_{ti}\|_{H_t^{-1}} \\
&\quad = \sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left\|\mathbb{E}_{\mathbf{w}_{t+1}}[x_{tj}] - \mathbb{E}_{\mathbf{w}^\star}[x_{tj}]\right\|_{H_t^{-1}} \\
&\quad = \sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \left\| \sum_{j \in S_t} \left(p_t(j|S_t, \mathbf{w}_{t+1}) - p_t(j|S_t, \mathbf{w}^\star)\right) x_{tj} \right\|_{H_t^{-1}},
\end{aligned}$$

where the last equality holds due to the setting of $x_{t0} = \mathbf{0}$. By the mean value theorem, there exists $\boldsymbol{\xi}_t = (1-c)\mathbf{w}^\star + c\mathbf{w}_{t+1}$ for some $c \in (0,1)$ such that

$$\begin{aligned}
&\left\| \sum_{j \in S_t} \left(p_t(j|S_t, \mathbf{w}_{t+1}) - p_t(j|S_t, \mathbf{w}^\star)\right) x_{tj} \right\|_{H_t^{-1}} = \left\| \sum_{j \in S_t} \nabla p_t(j|S_t, \boldsymbol{\xi}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}^\star) x_{tj} \right\|_{H_t^{-1}} \\
&\leqslant \sum_{j \in S_t} \left|\nabla p_t(j|S_t, \boldsymbol{\xi}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}^\star)\right| \|x_{tj}\|_{H_t^{-1}} \\
&= \sum_{j \in S_t} \left| \left(p_t(j|S_t, \boldsymbol{\xi}_t) x_{tj} - p_t(j|S_t, \boldsymbol{\xi}_t) \sum_{k \in S_t} p_t(k|S_t, \boldsymbol{\xi}_t) x_{tk} \right)^\top (\mathbf{w}_{t+1} - \mathbf{w}^\star) \right| \|x_{tj}\|_{H_t^{-1}} \\
&\leqslant \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \left|x_{tj}^\top (\mathbf{w}_{t+1} - \mathbf{w}^\star)\right| \|x_{tj}\|_{H_t^{-1}} \\
&+ \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|x_{tj}\|_{H_t^{-1}} \sum_{k \in S_t} p_t(k|S_t, \boldsymbol{\xi}_t) \left|x_{tk}^\top (\mathbf{w}_{t+1} - \mathbf{w}^\star)\right| \\
&\leqslant \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_t} \|x_{tj}\|_{H_t^{-1}}^2 \\
&+ \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|x_{tj}\|_{H_t^{-1}} \sum_{k \in S_t} p_t(k|S_t, \boldsymbol{\xi}_t) \|x_{tk}\|_{H_t^{-1}} \|\mathbf{w}_{t+1} - \mathbf{w}^\star\|_{H_t} \\
&\leqslant \beta_{t+1}(\delta) \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|x_{tj}\|_{H_t^{-1}}^2 + \beta_{t+1}(\delta) \left(\sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|x_{tj}\|_{H_t^{-1}} \right)^2 \\
&\leqslant 2\beta_{t+1}(\delta) \sum_{j \in S_t} p_t(j|S_t, \boldsymbol{\xi}_t) \|x_{tj}\|_{H_t^{-1}}^2 \\
&\leqslant 2\beta_{t+1}(\delta) \max_{j \in S_t} \|x_{tj}\|_{H_t^{-1}}^2,
\end{aligned}$$

where the fourth inequality holds by Lemma 1 and the second-to-last inequality holds due to Jensen's inequality. Hence, we have

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \|\bar{x}_{ti} - \tilde{x}_{ti}\|_{H_t^{-1}} &\leqslant 2 \sum_{t=1}^{T} \beta_{t+1}(\delta) \sum_{i \in S_t} p_t(i|S_t, \mathbf{w}^\star) \max_{j \in S_t} \|x_{tj}\|_{H_t^{-1}}^2 \\
&\leqslant 2\beta_{T+1}(\delta) \sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{H_t^{-1}}^2 \\
&\leqslant \frac{4}{\kappa} \beta_{T+1}(\delta) d \log\left(1 + \frac{T}{d\lambda}\right), \qquad \text{(H.2)}
\end{aligned}$$

where the last inequality holds by Lemma E.2. Using similar reasoning, we can bound the second term of (H.1). By the mean value theorem, there exists $\boldsymbol{\xi}'_t = (1-c')\mathbf{w}^\star + c'\mathbf{w}_{t+1}$ for some $c' \in (0,1)$

such that

$$\begin{aligned}
&\sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}^\star)-p_t(i|S_t,\mathbf{w}_{t+1})\right)\|\tilde{x}_{ti}\|_{H_t^{-1}} = \sum_{i\in S_t}\nabla p_t(i|S_t,\boldsymbol{\xi}_t')^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&= \sum_{i\in S_t}\left(p_t(i|S_t,\boldsymbol{\xi}_t')x_{ti} - p_t(i|S_t,\boldsymbol{\xi}_t')\sum_{k\in S_t}p_t(k|S_t,\boldsymbol{\xi}_t')x_{tk}\right)^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&\leqslant \beta_{t+1}(\delta)\sum_{i\in S_t}p_t(i|S_t,\boldsymbol{\xi}_t')\|x_{ti}\|_{H_t^{-1}}\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&+ \beta_{t+1}(\delta)\sum_{i\in S_t}p_t(i|S_t,\boldsymbol{\xi}_t')\|\tilde{x}_{ti}\|_{H_t^{-1}}\sum_{k\in S_t}p_t(k|S_t,\boldsymbol{\xi}_t')\|x_{tk}\|_{H_t^{-1}}\\
&\leqslant \beta_{t+1}(\delta)\max_{i\in S_t}\|x_{ti}\|_{H_t^{-1}}\|\tilde{x}_{ti}\|_{H_t^{-1}} + \beta_{t+1}(\delta)\max_{i\in S_t}\|\tilde{x}_{ti}\|_{H_t^{-1}}\max_{k\in S_t}\|x_{tk}\|_{H_t^{-1}}.
\end{aligned}$$

