A Contextual Bandit-based Recommendation Process for Maximizing Purchase Probability under a Multi-Choice Behavior Model

Theja Tulabandhula[a], Truong Vu[a]

[a] *University of Illinois Chicago, Chicago, IL 60607, United States*

Abstract

We examine a dynamic recommendation set optimization problem in which a decision-maker/platform presents a customer with a subset of products (i.e., a recommendation set) and observes the customer's response in each round. Customers can choose multiple products in each interaction, based on a behavior model that generalizes the contextual Multinomial Logit (MNL) model. In the contextual MNL model of customer behavior, a collection of features characterizes each of the products that can vary over time, and the product's mean utility is linear in the values of these features and an unknown parameter. Similar parameterization exists in our generalization as well. The platform's objective is to learn the customer behavior while maximizing the probability of purchase across customer interactions in a given time horizon T. For this more realistic customer behavior, we show how we can build on results for the MNL contextual bandit setting to provide a online recommendation algorithm with a provable regret upper bound that scales as $O(d\sqrt{T} + \kappa)$, where d is the dimension of the product feature vectors and κ is a problem-dependent constant that may have an exponential dependency on d. We complement this theoretical result with simulations that show the value of our recommendation algorithm for online learning under the contextual multi-choice customer behavior setting. Moreover, we also show through numerical tests that our technique performs robustly for changing κ and d values, providing supporting evidence for the form of the regret upper bound.

Email addresses: theja@uic.edu (Theja Tulabandhula), tvu25@uic.edu (Truong Vu)

November 1, 2025

Keywords: Multi-choice customer behavior, Contextual bandits, Dynamic recommendations, Marketing, Purchase probability

1. Introduction

Our work introduces a novel algorithmic framework designed to enhance recommendation systems in a bandit feedback setting. It specifically addresses the challenge of modeling customer choices in multi-choice scenarios and optimizing recommendation sets by leveraging a contextual bandit approach. Our methodology is rooted in the contextual multinomial logit (MNL) model, which is a popular model that predicts customer purchase probabilities under the single choice setting. The core contribution of this work lies in the development of the online recommendation setting under multi-choice and its corresponding algorithm, which integrates local variance and curvature information for optimizing recommendation sets. This approach not only improves the recommendations but also dynamically adapts to evolving customer preferences over time, leading to strong theoretical guarantees on its performance.

Related Work: Contextual bandit frameworks have been extensively used in online learning and recommendation problems, particularly for modeling user-item interactions under uncertainty. Foundational results on contextual and generalized linear bandits (Filippi et al. (2010); Abbasi-Yadkori et al. (2011); Auer et al. (2002a)) established regret bounds for single-choice decision settings. More recent advances have refined these methods for logistic and multinomial logit (MNL) choice models (Faury et al. (2020); Agrawal et al. (2023)), providing tighter confidence sets and improved regret guarantees. In parallel, the marketing and operations research literature has studied assortment optimization and multi-purchase behaviors (Kök et al. (2015); Tulabandhula et al. (2023)), demonstrating that real-world customers often select multiple items within a single session. However, while the contextual MNL bandit model is well understood for single-choice cases, extending it to realistic multi-choice decision processes — where customers may choose multiple complementary products — has received limited algorithmic attention, especially with theoretical performance guarantees.

Our Contributions: this paper bridges that gap by introducing a new multi-choice contextual bandit formulation that generalizes the classical MNL model to accommodate multiple selections per interaction. Our main contributions are as follows

1. Modeling Framework: We propose a generalized contextual multi-choice model that captures complex customer behavior where users may purchase multiple products simultaneously. This model extends the standard contextual MNL framework by introducing an interaction matrix Θ that captures pairwise preferences between items.
2. Algorithmic Development: We develop an online learning algorithm that transforms the multi-choice recommendation problem into an equivalent single-choice MNL instance over pseudo-products. This allows us to reuse efficient MNL-based estimation and exploration methods while addressing the combinatorial nature of the original problem.
3. Theoretical Analysis: We derive a provable regret bound that scales as $O(d\sqrt{T} + \kappa)$, where d is the feature dimension and κ reflects problem-dependent complexity. The bound demonstrates sublinear regret and robustness to parameter variations.
4. Empirical Validation: Through extensive simulations, we show that our algorithm consistently outperforms existing contextual MNL bandit baselines under multi-choice customer behaviors and remains stable across different parameter regimes.

