Cascading Contextual Assortment Bandits

Hyun-jun Choi
Seoul National University
Seoul, South Korea
nschj1@snu.ac.kr

Rajan Udwani
UC Berkeley
Berkeley, CA, USA
rudwani@berkeley.edu

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

Abstract

We present a new combinatorial bandit model, the *cascading contextual assortment bandit*. This model serves as a generalization of both existing cascading bandits and assortment bandits, broadening their applicability in practice. For this model, we propose our first UCB bandit algorithm, `UCB-CCA`. We prove that this algorithm achieves a T-step regret upper-bound of $\tilde{\mathcal{O}}(\frac{1}{\kappa} d\sqrt{T})$, sharper than existing bounds for cascading contextual bandits by eliminating dependence on cascade length K. To improve the dependence on problem-dependent constant κ, we introduce our second algorithm, `UCB-CCA+`, which leverages a new Bernstein-type concentration result. This algorithm achieves $\tilde{\mathcal{O}}(d\sqrt{T})$ without dependence on κ in the leading term. We substantiate our theoretical claims with numerical experiments, demonstrating the practical efficacy of our proposed methods.

1 Introduction

Sequential interactions between users and a recommender agent are often modeled as the multi-armed bandit problem or one of its variants [15]. In practice, a user typically encounters multiple items per round of interaction rather than a solitary item. Two popular models that capture this aspect are the cascading bandit [13; 14; 18] and the assortment bandit [4; 5; 7; 22], also often known as multinomial logistic bandits.

In the cascading bandit problem [13; 14; 18; 29; 25; 28], the agent selects a cascade of K items from a total of N items each round. These selected items are sequentially presented one at a time to a user. If the user clicks on a presented item, the cascading round ends. If not, the agent proceeds to reveal the next item from the cascade. This process continues until either the user clicks on an item or all K items in the cascade have been presented without a click. Once a round ends, a next round commences with a newly selected list of K items.

In the assortment bandit problem [4; 5; 7; 8; 9; 7; 22; 23], the agent presents an assortment of M items all at once, then receives user choice feedback on the assortment. A user may opt for one of the M items presented or choose none at all, concluding the round in either case. Both cascading and assortment bandit problems are significant combinatorial variations of the multi-armed bandit problem and have been extensively examined both theoretically and in practice.

However, a more commonly encountered scheme in real-world applications is a generalization of these two settings, where a *cascade of assortments* is sequentially revealed in each round. This approach is evident in video streaming services, where assortments of recommended contents are revealed as users scroll through webpages or mobile applications. Similar experiences can also be found in various online retail services and search engines. To address this, we propose a new interactive model, which we term *cascading assortment bandits*. In the cascading assortment bandits, the agent chooses a cascade of assortments that consists of K assortments with each assortment containing M items. The agent reveals one assortment at a time in the cascade. The cascade concludes if the user clicks on one of the items contained in a given assortment. If not, the agent proceeds

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

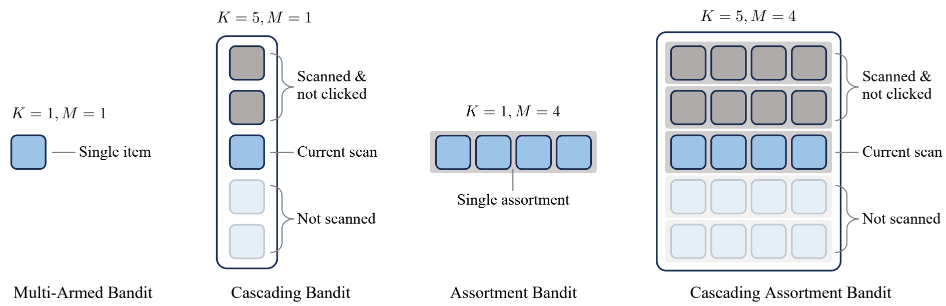

Figure 1: Comparisons between the cascading assortment bandit and the other combinatorial bandits. The cascading assortment bandit subsumes the multi-armed bandit ($K = 1, M = 1$), the cascading bandit ($K > 1, M = 1$), and the assortment bandit ($K = 1, M > 1$).

to unveil the subsequent assortment in the cascade. This cycle continues until either an assortment receives a click from the user or the agent depletes the pre-selected assortments.

The cascading assortment bandit problem is the strict generalization of both the cascading bandits model and the assortment bandits model. It is also a generalization of the simple multi-armed bandit problem. That is, if $K = 1$ and $M = 1$, the problem is the simple multi-armed bandit. If $K > 1$ and $M = 1$, the problem corresponds to the cascading bandit. If $K = 1$ and $M > 1$, we recover the assortment bandit problem. The illustrations on comparisons between the cascading assortment bandit and the other combinatorial bandits are presented in Figure 1.

In order to accommodate the generalization of the interactive model across items and assortments, we also incorporate the feature information of items and parametrization of a click model in the cascading assortment bandit model. Hence, we name the model as **cascading contextual assortment bandit** (see Section 2.2 for the formal definition of the problem setting).[1] Under this newly proposed combinatorial bandit model, we posit the following question:

Can we design a provably efficient algorithm for cascading contextual assortment bandits?

To address the question at hand, we first have to overcome the technical challenges inherent in each special case of our problem setting: the cascading contextual bandit and the contextual assortment bandit. Firstly, in the cascading contextual bandit [18; 25], a longstanding issue has been the suboptimal dependence on the cascade length, K. Intuitively, one would expect that as K increases in the cascading model, the regret should either diminish or at least remain constant; performance deterioration should not occur. However, all existing regret bounds for cascading contextual bandits scale proportionally to K [18; 29; 17; 28; 25]. This finding is not just counter-intuitive, but also suboptimal (for further discussions, refer to Section2.4). As a result, (i) **eradicating the suboptimal dependence on cascade length** K has been recognized as an open problem, even within the cascading contextual bandit setting.[2]

Further, in the context of assortment bandits, there is a widely recognized suboptimal dependence on the problem-specific constant κ, as demonstrated in the existing assortment bandit literature [9; 22; 23]. This problem-specific constant κ (in Assumption 4.2) represents the curvature of the multinomial logit (MNL) function. Recent studies [24; 3] have demonstrated an improved dependence on κ, albeit only multiplied by logarithmic factors. However, this improvement comes at the expense of an increased dependence on the assortment size M, a conclusion that is both counter-intuitive and suboptimal. Thus, (ii) **decreasing the κ dependence without escalating the dependence on M** still poses an unresolved issue. While addressing either of the two challenges (i) and (ii) can be daunting individually, tackling both issues simultaneously poses an even greater challenge in both our algorithm design and regret analysis.

[1]Note that a non-contextual version of the cascading assortment bandit is a special case of the cascading contextual assortment bandit with a one-hot encoded feature vector for each item. Hence, when we aim to provide efficient algorithms for cascading contextual assortment bandit, we also address the non-contextual cascading assortment bandit which has not been studied previously.

[2]A concurrent work [20] addresses this suboptimal dependence on K under the linear model assumption. Our work tackles this challenge under the MNL model in a more general setting.

Table 1: Comparisons of algorithms for contextual cascade and assortment bandits as well as for cascading contextual assortment bandits. N is the number of ground items, K is a length of cascade, d is a dimension of feature vectors and T is total rounds. κ is a problem-dependent parameter for the MNL model. See Appendix A for more discussions.

Algorithm	Context	Cascade	Assortment	Click Model	Regret Bound
`CombCascade` [14]	×	◯	×	×	$\tilde{\mathcal{O}}(\sqrt{KNT})$
`C`^{3}`-UCB` [18]	◯	◯	×	Linear	$\tilde{\mathcal{O}}(d\sqrt{KT})$
`EE-MNL` [5]	×	×	◯	MNL	$\tilde{\mathcal{O}}(\sqrt{NT})$
`TS-MNL` [22]	◯	×	◯	MNL	$\tilde{\mathcal{O}}(\frac{1}{\kappa}d^{3/2}\sqrt{T})$
`UCB-MNL` [23]	◯	×	◯	MNL	$\tilde{\mathcal{O}}(\frac{1}{\kappa}d\sqrt{T})$
`LinTS-Cascade` [28]	◯	◯	×	Linear	$\tilde{\mathcal{O}}(d^{3/2}K\sqrt{T})$
`CascadeWOFUL` [25]	◯	◯	×	Linear	$\tilde{\mathcal{O}}(\sqrt{d^2T + dTK})$
`VAC`^{2}`-UCB` [20]	◯	◯	×	Linear	$\tilde{\mathcal{O}}(d\sqrt{T})$
`UCB-CCA` (Algorithm 1)	◯	◯	◯	MNL	$\tilde{\mathcal{O}}(\frac{1}{\kappa}d\sqrt{T})$
`UCB-CCA+` (Algorithm 2)	◯	◯	◯	MNL	$\tilde{\mathcal{O}}(d\sqrt{T})$

To this end, we design novel upper confidence bound (UCB) algorithms for contextual cascading assortment bandits, tackling both technical challenges. We show that our proposed algorithms achieve provable guarantees on regret performances overcoming the longstanding technical challenges. Our regret bounds show sharper results than those of the existing contextual cascading bandits or assortment bandits. We corroborate our theoretical claims through numerical experiments, thus ensuring that both our newly proposed bandit framework and the proposed algorithms establish provable efficiency and practical applicability.

Our main contributions are summarized as follows.

- We formulate a general combinatorial bandit model, named *cascading contextual assortment bandit* that encompasses the existing cascading bandits and assortment bandits. This novel problem setting is observed in many practical applications.
- We first propose a UCB bandit algorithm `UCB-CCA` for the cascading contextual assortment bandit and establish the T-step regret upper-bound of $\tilde{\mathcal{O}}(\frac{1}{\kappa}d\sqrt{T})$ (in Theorem 4.3). This regret bound is tighter than the existing bounds for cascading contextual bandits, where we not only remove the longstanding, unnecessary dependence on K but also establish the result without dependence on M.
- While `UCB-CCA` is an efficient algorithm achieving both the statistical efficiency and practical performances (shown in Section 7), its regret bound includes dependence on the inverse of a problem-dependent constant κ, which can be potentially large in the worst case. To improve the dependence on κ, we propose our second algorithm `UCB-CCA+`, which exploits a new Bernstein-type concentration result, taking into account the effects of the local curvature of the MNL model. We prove that `UCB-CCA+` achieves $\tilde{\mathcal{O}}(d\sqrt{T})$ without the dependence on κ in the leading term (only scaling with logarithmic factors), hence significantly improving the regret of `UCB-CCA` without increasing the other dependencies. Hence, we successfully solve the two technical challenges (i) and (ii) mentioned above.
- As an independent contribution, we prove that a greedy algorithm for the cascading assortment optimization problem gives a 0.5 approximation of the optimal solution (discussed in Section 6). To our best knowledge, this is the first rigorous result showing the approximation guarantee even for the contextual cascading bandit problem, instead of simply assuming access to an approximation optimization oracle.
- We evaluate our proposed methods in numerical experiments and show that the practical performances support our theoretical claims. Hence, our proposed algorithms along with our newly proposed bandit model establish provable efficiency and practical applicability.

2 Preliminaries

2.1 Notation

Define $[n]$ as a set of positive integers from 1 to n. Let $|\cdot|$ be the length of a sequence or the cardinality of a set. For a vector $x \in \mathbb{R}^d$, we denote the ℓ_2-norm of x as $||x||_2$ and the V-weighted norm of x for a positive-definite matrix V as $||x||_V = \sqrt{x^\top V x}$. The determinant and trace of a matrix V are $\det(V)$ and trace(V), respectively. $\lambda_{\min}(V)$ denotes the minimum eigenvalue of a matrix V.

2.2 Cascading Contextual Assortment Bandit Problem

Consider $[N]$, a set of N items. Let $\mathcal{A}$ be a set of candidate assortments of items with size M, i.e., $\mathcal{A} := \{A \subseteq [N] : |A| = M\}$. A cascade S is an ordered sequence of K assortments chosen from $\mathcal{A}$ where all the items in these K assortments are distinct. Then, the set of all feasible cascades $\mathcal{S}$ can be defined as follows.

$$\mathcal{S} := \big\{S = \big(A_1, ..., A_K\big) \mid A_k \in \mathcal{A} \text{ for all } k \in [K], \cap_{k=1}^{K} A_k = \emptyset\big\}$$

At round t, feature vectors $\{x_{ti} \in \mathbb{R}^d, i \in [N]\}$ for every item are revealed to the decision-making agent. Each feature vector x_{ti} may contain the contextual information of the user at round t and the item i. After observing this contextual information, at round t, the agent recommends a cascade $S_t = (A_{tk})_{k\in[K]}$ to the user, where $A_{tk} \in \mathcal{A}$ represents the k-th assortment of the cascade at round t. The user scans the assortments in S_t one by one. If the items in A_{tk} do not attract the user, the user moves on to the next assortment $A_{t,k+1}$. The user stops at the O_t-th assortment when the user is attracted by an item in the O_t-th assortment and clicks on the item.

After the user clicks on the item, the agent observes a sequence of user choices $y_t = (y_{tk})_{k\in[O_t]}$ where a binary vector $y_{tk} = (y_{tk0}, y_{tk1}, ..., y_{tkM})$ represents user choices on assortment A_{tk}. Let $y_{tkm} = 1$ if the m-th item i_m in A_{tk} is clicked by the user, and $y_{tkm} = 0$ for items that are not clicked on. For each assortment, there is an *outside option*. That is, there is a probability that the user may not click any of the items in A_{tk}. If the user does not choose any items, $y_{tk0} = 1$ and $y_{tkm} = 0$ for all $m \in [M]$. The user choice for each assortment is given by the multinomial logit (MNL) choice model [21]. For this MNL model, there is an *unknown* time-invariant parameter $\theta^* \in \mathbb{R}^d$. We define the true weight of item i in round t as $w^*_{ti} := x_{ti}^\top \theta^*$. Also, we let the vector representation of the weights be defined as $w^*_t := (w^*_{ti})_{i\in[N]}$ for convenience.

Under this model, the user's click probability of the m-th item in A_{tk} and the probability of the outside option in A_{tk} is given respectively by

$$p_t(i_m|A_{tk}, w^*_t) = \frac{\exp(w^*_{ti_m})}{1 + \sum_{j\in A_{tk}} \exp(w^*_{tj})} \quad \text{and} \quad p_t(i_0|A_{tk}, w^*_t) = \frac{1}{1 + \sum_{j\in A_{tk}} \exp(w^*_{tj})}$$

where item i_0 represents the outside option. The user choice y_{tk} is sampled from the multinomial distribution, $y_{tk} \sim$ MNL$\big\{1, \big(p_t(i_m|A_{tk}, w^*_t)\big)_{m=0}^{M}\big\}$, where the argument 1 indicates that y_{tk} is a single-trial sample. Hence, $\sum_{m=1}^{M} y_{tkm}$ is always 1. Also, we denote measurement noise as $\epsilon_{tkm} := y_{tkm} - p_t(i_m|A_{tk}, w^*_t)$. Since ϵ_{tkm} is bounded in $[0, 1]$, ϵ_{tkm} is σ^2-sub-Gaussian with $\sigma^2 = 1/4$. It is important to note that ϵ_{tkm} across items in the same assortment is not independent due to the substitution effect in the MNL model.

The expected reward function of a combinatorial action S_t based on w^*_t is given by

$$f(S_t, w^*_t) = \sum_{k=1}^{K} \left\{ \prod_{\dot{k}=1}^{k-1} p_t(i_0|A_{t\dot{k}}, w^*_t) \right\} \sum_{i\in A_{tk}} p_t(i|A_{tk}, w^*_t) = 1 - \prod_{k=1}^{|S_t|} p_t(i_0|A_{tk}, w^*_t).$$

The formulation above is also known as the cascade model with disjunctive objective, where the user stops at the *first attractive* item [13; 14; 18].

2.3 α-Approximation Oracle and α-Regret

The exact combinatorial optimization to compute an optimal cascade of assortments can be computationally expensive. Therefore, we allow for approximate optimization. We assume that the agent has access to an α-optimization oracle to compute a α-approximation solution of the cascade optimization problem with $\alpha \leq 1$. For approximate optimization, we prove that a greedy selection for the cascading assortment optimization problem gives a 0.5 approximation of the optimal solution, which may be of independent interest (see Section 6).

