Multinomial Logit Bandit with Linear Utility Functions

Mingdong Ou, Nan Li, Shenghuo Zhu, Rong Jin,

Alibaba Group, Hang Zhou, China

{mingdong.omd, nanli.ln, shenghuo.zhu, jinrong.jr}@alibaba-inc.com,

Abstract

Multinomial logit bandit is a sequential subset selection problem which arises in many applications. In each round, the player selects a K-cardinality subset from N candidate items, and receives a reward which is governed by a *multinomial logit* (MNL) choice model considering both item utility and substitution property among items. The player's objective is to dynamically learn the parameters of MNL model and maximize cumulative reward over a finite horizon T. This problem faces the exploration-exploitation dilemma, and the involved combinatorial nature makes it non-trivial. In recent years, there have developed some algorithms by exploiting specific characteristics of the MNL model, but all of them estimate the parameters of MNL model separately and incur a regret no better than $\tilde{O}(\sqrt{NT})$ which is not preferred for large candidate set size N. In this paper, we consider the *linear utility* MNL choice model whose item utilities are represented as linear functions of d-dimension item features, and propose an algorithm, titled **LUMB**, to exploit the underlying structure. It is proven that the proposed algorithm achieves $\tilde{O}(dK\sqrt{T})$ regret which is free of candidate set size. Experiments show the superiority of the proposed algorithm.

1 Introduction

In traditional stochastic multi-armed bandit (MAB) [Bubeck and Cesa-Bianchi, 2012], the player selects *one* from N items and receives a reward corresponding to that item in each round. The objective is to maximize cumulative reward over a finite horizon of length T, or alternatively, minimize the regret relative to an oracle. Typically, algorithms are designed based on appropriate exploration-exploitation tradeoff which allows the player to identify the best item through exploration whilst not spending too much on sub-optimal ones, and the family of upper confidence bound (UCB) algorithms [Auer, 2002; Chu *et al.*, 2011] and Thompson sampling (TS) [Thompson, 1933; Agrawal and Goyal, 2012] are representative examples. This paper studies a combinato-

rial variant of MAB, where in each round, the player offers a subset of cardinality K to a user, and receives the reward associated with one of the items in the selected subset[1]. The player faces the problem of determining which subset of items to present to users who arrive sequentially, whilst not knowing user preference. Similar to MAB, we need to solve the exploration-exploitation tradeoff in this problem. However, a naive translation of this problem to MAB is prohibitive, since the number of possible K-cardinality subsets is exponentially large and cannot be efficiently explored within a reasonable sized time horizon. To tackle this issue, different strategies have been proposed in the literature, *e.g.*, [Kveton *et al.*, 2015; Lagrée *et al.*, 2016; Agrawal *et al.*, 2017a].

In recent literature, the *multinomial logit* (MNL) choice model [Luce, 2005; Plackett, 1975] which is widely used in economics is utilized to model user choice behavior by specifying the probability that a user selects an item given the offered set, and above exploration-exploitation problem is referred as the MNL-Bandit problem [Rusmevichientong *et al.*, 2010; Agrawal *et al.*, 2017a; 2017b; Cheung and Simchi-Levi, 2017]. Unlike other combinatorial bandit problems, MNL-Bandit problem considers the substitution property among items and leads to non-monotonic reward function. By exploiting specific characteristics of the MNL model, UCB-style [Agrawal *et al.*, 2017a] and TS-style [Agrawal *et al.*, 2017b] algorithms have been developed to dynamically learn the parameters of the MNL model which are a priori unknown, achieving a regret of $\tilde{O}(\sqrt{NT})$ under a mild assumption. Also, a lower regret bound is established in [Agrawal *et al.*, 2017a], by showing that any algorithm based on the MNL choice model must incur a regret of $\Omega(\sqrt{NT/K})$. It is easy to find that the regret depends on the number of candidate items N, making them less preferred in many large scale applications such as online advertising.

In this paper, we use the *linear utility* MNL choice model (formulated in Section 3) to model user choice behavior given a set of items, rather than traditional MNL model. Specifically, it is assumed that each item in candidate set is described by a d-dimension feature vector, and item utilities

---

[1] This problem is known as dynamic assortment selection in the literature [Caro and Gallien, 2007; Rusmevichientong *et al.*, 2010], where the selected subset of items forms an assortment.

of the MNL model can be formulated as linear functions of item features. Based on this, the problem of estimating item utilities (*i.e.*, parameters of the MNL model) is changed to estimating underlying model parameters of linear functions. Since the number of parameters is irrelevant with the number of items, it is possible to achieve more efficient solution when the number of items is large. By taking the UCB approach, we propose an algorithm, titled LUMB (which is short for Linear Utility MNL-Bandit), to dynamically learn the parameters and narrow the regret. The main contributions of this work include:

- To the best of our knowledge, this is the first to use linear utility MNL choice model in sequential subset selection problem. Also, an UCB-style algorithm LUMB is proposed to learn the model parameters dynamically.

