Fixed-Budget Pure Exploration in Multinomial Logit Bandits

Boli Fang
Department of Computer Science, Indiana University
bfang@iu.edu

Abstract

In this paper we investigate pure exploration problem in Multinomial Logit bandit(MNL-bandit) under fixed budget settings, a problem motivated by real-time applications in online advertising and retailing. Given an MNL-bandit instance and a fixed exploration budget, our goal is to minimize the misidentification error of the optimal assortment. Towards such an end we propose an algorithm that achieve gap-dependent complexities, and complement our investigation with a discussion on recent studies and a minimax lower bound on misidentification probability. To the best of our knowledge, our paper is the first to address the recently proposed open problem of fixed-budget pure exploration problem for MNL-bandits.

1 Introduction

In this paper we investigate the Multinomial Logit Bandit(MNL-Bandit) problem [Rusmevichientong *et al.*, 2010; Agrawal *et al.*, 2016; Agrawal *et al.*, 2019], an important problem motivated by online revenue management, retail and advertising. The seller needs to select a subset(assortment) of items from a set of items available in storage so as to maximize the expected revenue from the item that the customer will purchase. The probability that the buyer will purchase items from the assortments follows the MNL choice model, one of the most widely used discrete choice models [Luce, 1959; Train, 2009]. Once the buyer makes a decision on which item to purchase or not to purchase any item at all, the seller receives an income corresponding to the value of the item. The goal of the seller is to learn the parameters of the MNL choice model and in turn maximize the expected revenue generated through buyer purchase by sequentially offering assortments. Such a model effectively reflects the dynamics in practical interactions such as retailing/advertisements, where the seller seeks to maximize the likelihood of customer purchase/advertisement links while showcasing only a limited number of merchandise/advertisements.

Prior work on MNL-bandits have centered on regret minimization as well as pure exploration with fixed-confidence(i.e. under a PAC setting). The problem of pure exploration under the *fixed-budget* setting, however, remains unexplored and has been identified as an open problem by [Karpov and Zhang, 2020]. Investigation on the pure exploration problem under fixed-budget setup is particularly helpful as it provides additional insight for the optimal algorithmic performance on all MNL-bandit instances in addition to the regret minimization objective in prior work. The central challenge in our setting is that the algorithm not only needs to learn the optimal preference parameters corresponding to the buyer, but also the corresponding optimal expected revenue. Most of existing algorithms for combinatorial bandits/subset selection, as pointed out by [Yang, 2021], do not automatically solve the MNL-bandit problem as the MNL-bandit has limited feedback compared to the usual settings such as semi-bandit/full-bandit setups. Additionally, unlike the case for regular Multi-armed Bandits where the 'importance' of all arms is independent from other arms, the pairwise 'importance' order between any two items in the MNL-bandit problem can be affected by the 'importance' of other items. Furthermore, whereas the seller may offer any assortment an arbitrary number of times in *fixed-confidence* algorithms, such flexibility is not possible in *fixed-budget* algorithms due to the limited number of assortment the seller can offer.

Towards these objectives, we propose in this paper an algorithm based on successive accept-reject principles to tackle the fixed-budget pure exploration problem in MNL-bandits. Our contributions can be summarized as follows:

- We propose FB-MNL-SAR, an algorithm based on the successive accept-reject(SAR) principles, to solve the fixed-budget pure exploration in Multinomial Logit Bandits(MNL-Bandits). To the best of our knowledge, this algorithm is the first to address the issue of fixed-budget exploration in MNL-Bandits.
- Based on the FB-MNL-SAR algorithm, We also derive *instance-dependent* upper bound on the misidenification probability that the final assortment deviates from the optimal assortment.
- We complement our analysis with results on instance-independent lower bounds, as well as a discussion on the near optimality of the FB-MNL-SAR algorithm with respect to worst-case/instance-independent objectives.

2 Related Work

There is a plethora of literature on multi-armed bandits and combinatorial bandits, and we refer interested readers to [Lattimore and Szepesvári, 2020] for a comprehensive overview. In particular, the subset-selection problem [Chen *et al.*, 2017; Chen *et al.*, 2014; Chen *et al.*, 2016; Rejwan and Mansour, 2020], in which an algorithm chooses a subset of arms from an existing universe of arms, is the most relevant to our MNL-bandit problem. Prior research has uncovered gap-independent and gap-dependent bounds under the subset selection settings, although those results are not directly applicable towards our problem because the MNL-bandit receives substantially more limited feedback in the form of the single purchased item reward as compared to the rewards corresponding to all items in the assortment.