Formally, for a given an α-optimization oracle and a weight parameter w, the oracle outputs an approximately optimal cascade $\hat{S}^* = \mathbb{O}^\alpha(w) \in \mathcal{S}$ satisfying $f(\hat{S}^*, w) \geq \alpha f(S^*, w)$ where $S^* \in \arg\max_{S \in \mathcal{S}} f(S, w)$ is an optimal assortment without approximation. The instantaneous α-regret of cascade S_t in round t is defined as $\mathcal{R}^\alpha(t, S_t) := \mathbb{E}[\alpha f(S_t^*, w_t^*) - f(S_t, w_t^*)]$ where $S_t^* \in \arg\max_{S \in \mathcal{S}} f(S, w_t^*)$ is an true optimal assortment. Then, the goal of the agent is to minimize the cumulative α-regret defined as

$$\mathcal{R}^\alpha(T) := \sum_{t=1}^{T} \mathcal{R}^\alpha(t, S_t) = \sum_{t=1}^{T} \mathbb{E}\big[\alpha f(S_t^*, w_t^*) - f(S_t, w_t^*)\big].$$

2.4 Suboptimal Dependence on Problem Dependent Parameters

In this subsection, we discuss the main technical challenges faced in the regret analysis of our problem setting. In particular, the suboptimal dependence on cascade length K has been a long-standing open problem even in contextual cascading bandits.

2.4.1 Dependence on Length of Cascade K

The previous literature on the contextual cascading bandits [18; 17; 26; 25] bounds the instantaneous regret in each round, utilizing the monotonicity and Lipschitz continuity of the expected reward function f. A simple adaptation of the previously known techniques to our problem would result in the following upper bound for the instantaneous regret $\mathcal{R}^\alpha(t, S_t)$ for $S_t = (A_{t1}, A_{t2}, ..., A_{tK})$.

$$\mathcal{R}^\alpha(t, S_t) \leq \sum_{k=1}^{K} \sum_{i \in A_{tk}} \beta_t ||x_{ti}||_{V_{t-1}^{-1}} \tag{1}$$

where β_t is a suitable confidence radius chosen by an algorithm, and V_t is a positive definite gram matrix. Then, the dependence on the length of cascade K and the assortment size M would appear in the regret bound after summing the right-hand side of Eq.(1) over the time horizon and applying the Cauchy-Schwarz inequality. Because of this reason, even when $M = 1$, there still exists dependence on K which appears in the regret bounds of all previous contextual cascading bandits (see Table 1).

To overcome this challenge, we present a new Lipschitz continuity of the expected reward function to derive the regret bound independent of M and K by replacing the summation with the maximum over assortments and a cascade (see Section 4.3 for more details).

2.4.2 Dependence on Worst-Case Scanning Probability

Analogous to the existing algorithms for the contextual cascading bandits [18; 26], the gram matrix V_t contains the rank-1 matrices of observed items accumulated up to round t, i.e., $V_t = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} x_{\tau i} x_{\tau i}^\top + \lambda I$. However, there exists an out-of-control issue, that is, the summation of the rank-1 matrices over $O_t + 1$ to $|S_t|$ in Eq.(1) is not included in the gram matrix V_t. Note that this issue also arises in cascading contextual assortment bandits. Adapting a technique used in the existing literature [18] to mitigates this issue, let p_{t,S_t} be the probability of examining all assortments in S_t and $p^* = \min_{t \in [T]} \min_{S \in \mathcal{S}} p_{t,S_t}$ be the worst-case probability of examining a cascade over all rounds and all feasible cascades. A simple adaptation of the existing methods would result in the following bound for the expected instantaneous regret.

$$\mathbb{E}\left[\mathcal{R}^\alpha(t, S_t)\right] = \mathbb{E}\left[\mathcal{R}^\alpha(t, S_t)\mathbb{E}\left[\frac{1}{p_{t,S_t}} \mathbb{1}\left\{O_t = |S_t|\right\} \mid S_t\right]\right] \leq \frac{1}{p^*} \mathbb{E}\big[\mathcal{R}^\alpha(t, S_t)\mathbb{1}\{O_t = |S_t|\}\big].$$

This concedes the dependence on p^*, which can be exponentially small in the worst case. We overcome this challenge by designing an algorithm that offers the assortment containing the most uncertain item as the first assortment in a cascade. We discuss this salient feature of the proposed algorithm in more detail in Section 3.2.

Algorithm 1 `UCB-CCA`

Input: confidence radius β_t and ridge penalty parameter $\lambda \geq 1$

for $t = 1, \dots, T$ **do**
  Observe x_{ti} for all $i \in [N]$
  Compute $u_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \beta_{t-1} ||x_{ti}||_{V_{t-1}^{-1}}$ for all $i \in [N]$
  Compute a candidate cascade $S'_t \leftarrow (A'_{tk})_{k \in [K]} = \mathbb{O}^\alpha(u_t)$
  Find optimistic exposure assortment index k^* in $(k^*, i^*) = \underset{k \in [K], i \in A'_{tk}}{\text{argmax}} ||x_{ti}||_{V_{t-1}^{-1}}$
  Optimistic exposure swap $S_t \leftarrow (A_{tk})_{k \in [K]}$ where $A_{tk} := \begin{cases} A'_{tk^*} & \text{if } k = 1 \\ A'_{t1} & \text{if } k = k^* \\ A'_{tk} & \text{otherwise} \end{cases}$
  Offer S_t, and observe user feedback O_t and $y_t = (y_{tk})_{k \in [O_t]}$
  Update $V_t \leftarrow V_{t-1} + \sum_{k=1}^{O_t} \sum_{i \in A_{tk}} x_{ti} x_{ti}^\top$
  Compute the regularized MLE $\hat{\theta}_t$ by solving $\nabla_\theta \left[\ell_t(\theta) + \frac{\lambda}{2} ||\theta||_2^2 \right] = \mathbf{0}$
end for

3 Algorithm: `UCB-CCA`

3.1 Upper Confidence Bounds and Confidence Set

`UCB-CCA` utilizes the upper confidence bounds (UCB) technique [6; 1; 16] to compute an optimistic action based on optimistic estimates of each item's weight, $u_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta) ||x_{ti}||_{V_{t-1}^{-1}}$ for all $i \in [N]$. The confidence radius $\beta_t(\delta)$ is specified to maintain a high-probability confidence set $C_t(\delta)$ for the unknown parameter θ^*, although the algorithm does not explicitly compute $C_t(\delta)$.

$$C_t(\delta) := \left\{ \theta \in \mathbb{R}^d : ||\hat{\theta}_t - \theta||_{V_t} \leq \beta_t(\delta) \right\}.$$

Setting a proper confidence radius $\beta_t(\delta)$ can guarantee that θ^* lies within the confidence set with probability $1 - \delta$. On the event that $\theta^* \in C_t(\delta)$, the UCB weight u_{ti} serves as an upper bound of a true weight $w_{ti}^* := x_{ti}^\top \theta^*$ for every item $i \in [N]$. We denote the UCB weight vector as $u_t = (u_{ti})_{i \in [N]}$ for convenience.

3.2 Optimistic Exposure Swapping

A distinctive element of `UCB-CCA` is what we call *optimistic exposure swapping*, a procedure crucial for eliminating dependence on the worst-case scanning probability, as elaborated in Section 2.4.2. This technique strategically positions the assortment containing the item with the highest uncertainty among the top MK items in the first slot of the cascade of assortments.

In each round t, the α-approximate oracle $\mathbb{O}^\alpha(u_t)$ outputs a *candidate* cascade S'_t, determined by the UCB weights u_t. It is important to note that S'_t is not immediately presented to the user. Instead, after S'_t is derived using the optimization oracle $\mathbb{O}^\alpha(u_t)$, the algorithm identifies the index k^* of an assortment that includes the item with the largest $||x_{ti}||_{V_{t-1}^{-1}}$ in S'_t.

Subsequently, the algorithm swaps the positions: the assortment A'_{tk^*} is moved to the top of S_t, becoming A_{t1}, and the initially top assortment A'_{t1} in S'_t is relocated to the k^*-th position of S_t, now A_{tk^*}. The positions of the other assortments remain the same, that is, $A_{tk} = A'_{tk}$ for all $k \in [K] \setminus \{1, k^*\}$.[3] This procedure is viable as the expected reward is unaffected by the display order of assortments in the cascade, as shown in Lemma 4.5.

3.3 Regularized Maximum Likelihood Estimation

`UCB-CCA` computes a regularized maximum likelihood estimate of the unknown parameter θ^*. The negative log-likelihood is given by $\ell_t(\theta) = -\sum_{\tau=1}^{t-1} \sum_{k=1}^{O_\tau} \sum_{m=0}^{M} y_{\tau km} \log p_\tau(i_m | A_{\tau k}, w_\tau)$, where

[3] While the swapping occurs between the first and the k^*-th positions for specificity, it is sufficient to place A'_{tk^*} at the top position of S_t. The sequence of the remaining assortments is not critical.

$w_t = (w_{ti})_{i\in[N]}$ is a weight vector, and its element is $w_{ti} = x_{ti}^\top\theta$. For penalty parameter $\lambda \geq 1$, the ℓ_2-regularized MLE is given by

$$\hat{\theta}_t = \underset{\theta}{\text{argmin}}\left[\ell_t(\theta) + \frac{\lambda}{2}||\theta||_2^2\right] = \underset{\theta}{\text{argmin}}\left[-\sum_{\tau=1}^{t-1}\sum_{k=1}^{O_\tau}\sum_{m=0}^{M} y_{\tau km}\log p_\tau(i_m|A_{\tau k}, w_\tau) + \frac{\lambda}{2}||\theta||_2^2\right]. \tag{2}$$

4 Regret Analysis of UCB-CCA

4.1 Regularity Condition

Assumption 4.1. $||x||_{ti} \leq 1$ for all round t and items $i \in [N]$, and also $||\theta^*|| \leq 1$.

Assumption 4.2. There exists $\kappa > 0$ such that for all $t \in [T]$, any assortment $A \in \mathcal{A}$, and any item $i \in A$, $\inf_{\theta\in\mathbb{R}^d} p_t(i|A, w)p_t(i_0|A, w) \geq \kappa$, where $w = (w_i)_{i\in[N]}$ and $w_i = x_i^\top\theta$.

Discussion of Assumptions. Assumption 4.1 makes the regret bound independent on the scale of the feature vector and parameter. This is the standard assumption used in the contextual bandit literature [1; 18; 22]. Assumption 4.2 is the standard regularity assumption in the contextual assortment bandit literature [8; 27; 22; 7; 23], adapted from the standard assumption for the link function in the generalized linear contextual bandit literature [16] to ensure that the Fisher information matrix is non-singular.

4.2 Regret Bound of UCB-CCA

Theorem 4.3 (α-regret upper bound of UCB-CCA)**.** *Suppose Assumptions 4.1 and 4.2 hold, and we run UCB-CCA for total T rounds with $\beta_t = \frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{tKM}{d\lambda}\right)+4\log t} + \frac{\sqrt{\lambda}}{\kappa}$ and with $\lambda \geq 1$, Then, the α-regret of UCB-CCA is upper-bounded by*

$$\mathcal{R}^\alpha(T) \leq \left(\frac{K}{K+1}\right)^{K+1}\left[\frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{TKM}{d\lambda}\right)+4\log T} + \frac{\sqrt{\lambda}}{\kappa}\right]\sqrt{2dT\ln\left(1+\frac{TKM}{\lambda d}\right)}.$$

Discussion of Theorem 4.3. Theorem 4.3 establishes that UCB-CCA achieves a regret bound of $\tilde{\mathcal{O}}\big(\frac{d}{\kappa}\sqrt{T}\big)$. Notably, this regret bound removes dependence on p^* completely and removes polynomial dependence on K, achieving the best-known bound in contextual cascading bandits [18; 25; 20], a special case of our problem setting. Apart from the generalization we consider in this work, a key distinction between our work and the previous contextual cascading bandit models lies in our adoption of the MNL model, as opposed to the linear model assumed by the existing literature. This model choice introduces a dependence on the parameter κ within the regret bound of UCB-CCA, which is an aspect we address and refine in subsequent sections. It is essential to highlight that our work tackles a more general problem yet achieves improved bounds concerning the key problem parameters previously considered suboptimal. The factor comprised of the length of the cascade, $(K/(K+1))^{K+1}$, in Theorem 4.3 is also notable, which is bounded above by 1 regardless of the value of K. Consequently, the regret bound does not increase polynomially with K, ensuring scalability.

4.3 Proof Outline

In this subsection, we present the proof sketch of Theorem 4.3. One of the key components of the regret analysis that enables carving off the dependence on K is the following lemma.

Lemma 4.4 (Maximal Lipschitz continuity)**.** *Suppose $u_{ti} \geq w_{ti}^*$ for all $i \in [N]$. Then*

$$f(S_t, u_t) - f(S_t, w_t^*) \leq \left(\frac{K}{K+1}\right)^{K+1}\max_{A_{tk}\in S_t}\max_{i\in A_{tk}}(u_{ti} - w_{ti}^*). \tag{3}$$

Lemma 4.4 demonstrates that the difference between the reward functions under the UCB parameter and the true parameter is upper bounded by the maximal difference between the UCB and true parameters. This implies that the regret bound remains unaffected by increases in the cascade length K or the assortment size M.

Eliminating the dependence on p^* is another key element of our analysis. To this end, we first show that the order of assortments in the cascade model with the disjunctive objective does not affect the expected reward. We formalize this property in the following lemma.

Lemma 4.5. *Let p_k be the probability that the user clicks on any item in A_k. Given a collection of assortments $\{A_1, \cdots, A_K\}$ with probabilities $\{p_1, \cdots, p_K\}$, their order of display does not matter. Further, for every permutation $\rho : [K] \to [K]$, we have*

$$\sum_{k\in[K]} p_k \prod_{\dot{k}<k} (1 - p_{\dot{k}}) = 1 - \prod_{k\in[K]} (1 - p_k) = \sum_{k\in[K]} p_{\rho^{-1}(k)} \prod_{\dot{k}<k} \left(1 - p_{\rho^{-1}(\dot{k})}\right).$$

Consolidating these key results, we proceed to bound the cumulative regret. We begin by leveraging the monotonicity of the expected reward function and the definition of the α-approximate optimization oracle to bound the cumulative regret.

$$\begin{aligned}\mathcal{R}^\alpha(T) &\le \mathbb{E}\left[\sum_{t=1}^T f(S_t, u_t) - f(S_t, w_t^*)\right] \le C_K \mathbb{E}\left[\sum_{t=1}^T \max_{A_{tk}\in S_t} \max_{i\in A_{tk}} (u_{ti} - w_{ti}^*)\right] \\ &\le 2C_K \mathbb{E}\left[\beta_T \sum_{t=1}^T \max_{A_{tk}\in S_t} \max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}\right] = 2C_K \mathbb{E}\left[\beta_T \sum_{t=1}^T \max_{k\in[O_t]} \max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}\right]. \end{aligned} \quad (4)$$

The second inequality is from Lemma 4.4, letting $C_K := (K/(K+1))^{K+1}$. The third inequality is given by the concentration of the UCB weights (see Lemma B.5). Note that the assortment including the item with the largest value of $||x_{ti}||_{V_t^{-1}}$ is always examined by the user since it is included in the first assortment of S_t by the optimistic exposure swapping technique as described in Section 3.2. Note that a change in the order of assortments incurred by the optimistic exposure swapping does not affect the expected reward which is shown in Lemma 4.5. Hence, for every round $t \in [T]$, we obtain

$$\max_{A_{tk}\in S_t} \max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}} = \max_{k\in[O_t]} \max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}. \quad (5)$$

Therefore, the last equality in Eq.(4) is given by Eq.(5). Then, we can apply the maximal version of elliptical potential lemma (see Lemma B.8) to bound the cumulative regret.

5 Improved Dependence on κ

While UCB-CCA achieves the regret bound of $\tilde{\mathcal{O}}\big(\frac{1}{\kappa} d\sqrt{T}\big)$ improving dependence on K, the bound includes the problem-dependent constant κ. This implies a potential risk of the regret bound becoming large when κ becomes very small. In order to circumvent this challenge, we propose a new *optimism in the face of uncertainty (OFU)* algorithm, UCB-CCA+, which exhibits a regret bound that is independent of κ in the leading term. The pseudocode of UCB-CCA+ is detailed in Algorithm 2.

5.1 Algorithm: UCB-CCA+

5.1.1 Confidence Set

UCB-CCA+ computes a regularized MLE $\hat{\theta}_t$, following the same procedure described in Section 3.3. Then, the algorithm constructs a new confidence set centered around $\hat{\theta}_t$ utilizing a Bernstein-type tail inequality for self-normalized martingales [11; 3]. Nevertheless, a simple adaptation of the previous approaches may incur increased dependence on M. Hence, a more intricate analysis and refined algorithmic strategy are imperative to effectively address this challenge.

