Fully Gap-Dependent Bounds for Multinomial Logit Bandit

Jiaqi Yang
Tsinghua University
`yangjq17@gmail.com`

Abstract

We study the multinomial logit (MNL) bandit problem, where at each time step, the seller offers an assortment of size at most K from a pool of N items, and the buyer purchases an item from the assortment according to a MNL choice model. The objective is to learn the model parameters and maximize the expected revenue. We present (i) an algorithm that identifies the optimal assortment S^* within $\widetilde{O}(\sum_{i=1}^{N} \Delta_i^{-2})$ time steps with high probability, and (ii) an algorithm that incurs $O(\sum_{i \notin S^*} K\Delta_i^{-1} \log T)$ regret in T time steps. To our knowledge, our algorithms are the *first* to achieve gap-dependent bounds that *fully* depends on the suboptimality gaps of *all* items. Our technical contributions include an algorithmic framework that relates the MNL-bandit problem to a variant of the top-K arm identification problem in multi-armed bandits, a generalized epoch-based offering procedure, and a layer-based adaptive estimation procedure.

1 Introduction

The multinomial logit bandit (MNL-bandit) problem is an important problem in online revenue management and has attracted much attention from both operations research and online learning literature (Kök and Fisher, 2007; Rusmevichientong et al., 2010; Sauré and Zeevi, 2013; Agrawal et al., 2016, 2017; Chen and Wang, 2018; Agrawal et al., 2019; Wang et al., 2018). In MNL-bandit, at each time step, the seller offers an assortment of size at most K from the pool of N homogeneous items and the buyer purchases an item from the assortment according to the MNL choice model, which is arguably the simplest and most widely used discrete choice model (Train, 2009; Luce, 2012; Soufiani et al., 2013) and has deep theoretical foundations (McFadden, 1973). The objective of the seller is to learn the model parameters and maximize the expected revenue through sequentially offering the assortments. MNL-bandit captures the essence of many real-world applications, such as retailing, where the retailer presents a limited number of products on the shelf and the customer purchases an item according to the choice model, and online advertising, where the ad platform displays a limited number of ads and the user clicks one ad according to the choice model.

In this paper, we study the PAC (probably approximately correct) exploration problem and the regret minimization problem in MNL-bandit, with a focus on proving *fully gap-dependent* sample complexity and regret bounds that depend on the suboptimality gaps of *all* items (detailed in Section 2). There are strong practical motivations to study these bounds, because they adapt to every MNL-bandit instance and thus lead to better performances on good practical instances. Unfortunately, there is a lack of studies on such bounds in previous MNL-bandit literature, and bounds in other bandits problems focusing on subset selection do not directly translate to our setting due to the limited feedback issue. We review these in Section 1.1 after introducing our challenge, results, and technical contributions.

A central challenge in obtaining fully gap-dependent bounds for MNL-bandit is that the partial order between two items can be *interfered* by other items. We recall an important reason why we have such bounds in other bandits settings is that we can obtain pairwise partial orders between arms to early decide on the optimality of some arms. Arms being decided need no longer be explored and stop contributing to the bounds. For example, in the top-K arm identification problem, we obtain the partial order between two arms by comparing their estimations (which is independent of the estimations of other items) and we decide on arms by

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

whether they could have top-K means. While in our setting, the comparison of two items can be interfered by other items through changing their weights in the revenue function. We underscore that the changing of weights is done by changing the denominator of the *fractional* revenue function, which contains model parameters of *other* items.

Results We define the gaps for items and the problems in Section 2. Our definitions match the intuition and naturally extend the definitions in other bandits settings.

Our main results are three MNL-bandit algorithms with fully gap-dependent guarantees. For the PAC exploration problem, we present a δ-PAC algorithm with sample complexity $\widetilde{O}(\sum_{i=1}^{N} \Delta_i^{-2})$ and a (δ, ε)-PAC algorithm with similar guarantee. For the regret minimization problem, we present an algorithm with $O(\sum_{i \in [N] \setminus S^*} \frac{K \log T}{\Delta_i})$ regret bound.