- An upper regret bound $O\big(dK\sqrt{T}\,(\log T)^2\big)$ is established for the proposed LUMB algorithm. This regret bound is free of candidate item set size, which means that LUMB can be applied to large item set.

- Empirical studies demonstrate the superiority of the proposed LUMB algorithm over existing algorithms.

The rest of this paper is organized as follows. Section 2 briefly introduces related work. Section 3 and 4 present problem formulation and the LUMB algorithm. Section 5 establishes regret analysis. Section 6 summarizes the experiments, and Section 7 concludes this work with future directions.

2 Related Work

Classical bandit algorithms aim to find the best arm with exploration-exploitation strategy. Auer [2002] first proposes a UCB approach in linear payoff setting. Dani *et al.* [2008] and Abbasi-Yadkori *et al.* [2011] propose improved algorithms which bound the linear parameters directly. Agrawal and Goyal [2013] propose a Thompson sampling approach. However, because the reward of a subset is not a linear function of item features in the subset, these works cannot be directly applied to our problem.

Another class of bandit works related to our work is combinatorial bandit where the player selects a subset of arms and receive a collective reward in each round. Researchers study the problem mainly on two settings, stochastic setting [Gai *et al.*, 2012; Russo and Van Roy, 2014; Kveton *et al.*, 2015] and adversarial setting [Cesa-Bianchi and Lugosi, 2012; Audibert *et al.*, 2013]. Gai *et al.* [2012] first learn the problem in linear reward setting and Kveton *et al.* [2015] prove a tight regret bound. It is generalized to non-linear rewards in [Chen *et al.*, 2016; 2013]. Wen *et al.* [2015] and Wang *et al.* [2017] propose contextual algorithms which can handle large item sets. However, these works imply that the reward is monotonic which is not satisfied in MNL-Bandit (Section 3). In practice, as clarified in [Cheung and Simchi-Levi, 2017], the low-reward item may divert the attention of user and lead to lower subset reward.

Rusmevichientong *et al.* [2010] solve the MNL-Bandit problem and achieve instance-dependent upper regret bound $O\left(logT\right)$, and Sauré and Zeevi [2013] extend to a wider class of choice models. Agrawal *et al.* [2017a] propose a

UCB approach and achieve instance-independent upper regret bound $O\left(\sqrt{NT}\right)$. Agrawal *et al.* [2017a] propose a Thompson sampling approach with better empirical performance. Recently, some works begin to study variants of classic MNL-Bandit problem. Some works learn the personalized MNL-Bandit problem [Kallus and Udell, 2016; Bernstein *et al.*, 2017; Golrezaei *et al.*, 2014]. Cheung and Simchi-Levi [2017] learn the problem with resource constraints. However, as clarified in Section 1, these works model item utility separately which is not feasible for large item sets.

3 Problem Formulation

Suppose there is a candidate item set, $\mathcal{S} = \{1, \cdots, N\}$, to offer to users. Each item i corresponds to a reward $r_i \geq 0$ and a feature vector $\mathbf{x}_i \in \mathbf{R}^d$. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be the feature matrix. In MNL-Bandit, at each time step t, the player selects a subset $\overline{S}_t \in \mathcal{C}\left(\mathcal{S}\right) = \{S | S \subseteq \mathcal{S}, |S| \leq K\}$ and observes the user choice $c_t \in \overline{S}_t \cup \{0\}$, where $c_t = 0$ represents that the user chooses nothing from $\overline{S}_t$. The objective is to design a bandit policy to approximately maximize the expected cumulative reward of chosen items, *i.e.*, $\mathbb{E}\left(\sum_t r_{c_t}\right)$.

According to above setting, in each time step t, a bandit policy is only allowed to exploit the item features $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, historical selected subsets, $\{\overline{S}_\tau | \tau < t\}$ and the user choice feedbacks, $\{c_\tau | \tau < t\}$.

The user choice follows the MNL model [Luce, 2005; Plackett, 1975]. MNL assumes substitution property among items in a selected subset. Specifically, for item i, the larger the utilities of the other items, the smaller the chosen probability of item i is. The choice probability follows a multinomial distribution,

$$p_i\left(\overline{S}_t, \mathbf{v}\right) = \begin{cases} \dfrac{v_i}{1 + \sum_{j \in \overline{S}_t} v_j}, & i \in \overline{S}_t \\ \dfrac{1}{1 + \sum_{j \in \overline{S}_t} v_j}, & i = 0 \\ 0, & \text{otherwise}, \end{cases} \quad (1)$$

where $v_i \geq 0$ is item utility which is a priori unknown to the player. The choice probability of item i in selected subset, $\overline{S}_t$, is linear with its utility, v_i. Besides, it is possible that nothing is chosen which is realized by adding a virtual item 0 with utility 1. With the MNL model, we have the expected reward under given utility vector, $\mathbf{v} = [v_1, \cdots, v_N]$,

$$R\left(\overline{S}_t, \mathbf{v}\right) = \sum_{i \in \overline{S}_t} p_i\left(\overline{S}_t, \mathbf{v}\right) r_i = \frac{\sum_{i \in \overline{S}_t} v_i r_i}{1 + \sum_{i \in \overline{S}_t} v_i}. \quad (2)$$

