Thompson Sampling for Multinomial Logit Contextual Bandits

Min-hwan Oh
Columbia University
New York, NY
m.oh@columbia.edu

Garud Iyengar
Columbia University
New York, NY
garud@ieor.columbia.edu

Abstract

We consider a dynamic assortment selection problem where the goal is to offer a sequence of assortments that maximizes the expected cumulative revenue, or alternatively, minimize the expected regret. The feedback here is the item that the user picks from the assortment. The distinguishing feature in this work is that this feedback is given by a multinomial logit choice model. The utility of each item is a dynamic function of contextual information of both the item and the user. We refer to this problem as the multinomial logit contextual bandit. We propose two Thompson sampling algorithms for this multinomial logit contextual bandit. Our first algorithm maintains a posterior distribution of the unknown parameter and establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$[1] Bayesian regret over T rounds with d dimensional context vector. The second algorithm approximates the posterior by a Gaussian distribution and uses a new optimistic sampling procedure to address the issues that arise in worst-case regret analysis. This algorithm achieves $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case (frequentist) regret bound. The numerical experiments show that the practical performance of both methods is in line with the theoretical guarantees.

1 Introduction

In the stochastic multi-armed bandit (MAB) problem [10, 27], the learning agent selects one of N actions (or items) and receives a revenue feedback corresponding to the chosen action in each round. The objective is to maximize the cumulative revenue over a finite horizon of length T, or alternatively, to minimize the cumulative regret defined as the difference in cumulative revenues of the optimal strategy and the agent's strategy. The main challenge in MAB problems is to appropriately balance the trade-off between exploitation, i.e., pulling the best empirical arm, and exploration, i.e., experimenting with arms which are not sufficiently pulled. The balancing strategies for this exploration-exploitation trade-off typically fall into two categories: upper confidence bound (UCB) methods [9, 18] and Thompson sampling (TS) based methods [42]. (Besides UCB and TS, one may also consider ϵ-greedy approach [24].)

UCB methods maintain a confidence set for the unknown true parameter, and in each step, choose the most optimistic parameter from this set, and pull the optimal arm corresponding to this optimistic parameter value. The confidence set is updated based on the revenue feedback which is revealed after an arm is pulled. TS assumes a prior distribution over the parameters defining the reward distribution. At each step, a parameter value is sampled from the posterior distribution, and an optimal arm corresponding to a sampled parameter is pulled. Upon observing the reward for each round, the posterior distribution is updated via Bayes rule. TS has been successfully applied in a wide range of settings [40, 13, 38].

[1] $\widetilde{\mathcal{O}}$ suppresses logarithmic dependence.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

While UCB algorithms have simple implementations and good theoretical regret bounds [29], TS has been shown to achieve better empirical performance in many simulated and real-world settings without sacrificing simplicity [13, 23]. In order to bridge this gap, many recent studies have been focused on the analysis of worst-case regret and Bayesian regret in TS approaches for both contextual bandits and reinforcement learning settings [5, 7, 38, 3]. The main technical difficulty in analyzing regret in the TS lies in controlling the deviation introduced by the randomness in the algorithm.

In this paper, we consider a dynamic assortment selection with contextual information, which is a combinatorial variant of the contextual bandit problem. The goal is to offer a sequence of assortments of at most K items from a set of N possible items that minimize regret. The feedback here is the particular item chosen by the user from the offered assortment. This problem arises in many real-world applications such as online retailing, streaming services, news feed, online advertising, etc. We assume that the item choice is given by a multinomial logit (MNL) choice model [33]. This is one of the most widely used models in dynamic assortment optimization literature [12, 37, 39, 6, 7, 14]. The utility of each item that defines the MNL choice probability is assumed to be a linear function of a d-dimensional contextual information, or a set of d features. This contextual information can be a combined information of *both* the item and the user, and is allowed to change over time.

The MNL contextual bandit is a multinomial generalization of generalized linear contextual bandits [23, 30], particularly logistic bandits, that reduces to generalized linear bandits when the assortment contains a single item. However, this extension is non-trivial since the MNL model cannot be expressed in the form of a generalized linear model [15]; hence, the results of generalized linear bandits do not directly apply. Also, in contrast to the standard contextual bandit problems, in the MNL contextual bandit, the item choice (feedback) is a function of the entire offered assortment. Thus, regret analysis is more complicated. Furthermore, we allow the context vector to vary arbitrarily in time; thus, offering the same assortment repeatedly several times to learn the parameter values [6, 7] is no longer an effective strategy.

We propose two Thompson sampling algorithms for this multinomial logit contextual bandit. To our knowledge, these are the first TS algorithms for this problem.

(a) The first algorithm maintains a posterior distribution of the true parameter and establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$ Bayesian regret.

(b) The second algorithm approximates the posterior by a Gaussian distribution and uses a new optimistic sampling procedure to address the issues that arise in worst-case regret analysis. We establish $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case (frequentist) regret bound for this algorithm.

The additional $\sqrt{d}$ factor in the regret of the second algorithm is due to the deviation from the random sampling in TS which is addressed in the worst-case regret analysis and is consistent with the results in TS methods for linear bandits [5, 3]. Both regret bounds are free of candidate item set size N, which implies that our TS algorithms can be applied to a large item set. The TS algorithms we propose are efficient to implement as long as the assortment optimization step is solved efficiently, for which our TS algorithms can exploit efficient polynomial-time algorithms [36, 20], which is a significant advantage over the previously proposed UCB method in [15] which computes the confidence bound for *each assortment* (i.e., for each of the total N choose K assortments). Furthermore, the numerical experiments show that the practical performance of the proposed methods is in line with the theoretical guarantees.

2 Related Work

The MNL model [34, 33, 32] is one of the most widely used choice models for assortment selection problems. The problem of computing the optimal assortment (*static* assortment optimization problem), when the MNL parameters, i.e., user preferences, are known a priori, is well-studied [41, 21, 22]. Our work belongs to the literature on *dynamic* assortment optimization. [12] consider the setting where the demand for items in an assortment is independent. [37] and [39] consider the problem of minimizing regret under the MNL choice model and present an "explore first then exploit later" approach. [37] showed $\mathcal{O}(N^2 \log^2 T)$ regret bound, where N is the number of total candidate items. [39] later improved the bound to $\mathcal{O}(N \log T)$. However, these methods require a priori knowledge of "separability" between the true optimal assortment and the other sub-optimal alternatives.

More recent work by [6, 7, 16, 14, 15] also incorporated MNL models into dynamic assortment optimization and formulated the problem into an online regret minimization problem without requiring a priori knowledge on separability. [6] proposed UCB-style algorithm which shows $\widetilde{\mathcal{O}}(\sqrt{NT})$ regret bound. [7] achieve the same order of the regret bound $\widetilde{\mathcal{O}}(\sqrt{NT})$ using TS approach with improved empirical performance. [14] show a matching lower bound of $\Omega(\sqrt{NT})$. All of this previous work on MNL bandits assumes each item is associated with a unique parameter, i.e., one cannot learn across items. In our proposed MNL contextual bandits, the utility of item i at round t is of the form $x_{ti}^\top \theta^*$ some fixed but unknown *utility parameter* θ^*; hence, we can learn across items. When the feature dimension $d \ll \sqrt{N}$, learning across items allows one to reduce the regret bound from $\widetilde{\mathcal{O}}(\sqrt{NT})$ to $\widetilde{\mathcal{O}}(d\sqrt{T})$. However, one cannot directly incorporate (time-varying) contextual information into the previous work (see, e.g. [6, 7]) since these methods require that the same assortment be offered repeatedly for a random number of rounds until an outside choice (no purchase) is observed. [15] proposed a UCB method which establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound for the MNL contextual bandit similar to our settings. Apart from the fact that their method is UCB based, there is another fundamental difference between [15] and our work. [15] enumerates the exponentially many (N choose K) assortments and builds confidence bounds for each of them. In contrast, our methods only maintain uncertainty for each of the N different items.

It is also worth mentioning work in the personalized MNL-bandit problem [25, 17, 11]. These works consider each item utility separately and learn N different parameters; hence there is no generalization across different items, which is different from our setting. Perhaps, the most related one among these personalized MNL bandit methods is [17], which proposed a TS algorithm for their problem. However, they only provide the Bayesian regret which is relatively easier to control compared to the worst-case regret (we discuss this aspect in Section 5), and again their method (as well as other personalized MNL bandit methods) still considers learning N separate parameters for each of the items; hence it is not scalable for a large item set (i.e., large N).

Linear contextual bandits [2, 9, 19, 36, 1, 18, 5] have been widely studied. [23] and [30] extend the linear contextual bandit to scalar, monotone, generalized linear bandit using a UCB-type approach. In most of these linear bandits or generalized linear bandits, balancing exploitation and exploration can be done simply by taking an action that maximizes the sum of mean reward and the variance. [5] define TS for linear contextual bandit as a Bayesian algorithm where a Gaussian prior over θ^* is updated according to the observed rewards, a random sample is drawn from the posterior, and the corresponding optimal arm is selected at each step. They show $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case regret bound. Following the work of [5], [3] show that the TS does not need to sample from an actual Bayesian posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret and provide an alternative proof of TS achieving the same regret bound $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. However, these results in (generalized) linear contextual bandits (either UCB or TS) do not apply directly to our MNL contextual bandit problem, since the choice probability of an item in an assortment is non-linear and non-monotone in the MNL parameter θ^*. It is also worthwhile to mention a line of work in other combinatorial bandit problems [35, 43, 26] mostly with semi-bandit feedback or cascading feedback. Our work is distinct from these combinatorial bandit problems since in cascading or semi-bandit settings, the mapping from the item context to the user feedback is still independent of other items in an offered set; hence it does not take substitution effect into account. On the other hand, MNL choice feedback is a function of the entire assortment which makes our analysis more challenging.

3 Problem Formulation

3.1 Notations

For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its ℓ_2-norm and $x^\top$ its transpose. The weighted ℓ_2-norm associated with a positive-definite matrix V is defined by $\|x\|_V := \sqrt{x^\top V x}$. The minimum and maximum singular values of a matrix V are written as $\lambda_{\min}(V)$ and $\|V\|$, respectively. The trace of a matrix V is trace(V). For two symmetric matrices V and W of the same dimensions, $V \succeq W$ means that $V - W$ is positive semi-definite. We define $[n]$ for a positive integer n to be a set containing positive integers up to n, i.e., $\{1, 2, ..., n\}$. Finally, we define $\mathcal{S}$ to be the set of candidate assortments with size constraint at most K, i.e., $\mathcal{S} = \{S \subset [N] : |S| \leq K\}$.