First, we define $g_t(\theta) := \sum_{\tau=1}^t \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} p_\tau(i|A_{\tau k}, w_\tau) x_{\tau i} + \lambda_t \theta$ where $w_t = (w_{ti})_{i\in[N]}$ and $w_{ti} = x_{ti}^\top \theta$. We also denote the partial derivative of $p_t(i|A_{tk}, w_t)$ with respect to w_{ti} as $\dot{p}_t(i|A_{\tau k}, w_\tau) := p_t(i|A_{tk}, w_t) p_t(i_0|A_{tk}, w_t)$. Additionally, we define the new design matrix containing local information, denoted as $H_t(\theta) := \sum_{\tau=1}^t \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, w_\tau) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d$, and, for convenience, $H_t := \sum_{\tau=1}^t \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, w_\tau^*) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d$. Then, the algorithm constructs a confidence set as below:

$$B_t(\delta) := \left\{\theta \in \mathbb{R}^d : ||g_t(\hat{\theta}_t) - g_t(\theta)||_{H_t^{-1}(\theta)} \le \gamma_t(\delta)\right\} \quad (6)$$

Algorithm 2 UCB-CCA+

Input: confidence radius γ_t and ridge penalty parameter $\lambda \geq 1$

for $t = 1, \ldots, T$ **do**

  Observe x_{ti} for all $i \in [N]$

  Construct a confience set $B_{t-1}(\delta)$ as defined in Eq.(6)

  Compute a candidate cascade $(S'_t = (A'_{tk})_{k\in[K]}, \theta_t) = \arg\max_{S\in\mathcal{S},\theta\in B_t(\delta)} f(S, w_t)$

  Find optimistic exposure assortments $L_{t,H}$ and $L_{t,V}$ (and their positions h and v) in S'_t

  $S_t \leftarrow (A_{tk})_{k\in[K]}$ where $A_{tk} = \begin{cases} L_{t,H} & \text{if} \quad k = 1 \\ L_{t,V} & \text{if} \quad k = 2 \\ A'_{t1} & \text{if} \quad k = h \\ A'_{t2} & \text{if} \quad k = v \\ A'_{tk} & \text{otherwise} \end{cases}$

  Offer S_t and observe O_t, $y_t = (y_{tk})_{k\in[O_t]}$

  Update $H_t \leftarrow H_{t-1} + \sum_{k=1}^{O_t} \sum_{i\in A_{tk}} \dot{p}_t(i|A_{tk}, w_t^*) x_{ti} x_{ti}^\top$

  Update $V_t \leftarrow V_{t-1} + \sum_{k=1}^{O_t} \sum_{i\in A_{tk}} x_{ti} x_{ti}^\top$

  Compute the regularized MLE $\hat{\theta}_t$ by solving $\nabla_\theta \left[\ell_t(\theta) + \frac{\lambda}{2}||\theta||_2^2\right] = \mathbf{0}$

end for

where the confidence radius $\gamma_t(\delta)$ is suitably specified to ensure that the true parameter θ^* lies in the confidence set $B_t(\delta)$ with high probability. On the event of $\theta^* \in B_t(\delta)$, The following lemma bounds the weighted ℓ_2-norm of the difference between θ and θ^*.

Lemma 5.1. *Suppose $\theta^* \in B_t(\delta)$. Then, for any $\theta \in B_t(\delta)$, we have $||\theta - \theta^*||_{H_t} \leq 6\gamma_t(\delta)$.*

5.1.2 Doubly Optimistic Exposure Swapping

UCB-CCA+ still faces the challenge outlined in Section 2.4.2. The complication is exacerbated for UCB-CCA+ as the algorithm concurrently updates two gram matrices, H_t and V_t, which only contain the information of the observed items. Building upon the technique of optimistic exposure swapping detailed in Section 3.2, in each round t, UCB-CCA+ assigns the assortment $L_{t,H}$ — containing the item with the largest uncertainty with respect to H_{t-1} among the top KM items — to the first slot of cascade S_t. Similarly, the algorithm places $L_{t,V}$ — with the item that has the largest uncertainty with respect to V_{t-1} — in the second slot of S_t.

5.2 Regret Analysis of UCB-CCA+

Theorem 5.2 (Regret upper bound of UCB-CCA+)**.** *Suppose Assumptions 4.1 and 4.2 hold, and we run UCB-CCA+ for total T rounds with $\lambda \geq 1$ and $\gamma_t(\delta) := \frac{3\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{(\lambda_t + KMt/d)^{d/2}\lambda_t^{-d/2}}{\delta}\right) + \frac{2d}{\sqrt{\lambda_t}} \log 2$ with $\delta = \frac{1}{t^2}$. Then, the regret of UCB-CCA+ is upper-bounded by*

$$\mathcal{R}^\alpha(T) \leq C_1 \gamma_T(\delta) \sqrt{2dT \log\left(1 + \frac{KMT}{d\lambda}\right)} + \frac{C_2}{\kappa} \gamma_T(\delta)^2 d \log\left(1 + \frac{KMT}{d\lambda}\right)$$

where $C_1 = 36$ and $C_2 = 216(1 + Me)$.

Discussion of Theorem 5.2. Theorem 5.2 establishes the regret bound of $\tilde{\mathcal{O}}(d\sqrt{T})$. The leading term of the regret bound is independent of κ. Although the second term exhibits dependence on κ, the term only scales logarithmically in T, whose comparative effect diminishes as t increases compared to the leading term. Hence, the worst-case regret guarantee of UCB-CCA+ improves from that of UCB-CCA. The comprehensive proof of Theorem 5.2 is provided in the appendix.

6 Approximation Algorithm for Optimal Combinatorial Action

In this section, we show that computing the optimal cascade is, in general, a weakly NP-hard problem and give a (fast) polynomial time algorithm that the agent can use to compute a 0.5-approximation to the optimal cascade for any weight w. We assume that the MNL weights are given and consider the problem of finding the cascade that maximizes the expected reward. This is an optimization problem on selecting a sequence of K assortments with size M each from a ground set of N items. The number of feasible cascades is $O\left(\binom{N}{M}^K\right)$. In fact, we establish the following hardness result.

Lemma 6.1. *For general M, the optimization problem is weakly NP-hard even for $K = 2$.*

The hardness of our problem follows from the hardness result of [19] for a related setting of unconstrained cascade optimization where the size of each assortment can be arbitrary. In light of this hardness, we turn our attention to finding fast approximation algorithms for the problem. While there has been a lot of recent work on the unconstrained cascade optimization problem (see [19; 12; 10] and the references therein), none of the previous algorithms apply to our constrained setting (where the size of each assortment is M).

We consider the following greedy approach for the problem. Consider arbitrary weights at round t, i.e. $w_t = (w_{ti})_{i\in[N]}$, is given. We order the items in $[N]$ in decreasing order of given weights and consider the following cascade of assortments.

$$D_1 = \{1, 2, ..., M\}, D_2 = \{M + 1, M + 2, ..., 2M\}, \cdots, D_K = \{(K - 1)M + 1, \cdots, KM\}$$

We call these assortments "decreasing order assortments". Let OPT denote the value of the optimal solution to the problem. We establish that showing these assortments in any possible sequences gives a good approximation to the problem.

Lemma 6.2. *Let* OPT *denote the overall click probability in the optimal solution. The decreasing order assortments $D_1, ..., D_K$, shown in any order, have overall click probability at least* $0.5\,\mathrm{OPT}$.

We present the proof in Appendix E. We also show that our algorithm is optimal for $M = 1$, which captures the classic cascade optimization problem.

7 Numerical Experiments

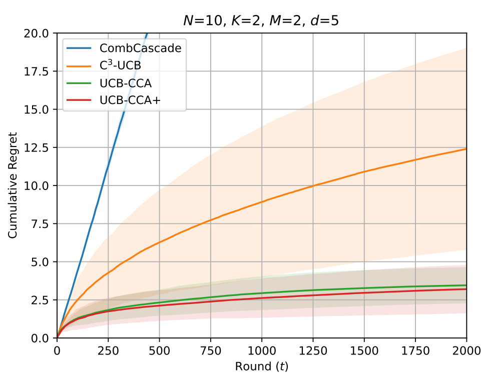

Figure 2: $N = 10$, $K = 2$, $M = 2$, and $d = 5$.

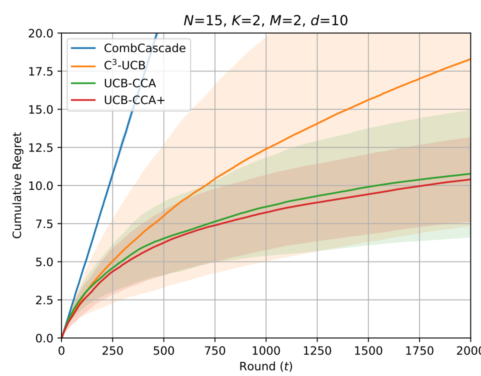

Figure 3: $N = 15$, $K = 2$, $M = 2$, and $d = 10$.

In this section, we evaluate the performances of our proposed algorithms `UCB-CCA` and `UCB-CCA+` in numerical experiments and compare their performances with the existing combinatorial bandit algorithms `CombCascade` and C^3-`UCB`. For simulations, we generate a random sample of the unknown time-invariant parameter θ^* from $\mathcal{N}(0, 1)$ at the beginning of the simulation. We sample N feature vectors from $\mathcal{N}(0, 1)$ in each round t. At each round t, the oracle computes a sequence of assortments in decreasing order, forming a cascade S_t based on given w_t. We assess the cumulative regret of `UCB-CCA`, `UCB-CCA+`, `CombCascade`, and C^3-`UCB`. Note that a user's choice for `UCB-CCA` and `UCB-CCA+` is determined by the MNL logit choice model, whereas C^3-`UCB` utilizes a linear model and `CombCascade` is a non-contextual model. Figure 2 and Figure 3 indicate that both `UCB-CCA` and `UCB-CCA+` significantly outperform C^3-`UCB` and `CombCascade`. We also observe that `UCB-CCA+` shows a slight performance advantage over `UCB-CCA`, although the difference is not statistically significant. While `UCB-CCA+` has a sharper worst-case regret guarantee, `UCB-CCA` can provide favorable practical performances, with simpler implementation and computational efficiency.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00222663, No. 2022R1C1C1006859, No. 2022R1A4A103057912, No. 2021M3E5D2A01024795) and by Creative-Pioneering Researchers Program through Seoul National University and by Naver.

References

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

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

[3] Priyank Agrawal, Theja Tulabandhula, and Vashist Avadhanula. A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, 2023.

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

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

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

[7] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *Journal of machine learning research*, 2020.

[8] Wang Chi Cheung and David Simchi-Levi. Assortment optimization under unknown multinomial logit choice models. *arXiv preprint arXiv:1704.00108*, 2017.

[9] 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.

[10] Elaheh Fata, Will Ma, and David Simchi-Levi. Multi-stage and multi-customer assortment optimization with inventory constraints. *arXiv preprint arXiv:1908.09808*, 2019.

[11] 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.

[12] Pin Gao, Yuhang Ma, Ningyuan Chen, Guillermo Gallego, Anran Li, Paat Rusmevichientong, and Huseyin Topaloglu. Assortment optimization and pricing under the multinomial logit model with impatient customers: Sequential recommendation and selection. *Operations research*, 69(5):1509–1532, 2021.

[13] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015.

[14] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

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

[16] 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.

[17] Shuai Li and Shengyu Zhang. Online clustering of contextual cascading bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[18] Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pages 1245–1253. PMLR, 2016.

[19] Nan Liu, Yuhang Ma, and Huseyin Topaloglu. Assortment optimization under the multinomial logit model with sequential offerings. *INFORMS Journal on Computing*, 32(3):835–853, 2020.

[20] 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.

[21] Daniel McFadden. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.

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

[23] 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.

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

[25] Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, and R Srikant. Minimax regret for cascading bandits. *arXiv preprint arXiv:2203.12577*, 2022.

[26] Kun Wang. Conservative contextual combinatorial cascading bandit. *IEEE Access*, 9:151434–151443, 2021.

[27] Xue Wang, Mike Mingcheng Wei, and Tao Yao. Online assortment optimization with high-dimensional data. *Available at SSRN 3521843*, 2019.

[28] Zixin Zhong, Wang Chi Chueng, and Vincent YF Tan. Thompson sampling algorithms for cascading bandits. *The Journal of Machine Learning Research*, 22(1):9915–9980, 2021.

[29] 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 32nd Conference on Uncertainty in Artificial Intelligence*, 2016.

A More Discussions on the Related Work

Table 1 summarizes comparisons between our work and the previous works that propose algorithms in various combinatorial bandit settings: cascading bandits and assortment bandits.

For cascading bandits, there are two major objectives. One is called a *disjunctive* objective where the agent receives a positive reward when at least one item in the recommended sequence of items K is attractive. The other one is a *conjunctive* objective where the agent receives a positive reward when all the items are attractive. Kveton et al. [13] first introduced the multi-armed cascading bandits with disjunctive objective, and Kveton et al. [14] proposed the cascading bandits with conjunctive objective. There is another difference between previous studies with these two objectives, which is the definition of the feasible set. The feasible set is an arbitrary subset of ground items in Kveton et al. [14], whereas it is a uniform matroid in Kveton et al. [13]. These previous studies propose UCB-type algorithms and derive both gap-dependent and gap-independent regret upper bounds.

Li et al. [18] generalize the above two models with contextual information. In this model, each item has its own weight that represents its attractiveness to the user. The weight of an item is assumed to have a linear relation with the feature of the item. The agent learns the unknown parameter from the feedback while maximizing the reward function for cascades. Li et al. [18] propose a UCB-type algorithm, referred as `C`^{3}`-UCB`, and prove that the T-step regret of `C`^{3}`-UCB` is upper bounded by $\tilde{\mathcal{O}}(d\sqrt{TK})$ where d is the dimension of the contextual information vector, K is the length of cascade, and p^* is the minimum probability that the user examines all the items in the offered cascade for any cascade and any round. The regret bound studied in the aforementioned cascading bandit literature [13; 14; 18] is dependent on the length of the cascade (K) which is counter-intuitive results. The latest work by Vial et al. [25] removes the dependence on K from their T-step regret upper bound, i.e. $\tilde{\mathcal{O}}(\sqrt{TN})$ where N is the number of ground items, in the tabular case where there is no assumption on the structure of the weight. In the linear case, however, the T-step regret upper bound of their algorithm, referred as `CascadeWOFUL`, is $\tilde{\mathcal{O}}(\sqrt{Td(d+K)})$ which still scales with K.

There are many recent works on assortment bandits [4; 5; 8; 9; 22; 23; 7] using the multinomial choice model While, in the cascade bandits, the agent offers a cascade and a user examines it one by one, in assortment bandits, a user receives an assortment from the agent and examines all the items in a given assortment at once. Due to this difference, the agent receives feedback of only examined items in cascade bandits, but of all items in assortment bandits.

Agrawal et al. [4] and Agrawal et al. [5] propose Thompson sampling and UCB-type algorithms, respectively, in a non-contextual setting. Both show that their regret upper-bound is $\tilde{\mathcal{O}}(\sqrt{NT})$. Oh and Iyengar [22, 23] and Chen et al. [7] incorporate the contextual information into the MNL assortment bandits. They introduce the unknown time-invariant learning parameter which represents the utility of the item, i.e. $x_{ti}^\top \theta$ where x_{ti} is the contextual vector of item i at round t and θ is the parameter that the agent is learning from the feedback. Oh and Iyengar [22] propose Thompson Sampling algorithm and derive $\tilde{\mathcal{O}}(d\sqrt{T})$. Oh and Iyengar [23] propose UCB-type algorithm, referred as `UCB-MNL`, and get a same regret bound $\tilde{\mathcal{O}}(d\sqrt{T})$.

B Proof of Theorem 4.3: α-Regret Analysis

Consider weights $w, w' \in [0,1]^{[N]}$. We denote $w \geq w'$ if $w_i \geq w'_i$ holds for all $i \in [N]$.