The MNL-bandit problem is first introduced in [Rusmevichientong *et al.*, 2010; Sauré and Zeevi, 2013] as the *Dynamic Assortment Selection Problem*, under which the seller has to learn the buyer's preference over items in the selling horizon. Algorithms based on Upper-confidence bound(UCB) [Agrawal *et al.*, 2019] and Thompson Sampling(TS) [Agrawal *et al.*, 2017] have been demonstrated to achieve a tight instance-independent cumulative regret bound of $\tilde{O}(\sqrt{NT})$, matching a regret bound of $\tilde{\Omega}(\sqrt{NT})$ for the MNL-bandit problem in our setting. Some previous pieces of work [Agrawal *et al.*, 2016; Agrawal *et al.*, 2019] have considered a reduction of MNL-bandits to the regular bandit problem by modeling each assortment as an arm to be offered, but such modeling creates an exponential number of arms without using the inherent relationships between the items and hence do not give good bounds. Variations of the MNL-bandits in dueling settings have also been studied [Chen *et al.*, 2018; Saha and Gopalan, 2019]. More recently, [Yang, 2021; Karpov and Zhang, 2020] have independently investigated the problem of fixed-confidence pure exploration in MNL bandits, and obtained the first upper bounds on sample complexity for fixed-confidence pure exploration problems. The problem of fixed-budget pure exploration has been identified by [Karpov and Zhang, 2020] as an open problem with theoretical value.

3 Problem Setting and Preliminaries

We adopt the concepts/notations used in [Yang, 2021] and [Karpov and Zhang, 2020], and consider the Multinomial Logit choice model because of its simplicity, applicability and wide range of applications as pointed out by prior work [Luce, 1959; McFadden, 1973; Soufiani *et al.*, 2013; Train, 2009] . At each time step, a buyer decides which item to purchase(or whether or not to purchase) out of an assortment(the size of which is bounded by a capacity constraint parameter K) offered by the seller from the list of available items.

More specifically, a MNL-bandit instance can be expressed as a quadruple $\mathcal{I} = (N, K, \mathbf{r}, \mathbf{v})$. Given a set of N items $[N] = \{1, 2, ...N\}$, the seller offers an assortment $S \subseteq [N]$ such that $|S| \leq K$, and the buyer purchases item $i \in S \cup \{0\}$, after which the buyer obtains a reward $r_i \in [0, 1]$ corresponding to the item i. Here the item $i = 0$ stands for the 'no purchase' decision, and $\mathbf{r}, \mathbf{v}$ respectively denote the reward and preference parameter vectors. The probability for the buyer to choose item $i \in S \cup \{0\}$ can be expressed as

$$P_S^{\mathbf{v}}(i) = \frac{v_i}{v_0 + \sum_{j \in S} v_j},$$

where $v_i \in [0, 1]$ is the preference parameter for the item i. Without loss of generality, we set $v_0 = 1$ and $r_0 = 0$, such that the 'no purchase' decision yields 0 reward and is often the most frequent outcome as suggested by common convention (Agrawal et al. 2016, 2017, 2019). It follows that the expected revenue generated by the assortment S is:

$$R(S, \mathbf{v}) = \mathbb{E}_{i \sim P_S^{\mathbf{v}}}[r_i] = \sum_{i \in S} \frac{r_i v_i}{1 + \sum_{j \in S} v_j}.$$

We denote the optimal assortment with respect to preference vector $\mathbf{v}$ as

$$S_{\mathbf{v}} = \arg \max_{S \subseteq [N], |S| \leq K} R(S, \mathbf{v}),$$

and the corresponding optimal revenue $\theta_{\mathbf{v}} = R(S_{\mathbf{v}}, \mathbf{v})$. Given the definitions above, our goal is to find an assortment S satisfying the constraint $|S| \leq K$, within a known time horizon T time steps, such that the misidentification probability

$$e_T = \mathbb{P}[S \neq S_{\mathbf{v}}]$$

is minimized. Our definition follows naturally in light of recent parallel developments in *fixed budget* bandits, where the objective is to output an arm(or action) within a fixed number of exploration epochs. Without loss of generality as in the case of [Yang, 2021], we assume that the optimal assortment $S_{\mathbf{v}}$ is unique given the preference vector $\mathbf{v}$, so that the hardness complexity of our problem will be bounded as will be discussed in more details in subsequent sections.

4 Algorithmic Framework

To facilitate analysis of our fixed-budget optimal assortment problem, we start off by relating the MNL-bandit problem to the positive top-K item identification (PTop-K) problem.

Lemma 1. *[Rusmevichientong* et al.*, 2010; Yang, 2021; Karpov and Zhang, 2020] Consider the advantage score of each item i given a preference vector* $\mathbf{v}$*:*

$$\eta_i = (r_i - \theta_{\mathbf{v}})v_i.$$

Then the optimal assortment $S_{\mathbf{v}}$ is the set of all items such each item i in the set has score $\eta_i > 0$, and η_i is among the top K of all scores $\{\eta_i\}$.

Lemma 1 indicates the connection between MNL bandit and top-K/thresholding arm identification problems based on pairwise partial order relation of advantage scores, thereby making it possible to derive corresponding gap-dependent bounds on misclassification probability using similar techniques. Using a similar set of notations as those in [Karpov and Zhang, 2020], we assume without loss of generality that $\eta_1 \geq \eta_2 \geq ... \geq \eta_N$ for the N items in consideration(an observation which is unknown to the player).

Moreover, without loss of generality and in regards to the assortment uniqueness described in previous sections, we assume that all the items have unique values of advantage scores, i.e. $\eta_i > \eta_{i+1}$ for all items $i \in [N-1]$.