3.2 MNL Contextual Bandits

We formulate the problem of the MNL contextual bandit as follows. The decision-making agent can choose an assortment as a subset of the item set containing N distinct items, indexed by $i \in [N]$. At round t, feature vectors $x_{ti} \in \mathbb{R}^d$ for every item $i \in [N]$ are revealed to the agent. Each feature vector combines the information of the user and the corresponding item i. For example, suppose the user at round t is characterized by a feature vector v_t and the item i has a feature vector w_{ti} (note that we allow feature vectors for an item and a user to change over time), then we can use $x_{ti} = \text{vec}(v_t w_{ti}^\top)$, the vectorized outer-product of v_t and w_{ti}, as the combined feature vector of item i a at round t. If v_t is not available, we can use item dependent features only $x_{ti} = w_{ti}$. Given this contextual information, at every round t, the agent selects an assortment $S_t \in \mathcal{S}$ and observes the user choice represented as a binary vector $y_t \in \{0,1\}^{|S_t|}$ where $y_{ti} = 1$ if the i-th item in assortment S_t is chosen by the user and $y_{tj} = 0$ for all non-chosen items $j \in S_t$. Note that $\sum_{i \in S_t} y_{ti} \leq 1$ and we allow an "outside option" ($i = 0$) which means the user does not choose any items offered in S_t, i.e., $y_{ti} = 0$ for all $i \in S_t$. This user choice is given by the MNL choice model. Under this model, the probability that a user chooses item $i \in S_t$ is given by,

$$p_{ti}(S_t, \theta^*) = \frac{\exp\{x_{ti}^\top \theta^*\}}{1 + \sum_{j \in S_t} \exp\{x_{tj}^\top \theta^*\}}$$

where $\theta^* \in \mathbb{R}^d$ is an unknown time-invariant parameter and 1 in the denominator accounts for the outside option with $p_{t0}(S_t, \theta^*) = 1/(1 + \sum_{j \in S_t} \exp\{x_{tj}^\top \theta^*\})$. Then, the choice response variable $y_t = (y_{t0}, y_{t1}, ..., y_{tK})$ is a sample from this multinomial distribution:

$$y_t \sim \text{multinomial}\big(1, p_{t0}(S_t, \theta^*), p_{t1}(S_t, \theta^*), ..., p_{tK}(S_t, \theta^*)\big)$$

where 1 represents y_t is a single-trial sample. Also, we define noise $\epsilon_{ti} := y_{ti} - p_{ti}(S_t, \theta^*)$. Since ϵ_{ti} is bounded in $[0,1]$, ϵ_{ti} is σ^2-sub-Gaussian with $\sigma^2 = 1/4$. It is important to note that ϵ_{ti} is not independent across $i \in S_t$ due to the substitution effect in the MNL model.

The revenue parameter for each item i is also revealed at round t, denoted by r_{ti}. Note that r_{ti} is the revenue incurred by item i if item i is chosen by the user at round t. Without loss of generality, we assume $|r_{ti}| \leq 1$ for all i and t. Then, the expected revenue corresponding to assortment S_t is given by

$$R_t(S_t, \theta^*) = \sum_{i \in S_t} \frac{r_{ti} \exp\{x_{ti}^\top \theta^*\}}{1 + \sum_{j \in S_t} \exp\{x_{tj}^\top \theta^*\}}.$$

Let S_t^* be the offline optimal assortment at round t under full information when θ^* is known, i.e., if the true MNL probabilities $p_{ti}(S, \theta^*)$ are known a priori:

$$S_t^* = \arg\max_{S \in \mathcal{S}} R_t(S, \theta^*).$$

Consider a planning horizon T, where assortments can be offered at rounds $t = 1, ..., T$. The agent does not know the value of θ^* (hence $p_{ti}(S, \theta^*)$ is not known) and can only make sequential assortment decisions, $S_1, ..., S_T$ at rounds $1, ..., T$ respectively. Hence, the main challenge is how to construct an algorithm that simultaneously learns the unknown parameter θ^* and sequentially makes the decisions on offered assortments based on past choices and observed responses to maximize cumulative expected revenues over the planning horizon. The performance of an algorithm is usually measured by the regret, which is the gap between the expected revenue generated by the assortment chosen by the algorithm and that of the offline optimal assortment. We define the (worst-case) cumulative expected regret as

$$\mathcal{R}(T, \theta^*) = \sum_{t=1}^{T} \mathbb{E}\big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \mid \theta^*\big]$$

where $R_t(S_t^*, \theta^*)$ is the expected revenue corresponding to the offline optimal assortment at round t, and the expectation is taken over random parameters and possible randomization in a learning algorithm. When it is clear that we condition on a fixed θ^*, we denote $\mathcal{R}(T) := \mathcal{R}(T, \theta^*)$ in the rest of the paper. In Bayesian settings, i.e., when θ^* is randomly generated or the learning agent has a prior belief in θ^*, the Bayesian cumulative regret [38] over T horizon is defined as

$$\mathcal{R}_{\text{Bayes}}(T) = \mathbb{E}_{\theta^*}\left[\mathcal{R}(T, \theta^*)\right] = \sum_{t=1}^{T} \mathbb{E}\big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*)\big]$$

where the expectation is taken also over the distribution of θ^*. In other words, $\mathcal{R}_{\text{Bayes}}(T)$ is a weighted average of $\mathcal{R}(T, \theta^*)$ under the prior on θ^*.

3.3 Assumptions

We introduce general assumptions on the structure of the problem.

Assumption 1. $\|x_{ti}\| \leq 1$ *for all* t *and* i*. Also,* $\|\theta^*\| \leq 1$.

This assumption is used to make the regret bounds scale-free for convenience and is in fact standard in the bandit literature. If $\|x_{ti}\| \leq C$ and $\|\theta^*\| \leq C$ for some constant C instead, then our regret bounds would increase by a factor of C.

Assumption 2. *There exists* $\kappa > 0$ *such that for every item* $i \in S$ *and any* $S \in \mathcal{S}$ *and all round* t $\inf_{S \in \mathcal{S}, \theta \in \mathbb{R}^d} p_{ti}(S, \theta) p_{t0}(S, \theta) \geq \kappa$.

Note that this is equivalent to a standard assumption in generalized linear contextual bandit literature [23, 30] to ensure the Fisher information matrix is invertible and is adapted to suit our MNL setting. We discuss the need for this assumption in detail in Appendix A.

4 Algorithm: TS-MNL

In this section, we describe TS-MNL, our first TS algorithm for the MNL contextual bandit problem, and present its Bayesian regret bound. We first provide the definition of the posterior distribution Q_t on the unknown parameter θ^*. At the beginning of the learning phase, the agent knows that θ^* is distributed according to Q_0, the prior distribution. Now, at each round t, the agent has access to the observations up to round t, $\mathcal{D}_t = \{X_\tau, y_\tau\}_{\tau=1}^{t-1}$ where $X_\tau = \{x_{\tau i}\}_{i \in S_\tau}$. Then the agent combines Q_0 and $\mathcal{D}_t$ to define the posterior distribution $Q_t(\theta)$:

$$Q_t(\theta) \propto Q_0(\theta) p(\mathcal{D}_t|\theta), \quad \text{where } p(\mathcal{D}_t|\theta) = \prod_{\tau=1}^{t-1} \prod_{i \in S_\tau} (p_{\tau i}(S_\tau, \theta))^{y_{\tau i}} \tag{1}$$

and the "$\propto$" notation hides the partition function $\int_\phi Q_0(\phi) p(\mathcal{D}_t|\phi) d\phi$ in the denominator. In other words, the posterior distribution is proportional to the product of the prior distribution and the likelihood function. Note that there is no conjugate prior for the MNL model. Hence, sampling from Q_t is intractable. In order to overcome this intractability, one may draw an approximate sampling using Markov chain Monte Carlo [8]. For ease of exposition, we assume the following in this section and in the Bayesian regret analysis. We will later provide a remedy for this intractability in the modification of our algorithm for the worst-case regret analysis.

Assumption 3. *We can sample from* $Q_t(\theta)$.

In each round t, TS-MNL algorithm consists of three major steps. First, it randomly samples a parameter $\widetilde{\theta}_t$ from the posterior distribution Q_t. Second, it computes the assortment choice S_t under this sampled parameter $\widetilde{\theta}_t$. Finally, S_t is offered to the user and feedback y_t is observed. The pseudocode of TS-MNL is presented in Algorithm 1.

Algorithm 1 TS-MNL

1: **Input**: prior distribution Q_0
2: **for** all $t = 1$ to T **do**
3: Observe x_{ti} and r_{ti} for all $i \in [N]$
4: Sample $\widetilde{\theta}_t$ from the posterior distribution Q_t in Eq.(1)
5: Compute $S_t = \arg\max_{S \in \mathcal{S}} R_t(S, \widetilde{\theta}_t)$
6: Offer S_t and observe y_t (user choice at round t)
7: **end for**

Combinatorial Optimization. Algorithm 1 has the combinatorial optimization step in Line 5. There are efficient polynomial-time algorithms available to solve this combinatorial optimization problem [37, 20] for given utility estimates under the sampled parameter. In particular, we can use the solution of the linear programming (LP) formulation presented in [20] for this optimization step.

4.1 Bayesian Regret of TS-MNL

We state the Bayesian cumulative regret bound for Algorithm 1 in Theorem 1. We also provide an overview of establishing the regret bound.

Theorem 1. *Suppose we run* TS-MNL *(Algorithm 1) for a total of T rounds with assortment size constraint K. Then the Bayesian regret of the algorithm is upper-bounded by*

$$\begin{aligned}\mathcal{R}_{\textit{Bayes}}(T) &\leq \mathcal{O}(1) + \left[\frac{1}{\kappa}\sqrt{2d\log\left(1+\frac{TK}{d^2}\right)+2\log T} + \frac{\sqrt{d}}{\kappa}\right]\cdot\sqrt{2dT\log\left(1+\frac{TK}{d^2}\right)} \\ &= \mathcal{O}\left(d\sqrt{T}\log\left(1+\frac{TK}{d^2}\right)\right).\end{aligned}$$

Theorem 1 establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$ Bayesian regret. [15] established the lower bound $\Omega(d\sqrt{T}/K)$ for MNL contextual bandits under almost identical settings. When K is small and fixed (which is typically true in many applications), Theorem 1 demonstrates that TS-MNL is almost optimal. Furthermore, the regret bound is completely free of N; hence TS-MNL is applicable to the case of a large number of items (large N). Also, if $K \leq d^2$, the regret bound becomes free of K. In Section 6, we introduce modifications to TS-MNL for the worst-case regret analysis which include the explicit use of regularized MLE for parameter estimation and sampling from the Gaussian distribution instead of maintaining the actual posterior to overcome the intractability. The concentration results derived for the Bayesian regret analysis in this section serve as a building block for the worst-case regret analysis for the modified algorithm.

The proof outline of Theorem 1 is motivated by [38, 43]. Given $\mathcal{F}_t$ which contains all available information up to round t, $\widetilde{\theta}_t$ and θ^* are i.i.d. with the posterior distribution Q_t in the Bayesian perspective. Also, the optimization step is a fixed combinatorial optimization and $\{x_{ti}\}_{i\in[N]}$ are fixed given $\mathcal{F}_t$. Hence, conditioning on $\mathcal{F}_t$, S_t and S_t^* are also i.i.d. Therefore, the expected regret pertaining to the random sampling is 0.; Then, we control the estimation error of θ^* for which we utilize the finite-sample concentration results for MNL parameter. The proofs are left to Appendix B.