Lemma B.1. *Given weights $w, w' \in [0,1]^{[N]}$ such that $w \geq w'$, we have that $f(S_t, w)$ is increasing with respect to w, that is, if $w \geq w'$, then for any $S_t \in \mathcal{S}$, it holds that $f(S_t, w) \geq f(S_t, w')$.*

Proof. It is easy to prove because of the structure of the expected reward fuction. we know that

$$f(S_t, w) = 1 - \prod_{k=1}^{|S_t|} p_t(i_0 | A_{tk}, w).$$

If w increases, then $p_t(i_0|A_{tk}, w)$ decrease and $f(S_t, w)$ increases. □

Lemma B.2. *(Restatement of Lemma 4.4) Suppose $u_{ti} \geq w^*_{ti}$ for all $i \in [N]$. Then,*

$$f(S_t, u_t) - f(S_t, w^*_t) \leq \left(\frac{K}{K+1}\right)^{K+1} \max_{A_{tk} \in S_t} \max_{i \in A_{tk}} \left(u_{ti} - w^*_{ti}\right). \tag{7}$$

Proof. By the mean value theorem,

$$\begin{aligned} f\left(S_t, u_t\right) - f\left(S_t, w^*_t\right) &= \nabla_\theta f\left(S_t, \bar{w}\right)\left(\theta_t - \theta^*\right) \\ &= \left\{\prod_{A_{t\dot{k}} \in S_t} p_t(i_0 | A_{t\dot{k}}, \bar{w})\right\} \sum_{A_{tk} \in S_t} \sum_{i \in A_{tk}} p_t(i|A_{tk}, \bar{w}) x_{ti}^\top (\theta_t - \theta^*) \\ &\leq \left(\frac{K}{K+1}\right)^{K+1} \max_{A_{tk} \in S_t} \max_{i \in A_{tk}} \left(u_{ti} - w^*_{ti}\right) \end{aligned}$$

We can simplify $\left\{\prod_{A_{t\dot{k}} \in S_t} p_t(i_0|A_{t\dot{k}}, \bar{w})\right\} \sum_{A_{tk} \in S_t} \sum_{i \in A_{tk}} p_t(i|A_{tk}, \bar{w})$ on the second equality as follows if we denote $P_{tk} := \sum_{i \in A_{tk}} p_t(i|A_{tk}, \bar{w})$ for convenience:

$$\prod_{\dot{k} \in [K]} (1 - P_{t\dot{k}}) \sum_{k \in [K]} P_{tk}.$$

We can easily see that this expression is maximized as $\left(\frac{K}{K+1}\right)^{K+1}$ when $P_{tk} = \frac{1}{K+1}$ for all $k \in [K]$, since $0 < P_{tk} < 1$. □

Lemma B.3 (Lemma 4 in Oh and Iyengar [22])**.** *Let $\beta_t(\delta) = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tKM}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \frac{\sqrt{\lambda}}{\kappa}$. Then for any $\delta \geq 0$, we have*

$$||\hat{\theta}_t - \theta^*||_{V_t} \leq \beta_t(\delta)$$

with a probability at least $1 - \delta$ for all round t.

Proof. We adapt the proof of Lemma 4 in Oh and Iyengar [22] to our setting. We first denote the probability that a customer clicks the item i in A_{tk} as below:

$$p_t(i|A_{tk}, \theta) = \frac{\exp\left(x_{ti}^\top \theta\right)}{1 + \sum_{j \in A_{tk}} \exp\left(x_{tj}^\top \theta\right)}$$

Then, we define the function $G_t(\theta)$ (where θ is a parameter),

$$G_t(\theta) := \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left[\left(p_\tau(i|A_{\tau k}, \theta) - p_\tau(i|A_{\tau k}, \theta^*)\right) x_{\tau i}\right] + \lambda\left(\theta - \theta^*\right) \tag{8}$$

$G_t(\theta)$ represent the difference in the gradients of the ridge penalized maximum likelihood evaluated at θ and at θ^*. Note that $\hat{\theta}$ can be obtained by minimizing Eq.(2). Therefore, the following equation is satisfied.

$$\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} \left(p_\tau(i_m|A_{\tau k}, \hat{\theta}) - y_{\tau k m}\right) x_{\tau i_m} + \lambda\hat{\theta} = 0 \tag{9}$$

Now we have

$$\begin{aligned} G_t(\hat{\theta}) &= \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left[\left(p_\tau(i|A_{\tau k}, \hat{\theta}) - p_\tau(i|A_{\tau k}, \theta^*)\right) x_{\tau i}\right] + \lambda\left(\hat{\theta} - \theta^*\right) \\ &= \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} \left(p_\tau(i_m|A_{\tau k}, \hat{\theta}) - y_{\tau k m}\right) x_{\tau i_m} + \lambda\hat{\theta} \\ &\quad + \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} \left(p_\tau(i_m|A_{\tau k}, \theta^*) - y_{\tau k m}\right) x_{\tau i_m} - \lambda\theta^* \\ &= 0 + \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} \epsilon_{\tau k m} x_{\tau i_m} - \lambda\theta^* \end{aligned}$$

where the last equality is from Eq.(9) and the definition of $\epsilon_{tkm} = y_{tkm} - p_t(i_m|A_{tk}, \theta^*)$. For convenience, we define $Z_t := \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} \epsilon_{\tau k m} x_{\tau i_m}$.

For any parameters $\theta_1, \theta_2 \in \mathbb{R}^d$, by mean value theorem, there exists $\bar{\theta} = c\theta_1 + (1-c)\theta_2$ with some $c \in (0,1)$ such that

$$
\begin{aligned}
& G_t(\theta_1) - G_t(\theta_2) \\
&= \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left[\left(p_\tau(i|A_{\tau k}, \theta_1) - p_\tau(i|A_{\tau k}, \theta_2) \right) x_{\tau i} \right] + \lambda(\theta_1 - \theta_2) \\
&= \left[\left(\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} \nabla_j p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau j}^\top \right) + \lambda I_d \right] + \lambda(\theta_1 - \theta_2) \\
&= \left[\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \left(\sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \sum_{i,j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau j}^\top \right) \right] (\theta_1 - \theta_2) \\
&\quad + \lambda I_d (\theta_1 - \theta_2)
\end{aligned}
$$

where I_d is a $d \times d$ identity matrix. We define the matrix $H_{\tau k}$ as

$$
H_{\tau k} := \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau j}^\top
$$

We can see that $H_{\tau k}$ is positive semi-definite, since $H_{\tau k}$ is a Hessian of a negative log-likelihood which is convex.

Also, notice that

$$
(x_i - x_j)(x_i - x_j)^\top = x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top \succeq 0
$$

which implies $x_i x_i^\top + x_j x_j^\top \succeq x_i x_j^\top + x_j x_i^\top$. Therefore, it follows that

$$
\begin{aligned}
H_{\tau k} &= \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau j}^\top \\
&= \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \frac{1}{2} \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) \left(x_{\tau i} x_{\tau j}^\top + x_{\tau j} x_{\tau i}^\top \right) \\
&\succeq \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \frac{1}{2} \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) \left(x_{\tau i} x_{\tau i}^\top + x_{\tau j} x_{\tau j}^\top \right) \\
&= \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top - \sum_{i \in A_{\tau k}} \sum_{j \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(j|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top \\
&= \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) \left(1 - \sum_{j \in A_{\tau k}} p_\tau(j|A_{\tau k}, \bar{\theta}) \right) x_{\tau i} x_{\tau i}^\top \\
&= \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(i_0|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top
\end{aligned}
$$

where $p_\tau(i_0|A_{\tau k}, \theta) := \frac{1}{1 + \sum_{j \in A_{\tau k}} x_{\tau i}^\top \theta}$ is the click probability of the outside option with respect to θ at round τ. Now,

$$
\begin{aligned}
G_t(\theta_1) - G_t(\theta_2) &= \left[\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} H_{\tau k} + \lambda I_d \right] (\theta_1 - \theta_2) \\
&\geq \left[\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \bar{\theta}) p_\tau(i_0|A_{\tau k}, \bar{\theta}) x_{\tau i} x_{\tau i}^\top + \lambda I_d \right] (\theta_1 - \theta_2) \\
&:= \mathcal{H}(\bar{\theta})(\theta_1 - \theta_2).
\end{aligned}
$$

$p_\tau(i|A_{\tau k},\bar\theta)p_\tau(i_0|A_{\tau k},\bar\theta)$ is lower-bounded by κ from Assumption 4.2. Then we have

$$(\theta_1-\theta_2)^\top(G_t(\theta_1)-G_t(\theta_2)) \geq (\theta_1-\theta_2)^\top(\kappa V_t)(\theta_1-\theta_2) > 0$$

for any $\theta_1 \neq \theta_2$. . From the definition on Eq.(8), $G_t(\theta^*)=0$. Hence, for any $\theta\in\mathbb{R}^d$, we have

$$\begin{aligned}
||G_t(\theta)||^2_{V_t^{-1}} &= ||G_t(\theta)-G_t(\theta^*)||^2_{V_t^{-1}} \\
&= (G_t(\theta)-G_t(\theta^*))^\top V_t^{-1}(G_t(\theta)-G_t(\theta^*)) \\
&\geq (\theta-\theta^*)^\top \mathcal{H}(\bar\theta)V_t^{-1}\mathcal{H}(\bar\theta)(\theta-\theta^*) \\
&\geq \kappa^2(\theta-\theta^*)^\top V_t(\theta-\theta^*) \\
&= \kappa^2||\theta-\theta^*||^2_{V_t}
\end{aligned}$$

where the last inequality is from $\mathcal{H}(\bar\theta)\succeq\kappa V_t$. Using $G_t(\hat\theta)=Z_t-\lambda\theta^*$, we have

$$\kappa||\hat\theta-\theta^*||_{V_t} \leq ||G_t(\hat\theta)||_{V_t^{-1}} \leq ||Z_t||_{V_t^{-1}} + \lambda||\theta^*||_{V_t^{-1}}$$

Recall that $Z_t=\sum_{\tau=1}^t\sum_{k\in O_\tau}\sum_{i_m\in A_{\tau k}}\epsilon_{\tau k m}x_{\tau i_m}$ and $\epsilon_{\tau k m}$ is sub-Gaussian with parameter σ, then we can apply Theorem 1 in Abbasi-Yadkori et al. [1]:

$$||Z_t||^2_{V_t^{-1}} \leq 2\sigma^2\log\left(\frac{\text{det}(V_t)^{1/2}\text{det}(V)^{-1/2}}{\delta}\right)$$

with probability at least $1-\delta$. Then we combine with Lemma B.9:

$$\begin{aligned}
||Z_t||^2_{V_t^{-1}} &\leq 2\sigma^2\left[\frac{d}{2}\ln\left(\frac{\text{trace}V+tKM}{d}\right)-\frac{1}{2}\ln\det V+\ln\frac{1}{\delta}\right] \\
&\leq 2\sigma^2\left[\frac{d}{2}\ln\left(d\lambda+tKM\right)d-\frac{1}{2}\ln\lambda^d+\ln\frac{1}{\delta}\right] \\
&= 2\sigma^2\left[\frac{d}{2}\ln\left(\lambda+\frac{tKM}{d}\right)-\frac{d}{2}\ln\lambda+\ln\frac{1}{\delta}\right] \\
&= 2\sigma^2\left[\frac{d}{2}\ln\left(1+\frac{tKM}{d\lambda}\right)+\ln\frac{1}{\delta}\right]
\end{aligned}$$

where the first inequality is from the fact that $V=\lambda I$. Next we need to bound $\lambda||\theta^*||_{V_t^{-1}}$. We have

$$||\theta^*||^2_{V_t^{-1}} \leq \frac{||\theta^*||_2^2}{\lambda_{\min}(V_t)} \leq \frac{||\theta^*||_2^2}{\lambda_{\min}(V)} \leq \frac{||\theta^*||_2^2}{\lambda}.$$

By Assumption 4.1 that $||\theta^*||_2^2\leq 1$, $\lambda||\theta^*||_{V_t^{-1}}\leq\sqrt{\lambda}$. Recall that $\sigma=\frac{1}{2}$ in our problem. Combining the bound of $||Z_t||_{V_t^{-1}}$ and $\lambda||\theta^*||_{V_t^{-1}}$, we have

$$||\hat\theta_t-\theta^*||_{V_t} \leq \frac{1}{2\kappa}\sqrt{d\ln\left(1+\frac{tKM}{d\lambda}+2\ln\frac{1}{\delta}\right)}+\frac{\sqrt{\lambda}}{\kappa}$$

with probability at least $1-\delta$. □

Thus $\hat\theta_t$ lies in the ellipsoid centered at θ^* with confidence radius $\beta_t(\delta)$ under V_t norm. Building on this, we can define an upper confidence bound of the true weight for each base arm i by

$$u_{ti} = x_{ti}^\top\hat\theta_{t-1}+\beta_{t-1}||x_{ti}||_{V_{t-1}^{-1}}$$

Recall that we define the high-probability concentration event $\mathcal{E}_t(\delta):=\{||\hat\theta_t-\theta^*||_{V_t}\leq\beta_t(\delta)\}$. The fact that u_{ti} is an upper confidence bound of true weight $w^*_{ti}=x_{ti}^\top\theta^*$ is proved in the following Lemma B.4 and Lemma B.5.

Lemma B.4 (Optimism)**.** *On event $\mathcal{E}_t(\delta)$, for every item $i \in [N]$, we have*

$$u_{ti} \geq w_{ti}^* \,. \tag{10}$$

Proof. Recall that $w_{ti}^* = x_{ti}^\top \theta^*$. By Hölder's inequality,

$$\begin{aligned}
\left| x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^* \right| &= \left| \left[V_{t-1}^{1/2} \left(\hat{\theta}_{t-1} - \theta^* \right) \right]^\top \left(V_{t-1}^{-1/2} x_{ti} \right) \right| \\
&\leq ||V_{t-1}^{1/2} \left(\hat{\theta}_{t-1} - \theta^* \right) ||_2 ||V_{t-1}^{-1/2} x_{ti}||_2 \\
&= ||\hat{\theta}_{t-1} - \theta^*||_{V_{t-1}} ||x_{ti}||_{V_{t-1}^{-1}} \\
&\leq \beta_{t-1}(\delta) ||x_{ti}||_{V_{t-1}^{-1}} . \qquad (11)
\end{aligned}$$

From the last inequality above, we have

$$-\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} \leq x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^* .$$

Add $\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}$, then we have

$$0 \leq \left(x_{ti}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} \right) - x_{ti}^\top \theta^* = u_{ti} - w_{ti}^* .$$

□

Lemma B.5 (Concentration of UCB weights)**.** *On event $\mathcal{E}_t(\delta)$, for every item $i \in [N]$, we have*

$$u_{ti} - w_{ti}^* \leq 2\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} . \tag{12}$$

Proof. From inequality (11), we have

$$\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} \geq x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^* .$$

Adding $\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}$ to both sides gives,

$$2\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} \geq \left(x_{ti}^\top \hat{\theta}_{t-1} + \beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}} \right) - x_{ti}^\top \theta^* = u_{ti} - w_{ti}^*$$

□

Lemma B.6. *For any round t and action S_t, we have*

$$\mathcal{R}^\alpha \left(t, S_t \right) \leq 2 \left(\frac{K}{K+1} \right)^{K+1} \max_{k \in [|S_t|]} \max_{i \in A_{tk}} \beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}$$

Proof. Let $S^{u_t} = \arg\max_{S \in \mathcal{S}} f\left(S, u_t\right)$ and recall that $S_t^* = \arg\max_{S \in \mathcal{S}} f\left(S, w_t^*\right)$. Then

$$f\left(S_t, u_t\right) \geq \alpha f\left(S^{u_t}, u_t\right) \geq \alpha f\left(S_t^*, u_t\right) \geq \alpha f\left(S_t^*, w_t^*\right) .$$

The first inequality is by the definition of α-approximate oracle. The second inequality comes from the fact that S^{u_t} has the maximum expected reward when u_t is given. Combining Lemma B.4 which states $u_t \geq w_t^*$ and Lemma B.1 which is about the monotonicity of the expected reward function, we can obtain the last inequality. Then we can bound the $\mathcal{R}^\alpha\left(t, S_t\right)$ with the expected reward difference of S_t between given u_t and w_t^* as follows:

$$\mathcal{R}^\alpha \left(t, S_t \right) = \alpha f\left(S_t^*, w_t^*\right) - f(S_t, w_t^*) \leq f\left(S_t, u_t\right) - f(S_t, w_t^*).$$

By Lemma B.2 and Lemma B.5,

$$\begin{aligned}
\mathcal{R}^\alpha \left(t, S_t \right) &\leq f\left(S_t, u_t\right) - f(S_t, w_t^*) \\
&\leq \left(\frac{K}{K+1} \right)^{K+1} \max_{k \in [|S_t|]} \max_{i \in A_{tk}} \left(u_{ti} - w_{ti}^* \right) \\
&\leq 2 \left(\frac{K}{K+1} \right)^{K+1} \max_{k \in [|S_t|]} \max_{i \in A_{tk}} \beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}
\end{aligned}$$

□

Lemma B.7. *Let $x_i \in \mathbb{R}^d$ and $I \in \mathbb{R}^{d\times d}$ be an identity matrix. Then we have*

$$\det\left(I + \sum_{i=1}^{n} x_i x_i^\top\right) \geq 1 + \sum_{i=1}^{n} ||x_i||_2^2.$$