Then, by applying the AM-GM inequality to each term, we obtain

$$\begin{aligned}
&\beta_{t+1}(\delta)\max_{i\in S_t}\|x_{ti}\|_{H_t^{-1}}\|\tilde{x}_{ti}\|_{H_t^{-1}} + \beta_{t+1}(\delta)\max_{i\in S_t}\|\tilde{x}_{ti}\|_{H_t^{-1}}\max_{k\in S_t}\|x_{tk}\|_{H_t^{-1}}\\
&\leqslant \beta_{t+1}(\delta)\max_{i\in S_t}\frac{\|x_{ti}\|^2_{H_t^{-1}}+\|\tilde{x}_{ti}\|^2_{H_t^{-1}}}{2} + \beta_{t+1}(\delta)\frac{\left(\max_{i\in S_t}\|\tilde{x}_{ti}\|_{H_t^{-1}}\right)^2+\left(\max_{k\in S_t}\|x_{tk}\|_{H_t^{-1}}\right)^2}{2}\\
&= \beta_{t+1}(\delta)\max_{i\in S_t}\frac{\|x_{ti}\|^2_{H_t^{-1}}+\|\tilde{x}_{ti}\|^2_{H_t^{-1}}}{2} + \beta_{t+1}(\delta)\frac{\max_{i\in S_t}\|\tilde{x}_{ti}\|^2_{H_t^{-1}}+\max_{k\in S_t}\|x_{tk}\|^2_{H_t^{-1}}}{2}\\
&\leqslant 2\beta_{t+1}(\delta)\max\left\{\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}},\max_{i\in S_t}\|\tilde{x}_{ti}\|^2_{H_t^{-1}}\right\}
\end{aligned}$$

where the equality holds since $(\max_i a_i)^2 = \max_i a_i^2$ for any $a_i \geqslant 0$. Thus, by Lemma H.3 (or Lemma E.2), we get

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i\in S_t}\left(p_t(i|S_t,\mathbf{w}^\star)-p_t(i|S_t,\mathbf{w}_{t+1})\right)\|\tilde{x}_{ti}\|_{H_t^{-1}}\\
&\qquad\leqslant 2\beta_{t+1}(\delta)\max\left\{\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}},\max_{i\in S_t}\|\tilde{x}_{ti}\|^2_{H_t^{-1}}\right\} \leqslant \frac{4}{\kappa}\beta_{T+1}(\delta)d\log\left(1+\frac{T}{d\lambda}\right),,
\end{aligned}\tag{H.3}$$

Finally, we bound the third term of (H.1). By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|_{H_t^{-1}} &\leqslant \sqrt{\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}_{t+1})}\sqrt{\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|^2_{H_t^{-1}}}\\
&\leqslant \sqrt{T}\sqrt{2d\log\left(1+\frac{T}{d\lambda}\right)},
\end{aligned}\tag{H.4}$$

where the last inequality holds by Lemma H.3. Plugging (H.2), (H.3), and (H.4) into (H.1), we get

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i\in S_t}p_t(i|S_t,\mathbf{w}^\star)\|x_{ti}-\mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w}^\star)}\left[x_{tj}\right]\|_{H_t^{-1}}\\
&\qquad\leqslant \sqrt{T}\sqrt{2d\log\left(1+\frac{T}{d\lambda}\right)} + \frac{8}{\kappa}\beta_{T+1}(\delta)d\log\left(1+\frac{T}{d\lambda}\right)
\end{aligned}$$

Thus, we can bound the term (c) as follows:

$$\sum_{t=1}^{T}\nabla\tilde{Q}(\mathbf{u}_t^\star)^\top(\mathbf{u}_t-\mathbf{u}_t^\star) \leqslant 2\beta_T(\delta)\sqrt{T}\sqrt{2d\log\left(1+\frac{T}{d\lambda}\right)} + \frac{16}{\kappa}\beta_T(\delta)\beta_{T+1}(\delta)d\log\left(1+\frac{T}{d\lambda}\right),\tag{H.5}$$

Now, we bound the term (D). Define $Q : \mathbb{R}^{|S_t|} \to \mathbb{R}$, such that for all $\mathbf{u} = (u_1, \ldots, u_{|S_t|}) \in \mathbb{R}^{|S_t|}$, $Q(\mathbf{u}) = \sum_{i=1}^{|S_t|} \frac{\exp(u_i)}{v_0 + \sum_{j=1}^{|S_t|} \exp(u_j)}$. Then, we have $\left|\frac{\partial^2 \tilde{Q}}{\partial i \partial j}\right| \leqslant \left|\frac{\partial^2 Q}{\partial i \partial j}\right|$ since $r_{ti} \in [0, 1]$. By following the similar reasoning from Equation (E.3) to (E.5) in Section E.1, we have

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}(\mathbf{u}_t - \mathbf{u}_t^\star)^\top \nabla^2 \tilde{Q}(\bar{\mathbf{u}}_t)(\mathbf{u}_t - \mathbf{u}_t^\star) &= \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}(u_{ti} - u_{ti}^\star)\frac{\partial^2 \tilde{Q}}{\partial i \partial j}(u_{tj} - u_{tj}^\star) \\
&\leqslant \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S_t}\sum_{j\in S_t}|u_{ti} - u_{ti}^\star|\left|\frac{\partial^2 Q}{\partial i \partial j}\right||u_{tj} - u_{tj}^\star| \\
&\leqslant 10\beta_T(\delta)^2 \sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}}. \qquad \text{(H.6)}
\end{aligned}$$

where the first inequality holds because $\left|\frac{\partial^2 \tilde{Q}}{\partial i \partial j}\right| \leqslant \left|\frac{\partial^2 Q}{\partial i \partial j}\right|$. Combining (H.5) and (H.6), we derive that

$$\begin{aligned}
\mathbf{Reg}_T(\mathbf{w}^\star) &\leqslant 2\beta_T(\delta)\sqrt{T}\sqrt{2d\log\left(1 + \frac{T}{d\lambda}\right)} + \frac{16}{\kappa}\beta_T(\delta)\beta_{T+1}(\delta)d\log\left(1 + \frac{T}{d\lambda}\right) \\
&\quad + 10\beta_T(\delta)^2\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} \\
&= \tilde{\mathcal{O}}\left(d\sqrt{T} + \frac{1}{\kappa}d^2\right),
\end{aligned}$$

where $\beta_T(\delta) = \mathcal{O}\left(\sqrt{d}\log T \log K\right)$. This concludes the proof of Theorem 4. □

Remark H.1. *If the boundedness assumption on the parameter is relaxed to $\|\mathbf{w}\|_2 \leqslant B$, since $\beta_t(\delta) = \mathcal{O}\left(B\sqrt{d}\log t \log K + B^{3/2}\sqrt{d\log K}\right)$ (refer Corollary F.1), we have $\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(B^{3/2}d\sqrt{T} + \frac{1}{\kappa}d^2\right)$.*

H.2 Proofs of Lemmas for Theorem 4

H.2.1 Proof of Lemma H.2

Proof of Lemma H.2. We prove the result by first showing that for any $i \in S_t$, we have $r_{ti} \geqslant \tilde{R}_t(S_t)$. This can be proven similarly to Lemma G.1. Suppose that there exists $i \in S_t$ for which $r_{ti} < \tilde{R}_t(S_t)$. Removing item i from the assortment S_t results in a higher expected revenue. Consequently, $S_t \neq \operatorname{argmax}_{S\in\mathcal{S}} \tilde{R}_t(S)$, which contradicts the optimality of S_t. Thus, we have

$$r_{ti} \geqslant \tilde{R}_t(S_t), \quad \forall i \in S_t.$$

If we increase α_{ti} to α'_{ti} for all $i \in S_t$, the probability of selecting the outside option decreases. In other words, the sum of probabilities of choosing any $i \in S_t$ increases. Since $r_{ti} \geqslant \tilde{R}_t(S_t)$ for all $i \in S_t$, this results in an increase in revenue. Hence, we get

$$\tilde{R}_t(S_t) = \frac{\sum_{i\in S_t}\exp(\alpha_{ti})r_{ti}}{v_0 + \sum_{j\in S_t}\exp(\alpha_{tj})} \leqslant \frac{\sum_{i\in S_t}\exp(\alpha'_{ti})r_{ti}}{v_0 + \sum_{j\in S_t}\exp(\alpha'_{tj})}.$$