5 Worst-Case Regret

Algorithm 1 is still valid under a frequentist setting, i.e., when the true parameter is not a random variable but a fixed parameter. However, when analyzing the worst-case regret (also known as frequentist regret) for the algorithm, the main technical difficulty lies in controlling the deviation in performance due to the random sampling of the algorithm. Note that in Bayesian regret analysis, controlling this sampling deviation is not addressed because of the assumption that $\widetilde{\theta}_t$ and θ^* are i.i.d. conditioning on $\mathcal{F}_t$. However, this does not hold anymore when θ^* is fixed; hence the worst-case regret analysis needs to ensure that the deviation due to sampling is small enough. To see this, we decompose the worst-case immediate regret into a few components.

$$\begin{aligned}\mathcal{R}(t) &= \mathbb{E}[R_t(S_t^*,\theta^*) - R_t(S_t,\theta^*)] \\ &= \mathbb{E}[R_t(S_t^*,\theta^*) - R_t(S_t^*,\widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t^*,\widetilde{\theta}_t) - R_t(S_t,\widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t,\widetilde{\theta}_t) - R_t(S_t,\theta^*)] \\ &\leq \mathbb{E}[R_t(S_t^*,\theta^*) - R_t(S_t^*,\widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t,\widetilde{\theta}_t) - R_t(S_t,\theta^*)] \qquad (2)\end{aligned}$$

The inequality comes from the fact that our assortment choice at round t, S_t, is optimal under $\widetilde{\theta}_t$; hence $R_t(S_t^*,\widetilde{\theta}_t) \leq R_t(S_t,\widetilde{\theta}_t)$. The second term $\mathbb{E}[R_t(S_t,\widetilde{\theta}_t) - R_t(S_t,\theta^*)]$ in Eq.(2) is relatively easier to control. We can show that the term can be bounded by combining the upper-bound for the estimation error $|x^\top(\hat{\theta}_t - \theta^*)|$ and the concentration of the sampling probability of $\widetilde{\theta}_t$. However, controlling the first term $\mathbb{E}[R_t(S_t^*,\theta^*) - R_t(S_t^*,\widetilde{\theta}_t)]$ in Eq.(2) is more challenging in frequentist analysis. First, note that $\mathbb{E}[R_t(S_t^*,\theta^*) - R_t(S_t^*,\widetilde{\theta}_t)] = 0$ in the Bayesian regret by the assumption that θ^* and $\hat{\theta}_t$ are i.i.d. conditioning on $\mathcal{F}_t$ as mentioned earlier. However, this is no longer true in the worst-case regret analysis. In the worst-case regret analysis of TS, this term is controlled by showing that a sampled parameter is optimistic frequently enough. In other words, we need to lower-bound the probability of the sampled parameter being optimistic, i.e., $\mathbb{P}\big(R_t(S_t^*,\widetilde{\theta}_t) \geq R_t(S_t^*,\theta^*) \mid \mathcal{F}_t\big) \geq p$ for some parameter free $p > 0$.

To describe the challenge in our MNL contextual bandit problem, we present the following lemma which shows that the expected revenue for the optimal assortment is monotonically increasing with an increase in the utility estimates.

Lemma 1 ([6], Lemma 4.2)**.** *Suppose S_t^* is the optimal assortment under the true parameter θ^* at round t, i.e., $S_t^* = \arg\max_{S \in \mathcal{S}} R_t(S, \theta^*)$. Also suppose that $x_{ti}^\top \theta^* \leq x_{ti}^\top \theta'$ for all $i \in S_t^*$. Then $R_t(S_t^*, \theta^*) \leq R_t(S_t^*, \theta')$.*

Note that Lemma 1 shows the monotonicity of expected revenue only for the optimal assortment and it does not claim that the expected revenue is generally a monotone function for all assortments. This lemma implies that we can lower-bound the probability of having an optimistic expected revenue under the sampled parameter.

$$\mathbb{P}\left(R_t(S_t^*, \widetilde{\theta}_t) \geq R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\right) \geq \mathbb{P}\left(x_{ti}^\top \widetilde{\theta}_t \geq x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right)$$

However, this makes the probability of being optimistic exponentially small in the size of the assortment S_t^*, i.e., exponentially small in $\mathcal{O}(K)$, which in turn results in exponential dependence on $\mathcal{O}(K)$ in the worst-case regret bound. In order to overcome such an issue, we adopt a few modifications in the algorithm which we discuss in the following section.

6 TS-MNL with Optimistic Sampling

Sampling from Gaussian Distribution. We modify our TS algorithm to a generic randomized algorithm constructed on the regularized MLE rather than sampling from an actual Bayesian posterior. [3] show that TS does not need to sample from an actual posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret. Specifically, instead of sampling from the posterior Q_t, we sample $\widetilde{\theta}_t$ from Gaussian distribution $\mathcal{N}\big(\hat{\theta}_t, \alpha_t^2 V_t^{-1}\big)$ where $\hat{\theta}_t$ is the regularized MLE, the minimizer of Eq.(3), and α_t is the confidence radius. This way, we ensure tractability of the sampling distribution. Furthermore, this Gaussian approximation allows us to adopt optimistic sampling (which we discuss below) in an efficient manner.

Optimistic Sampling. The optimistic sampling we present here is a key ingredient in avoiding the theoretical challenges present in the worst-case regret analysis. For optimistic sampling, instead of drawing a single sample $\widetilde{\theta}_t$, we draw M independent samples $\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$ from $\mathcal{N}\big(\hat{\theta}_t, \alpha_t^2 V_t^{-1}\big)$ (the exact value of M is specified in Theorem 2). Then we compute the optimistic utility estimate $\widetilde{u}_{ti}$ for each $i \in [N]$:

$$\widetilde{u}_{ti} = \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)}.$$

We define $\widetilde{R}_t(S)$ to be the expected revenue of assortment S based on $\widetilde{u}_{ti}$:

$$\widetilde{R}_t(S) = \frac{\sum_{i \in S} r_{ti} \exp\{\widetilde{u}_{ti}\}}{1 + \sum_{j \in S} \exp\{\widetilde{u}_{tj}\}}$$

Note that this optimistic sampling scheme is different from that proposed in [7]. The setting in [7] is non-contextual, and they use a 1-dimensional Gaussian random variable to correlate the samples of the utility of the K items in order to ensure the probability that all samples are simultaneously optimistic is a constant. This correlated sampling reduces the overall variance severely, hence they propose taking K samples instead of a single sample to increase the variance. In contrast, we take multiple samples of the multivariate Gaussian distribution to directly ensure that the probability of an optimistic sample is sufficiently large.

The pseudocode of the modified algorithm is presented in Algorithm 2. As before, we can utilize the LP solution [20] for the optimization step in Line 6. The modified algorithm now explicitly maintains the matrix V_t and computes the regularized MLE $\hat{\theta}_t$. Note that α_T can be replaced by $\alpha_t = \mathcal{O}\left(\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t}\right)$ at round t, if the planning horizon T is not known and the analysis holds for either case.

Algorithm 2 TS-MNL with Optimistic Sampling

Input: sample size M, confidence radius α_T, penalty parameter λ
for all $t = 1$ to T **do**
  Observe x_{ti} and r_{ti} for all $i \in [N]$
  Sample $\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$ independently from $\mathcal{N}(\hat{\theta}_t, \alpha_T^2 V_t^{-1})$
  Compute $\widetilde{u}_{ti} = \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)}$ for all $i \in [N]$
  Compute $S_t = \arg\max_{S \in \mathcal{S}} \widetilde{R}_t(S)$
  Offer S_t and observe y_t (user choice at round t)
  Update $V_{t+1} \leftarrow V_t + \sum_{i \in S_t} x_{ti} x_{ti}^\top$
  Compute the regularized MLE $\hat{\theta}_t$ by minimizing

$$-\sum_{\tau=1}^{t} \sum_{i \in S_\tau} y_{\tau i} \log p_{\tau i}(S_\tau, \theta) + \frac{\lambda}{2}\|\theta\|^2. \tag{3}$$

end for

6.1 Worst-Case Regret of TS-MNL with Optimistic Sampling

Theorem 2. *Suppose we run* TS-MNL *with "optimistic sampling" (Algorithm 2) for a total of T rounds with optimistic sample size $M = \lceil 1 - \frac{\log K}{\log(1-1/(4\sqrt{e\pi}))} \rceil$, the penalty parameter $\lambda \geq 1$ and assortment size constraint K. Then the worst-case regret of the algorithm is upper-bounded by*

$$\mathcal{R}(T) \leq \mathcal{O}(1) + 16\sqrt{e\pi}\beta_T \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T} \right) + (\alpha_T + \beta_T) \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}$$

where $\alpha_T = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{TK}{d\lambda}\right) + 4 \log T} + \frac{\sqrt{\lambda}}{\kappa}$ and $\beta_T = \alpha_T \sqrt{2d \log(MT)}$.

Theorem 2 establishes $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case regret, which matches the regret bounds of TS methods for linear contextual bandits [5, 3] up to logarithmic factor. The regret bound shows no dependence on N, and has an additional $\mathcal{O}(\sqrt{\log \log K})$ dependence due to optimistic sampling which is very small for any reasonable assortment size K. Compared to Theorem 1, the additional factor $\sqrt{d}$ comes from the deviation of the random sampling which is addressed in the worst-case regret analysis.

The proof of Theorem 2 utilizes the anti-concentration property of the maximum of Gaussian random variables for ensuring frequent optimism. In particular, we show in the following lemma that the proposed optimistic sampling can ensure a constant probability of optimism.

Lemma 2. *Suppose $\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t} + \frac{\sqrt{\lambda}}{\kappa}$ and we take optimistic samples of size $M = \lceil 1 - \frac{\log K}{\log(1-1/(4\sqrt{e\pi}))} \rceil$. Then we have*

$$\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\right) \geq \frac{1}{4\sqrt{e\pi}}.$$

The inverse of the lower-bounding probability $4\sqrt{e\pi}$ can be interpreted as the expected time between any two optimistic assortment selections. In other words, our modified algorithm is optimistic at least with a constant frequency. Then, using this frequent optimism, we can ensure that the cumulative regret due to the random sampling can be bounded. Along with this result, we show the concentrations of both regularized MLE and TS samples to establish the regret bound in Theorem 2. The proofs are left to Appendix D.

7 Numerical Study

In this section, we perform numerical evaluations to analyze two variants of our proposed algorithm: TS-MNL with optimistic sampling (Algorithm 2) and TS-MNL with the Gaussian approximation for the posterior distribution. We perform both synthetic experiments as well as simulated experiments using a real-world dataset: *MovieLens* dataset.[2] We simulated instances of the MNL contextual bandit problem with varying parameter values.

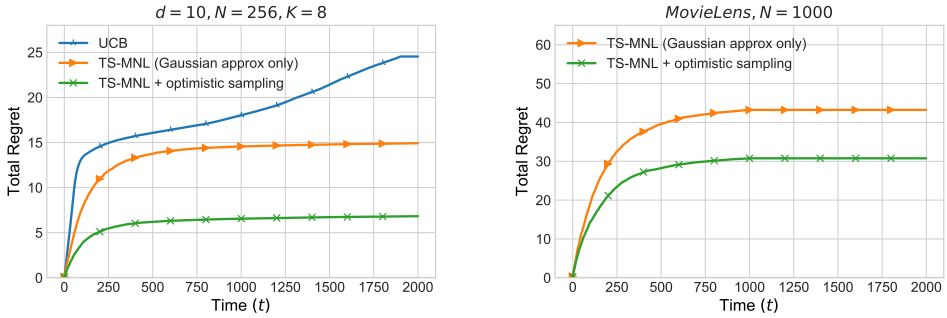

Figure 1: Regret growth with T for a UCB method and TS-MNL variants on MNL contextual bandits.

We report the worst-case cumulative expected regret for each of the experiments. For the synthetic experiments, we randomly draw θ^* for each instance and hence we can directly compute the expected regret using θ^*. For the experiments using MovieLens dataset, we use offline regression using the entire dataset to estimate the unknown parameter θ^* and compare with the estimates from online experiments. The details of the experimental setup and additional experimental results are presented in Appendix G.