Proof. Let the eigenvalues of $\sum_{i=1}^{n} x_i x_i^\top$ be $\lambda_1, \ldots, \lambda_d$ where $\lambda_j \geq 0$ for all $1 \leq j \leq d$. Then we have

$$\begin{aligned}
\det\left(I + \sum_{i=1}^{n} x_i x_i^\top\right) &= \prod_{j=1}^{d}(1+\lambda_j) \geq 1 + \sum_{j=1}^{d}\lambda_j = 1 - d + \sum_{j=1}^{d}(1+\lambda_j) \\
&= 1 - d + \text{trace}\left(I + \sum_{i=1}^{n} x_i x_i^\top\right) = 1 - d + d + \sum_{i=1}^{n} ||x_i||_2^2 \\
&= 1 + \sum_{i=1}^{n} ||x_i||_2^2
\end{aligned}$$

□

Lemma B.8 (Maximal elliptical potential)**.**

$$\sum_{\tau=1}^{t} \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_\tau^{-1}}^2 \leq 2\ln\left(\frac{\det(V_t)}{\lambda^d}\right)$$

Proof. Let $\lambda_{\min}(V_\tau)$ be the minimum eigenvalue of V_τ. Since $\lambda \geq 1$ and $||x_{\tau i}||_{V_{\tau-1}}^2 \leq \frac{||x_{\tau i}||_2^2}{\lambda_{\min}(V_{\tau-1})} \leq \frac{1}{\lambda}$, we have

$$\max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_{\tau-1}^{-1}}^2 \leq 1.$$

Using the fact that $z \leq 2\ln(1+z)$ for any $z \in [0,1]$, we have

$$\begin{aligned}
\sum_{\tau=1}^{t} \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_\tau^{-1}}^2 &\leq 2\sum_{\tau=1}^{t}\ln\left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_\tau^{-1}}^2\right) \\
&= 2\ln\prod_{\tau=1}^{t}\left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_\tau^{-1}}^2\right). \qquad (13)
\end{aligned}$$

Now we upper bound $\prod_{\tau=1}^{t}\left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_\tau^{-1}}^2\right)$ from $\det(V_t)$.

$$\begin{aligned}
\det(V_t) &= \det\left(V_{t-1} + \sum_{k=1}^{O_t}\sum_{i\in A_{tk}} x_{ti}x_{ti}^\top\right) \\
&= \det(V_{t-1})\det\left(I + V_{t-1}^{-1/2}\sum_{k=1}^{O_t}\sum_{i\in A_{tk}} x_{ti}x_{ti}^\top V_{t-1}^{-1/2}\right) \\
&= \det(V_{t-1})\det\left(I + \sum_{k=1}^{O_t}\sum_{i\in A_{tk}} \left(V_{t-1}^{-1/2}x_{ti}\right)\left(V_{t-1}^{-1/2}x_{ti}\right)^\top\right) \\
&\geq \det(V_{t-1})\left(1 + \sum_{k=1}^{O_t}\sum_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}^2\right) \\
&\geq \det(\lambda I)\prod_{\tau=1}^{t}\left(1 + \sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}} ||x_{\tau i}||_{V_{\tau-1}^{-1}}^2\right) \\
&\geq \det(\lambda I)\prod_{\tau=1}^{t}\left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||x_{\tau i}||_{V_{\tau-1}^{-1}}^2\right).
\end{aligned}$$

The second equality above is from the fact that $V + U = V^{1/2}(I + V^{-1/2}UV^{-1/2})V^{1/2}$ for a symmetric positive definite matrix V. The first inequality above can be obtained by applying Lemma B.7. Applying the first inequality repeatedly, we can get the second inequality above. Thus, we have

$$\prod_{\tau=1}^{t}\left(1+\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||x_{\tau i}||^2_{V_{\tau-1}^{-1}}\right)\le\frac{\det(V_t)}{\det(\lambda I)}. \tag{14}$$

Then applying Eq.(14) to Eq.(13), we complete the proof as follows:

$$\sum_{\tau=1}^{t}\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||x_{\tau i}||^2_{V_\tau^{-1}}\le 2\ln\frac{\det(V_t)}{\det(\lambda I)}\le 2\ln\frac{\det(V_t)}{\lambda^d}$$

where the last inequality is from Lemma B.9. □

Lemma B.9 (Lemma 10 in Abbasi-Yadkori et al. [1]). *det* (V_t) *is increasing with respect to* t *and*

$$\mathit{det}\,(V_t)\le\left(\lambda+\frac{tKM}{d}\right)^d$$

Proof. We first prove that det (V_t) is increasing with respect to t. For any symmetric positive definite matrix $\tilde{V}\in\mathbb{R}^{d\times d}$ and column vector $x\in\mathbb{R}^{d\times 1}$, we can see $\det(\tilde{V}+xx^\top)\ge\det(\tilde{V})$ as follows:

$$\begin{aligned}\det(\tilde{V}+xx^\top)&=\det(\tilde{V})\det(I+\tilde{V}^{-1/2}xx^\top\tilde{V}^{-1/2})\\&=\det(\tilde{V})\det(1+||\tilde{V}^{-1/2}x||^2)\\&\ge\det(\tilde{V}).\end{aligned}$$

The second equality above is due to Sylvester's determinant theorem, which states that $\det(I+AB)=\det(I+BA)$.

Next, we prove the inequality in Lemma B.9. Let $\lambda_1,\ldots,\lambda_d$ be the eigenvalues of $V_t\in\mathbb{R}^{d\times d}$. Then

$$\begin{aligned}\det(V_t)&=\lambda_1\lambda_2\cdots\lambda_d\\&\le\left(\frac{\lambda_1+\cdots+\lambda_d}{d}\right)^d=\left(\frac{\mathrm{trace}(V_t)}{d}\right)^d.\end{aligned}$$

The second inequality above is from the AM-GM inequality. Now we need to bound $\mathrm{trace}(V_t)$ as follows:

$$\begin{aligned}\mathrm{trace}(V_t)&=\mathrm{trace}(V)+\sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}}\mathrm{trace}(x_{\tau i}x_{\tau i}^\top)\\&=d\lambda+\sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}}||x_{\tau i}||_2^2\\&\le d\lambda+tKM\end{aligned}$$

The second inequality is due to Assumption 4.1 that $||x_{ti}||\le 1$. Thus, $\det(V_t)\le\left(\lambda+\frac{tKM}{d}\right)^d$. □

Now, we can prove Theorem 4.3.

Proof of Theorem 4.3. Suppose B.3 holds for all round t. Then, with probability $1-\delta$, we have

$$
\begin{aligned}
\mathcal{R}^{\alpha}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\mathcal{R}^{\alpha}(t,S_t)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} 2\left(\frac{K}{K+1}\right)^{K+1} \max_{k\in[|S_t|]}\max_{i\in A_{tk}} \beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}\right] &(15)\\
&\leq \mathbb{E}\left[2\left(\frac{K}{K+1}\right)^{K+1} \beta_T(\delta)\sum_{t=1}^{T}\max_{k\in[|S_t|]}\max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}\right] &(16)\\
&= \mathbb{E}\left[2\left(\frac{K}{K+1}\right)^{K+1} \beta_T(\delta)\sum_{t=1}^{T}\max_{k\in[O_t]}\max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}\right] &(17)
\end{aligned}
$$

Eq.(15) comes from Lemma B.6. Using the fact that $\beta_t(\delta)$ is increasing with respect to t, Eq.(16) is satisfied. Note that the upper bound of $\mathcal{R}^{\alpha}(t,S_t)$ is in terms of all assortments of S_t in Eq.(16). The previous work [18] mentioned that it is hard to estimate an upper bound for $2\max_{k\in[K]}\max_{i\in A_{tk}}\beta_{t-1}(\delta)||x_{ti}||_{V_{t-1}^{-1}}$ because V_t only contains information of observed assortments. We cope with this by max operations and the property of `UCB-CCA` that the largest item in S_t in terms of $||x_{ti}||_{V_{t-1}^{-1}}$ is always in the first assortment. Then, we have

$$\max_{k\in[|S_t|]}\max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}} = \max_{k\in[O_t]}\max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}$$

which is stated in Eq.(5), and thus Eq.(17) is given by the equation above.

We complete the remain part as follows:

$$
\begin{aligned}
\mathcal{R}^{\alpha}(T) &\leq \mathbb{E}\left[2C_\kappa\beta_T(\delta)\sum_{t=1}^{T}\max_{k\in[O_t]}\max_{i\in A_{tk}} ||x_{ti}||_{V_{t-1}^{-1}}\right] \\
&\leq \mathbb{E}\left[2C_\kappa\beta_T(\delta)\sqrt{\sum_{t=1}^{T}1^2\sum_{t=1}^{T}\max_{k\in[O_t]}\max_{i\in A_{tk}} ||x_{ti}||^2_{V_{t-1}^{-1}}}\right] \\
&\leq \mathbb{E}\left[2C_\kappa\left(\frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{TKM}{d\lambda}\right)+2\log\frac{1}{\delta}}+\frac{\sqrt{\lambda}}{\kappa}\right)\sqrt{T\cdot 2\ln\left(\frac{\det(V_t)}{\lambda^d}\right)}\right] \\
&\leq 2\sqrt{2}C_\kappa\left(\frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{TKM}{d\lambda}\right)+2\log\frac{1}{\delta}}+\frac{\sqrt{\lambda}}{\kappa}\right)\sqrt{Td\ln\left(1+\frac{TKM}{d\lambda}\right)}.
\end{aligned}
$$

The first inequality above is by applying Cauchy-Schwartz inequality. The second inequality comes from the definition of $\beta_t(\delta)$ in Lemma B.3 and Lemma B.7. The last inequality is from the upper bound of $\det(V_t)$ in Lemma B.9

Since we set $\delta = \frac{1}{t^2}$, we have

$$\mathcal{R}^{\alpha}(T) = \leq 2\sqrt{2}C_\kappa\left(\frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{TKM}{d\lambda}\right)+4\log T}+\frac{\sqrt{\lambda}}{\kappa}\right)\sqrt{Td\ln\left(1+\frac{TKM}{d\lambda}\right)}.$$

□

C Proof of Theorem 5.2

Since `UCB-CCA+` is a parametric-based algorithm, we redefine our notation for convenience. We denote the user click probability of m-th item in A_{tk} and probability of the outside option in A_{tk} as:

$$p_t(i_m|A_{tk},\theta) = \frac{\exp\left(x_{ti_m}^{\top}\theta\right)}{1+\sum_{j\in A_{tk}}\exp\left(x_{tj}^{\top}\theta\right)}.$$

We also denote the expected reward function as:

$$f(S_t, \theta^*) = \sum_{k=1}^{K} \left\{ \prod_{\dot{k}=1}^{k-1} p_t(i_0 | A_{t\dot{k}}, \theta^*) \right\} \sum_{i \in A_{tk}} p_t(i | A_{tk}, \theta^*).$$

We define the first and second derivative of $p_t(i|A_{tk}, \theta)$ with respect to the arm i:

$$\begin{aligned} \dot{p}_t(i|A_{tk}, \theta) &:= p_t(i|A_{tk}, \theta) p_t(i_0|A_{tk}, \theta) \\ \ddot{p}_t(i|A_{tk}, \theta) &:= p_t(i|A_{tk}, \theta) p_t(i_0|A_{tk}, \theta)(1 - 2p_t(i_0|A_{tk}, \theta)). \end{aligned}$$

Additionally, we define the new design matrix containing local information:

$$H_t(\theta) := \sum_{\tau=1}^{t} \sum_{k \in [O_\tau]} \sum_{i \in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d \tag{18}$$

$$H_t := \sum_{\tau=1}^{t} \sum_{k \in [O_\tau]} \sum_{i \in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta^*) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d. \tag{19}$$

The negative log-likelihood function under any parameter θ is given by:

$$\ell_t(\theta) := -\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} y_{\tau k m} \log p_\tau(i_m | A_{\tau k}, \theta)$$

where $y_{\tau km}$ is a user choice random variable on the m-th item i_m in the k-th assortment $A_{\tau k}$ at round τ. And the regularized maximum likelihood estimates $\hat{\theta}_t$ is given by minimizing the regularized log-likelihood function:

$$\hat{\theta}_t = \operatorname*{argmin}_\theta \left[\ell_t(\theta) + \frac{\lambda_t}{2} ||\theta||_2^2 \right]$$

where the penalty parameter $\lambda_t \geq 1$. To obtain $\hat{\theta}_t$, we take the gradient of the above regularized log-likelihood function with respect to θ:

$$\nabla_\theta \left[\ell_t(\theta) + \frac{\lambda_t}{2} ||\theta||_2^2 \right] = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \{ p_\tau(i|A_{\tau k}, \theta) - y_{\tau k m} \} x_{\tau i} + \lambda_t \theta. \tag{20}$$

And we denote the function $g_t(\theta)$ as follows:

$$g_t(\theta) := \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \theta) x_{\tau i} + \lambda_t \theta. \tag{21}$$

Since $\hat{\theta}_t$ is the minimizer of the regularized negative log-likelihood function, we can get $\hat{\theta}_t$ by setting Eq.(20) to 0. Then we can see that $g_t(\hat{\theta}_t)$ can be represented as a function of feature vector x_{ti} and user choice variable y_{tkm}:

$$\sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} y_{\tau k m} x_{\tau i} = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} p_\tau(i|A_{\tau k}, \hat{\theta}_t) + \lambda_t \hat{\theta}_t x_{\tau i} = g_t(\hat{\theta}_t) \tag{22}$$

Now, the following lemma illustrates the relationship between the function $g_t(\theta)$ and its input parameter θ.

Lemma C.1. *For any parameter $\theta_1, \theta_2 \in \mathbb{R}^d$, the following equality hold:*

$$\begin{aligned} g_t(\theta_1) - g_t(\theta_2) &= \mathbb{G}_t(\theta_1, \theta_2) \\ ||g_t(\theta_1) - g_t(\theta_2)||_{\mathbb{G}_t^{-1}(\theta_1, \theta_2)} &= ||\theta_1 - \theta_2||_{\mathbb{G}_t(\theta_1, \theta_2)} \end{aligned}$$

where $\mathbb{G}_t(\theta_1, \theta_2) = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left[\int_{v=0}^{1} \dot{p}_\tau(i|A_{\tau k}, v\theta_2 + (1-v)\theta_1) dv \right] x_{\tau i} x_{\tau i}^\top + \lambda_t I_d$ and I_d is d-dimensional identitiy matrix.

Proof. We first derive the first equality in Lemma C.1.

$$
\begin{aligned}
g_t(\theta_1) - g_t(\theta_2) &= \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}} \{p_\tau(i|A_{\tau k},\theta_1) - p_\tau(i|A_{\tau k},\theta_2)\}\, x_{\tau i} + \lambda_t(\theta_1-\theta_2)\\
&= \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}} \left[\int_{v=0}^{1} \dot p_\tau(i|A_{\tau k}, v\theta_2 + (1-v)\theta_1)dv\right] x_{\tau i}x_{\tau i}^\top(\theta_1-\theta_2) + \lambda_t(\theta_1-\theta_2)\\
&= \left(\sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}} \left[\int_{v=0}^{1} \dot p_\tau(i|A_{\tau k}, v\theta_2 + (1-v)\theta_1)dv\right] x_{\tau i}x_{\tau i}^\top + \lambda_t I_d\right)(\theta_1-\theta_2)\\
&= \mathbb{G}_t(\theta_1,\theta_2)(\theta_1-\theta_2)
\end{aligned}
$$

Since $\int_{v=0}^{1} \dot p_t(i|A_{tk}, v\theta_2+(1-v)\theta_1)dv \geq \kappa$ from Assumption 4.2 and the definition of $\dot p_t(i|A_{tk},\theta)$, $\mathbb{G}_t(\theta_1,\theta_2)(\theta_1-\theta_2)$ is positive definite. Thus, we can derive the second equality in Lemma C.1 from the first eqaulity in the same lemma. □

Theorem C.2 (Theorem 4 in Abeille et al. [2])**.** *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a stochastic process in such that x_t is $\mathcal{F}_t$ measurable. Let $\{\epsilon_t\}_{t=2}^{\infty}$ be a martingale difference sequence such that ϵ_{t+1} is $\mathcal{F}_{t+1}$ measurable. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\epsilon_{t+1}| \leq 1$ almost surely, and note $\sigma_t^2 := \mathbb{E}\left[\epsilon_{t+1}^2 \mid \mathcal{F}_t\right]$. Let $\{\lambda_t\}_{t=1}^{\infty}$ be a predictable sequence of non-negative scalars. Define: $\mathbb{H}_t := \sum_{\tau=1}^{t}\sigma_\tau^2 x_\tau x_t^2 + \lambda_t I_d$, $\mathbb{M}_t := \sum_{\tau=1}^{t}\epsilon_{\tau+1}x_\tau$ Then for any $\delta \in (0,1]$:*

$$
\mathbb{P}\left(\exists t\geq 1, ||\mathbb{M}_t||_{\mathbb{H}_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}}\log\left(\frac{\det(\mathbb{H}_t)^{\frac{1}{2}}\lambda_t^{-\frac{d}{2}}}{\delta}\right) + \frac{2}{\sqrt{\lambda_t}} d\log 2\right) \leq \delta
$$

Using Theorem C.2, we show that the following holds with high probability.