Together, these results provide a theoretically grounded and practically effective solution to dynamic recommendation problems where users can make multiple selections, extending contextual bandit theory toward more realistic decision-making environments.

2. Problem Setting, Choice Model and the Objective

2.1. Problem Setting

We study the multi-choice contextual bandit problem, which is defined as follows. Let $\mathcal{N} = \{1, 2, ..., N\}$ denote our product universe. In each round $t = 1, 2, \dots, T$, the platform chooses a recommendation set $Q_t \subset \mathcal{N}$ which has a cardinality of at most K, i.e. $|Q_t| \leq K$ and offers it to the arriving customer in that round. Each arriving customer is associated with feature vectors $\{x_{t,i}\}_{i \in \mathcal{N}}$ associated with each product i, where $x_{t,i} \in \mathbb{R}^d$. The customer then chooses at most two products based on these preferences from this set, which is represented by $I \subseteq Q_t$. Based on the customer's choice, the platform obtains a reward $r_{t,I}$. The choice behavior is driven by a probabilistic model with underlying parameters, which are a priori unknown to the platform. The platform's objective is to learn the underlying parameters while minimizing *regret* across these T interactions. Here, (cumulative) regret is the difference between what the platform actually chose to offer the customers across the T rounds of interaction versus what the platform would chose to offer them in hindsight.

2.2. The Multi-Choice Model

The choice behavior we consider is a generalization of the popular MNL choice behavior (Kök et al., 2015), and was introduced in Tulabandhula et al. (2023) that captures users choosing more than one distinct product from a given set of recommendations. In particular, we focus on customers choosing at most two products, although the behavior model generalizes to an arbitrary subset size. Thus, in our setting, the probability of the customer arriving at step t with contextual vectors $\{x_{t,i}\}_{i\in\mathcal{N}}$ choosing no products, one product and two products is as follows. First, the probability of choosing no product is:

$$\mathbb{P}(\text{choosing no product} \,|\, Q_t) = \frac{1}{1 + \sum\limits_{l\in Q_t} \exp^{\theta^T x_{t,l}} + \sum\limits_{(k,l)\in Q_t^2} \exp^{x_{t,k}^T \Theta x_{t,l}}}, \qquad (1)$$

where $\theta \in \mathbb{R}^d$ and $\Theta \in \mathbb{R}^{d\times d}$ are two unknown (to the platform) time-invariant parameters, and $Q_t^2 := \{(k,l)|k < l \text{ and } k,l \in Q_t\}$. Next, the probability of choosing one product, say i, is:

$$\mathbb{P}(\text{choosing } i \,|\, Q_t) = \frac{\exp^{\theta^T x_{t,i}}}{1 + \sum\limits_{l\in Q_t} \exp^{\theta^T x_{t,l}} + \sum\limits_{(k,l)\in Q_t^2} \exp^{x_{t,k}^T \Theta x_{t,l}}}. \qquad (2)$$

Finally, when the customer chooses two products (i,j), then the probability is:

$$\mathbb{P}(\text{choosing } (i,j) \,|\, Q_t) = \frac{\exp^{x_{t,i}^T \Theta x_{t,j}}}{1 + \sum\limits_{l\in Q_t} \exp^{\theta^T x_{t,l}} + \sum\limits_{(k,l)\in Q_t^2} \exp^{x_{t,k}^T \Theta x_{t,l}}}. \qquad (3)$$