Figure 1 shows the performances averaged over 40 independent instances for each experiment. For comparison, we evaluate the performances of our TS-MNL algorithms along with the performances of the UCB method proposed in [15]. The performances of the proposed two variants of TS-MNL are observed to be superior to that of the UCB method on the synthetic data in our experiments, which is consistent with the other empirical evidence of TS methods in the literature. The experiments with MovieLens dataset (and the additional experiments shown in Appendix G) suggest that our methods can be used and effective for problem instances with a large number of items, i.e., large N. Furthermore, TS-MNL with optimistic sampling consistently performs better than TS-MNL with Gaussian approximation only. The results of these experiments support our theoretical analysis: TS-MNL with optimistic sampling takes advantage of the MNL structure and can guarantee a worst-case statistical efficiency.

8 Discussions

In this paper, we study the dynamic assortment selection problem under an MNL model with contextual information. We propose two TS algorithms for the MNL contextual bandits which learn the parameters of the underlying choice model while simultaneously maximizing the cumulative revenue. We provide their theoretical performance bounds and show attractive numerical performances in our experiments. We also discuss the challenges which arise in worst-case regret analysis for this combinatorial action selection problem under the MNL model. We believe that these challenges are potentially present in many other problems involving combinatorial action selections with context/feature information beyond the MNL model. To our knowledge, the worst-case regret analysis in this work is the first frequentist regret guarantee for contextual bandits with combinatorial action selection of any kind. We believe that our proposed optimistic sampling framework can be useful for other combinatorial contextual bandit problems.

[2] https://grouplens.org/datasets/movielens/

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, pages 3–11, 1999.

[3] Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

[4] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

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

[6] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: a dynamic learning approach to assortment selection. *arXiv preprint arXiv:1706.03880*, 2017.

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

[8] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

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

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

[11] Fernando Bernstein, Sajad Modaresi, and Denis Sauré. A dynamic clustering approach to data-driven assortment personalization. *Management Science*, 2018.

[12] Felipe Caro and Jérémie Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

[13] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

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

[15] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information. *arXiv preprint arXiv:1810.13069*, 2018.

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

[17] 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.

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

[19] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, page 355–366, 2008.

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

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

[22] Antoine Désir, Vineet Goyal, and Jiawei Zhang. Near-optimal algorithms for capacity constrained assortment optimization. *Available at SSRN 2543309*, 2014.

[23] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[24] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[25] Nathan Kallus and Madeleine Udell. Dynamic assortment personalization in high dimensions. *arXiv preprint arXiv:1610.05604*, 2016.

[26] 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, 2015.

[27] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

[28] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[29] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

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

[31] Shuai Li, Tor Lattimore, and Csaba Szepesvari. Online learning to rank with features. In *International Conference on Machine Learning*, pages 3856–3865, 2019.

[32] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

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

[34] Robin L Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.

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

[36] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

[38] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

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

[40] Malcolm Strens. A bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pages 943–950, 2000.

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

[42] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[43] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122, 2015.

Appendices for Thompson Sampling for MNL Contextual Bandits

A Regularized Maximum Likelihood Estimation for MNL Model

We briefly discuss regularized maximum likelihood estimation (MLE) for MNL model – specifically the estimation of the unknown parameter θ^* of the MNL model with the rigde penalty. First, recall that $y_t \in \{0,1\}^{|S_t|}$ is the user choice where y_{ti} is the i-th component of y_t. Then, the likelihood function under parameter θ is then given by

$$\mathcal{L}(\mathcal{D}_n|\theta) = \prod_{t=1}^{n} \prod_{i \in S_t} (p_{ti}(S_t, \theta))^{y_{ti}}$$

where $\mathcal{D}_n = \{X_t, S_t, y_t\}_{t=1}^n$ and $X_t = \{x_{ti}\}_{i\in[N]}$. Taking the negative logarithm gives

$$\ell_n(\theta) = -\log \mathcal{L}(\mathcal{D}_n|\theta) = -\sum_{t=1}^{n} \sum_{i \in S_t} y_{ti} \log p_{ti}(S_t, \theta)$$

which is known as the cross-entropy error function for the multi-class classification problem. Now, the ridge penalized maximum likelihood estimation for MNL model is given by the following minimization problem:

$$\hat{\theta} = \arg\min_{\theta} \left[\ell_n(\theta) + \frac{\lambda}{2} \|\theta\|^2 \right] \tag{4}$$

with the penalty parameter $\lambda \geq 1$.

Taking the gradient of this penalized log-likelihood function with respect to θ, we obtain

$$\nabla_\theta \left[\ell_n(\theta) + \frac{\lambda}{2} \|\theta\|_2^2 \right] = \sum_{t=1}^{n} \sum_{i \in S_t} (p_{ti}(S_t, \theta) - y_{ti}) x_{ti} + \lambda\theta. \tag{5}$$

Instead of using the regularized MLE for the parameter estimation, one could consider using the MLE without regularization. For this, however, one may consider performing a random initialization (random exploration) to ensure that the matrix V_t is invertible. This necessity comes from the classical likelihood theory [28]: as the sample size n goes to infinity, we know the MLE $\hat{\theta}_n^{\text{ML}}$ is asymptotically normal, with $\hat{\theta}_n^{\text{ML}} - \theta^* \to \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1})$ where $\mathcal{I}_{\theta^*}$ is the Fisher information matrix. In the MNL model, $\mathcal{I}_{\theta^*}$ is lower bounded by $\sum_t \sum_{i\in S_t} p_{ti}(S_t, \theta^*) p_{t0}(S_t, \theta^*) x_{ti} x_{ti}^\top$ (see Lemma 4). Hence, if $p_{ti}(S_t, \theta^*) p_{t0}(S_t, \theta^*) \geq \kappa > 0$, then we can ensure that $\mathcal{I}_{\theta^*}$ is invertible and prevent asymptotic variance of $x^\top \hat{\theta}_n^{\text{ML}}$ from going to infinity for any x. When performing random exploration instead of the regularization, the length of such exploration needs to be specified to ensure that the minimum eigenvalue of the matrix V_t is large enough — we discuss in detail in Appendix F.

B Proof of Theorem 1: Bayesian Regret Analysis

Let $\mathcal{F}_t$ denote the filtration which contains all available information up to round t. Recall that $\widetilde{\theta}_t$ is independently drawn from the posterior distribution Q_t in Algorithm 1 and also note that in our Bayesian setting the posterior belief in θ^* is distributed as Q_t conditioning on $\mathcal{F}_t$. Therefore, conditioning on $\mathcal{F}_t$, $\widetilde{\theta}_t$ and θ_t^* are i.i.d. with Q_t. Also note that our optimization oracle is a fixed combinatorial optimization algorithm and $\{x_{ti}\}_{i\in[N]}$ are fixed given $\mathcal{F}_t$. Hence, conditioning on $\mathcal{F}_t$, S_t and S^* are also i.i.d.

B.1 Confidence Bound for Expected Revenue

We define a upper confidence expected revenue as

$$U_t(S, \hat{\theta}_t) = \frac{\sum_{i\in S} r_{ti} \exp\left\{ x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} \right\}}{1 + \sum_{j\in S} \exp\left\{ x_{tj}^\top \hat{\theta}_t + \alpha_t \|x_{tj}\|_{V_t^{-1}} \right\}}$$

where $\alpha_t > 0$ is the confidence width and its value is specified later (Lemma 4). Also, we define $V_t = \sum_{\tau=1}^{t} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$. Note that this upper confidence expected revenue U_t is constructed for the sake of the analysis presented in this section and does not affect the proposed algorithm (or its assortment selection). We first decompose the immediate regret using U_t.

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(t) \mid \mathcal{F}_t] &= \mathbb{E}\big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\big] \\
&= \mathbb{E}\big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\big] + \mathbb{E}\big[U_t(S_t^*, \hat{\theta}_t) - U_t(S_t, \hat{\theta}_t) \mid \mathcal{F}_t\big] \\
&\quad + \mathbb{E}\big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\big].
\end{aligned}$$

Notice that $\mathbb{E}\big[U_t(S_t^*, \hat{\theta}_t) - U_t(S_t, \hat{\theta}_t) \mid \mathcal{F}_t\big] = 0$ since conditioning on $\mathcal{F}_t$, S_t and S^* are i.i.d. and U_t is a deterministic function. Hence, for the Bayesian cumulative regret, we are left bound the two quantities $\mathcal{R}^1_{\text{Bayes}}(T)$ and $\mathcal{R}^2_{\text{Bayes}}(T)$ as the following:

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{R}(t) \mid \mathcal{F}_t] = \underbrace{\sum_{t=1}^{T} \mathbb{E}\big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\big]}_{\mathcal{R}^1_{\text{Bayes}}(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\big]}_{\mathcal{R}^2_{\text{Bayes}}(T)}$$

In the following sections, we present the upper-bounds for $\mathcal{R}^1_{\text{Bayes}}(T)$ and $\mathcal{R}^2_{\text{Bayes}}(T)$. Then we combine the results to establish the Bayesian cumulative regret for TS-MNL (Algorithm 1).

B.2 Bounding $\mathcal{R}^1_{\text{Bayes}}(T)$

Before we present the upper bound for $\mathcal{R}^1_{\text{Bayes}}(T)$, we introduce the following lemma which utilizes the structure of the MNL model. Lemma 3 shows that the expected revenue R_t (and hence U_t) has a Lipschitz property, i.e., Lemma 3 ensures that we can control the difference between expected revenues by bounding with maximum difference in utilities.

Lemma 3. *For any two utility parameters $u_t = [u_{t1}, ..., u_{tN}]$ and $u'_t = [u'_{t1}, ..., u'_{tN}]$, we have*

$$\frac{\sum_{i \in S} r_{ti} \exp(u_{ti})}{1 + \sum_{j \in S} \exp(u_{tj})} - \frac{\sum_{i \in S} r_{ti} \exp(u'_{ti})}{1 + \sum_{j \in S} \exp(u'_{tj})} \leq \max_{i \in S} |u_{ti} - u'_{ti}|.$$

In particular, if $u_{ti} \geq u'_{ti}$ for all i, then

$$\frac{\sum_{i \in S} r_{ti} \exp(u_{ti})}{1 + \sum_{j \in S} \exp(u_{tj})} - \frac{\sum_{i \in S} r_{ti} \exp(u'_{ti})}{1 + \sum_{j \in S} \exp(u'_{tj})} \leq \max_{i \in S} (u_{ti} - u'_{ti}).$$

Note that in the statement of Lemma 3 we use the explicit form of expected revenues (with generic utility parameters) in order to accommodate both R_t and U_t. Now, Lemma 4 below shows that the true parameter θ^* lies within an ellipsoid centered at $\hat{\theta}_t$ with confidence radius α_t. This is the result for the non-i.i.d. finite-sample confidence bound for the MNL parameter.

Lemma 4. *Define $\alpha_t = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t} + \frac{\sqrt{\lambda}}{\kappa}$. If $\hat{\theta}_t$ is the solution to the regularized MLE in Eq.*(4) *at round t, then*

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t$$

holds for all t with a probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$.

If θ^* is indeed within the confidence region for all t, i.e., if the high probability event of Lemma 4 holds, then one can show that $x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} \geq x_{ti}^\top \theta^*$ for all i. Hence, $U_t(S_t^*, \hat{\theta}_t)$ is greater than $R_t(S_t^*, \theta^*)$. Then, $\mathcal{R}^1_{\text{Bayes}}(T)$ can be upper-bounded by 0. However, there is a small probability of failure for the confidence region which we need to take into consideration. The following lemma formally state the result.