This concludes the proof. □

H.2.2 Proof of Lemma H.3

Proof of Lemma H.3. For notational simplicity, let $\mathbb{E}_{\mathbf{w}}[x_{tj}] = \mathbb{E}_{j\sim p_t(\cdot|S_t,\mathbf{w})}[x_{tj}]$. Let $x_{t0} = \mathbf{0}$. We can rewrite $\mathcal{G}_s(\mathbf{w})$ in the following way:

$$
\begin{aligned}
&\mathcal{G}_s(\mathbf{w}_{s+1}) \\
&= \sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1}) x_{si}x_{si}^\top - \sum_{i\in S_s}\sum_{j\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})p_s(j|S_s,\mathbf{w}_{s+1}) x_{si}x_{sj}^\top \qquad \text{(H.7)} \\
&= \sum_{i\in S_s\cup\{0\}} p_s(i|S_s,\mathbf{w}_{s+1}) x_{si}x_{si}^\top - \sum_{i\in S_s\cup\{0\}}\sum_{j\in S_s\cup\{0\}} p_s(i|S_s,\mathbf{w}_{s+1})p_s(j|S_s,\mathbf{w}_{s+1}) x_{si}x_{sj}^\top \\
&= \mathbb{E}_{\mathbf{w}_{s+1}}[x_{si}x_{si}^\top] - \mathbb{E}_{\mathbf{w}_{s+1}}[x_{si}]\left(\mathbb{E}_{\mathbf{w}_{s+1}}[x_{si}]\right)^\top \\
&= \mathbb{E}_{\mathbf{w}_{s+1}}\left[(x_{si} - \mathbb{E}_{\mathbf{w}_{s+1}}[x_{sm}])(x_{si} - \mathbb{E}_{\mathbf{w}_{s+1}}[x_{sm}])^\top\right] \\
&= \mathbb{E}_{\mathbf{w}_{s+1}}[\tilde{x}_{si}\tilde{x}_{si}^\top] = \sum_{i\in S_s\cup\{0\}} p_s(i|S_s,\mathbf{w}_{s+1})\tilde{x}_{si}\tilde{x}_{si}^\top \succeq \sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})\tilde{x}_{si}\tilde{x}_{si}^\top.
\end{aligned}
$$

This means that

$$
H_{t+1} = H_t + \mathcal{G}_t(\mathbf{w}_{t+1}) \succeq H_t + \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}_{t+1})\tilde{x}_{ti}\tilde{x}_{ti}^\top. \qquad \text{(H.8)}
$$

Hence, we can derive that

$$
\det(H_{t+1}) \geqslant \det(H_t)\left(1 + \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|^2_{H_t^{-1}}\right).
$$

Since $\lambda \geqslant 1$, for all $t \geqslant 1$ we have $\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}_{t+1})\|\tilde{x}_{ti}\|^2_{H_t^{-1}} \leqslant \frac{1}{\lambda}\max_{i\in S_t}\|\tilde{x}_{ti}\|_2 \leqslant 1$. Then, using the fact that $z \leqslant 2\log(1+z)$ for any $z \in [0,1]$, we get

$$
\begin{aligned}
\sum_{s=1}^{t}\sum_{i\in S_s} p_s(i|S_s,\mathbf{w}_{s+1})\|\tilde{x}_{si}\|^2_{H_s^{-1}} &\leqslant 2\sum_{s=1}^{t}\log\left(1 + p_s(i|S_s,\mathbf{w}_{s+1})\|\tilde{x}_{si}\|^2_{H_s^{-1}}\right) \\
&\leqslant 2\sum_{s=1}^{t}\log\left(\frac{\det(H_{s+1})}{\det(H_s)}\right) \\
&= 2\log\left(\frac{\det(H_{t+1})}{\det(H_1)}\right) \\
&\leqslant 2d\log\left(\frac{\mathrm{tr}(H_{t+1})}{d\lambda}\right) \leqslant 2d\log\left(1 + \frac{t}{d\lambda}\right).
\end{aligned}
$$

This proves the first inequality.

To show the second inequality, we come back to equation (H.8). By the definition of κ, we get

$$
\begin{aligned}
H_{t+1} = H_t + \mathcal{G}_t(\mathbf{w}_{t+1}) &= H_t + \sum_{i\in S_t} p_t(i|S_t,\mathbf{w}_{t+1})\tilde{x}_{ti}\tilde{x}_{ti}^\top \\
&\succeq H_t + \kappa\sum_{i\in S_t}\tilde{x}_{ti}\tilde{x}_{ti}^\top.
\end{aligned}
$$

Thus, we obtain that

$$
\det(H_{t+1}) \geqslant \det(H_t)\left(1 + \kappa\sum_{i\in S_t}\|\tilde{x}_{ti}\|^2_{H_t^{-1}}\right) \geqslant \det(H_t)\left(1 + \kappa\max_{i\in S_t}\|\tilde{x}_{ti}\|^2_{H_t^{-1}}\right).
$$

Since $\lambda \geqslant 1$, for all $t \geqslant 1$ we have $\kappa \max_{i \in S_t} \|\tilde{x}_{ti}\|^2_{H_t^{-1}} \leqslant \frac{\kappa}{\lambda}\|\tilde{x}_{ti}\|_2 \leqslant \kappa$. We then reach the conclusion in the same manner:

$$\begin{aligned}
\sum_{s=1}^{t} \max_{i \in S_s} \|\tilde{x}_{si}\|^2_{H_s^{-1}} &\leqslant \frac{2}{\kappa} \sum_{s=1}^{t} \log\left(1 + \kappa \max_{i \in S_s} \|\tilde{x}_{si}\|^2_{H_s^{-1}}\right) \\
&\leqslant \frac{2}{\kappa} \sum_{s=1}^{t} \log\left(\frac{\det(H_{s+1})}{\det(H_s)}\right) \\
&= \frac{2}{\kappa} \log\left(\frac{\det(H_{t+1})}{\det(H_1)}\right) \\
&\leqslant \frac{2}{\kappa} d \log\left(\frac{\operatorname{tr}(H_{t+1})}{d\lambda}\right) \leqslant \frac{2}{\kappa} d \log\left(1 + \frac{t}{d\lambda}\right).
\end{aligned}$$