These assumptions are also essential in prior work using 'assortment-level' complexities, as it guarantees that the hardness of the problem does not go to infinity and that the best assortment is unique. They are also key in establishing the validity of our algorithmic estimations as follows, as demonstrated by subsequent sections.

4.1 A Two-phase Successive Accept-Reject Algorithm

To output an assortment satisfying these two requirements, we design a two-phase algorithm that estimates the items with top-K advantage scores before choosing the items with the positive advantage scores. The high-level intuition of our approach is that any item which belongs to the optimal assortment must have advantage scores that are in the top-K amongst all items; given these top-K items, we subsequently choose items with positive advantage scores, and return the corresponding item set as the optimal assortment.

Towards the respective objectives of the 2 phases, we adopt the Successive Accept-Reject framework commonly used in previous work [Yang, 2021; Bubeck *et al.*, 2013; Chen *et al.*, 2017; Rejwan and Mansour, 2020]. Our algorithm consists of two subroutines: the first estimates the Top-K items and the second subroutine estimates the positive advantage-score item out of the top-K. The T time steps are divided into $N+K$ rounds, with the first subroutine using N rounds and the second using K rounds.

In the first subroutine Top-K-EST, the high level idea is to iteratively accept/reject items based on the upper/lower confidence estimates of the advantage scores $\hat{\eta}_i$, $\check{\eta}_i$, which are computed from the upper/lower confidence estimates of the optimal revenue $\hat{\theta}_{\mathbf{v}}, \check{\theta}_{\mathbf{v}}$. The confidence estimates of the optimal revenue $\hat{\theta}_{\mathbf{v}}, \check{\theta}_{\mathbf{v}}$, in turn, are constructed from the upper/lower confidence estimates of the preference vector $\hat{\mathbf{v}}, \check{\mathbf{v}}$ by considering the maximal revenue with respect to the vectors $\hat{\mathbf{v}}, \check{\mathbf{v}}$. Assume, without loss of generality, that $\eta_1 \geq \eta_2 \geq ... \geq \eta_N$ for all elements in the set $\{\eta_i\}_{i=1}^N$. More specifically, in addition to the reward gaps defined in previous section, we define the *top-K advantage gap* of an item i as

$$\Delta_i^{(K)} = \begin{cases} \eta_i - \eta_{K+1} & \text{if } i \leq K, \\ \eta_K - \eta_i & \text{if } i \geq K+1, \end{cases}$$

We also define the estimate of *top-M advantage gap*, with respect to the upper ($\hat{\eta}_i$) and lower ($\check{\eta}_i$) estimates of advantage scores η_i's, as

$$\tilde{\Delta}_{\sigma_\tau(i)}^{(M)} = \begin{cases} \check{\eta}_{\sigma_\tau(i)} - \hat{\eta}_{\sigma_\tau(M+1)} & \text{if } i \leq M, \\ \check{\eta}_{\sigma_\tau(M)} - \hat{\eta}_{\sigma_\tau(i)} & \text{if } i \geq M+1. \end{cases}$$

We adopt the notations used in [Bubeck *et al.*, 2013], and define σ_τ as the bijection from $[N+1-\tau]$ to the pending set P_τ such that $\hat{\eta}_{\sigma_\tau(1)} \geq \hat{\eta}_{\sigma_\tau(2)} \geq ... \geq \hat{\eta}_{\sigma_\tau(N+1-\tau)}$. At each round, we consider an item i with the largest value of $\tilde{\Delta}_{\sigma_\tau(i)}^{(M)}$. We remove item i from the pending set. If $\check{\eta}_i \geq \hat{\eta}_{\sigma_\tau(M+1)}$, we accept item i, and decrease M by 1; if $\check{\eta}_{\sigma_\tau(M)} \geq \hat{\eta}_i$, we reject item i. The algorithm repeats such a procedure until all the items are either accepted or rejected. Details of the top-K item estimation subroutine Top-K-EST can be found in Algorithm 1.

Algorithm 1 Top-K-EST

Input: Items $\mathcal{I} = [N]$, Capacity Parameter K, N Exploration Epochs $\{T_\tau\}_{\tau=1}^N$.