Lemma 5. *Let the upper confidence expected revenue $U_t(S_t^*, \hat{\theta}_t)$ be defined with the confidence width $\alpha_t = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t} + \frac{\sqrt{\lambda}}{\kappa}$. Then, we have*

$$\sum_{t=1}^{T} \mathbb{E}\big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\big] = \mathcal{O}(1).$$

B.3 Bounding $\mathcal{R}^2_{\text{Bayes}}(T)$

This portion of the regret is controlled by the concentration of the upper-confidence expected revenue $U_t(S_t, \hat{\theta}_t)$ to the true expected revenue $R_t(S_t, \theta^*)$. We can first use Lemma 3 to upper-bound $\mathcal{R}^2_{\text{Bayes}}(T)$ by the expected maximum difference in utilities. Now, suppose that θ^* resides within the confidence region with the radius α_t for all rounds t (Lemma 4). Then the same holds for the radius α_T since $\alpha_T \geq \alpha_t$. Using this fact and Cauchy-Schwartz inequality, we can further bound $\mathcal{R}^2_{\text{Bayes}}(T)$ by Eq.(6).

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\big] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t} \left(x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} - x_{ti}^\top \theta^*\right) \mid \mathcal{F}_t\right] \\ &\leq 2\alpha_T \sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \qquad (6)\end{aligned}$$

Then, we are left to control the sum of the expectations in Eq.(6). Specifically, we provide a worst-case bound on $\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}$ for any realization of random variables in Lemma 6, which presents a self-normalized bound.

Lemma 6. *Define $V_T = V + \sum_{t=1}^{T} \sum_{i \in S_t} x_{ti} x_{ti}^\top$ where $V = \lambda I_d$. Then we have*

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}.$$

Combining the results of Lemma 6 and Eq.(6), we have

$$\sum_{t=1}^{T} \mathbb{E}\big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\big] \leq 2\alpha_T \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \mathcal{O}(1)$$

where $\alpha_T = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{TK}{d\lambda}\right) + 4 \log T} + \frac{\sqrt{\lambda}}{\kappa}$ and $\mathcal{O}(1)$ comes from the failure event of the concentration of $\hat{\theta}_t$ in Lemma 4.

B.4 Combining $\mathcal{R}^1_{\text{Bayes}}(T)$ and $\mathcal{R}^2_{\text{Bayes}}(T)$

Combining the bounds for $\mathcal{R}^1_{\text{Bayes}}(T)$ and $\mathcal{R}^2_{\text{Bayes}}(T)$, we have

$$\mathcal{R}_{\text{Bayes}}(T) \leq \mathcal{O}(1) + \left[\frac{1}{\kappa}\sqrt{2d \log\left(1 + \frac{TK}{d\lambda}\right) + 2 \log T} + \frac{\sqrt{\lambda}}{\kappa}\right] \cdot \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}.$$

For completeness, we choose $\lambda = d$ to get the regret bound shown in Theorem 1 which gives the Bayesian regret $\mathcal{R}_{\text{Bayes}}(T) = \mathcal{O}\left(d\sqrt{T} \log\left(1 + \frac{TK}{d^2}\right)\right)$. Since Algorithm 1 itself does not use the regularized MLE for parameter estimation, one may optimize over the choice of λ in the regret bound.

C Proofs of Lemmas for Theorem 1

C.1 Proof of Lemma 3

Proof. By the mean value theorem, there exists $\bar{u}_{ti} := (1 - c)u_{ti} + cu'_{ti}$ for some $c \in (0, 1)$ with

$$\begin{aligned}&\frac{\sum_{i \in S} r_{ti} \exp(u_{ti})}{1 + \sum_{j \in S} \exp(u_{tj})} - \frac{\sum_{i \in S} r_{ti} \exp(u'_{ti})}{1 + \sum_{j \in S} \exp(u'_{tj})} \\ &= \sum_{i \in S} r_{ti} p_{ti}(S, \bar{u}_t)(u_{ti} - u'_{ti}) - R_t(S, \bar{u}_t) \cdot \sum_{i \in S} p_{ti}(S, \bar{u}_t)(u_{ti} - u'_{ti}) \\ &= \sum_{i \in S} \big(r_{ti} - R_t(S, \bar{u}_t)\big) p_{ti}(S, \bar{u}_t)(u_{ti} - u'_{ti}) \\ &\leq \max_{i \in S} |u_{ti} - u'_{ti}|\end{aligned}$$

where the inequality is from $|r_{ti}| \leq 1$, and $p_{ti}(S, \bar{u}_t) \leq 1$ is a multinomial probability (and hence $R_t(S, \bar{u}_t) \leq 1$). □

C.2 Proof of Lemma 4

Proof. We first define the function $G_n(\theta)$ which we use throughout the proof:

$$G_n(\theta) = \sum_{t=1}^{n} \sum_{i \in S_t} \left[\left(p_{ti}(S_t, \theta) - p_{ti}(S_t, \theta^*) \right) x_{ti} \right] + \lambda(\theta - \theta^*)$$

$G_n(\theta)$ is the difference in the gradients of the ridge penalized maximum likelihood in Eq.(5) evaluated at θ and at θ^*. Notice that $G_n(\hat{\theta}) = \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti} - \lambda\theta^*$ since the choice of $\hat{\theta}$ is given by the ridge penalized maximum likelihood. To see that, first note that $\hat{\theta}$ is the minimizer of Eq.(4); hence is given by the solution to the following equation:

$$\sum_{t=1}^{n} \sum_{i \in S_t} \left(p_{ti}(S_t, \hat{\theta}) - y_{ti} \right) x_{ti} + \lambda\hat{\theta} = 0 \tag{7}$$

Therefore, it follows that

$$\begin{aligned}
G_n(\hat{\theta}) &= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_{ti}(S_t, \hat{\theta}) - p_{ti}(S_t, \theta^*) \right) x_{ti} + \lambda(\hat{\theta} - \theta^*) \\
&= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_{ti}(S_t, \hat{\theta}) - y_{ti} \right) x_{ti} + \lambda\hat{\theta} + \sum_{t=1}^{n} \sum_{i \in S_t} \left(y_{ti} - p_{ti}(S_t, \theta^*) \right) x_{ti} - \lambda\theta^* \\
&= 0 + \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti} - \lambda\theta^*
\end{aligned}$$

where the last equality is from (7) and the definition of $\epsilon_{ti} = y_{ti} - p_{ti}(S_t, \theta^*)$. For convenience, we define $Z_n := \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti}$. Hence, $G_n(\hat{\theta}) = Z_n - \lambda\theta^*$. Also, we will denote $p_{ti}(\theta) := p_{ti}(S_t, \theta)$ when it is clear that S_t is the assortment chosen at round t.

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

$$\begin{aligned}
G_n(\theta_1) - G_n(\theta_2) &= \sum_{t=1}^{n} \sum_{i \in S_t} \Big[\left(p_{ti}(\theta_1) - p_{ti}(\theta_2) \right) x_{ti} \Big] + \lambda(\theta_1 - \theta_2) \\
&= \left[\left(\sum_{t=1}^{n} \sum_{i \in S_t} \sum_{j \in S_t} \nabla_j p_{ti}(\bar{\theta}) x_{ti} x_{tj}^\top \right) + \lambda I_d \right] (\theta_1 - \theta_2) \\
&= \left[\sum_{t=1}^{n} \left(\sum_{i \in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top \right) + \lambda I_d \right] (\theta_1 - \theta_2)
\end{aligned}$$

where I_d is a $d \times d$ identitiy matrix. We define the matrix H_t as

$$H_t := \sum_{i \in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i,j \in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top$$

Notice H_t is a Hessian of a negative log-likelihood which is convex. Hence, H_t is positive semidefinite. Also note 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

$$\begin{aligned}
H_t &= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{tj}^\top + x_{tj} x_{ti}^\top\right) \\
&\succeq \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{ti}^\top + x_{tj} x_{tj}^\top\right) \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) \left(1 - \sum_{j\in S_t} p_{tj}(\bar{\theta})\right) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top
\end{aligned}$$

where $p_{t0}(\bar{\theta})$ is the probability of choosing the outside option. Now,

$$\begin{aligned}
G_n(\theta_1) - G_n(\theta_2) &= \left[\sum_{t=1}^n H_t + \lambda I_d\right](\theta_1 - \theta_2) \\
&\geq \left[\sum_{t=1}^n \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top + \lambda I_d\right](\theta_1 - \theta_2) \\
&:= \mathcal{H}(\bar{\theta})(\theta_1 - \theta_2).
\end{aligned}$$

Consider some $\bar{\theta} \in \mathbb{R}^d$. From Assumption 2, $p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta})$ is lower-bounded by κ. Then we have

$$(\theta_1 - \theta_2)^\top (G_n(\theta_1) - G_n(\theta_2)) \geq (\theta_1 - \theta_2)^\top (\kappa V_n)(\theta_1 - \theta_2) > 0$$

for any $\theta_1 \neq \theta_2$. Therefore, $G_n(\theta)$ is an injection from $\mathbb{R}^d$ to $\mathbb{R}^d$, and so G^{-1} is a well-defined function. By the definition, $G_n(\theta^*) = 0$. Hence, for any $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|G_n(\theta)\|_{V_n^{-1}}^2 &= \|G_n(\theta) - G_n(\theta^*)\|_{V_n^{-1}}^2 \\
&= \left(G_n(\theta) - G_n(\theta^*)\right)^\top V_n^{-1} \left(G_n(\theta) - G_n(\theta^*)\right) \\
&\geq (\theta - \theta^*)^\top \mathcal{H}(\bar{\theta}) V_n^{-1} \mathcal{H}(\bar{\theta})(\theta - \theta^*) \\
&\geq \kappa^2 (\theta - \theta^*)^\top V_n (\theta - \theta^*) \\
&= \kappa^2 \|\hat{\theta} - \theta^*\|_{V_n}^2
\end{aligned}$$

where the last inequality is from $\mathcal{H}(\bar{\theta}) \succeq \kappa V_n$. Now, recall for $\hat{\theta}$ which is the solution to Eq.(7), $G_n(\hat{\theta}) = Z_n - \lambda\theta^*$ where $Z_n = \sum_{t=1}^n \sum_{i\in S_t} \epsilon_{ti} x_{ti}$. Hence, we have

$$\kappa \|\hat{\theta} - \theta^*\|_{V_n} \leq \|G_n(\hat{\theta})\|_{V_n^{-1}} \leq \|Z_n\|_{V_n^{-1}} + \lambda \|\theta^*\|_{V_n^{-1}}$$

Then we can use Theorem 1 in [1], which states if the noise ϵ_{ti} is sub-Gaussian with parameter σ (with $\sigma = \frac{1}{2}$ in our problem), then

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

with probability at least $1 - \delta$. Then we combine with Lemma 9. So it follows that

$$\|Z_n\|_{V_n^{-1}}^2 \leq 2\sigma^2 \left[\frac{d}{2} \log\left(\frac{\text{trace}(V) + nK}{d}\right) - \frac{1}{2}\log\det(V) + \log\frac{1}{\delta}\right].$$

Since $V = \lambda I_d$, it follows that

$$\begin{aligned}\|Z_n\|^2_{V_n^{-1}} &\leq 2\sigma^2 \left[\frac{d}{2}\log\left(\frac{d\lambda + nK}{d}\right) - \frac{1}{2}\log\lambda^d + \log\frac{1}{\delta}\right] \\ &= 2\sigma^2 \left[\frac{d}{2}\log\left(\lambda + \frac{nK}{d}\right) - \frac{d}{2}\log\lambda + \log\frac{1}{\delta}\right] \\ &= 2\sigma^2 \left[\frac{d}{2}\log\left(1 + \frac{nK}{d\lambda}\right) + \log\frac{1}{\delta}\right].\end{aligned}$$

Then for $\|\theta^*\|_{V_n^{-1}}$, we have

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

Hence, $\lambda\|\theta^*\|_{V_n^{-1}} \leq \sqrt{\lambda}$ since $\|\theta^*\| \leq 1$. Combining the results and using the fact that $\sigma = \frac{1}{2}$ for our problem, we have that

$$\|\hat{\theta}_n - \theta^*\|_{V_n} \leq \frac{1}{2\kappa}\sqrt{d\log\left(1 + \frac{nK}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \frac{\sqrt{\lambda}}{\kappa}.$$

with probability at least $1 - \delta$.