This proves the second inequality. □

I Proof of Proposition 1

I.1 Main Proof of Proposition 1

Proof of Proposition 1. We construct the parameters and features in the same way as in Section D.1. Recall that, by Proposition D.1, it is sufficient to bound $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V)$. We denote $\tilde{U}_t$ as the unique $U^\star \in \mathcal{V}_{d/4}$ in $\tilde{S}_t$. We define a function $\mu : \mathbb{R} \to [0, 1]$ such that for any $z \in \mathbb{R}$, $\mu(z) = \frac{K \exp(z)}{v_0 + K \exp(z)}$. Since all items in $S^\star$ (and in $\tilde{S}_t$) are identical, we can express $\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V)$ (and $\sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V)$) as follows:

$$\begin{aligned}
\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) &= \frac{K \exp(x_V^\top \mathbf{w}_V)}{v_0 + K \exp(x_V^\top \mathbf{w}_V)} = \mu(x_V^\top \mathbf{w}_V), \\
\sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) &= \frac{K \exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)}{v_0 + K \exp(x_{\tilde{U}_t}^\top \mathbf{w}_V)} = \mu(x_{\tilde{U}_t}^\top \mathbf{w}_V).
\end{aligned}$$

Given $\mathbf{w}_V$, we define $\kappa_t^\star(\mathbf{w}_V) := \sum_{i \in S^\star} p_t(i|S^\star, \mathbf{w}_V) p_t(0|S^\star, \mathbf{w}_V)$. Since the context vectors $\{x_{ti}\}$ are constructed to be invariant across rounds t, $\kappa_t^\star(\mathbf{w}_V)$ is also independent of t, i.e., $\kappa_1^\star(\mathbf{w}_V) = \cdots = \kappa_T^\star(\mathbf{w}_V)$. Therefore, we omit the index t for simplicity. Note that, in our instance construction, $\kappa^\star(\mathbf{w}_V) = \mu(x_V^\top \mathbf{w}_V)\left(1 - \mu(x_V^\top \mathbf{w}_V)\right) = \dot{\mu}(x_V^\top \mathbf{w}_V)$.

We hypothesize that for all $V \in \mathcal{V}_{d/4}$, the regret is dominated by $d\sqrt{\kappa^\star(\mathbf{w}_V)T}$. If this assumption does not hold, then by definition, there exists some $V \in \mathcal{V}_{d/4}$ such that $\mathbb{E}_V\left[\mathbf{Reg}_T(\mathbf{w}_V)\right] = \Omega\left(d\sqrt{\kappa^\star(\mathbf{w}_V)T}\right)$, thereby completing the proof.

Hypothesis. There exists a constant $C > 0$ such that:

$$\mathbb{E}_V\left[\mathbf{Reg}_T(\mathbf{w}_V)\right] \leqslant C \cdot d\sqrt{\kappa^\star(\mathbf{w}_V)T}, \quad \forall V \in \mathcal{V}_{d/4}. \tag{I.1}$$

Additionally, we set $\mathbf{w}^\star = \operatorname{argmax}_{\mathbf{w}_V} \kappa^\star(\mathbf{w}_V)$, thus, we have $\kappa^\star = \kappa^\star(\mathbf{w}^\star) = \max_{\mathbf{w}_V} \kappa^\star(\mathbf{w}_V)$.

To establish an instance-dependent lower bound for $\mu(x_V^\top \mathbf{w}_V) - \mu(x_{\tilde{U}_t}^\top \mathbf{w}_V)$, we use the following lemma in place of Lemma D.1:

Lemma I.1. *Suppose $\epsilon \in (0, 1/d\sqrt{d})$ and define $\delta := d/4 - |\tilde{U}_t \cap V|$. Then, we have*

$$\mu(x_V^\top \mathbf{w}_V) - \mu(x_{\tilde{U}_t}^\top \mathbf{w}_V) \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\delta\epsilon}{\sqrt{d}}.$$

For any $j \in V$, define random variables $\tilde{M}_j := \sum_{t=1}^{T} \mathbf{1}\{j \in \tilde{U}_t\}$. Then, by Lemma I.1, for all $V \in \mathcal{V}_{d/4}$, we get

$$\mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\epsilon}{\sqrt{d}} \left(\frac{dT}{4} - \sum_{j \in V} \mathbb{E}_V[\tilde{M}_j]\right). \tag{I.2}$$

Additionally, we define $\mathcal{V}_{d/4}^{(j)} := \{V \in \mathcal{V}_{d/4} : j \in V\}$ and $\mathcal{V}_{d/4-1} := \{V \subseteq [d] : |V| = d/4 - 1\}$. By averaging both sides of Equation (I.2) with respect to all $V \in \mathcal{V}_{d/4}$, and by following reasoning similar to that in the proof of Theorem 1, we get

$$\begin{aligned}
&\frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \\
&\qquad \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\epsilon}{\sqrt{d}} \left(\frac{dT}{6} - \max_{V \in \mathcal{V}_{d/4-1}} \sum_{j \notin V} \left| \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] - \mathbb{E}_V[\tilde{M}_j] \right| \right). \qquad \text{(I.3)}
\end{aligned}$$

For simplicity, let $P = \mathbb{P}_V$ and $Q = \mathbb{P}_{V \cup \{j\}}$. Then, we can bound the term $\left| \mathbb{E}_{V \cup \{j\}}[\tilde{M}_j] - \mathbb{E}_V[\tilde{M}_j] \right|$ in (I.3) for any $V \in \mathcal{V}_{d/4-1}$.

$$\begin{aligned}
\left| \mathbb{E}_P[\tilde{M}_j] - \mathbb{E}_Q[\tilde{M}_j] \right| &\leqslant \sum_{t=0}^{T} t \cdot \left| P[\tilde{M}_j = t] - Q[\tilde{M}_j = t] \right| \\
&\leqslant T \cdot \|P - Q\|_{\mathrm{TV}} \leqslant T \cdot \sqrt{\frac{1}{2} \mathrm{KL}(P \| Q)}, \qquad \text{(I.4)}
\end{aligned}$$

where $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$ is the total variation distance between P and Q, $\mathrm{KL}(P\|Q) = \int (\log \mathrm{d}P/\mathrm{d}Q) \mathrm{d}P$ is the Kullback-Leibler (KL) divergence between P and Q, and the last inequality holds by Pinsker's inequality. We can derive the instance-dependent bound for the KL divergence term in (I.4) by the following lemma:

Lemma I.2. *For any $V \in \mathcal{V}_{d/4-1}$ and $j \in [d]$, there exists a positive constant $C_{\mathrm{KL}} > 0$ such that*

$$\sum_{j=1}^{d} \mathrm{KL}(P_V \| Q_{V \cup \{j\}}) \leqslant C_{\mathrm{KL}} \cdot \epsilon^2 \kappa^\star(\mathbf{w}_V) T.$$

Plugging (I.4) into (I.3), we get

$$\begin{aligned}
&\frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \\
&\qquad \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\epsilon}{\sqrt{d}} \left(\frac{dT}{6} - T \sum_{j=1}^{d} \sqrt{\frac{1}{2} \mathrm{KL}(P\|Q)} \right) \\
&\qquad \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\epsilon}{\sqrt{d}} \left(\frac{dT}{6} - T\sqrt{d} \cdot \sqrt{\frac{1}{2} \sum_{j=1}^{d} \mathrm{KL}(P\|Q)} \right) \\
&\qquad \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{\epsilon}{\sqrt{d}} \left(\frac{dT}{6} - T\sqrt{d} \cdot \sqrt{\frac{1}{2} C_{\mathrm{KL}} \cdot \epsilon^2 \kappa^\star(\mathbf{w}_V) T} \right),
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwartz inequality, and the last inequality is by Lemma I.2.