When $K = 1$, our MNL-bandit setting becomes the multi-armed bandit setting and our bounds recover their instance-optimal sample complexity and regret bounds (Lai and Robbins, 1985; Auer et al., 2002; Slivkins et al., 2019; Lattimore and Szepesvári, 2020). When $K \geq 2$, our regret bound recovers the $O(\frac{N^2 \log NT}{\Delta})$ global gap-dependent regret bound in (Agrawal et al., 2019), because by definition we have $\Delta_i \geq \Delta$, where $\Delta = \theta^* - \max_{S \subseteq [N]: |S| \leq K} \{R(S, \boldsymbol{v}) : R(S, \boldsymbol{v}) \neq \theta^*\}$ is the gap between the optimal and second-best assortments. We compare our sample complexity bound with a previous gap-independent bound in Section 3.4 after presenting the theorems.

Technical Contributions We present our three techniques under the context of the PAC exploration problem in Section 3 and we extend them to the regret minimization problem in Section 4.

Our first technique is an algorithmic framework in Section 3.1, which resolves our central challenge by the relation in Proposition 1. The relation suggests we obtain the pairwise partial order of two items by comparing the confidence intervals of their advantage scores (Definition 2) and early decide the items according to whether they could have positive and top-K advantage scores. Since this early decision rule is similar to that of the top-K arm identification problem, we modify the successive accept-reject algorithm for the latter problem to obtain an algorithm with fully gap-dependent guarantees. We add a caveat that the framework itself does *not* conclude, because estimating the advantage score is not a trivial job. As we show in Lemma 3.2, a naive estimation procedure using only methods in previous work could lead to two extra K factors in the sample complexity bound.

Our second technique eliminates an extra K factor by removing some dependency in the naive procedure, as we present in Section 3.2. An anomaly in the naive procedure is that we need to explore accepted items even though we do not need their scores. To remove this dependency on accepted items, we define a reduced revenue function that requires estimating a *ratio* of the model parameters. However, previous work only showed how to estimate the parameters themselves using the epoch-based offering procedure (Agrawal et al., 2016, 2019), with which we have to separately estimate the numerator and denominator of the ratio and suffer a huge error in the estimation. To resolve this, we generalize the epoch-based offering procedure to *directly* estimate the ratio.

Our third technique eliminates another extra K factor by a layer-based adaptive estimation procedure in Section 3.3. By carefully examining the error sources in the estimations of advantage scores, we find that the number of explorations for each item should adapt to the error it incurs, but exact adaption requires full knowledge of the unknown model parameters. So we surrogate by putting items with similar adaption requirements into the same layer and handling them altogether. We emphasize that the layers are still *unknown* and they could *vary* from phase to phase. We highlight that our surrogate method re-estimates the layers for *each* phase while only paying the sample complexity cost *once*.

We remark that our first technique indeed provides a systematic way to apply fractional programming (the method that proves our relation proposition) to online learning settings. Thus it may be of independent interests. Our second and third techniques utilizes the delicate structure of the MNL model, which could inspire future studies on MNL-bandit and other bandits with MNL model.

1.1 Related Work

MNL-bandit was first studied in (Rusmevichientong et al., 2010; Sauré and Zeevi, 2013), where the algorithms required the knowledge of the global suboptimality gap Δ in advance. Upper confidence bound-type algorithm and Thompson sampling were shown to achieve an $\widetilde{O}(\sqrt{NT})$ minimax regret bound (Agrawal et al., 2016, 2017). A matching $\Omega(\sqrt{NT})$ regret lower bound was shown in (Chen and Wang, 2018). The first gap-dependent $O(\frac{N^2 \log NT}{\Delta})$ regret bound was shown in (Agrawal et al., 2019). All bounds we mentioned are regret bounds, since no previous literature discussed the PAC exploration problem. Although there was a reduction from the MNL-bandit to the multi-armed bandit (Agrawal et al., 2016, 2019), that reduction involves exponentially many arms and thus does not give good gap-dependent bounds.