□

C.3 Proof of Lemma 5

Proof. First, define event $\hat{\mathcal{E}}_t = \{\|\theta^* - \hat{\theta}_t\|_{V_t} \leq \alpha_t\}$, i.e. the regularized MLE estimate concentrates properly to θ^* in rounds t. From Lemma 4, this concentration event holds with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$ for each round t. On $\hat{\mathcal{E}}_t$, we show $x_{ti}^\top\theta^* \leq x_{ti}^\top\hat{\theta}_t + \alpha_t\|x_{ti}\|_{V_t^{-1}}$ for all i.

$$\begin{aligned}|x_{ti}^\top\hat{\theta}_t - x_{ti}^\top\theta^*| &= \left|\left[V_t^{-1/2}(\hat{\theta}_t - \theta^*)\right]^\top (V_t^{-1/2}x_{ti})\right| \\ &\leq \left\|V_t^{-1/2}(\hat{\theta}_t - \theta^*)\right\| \left\|V_t^{-1/2}x_{ti}\right\| \\ &= \|\hat{\theta}_t - \theta^*\|_{V_t}\|x_{ti}\|_{V_t^{-1}} \\ &\leq \alpha_t\|x_{ti}\|_{V_t^{-1}}\end{aligned}$$

where the first inequality is by Hölder's inequality. Hence, it follows that

$$x_{ti}^\top\theta^* - \left(x_{ti}^\top\hat{\theta}_t + \alpha_t\|x_{ti}\|_{V_t^{-1}}\right) \leq 0$$

for all i. Hence, using the restricted monotonicity in Lemma 1, if event $\hat{\mathcal{E}}_t$ holds, then we have

$$R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \leq 0.$$

Then we have

$$\begin{aligned}\mathbb{E}\big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\big] &\leq \mathbb{E}\left[\Big(R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t)\Big)\mathbb{1}(\hat{\mathcal{E}}_t) \mid \mathcal{F}_t\right] + \mathbb{E}\left[\mathbb{1}(\hat{\mathcal{E}}_t^c) \mid \mathcal{F}_t\right] \\ &\leq 0 + \mathcal{O}(t^{-2}).\end{aligned}$$

Therefore, summing over all $t \leq T$, we have

$$\sum_{t=1}^{T}\mathbb{E}\big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\big] \leq 0 + \sum_{t=1}^{T}\mathcal{O}(t^{-2}) = \mathcal{O}(1).$$

□

C.4 Proof of Lemma 6

The proof of Lemma 6 requires the following three technical lemmas.

Lemma 7. *Let $x_{ti} \in \mathbb{R}^d$. Then we have*

$$\det\left(I + \sum_{i=1}^{n} x_{ti} x_{ti}^\top\right) \geq 1 + \sum_{i=1}^{n} \|x_{ti}\|_2^2 \tag{8}$$

Proof. Let $\lambda_1, \lambda_2, ..., \lambda_d$ be the eigenvalues of $\sum_{i=1}^{n} x_{ti} x_{ti}^\top$. Since $\sum_{i=1}^{n} x_{ti} x_{ti}^\top$ is positive semi-definite, $\lambda_j \geq 0$ for all j. Hence,

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

□

Lemma 8. *Suppose $\|x_{ti}\| \leq 1$ for all t and i. Define $V_t = V + \sum_{\tau=1}^{t} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$ with $V = \lambda I_d$. If $\lambda \geq 1$, then*

$$\sum_{\tau=1}^{t} \max_{i \in S_\tau} \|x_{\tau i}\|_{V_\tau^{-1}}^2 \leq 2 \log\left(\frac{\det(V_t)}{\lambda_{\min}(V)^d}\right).$$

Proof.

$$\begin{aligned}
\det(V_t) &= \det\left(V_{t-1} + \sum_{i \in S_t} x_{ti} x_{ti}^\top\right) \\
&= \det(V_{t-1}) \det\left(I + \sum_{i \in S_t} V_{t-1}^{-1/2} x_{ti} (V_{t-1}^{-1/2} x_{ti})^\top\right) \\
&\geq \det(V_{t-1}) \left(1 + \sum_{i \in S_t} \|x_{ti}\|_{V_{t-1}^{-1}}^2\right) \\
&\geq \det(V) \prod_{\tau=1}^{t} \left(1 + \sum_{i \in S_\tau} \|x_{\tau i}\|_{V_\tau^{-1}}^2\right) \\
&\geq \det(V) \prod_{\tau=1}^{t} \left(1 + \max_{i \in S_\tau} \|x_{\tau i}\|_{V_\tau^{-1}}^2\right)
\end{aligned} \tag{9}$$

The first inequality comes from Lemma 7. The second inequality comes from applying the first inequality repeatedly.

Let $\lambda_{\min}(V_t)$ be the minimum eigenvalue of V_t. We have

$$\max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}^2 \leq \max_{i \in S_t} \frac{\|x_{ti}\|^2}{\lambda_{\min}(V_t)} \leq \frac{1}{\lambda_{\min}(V)} = \frac{1}{\lambda}.$$

Since $\lambda \geq 1$, using the fact that $z \leq 2\log(1+z)$ for any $z \in [0,1]$, we have

$$\begin{aligned}\sum_{\tau=1}^{t} \max_{i \in S_\tau} \|x_{\tau i}\|^2_{V_\tau^{-1}} &\leq 2\sum_{\tau=1}^{t} \log\left(1 + \max_{i \in S_\tau} \|x_{\tau i}\|^2_{V_\tau^{-1}}\right) \\ &= 2\log \prod_{\tau=1}^{t} \left(1 + \max_{i \in S_\tau} \|x_{\tau i}\|^2_{V_\tau^{-1}}\right) \\ &\leq 2\log\left(\frac{\det(V_t)}{\det(V)}\right)\end{aligned}$$

The last inequality is from (9)

□

Lemma 9. *Suppose* $\|x_{ti}\| \leq 1$ *for all* t*. Then* $\det(V_t)$ *is increasing with respect to* t *and*

$$\det(V_t) \leq \left(\frac{\mathit{trace}(V) + tK}{d}\right)^d \tag{10}$$

Proof. For any symmetric positive definite matrix $\widetilde{V} \in \mathbb{R}^{d \times d}$ and column vector $x \in \mathbb{R}^d$, we have

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

The second equality above is due to Sylvester's determinant theorem, which states that $\det(I+BA) = \det(I+AB)$. Let $\lambda_1, ..., \lambda_d > 0$ be the eigenvalues of V_t. Then

$$\begin{aligned}\det(V_t) &\leq \left(\frac{\lambda_1 + ... + \lambda_d}{d}\right)^d \\ &= \left(\frac{\text{trace}(V_t)}{d}\right)^d \\ &= \left(\frac{\text{trace}(V) + \sum_{\tau=1}^{t}\sum_{i \in S_\tau} \text{trace}(x_{\tau i} x_{ti}^\top)}{d}\right)^d \\ &= \left(\frac{\text{trace}(V) + \sum_{\tau=1}^{t}\sum_{i \in S_\tau} \|x_{\tau i}\|^2}{d}\right)^d \\ &\leq \left(\frac{\text{trace}(V) + tK}{d}\right)^d.\end{aligned}$$

□

Proof of Lemma 6. Combining Lemma 8 and Lemma 9, we have that

$$\begin{aligned}\sum_{\tau=1}^{t} \max_{i \in S_\tau} \|x_{\tau i}\|^2_{V_\tau^{-1}} &\leq 2\log\left(\frac{\det(V_t)}{\det(V)}\right) \\ &\leq 2\log\left[\left(\frac{\text{trace}(V) + tK}{d}\right)^d \frac{1}{\det(V)}\right] \\ &\leq 2d\log\left(1 + \frac{tK}{d\lambda}\right).\end{aligned}$$

Then applying the Cauchy-Schwarz inequality, we have

$$\sum_{\tau=1}^{t} \max_{i \in S_\tau} \|x_{\tau i}\|_{V_\tau^{-1}} = \sqrt{2dt\log\left(1 + \frac{tK}{d\lambda}\right)}.$$

□

D Proof of Theorem 2: Worst-case Regret Analysis

We first decompose the cumulative regret, similar to the procedure in previous sections but this time using $\widetilde{R}_t(S_t)$. In the following sections, we derive the bounds for $\mathcal{R}_1(t)$ and $\mathcal{R}_2(t)$ separately.

$$\mathcal{R}(T) = \underbrace{\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta^*) - \widetilde{R}_t(S_t)]}_{\mathcal{R}_1(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[\widetilde{R}_t(S_t) - R_t(S_t, \theta^*)]}_{\mathcal{R}_2(T)}$$

D.1 Bounding $\mathcal{R}_2(T)$.

We can control $\mathcal{R}_2(T)$ by showing that both MLE $\hat{\theta}_t$ and TS parameters $\{\widetilde{\theta}_t\}$ concentrate appropriately. To show each of these concentration results, we first further decompose $\mathcal{R}_2(T)$:

$$\mathcal{R}_2(T) = \sum_{t=1}^{T} \mathbb{E}[\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t)] + \sum_{t=1}^{T} \mathbb{E}[R_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*)]. \tag{11}$$

The second term deals with the estimation error and can be bounded by the concentration of $\hat{\theta}_t$ in Lemma 4 and the Lipschitz-like property in Lemma 3, i.e., with probability $1 - \mathcal{O}(t^{-2})$, we have

$$R_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \leq \max_{i \in S_t} \left| x_{ti}^\top (\hat{\theta}_t - \theta^*) \right| \leq \alpha_t \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}. \tag{12}$$

The first term in Eq.(11) deals with the random sampling of $\{\widetilde{\theta}_t^{(j)}\}$. Again, we can bound the difference in expected revenue by the difference in utility estimates using Lemma 3: $\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \leq \max_{i \in S_t} (\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t)$. Then we are left to show that $\widetilde{u}_{ti}$ concentrates appropriately for all $i \in [N]$. The following lemma ensures the concentration of $\widetilde{u}_{ti}$.

Lemma 10. *Let* $\beta_t = \alpha_t \min\left(\sqrt{4d\log(Mt)}, \sqrt{2\log(2M)} + \sqrt{4\log(Nt)}\right)$. *Then for all* $i \in [N]$,

$$\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t \leq \beta_t \|x_{ti}\|_{V_t^{-1}}.$$

with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$.