Initialization: Accepted Item Set $A_0 = \emptyset$, Pending Item Set $P_0 = [N]$, Estimated upper/lower confidence values of preference factors v_i $\hat{v}_i = 1, \check{v}_i = 0$, Counter of maximum number of items to be accepted $M = K$. Estimated gaps $\tilde{\Delta}_{\sigma_0(i)}^{(M)} = 1$ for all items i.

```
for $\tau = 1, 2, ..., N$ do
    for all $i \in P_\tau$ do
        Offer singleton assortment $\{i\}$ $(T_\tau - T_{\tau-1})$ times;
        $x_i^{(\tau)} \leftarrow$ the ratio of $\{i\}$ being rejected out of a total of $T_\tau$ offerings;
        $v_i^{(\tau)} \leftarrow \min\{\frac{1}{x_i^{(\tau)}} - 1, 1\}$; {Denote the vector of $v_i$'s as $\mathbf{v}$}
        $\check{v}_i \leftarrow \max\{v_i^{(\tau)} - \frac{\tilde{\Delta}_{\sigma_\tau(i)}^{(M)}}{8K}, 0\}, \hat{v}_i \leftarrow \min\{v_i^{(\tau)} + \frac{\tilde{\Delta}_{\sigma_\tau(i)}^{(M)}}{8K}, 1\}$; {Corresponding vectors $\check{\mathbf{v}}, \hat{\mathbf{v}}$}
        $\check{\theta} \leftarrow \max_{S \subseteq A \cup P} R(S, \check{\mathbf{v}}), \hat{\theta} \leftarrow \max_{S \subseteq A \cup P} R(S, \hat{\mathbf{v}})$; {Estimates of optimal revenue}
        $\check{\eta}_i \leftarrow \min\{\check{v}_i(r_i - \hat{\theta}), \hat{v}_i(r_i - \hat{\theta})\}, \hat{\eta}_i \leftarrow \max\{\check{v}_i(r_i - \check{\theta}), \hat{v}_i(r_i - \check{\theta})\}$; {Estimates of advantage scores}
    end for
    Sort $\{\check{\eta}_i : i \in P_\tau\}, \{\hat{\eta}_i : i \in P_\tau\}$ by decreasing order;
    Compute estimates $\tilde{\Delta}_{\sigma_\tau(i)}^{(M)}$'s and order them by $\sigma_\tau(i)$.
    Remove $i \leftarrow \arg\max \tilde{\Delta}_{\sigma_\tau(i)}^{(M)}$ from Pending Set $P_{\tau-1}$, $P_\tau = P_{\tau-1} \backslash \{i\}$.
    if $\check{\eta}_{\sigma_\tau(i)} > \hat{\eta}_{\sigma_\tau(M+1)}$ then
        $A_\tau \leftarrow A_{\tau-1} \cup \{i\}$.
        $M = M - 1$.
    else if $\check{\eta}_{\sigma_\tau(M)} > \hat{\eta}_{\sigma_\tau(i)}$ then
        Reject item $i$.
    end if
end for
```

Return: Estimated Top-K advantage score item set A_N.

After running the Top-K-EST subroutine, we use an additional K epochs for the subroutine in the 2nd phase to obtain the positive advantage-score items from the top-K item set returned by the first phase. Again, at each round after building up confidence intervals, the rules are to accept pending items with positive scores, and to reject pending items with negative scores. When phase 2 terminates, the subset of K items with estimated positive advantage scores is returned as the optimal assortment. To reduce the randomness involved in estimators, a key difference between Positive-EST and Top-K-EST is that the estimator being considered is no longer the advantage score used in Top-K-EST, but the *gap* $\Delta_i^{\text{gap}} = |r_i - \theta_{\mathbf{v}}|$ between individual item reward r_i and the optimal revenue $\theta_{\mathbf{v}}$. Similar to the case of top-M advantage gap, we define the corresponding estimates of the gap Δ_i^{gap} as follows:

$$\tilde{\Delta}_i^{\text{gap}} = \begin{cases} r_i - \hat{\theta}_{\mathbf{v}} & \text{if } r_i \geq \hat{\theta}_{\mathbf{v}}, \\ 0 & \text{if } \hat{\theta}_{\mathbf{v}} \geq r_i \geq \check{\theta}_{\mathbf{v}}, \\ \check{\theta}_{\mathbf{v}} - r_i & \text{if } \check{\theta}_{\mathbf{v}} \geq r_i, \end{cases}$$

The flowchart Positive-EST provides an overview of the estimation procedure at phase 2.

Algorithm 2 Positive-EST

```
Input: Estimated Top-K Items $\mathcal{I}$, $K$ Exploration Epochs $\{T_\kappa\}_{\kappa=N+1}^{N+K}$.
Initialization: Accepted Item Set $B_0 \leftarrow \emptyset$, Pending Item Set $P_0 = \mathcal{I}$. $\tilde{\Delta}_i^{\text{gap}} = 0$ for all $i$.
for $\kappa = N+1, ..., N+K$ do
    for all $i \in P_\kappa$ do
        Offer singleton assortment $\{i\}$ $(T_\kappa - T_{\kappa-1})$ times;
        $x_i^{(\kappa)} \leftarrow$ the ratio of $\{i\}$ being rejected out of a total of $T_\kappa$ offerings, including those in subroutine Top-K-EST;
        $v_i^{(\kappa)} \leftarrow \min\{\frac{1}{x_i^{(\kappa)}} - 1, 1\}$; {Denote the vector of $v_i$'s as $\mathbf{v}$}
        $\check{v}_i \leftarrow \max\{v_i^{(\kappa)} - \frac{\tilde{\Delta}_i^{\text{gap}}}{8K}, 0\}, \hat{v}_i \leftarrow \min\{v_i^{(\kappa)} + \frac{\tilde{\Delta}_i^{\text{gap}}}{8K}, 1\}$; {Corresponding vectors $\check{\mathbf{v}}, \hat{\mathbf{v}}$}
        $\check{\theta} \leftarrow \max_{S \subseteq A \cup P} R(S, \check{\mathbf{v}}), \hat{\theta} \leftarrow \max_{S \subseteq A \cup P} R(S, \hat{\mathbf{v}})$; {Estimates of optimal revenue}
    end for
    Compute $\tilde{\Delta}_i^{\text{gap}}$ for each pending item $i \in P_\tau$.
    Remove $i \leftarrow \arg\max \tilde{\Delta}_i^{\text{gap}}$ from Pending Set $P_\tau$.
    if $r_i \geq \hat{\theta}_{\mathbf{v}}$ then
        $B \leftarrow B \cup \{i\}$.
    else if $r_i \leq \check{\theta}_{\mathbf{v}}$ then
        Reject item $i$.
    else
        Accept or reject $i$ with equal probability.
    end if
end for
Return: Estimated Top-K Positive Revenue Item Set $B$.
```