Note that the expected reward is non-monotonic, that is, both the addition of low-reward item to selected subset and increment on utility of low-reward item may lead to lower reward. The expected cumulative reward is

$$\mathbb{E}\left(\sum_t r_{c_t}\right) = \sum_t \mathbb{E}\left(r_{c_t}\right) = \sum_t \mathbb{E}\left(R\left(\overline{S}_t, \mathbf{v}\right)\right). \quad (3)$$

Since direct analysis of (3) is not tractable when $\mathbf{v}$ is unknown, we analyze the regret instead,

$$Reg\left(T, \mathbf{v}\right) = \sum_{t=1}^{T} \left(R\left(S^*, \mathbf{v}\right) - \mathbb{E}\left(R\left(\overline{S}_t, \mathbf{v}\right)\right)\right), \quad (4)$$

where T is the length of time horizon and S^* is the optimal subset,

$$S^* = \arg\max_{S \in \mathcal{C}(\mathcal{S})} R(S, \mathbf{v}) .$$

Naturally, the objective is to approximately minimize the expected cumulative regret, $Reg(T, \mathbf{v})$, with appropriate bandit policy. Especially, after enough time steps, an appropriate solution should almost achieve the subsets with highest reward, which implies that the cumulative regret, $Reg(T, \mathbf{v})$, should be sub-linear with T. As each item corresponds to an utility which needs to be estimated separately, this makes the lower cumulative regret bound relevant with item number and will be not feasible for large item sets.

Therefore, linear item utility is introduced where item utility is a linear function of item feature,

$$v_i = \boldsymbol{\theta}^{*\top} \mathbf{x}_i , \tag{5}$$

where $\boldsymbol{\theta}^*$ is a linear parameter vector unknown to the player. Thus, estimating item utilities will be changed to estimating utility function parameters which can exploit the correlation between items on features, then it is potential to achieve regret bound free of item number, N.

4 Algorithm

We propose an algorithm, called *Linear Utility MNL-Bandit* (LUMB), which proposes a UCB approach to sequentially estimate linear utility function and approach highest reward. LUMB first estimates the linear parameters of utility function, then constructs the UCB of item utility and subset reward, finally offers the subset with highest reward UCB. Algorithm 1 clarifies the detail of LUMB.

4.1 Estimation of Linear Utility Function

As the choice probability of an item is non-linear with the parameters of utility function, it is difficult to estimate the parameters directly with user choice feedback. Instead, we split the time horizon into epochs like [Agrawal *et al.*, 2017a]. Let L be the number of epochs. In each epoch l, the selected subset, $S_l \in \mathcal{C}(\mathcal{S})$, is offered repeatedly until that the user chooses nothing from the offered subset. Then, we can obtain the chosen times of each item $i \in S_l$,

$$\hat{v}_{i,l} = \sum_{t \in \mathcal{E}_l} \mathbb{I}(c_t = i) , \tag{6}$$

$$\text{s.t.} \quad \mathbb{I}(c_t = i) = \begin{cases} 1, & c_t = i \\ 0, & c_t \neq i , \end{cases}$$

where $\mathcal{E}_l$ is the set of time steps in epoch l, c_t is the chosen item in time step t.

It can be proven that $\mathbb{E}(\hat{v}_{i,l}) = v_i$ (Lemma 1), which means the empirical average of $\hat{v}_{i,l}$ is almost equal to real item utility and irrelevant to other items in the subset. Thus, the estimation of utility function can be simply formulated as a linear regression which directly approaches empirical samples of $\hat{v}_{i,l}$. Specifically,

$$\boldsymbol{\theta_l} = \arg\min_{\boldsymbol{\theta}} \sum_{\tau \leq l} \sum_{i \in S_\tau} \|\boldsymbol{\theta}^\top \mathbf{x}_i - \hat{v}_{i,l}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 ,$$