Lemma C.3. *With $\hat\theta_t$ as the regularized maximum log-likelihood estimate as defined in Eq.*(2)*, the following inequality holds with probability at least $1-\delta$:*

$$
\forall t \geq 1, ||g_t(\hat\theta_t) - g_t(\theta^*)||_{H_t^{-1}} \leq \gamma_t(\delta), \tag{23}
$$

where the confidence radius $\gamma_t(\delta) := \frac{3\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}}\log\left(\frac{(\lambda_t + tKM/d)^{d/2}\lambda_t^{-d/2}}{\delta}\right) + \frac{2d}{\sqrt{\lambda_t}}\log 2.$

Proof. In Eq.(22), the following holds:

$$
g_t(\hat\theta_t) = \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}} y_{\tau k m}x_{\tau i}.
$$

Subtract $g_t(\theta^*) = \sum_{\tau=1}^{t-1}\sum_{k=1}^{O_\tau}\sum_{i_m\in A_{\tau k}} p_\tau(i_m|A_{\tau k},\theta^*) + \lambda_t\theta^*$ from both sides of the above equation, then we get:

$$
\begin{aligned}
g_t(\hat\theta_t) - g_t(\theta^*) &= \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i_m\in A_{\tau k}} \{y_{\tau k m} - p_\tau(i_m|A_{\tau k},\theta^*)\}\, x_{\tau i_m} - \lambda_t\theta^*\\
&= \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i_m\in A_{\tau k}} \epsilon_{\tau k m}x_{\tau i_m} - \lambda_t\theta^*\\
&=: M_t - \lambda_t\theta^*.
\end{aligned}
$$

Thus, we have:

$$
||g_t(\hat\theta_t) - g_t(\theta^*)||_{H_t^{-1}} \leq ||M_t||_{H_t^{-1}} + \lambda_t||\theta^*||_{H_t^{-1}}. \tag{24}
$$

Now, we need to bound $\lambda_t||\theta^*||_{H_t^{-1}}$. We have

$$
||\theta^*||_{H_t^{-1}}^2 \leq \frac{||\theta^*||_2^2}{\lambda_{\min}(H_t)} \leq \frac{||\theta^*||_2^2}{\lambda_{\min}(\lambda_t I_d)} \leq \frac{||\theta^*||_2^2}{\lambda_t}.
$$

By Assumption 4.1 that $||\theta^*||_2^2 \leq 1$, $\lambda_t ||\theta^*||_{H_t^{-1}} \leq \sqrt{\lambda_t}$. We can rewritten Equation (24) as follows:

$$||g_t(\hat{\theta}_t) - g_t(\theta^*)||_{H_t^{-1}} \leq ||M_t||_{H_t^{-1}} + \sqrt{\lambda_t}. \tag{25}$$

Since the reward at any round and for any cascade is constrained to be no more than 1, given the filtration $\mathcal{F}_t$, $\epsilon_{\tau km}$ which is upper bounded by 1 behaves as a martingale difference. To apply Theorem C.2, we calculate $\forall \tau \geq 1$:

$$\begin{aligned}
\mathbb{E}\left[\epsilon_{\tau km}^2 \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\{y_{\tau km} - p_\tau(i_m|A_{\tau k}, \theta^*)\}^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{V}\left[y_{\tau km} \mid \mathcal{F}_t\right] = p_\tau(i_m|A_{\tau k}, \theta^*)(1 - p_\tau(i_m|A_{\tau k}, \theta^*)) \\
&=: \dot{p}_\tau(i_m|A_{\tau k}, \theta^*).
\end{aligned}$$

Then, setting $\mathbb{H}_t$ as H_t and $\mathbb{M}_t$ as M_t, we obtain:

$$1 - \delta \leq \mathbb{P}\left(\forall t \geq 1, ||M_t||_{H_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{\det(H_t)^{\frac{1}{2}} \lambda_t^{-\frac{d}{2}}}{\delta}\right) + \frac{2}{\sqrt{\lambda_t}} d \log 2\right). \tag{26}$$

Now, we need to bound $\det(H_t)$.

$$\begin{aligned}
\det(H_t) &= \det\left(\sum_{\tau=1}^{t} \sum_{k \in [O_\tau]} \sum_{i \in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta^*) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d\right) \\
&\leq \det\left(\sum_{\tau=1}^{t-1} \sum_{k \in [O_\tau]} \sum_{i \in A_{\tau k}} x_{\tau i} x_{\tau i}^\top + \lambda_t I_d\right) \\
&= \det(V_t) \leq \left(\lambda_t + \frac{tKM}{d}\right)^d.
\end{aligned}$$

The first inequality is from the fact that $\dot{p}_\tau(i|A_{\tau k}, \theta^*) \leq 1$ and the last inequality is obtained by using Lemma B.9. We substitute the above results into Eq.(26). This simplify Eq.(26) as:

$$1 - \delta \leq \mathbb{P}\left(\forall t \geq 1, ||M_t||_{H_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{(\lambda_t + tKM/d)^{\frac{d}{2}} \lambda_t^{-\frac{d}{2}}}{\delta}\right) + \frac{2}{\sqrt{\lambda_t}} d \log 2\right).$$

Thus, the following holds with probability at least $1 - \delta$ by combining the above result and Equation (25):

$$||g_t(\hat{\theta}_t) - g_t(\theta^*)||_{H_t^{-1}} \leq \frac{3\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{(\lambda_t + tKM/d)^{\frac{d}{2}} \lambda_t^{-\frac{d}{2}}}{\delta}\right) + \frac{2}{\sqrt{\lambda_t}} d \log 2.$$

□

Lemma C.4 (Lemma 12 in [3]). *For an assortment A_{tk} and $\theta_1, \theta_2 \in \mathbb{R}_d$, the following holds:*

$$\sum_{i \in A_{tk}} \int_{v=0}^{1} \dot{p}_t(i|A_{tk}, v\theta_2 + (1-v)\theta_1) \cdot dv \geq \sum_{i \in A_{tk}} \dot{p}_t(i|A_{tk}, \theta_1)(1 + |x_{ti}^\top \theta_1 - x_{ti}^\top \theta_2|)^{-1}.$$

Lemma C.5. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, the followings hold:*

$$\begin{aligned}
\mathbb{G}_t(\theta_1, \theta_2) &\succeq \frac{1}{3} H_t(\theta_1) \\
\mathbb{G}_t(\theta_1, \theta_2) &\succeq \frac{1}{3} H_t(\theta_2)
\end{aligned}$$

where $\mathbb{G}_t(\theta_1, \theta_2) = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left[\int_{v=0}^{1} \dot{p}_\tau(i|A_{\tau k}, v\theta_2 + (1-v)\theta_1) dv\right] x_{\tau i} x_{\tau i}^\top + \lambda_t I_d$ and I_d is d-dimensional identitiy matrix.

Proof.

$$\begin{aligned}\sum_{i\in A_{tk}}\int_{v=0}^{1}\dot{p}_t(i|A_{tk},v\theta_2+(1-v)\theta_1)\cdot dv &\geq \sum_{i\in A_{tk}}\dot{p}_t(i|A_{tk},\theta_1)(1+|x_{ti}^\top\theta_1-x_{ti}^\top\theta_2|)^{-1}\\ &\geq \sum_{i\in A_{tk}}\dot{p}_t(i|A_{tk},\theta_1)(1+||x_{ti}||_2||\theta_1-\theta_2||_2)^{-1}\\ &\geq \sum_{i\in A_{tk}}\frac{\dot{p}_t(i|A_{tk},\theta_1)}{3}.\end{aligned}$$

The first inequality is from Lemma C.4, the second inequality is obtained by applying Cachy-Schwarz inequality and the last inequality is from our Assumption 4.1. The definition of $\mathbb{G}_t(\theta_1,\theta_2)$ is as follows:

$$\begin{aligned}\mathbb{G}_t(\theta_1,\theta_2) &= \sum_{\tau=1}^{t}\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}}\left[\int_{v=0}^{1}\dot{p}_\tau(i|A_{\tau k},v\theta_2+(1-v)\theta_1)dv\right]x_{\tau i}x_{\tau i}^\top+\lambda_t I_d\\ &\succeq \frac{1}{3}\sum_{\tau=1}^{t}\sum_{k\in[O_\tau]}\sum_{i\in A_{\tau k}}\dot{p}_t(i|A_{tk},\theta_1)x_{\tau i}x_{\tau i}^\top+\lambda_t I_d\\ &= \frac{1}{3}\left\{\sum_{\tau=1}^{t}\sum_{k\in[O_\tau]}\sum_{i\in A_{\tau k}}\dot{p}_t(i|A_{tk},\theta_1)x_{\tau i}x_{\tau i}^\top+(1+2)\lambda_t I_d\right\}\\ &\succeq \frac{1}{3}\left\{\sum_{\tau=1}^{t}\sum_{k\in[O_\tau]}\sum_{i\in A_{\tau k}}\dot{p}_t(i|A_{tk},\theta_1)x_{\tau i}x_{\tau i}^\top+\lambda_t I_d\right\}\\ &= \frac{1}{3}H_t(\theta_1).\end{aligned}$$

For deriving the second relationship in Lemma C.5, given that θ_1 and θ_2 possess symmetrical roles in $\int_{v=0}^{1}\dot{p}_\tau(i|A_{\tau k},v\theta_2+(1-v)\theta_1)dv$, we substitute two parameters and can get the result. □

Lemma C.6 (Restatement of Lemma 5.1)**.** *Suppose $\theta^*\in B_t(\delta)$. Then, for any $\theta\in B_t(\delta)$, we have*

$$||\theta-\theta^*||_{H_t}\leq 6\gamma_t(\delta). \tag{27}$$

where the confidence radius $\gamma_t(\delta):=\frac{3\sqrt{\lambda_t}}{2}+\frac{2}{\sqrt{\lambda_t}}\log\left(\frac{(\lambda_t+tKM/d)^{d/2}\lambda_t^{-d/2}}{\delta}\right)+\frac{2d}{\sqrt{\lambda_t}}\log 2.$

Proof.

$$\begin{aligned}||\theta-\theta^*||_{H_t} &\leq \sqrt{3}||\theta-\theta^*||_{\mathbb{G}_t(\theta,\theta^*)}\\ &= \sqrt{3}||g_t(\theta^*)-g_t(\theta)||_{\mathbb{G}_t^{-1}(\theta,\theta^*)}\\ &\leq \sqrt{3}\left\{||g_t(\theta^*)-g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta,\theta^*)}+||g_t(\hat{\theta}_t)-g_t(\theta)||_{\mathbb{G}_t^{-1}(\theta,\theta^*)}\right\}\\ &\leq \sqrt{3}\left\{\sqrt{3}||g_t(\theta^*)-g_t(\hat{\theta}_t)||_{H_t^{-1}}+\sqrt{3}||g_t(\hat{\theta}_t)-g_t(\theta)||_{H_t^{-1}}\right\}\\ &= 3\left\{||g_t(\theta^*)-g_t(\hat{\theta}_t)||_{H_t^{-1}}+||g_t(\hat{\theta}_t)-g_t(\theta)||_{H_t^{-1}}\right\}\\ &= 6\gamma_t(\delta).\end{aligned}$$

The first and third inequality is from Lemma C.5. The first equality is from Lemma C.1. From triangle inequality, we obtain the second inequality. Since $\theta\in B_t(\delta)$, the last equality holds. □

The above lemma shows the upper bound on the difference between an arbitrary θ and θ^*, and is pivotal for determining the overall regret bound.

Lemma C.7. *For the cascade S_t chosen by* `UCB-CCA+` *and any $\theta_t \in B_t(\delta)$, the following holds with probability at least $1-\delta$:*

$$f(S_t,\theta_t) - f(S_t,\theta^*) \le 36\,\gamma_t(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \sqrt{\dot{p}_t(i|A_{tk},\theta^*)}||x_{ti}||_{H_{t-1}^{-1}} + \frac{216}{\kappa}\gamma_t^2(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} ||x_{ti}||^2_{V_{t-1}^{-1}}$$

Proof. We begin by performing a second Taylor expansion of the expected reward function:

$$\begin{aligned}
&f(S_t,\theta_t) - f(S_t,\theta^*)\\
&= \nabla_\theta f(S_t,\theta^*)(\theta_t-\theta^*) + (\theta_t-\theta^*)^\top \left[\int_{v=0}^1 (1-v)\nabla_\theta^2 f\left(S_t,\bar{\theta}\right)dv\right](\theta_t-\theta^*)\\
&\le \sum_{A_{tk}\in S_t}\sum_{i\in A_{tk}} \left\{ \prod_{\substack{A_{t\dot{k}}\in\\ S_t\setminus\{A_{tk}\}}} p_t(i_0|A_{t\dot{k}},\theta^*)\right\} \dot{p}_t(i|A_{tk},\theta^*) x_{ti}^\top(\theta_t-\theta^*)\\
&+ \int_{v=0}^1 (1-v)\sum_{A_{tk}\in S_t}\sum_{i\in A_{tk}} \left\{ \prod_{\substack{A_{t\dot{k}}\in\\ S_t\setminus\{A_{tk}\}}} p_t(i_0|A_{t\dot{k}},\bar{\theta})\right\} \ddot{p}_t(i|A_{tk},\bar{\theta})dv(x_{ti}^\top(\theta_t-\theta^*))^2 \qquad (28)
\end{aligned}$$

where $\bar{\theta} := \theta^* + v(\theta_t-\theta^*)$.