Remark 1. *Lemma 10 shows that the confidence radius* β_t *is larger than* α_t *by the factor of at most* $\sqrt{2d\log(Mt)}$. *The additional* $\sqrt{d}$ *factor comes from the oversampling of TS, which also appears in other TS methods for linear contextual bandit problems [5, 3].* $\sqrt{\log M}$ *factor comes from drawing optimistic samples where* $M = \mathcal{O}(\log K)$*; hence the marginal increase of the regret bound due to optimistic sampling is very small.*

Hence for the first term in Eq.(11), we have $\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \leq \beta_t \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}$ with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$. We combine with Eq.(12) to derive the bound for $\mathcal{R}_2(T)$:

$$\mathcal{R}_2(T) \leq \sum_{t=1}^{T} (\alpha_t + \beta_t) \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} + \sum_{t=1}^{T} \mathcal{O}\left(t^{-2}\right) \tag{13}$$

D.2 Bounding $\mathcal{R}_1(T)$.

As discussed in Section 5, a sufficient condition for ensuring the success of TS is to show the probability of TS samples being optimistic is high enough. The following lemma lower-bounds the probability that the expected revenue under sampled parameters is higher than the optimal expected revenue under the true parameter. The proof utilizes the anti-concentration property of Gaussian distribution.

Lemma 2 (restate). *Suppose* $\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t$ *and we take* $M = \lceil 1 - \frac{\log K}{\log\left(1 - 1/(4\sqrt{e\pi})\right)} \rceil$ *samples. Then we have*

$$\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) \geq \frac{1}{4\sqrt{e\pi}}. \tag{14}$$

Using this frequent optimistic sampling, we can ensure that the regret due to the oversampling is not too large.

Lemma 12. *Let $\widetilde{p} = \frac{1}{4\sqrt{e\pi}}$. Then, we have*

$$\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)] \leq \frac{4\beta_T}{\widetilde{p}} \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T} \right) + \mathcal{O}(1)$$

D.3 Combining the results

Applying Lemma 6 to the bound for $\mathcal{R}_2(T)$ in Eq.(13) and combining with Lemma 12, we have the final bound for the worst-case cumulative regret.

$$\mathcal{R}(T) \leq (\alpha_T + \beta_T)\sqrt{2dT \log(T/d)} + 16\sqrt{e\pi}\beta_T \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T} \right) + \mathcal{O}(1)$$

E Proofs of Lemmas for Theorem 2

E.1 Proof of Lemma 10

Proof. Given $\mathcal{F}_t$, each of Gaussian random variable $x_{ti}^\top \widetilde{\theta}_t^{(j)}$ has mean $x_{ti}^\top \hat{\theta}_t$ and standard deviation $\alpha_t \|x_{ti}\|_{V_t^{-1}}$.

$$\begin{aligned}
|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| &= \alpha_t \|x_{ti}\|_{V_t^{-1}} \frac{\left|\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t\right|}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \left| \frac{x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \right| \\
&= \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j |Z_j|
\end{aligned}$$

where each Z_j is a standard normal random variable. Using the result from Lemma 13, we have $\max_j |Z_j| \leq \sqrt{2\log(2M)} + \sqrt{4\log t}$ with probability at least $1 - \frac{1}{t^2}$. Then, for all $i \in [N]$,

$$|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| \leq \left(\sqrt{2\log(2M)} + \sqrt{4\log(Nt)} \right) \alpha_t \|x_{ti}\|_{V_t^{-1}}$$

with probability at least $1 - \frac{1}{t^2}$. Alternatively, let $m = \arg\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)}$. Then we can write

$$\begin{aligned}
|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| &= \left| \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t \right| \\
&= \left| x_{ti}^\top (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right| \\
&= \left| x_{ti}^\top V_t^{-1/2} V_t^{1/2} (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right| \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \left\| \alpha_t^{-1} V_t^{1/2} (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right\| \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \left\| \alpha_t^{-1} V_t^{1/2} (\widetilde{\theta}_t^{(j)} - \hat{\theta}_t) \right\| \\
&= \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \|\zeta_j\|
\end{aligned}$$

where each element in $\zeta_j \in \mathbb{R}^d$ is a univariate standard normal variable $\mathcal{N}(0,1)$. Hence, each $\|\zeta_j\| \leq \sqrt{4d\log t}$ with probability at least $1 - \frac{1}{t^2}$. Using the union bound for all $j \in \{1, ..., M\}$, we have with probability at least $1 - \frac{1}{t^2}$

$$|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| \leq \sqrt{4d\log(Mt)} \alpha_t \|x_{ti}\|_{V_t^{-1}}.$$

□

Lemma 13. *Let $Z_i \sim \mathcal{N}(0,1), i = 1, ..., n$ be a standard Gaussian random variable. Then we have*

$$\mathbb{P}\left(\max_i |Z_i| \leq \sqrt{2\log(2n)} + \sqrt{2\log\frac{1}{\delta}}\right) \geq 1 - \delta.$$

Proof. Using the Chernoff bound, for each Z_i, we have

$$\mathbb{P}(|Z_i| > \epsilon) \leq 2e^{-\epsilon^2/2}.$$

Applying the union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\max_i |Z_i| > \sqrt{2\log(2n)} + \epsilon\right) &\leq 2n \exp\left(-(\sqrt{2\log(2n)} + \epsilon)^2/2\right) \\
&= 2n \exp(-\log(2n) - \epsilon\sqrt{2\log(2n)} - \epsilon^2/2) \\
&\leq e^{-\epsilon\sqrt{2\log(2n)}} e^{-\epsilon^2/2} \\
&\leq e^{-\epsilon^2/2}.
\end{aligned}$$

Letting $\delta = e^{-\epsilon^2/2}$, we have the result. □

E.2 Proof of Lemma 2

Proof. Given $\mathcal{F}_t$, each of Gaussian random variable $x_{ti}^\top \widetilde{\theta}_t^{(j)}$ has mean $x_{ti}^\top \hat{\theta}_t$ and standard deviation $\alpha_t \|x_{ti}\|_{V_t^{-1}}$. Hence, for each $i \in S_t^*$, we have

$$\begin{aligned}
\mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^* \mid \mathcal{F}_t\right) &= 1 - \mathbb{P}\left(x_{ti}^\top \widetilde{\theta}_t^{(j)} \leq x_{ti}^\top \theta^*, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right) \\
&= 1 - \mathbb{P}\left(\frac{x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \leq \frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}}, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right) \\
&= 1 - \mathbb{P}\left(Z_j \leq \frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}}, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right)
\end{aligned}$$

where Z_j is a standard normal random variable. By the assumption, we have $|x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t| \leq \alpha_t \|x_{ti}\|_{V_t^{-1}}$ for all i, Hence, we can bound the RHS term within the probability.

$$\frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \leq \frac{\alpha_t \|x_{ti}\|_{V_t^{-1}}}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} = 1$$

Then, it follows that

$$\mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^* \mid \mathcal{F}_t\right) \geq 1 - \left(\mathbb{P}(Z \leq 1)\right)^M. \tag{15}$$

Now, since $S_t = \arg\max_S \widetilde{R}_t(S)$, we have $\widetilde{R}_t(S_t) \geq \widetilde{R}_t(S_t^*)$. Then combining with Lemma 1, we can lower-bound the probability of having an expected revenue optimistic under the sampled parameter (the second inequality below).

$$\begin{aligned}
\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) &\geq \mathbb{P}\left(\widetilde{R}_t(S_t^*) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) \\
&\geq \mathbb{P}\left(\widetilde{u}_{ti} > x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right) \\
&= \mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right) \\
&\geq 1 - K\left(\mathbb{P}(Z \leq 1)\right)^M
\end{aligned}$$

where the last inequality comes from Eq.(15) and the union bound. Using the anti-concentration inequality in Lemma 15, we have $\mathbb{P}(Z \leq 1) \leq 1 - \frac{1}{4\sqrt{e\pi}}$. Hence, it follows that

$$\begin{aligned}\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) &\geq 1 - K\left(1 - \frac{1}{4\sqrt{e\pi}}\right)^M \\ &\geq 1 - \left(1 - \frac{1}{4\sqrt{e\pi}}\right) \\ &= \frac{1}{4\sqrt{e\pi}}\end{aligned}$$

where the second inequality comes from our choice of $M = \lceil 1 - \frac{\log K}{\log\left(1 - 1/(4\sqrt{e\pi})\right)} \rceil$ which implies $\left(1 - \frac{1}{4\sqrt{e\pi}}\right)^M \leq \frac{1}{K}\left(1 - \frac{1}{4\sqrt{e\pi}}\right)$.

□

E.3 Proof of Lemma 12

Proof. The proof is inspired by the techniques used for Theorem 1 in [3]. First, we define $\widetilde{\Theta}_t$ the set of parameter samples for which the expected revenue concentrate appropriately to the expected revenue based on the MLE parameter. Also, we define the set of optimistic parameter samples $\widetilde{\Theta}_t^{\text{opt}}$ which coinciding with $\widetilde{\Theta}_t$.

$$\begin{aligned}\widetilde{\Theta}_t &:= \left\{\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M : \widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \leq \beta_t \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}\right\} \\ \widetilde{\Theta}_t^{\text{opt}} &:= \left\{\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M : \widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*)\right\} \cap \widetilde{\Theta}_t\end{aligned}$$

Define the event $\mathcal{E}_t$ that both $x_{ti}^\top \hat{\theta}_t$ and $\widetilde{u}_{ti}$ are concentrated around their respective means.

$$\mathcal{E}_t = \{x_{ti}^\top \hat{\theta}_t - x_{ti}^\top \theta^* \leq \alpha_t \|x_{ti}\|_{V_t^{-1}}, \forall i\} \cap \{\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t \leq \beta_t \|x_{ti}\|_{V_t^{-1}}, \forall i\}.$$

For any $\widetilde{\theta}_t^{1:M} := \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M \in \widetilde{\Theta}_t^{\text{opt}}$, we have

$$\left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \mathbb{1}(\mathcal{E}_t) \leq \left(R_t(S_t^*, \theta_t^*) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \widetilde{R}_t(S_t, \theta_t^{1:M})\right) \mathbb{1}(\mathcal{E}_t)$$

where $\widetilde{R}_t(S_t, \theta_t^{1:M})$ is the optimistic expected revenue under the sampled parameters $\theta_t^{1:M}$. Then we can bound $R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)$ by the expectation over any random choice $\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}$

$$\begin{aligned}R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t) &\leq \mathbb{E}\left[\left(\widetilde{R}_t(S_t) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \widetilde{R}_t(S_t, \theta_t^{1:M})\right) \mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}\right] \\ &= \mathbb{E}\left[\sup_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \left(\widetilde{R}_t(S_t) - \widetilde{R}_t(S_t, \theta_t^{1:M})\right) \mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}\right] \\ &\leq \mathbb{E}\left[\sup_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \max_{i \in S_t} \left|\widetilde{u}_{ti} - x_{ti}^\top \theta_t^{(j)}\right| \mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}\right] \\ &\leq 2\beta_t \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t)\end{aligned}$$

where the last inequality is from the definition of the set $\widetilde{\Theta}_t$ and $S_t(\widetilde{\theta}_t^{1:M})$ stands for the optimal assortment under the sampled parameters $\widetilde{\theta}_t^{1:M} = \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$.