There is a line of work in multi-armed bandits and combinatorial multi-armed bandits that studies the subset selection problem, where an algorithm learns a subset to maximize a reward function. Near-optimal fully gap-dependent regret and sample complexity bounds have been proved in those settings (Bubeck et al., 2012; Chen et al., 2017, 2016a, 2013, 2014, 2016b; Rejwan and Mansour, 2020). While MNL-bandit can be seen as a subset selection problem, the major difference is that the feedback in our setting is much more *limited.* In their settings, by selecting a subset (some called "super arm"), the player gets feedback from all arms in the subset. In our setting, the seller can obtain feedback only from the purchased item in the subset.

Some recent paper studies the MNL choice model under the dueling bandits framework (Chen et al., 2018; Saha and Gopalan, 2019), proving fully gap-dependent bounds. Their setting can be seen as a simplification of ours through assuming all items have the same reward $r_i \equiv 1$ and removing the "no purchase" decision. In their setting, the optimal assortment simply consists of the items with largest model parameters, so their focus is to learn the order of the parameter. In our setting, the optimal assortment depends on the model parameters in a more complicated manner, so we need to learn the parameters themselves.

2 Preliminaries

Notations We define $\alpha \wedge \beta = \min\{\alpha, \beta\}, \alpha \vee \beta = \max\{\alpha, \beta\}$. For any two expressions α and β, if there exists a constant $C > 0$ in digits such that $\alpha \le C \cdot \beta$, we write $\alpha \lesssim \beta$. If $\beta \lesssim \alpha$, we write $\alpha \gtrsim \beta$. If $\alpha \lesssim \beta$ and $\beta \lesssim \alpha$, we write $\alpha \asymp \beta$. The notions $\widetilde{O}$ and $\widetilde{\Theta}$ suppress the logarithmic terms and the relatively small gap-independent terms in sample complexity bounds, and the logarithmic terms in regret bounds. We use both $|A|$ and $\#A$ to denote the size of a set A. For two disjoint sets A, B that $A \cap B = \emptyset$, we use $A \sqcup B = A \cup B$ to denote their union.

Settings and Problems We define the MNL choice model with parameter v_i for each item $i \in [N] \cup \{0\}$, where item $i = 0$ stands for the "no purchase" decision. In this model, when the seller offers an assortment S, the buyer purchases item $i \in S \cup \{0\}$ with probability $P_S^{\boldsymbol{v}}(i) = \frac{v_i}{v_0 + \sum_{j \in S} v_j}$. Note that "no purchase" decision is always available to the buyer.

We define an MNL-bandit instance as a quadruple $\mathcal{I} = (N, K, \boldsymbol{r}, \boldsymbol{v})$, where the reward of item $i \in [N]$ is $r_i \in [0, 1]$ and its MNL model parameter is $v_i \in [0, 1]$. The seller knows $N, K, \boldsymbol{r}$, but does not know $\boldsymbol{v}$. At each time step $t = 1, 2, \ldots$, the seller offers an assortment $S_t \subseteq [N]$ under the capacity constraint $|S_t| \le K$ and receive the buyer's purchase decision $c_t \sim P_{S_t}^{\boldsymbol{v}}$. As a result, the seller's revenue is $R(S, \boldsymbol{v}) = \mathbb{E}_{i \sim P_S^{\boldsymbol{v}}}[r_i] = \frac{\sum_{i \in S} v_i r_i}{1 + \sum_{i \in S} v_i}$, where we assume that $r_0 = 0$. We adopt a common convention that $v_0 = 1$, which means the "no purchase" decision is the most frequent outcome (Agrawal et al., 2016, 2017, 2019). We use $S^* = \arg\max_{S \subseteq [N]: |S| \le K} R(S, \boldsymbol{v})$ to denote the optimal assortment and $\theta^* = R(S^*, \boldsymbol{v})$ to denote its revenue. Next we formally define the suboptimality gap for each item.