By setting $\epsilon = \sqrt{\frac{d}{72 C_{\mathrm{KL}} \cdot \kappa^\star(\mathbf{w}_V) T}}$, we have

$$\begin{aligned}
&\frac{1}{|\mathcal{V}_{d/4}|} \sum_{V \in \mathcal{V}_{d/4}} \mathbb{E}_V \sum_{i \in S^\star} p(i|S^\star, \mathbf{w}_V) - \sum_{i \in \tilde{S}_t} p(i|\tilde{S}_t, \mathbf{w}_V) \\
&\qquad \geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3} \cdot \frac{1}{\sqrt{d}} \cdot \sqrt{\frac{d}{72 C_{\mathrm{KL}} \cdot \kappa^\star(\mathbf{w}_V) T}} \cdot \frac{dT}{12} \\
&\qquad = \Omega\left(d \sqrt{\kappa^\star(\mathbf{w}_V) T} \right).
\end{aligned}$$

Recall that by construction, $\kappa^\star = \max_{\mathbf{w}_V} \kappa^\star(\mathbf{w}_V)$. Thus, by taking the maximum over $\mathbf{w}_V$, we complete the proof of Proposition 1. □

I.2 Proofs of Lemmas for Proposition 1

I.2.1 Proof of Lemma I.1

Proof of Lemma I.1. For simplicity, let $x = x_V$ and $\hat{x} = x_{\tilde{U}_t}$. Then, by the mean value theorem, we have

$$\begin{aligned}
\mu(x^\top \mathbf{w}_V) - \mu(\hat{x}^\top \mathbf{w}_V) &= \int_{v=1}^{1} \dot{\mu}\left(x^\top \mathbf{w}_V + v(\hat{x}-x)^\top \mathbf{w}_V\right)(x-\hat{x})^\top \mathbf{w}_V \\
&\geqslant \frac{\dot{\mu}(x^\top \mathbf{w}_V)}{1+|(x-\hat{x})^\top \mathbf{w}_V|}(x-\hat{x})^\top \mathbf{w}_V \\
&\geqslant \frac{\dot{\mu}(x^\top \mathbf{w}_V)}{3}(x-\hat{x})^\top \mathbf{w}_V \\
&= \frac{\kappa^\star(\mathbf{w}_V)}{3}(x-\hat{x})^\top \mathbf{w}_V \\
&\geqslant \frac{\kappa^\star(\mathbf{w}_V)}{3}\cdot\frac{\delta\epsilon}{\sqrt{d}},
\end{aligned}$$

where in the first equality, $\bar{x}$ is the convex combination of x and $\hat{x}$, the first inequality holds by Lemma I.3, the second inequality holds by the bounded assumption (Assumption 1), and the last inequality holds by the definition of $\delta = d/4 - |\tilde{U}_t \cap V|$. □

I.2.2 Proof of Lemma I.2

Proof of Lemma I.2. Consider a fixed round t, an assortment $\tilde{S}_t$, and the set $\tilde{U}_t$. Let $U = \tilde{U}_t$. Define $m_j(\tilde{S}_t) := \sum_{x_U \in \tilde{S}_t} \mathbf{1}\{j \in U\}/K$, which captures the average presence of $j \in U$ across the assortment $\tilde{S}_t$. For simplicity, we denote $p = \frac{K\exp(x_U^\top \mathbf{w}_V)}{v_0 + K\exp(x_U^\top \mathbf{w}_V)}$ and $q = \frac{K\exp(x_U^\top \mathbf{w}_{V\cup\{j\}})}{v_0 + K\exp(x_U^\top \mathbf{w}_{V\cup\{j\}})}$. Then, we get

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_V(\cdot|\tilde{S}_t)\|\mathbb{P}_{V\cup\{j\}}(\cdot|\tilde{S}_t)\right) &\leqslant \chi^2\left(\mathbb{P}_V(\cdot|\tilde{S}_t)\|\mathbb{P}_{V\cup\{j\}}(\cdot|\tilde{S}_t)\right) \\
&= \chi^2(\mathrm{Bernoulli}\,(p)\,\|\,\mathrm{Bernoulli}\,(q)) \\
&= \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q} \\
&= \frac{(p-q)^2}{q(1-q)} \\
&= \frac{\left(\mu(x_U^\top \mathbf{w}_V) - \mu(x_U^\top \mathbf{w}_{V\cup\{j\}})\right)^2}{\dot{\mu}(x_U^\top \mathbf{w}_{V\cup\{j\}})} \\
&= \frac{\left(\dot{\mu}(x_U^\top \bar{\mathbf{w}}_{V,j})\right)^2}{\dot{\mu}(x_U^\top \mathbf{w}_{V\cup\{j\}})}\left(x_U^\top\left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right)\right)^2,
\end{aligned}$$

where in the first inequality, we used $\mathrm{KL} \leqslant \chi^2$ (Tsybakov [50], Chapter 2), where χ^2 is a chi-square divergence, the second equality holds by the expression of the chi-square divergence for Bernoulli random variables, and the last equality holds by the mean value theorem, where $\bar{\mathbf{w}}_{V,j}$ is a convex combination of $\mathbf{w}_V$ and $\mathbf{w}_{V\cup\{j\}}$. Then, by Lemma I.4, we can further bound the last term as follows:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_V(\cdot|\tilde{S}_t)\|\mathbb{P}_{V\cup\{j\}}(\cdot|\tilde{S}_t)\right) &\leqslant \dot{\mu}(x_U^\top \mathbf{w}_V)e^{3\left|x_U^\top(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}})\right|}\left(x_U^\top\left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right)\right)^2 \\
&\leqslant \dot{\mu}(x_U^\top \mathbf{w}_V)e^{3/d^2}\left(x_U^\top\left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right)\right)^2 \\
&\leqslant \dot{\mu}(x_U^\top \mathbf{w}_V)e^{1/4}\left(x_U^\top\left(\mathbf{w}_V - \mathbf{w}_{V\cup\{j\}}\right)\right)^2 \\
&\leqslant \dot{\mu}(x_U^\top \mathbf{w}_V)e^{1/4}\cdot\frac{m_j(\tilde{S}_t)\epsilon^2}{d},
\end{aligned}$$

where the second-to-the last inequality holds based on the assumption that $d \geqslant 4$, and the last inequality holds since $m_j(\tilde{S}_t) \leqslant 1$.