2.3. Rewards and the Platform's Objective

The platform is motivated to offer such a recommendation set that the customer's propensity to make a successful selection is high. The total purchase probability of the customer when shown the recommendation set Q_t, is defined as:

$$\mu(Q_t) := \sum_{i \in Q_t} \mathbb{P}(\text{choosing } i \,|\, Q_t) + \sum_{(i,j) \in Q_t^2} \mathbb{P}(\text{choosing } (i,j) \,|\, Q_t). \tag{4}$$

The platform does not know θ and Θ a priori. It sequentially makes recommendation set decisions, $Q_1, Q_2, \ldots, Q_T$ and realizes the sum $\sum\limits_{t=1}^{T} \mu(Q_t)$, which represents how well the platform did over the interaction horizon. The goal of the platform is to maximize this cumulative reward signal over time period T when compared to the best decisions in hindsight. That is, it wants to minimize the expected regret:

$$R_T = \sum_{t=1}^{T} \mathbb{E}\left[\mu(Q_t^*) - \mu(Q_t)\right], \tag{5}$$

where Q_t^* and Q_t are the optimal recommendation set and the actually offered recommendation set for the customer interacting at time t. The explicit form of the objective will be discussed below. For now, we remark that the regret depends on the true unknown parameters, the randomness in the choice realizations, and possible randomness in the platform's algorithm that makes decisions.

3. Solution Approach and the Algorithm

Our solution approach relies on a natural transformation of the multi-choice problem instance to the single-choice problem instance. While this does not change the optimization complexity faced by the platform in each interaction round, it allows for a natural reuse of the estimation sub-routine for the unknown parameters using the single-choice machinery. We first discuss the transformation and then introduce our algorithm. Its regret upper bound will be discussed in Section 4.

3.1. Transforming Multi-Choice Instance to Single-Choice

We do this transformation in two parts: (a) we first show how we can unify single and two-choice outcomes into a larger set of two-choice outcomes, and (b) we show how the multi-choice problem can be mapped to a larger single-choice problem instance.

Without loss of generality, we can assume that the set I chosen by the customer is always of size 2 or empty (no purchase). This is because, we can consider a dummy product $(N+1)$ such that when the customer choosing single product i, we can think if it as them choosing product $(i, N+1)$. This leads to the following expressions for the purchase probabilities:

$$\mathbb{P}(\text{choosing no product} \,|\, Q_t) = \frac{1}{1 + \sum\limits_{(k,l)\in Q_t^2} \exp^{x_{t,k}^T \Theta x_{t,l}}}, \tag{6}$$

and

$$\mathbb{P}(\text{choosing } I = (i,j) \,|\, Q_t) = \mu_I(Q_t, \Theta) = \frac{\exp^{x_{t,i}^T \Theta x_{t,j}}}{1 + \sum\limits_{(k,l)\in Q_t^2} \exp^{x_{t,k}^T \Theta x_{t,l}}}, \tag{7}$$

where we have redefined the parameter Θ to include both the original unknown parameters (this is straightforward and omitted). Along with Θ, $\{x_{t,i}\}$ and $\mathcal{Q}_t^2$ are also redefined as follows. First we redefine $\mathcal{N}$ to include the dummy product with $x_{t,N+1} \in \mathbb{R}^{d+1}$ being the one-hot encoded vector with a 1 in the $(d+1)^{th}$ coordinate. All $\{x_{t,i}\}$ are also embedded in $\mathbb{R}^{d+1}$ by adding an additional coordinate to each with value 0. Θ is defined as a matrix in $\mathbb{R}^{(d+1)\times(d+1)}$ that is obtained by augmenting the original Θ with a $(d+1)^{th}$ additional column representing the original θ, and and an additional row of length $d+1$ being all zeros. $\mathcal{Q}_t^2$ now includes unique pairs of products including the dummy-product pairings if any, viz., $(1, N+1), (2, N+1), \ldots$ based on the recommendation set Q_t.