Combining the two subroutines executed the two phases as well as the estimation procedure of the preference vector, the FB-MNL-SAR algorithm is described in details in Algorithm 3. Theorem 1 provides the theoretical guarantee of our 2-phase algorithm algorithm.

Theorem 1. *Given an exploration budget $T > N$, the 2-stage FB-MNL-SAR algorithm returns the optimal assortment S with misidentification probability*

Algorithm 3 FB-MNL-SAR

```
Input: Items $\mathcal{I} = [N]$, Capacity Parameter $K$, Exploration Budget $T$.
Initialization: Accepted item set $A_1 = \emptyset$, Pending item set $P_1 = [N]$.
Compute $T_\tau = \lceil \frac{1}{\overline{\log}(N+K)} \frac{T-N}{N+K+1-\tau} \rceil$ for $\tau \in [N+K-1]$, $T_0 = 0$.
Obtain Top-K items $A \leftarrow$ Top-K-EST$(\mathcal{I}, K, \{T_\tau\}_{\tau=1}^{N})$.
Obtain $S \leftarrow$ Positive-EST$(A, \{T_\kappa\}_{\kappa=N+1}^{N+K})$.
Return: The best assortment $S$.
```

$$e_T = O\Big(KN^2 \exp\Big(-\frac{(T-N)}{K^2 \overline{\log}(N+K) H_1^{(K)}}\Big) + K^3 \exp\Big(-\frac{(T-N)}{K^2 \overline{\log}(N+K) H_2^{gap}}\Big)\Big),$$

where

$$\overline{\log}(N+K) = \sum_{i=1}^{N+K} \frac{1}{i},$$

$$H_1^{(K)} = \max_{i \in [N]} \frac{i}{(\Delta_i^{(K)})^2}, \quad H_2^{gap} = \max_{i \in [N]} \frac{i}{(\Delta_i^{gap})^2},$$

are respective intrinsic hardness quantities of our MNL Bandit instance in consideration.

Proof. Notice that conditioned on the correctness of subroutines in phase 1 and phase 2, the misidentification probability e_T can be expressed as

$$\begin{aligned} e_T &= \mathbb{P}[S \neq S_{\mathbf{v}} | \text{Top-K-EST wrong}]\mathbb{P}[\text{Top-K-EST wrong}] \\ &+ \mathbb{P}[S \neq S_{\mathbf{v}} | \text{Top-K-EST right}]\mathbb{P}[\text{Top-K-EST right}] \\ &\leq \mathbb{P}[\text{Top-K-EST wrong}] + \mathbb{P}[S \neq S_{\mathbf{v}} | \text{Top-K-EST right}] \\ &\leq \mathbb{P}[\text{Top-K-EST wrong}] \\ &+ \mathbb{P}[\text{Positive-EST wrong} | \text{Top-K-EST right}]. \end{aligned}$$

Our last inequality holds by the rule of total probability, the definition of conditional probability, and the observation that the FB-MNL-SAR algorithm returns the correct optimal assortment with certainty if all the top-K and the positive items in the top-K subset are identified correctly. Therefore, it suffices to bound the misclassification probability incurred by the 1st Top-K-EST and the 2nd Positive-EST subroutine.

We first of all present a lemma that the optimal revenue $\theta_{\mathbf{v}} = R(S_{\mathbf{v}}, \mathbf{v})$ does fall into the confidence interval $[\check{\theta}, \hat{\theta}]$ we have constructed from corresponding lower and upper $(\check{v}_i, \hat{v}_i)$ estimates of the preference factor v_i at every epoch:

Lemma 2. *At each epoch τ, we have $\check{\theta}_{\mathbf{v}} \leq \theta_{\mathbf{v}} \leq \hat{\theta}_{\mathbf{v}}$ in both the Top-K-EST and the Positive-EST subroutines.*

By the observation that $0 \leq \check{v}_i \leq v_i \leq \hat{v}_i \leq 1$ and simple algebra, we obtain the following corollary that demonstrates the validity of confidence intervals constructed for advantage scores:

Corollary 1. *Per the lower/upper estimates of the optimal revenue $\theta_{\mathbf{v}}$ produced by line 9 of algorithm 1, the advantage score $\eta_i \in [\check{\eta}_i, \hat{\eta}_i]$ for each item i.*

Given the corollary above which demonstrates the correctness of our estimation, we then establish the following lemma that builds the relationship between estimated $\theta_{\mathbf{v}}$ and the quantitatively.