Now, let us consider t varying over the round $t \in [T]$. Then, we obtain

$$\begin{aligned}
\sum_{j=1}^{d} \mathrm{KL}(P_V \| Q_{V\cup\{j\}}) &= \sum_{j=1}^{d}\sum_{t=1}^{T} \mathbb{E}_V\left[\mathrm{KL}\left(\mathbb{P}_V(\cdot|\tilde{S}_t)\|\mathbb{P}_{V\cup\{j\}}(\cdot|\tilde{S}_t)\right)\right] \\
&\leqslant e^{1/4}\cdot\frac{\epsilon^2}{d}\sum_{j=1}^{d}\sum_{t=1}^{T}\mathbb{E}_V\left[\dot{\mu}(x_{\tilde{U}_t}^\top \mathbf{w}_V)\cdot m_j(\tilde{S}_t)\right] \\
&= e^{1/4}\cdot\frac{\epsilon^2}{d}\mathbb{E}_V\left[\sum_{t=1}^{T}\left(\dot{\mu}(x_{\tilde{U}_t}^\top \mathbf{w}_V)\sum_{j=1}^{d} m_j(\tilde{S}_t)\right)\right] \\
&\leqslant e^{1/4}\cdot\frac{\epsilon^2}{4}\mathbb{E}_V\left[\sum_{t=1}^{T}\dot{\mu}(x_{\tilde{U}_t}^\top \mathbf{w}_V)\right] \\
&\leqslant e^{1/4}\cdot\frac{\epsilon^2}{4}\left(\dot{\mu}(x_V^\top \mathbf{w}_V)T + \mathbb{E}_V\left[\mathbf{Reg}_T(\mathbf{w}_V)\right]\right) \\
&= e^{1/4}\cdot\frac{\epsilon^2}{4}\left(\kappa^\star(\mathbf{w}_V)T + \mathbb{E}_V\left[\mathbf{Reg}_T(\mathbf{w}_V)\right]\right) \\
&\leqslant C_{\mathrm{KL}}\cdot\frac{\epsilon^2}{2}\left(\kappa^\star(\mathbf{w}_V)T + d\sqrt{\kappa^\star(\mathbf{w}_V)T}\right),
\end{aligned}$$

the equality follows from the chain rule of relative entropy (cf. Exercise 14.11 of Lattimore and Szepesvári [32]), the second inequality holds because $\sum_{j=1}^{d} m_j(\tilde{S}_t) \leqslant \frac{d}{4}$, second-to-the last inequality holds by Lemma I.5, and the last inequality holds by the hypothesis (Equation (I.1)) with $C_{\mathrm{KL}} > 0$.

Furthermore, by the definition of κ (Assumption 2), we know that $\kappa^\star(\mathbf{w}_V) \geqslant \kappa$. Hence, given the assumption that $T \geqslant d^2/\kappa \geqslant d^2/\kappa^\star(\mathbf{w}_V)$, we derive that

$$\sum_{j=1}^{d} \mathrm{KL}(P_V \| Q_{V\cup\{j\}}) \leqslant C_{\mathrm{KL}}\cdot\epsilon^2\kappa^\star(\mathbf{w}_V)T,$$

which concludes the proof. □

I.3 Technical Lemmas for Proposition 1

Lemma I.3 (Lemma 7 of Abeille et al. [4])**.** *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 $z_1, z_2 \in \mathcal{Z}$:*

$$\int_{v=0}^{1} \dot{f}\left(z_1 + v(z_2 - z_1)\right) dv \geqslant \frac{\dot{f}(z)}{1 + |z_1 - z_2|}, \quad \textit{for } z \in \{z_1, z_2\}.$$

Lemma I.4 (Lemma 9 of Abeille et al. [4])**.** *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 $z_1, z_2 \in \mathcal{Z}$:*

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

Lemma I.5 (Lemma 11 of Perivier and Goyal [44])**.** *Let $\kappa_t^\star := \sum_{i\in S_t^\star} p_t(i|S_t^\star, \mathbf{w}^\star)p_t(0|S_t^\star, \mathbf{w}^\star)$. Then, we have*

$$\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)p_t(0|S_t, \mathbf{w}^\star) \leqslant \sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star).$$

J Proof of Proposition 2

In this section, we provide the proof of Proposition 2. We introduce useful lemmas to support the proof.

Lemma J.1 (Theorem 4 of Tran-Dinh et al. [49]). *Let $f : \mathbb{R}^K \to \mathbb{R}$ be a M_f-self-concordant-like function and let $x, y \in \mathrm{dom}(f)$, then:*

$$e^{-M_f\|y-x\|_2}\nabla^2 f(x) \preceq \nabla^2 f(y).$$

Proof of Proposition 2. We begin the proof with Equation (E.2) from the proof of Theorem 2:

$$\sum_{t=1}^{T} \nabla Q(\mathbf{u}_t^\star)^\top(\mathbf{u}_t - \mathbf{u}_t^\star) \leqslant 2\beta_T(\delta) \sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)p_t(0|S_t, \mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}}. \tag{J.1}$$

We then divide the total rounds into two disjoint sets, J_1 and J_2, such that $J_1 \bigcup J_2 = [T]$. Specifically, let $J_1 = \{t \in [T] | \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)p_t(0|S_t, \mathbf{w}^\star) \geqslant \sum_{i\in S_t} p_t(i|S_t, \mathbf{w}_{t+1})p_t(0|S_t, \mathbf{w}_{t+1})\}$ and $J_2 = [T]\backslash J_1$. For a better presentation, we also define that:

$$g_t(S; \mathbf{w}) := \sum_{i\in S} p_t(i|S, \mathbf{w})p_t(0|S, \mathbf{w})\|x_{ti}\|_{H_t^{-1}}.$$

Then, we can rewrite the right-hand side of (J.1) as follows:

$$\begin{aligned}
&2\beta_T(\delta) \sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)p_t(0|S_t, \mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}} \\
&\qquad = 2\beta_T(\delta) \sum_{t=1}^{T} g_t(S_t; \mathbf{w}^\star) \\
&\qquad = 2\beta_T(\delta) \sum_{t\in J_1} g_t(S_t; \mathbf{w}^\star) + 2\beta_T(\delta) \sum_{t\in J_2} g_t(S_t; \mathbf{w}^\star).
\end{aligned}$$