Algorithm 1 Linear Utility MNL-Bandit

1: **Inputs:**
  $\mathcal{S} = \{1, \cdots, N\}, \mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N],$
  $\mathbf{r} = [r_1, \cdots, r_N]$
2: **Initialize:**
  $\theta_0 \leftarrow [0]^d, \mathbf{A_0} \leftarrow \mathbf{I}_d, \mathbf{b_0} \leftarrow [0]^d,$
  $v_{i,0}^{\text{UCB}} \leftarrow \dfrac{\sqrt{2} + \alpha}{\lambda} \|\mathbf{x}_i\|, \forall i \leq N$
  $t \leftarrow 1, l \leftarrow 1, \mathcal{E}_1 \leftarrow \emptyset, c_0 \leftarrow 0$
3: **repeat**
4:   **if** $c_{t-1} = 0$ **then**
5:     Compute $S_l \leftarrow \arg\max\limits_{S \in \mathcal{S}} R(S, \mathbf{v}_{l-1}^{\text{UCB}})$
6:   **end if**
7:   Offer subset S_l, observe the user choice c_t.
8:   **if** $c_t = 0$ **then**
9:     **for** $i \in S_l$ **do**
10:       compute $\hat{v}_{i,l} \leftarrow \sum_{t \in \mathcal{E}_l} \mathbb{I}(c_t = i)$
11:     **end for**
12:     update $\mathbf{b}_l \leftarrow \mathbf{b}_{l-1} + \sum_{i \in S_l} \hat{v}_{i,l} \mathbf{x}_i$
13:     update $\mathbf{A}_l \leftarrow \mathbf{A}_{l-1} + \sum_{i \in S_l} \mathbf{x}_i \mathbf{x}_i^\top$
14:     update $\theta_l \leftarrow \mathbf{A}_l^{-1} \mathbf{b}_l$
15:     **for** $i \in \mathcal{S}$ **do**
16:       $v_{i,l}^{\text{UCB}} \leftarrow \theta_l^\top \mathbf{x}_i + (\sqrt{2} + \alpha) \sqrt{\mathbf{x}_i^\top \mathbf{A}_l^{-1} \mathbf{x}_i}$
17:     **end for**
18:     $l \leftarrow l + 1$
19:     $\mathcal{E}_l \leftarrow \emptyset$
20:   **else**
21:     $\mathcal{E}_l \leftarrow \mathcal{E}_l \cup t$
22:   **end if**
23:   $t \leftarrow t + 1$
24: **until** $t < T$

---

where λ is a constant regularization coefficient. Then, we can obtain close-form solution

$$\mathbf{A}_l = \lambda \mathbf{I}_d + \sum_{\tau \leq l} \sum_{i \in S_\tau} \mathbf{x}_i \mathbf{x}_i^\top , \tag{7}$$

$$\mathbf{b}_l = \sum_{\tau \leq l} \sum_{i \in S_\tau} \hat{v}_{i,\tau} \mathbf{x}_i , \tag{8}$$

$$\theta_l = \mathbf{A}_l^{-1} \mathbf{b}_l , \tag{9}$$

where $\mathbf{I}_d$ is a d-by-d identity matrix.

4.2 Construction of Upper Confidence Bound

We construct the UCB of item utility, which is proven in Lemma 2, as

$$v_{i,l}^{\text{UCB}} = \theta_l^\top \mathbf{x}_i + (\sqrt{2} + \alpha) \sigma_{i,l} , \tag{10}$$

$$\text{s.t.} \quad \sigma_{i,l} = \sqrt{\mathbf{x}_i^\top \mathbf{A}_l^{-1} \mathbf{x}_i}$$

where α is constant. Let $\mathbf{v}_l^{\text{UCB}} = [v_{1,l}^{\text{UCB}}, \cdots, v_{N,l}^{\text{UCB}}]$. Then, the UCB of the highest reward, $R(S^*, \mathbf{v})$, is constructed as the highest reward with $\mathbf{v}_l^{\text{UCB}}$ (Lemma 5). The corresponding subset is

$$S_{l+1} = \arg\max_{S \in \mathcal{C}(\mathcal{S})} R(S, \mathbf{v}_l^{\text{UCB}}) , \tag{11}$$

and we offer the subset S_{l+1} in epoch $l+1$.

It is hard to get S_{l+1} by directly solving the above optimization problem. According to [Davis *et al.*, 2013], the above optimization problem can be translated to a linear program problem,

$$\max \quad \sum_{i=1}^{N} r_i w_i , \qquad (12)$$

$$\text{s.t.} \quad w_0 + \sum_{i=1}^{N} w_i = 1, \quad \sum_{i=1}^{N} \frac{w_i}{v_i^{\text{UCB}}} \le K w_0 ,$$

$$\forall i \in \mathcal{S}, 0 \le \frac{w_i}{v_i^{\text{UCB}}} \le w_0 .$$

Then, $S_{l+1} = \{i | w_i > 0, i > 0\}$.

5 Regret Analysis

In this section, we analyze the upper regret bound of Algorithm 1 to theoretically identify the convergence performance. Without loss of generality, we first declare the assumption in the following analysis.

Assumption 1. $\forall i \in \mathcal{S}, r_i \le 1, \|\mathbf{x}_i\| \le 1, \|\boldsymbol{\theta}^*\| \le 1$.

According to the assumption, we have that $\forall i \in \mathcal{S}, v_i \le 1$. Moreover, we let $\lambda = 1$. Then, we give the upper bound of regret in Theorem 1 in advance which is proven in Section 5.3. We can achieve result similar to Theorem 1 when the parameters in Assumption 1 are bounded by finite constants.