Lemma 3. *Within both Top-K-EST and Positive-EST subroutines, the upper and lower estimates of the advantage scores $\eta_i^{(\tau)}$ for item i at round τ satisfy:*

$$\hat{\eta}_i^{(\tau)} - \check{\eta}_i^{(\tau)} \leq 2 \sum_{j \in S} |\hat{v}_j - \check{v}_j|.$$

Correctness and Probability of Top-K-EST

For the error contributed by misidentification in Top-K-EST subroutine at phase 1, On a high-level, we can apply similar reasoning as that in [Bubeck *et al.*, 2013] to demonstrate that our Top-K-EST subroutine returns the Top-K items as long as the η_i's are well-estimated to the degree specified by event F. Similar to the proof in [Bubeck *et al.*, 2013], we will consider the event

$$F = \{\forall i \in [N], \forall \tau \in [N], |\hat{\eta}_i - \check{\eta}_i| \leq \frac{\Delta^{(K)}_{(N+1-\tau)}}{4}\}.$$

Since there is at least 1 gap between the $K/K-1$-th largest advantage score and the $(N+1-\tau)$-th largest advantage score, the following relationship holds:

$$\tilde{\Delta}^{(M)}_{\sigma_\tau(i)} \leq \Delta^{(M)}_{\sigma_\tau(i)} \leq \Delta^{(K)}_{(N+1-\tau)}$$

Since η_i can be arbitrarily close to either one of the two end points of its confidence interval $[\check{\eta}_i, \hat{\eta}_i]$, we set the length of the confidence interval for v_i, by our observation between the relationships of estimates in Lemma 3, as reflected by line 6 in computing our upper/lower estimates of v_i:

$$\hat{v}_i - \check{v}_i \leq \frac{\tilde{\Delta}^{(M)}_{\sigma_\tau(i)}}{4K}.$$

Notice that the items that are eliminated out of the pending set have either very high advantage scores or very low advantage scores with respect to the K-th or $(K+1)$-th largest advantage scores per our construction of $\Delta_i^{(K)}$, and the event F is a superset of the event regarding actual estimation of η_i's. Therefore, it suffices to show that conditioned on F, the Top-K-EST subroutine identifies all top-K items correctly, and that $\bar{F}$ occurs with small probability.

By Chernoff-Hoeffding inequality, we are able to establish the following inequality between estimated rejection ratio $x_i^{(\tau)}$ and the actual rejection ratio x_i as determined by v_i for a given number $t \geq 0$:

$$\mathbb{P}[|x_i^{(\tau)} - x_i| \geq t] \leq 2\exp\left(-T_\tau t^2\right).$$

We thus consider the event where all the x_i's are well-estimated, defined as follows:

$$\mathcal{X} = \{\forall i \in [N], \forall \tau \in [N] : |x_i^{(\tau)} - x_i| < \frac{\Delta^{(K)}_{N+1-\tau}}{16K}\}.$$

Conditioned on $\mathcal{X}$, it holds that

$$x_i^{(\tau)} \geq x_i - \frac{\Delta^{(K)}_{N+1-\tau}}{16K}.$$

Together with the observations that $\Delta^{(K)}_{N+1-\tau} \leq 2$ for all τ, $K \geq 0$, and $x_i = \frac{1}{1+v_i} \geq \frac{1}{2}$, we can conclude that given the event $\mathcal{X}$,

$$|v_i^{(\tau)} - v_i| = |\frac{x_i - x_i^{(\tau)}}{x_i x_i^{(\tau)}}| \leq \frac{|x_i - x_i^{(\tau)}|}{1/2 \cdot 3/8} = \frac{16}{3}|x_i - x_i^{(\tau)}|.$$

We also notice that $|v_i^{(\tau)} - v_i| = \frac{1}{2}|\hat{v}_i^{(\tau)} - \check{v}_i^{(\tau)}|$. Using a union bound and the equations we derived in previous sections, it follows that the probability of some items are *not well estimated* by the Top-K-EST subroutine is at most:

$$\begin{aligned}
\mathbb{P}[\bar{F}] &\leq \sum_{i\in[N]} \sum_{\tau\in[N]} \mathbb{P}[|\hat{\eta}_i^{(\tau)} - \check{\eta}_i^{(\tau)}| \geq \frac{\Delta^{(K)}_{N+1-\tau}}{4}] \\
&\leq \sum_{i\in[N]} \sum_{\tau\in[N]} \mathbb{P}[\sum_{i\in S} |\hat{v}_i^{(\tau)} - \check{v}_i^{(\tau)}| \geq \frac{\Delta^{(K)}_{N+1-\tau}}{8}] \\
&\leq \sum_{i\in[N]} \sum_{\tau\in[N]} K\mathbb{P}[|\hat{v}_i^{(\tau)} - \check{v}_i^{(\tau)}| \geq \frac{\Delta^{(K)}_{N+1-\tau}}{8K}] \\
&\leq \sum_{i\in[N]} \sum_{\tau\in[N]} K\mathbb{P}[|v_i^{(\tau)} - v_i| \geq \frac{\Delta^{(K)}_{N+1-\tau}}{16K}] \\
&\leq \sum_{i\in[N]} \sum_{\tau\in[N]} K\mathbb{P}[|x_i^{(\tau)} - x_i| \geq \frac{3\Delta^{(K)}_{N+1-\tau}}{256K}] \\
&\leq \sum_{i\in[N]} \sum_{\tau\in[N]} 2K\exp(-2T_\tau(\frac{3\Delta^{(K)}_{N+1-\tau}}{256K})^2) \\
&\leq 2KN^2\exp(-\frac{9(T-N)}{16394K^2\overline{\log}(N+K)H_1^{(K)}}).
\end{aligned}$$