Next, we do another natural, but key transformation, that makes this generalized choice behavior setting an instance of the canonical single choice contextual MNL model. The advantage of doing this transformation is that it lets us import the algorithmic and analytic machinery verbatim from the contextual MNL case and apply it here. The transformation involves mapping every pair of products that the customer can choose into singleton *pseudo-products*. That is, for any instance with $\mathcal{N}$ products (we assume this includes the dummy product moving forward) under the generalized choice behavior setting, we can define a new instance under the MNL model with $\mathcal{M}$ pseudo-products. Here, each pseudo-product in the new instance represents one tuple in the set $\{(k,l)|k < l \text{ and } k,l \in \mathcal{N}\}$. With a slight abuse of notation, we will denote the original problem instance using $\mathcal{N}$ and the transformed problem instance using $\mathcal{M}$.

The cardinality of product universe $\mathcal{M}$ is $\binom{N+1}{2}$. Next, each pseudo-product $m \in \mathcal{M}$ is associated with a vector $z_{t,m} \in \mathbb{R}^{(d+1)^2}$ that is equal to $\text{vec}(x_{t,i}x_{t,j}^T)$ for some original product pair (i,j). Here, the notation $\text{vec}(x_{t,i}x_{t,j}^T)$ stands for the vectorization of the rank-1 matrix $x_{t,i}x_{t,j}^T$. Further, the parameter of the new MNL model instance, represented by θ (by repurposing the symbol from before) is defined to be equal to $\text{vec}(\Theta) \in \mathbb{R}^{(d+1)^2}$. In other words, we now have an equivalent MNL problem instance over product universe $\mathcal{M}$ with contextual vectors $\{z_{t,m}\}$ for $m \in \mathcal{M}$, where $|\mathcal{M}| = M = \binom{N+1}{2}$ and the unknown MNL parameter of interest is $\theta \in \mathbb{R}^{(d+1)^2}$. Note that this reduction does not completely solve the platform's problem because the platform still needs to create recommendation sets in the original product universe $\mathcal{N}$, which turns out to be an NP-hard problem even in the unconstrained setting Tulabandhula et al. (2023) (more on this will be discussed in the next section).

Thus, for the rest of the paper, we will use both the transformed representation of the problem instance $\mathcal{M}$ and the original representation $\mathcal{N}$ as necessary. In other words, although our focus is on the generalized multi-choice behavior given a product universe, we will be equivalently consider the transformed problem instance under single-choice (MNL) with a derived pseudo-product universe $\mathcal{M}$ to discuss the development of an online learning algorithm, and make it clear the parts of the algorithm where switching to the original representation is necessary.

3.2. The Algorithm

At each round t, the attribute vectors (contexts) $\{x_{t,1}, x_{t,2}, \cdots, x_{t,N}\}$ are made available to the platform, which are suitably converted into attribute vectors for the pseudo-products as shown in UCB-MNL (see Alg 1 below). The algorithm calculates an estimate of the true parameter θ_* according to Eq (14). The algorithm keeps track of the confidence set $C_t(\delta)$ $(E_t(\delta))$ as defined in Eq (17) (Eq (15). Let the set $\mathcal{A}$ contain all feasible recommendation sets of $\mathcal{N}$ with cardinality up to K. The algorithm makes the following decision:

$$(\mathcal{Q}_t, \text{vec}(\Theta)_t) = \underset{A_t \in \mathcal{A}, \text{vec}(\Theta) \in C_t(\delta)}{\text{argmax}} \mu(\mathbf{X}_{A_t}^\top, \text{vec}(\Theta)). \tag{8}$$

```
Algorithm 1 UCB-MNL
Input: Product universe $\mathcal{M}$ of size $M$, parameters for estimation and uncertainty set construction (defined later), set $\mathcal{A}$ of all feasible recommendation sets of the original product universe $\mathcal{N}$ with cardinality at most $K$.
for $t \geq 1$ do
    Get $\{x_{t,1}, x_{t,2}, \cdots, x_{t,N+1}\}$, each in $\mathbb{R}^{d+1}$ and convert them to $\{z_{t,1}, z_{t,2}, \cdots, z_{t,M}\}$, each in $\mathbb{R}^{(d+1)^2}$.
    Estimate $\hat{\theta}_t$ defined by (14) using observed contexts and rewards thus far.
    Construct $C_t(\delta)$ as defined in Eq (17).
    Play $\mathcal{Q}_t = \text{argmax}_{A_t \in \mathcal{A}, \theta \in C_t(\delta)} \mu(\mathbf{Z}_{A_t}^\top \theta)$.
    Observe reward vector $\mathbf{r}_t$.
end
```

In the above, $\mathbf{Z}_{Q_t}$ is a design matrix whose columns are the attribute vectors ($z_{t,m}$) of the pseudo-products in instance $\mathcal{M}$ defined using the products in the recommendation set Q_t of instance $\mathcal{N}$.

We want $\max_{Q \subset \mathcal{N}, Q \subset \mathcal{A}} \frac{\sum_{i,j} \eta_{ij} u_i^Q u_j^Q}{1 + \sum_i \sum_j \eta_{ij} u_i^Q u_j^Q}$.

3.3. The Maximum Likelihood Estimate

BundleMVL uses the regularized maximum likelihood estimator to compute an estimate $\hat{\theta}_t$ of θ_*. Since $\{r_{t,m}\}_{m \in Q_t^2}$ follows a multinomial distribution, the regularized log-likelihood function, till the $(t-1)-$th round, under parameter θ could be written as:

$$\mathcal{L}_t^{\lambda_t}(\theta) = \sum_{s=1}^{t-1} \sum_{m \in Q_s^2} r_{s,m} \log(\mu_m(\mathbf{Z}_{Q_s}^T \theta)) - \frac{\lambda_t}{2} \|\theta\|_2^2 \tag{9}$$

$\mathcal{L}_t^{\lambda_t}(\theta)$ is concave in θ for $\lambda_t > 0$, and the maximum likelihood estimator is given by calculating the critical point of $\mathcal{L}_t^{\lambda_t}(\theta)$. Setting $\nabla_\theta \mathcal{L}_t^{\lambda_t}(\theta) = 0$, we get $\hat{\theta}$ as the solution of:

$$\sum_{s=1}^{t-1}\sum_{m\in Q_s^2}\Big[r_{s,m}z_{s,m}\left\{\frac{e^{z_{s,m}\hat{\theta}}}{1+\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}}} : \frac{e^{z_{s,m}\hat{\theta}}}{1+\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}}}\right\}\Big] \tag{10}$$

$$- e^{z_{s,m}\hat{\theta}}\left(\frac{\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}}}{(1+\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}})^2} : \frac{e^{z_{s,m}\hat{\theta}}}{1+\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}}}\right) \tag{11}$$

$$- \lambda_t\hat{\theta} = 0 \tag{12}$$

which implies

$$\sum_{s=1}^{t-1}\sum_{m\in Q_s^2} r_{s,m}z_{s,m} - \frac{\sum\limits_{s=1}^{t-1}\sum\limits_{m\subset Q_s^2}\sum\limits_{m\in Q_s^2} z_{s,m}e^{z_{s,m}\hat{\theta}}}{(1+\sum\limits_{m\in Q_s^2} e^{z_{s,m}\hat{\theta}})} - \lambda_t\hat{\theta} = 0, \tag{13}$$

and we have

$$\sum_{s=1}^{t-1}\sum_{m\subset Q_s^2} r_{s,m}z_{s,m} - \sum_{s=1}^{t-1}\sum_{m\in Q_s^2} |Q_s| z_{s,m}\mu_m(z_{s,m}\hat{\theta}) - \lambda_t\hat{\theta} = 0 \tag{14}$$

We put $g_t(\theta) := \sum\limits_{s=1}^{t-1}\sum\limits_{m\in Q_s^2} |Q_s| z_{s,m}\mu_m(z_{s,m}\theta) + \lambda_t\theta$,
$g_t(\hat{\theta}) := \sum\limits_{s=1}^{t-1}\sum\limits_{m\subset Q_s^2} r_{s,m}z_{s,m}$.