Theorem 1. *Following the process in Algorithm 1, let*

$$\beta = 2 \log_2 T,$$

$$\alpha = \beta \sqrt{2 \log \left(2\sqrt{T} \left(1 + \frac{T}{d} \right)^{d/2} \right)},$$

then the upper bound of $Reg(T, \mathbf{v})$ is

$$O \left(dK\sqrt{T} (\log T)^2 \right) . \qquad (13)$$

The proof is separated into three steps. We first prove the correctness of the constructed UCB of utility and the cumulative deviation between UCB of utility and real utility which is sublinear with respect to T. Then we prove that the deviation between UCB of reward and real reward can be bounded by deviation of utility, finally the upper bound of cumulative regret can be proved by combining the above two results.

5.1 Analysis of Utility

We first prove the distribution of $\hat{v}_{i,l}$.

Lemma 1. *With the definition in Eq. (6), $\forall l \le L, i \in S_l$, $\hat{v}_{i,l}$ follows geometric distribution, that is*

$$\mathbb{P}(\hat{v}_{i,l} = \beta) = \frac{1}{1+v_i} \left(\frac{v_i}{1+v_i} \right)^{\beta}, \forall \beta \ge 0 , \qquad (14)$$

$$\mathbb{E}(\hat{v}_{i,l}) = v_i . \qquad (15)$$

Because of the space limitation, proof will be attached in a longer version. According to above Lemma, the deviation between $\hat{v}_{i,l}$ and real utility is unbounded. This makes the prove of utility UCB difficult. Fortunately, the probability $\mathbb{P}(\hat{v}_{i,l} > \beta)$ decays exponentially when β increases. We bound $\hat{v}_{i,l}$ in a relative small interval with high probability. Then, we can prove the utility UCB as below.

Lemma 2. *With definition of $v_{i,l}^{\text{UCB}}$ in Eq. (10) and definition of $\hat{v}_{i,l}$ in Eq. (6), if $\beta \ge 2 \log_2 T, \forall \tau \le l, j \in S_l, \hat{v}_{j,\tau} \le \beta$, then $\forall i \in \mathcal{S}, l \le L$,*

$$0 \le v_{i,l}^{\text{UCB}} - v_i \le 2 \left(\sqrt{2} + \alpha \right) \sigma_{i,l} , \qquad (16)$$

with probability at least

$$1 - \left(1 + \frac{T}{d} \right)^{d/2} \exp \left(-\frac{\alpha^2}{2\beta^2} \right) .$$

Proof. Let $\Delta_{i,l} = |\boldsymbol{\theta}_l^{\top} \mathbf{x}_i - v_i|$, we just need to prove

$$\Delta_{i,l} \le \left(\sqrt{2} + \alpha \right) \sigma_{i,l} .$$

According to Lemma 1, when $\hat{v}_{i,l} \le \beta$,

$$\mathbb{E}(\hat{v}_{i,l} - v_i) = -(1+\beta) \frac{p_i^{\beta+1}}{1 - p_i^{\beta+1}} ,$$

$$\text{s.t.} \quad p_i = \frac{v_i}{1+v_i} .$$

Note that the result of $\mathbb{E}(\hat{v}_{i,l} - v_i)$ is irrelevant with l. Let $\epsilon_i = \mathbb{E}(\hat{v}_{i,l} - v_i)$.

$$\Delta_{i,l} \le |\mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \sum_{\tau \le l} \sum_{j \in S_\tau} \mathbf{x}_j (\hat{v}_{j,\tau} - v_j - \epsilon_j)|$$

$$+ |\mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \sum_{\tau \le l} \sum_{j \in S_\tau} \mathbf{x}_j \epsilon_j - \mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \boldsymbol{\theta}^*| .$$

We prove the bound of two parts respectively. Let $u_i = \hat{v}_{i,\tau} - v_i - \epsilon_i, \mathbf{s}_l = \sum_{\tau \le l} \sum_{i \in S_\tau} \mathbf{x}_i u_i$, it is easy to prove that

$$\mathbb{E}(\exp(\gamma u_i)) \le \exp(\frac{\gamma^2 \beta^2}{2}) ,$$

then, with Lemma 9 in [Abbasi-Yadkori *et al.*, 2011], we can prove that with probability

$$1 - \left(1 + \frac{T}{d} \right)^{d/2} \exp \left(-\frac{\alpha^2}{2\beta^2} \right),$$

we have that

$$|\mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \sum_{\tau \le l} \sum_{i \in S_\tau} \mathbf{x}_i (\hat{v}_{i,\tau} - v_i - \epsilon_i)| \le \alpha \sigma_{i,l} . \qquad (17)$$

Moreover, with Cauchy–Schwarz inequality, the other part is bounded as

$$|\mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \sum_{\tau \le l} \sum_{j \in S_\tau} \mathbf{x}_j \epsilon_j - \mathbf{x}_i^{\top} \mathbf{A}_l^{-1} \boldsymbol{\theta}^*| \le \sqrt{2} \sigma_{i,l} . \qquad (18)$$

The lemma can be proven by combining Eq. (17) and Eq. (18). $\qquad \square$

Moreover, we prove the bound of cumulative utility deviation.