where the last inequality holds because $0 \leq \tau \leq N$ and

$$\begin{aligned}
T_\tau(\Delta^{(K)}_{N+1-\tau})^2 &\geq \frac{T-N}{\overline{\log}(N+K+1-\tau)(\Delta^{(K)}_{N+1-\tau})^{-2}} \\
&\geq \frac{T-N}{\overline{\log}(N+K)H_1^{(K)}}.
\end{aligned}$$

The following lemma shows the correctness of Top-K advantage score items thus retrieved.

Lemma 4. *Denote the set of items with top-K advantage scores as S_1. Then at any round τ in the Top-K-EST subroutine, $A_\tau \subseteq S_1 \subseteq A_\tau \cup P_\tau$.*

By Lemma 4, we have shown that conditioned on the event F, the Top-K-EST subroutine will always terminate with the correct Top-K output, because there is always at least 1 item that will be accepted or rejected in each epoch in both of the subroutines. Since there is a finite number of rounds of at most $N-1$ and we are keeping an adaptive counter M to record the number of additional items to be included, the item set returned is guaranteed to have size equal to K, containing all Top-K items when the event F holds.

Correctness and Probability of Positive-EST

Similarly, for the error contributed by misidentification in the Positive-EST subroutine at phase 2, we consider the event G such that the team-level revenue $\theta_{\mathbf{v}}$ is well-estimated in a reasonably small interval at all rounds:

$$G = \{\forall i \in [N], \forall \tau = N+1, ..., N+K : |\hat{\theta}_{\mathbf{v}} - \check{\theta}_{\mathbf{v}}| \leq \Delta_i^{\text{gap}}\},$$

Notice that conditioned on event G, if any error is produced by Positive-EST, then we have $\check{\theta}_{\mathbf{v}} \leq r_i \leq \hat{\theta}_{\mathbf{v}}$. This is because item i is a positive item if $r_i > \hat{\theta}_{\mathbf{v}} > \theta_{\mathbf{v}}$, and item i is a negative item if $r_i < \check{\theta}_{\mathbf{v}} < \theta_{\mathbf{v}}$ per Lemma 2 and the observation that $\tilde{\Delta}_i^{\text{gap}} \leq \Delta_i^{\text{gap}}$ for all i. When $\check{\theta}_{\mathbf{v}} \leq r_i \leq \hat{\theta}_{\mathbf{v}}$, $|r_i - \theta_{\mathbf{v}}| \leq |\check{\theta}_{\mathbf{v}} - \hat{\theta}_{\mathbf{v}}|$, and $|\hat{\theta}_{\mathbf{v}} - \check{\theta}_{\mathbf{v}}| > \Delta_i^{\text{gap}}$. Hence, the probability that each item i in the top-K item set gets misclassified across all stages τ does not exceed

$$\mathbb{P}[\bar{G}] = \mathbb{P}[\exists i \in \{\text{top-K-EST}\}, \tau = N+1, ..., N+K : \\ |\hat{\theta}_{\mathbf{v}}^{(\tau)} - \check{\theta}_{\mathbf{v}}^{(\tau)}| \geq \Delta_i^{\text{gap}}].$$

Again, by Chernoff-Hoeffding, union bound, and pigeonhole principle, it follows that the probability of some items not well estimated is at most:

$$\begin{aligned}
&\sum_{i \in \text{Top-K-EST}} \sum_{\tau=N+1}^{N+K} \mathbb{P}[|\hat{\theta}_{\mathbf{v}}^{(\tau)} - \check{\theta}_{\mathbf{v}}^{(\tau)}| \geq \Delta_i^{\text{gap}}] \\
&\leq \sum_{i \in \text{Top-K-EST}} \sum_{\tau=N+1}^{N+K} K\mathbb{P}[|\hat{v}_i^{(\tau)} - \check{v}_i^{(\tau)}| \geq \frac{\Delta_i^{\text{gap}}}{K}] \\
&\leq \sum_{i \in \text{Top-K-EST}} \sum_{\tau=N+1}^{N+K} K\mathbb{P}[|v_i^{(\tau)} - v_i| \geq \frac{\Delta_i^{\text{gap}}}{2K}] \\
&\leq \sum_{i \in \text{Top-K-EST}} \sum_{\tau=N+1}^{N+K} K\mathbb{P}[|x_i^{(\tau)} - x_i| \geq \frac{3\Delta_i^{\text{gap}}}{32K}] \\
&\leq \sum_{i \in \text{Top-K-EST}} \sum_{\tau=N+1}^{N+K} K\exp(-2T_\tau(\frac{3\Delta_i^{\text{gap}}}{32K})^2) \\
&\leq 2K^3 \exp(-\frac{9(T-N)}{512K^2\overline{\log}(N+K)H_2^{\text{gap}}}).
\end{aligned}$$

We now analyze the correctness of the 2nd subroutine Positive-EST. Again, conditioned on the event G, the subroutine Positive-EST returns an error at each round τ if and only if an item with non-positive advantage score has been accepted or an item with positive advantage score has been rejected. For either, conditioned on the event G, we can similarly use induction to conclude that the items returned have positive advantage scores amongst the top-K items.