When $\lambda_t > 0$, $\hat{\theta}_t$ is well-defined at the beginning of the interaction when no contexts have been observed, according to Eq (14). Consequently, unlike some earlier research, the regularization parameter λ_t renders UCB-MNL burn-in period free. e.g. Filippi et al. (2010).

3.4. The Confidence Set

The *in the face of uncertainty (OFU)* strategies (Auer et al., 2002b; Filippi et al., 2010; Faury et al., 2020) are followed by algorithm 1. Technical examination of *OFU* algorithms depends on two main components: the confidence set's design and how simple it is to select an action based on it.

By using the confidence set on θ_* such that $\theta_* \in C_t(\delta)$, $\forall t$ with probability at least $1-\delta$ (randomness is over customer selections), we derive $E_t(\delta)$ (described below) in Section **??**. $E_t(\delta)$ is utilized by UCB-MNL in Algorithm 1 to make decisions at each round (see Eq (8)):

$$E_t(\delta) \coloneqq \{\theta : \mathcal{L}_t^{\lambda_t}(\theta) - \mathcal{L}_t^{\lambda_t}(\hat{\theta}) \leq \beta_t^2(\delta)\}, \tag{15}$$

where $\beta_t(\delta) \coloneqq \gamma_t(\delta) + \frac{\gamma_t^2(\delta)}{\lambda_t}$, and

$$\gamma_t(\delta) \coloneqq \frac{\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\Big(\frac{(\lambda_t + LKt/d)^{d/2}\lambda_t^{-d/2}}{\delta}\Big) + \frac{2d}{\sqrt{\lambda_t}} \log(2). \tag{16}$$

A confidence set similar to $E_t(\delta)$ in Eq (15) was recently proposed in Abeille et al. (2021) and Agrawal et al. (2023) for the simpler logisitic bandit setting. Here, we expand its construction to include the BundleMVL setting. The set $E_t(\delta)$ is convex since the log-loss function is convex. This makes the decision step in Eq (8) a constraint convex optimization problem. However, it is difficult to prove bounds directly with $E_t(\delta)$. Therefore we leverage a result in Faury et al. (2020), where the authors proposed a new Bernstein-like tail inequality for self-normalized vectorial martingales (see Appendix A.1), to derive another confidence set on θ_*:

$$C_t(\delta) \coloneqq \{\theta : ||g_t(\theta) - g_t(\hat{\theta}_t)||_{\mathbf{H}_t^{-1}(\theta)} \leq \gamma_t(\delta)\}. \tag{17}$$

where

$$\mathbf{H}_t(\theta_1) \coloneqq \sum_{s=1}^{t-1} \sum_{m \in Q_s^2} \dot{\mu}_m(\mathbf{Z}_{Q_s}^T \theta_1) z_{s,m} z_{s,m}^\top + \lambda_t \mathbf{I}_d. \tag{18}$$

$\mu_m(\cdot)$ is the partial derivative of μ_m in the direction of the m_{th} component of the recommendation set and $\gamma_t(\delta)$ is defined in Eq (16). The value of $\gamma_t(\delta)$ is an outcome of the concentration result of Faury et al. (2020). As a consequence of this concentration, we have $\theta_* \in C_t(\delta)$ with probability at least $1-\delta$ (randomness is over customer choices). The Bernstein-like concentration inequality used here is similar to Theorem 1 of Abbasi-Yadkori et al. (2011) with the difference that we take into account local variance information (hence local curvature information of the reward function) in defining $\mathbf{H}_t$. The above discussion is formalized in Appendix A.1.