To bound $\sum_{t\in J_1} g_t(S_t; \mathbf{w}^\star)$, by the mean value theorem, we get

$$\sum_{i\in J_1} g_t(S_t; \mathbf{w}^\star) = \sum_{i\in J_1} g_t(S_t; \mathbf{w}_{t+1}) + \sum_{i\in J_1} \nabla_{\mathbf{w}} g_t(S_t; \bar{\mathbf{w}}_t)^\top(\mathbf{w}^\star - \mathbf{w}_{t+1}), \tag{J.2}$$

where $\bar{\mathbf{w}}_t$ is the convex combination of $\mathbf{w}^\star$ and $\mathbf{w}_{t+1}$. The first term of (J.2) can be bound by

$$\begin{aligned}
&\sum_{t\in J_1} g_t(S_t; \mathbf{w}_{t+1}) \\
&\leqslant \sqrt{\sum_{t\in J_1}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}_{t+1})p_t(0|S_t, \mathbf{w}_{t+1})} \sqrt{\sum_{t\in J_1}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}_{t+1})p_t(0|S_t, \mathbf{w}_{t+1})\|x_{ti}\|^2_{H_t^{-1}}} \\
&\leqslant \sqrt{\sum_{t\in J_1}\sum_{i\in S_t} p_t(i|S_t, \mathbf{w}^\star)p_t(0|S_t, \mathbf{w}^\star)} \cdot \tilde{\mathcal{O}}\left(\sqrt{d}\right) \leqslant \sqrt{\sum_{t=1}^{T} \kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star)} \cdot \tilde{\mathcal{O}}\left(\sqrt{d}\right),
\end{aligned} \tag{J.3}$$

where the second-to-the last inequality holds by the definition of J_1 and the elliptical potential lemma (Lemma E.2), and the last inequality holds by the following Lemma:

Moreover, the second term of (J.2) can be bounded as follows:

$$
\begin{aligned}
&\left|\sum_{i\in J_1} \nabla_{\mathbf{w}} g_t(S_t;\bar{\mathbf{w}}_t)^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\right| \\
&= \Bigg| \sum_{t\in J_1} p_t(0|S_t,\bar{\mathbf{w}}_t)\sum_{i\in S_t} p_t(i|S_t,\bar{\mathbf{w}}_t)x_{ti}^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\|x_{ti}\|_{H_t^{-1}} \\
&- 2p_t(0|S_t,\bar{\mathbf{w}}_t)\sum_{i\in S_t} p_t(i|S_t,\bar{\mathbf{w}}_t)\|x_{ti}\|_{H_t^{-1}}\sum_{j\in S_t} p_t(j|S_t,\bar{\mathbf{w}}_t)x_{tj}^\top(\mathbf{w}^\star-\mathbf{w}_{t+1})\Bigg| \\
&\leqslant \beta_T(\delta)\sum_{t\in J_1}\sum_{i\in S_t} p_t(i|S_t,\bar{\mathbf{w}}_t)\|x_{ti}\|^2_{H_t^{-1}} + 2\beta_T(\delta)\sum_{t\in J_1}\left(\sum_{i\in S_t} p_t(i|S_t,\bar{\mathbf{w}}_t)\|x_{ti}\|_{H_t^{-1}}\right)^2 \\
&\leqslant \beta_T(\delta)\sum_{t\in J_1}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} + 2\beta_T(\delta)\sum_{t\in J_1}\left(\max_{i\in S_t}\|x_{ti}\|_{H_t^{-1}}\right)^2 \\
&\leqslant 3\beta_T(\delta)\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{H_t^{-1}} = \tilde{\mathcal{O}}\left(d^{3/2}/\kappa\right), \qquad \text{(J.4)}
\end{aligned}
$$

where the first inequality holds by Lemma 1, and the last inequality holds by Lemma E.2. Combining (J.3) and (J.4), we can bound the term $\sum_{t\in J_1} g_t(S_t;\mathbf{w}^\star)$ as follows:

$$
\sum_{t\in J_1} g_t(S_t;\mathbf{w}^\star) \leqslant \tilde{\mathcal{O}}\left(\sqrt{d}\cdot\sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star)} + d^{3/2}/\kappa\right). \qquad \text{(J.5)}
$$

Now, we bound $\sum_{t\in J_2} g_t(S_t;\mathbf{w}^\star)$. We define $H_t^\star := \lambda\mathbf{I}_d + \sum_{s=1}^{t-1}\mathcal{G}_s(\mathbf{w}^\star)$. Recall that $\nabla^2\ell_s(\mathbf{w}) = \mathcal{G}_s(\mathbf{w})$ and ℓ_s is $3\sqrt{2}$-self-concordant-like function (Proposition C.1). Then, by Lemma J.1, we get $H_t \succeq H_t^\star e^{-3\sqrt{2}}$. With this fact, we can now bound the term $\sum_{t\in J_2} g_t(S_t;\mathbf{w}^\star)$:

$$
\begin{aligned}
\sum_{t\in J_2} g_t(S_t;\mathbf{w}^\star) &\leqslant \sqrt{\sum_{t\in J_2}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)}\sqrt{\sum_{t\in J_2}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|^2_{H_t^{-1}}} \\
&\leqslant \sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T}\sqrt{\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|^2_{H_t^{-1}}} \\
&\leqslant \sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T}\sqrt{e^{3\sqrt{2}}\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|^2_{(H_t^\star)^{-1}}} \\
&\leqslant \tilde{\mathcal{O}}\left(\sqrt{d}\cdot\sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T}\right), \qquad \text{(J.6)}
\end{aligned}
$$

where the second inequality follows from Lemma I.5, second-to-the last inequality holds because $H_t \succeq H_t^\star e^{-3\sqrt{2}}$, and the last inequality is obtained by a slight modification of Lemma E.2 with $\mathbf{w}_1,\ldots,\mathbf{w}_T = \mathbf{w}^\star$.

Now, by plugging (J.5) and (J.6) into (J.1), and using the fact that $\beta_T(\delta) = \tilde{\mathcal{O}}(\sqrt{d})$, we obtain that

$$
2\beta_T(\delta)\sum_{t=1}^{T}\sum_{i\in S_t} p_t(i|S_t,\mathbf{w}^\star)p_t(0|S_t,\mathbf{w}^\star)\|x_{ti}\|_{H_t^{-1}} = \tilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star)} + d^{3/2}/\kappa\right).
$$

Furthermore, by applying the same analysis as in the proof of Theorem 2 in order to bound the second-order term in the Taylor expansion (term (B) in (E.1)), we get

$$
\mathbf{Reg}_T(\mathbf{w}^\star) = \tilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\kappa_t^\star + \mathbf{Reg}_T(\mathbf{w}^\star)} + d^2/\kappa\right).
$$

By solving the above equation, we conclude the proof of Proposition 2. □

K Experiment Details and Additional Results

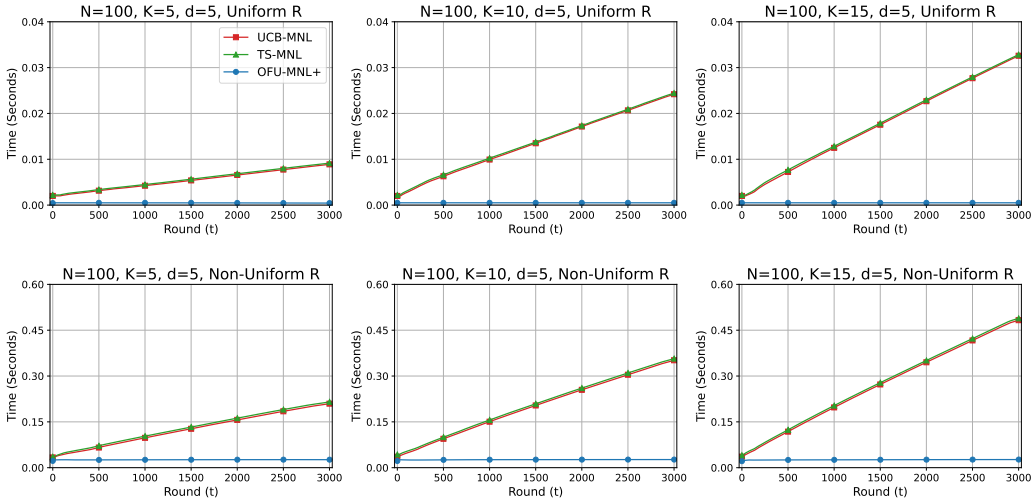

Figure K.1: Runtime per round under uniform rewards (first row) and non-uniform rewards (second row).