Then, we bound $\sum_{A_{tk}\in S_t}\sum_{i\in A_{tk}}\left\{\prod_{\substack{A_{t\dot{k}}\in\\ S_t\setminus\{A_{tk}\}}} p_t(i_0|A_{t\dot{k}},\theta^*)\right\}$ as follows:

$$\begin{aligned}
\sum_{A_{tk}\in S_t}\sum_{i\in A_{tk}} \left\{ \prod_{\substack{A_{t\dot{k}}\in\\ S_t\setminus\{A_{tk}\}}} p_t(i_0|A_{t\dot{k}},\theta^*)\right\} &\le \sum_{A_{tk}\in S_t}\sum_{i\in A_{tk}} \left\{\frac{1}{1+M\exp(-1)}\right\}^{K-1}\\
&= KM\left\{\frac{1}{1+M\exp(-1)}\right\}^{K-1}\\
&= K\left\{\frac{1}{1+M\exp(-1)}\right\}^{K-2}\cdot M\left\{\frac{1}{1+M\exp(-1)}\right\}\\
&\le K\left\{\frac{1}{1+\exp(-1)}\right\}^{K-2}\cdot \exp(1)\\
&\le \frac{(1+\exp(1))^2}{\exp(1)^3(\log(1+\exp(1))-1)}\cdot\exp(1)\\
&< \frac{(1+\exp(1))^2}{\exp(1)^2(\log(1+\exp(1))-1)} < 6. \qquad (29)
\end{aligned}$$

The first inequality is due to Assumption 4.1 and the definition of $p_t(i_0|A_{t\dot{k}},\theta^*)$. Since $K\left\{\frac{1}{1+\exp(-1)}\right\}^{K-2}$ is maximized as $\frac{(1+\exp(1))^2}{\exp(1)^3(\log(1+\exp(1))-1)}$ at $K = \frac{1}{\log(1+\exp)-1} \approx 3.19$, we can derive the third inequality. We apply Equation (29) to Equation (28) as follows:

$$\begin{aligned}
f(S_t,\theta_t) - f(S_t,\theta^*) &\le 6 \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \dot{p}_t(i|A_{tk},\theta^*) x_{ti}^\top(\theta_t-\theta^*) \\
&\quad + 6 \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \left[\int_{v=0}^{1} (1-v)\ddot{p}_t(i|A_{tk},\bar{\theta})dv\right] (x_{ti}^\top(\theta_t-\theta^*))^2 \\
&\le 6 \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \dot{p}_t(i|A_{tk},\theta^*) ||x_{ti}||_{H_{t-1}^{-1}} ||\theta_t-\theta^*||_{H_{t-1}} \\
&\quad + 6 \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \left[\int_{v=0}^{1} (1-v)\ddot{p}_t(i|A_{tk},\bar{\theta})dv\right] ||x_{ti}||^2_{H_{t-1}^{-1}} ||\theta_t-\theta^*||^2_{H_{t-1}} \\
&\le 36\, \gamma_t(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \dot{p}_t(i|A_{tk},\theta^*) ||x_{ti}||_{H_{t-1}^{-1}} + 216\, \gamma_t^2(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} ||x_{ti}||^2_{H_{t-1}^{-1}} \\
&\le 36\, \gamma_t(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} \sqrt{\dot{p}_t(i|A_{tk},\theta^*)} ||x_{ti}||_{H_{t-1}^{-1}} + \frac{216}{\kappa} \gamma_t^2(\delta) \max_{\substack{A_{tk}\in S_t\\ i\in A_{tk}}} ||x_{ti}||^2_{V_{t-1}^{-1}}
\end{aligned}$$

□

Lemma C.8 (Elliptical potential with local information)**.**

$$\sum_{t=1}^{T} \max_{\substack{k\in[O_t]\\ i\in A_{tk}}} ||\sqrt{\dot{p}_t(i|A_{tk},\theta^*)} x_{ti}||^2_{H_{t-1}^{-1}} \le 2d \log\left(1 + \frac{KMT}{d\lambda}\right)$$

Proof. We begin by demonstrating that the weighted ℓ_2 norm of the feature vector combined with the local information inside the above max operator is upper bounded by 1. For convenience, $\sqrt{\dot{p}_t(i|A_{tk},\theta^*)}x_{ti}$ is denoted by $\tilde{x}_{ti}$. So, we rewrite $H_t \coloneqq \sum_{\tau=1}^{t}\sum_{k\in[O_\tau]}\sum_{i\in A_{\tau k}} \tilde{x}_{\tau i}\tilde{x}_{\tau i}^\top + \lambda_t I_d$. Let $\lambda_{\min}(H_t)$ be the minimum eigenvalue of H_t. Since $\lambda_t \ge 1$ and $||\tilde{x}_{ti}||^2_{H_{t-1}^{-1}} \le \frac{||\tilde{x}_{ti}||_2^2}{\lambda_{\min}(H_{t-1})} \le \frac{1}{\lambda_t}$, we have:

$$\max_{k\in[O_t]} \max_{i\in A_{tk}} ||\tilde{x}_{ti}||^2_{H_{t-1}^{-1}} \le 1.$$

Using the fact that $z \le 2\ln(1+z)$ for any $z \in [0,1]$, we have

$$\begin{aligned}
\sum_{\tau=1}^{t} \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||\tilde{x}_{\tau i}||^2_{H_{\tau-1}^{-1}} &\le 2\sum_{\tau=1}^{t} \ln\left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||\tilde{x}_{\tau i}||^2_{H_{\tau-1}^{-1}}\right) \\
&= 2\ln \prod_{\tau=1}^{t} \left(1 + \max_{k\in[O_\tau]} \max_{i\in A_{\tau k}} ||\tilde{x}_{\tau i}||^2_{H_{\tau-1}^{-1}}\right). \qquad (30)
\end{aligned}$$

Now we upper bound $\prod_{\tau=1}^{t}\left(1+\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||\tilde{x}_{\tau i}||^2_{H^{-1}_{\tau-1}}\right)$ from $\det(H_t)$.

$$\begin{aligned}
\det(H_t) &= \det\left(H_{t-1}+\sum_{k=1}^{O_t}\sum_{i\in A_{tk}}\tilde{x}_{ti}\tilde{x}_{ti}^\top\right)\\
&= \det(H_{t-1})\det\left(I+H_{t-1}^{-1/2}\sum_{k=1}^{O_t}\sum_{i\in A_{tk}}\tilde{x}_{ti}\tilde{x}_{ti}^\top H_{t-1}^{-1/2}\right)\\
&= \det(H_{t-1})\det\left(I+\sum_{k=1}^{O_t}\sum_{i\in A_{tk}}\left(H_{t-1}^{-1/2}\tilde{x}_{ti}\right)\left(H_{t-1}^{-1/2}\tilde{x}_{ti}\right)^\top\right)\\
&\geq \det(H_{t-1})\left(1+\sum_{k=1}^{O_t}\sum_{i\in A_{tk}}||\tilde{x}_{ti}||^2_{H^{-1}_{t-1}}\right)\\
&\geq \det(\lambda I)\prod_{\tau=1}^{t}\left(1+\sum_{k=1}^{O_\tau}\sum_{i\in A_{\tau k}}||\tilde{x}_{\tau i}||^2_{H^{-1}_{\tau-1}}\right)\\
&\geq \det(\lambda I)\prod_{\tau=1}^{t}\left(1+\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||\tilde{x}_{\tau i}||^2_{H^{-1}_{\tau-1}}\right).
\end{aligned}$$

The second equality above is from the fact that $V+U=V^{1/2}(I+V^{-1/2}UV^{-1/2})V^{1/2}$ for a symmetric positive definite matrix V. The first inequality above can be obtained by applying Lemma B.7. Applying the first inequality repeatedly, we can get the second inequality above. Thus, we have

$$\prod_{\tau=1}^{t}\left(1+\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||\tilde{x}_{\tau i}||^2_{H^{-1}_{\tau-1}}\right)\leq\frac{\det(H_t)}{\det(\lambda I)}. \tag{31}$$

Then applying Eq.(31) to Eq.(30), we complete the proof as follows:

$$\sum_{\tau=1}^{t}\max_{k\in[O_\tau]}\max_{i\in A_{\tau k}}||\tilde{x}_{\tau i}||^2_{H^{-1}_{\tau-1}}\leq 2\ln\frac{\det(H_t)}{\det(\lambda I)}\leq 2\ln\frac{\det(H_t)}{\lambda^d}\leq 2\ln\frac{\det(V_t)}{\lambda^d}\leq 2d\log\left(1+\frac{KMT}{d\lambda}\right)$$

where the last inequality is from Lemma B.9. □

Proof of Theorem 5.2.

$$
\begin{aligned}
&\mathcal{R}^{\alpha}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mathcal{R}^{\alpha}(t, S_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} f(S_t, \theta_t) - f(S_t, \theta^*)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} 36\, \gamma_t(\delta) \max_{\substack{A_{tk}\in S_t \\ i\in A_{tk}}} \sqrt{\dot{p}_t(i|A_{tk}, \theta^*)}||x_{ti}||_{H_{t-1}^{-1}} + \sum_{t=1}^{T} \frac{216}{\kappa}\gamma_t^2(\delta) \max_{\substack{A_{tk}\in S_t \\ i\in A_{tk}}} ||x_{ti}||^2_{V_{t-1}^{-1}}\right] \quad (32) \\
&\leq \mathbb{E}\left[36\, \gamma_T(\delta) \sum_{t=1}^{T} \max_{\substack{A_{tk}\in S_t \\ i\in A_{tk}}} ||\sqrt{\dot{p}_t(i|A_{tk}, \theta^*)}x_{ti}||_{H_{t-1}^{-1}} + \frac{216}{\kappa}\gamma_T^2(\delta) \sum_{t=1}^{T} \max_{\substack{A_{tk}\in S_t \\ i\in A_{tk}}} ||x_{ti}||^2_{V_t^{-1}}\right] \\
&\leq \mathbb{E}\left[36\, \gamma_T(\delta) \sum_{t=1}^{T} \max_{\substack{k\in[O_t] \\ i\in A_{tk}}} ||\sqrt{\dot{p}_t(i|A_{tk}, \theta^*)}x_{ti}||_{H_{t-1}^{-1}}\right] \quad (33) \\
&\quad + \mathbb{E}\left[\frac{216}{\kappa p_t(i_0|A_{t1}, \theta^*)}\gamma_T^2(\delta) \sum_{t=1}^{T} \max_{\substack{k\in[O_t] \\ i\in A_{tk}}} ||x_{ti}||^2_{V_{t-1}^{-1}}\right] \quad (34) \\
&\leq \mathbb{E}\left[36\, \gamma_T(\delta) \sqrt{T \cdot \sum_{t=1}^{T} \max_{\substack{k\in[O_t] \\ i\in A_{tk}}} ||\sqrt{\dot{p}_t(i|A_{tk}, \theta^*)}x_{ti}||^2_{H_{t-1}^{-1}}}\right] \quad (35) \\
&\quad + \mathbb{E}\left[\frac{216(1+Me)}{\kappa}\gamma_T^2(\delta) \sum_{t=1}^{T} \max_{\substack{k\in[O_t] \\ i\in A_{tk}}} ||x_{ti}||^2_{V_t^{-1}}\right] \quad (36) \\
&\leq 36\gamma_T(\delta)\sqrt{T \cdot 2d\log\left(1 + \frac{KMT}{d\lambda_t}\right)} + \frac{216(1+Me)}{\kappa}\gamma_T^2(\delta)\left(d\log\left(1 + \frac{KMT}{d\lambda_t}\right)\right) \quad (37)
\end{aligned}
$$

The inequality (32) is from Lemma C.7. The inequality (34) can be obtained by *doubly optimistic exposure swapping* (see Section 5.1.2 and max operation). The Inequality (36) is from Cauchy-Schwarz inequality. The last inequality is from Lemma C.8, Lemma B.8 and Lemma B.9. ☐

D Convex Relaxation

In `UCB-CCA+`'s optimization step (see line 4 in Algorithm 2), it is particularly challenging to solve due to the confidence set $B_t(\delta)$ in Eq.(6) being a non-convex set. [2; 3] address this problem by designing a convex relaxation for the set $B_t(\delta)$ in simple logistic bandits and MNL bandits, respectively, and we extend this idea to our model. The following confidence set is a convex relation set for $B_t(\delta)$:

$$E_t(\delta) \coloneqq \left\{\theta \in \mathbb{R}^d : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \zeta_t^2(\delta)\right\}$$

where $\zeta_t(\delta) \coloneqq \gamma_t(\delta) + \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}}$ and $\gamma_t(\delta) \coloneqq \frac{3\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}}\log\left(\frac{(\lambda_t+KMt/d)^{d/2}\lambda_t^{-d/2}}{\delta}\right) + \frac{2d}{\sqrt{\lambda_t}}\log 2$. We exploit $E_t(\delta)$ instead of the original confidence set $B_t(\delta)$ for numerical experiments in Section 7. In this section, we justify our strategy of relaxing the confidence set based on the following lemmas.

Lemma D.1. $\forall t \geq 1, E_t(\delta) \supseteq B_t(\delta)$, *therefore* $\mathbb{P}(\forall t \geq 1, \theta^* \in E_t(\delta)) \geq 1 - \delta$.

Proof. We start by performing the second-order Taylor expansion of the log-likelihood function with respect to $\hat{\theta}_t$ as follows:

$$\begin{aligned}
&\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \\
&\quad = \nabla \mathcal{L}_t(\hat{\theta}_t)^\top (\theta - \hat{\theta}_t) + (\theta - \hat{\theta}_t)^\top \left(\int_{v=0}^{1} (1-v) \nabla^2 \mathcal{L}_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t)) dv \right) (\theta - \hat{\theta}_t) \quad (38) \\
&\quad = (\theta - \hat{\theta}_t)^\top \left(\int_{v=0}^{1} (1-v) \nabla^2 \mathcal{L}_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t)) dv \right) (\theta - \hat{\theta}_t) \\
&\quad = (\theta - \hat{\theta}_t)^\top \left(\int_{v=0}^{1} (1-v) H_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t)) dv \right) (\theta - \hat{\theta}_t) \quad (39) \\
&\quad \le (\theta - \hat{\theta}_t)^\top \mathbb{G}_t(\hat{\theta}_t - \theta)(\theta - \hat{\theta}_t) \quad (40) \\
&\quad = ||\theta - \hat{\theta}_t||^2_{\mathbb{G}_t(\hat{\theta}_t, \theta)} \\
&\quad = ||g_t(\theta) - g_t(\hat{\theta}_t)||^2_{\mathbb{G}_t^{-1}(\hat{\theta}_t, \theta)} \quad (41) \\
&\quad = ||g_t(\theta) - g_t(\hat{\theta}_t)||^2_{\mathbb{G}_t^{-1}(\theta, \hat{\theta}_t)}. \quad (42)
\end{aligned}$$

Equation (38) from the fact that $\nabla \mathcal{L}_t(\hat{\theta}_t) = 0$. Equation (39) is due to $\nabla^2 \mathcal{L}_t(\theta) = H_t(\theta)$. Equation (40) is derived through the following process:

$$\begin{aligned}
&\int_{v=0}^{1} (1-v) H_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t)) dv \\
&\quad = \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left(\int_{v=0}^{1} (1-v) \dot{p}_\tau(i | A_{\tau k}, \hat{\theta}_t + v(\theta - \hat{\theta}_t) dv \right) x_{\tau i} x_{\tau i}^\top + \lambda_\tau I_d \\
&\quad \preceq \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i \in A_{\tau k}} \left(\int_{v=0}^{1} \dot{p}_\tau(i | A_{\tau k}, \hat{\theta}_t + v(\theta - \hat{\theta}_t) dv \right) x_{\tau i} x_{\tau i}^\top + \lambda_\tau I_d \quad (43) \\
&\quad = \mathbb{G}_t(\hat{\theta}_t - \theta) \quad (44)
\end{aligned}$$

Equation (43) holds because $p_t(i|A, \theta,)$ is a strictly increasing function with respect to θ ($\dot{p}_t(i|A, \theta,)$). And Equation (44) is from the definition of $\mathbb{G}_t(\theta_1 - \theta_2)$ as defined in Lemma C.5. Equation (41) is obtained by applying Lemma C.1 Equation (42) is from the fact that $\mathbb{G}_t^{-1}(\hat{\theta}_t, \theta) = \mathbb{G}_t^{-1}(\theta, \hat{\theta}_t)$. Now, we apply Lemma D.2 to Equation (42), then we have:

$$\mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \le \left\{ \gamma_t(\delta) + \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}} \right\}^2 = \zeta_t^2(\delta)$$

for any $\theta \in B_t(\delta)$. This shows that if $\theta \in B_t(\delta)$, then $\theta \in E_t(\delta)$ and thus $B_t(\delta) \subset E_t(\delta)$. $\square$

Lemma D.2. *Let $\delta \in (0, 1]$. For all $\theta \in B_t(\delta)$:*

$$||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta, \hat{\theta}_t)} \le \gamma_t(\delta) + \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}} = \zeta_t(\delta)$$

Proof.

$$\begin{aligned}
\mathbb{G}_t(\theta, \hat{\theta}_t) &= \sum_{\tau=1}^{t1} \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \alpha_i(A_{tk}, \hat{\theta}_t, \theta) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d \\
&\geq \sum_{\tau=1}^{t1} \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta)(1 + |x_{\tau i}^\top \theta - x_{\tau i}^\top \hat{\theta}_t|)^{-1} x_{\tau i} x_{\tau i}^\top + \lambda_t I_d \\
&\geq \sum_{\tau=1}^{t1} \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta)(1 + ||x_{\tau i}||_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)} ||\theta - \hat{\theta}_t||_{\mathbb{G}_t(\theta,\hat{\theta}_t)})^{-1} x_{\tau i} x_{\tau i}^\top + \lambda_t I_d \\
&\geq (1 + \lambda_t^{-1/2} ||\theta - \hat{\theta}_t||_{\mathbb{G}_t(\theta,\hat{\theta}_t)})^{-1} \left(\sum_{\tau=1}^{t} \sum_{k\in[O_\tau]} \sum_{i\in A_{\tau k}} \dot{p}_\tau(i|A_{\tau k}, \theta) x_{\tau i} x_{\tau i}^\top + \lambda_t I_d \right) \\
&= (1 + \lambda_t^{-1/2} ||\theta - \hat{\theta}_t||_{\mathbb{G}_t(\theta,\hat{\theta}_t)})^{-1} H_t(\theta) \\
&= (1 + \lambda_t^{-1/2} ||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)})^{-1} H_t(\theta).
\end{aligned}$$

The first inequality is from Lemma C.4 and the second inequality is obtained by applying Cauchy-Schwarz inequality.