Lemma 3. *Following the process in Algorithm 1, the cumulative deviation between utility UCB and real utility can be bounded as*

$$\sum_{l=1}^{L} \sum_{i \in S_l} \sigma_{i,l} \le K \sqrt{70 dL \log L} . \qquad (19)$$

Proof is similar to the proof of Lemma 3 in [Chu *et al.*, 2011]. Because of the space limitation, the proof will be attached in a longer version. Lemma 3 shows that the bound of cumulative deviation is sub-linear with epoch number, and the average deviation in each epoch will vanish after enough epochs.

5.2 Analysis of Reward

We first estimate the deviation between estimated reward and real reward of S_l with the result of Lemma 3.

Lemma 4. *In each epoch l of Algorithm 1, if $\forall i \in S_l$,*

$$0 \le v_i^{\mathrm{UCB}} - v_i \le 2\left(\sqrt{2} + \alpha\right)\sigma_{i,l},$$

then the cumulative deviation between estimated reward and real reward of S_l is

$$\mathbb{E}\left(\sum_{t \in \mathcal{E}_l}\left(R\left(S_l, \mathbf{v}_l^{\mathrm{UCB}}\right) - r_{c_t}\right)\right) \le 2\left(\sqrt{2} + \alpha\right)\sum_{i \in S_l} r_i \sigma_{i,l}. \tag{20}$$

Because of the space limitation, proof will be attached in a longer version. This Lemma means that the deviation of subset reward is bounded by deviation of item utilities in the subset.

Lemma 5. *(Lemma 4.2 in [Agrawal* et al.*, 2017a])With the reward defined in Eq. (2), suppose there are two subsets, $\tilde{S}^{\mathrm{UCB}}$ and $\tilde{S}$, that*

$$\tilde{S}^{\mathrm{UCB}} = \arg\max_{S \in \mathcal{C}(\mathcal{S})} R\left(S, \mathbf{v}^{\mathrm{UCB}}\right),$$

$$\tilde{S} = \arg\max_{S \in \mathcal{C}(\mathcal{S})} R\left(S, \mathbf{v}\right).$$

If $\forall i \in \tilde{S}, v_i^{\mathrm{UCB}} \ge v_i$, then the rewards satisfy the inequality

$$R\left(\tilde{S}, \mathbf{v}\right) \le R\left(\tilde{S}, \mathbf{v}^{\mathrm{UCB}}\right) \le R\left(\tilde{S}^{\mathrm{UCB}}, \mathbf{v}^{\mathrm{UCB}}\right). \tag{21}$$

Lemma 5 shows that the estimated reward of subset S_l is an upper bound of real highest reward. Then we can easily bound the regret in each epoch with Lemma 4 and Lemma 5.

Lemma 6. *(Regret bound in a single epoch) In each epoch l of Algorithm 1, if $\forall i \in S_l \cup S^*$*

$$0 \le v_i^{\mathrm{UCB}} - v_i \le 2\left(\sqrt{2} + \alpha\right)\sigma_{i,l}, \tag{22}$$

then the regret of epoch l is

$$\mathbb{E}\left(\sum_{t \in \mathcal{E}_l}\left(R\left(S^*, \mathbf{v}\right) - r_{c_t}\right)\right) \le 2\left(\sqrt{2} + \alpha\right)\sum_{i \in S_l} r_i \sigma_{i,l}. \tag{23}$$

5.3 Upper Bound of Regret

We first prove a more general version of Theorem 1 with parameters, α and β.

Lemma 7. *Following the process in Algorithm 1, if $\beta \ge 2\log_2 T$, the cumulative regret, defined in Eq. (4), can be bounded,*

$$Reg\left(T, \mathbf{v}\right) \le 2TK\left(1 + \frac{T}{d}\right)^{d/2}\exp\left(-\frac{\alpha^2}{2\beta^2}\right) + \frac{T^2 K}{2^{\beta+2}}$$
$$+ 2\left(\sqrt{2} + \alpha\right)K\sqrt{70dT\log T}. \tag{24}$$

Proof. We obtain the bound of regret respectively in two situations: the item utility inequality in Lemma 2 is (or not) satisfied. We model the event that the item utility inequality in Lemma 2 is not satisfied as

$$\mathcal{A}_l = \mathcal{U}_l \cup \mathcal{B}_l, \tag{25}$$

$$\text{s.t.} \quad \mathcal{U}_l = \{v_{i,l}^{\mathrm{UCB}} > v_i + 2\left(\sqrt{2} + \alpha\right)\sigma_{i,l}$$
$$\text{or} \quad v_{i,l}^{\mathrm{UCB}} < v_i, \exists i \in S_l \cup S^*\},$$
$$\mathcal{B}_l = \{\hat{v}_{i,\tau} > \beta, \exists \tau \le l, i \in S_\tau\}.$$

Then, it is easy to bound the probability of $\mathcal{A}_l$,

$$\mathbb{P}\left(\mathcal{A}_l\right) \le 2K\left(1 + \frac{T}{d}\right)^{d/2}\exp\left(-\frac{\alpha^2}{2\beta^2}\right) + \frac{lK}{2^\beta}. \tag{26}$$