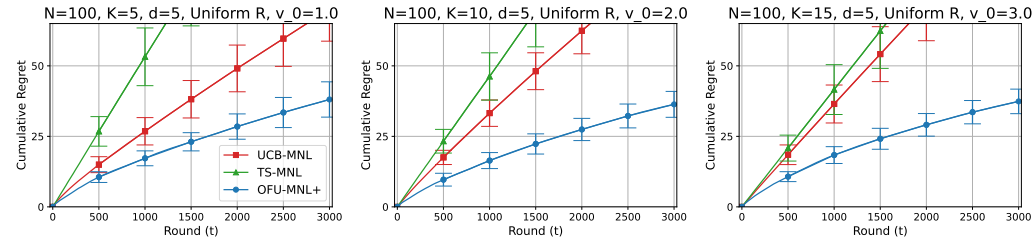

Figure K.2: Cumulative regret under uniform rewards with $v_0 = \Theta(K)$.

For each instance, we sample the true parameter $\mathbf{w}^\star$ from a uniform distribution in $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. For the context features x_{ti}, we sample each x_{ti} independently and identically distributed (i.i.d.) from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and clip it to range $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. Therefore, we ensure that $\|\mathbf{w}^\star\|_2 \leqslant 1$ and $\|x_{ti}\|_2 \leqslant 1$, satisfying Assumption 1. For each experimental configuration, we conducted 20 independent runs for each instance and reported the average cumulative regret (Figure 1) and runtime per round (Figure K.1) for each algorithm. The error bars in Figure 1 and K.2 represent the standard deviations (1-sigma error). We have omitted the error bars in Figure K.1 because they are minimal.

In the uniform reward setting where $r_{ti} = 1$, the combinatorial optimization step to select the assortment simply involves sorting items by their utility estimate. In contrast, in the non-uniform reward setting, rewards are sampled from a uniform distribution in each round, i.e., $r_{ti} \sim \mathrm{Unif}(0, 1)$. For combinatorial optimization in this setting, we solve an equivalent linear programming (LP) problem that is solvable in polynomial-time [47, 21]. The experiments are run on Xeon(R) Gold 6226R CPU @ 2.90GHz (16 cores).

Figure K.1 presents additional empirical results on the runtime per round. Our algorithm `OFU-MNL+` demonstrates a constant computation cost for each round, while the other algorithms exhibit a linear dependence on t. It is also noteworthy that the runtime for uniform rewards is approximately 10 times faster than that for non-uniform rewards. This difference arises because we use linear

programming (LP) optimization for assortment selection in the non-uniform reward setting, which is more computationally intensive.

Furthermore, Figure K.2 illustrates the cumulative regrets of the proposed algorithm compared to other baseline algorithms under uniform rewards with $v_0 = K/5$. Since v_0 is proportional to K, an increase in K does not improve the regret. This observation is also consistent with our theoretical results.

L Technical Errors in Agrawal et al. [5]

In this section, we discuss the technical errors in Agrawal et al. [5]. There are two main significant errors: There are mainly two significant errors:

1. Equation (16).

$$\alpha_i(\mathbf{X}_{S_t}, \theta_t, \theta^\star)x_{ti} := \mu_i(\mathbf{X}_{S_t}^\top \theta^\star) - \mu_i(\mathbf{X}_{S_t}^\top \theta_t), \tag{L.1}$$

where $\mathbf{X}_{S_t}$ is a design matrix whose columns are the attribute vectors x_{ti} of the items in the assortment S_t and $\mu_i(\mathbf{X}_{S_t}^\top \theta) = P_t(i|S_t, \theta)$.

It appears that the authors may have intended to derive this equation using a first-order exact Taylor expansion. However, in MNL bandits, this equation generally does not hold. Consider a counterexample where $x_{ti} = 0$, $x_{tj} \neq 0$ for $j \neq i$, and $\theta^\star \neq \theta_t$. Then, the left-hand side of (L.1) is equals to 0 (since $x_{ti} = 0$), but the right-hand side of (L.1) is not 0, because the denominators of each $\mu_i(\mathbf{X}_{S_t}^\top \theta^\star)$ and $\mu_i(\mathbf{X}_{S_t}^\top \theta_t)$ differ. This equation only holds in special cases, such as when $K = 1$, which corresponds to the logistic bandit case. Equation (16) in Agrawal et al. [5] serves as the foundation for the entire proof in their paper. Consequently, all subsequent results derived from it are also incorrect.

2. Cauchy-Schwarz inequality in the regret analysis (Page 46)
When using the Cauchy-Schwarz inequality on the regret before applying the elliptical potential lemma, they indeed incur an additional $\sqrt{K}$ factor:

$$\sum_{t=1}^{T} \min\left(\sum_{i\in S_t} \|x_{ti}\|_{\mathbf{J}_t^{-1}}, 1\right) \leqslant \sqrt{KT}\sqrt{\min\left(\sum_{i\in S_t} \|x_{ti}\|^2_{\mathbf{J}_t^{-1}}, 1\right)}.$$

Hence, their regret should actually be $\tilde{\mathcal{O}}(d\sqrt{KT} + d^2/\kappa)$, which is worse than the result in our Theorem 2 (upper bound under uniform rewards) by a factor of K.

M Limitations

To determine the optimistic assortment in line 5 of Algorithm 1, we must calculate the optimistic utility α_{ti} for each item $i \in [N]$. Consequently, the computational cost for each round scales polynomially with N, the number of items. If N is very large, or infinite, our proposed algorithm becomes intractable. It is important to note that much of the existing literature on contextual bandits [2, 10, 24, 1, 20, 6–8, 16, 40, 41, 4, 44] also requires enumerating all items to select the item for each round.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Abstract and Section 1

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section M in Appendix

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: Assumption 1 and 2

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: Secion 7 and Section K

   Guidelines:

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

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

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

Answer: [No]

Justification: We have included the code in the supplementary material. After our paper is accepted, we will provide open access to the code.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: Section 7 and Section K

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: Figure 1 and K.2. We omitted the error bars in Figure K.1 because they are too minimal to be significant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Section K

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification:

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: This paper focuses on theoretical results and therefore does not discuss societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: No risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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