The set $C_t(\delta)$ is non-convex, which follows from the non-linearity of $\mathbf{H}_t^{-1}(\theta)$. We use $C_t(\delta)$ directly to prove regret guarantees. We mention how the a convex set $E_t(\delta)$ is related to $C_t(\delta)$ and share many useful properties of $C_t(\delta)$. Till then, to maintain ease of technical flow and to compare it with the previous work Faury et al. (2020), we assume that the algorithm uses $C_t(\delta)$ as the confidence set. We highlight that for the confidence sets, $C_t(\delta)$ and $E_t(\delta)$, Algorithm UCB-MNL is identical except for the calculation in Eq (8). For later sections we also define the following norm inducing design matrix based on all the contexts observed till time $t-1$:

$$\mathbf{V}_t := \sum_{s=1}^{t-1} \sum_{m \subset Q_s^2} z_{s,m} z_{s,m}^\top + \lambda_t \mathbf{I}_d. \tag{19}$$

4. Analysis

We start with standard assumptions on the transformed MNL problem instance $\mathcal{M}$, which will have straightforward implications on instance $\mathcal{N}$.

Assumption 1 (Bounded parameters). $\theta \in \mathbb{M}$, where $\mathbb{M}$ is a compact subset of $\mathbb{R}^{(d+1)^2}$. $S := \max_{\theta \in \mathbb{M}} \|\theta\|$ is known to the platform. Further, $\|z_{t,m}\|_2 \leq 1$ (or equivalently, $\|\text{vec}(x_{t,i} x_{t,j}^T)\|_2 \leq 1$ in problem instance $\mathcal{N}$) for all values of t and m.

Assumption 2. There exists $\kappa > 0$ such that for every product $m \subset Q_t$ and $|I| = 2$ and for all $Q_t \subset \mathcal{N}$ and every round t:

$$\inf_{Q_t \subset \mathcal{N}} \mu_m(\mathbf{Z}_{Q_t}^T \theta)(1 - \mu_m(\mathbf{Z}_{Q_t}^T \theta)) \geq \frac{1}{\kappa}. \tag{20}$$

In each round t, the reward of the online platform is denoted by the vector r_t. Also, the *prediction error* of θ at $\mathbf{X}_{Q_t}$, defined as:

$$\Delta^{\text{pred}}(\mathbf{Z}_{Q_t}, \theta) := |\mu(\mathbf{Z}_{Q_t}^T \theta_*) - \mu(\mathbf{Z}_{Q_t}^T \theta)|. \tag{21}$$

$\Delta^{\text{pred}}\left(\mathbf{Z}_{Q_t}, \theta\right)$ represents the difference in perceived rewards due to the inaccuracy in the estimation of the parameter θ_*.

Remark 1 (Optimistic parameter search)**.** UCB-MNL *enforces optimism via an optimistic parameter search (e.g. in Abbasi-Yadkori et al. (2011)), which is in contrast to the use of an exploration bonus as seen in Faury et al. (2020); Filippi et al. (2010). Optimistic parameter search provides a cleaner description of the learning strategy. In non-linear reward models, both approaches may not follow similar trajectory but may have overlapping analysis styles (see Filippi et al. (2010) for a short discussion).*

Recall that

$$\mu_I(Q_t, \Theta) = \frac{\exp^{x_i^T \Theta x_j}}{1 + \sum\limits_{(k,l)\in Q} \exp^{x_k^T \Theta x_l}}, \tag{22}$$

With probability at least $1-\delta$ on the randomness of customer choices

$$R_T \leq C_1 \gamma_T(\delta) \sqrt{2(d+1)^2 \log(1 + \frac{LKT}{(d+1)^2 \lambda_T})T} + C_2 \kappa \gamma_T(\delta)^2 (d+1)^2 + \log(1 + \frac{KT}{(d+1)^2 \lambda_T}), \tag{23}$$

where the constants are determined later.