Let $\tilde{\mathcal{A}}_l$ be the complement of $\mathcal{A}_l$. Then, the regret can be splited into two parts, that is,

$$Reg\left(T, \mathbf{v}\right) = \mathbb{E}\left(\sum_{l=1}^{L}\sum_{t \in \mathcal{E}_l}\mathbb{I}\left(\mathcal{A}_{l-1}\right)\left(R\left(S^*, \mathbf{v}\right) - r_{c_t}\right)\right)$$
$$+ \mathbb{E}\left(\sum_{l=1}^{L}\sum_{t \in \mathcal{E}_l}\mathbb{I}\left(\tilde{\mathcal{A}}_{l-1}\right)\left(R\left(S^*, \mathbf{v}\right) - r_{c_t}\right)\right),$$

where $\mathbb{I}\left(\mathcal{A}\right)$ is an indicator random variable whose value is 1 when $\mathcal{A}$ happens, othewise 0. We first consider the situation that $\mathcal{A}_l$ happens,

$$\mathbb{E}\left(\sum_{l=1}^{L}\sum_{t \in \mathcal{E}_l}\mathbb{I}\left(\mathcal{A}_l\right)\left(R\left(S^*, \mathbf{v}\right) - r_{c_t}\right)\right)$$
$$\le 2TK\left(1 + \frac{T}{d}\right)^{d/2}\exp\left(-\frac{\alpha^2}{2\beta^2}\right) + \frac{T^2 K}{2^{\beta+2}}. \tag{27}$$

Then, we consider that $\mathcal{A}_l$ does not happen. According to Lemma 6 and Lemma 3,

$$\mathbb{E}\left(\sum_{l=1}^{L}\sum_{t \in \mathcal{E}_l}\mathbb{I}\left(\tilde{\mathcal{A}}_{l-1}\right)\left(R\left(S^*, \mathbf{v}\right) - r_{c_t}\right)\right)$$
$$\le 2\left(\sqrt{2} + \alpha\right)K\sqrt{70dT\log T}. \tag{28}$$

Finally, we can finish the proof by adding Eq. (27) and Eq. (28).

$\square$

With Lemma 7, Theorem 1 can be proven by setting

$$\beta = 2\log_2 T,$$

$$\alpha = \beta\sqrt{2\log\left(2\sqrt{T}\left(1 + \frac{T}{d}\right)^{d/2}\right)}.$$

As our method is in the similar framework of MNL-Bandit [Agrawal *et al.*, 2017a] whose lower bound is $\tilde{O}\left(\sqrt{T}\right)$, our regret bound matches the lower bound up to logarithmic terms with respect to T.

6 Experiments

In this section, we evaluate LUMB on synthetic data and compare it to three existing alternative algorithms. We demonstrate the superiority of LUMB on cumulative regret. Moreover, we show that the estimated linear parameters of utility function and utilities will asymptotically converge to the real value.

6.1 Setting

The synthetic data is generated randomly. N rewards are sampled from interval $(0, 1]$ uniformly. d-dimension parameter vector of utility function, $\boldsymbol{\theta}^*$, is sampled from $[0, 1]^d$ uniformly, then is normalized to 1. N d-dimension feature vectors are sampled from $[0, 1]^d$ uniformly. To follow the experiment setting in [Agrawal *et al.*, 2017b], feature vectors are normalized so that item utilities distribute uniformly on $[0, 1]$. Experiments are all performed on ten randomly generated data sets and the results show below are all average of results on these data sets.

Three alternative algorithms are compared:

- UCB-MNL [Agrawal *et al.*, 2017a]: This algorithm proposes a UCB approach with MNL choice model.

- Thompson-Beta [Agrawal *et al.*, 2017b]: This algorithm proposes a Thompson sampling approach with MNL choice model.

- Thompson-Corr [Agrawal *et al.*, 2017b]: This algorithm is a variant of Thompson-Beta which samples item utilities with correlated sampling.

6.2 Results

We conduct empirical experiments on synthetic data sets with $N = 1000, d = 10$. Subset size K is set to 10. Figure 1 shows the cumulative regret on the synthetic data sets, which is normalized by best reward, *i.e.*, $Reg\,(t, \mathbf{v})\,/R\,(S^*, \mathbf{v})$. Note that the axis of cumulative regret is plot in a logarithm style for the convenience of observing the trend of LUMB regret on time horizon. The cumulative regrets increase slower when time step increases. Besides, we can see that the cumulative regret of LUMB is much smaller than the alternative algorithms through the time horizon.

We evaluate the convergence of utility vector on synthetic data sets with $N = 1000, d = 10$. Figure 2 shows the deviation between estimated mean utility and real utility, which is normalized by the norm of real utility, *i.e.*, $\|\boldsymbol{\theta}_t^\top \mathbf{X} - \mathbf{v}\|/\|\mathbf{v}\|$. The deviation of LUMB decreases fast in the early stage and achieve smaller deviation compared to the alternative algorithms.