From Lemma 2, we have $\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) \geq \frac{1}{4\sqrt{e\pi}} =: \widetilde{p}$. Therefore it follows that

$$\begin{aligned}
\mathbb{P}\left(\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \mid \mathcal{F}_t, \mathcal{E}_t\right) &= \mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \text{ and } \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t, \mathcal{E}_t\right) \\
&\geq \mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) - \mathbb{P}\left(\widetilde{\theta}_t^{1:M} \notin \widetilde{\Theta}_t, \mathcal{E}_t\right) \\
&\geq \widetilde{p} - \mathcal{O}(t^{-1}) \\
&\geq \widetilde{p}/2.
\end{aligned}$$

Now, note that we can write

$$\begin{aligned}
\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \mathcal{E}_t\right] &\geq \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \mathbb{P}\left(\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \mid \mathcal{F}_t, \mathcal{E}_t\right) \\
&\geq \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \cdot \widetilde{p}/2
\end{aligned}$$

Therefore, combining the results, we have

$$\begin{aligned}
R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t) &\leq 2\beta_t \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t) \\
&\leq \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t) \\
&\leq \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right].
\end{aligned}$$

Summing over all t and taking the failure event into consideration, we have

$$\sum_{t=1}^{T} \left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \leq \sum_{t=1}^{T} \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right].$$

Here, the summation on the RHS contains an expectation, so we cannot directly apply Lemma 6. Instead, we use Lemma 14 to bound the sum of the expectations

$$\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)] \leq \sum_{t=1}^{T} \frac{4\beta_t}{\widetilde{p}} \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T}\right) + \mathcal{O}(1).$$

□

Lemma 14. *If $\lambda_{\min}(V_t) \geq \lambda$, then with probability $1 - \mathcal{O}(T^{-1})$ we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T}.$$

Proof. We rewrite the summation as follows.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \\
&= \sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} + \sum_{t=1}^{T} \left(\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}\right) \qquad (16)
\end{aligned}$$

The first summation can be bounded by using Lemma 6 and Cauchy-Schwarz inequality.

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}^2} \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} \qquad (17)$$

For the second summation in Eq.(16), we can apply Azuma-Hoeffding inequality (Lemma 16). Note that the second summation is a martingale by construction. Also recall that $\max_{i \in S_t} \|x_{ti}\| \leq 1$ for all t, hence we have

$$\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \frac{2}{\lambda_{\min}(V_t)} \leq \frac{2}{\lambda_{\min}(V)} = \frac{2}{\lambda}.$$

Therefore, $\frac{2}{\lambda}$ is an upper-bound for each element in the second summation. Now applying Azuma-Hoeffding inequality, we have

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \right) \leq \sqrt{\frac{8T}{\lambda} \log 2T} \tag{18}$$

with probability $1 - \mathcal{O}(T^{-1})$. Combining Eq.(17) and Eq.(18), we have the result. □

E.4 Other Lemmas

The following lemma is used to derive the concentration and anti-concentration inequalities for Gaussian random variables.

Lemma 15 (Abramowitz and Stegun 4)**.** *For a Gaussian random variable Z with mean μ and variance σ^2, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi}z} e^{-z^2/2} \leq \mathbb{P}\left(|Z - \mu| > z\sigma\right) \leq \frac{1}{\sqrt{\pi}z} e^{-z^2/2}. \tag{19}$$

Lemma 16 (Azuma-Hoeffding inequality)**.** *If a super-martingale $(Y_t; t \geq 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, ..., T$, then for any $a \geq 0$,*

$$\mathbb{P}(Y_T - Y_0 \geq a) \leq 2e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}$$

F Guarantees for Random Initialization

As we discussed briefly in Section A, TS-MNL can start with the random initialization phase where the agent randomly chooses an assortment S_t instead of using regularization in the parameter estimation. However, the length of the initialization T_0 needs to be specified in order to ensure a unique solution of MLE for a theoretical guarantee.

We maintain $V_{T_0} = \sum_{\tau=1}^{T_0} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$ while choosing assortments randomly during the random initialization. The initialization duration T_0 is chosen to ensure that $\lambda_{\min}(V_{T_0})$ is large enough so that V_{T_0} is invertible. The following proposition allows us to find such T_0.

Proposition 1. *Let $x_{\tau i}$ be drawn i.i.d. from some distribution with $\|x_{\tau i}\| \leq 1$ and $\mathbb{E}[x_{\tau i} x_{\tau i}^\top] \geq \sigma_0$. Define $V_{T_0} = \sum_{\tau=1}^{T_0} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$, where T_0 is the length of random initialization. Suppose we run a random initialization with assortment size K for duration T_0 which satisfies*

$$T_0 \geq \frac{1}{K} \left(\frac{C_1 \sqrt{d} + C_2 \sqrt{\log T}}{\sigma_0} \right)^2 + \frac{2B}{K\sigma_0}$$

for some positive, universal constants C_1 and C_2. Then, $\lambda_{\min}(V_{T_0}) \geq B$ with probability at least $1 - T^{-1}$.

The proposition is the adaptation of Proposition 1 in [30], modified for our multinomial setting. If we use $B = K$, then the proposition implies that we can have $\lambda_{\min}(V_{T_0}) \geq K$ with a high probability if we run the initialization for $\mathcal{O}(\sigma_0^{-2}(d + \log T))$ rounds. Similar to [23] and [30], the i.i.d. assumption on the context x_{ti} may be only needed to ensure that V_τ is invertible at the end of the initialization phase. Hence, after the initialization, x_{ti} can even be chosen adversarially as long as $\|x_{ti}\|$ is bounded.

G Numerical Study Details

G.1 Synthetic Experiments

For synthetic experiments, we first sample feature vectors x_i for each $i \in [N]$ in $d-1$ dimension with each entry from the standard Gaussian distribution. We then normalize this vectors and add an extra dimension with constant 1 for the intercept and divide by $\sqrt{2}$ so that the ℓ_2 norm of feature vectors is bounded, i.e., $\|x_i\| \leq 1$. Similarly, we sample the parameter θ^* from the d-dimensional standard multivariate Gaussian distribution but without the normalization. For each experimental instance, we draw new samples of $\{x_i\}$ and θ^*.

In Figure 1, we only showed the performance of the UCB algorithm proposed in [15] for $N = 256$. The UCB algorithm proposed in [15] constructs confidence bounds for each of (N choose K) assortments (as discussed in Section 2), the evaluation on a larger N causes a significant computational burden; hence we had to keep N at a reasonable size for evaluating the UCB method. In fact, even with $N = 256$, we could not use the original version of the UCB method in [15] due to the computational complexity. We instead use a greedy heuristic for solving the combinatorial optimization proposed as an alternative efficient approximation (see Algorithm 4 in [15]) although it does not have rigorous guarantees. However, it is important to note that even with such computational compromises for the UCB method, our TS methods still have better computational efficiency as well as superior performances on the statistical efficiency. Note that our proposed methods do not suffer from this issue and can be evaluated with a much larger N which is shown in the experiments in Figure 2 as well as the MovieLens experiment (with $N = 1000$) in Figure 1.

The left plot in Figure 2 shows the evaluations of TS-MNL with optimistic sampling with varying feature dimensions. The reported results are averaged over 40 independent instances. The results show that the performance of our algorithm is still attractive even with an increase in the feature dimension, which shows a better scalability in d than the theoretical guarantees, at most $d^{3/2}$ dependence on the worst-case regret bound.

Furthermore, the experiments in the right plot of Figure 2 show that even when the number of total items N increases, the empirical performances of our proposed algorithms remain the same as the performance in Figure 1 and are not hindered by such an increase in N. This observation is consistent with our established theoretical results and supports the claim that our methods can be used and effective for problem instances with very large N — as long as the combinatorial optimization step can be efficiently computed.

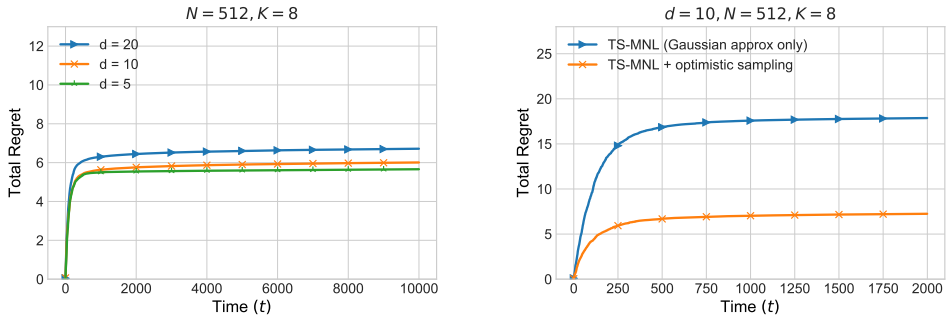

Figure 2: Experiments on varying feature dimensions and with an increased number of items

G.2 Experiments with MovieLens Dataset

Dataset. MovieLens datasets[3] contain the ratings of users for movies from the MovieLens website. The datasets come in different sizes and we use "MovieLens 20M" for our experiments. This dataset contains 20 million ratings of 2.7×10^4 movies by 1.38×10^5 users.

Feature extraction. We follow the experimental setup of [31]. For our experiments, we use $N = 1000$ movies with most ratings and 1.1×10^3 user with the most number of ratings. We

[3] https://grouplens.org/datasets/movielens/

randomly split the user set into two parts A_1 and A_2 with $|A_1| = 100$ and $|A_2| = 1000$. Then we use the matrix of the movie rating for users in A_1 to extract feature vectors with $d = 5$. Note that the MovieLens dataset does not come with movie or user features — it only contains ratings of the movies by the user as a matrix, which we denote as W. Hence, we construct features for our experiments using the collaborative filtering approach. We derive the features of movies using low-rank matrix factorization.

Splitting the user set into two parts A_1 and A_2 means dividing the rows of the matrix W into two matrices: one with 100 rows corresponding to A_1 and the other with 1000 rows corresponding to A_2. We define training matrix $W_{\text{train}} \in \mathbb{R}^{|A_1| \times N}$ and test matrix $W_{\text{test}} \in \mathbb{R}^{|A_2| \times N}$ corresponding to user sets A_1 and A_2 respectively. We use W_{train} to learn the features of items and W_{test} to evaluate our learning algorithms.

Let $W_{\text{train}} \approx U \Sigma V^\top$ be rank-d truncated SVD of W_{train}, where $U \in \mathbb{R}^{|A_1| \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$,and $V \in \mathbb{R}^{N \times d}$. Then the features of movies are the rows of $V\Sigma$. Note that the matrix V in the section is defined within this experimental setup only and is different from the gram matrix V_t used in the regret analysis or in the algorithm. We overload this term for the sake of consistency with terms typically used in matrix factorization literature.

Offline regression. We use the extracted features of movies, i.e., rows of $\bar{V}\Sigma$, and the mean score of each of the movies considered. The true parameter θ^* is computed by solving the linear system of N with respect to the rating matrix of W_{test}.

Evaluations. Once we extract the features and learn θ^*, we perform online evaluations. We set λ to be the same as the feature dimension d. Note that the dataset does not contain separate revenue information for different movies. Hence we assume that the revenue parameter is uniform across all movies, i.e. each user choice/click is equally weighted. Therefore, the combinatorial optimization step reduces to sorting items according to estimated utilities and choosing top K movies. The evaluation results show that two variants of TS-MNL are effective. In particular, TS-MNL with optimistic sampling shows more attractive performances in these sample results.