Corollary 1. *Setting the regularization parameter* $\lambda_T = \mathrm{O}(d^2 \log(KT))$*, where* K *is the maximum cardinality of the assortments to be selected, makes* $\gamma_T(\delta) = \mathrm{O}(d \log^{1/2}(KT))$*. The regret upper bound is given by* $\mathbf{R}_T = \mathrm{O}(d^2 \sqrt{T} \log(KT) + \kappa d^4 \log^2(KT))$*.*

Recall the expression for cumulative regret

$$\begin{aligned} \mathbf{R}_T &= \sum_{t=1}^{T} [\mu(\mathbf{Z}_{Q_t^*}^T \theta_*) - \mu(\mathbf{Z}_{Q_t}^T \theta_*)] \\ &= \sum_{t=1}^{T} \underbrace{[\mu(\mathbf{Z}_{Q_t^*}^T \theta_*) - \mu(\mathbf{Z}_{Q_t}^T \theta_t)]}_{\text{pessimism}} + \sum_{t=1}^{T} \underbrace{[\mu(\mathbf{Z}_{Q_t}^T \theta_t) - \mu(\mathbf{Z}_{Q_t}^T \theta_*)]}_{\text{prediction error}}, \end{aligned}$$

where *pessimism* is the additive inverse of the optimism (difference between the payoffs under true parameters and those estimated by UCB-MNL). Due to optimistic decision-making and the fact that $\theta_* \in C_t(\delta)$ (see Eq (8)), *pessimism* is non-positive, for all rounds. Thus, the regret is upper bounded by the sum of the prediction error for T rounds. We derive an the expression for prediction error upper bound for a single round t similar to Section 4.1 in Agrawal et al. (2023). For the regret calculation, we refer to Appendix A.4 and Appendix A.5 in Agrawal et al. (2023) with suitable modifications. The final step can be done by following Section 4.3 and Appendix A.6 in Agrawal et al. (2023).

5. Conclusion

This study has introduced a novel approach to leveraging contextual bandit models, specifically targeting the enhancement of recommendation systems for maximizing purchase probability within a multi-choice behavior scenario. Customers with time varying preferences are shown recommendation sets over a horizon of length T and they choose at most two products from these sets according to a generalization of the MNL customer choice model. We've employed a rigorous mathematical framework, supported by assumptions that allow reuse of the results for the contextual mutlinomial logit (MNL) problem, leading to the development of our algorithmic approach that helps the platform achieve provably sublinear regret. We support our recommendation process with numerical results that indicate that the way the problem parameters influence regret upper bound is reasonable.

References

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, *24*, 2312–2320.

Abeille, M., Faury, L., & Calauzènes, C. (2021). Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics* (pp. 3691–3699). PMLR.

Agrawal, P., Tulabandhula, T., & Avadhanula, V. (2023). A tractable online learning algorithm for the multinomial logit contextual bandit. *European Journal of Operational Research*, *310*, 737–750. URL: https://www.sciencedirect.com/science/article/pii/S0377221723001832. doi:https://doi.org/10.1016/j.ejor.2023.02.036.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, *47*, 235–256. URL: https://doi.org/10.1023/A:1013689704352. doi:10.1023/A:1013689704352.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002b). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*, 235–256.

Faury, L., Abeille, M., Calauzènes, C., & Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning* (pp. 3052–3060). PMLR.

Filippi, S., Cappe, O., Garivier, A., & Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems* (pp. 586–594).

Kök, A. G., Fisher, M. L., & Vaidyanathan, R. (2015). Assortment planning: Review of literature and industry practice. *Retail supply chain management: Quantitative models and empirical studies*, (pp. 175–236).

Tulabandhula, T., Sinha, D., Karra, S. R., & Patidar, P. (2023). Multi-purchase behavior: Modeling, estimation, and optimization. *Manufacturing & Service Operations Management*, *25*, 2298–2313.