Moreover, we evaluate the deviation of estimated linear parameter vector in Figure 2. The deviation is normalized by the norm of real parameters, *i.e.*, $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|/\|\boldsymbol{\theta}^*\|$. Note that the deviation also decreases fast in the early stage and asymptotically converges to zero finally. This demonstrates that LUMB can correctly estimate the linear parameters.

7 Conclusion

We study the sequential subset selection problem with linear utility MNL choice model, and propose a UCB-style algorithm, LUMB. Also, an upper regret bound,
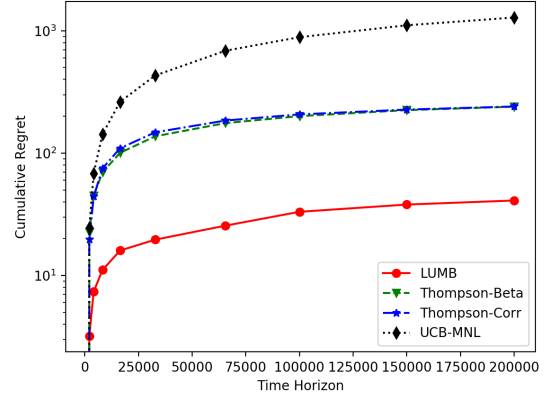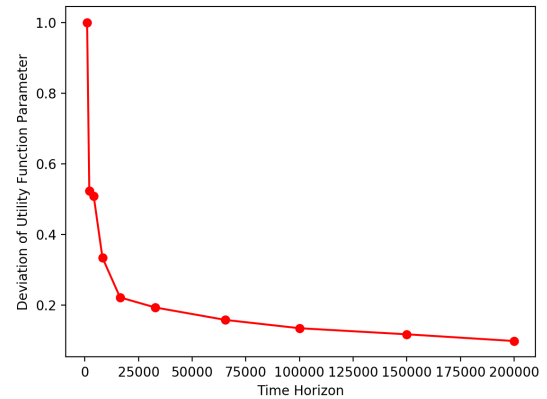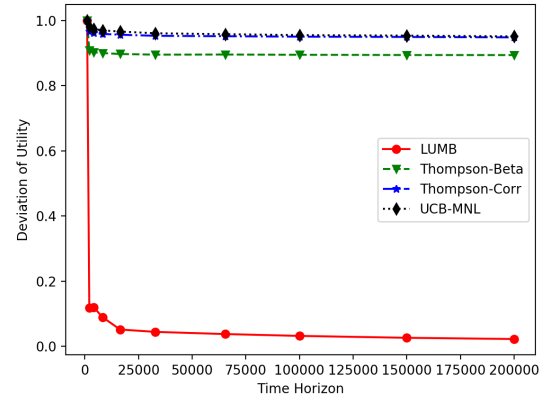

Figure 1: Cumulative regret along time horizon.

Figure 2: Deviation of item utility vector (upper figure) and linear utility function parameter vector (lower figure) on different time steps.

$O\left(dK\sqrt{T}\,(\log T)^2\right)$, is established, which is free of candidate item number. Experiments show the performance of LUMB. In the future work, we are interested in extending the idea to other choice models such as nested MNL.

References

[Abbasi-Yadkori *et al.*, 2011] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

[Agrawal and Goyal, 2012] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

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

[Agrawal *et al.*, 2017b] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*, 2017.

[Audibert *et al.*, 2013] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[Auer, 2002] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3(Nov):397–422, 2002.

[Bernstein *et al.*, 2017] Fernando Bernstein, Sajad Modaresi, and Denis Sauré. A dynamic clustering approach to data-driven assortment personalization. 2017.

[Bubeck and Cesa-Bianchi, 2012] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012.

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

[Cesa-Bianchi and Lugosi, 2012] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[Chen *et al.*, 2013] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *ICML*, pages 151–159, 2013.

[Chen *et al.*, 2016] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *NIPS*, pages 1659–1667, 2016.

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

[Chu *et al.*, 2011] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *IJCAI*, pages 208–214, 2011.

[Dani *et al.*, 2008] Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

[Davis *et al.*, 2013] James Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Technical Report, Cornell University*, pages 335–357, 2013.

[Gai *et al.*, 2012] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *TON*, 20(5):1466–1478, 2012.

[Golrezaei *et al.*, 2014] Negin Golrezaei, Hamid Nazerzadeh, and Paat Rusmevichientong. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014.

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

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

[Lagrée *et al.*, 2016] Paul Lagrée, Claire Vernade, and Olivier Cappe. Multiple-play bandits in the position-based model. In *NIPS*, pages 1597–1605. 2016.

[Luce, 2005] Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

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

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

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

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

[Thompson, 1933] William 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.

[Wang *et al.*, 2017] Yingfei Wang, Hua Ouyang, Chu Wang, Jianhui Chen, Tsvetan Asamov, and Yi Chang. Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *AAAI*, pages 2746–2753, 2017.

[Wen *et al.*, 2015] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *ICML*, pages 1113–1122, 2015.