This inequality gives:

$$\begin{aligned}
||g_t(\theta) - g_t(\hat{\theta}_t)||^2_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)} &\leq \left\{ 1 + \lambda_t^{-1/2} ||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)} \right\} ||g_t(\theta) - g_t(\hat{\theta}_t)||^2_{H_t(\theta)^{-1}} \\
&\leq \lambda_t^{-1/2} \gamma_t^2(\delta) ||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)} + \gamma_t^2(\delta).
\end{aligned}$$

Resolving this polynomial inequality with respect to $||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbb{G}_t^{-1}(\theta,\hat{\theta}_t)}$ by using the fact that $x^2 \leq bx + c \Rightarrow x \leq b + \sqrt{c}$ where $x \in \mathbb{R}$ and $b, c \in \mathbb{R}+$ (See Proposition 7 in Abeille et al. [2]).,
then we get the result. □

Lemma D.3. *Suppose $\theta^* \in B_t(\delta)$, the following holds for all $\theta \in E_t(\delta)$:*

$$||\theta - \theta^*||_{H_t} \leq 4\gamma_t(\delta) + 2\sqrt{2}\zeta_t(\delta).$$

Proof. We start by performing the second-order Taylor expansion of the log-likelihood function with respect to θ^* as follows:

$$\begin{aligned}
&\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta^*) \\
&\quad = \nabla \mathcal{L}_t(\theta^*)^\top (\theta - \theta^*) + (\theta - \theta^*)^\top \left(\int_{v=0}^1 (1 - v) \nabla^2 \mathcal{L}_t(\theta^* + v(\theta - \theta^*)) dv \right) (\theta - \theta^*) \\
&\quad = \nabla \mathcal{L}_t(\theta^*)^\top (\theta - \theta^*) + ||\theta - \theta^*||^2_{\tilde{\mathbb{G}}_t(\theta^*,\theta)} \\
&\quad \geq \nabla \mathcal{L}_t(\theta^*)^\top (\theta - \theta^*) + \frac{1}{4} ||\theta - \theta^*||^2_{H_t}
\end{aligned}$$

where $\tilde{\mathbb{G}}_t(\theta^*, \theta) = (\theta - \theta^*)^\top \left(\int_{v=0}^1 (1 - v) H_t(\theta^* + v(\theta - \theta^*)) dv \right) (\theta - \theta^*)$. The last inequality is from applying Lemma 8 in Abeille et al. [2]. Thus, we have:

$$\begin{aligned}
||\theta - \theta^*||^2_{H_t} &\leq 4|\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta^*)| + 4|\nabla \mathcal{L}_t(\theta^*)^\top (\theta - \theta^*)| \\
&\leq 8\zeta_t(\delta)^2 + 4|\nabla \mathcal{L}_t(\theta^*)^\top (\theta - \theta^*)| \\
&\leq 8\zeta_t(\delta)^2 + 4||\nabla \mathcal{L}_t(\theta^*)||_{H_t^{-1}} ||\theta - \theta^*||_{H_t} \\
&\leq 8\zeta_t(\delta)^2 + 4\gamma_t(\delta) ||\theta - \theta^*||_{H_t}. \qquad (45)
\end{aligned}$$

The above second inequality is due to $\theta, \theta^* \in E_t(\delta)$. The third inequality is by applying Cachy-Schwarz inequality. And Equation (45) holds from the following inequality:

$$\begin{aligned}
||\nabla \mathcal{L}_t(\theta^*)||_{H_t^{-1}} &= ||g_t(\theta^*) - \sum_{\tau=1}^{t} \sum_{k=1}^{O_\tau} \sum_{i_m \in A_{\tau k}} y_{\tau k m} x_{\tau i_m} ||_{H_t^{-1}} \qquad (46) \\
&= ||g_t(\theta^*) - g_t(\hat{\theta}_t)||_{H_t^{-1}} \leq \gamma_t(\delta). \qquad (47)
\end{aligned}$$

Equation (46) is from the definition of $\nabla\mathcal{L}_t(\theta^*)$ and $g_t(\theta^*)$. Equation (47) is from Equation (22). In conclusion, Equation (45) is a polynomial inequality in terms of $||\theta-\theta^*||_{H_t}$. Solving it yields the following result:

$$||\theta-\theta^*||_{H_t} \le 4\gamma_t(\delta)+2\sqrt{2}\zeta_t(\delta).$$

since $x^2 \le bx+c \Rightarrow x \le b+\sqrt{c}$ where $x\in\mathbb{R}$ and $b,c\in\mathbb{R}+$ (See Proposition 7 in Abeille et al. [2]).

Lastly, we set the penalty parameter $\lambda_t=\mathcal{O}(d\log(tKM))$, then we get the followings:

$$\gamma_t(\delta)=\mathcal{O}(d\log(tKM)),$$

$$\zeta_t(\delta)=\gamma_t(\delta)+\frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}}=\mathcal{O}(d\log(tKM)).$$

This implies that the following holds with probability at least $1-\delta$:

$$||\theta-\theta^*||_{H_t}=\mathcal{O}(d\log(tKM))$$

for any $\theta\in E_t(\delta)$. □

E 0.5 Approximation for Cascading Assortment Optimization

E.1 Proof of 0.5 Approximation Ratio

Let $\phi(i,A)=\frac{w_i}{1+w(A)}$, where $w(A)=\sum_{i\in A}w_i$, which we use throughout this section. Recall that N is the number of items in the ground set, and K and M denote the length of a cascade and assortment, respectively. Without loss of generality, $KM\le N$ (if not, add dummy items with MNL weight 0).

For $M=1$, we show that the problem is easy to solve optimally – simply pick the K highest probability items and show them in arbitrary order.

Lemma E.1. *For general M, the optimization problem is weakly NP-hard even for $K=2$.*

Proof. We use the hardness of unconstrained cascade optimization shown in [19] (Theorem 1). Given an instance of the unconstrained problem with $K=2$ and ground set $[N]$, consider an instance of our cardinality constrained problem with $M=N$ over expanded ground set $[N]\cup[N]_0$ where $[N]_0$ consists of $|N|$ dummy elements that each have MNL weight parameter 0. □

Lemma E.2. *For any M, given a collection of assortments $\{A_1,\cdots,A_K\}$ with success probabilities $\{p_1,\cdots,p_K\}$, their order of display does not matter. Further, for every permutation $\rho:[K]\to[K]$, we have,*

$$\sum_{k\in[K]}p_k\prod_{\dot{k}<k}(1-p_{\dot{k}})=1-\prod_{k\in[K]}(1-p_k)=\sum_{k\in[K]}p_{\rho^{-1}(k)}\prod_{\dot{k}<k}(1-p_{\rho^{-1}(\dot{k})}).$$

Proof. If the customer views an assortment A_k, a success occurs in this assortment independently with probability p_k. We can (independently) pre-sample these Bernoulli random variables for each assortment. Then, the sequence in which assortments are shown does not matter since each ordering leads to the same end result (success of failure) once the Bernoulli variables are fixed. Algebraically, the probability that at least one of these random variables succeeds is given by $1-\prod_{k\in[K]}(1-p_k)$. An alternative way to compute these probabilities is to examine the random variables one by one until a success is found. If we examine the random variables in the order given by ρ, we get an alternative expression for the probability that at least one of the random variables succeeds, given by, $\sum_{i\in[K]}p_{\rho^{-1}(k)}\prod_{\dot{k}<k}(1-p_{\rho^{-1}(\dot{k})})$. This completes the proof. □

Lemma E.3. *Let $\{A_k^*\}_{k\in[K]}$ denote the optimal solution. Then, $\cup_{i\in[K]}A_k^*$ is the set of KM items with highest value of MNL weights.*

Proof. Suppose not. Then, there is an item i in some assortment A_k^* and an item j that is not in any assortment such that $w_j>w_i$. Consider the assortment $A'_k=A_k^*\cup\{j\}\backslash\{i\}$. The probability of click is strictly higher in assortment A'_k than in assortment A_k^*. Therefore, is we keep all other assortment as is but replace assortment A_k^* with A'_k, we have a strictly better solution, contradiction. □

For $M = 1$, combining Lemma E.2 with Lemma E.3 shows that showing the M highest probability items is optimal.

Now, order the items in $[N]$ in decreasing order of MNL weights and consider the following assortment for general M.

$$D_1 = \{1, 2, \cdots, M\}, D_2 = \{M+1, M+2, \cdots, 2M\} \cdots, D_K = \{(K-1)M + 1, \cdots, KM\}.$$

Let OPT denote the overall click probability in the optimal solution.

Lemma E.4. *When $w(D_1) < 1$, we have,*

$$\text{OPT} \leq w(D_1) + w(D_2)\left(1 - w(D_1)\right) + \cdots + w(D_K) \prod_{k \in [K-1]} (1 - w(D_k)).$$

Proof. Given $w_1 + \cdots + w_M < 1$, we have. $w(D_k) < 1 \; \forall k \in [K]$. In fact,

$$\phi(A) := \sum_{i \in A} \phi(i, A) = \sum_{i \in A} \frac{w_i}{1 + w(A)} \leq w(A) \qquad \forall A \subseteq [N], |A| \leq M.$$

Given optimal solution $\{A_k^*\}_{k \in [K]}$, consider a hypothetical solution where for every $k \in [K]$, given that assortment k is shown, the (independent) probability of click in assortment k is $h_k := w(A_k^*)$ $(\geq \phi(A_k^*))$. We claim that this hypothetical solution, say H, has expected click probability at least as much as OPT. To see this, we couple the Bernoulli random variable for each assortment in H with the corresponding assortment in OPT so that whenever there is a success in assortment k in OPT, there is also a success in slab k in H (this is possible since $h_k \geq \phi(A_k^*)$). Thus,

$$H \geq \text{OPT}.$$

Now, it suffices to show that,

$$H \leq w(D_1) + w(D_2)(1 - w(D_1)) + \cdots + w(D_K) \prod_{k \in [K-1]} (1 - w(D_k)).$$

Observe that the RHS corresponds to a hypothetical solution with assortments $\{D_k\}_{k \in [K]}$ and success probabilities $\{w(D_k)\}_{k \in [K]}$ (instead of $\{\phi(D_k)\}_{k \in [K]}$). We now focus on the hypothetical scenario where the success probability of an assortment A equals $w(A)$ (instead of $\phi(A)$). We show that in this scenario, the assortments $\{D_k\}_{k \in [K]}$ are optimal and this proves the main claim. We proceed by setting up a contradiction. Suppose that assortments $\{D_k\}_{k \in [K]}$ are sub-optimal and consider the optimal partition of $\cup_{k \in [K]} D_k$ into assortments $\{E_k\}_{k \in [K]}$. We have,

$$1 - \prod_{k \in [K]} (1 - w(E_k)) > 1 - \prod_{k \in [K]} (1 - w(D_k)).$$

Since assortments $\{E_k\}_{k \in [K]}$ are distinct from $\{D_k\}_{k \in [K]}$, there exists assortments E_l and E_n such that $w(E_l) > w(E_n)$ but $w_i < w_j$ for items $i \in E_l, j \in E_n$. Consider a new set of assortments $\{F_k\}_{k \in [K]}$ defined as follows,

$$F_k = \begin{cases} E_k & \forall k \in [K] \backslash \{l, n\}, \\ E_l \cup \{j\} \backslash \{i\} & k = l, \\ E_n \cup \{i\} \backslash \{j\} & k = n. \end{cases}$$

Observe that assortment F_l has success probability $w(F_l) = w(E_l) + w_j - w_i$. Similarly, assortment F_n has success probability $w(F_n) = w(E_n) - w_j + w_i$. Using Lemma E.2, let assortments E_l, E_n and assortments F_l, F_n be the last two assortments shown in their respective sequences. Then, the following inequalities show that the overall probability of success in $\{F_k\}_{k \in [K]}$ is strictly higher than the optimal partition $\{E_k\}_{k \in [K]}$, contradicting the optimality of $\{E_k\}_{k \in [K]}$.

$$\begin{aligned} 1 - (1 - w(F_l))(1 - w(F_n)) &= w(F_l) + w(F_n) - w(F_l)w(F_n), \\ &= w(E_l) + w(E_n) - (w(E_l) + w_j - w_i)(w(E_n) - w_j + w_i), \\ &> 1 - (1 - w(E_l))(1 - w(E_n)), \end{aligned}$$

here the last inequality follows from the fact that $w_j > w_i$ and $w(E_l) > w(E_n)$. □

Lemma E.5. *The assortments $D_1, \cdots, D_K$, shown in any order, have overall click probability at least* $0.5\,\mathrm{OPT}$.

Proof. **Case 1:** Let $w_1 + w_2 + \cdots + w_M \geq 1$. Using Lemma E.2, we can show assortment D_1 first without loss of generality. Then, the probability of click in assortment D_1 is at least 0.5. Since probability of click overall is at least as much as the probability of click in D_1, we are done.

Case 2: Let $w_1 + \cdots + w_M < 1$. In this case, $1 - \sum_{i \in D_k} \phi(i, D_k) \leq 1 - 0.5 w(D_k)$ for all $k \in [K]$, and the overall click probability is at least,

$$1 - \prod_{k \in [K]} (1 - 0.5\, w(D_k)).$$

Now, observe that,

$$1 - \prod_{k \in [K]} (1 - 0.5\, w(D_k)) = 0.5\, w(D_1) + \cdots + 0.5\, w(D_K) \prod_{k \in [K-1]} (1 - 0.5\, w(D_k)).$$

Comparing this term by term with the upper bound on OPT in Lemma E.4 completes the proof. □

Remark: While the order of assortments does not matter here, if the customer was impatient and left early with some probability then the order would matter. In fact, in that setting it can be shown that displaying the slabs in the natural order $A_1, A_2, \cdots$ is 0.5 approximate.

F Limitations

While we study a more general version of combinatorial bandits, for the choice model, we adapt the MNL model. The MNL model is certainly one of the most popular options for modeling the outcomes of multi-class classification problems and certainly a practical and suitable extension of a simple linear model. However, it does have some drawbacks. For example, the MNL model relies on the Independence of Irrelevant Alternatives (IIA) assumption, and the utility functions in an MNL model are linear in parameters. In future work, we plan to address these challenges and extend to a more flexible choice model. However, for this work, we strongly believe that the current new model, proposed algorithms, and the regret analysis based on this newly proposed model provide more than sufficient contributions.

Consider MNL assortment bandits. Suppose a weight $w = w_1, \ldots, w_K \in \mathbb{R}^d$. For any given weight w, the expected reward function is as follows:

$$R(w) = \frac{\sum_{k \in [K]} \exp(w_k)}{1 + \sum_{k' \in [K]} \exp(w'_k)}.$$

At round t, the agent chooses the assortment $A_t = i_1, \ldots, i_K$. Let $w_t = x_{ti_1}^\top \theta_t, \ldots, x_{ti_K}^\top \theta_t$ and the optimal weight at round $w_t^* = x_{ti_1}^\top \theta^*, \ldots, x_{ti_K}^\top \theta^*$.

A second-order Taylor expansion gives that:

$$\begin{aligned}
R(w_t^*) &= R(w_t) + \nabla R(w_t)^\top (w_t^* - w_t) + \frac{1}{2}(w_t^* - w_t)^\top \nabla^2 R(\bar{w}_t)(w_t^* - w_t) \\
R(w_t^*) - R(w_t) &= +\nabla R(w_t)^\top (w_t^* - w_t) + \frac{1}{2}(w_t^* - w_t)^\top \nabla^2 R(\bar{w}_t)(w_t^* - w_t) \\
R(w_t) - R(w_t^*) &= -\nabla R(w_t)^\top (w_t^* - w_t) - \frac{1}{2}(w_t^* - w_t)^\top \nabla^2 R(\bar{w}_t)(w_t^* - w_t) \\
R(w_t) - R(w_t^*) &= \nabla R(w_t)^\top (w_t - w_t^*) - \frac{1}{2}(w_t^* - w_t)^\top \nabla^2 R(\bar{w}_t)(w_t^* - w_t) \qquad (48)
\end{aligned}$$

Thus, if we apply Equation (28), the cumulative regret can be represented as follows:

$$\sum_{t=1}^{T} \{R(w_t) - R(w_t^*)\} = \sum_{t=1}^{T} \nabla R(w_t)^\top (w_t - w_t^*) - \frac{1}{2}\sum_{t=1}^{T} (w_t^* - w_t)^\top \nabla^2 R(\bar{w}_t)(w_t^* - w_t)$$

We first consider the upper bound of $\sum_{t=1}^{T} \nabla R(w_t)^\top (w_t - w_t^*)$ in the right side.

$$\nabla R(w) = \begin{bmatrix} R_1(w) R_0(w) \\ \vdots \\ R_K(w) R_0(w) \end{bmatrix}$$

where $R_k(w) = \frac{\exp(w_k)}{1+\sum_{k' \in [K]} \exp(w'_k)}$ and $R_0(w) = \frac{1}{1+\sum_{k' \in [K]} \exp(w'_k)}$