□

Remarks. A direct adaptation of the fixed-confidence SAR algorithm described in [Yang, 2021] by setting a cut off to the number of rounds τ does not guarantee correctness of output, because it is uncertain whether the fixed-confidence algorithms will terminate within the T time steps. The main reason is that the fixed-confidence algorithms offer item assortment an *arbitrary* amount of times, and it may take an arbitrary long period of time to produce an assortment of size K from the original N items. Hence, the returned assortment can be incorrect with extra non-optimal items even when the algorithm estimates all the advantage scores correctly with high probability. A similar argument also holds for the variation of the successive-accept algorithm with sweep-line optimal revenue search described in [Karpov and Zhang, 2020].

5 Discussion on Lower Bound

Our algorithm is based on the successive accept-reject(SAR) principle, and can be seen as the fixed-budgeted counterpart to the fixed-confidence SAR algorithms proposed by [Yang, 2021]. Although the algorithms described in [Yang, 2021] center on minimizing sample complexity rather than the misclassification probability in our paper, both algorithms can be translated to a minimax(i.e. worse-case gap-independent) upper bound of $\tilde{O}(\sqrt{NT})$ ($\tilde{O}$ suppresses all poly-logarithmic factors) using a similar expectation argument in [Yang, 2021], in which a regret-minimizing algorithm is run for T rounds to generate T assortments and the returned assortment is chosen randomly out of the T assortments. This bound is tight to the existing $\Omega(\sqrt{NT})$ lower bound on cumulative regret up to a poly-logarithmic factor, suggesting the near-optimality of the SAR algorithm with respect to minimax bounds.

To complement our analysis of upper bound on e_T, we further show a worst-case minimax lower bound, as stated by Theorem 2 below.

Theorem 2. *Any exploration algorithm that solves an MNL-bandit instance in T rounds will incur a misclassification probability $e_T = \Omega(\sqrt{\frac{N}{TK^2}})$, where N, K are corresponding parameters of the MNL-bandit instance.*

6 Conclusions

We investigate in this paper the open problem of fixed-budget pure exploration for Multinomial Logit Bandits as proposed by [Karpov and Zhang, 2020], and propose the first instance-dependent algorithm/upper bound as well as the first minimax lower bound of the problem.

There are several valuable future directions to pursue. To begin with, while we have identified a minimax lower bound, finding a problem-dependent lower bound for the MNL bandit problem remains an open challenge. Additionally, since existing measures of complexity ([Agrawal *et al.*, 2016; Yang, 2021; Karpov and Zhang, 2020]) are not directly comparable, it will be insightful to develop a theoretical framework that covers all existing instance-dependent bounds on these problems. Furthermore, it is still an open question to design pure exploration algorithms for MNL-bandit problems involving additional nested structures(e.g. [Chen *et al.*, 2021]).

Acknowledgments

Boli Fang has been supported by NSF grant CCF-2006591.

References

[Agrawal *et al.*, 2016] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, 2016.

[Agrawal *et al.*, 2017] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*. PMLR, 2017.

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

[Bubeck *et al.*, 2013] Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. ICML, 2013.

[Chen *et al.*, 2014] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2014.

[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 *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS, 2016.

[Chen *et al.*, 2017] Jiecao Chen, Xi Chen, Qin Zhang, and Yuan Zhou. Adaptive multiple-arm identification. In *Proceedings of the 34th International Conference on Machine Learning*, pages 722–730. PMLR, 2017.

[Chen *et al.*, 2018] Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA, 2018.

[Chen *et al.*, 2021] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment selection under the nested logit models. *Production and Operations Management*, 30(1):85–102, 2021.

[Karpov and Zhang, 2020] Nikolai Karpov and Qin Zhang. Instance-sensitive algorithms for pure exploration in multinomial logit bandit. *CoRR*, abs/2012.01499, 2020.

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

[Luce, 1959] R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

[McFadden, 1973] D. McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press New York, New York, NY, USA, 1973.

[Rejwan and Mansour, 2020] Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2020.

[Rusmevichientong *et al.*, 2010] 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.

[Saha and Gopalan, 2019] Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019.

[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.

[Soufiani *et al.*, 2013] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Preference elicitation for general random utility models. *CoRR*, 2013.

[Train, 2009] Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge Books. Cambridge University Press, 2009.

[Yang, 2021] Jiaqi Yang. Fully gap-dependent bounds for multinomial logit bandit. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 199–207. PMLR, 2021.