Definition 1 (Suboptimality gap)**.** For every item $i \in [N]$, we define its suboptimality gap as

$$\Delta_i = \begin{cases} R(S^*, \boldsymbol{v}) - \max\limits_{|S| \le K: i \in S} R(S, \boldsymbol{v}), & i \notin S^* \\ R(S^*, \boldsymbol{v}) - \max\limits_{|S| \le K: i \notin S} R(S, \boldsymbol{v}), & i \in S^*. \end{cases}$$

Our definition has the same form as the suboptimality gaps in other bandits problems focusing on subset selection (Bubeck et al., 2013; Chen et al., 2014). Note that the bounds usually inversely depend on the gaps, so our definition matches the intuition that items with small Δ_i are more difficult to be separated from the optimal assortment and thus lead to worse bounds. We make the following uniqueness assumption, which is typically assumed when studying gap-dependent bounds in bandits literature (Bubeck et al., 2013; Chen et al., 2017; Karnin et al., 2013).

Our definition is related to the global gap Δ studied in previous literature (Rusmevichientong et al., 2010; Sauré and Zeevi, 2013; Agrawal et al., 2019), by that we have $\Delta_i \ge \Delta$ for every item i. We mention again that $\Delta = \theta^* - \max_{S \subseteq [N]: |S| \le K}\{R(S, \boldsymbol{v}) : R(S, \boldsymbol{v}) \neq \theta^*\}$ is the gap between the optimal and second-best assortments.

Assumption 1 (Uniqueness)**.** The optimal assortment S^* is unique.

Finally, we define the two problems we study. The first problem is defined in light of the PAC (probably approximately correct) learning framework and follows the definitions of the exploration problems in other bandits under the fixed-confidence setting (Jamieson and Nowak, 2014; Rejwan and Mansour, 2020). The second problem follows the regret definition in previous MNL-bandit literature (Agrawal et al., 2016, 2017, 2019; Chen et al., 2018; Chen and Wang, 2018).

Problem 1 (PAC Exploration)**.** An algorithm is (δ, ε)-PAC with sample complexity T, if it returns an assortment S that $\theta^* - R(S, \boldsymbol{v}) \le \varepsilon$ in T time steps with probability $1 - \delta$. If $\varepsilon = 0$, we say it is δ-PAC. The goal is to design δ-PAC and (δ, ε)-PAC algorithms with minimum sample complexity.

Problem 2 (Regret Minimization)**.** The goal is to design an algorithm that offers assortments over a known time horizon $T(\geq N)$ with minimum regret $\mathrm{Reg}_T = \sum_{t=1}^{T} R(S^*, \boldsymbol{v}) - \mathbb{E}[R(S_t, \boldsymbol{v})]$.

3 PAC Exploration

3.1 Algorithmic Framework with Fully Gap-Dependent Bounds

In this subsection, we introduce an algorithmic framework for which we can obtain fully gap-dependent sample complexity bounds and, as a direct application, present a δ-PAC algorithm with sample complexity $\widetilde{O}(\sum_{i=1}^{N} K^2 \Delta_i^{-2})$. Our framework is based on relating the MNL-bandit problem to the positive top-K item identification (PTOP-K) problem via the notion of advantage score.

Relate MNL-bandit to PTOP-K We first describe the goal of the PTOP-K problem, then relate it with the MNL-bandit problem. To describe the goal, we define the following function $\mathcal{F}$. Given a capacity constraint M and a set W where each $i \in W$ has a score $\xi_i \in \mathbb{R}$, we denote the subset containing elements with positive and top-M scores as

$$\begin{aligned} &\mathcal{F}(W, M, \xi) = \{i \in W : \xi_i > 0\} \\ &\cap \{i \in W : \xi_i \text{ is among the top } M \text{ of } \{\xi_j\}_{j \in W}\}. \end{aligned} \quad (1)$$

The goal of the PTOP-K problem is to identify the subset $\mathcal{F}([N], K, \boldsymbol{u})$ of items, where u_i is the specially constructed score defined with respect to each item $i \in [N]$ as follows.

Definition 2 (Advantage Score)**.** We define the *advantage score* of item i as $u_i = v_i(r_i - \theta^*)$.

Now we relate the MNL-bandit problem to the PTOP-K problem by the following proposition, which states that they share the *same* goal of identifying the optimal assortment $S^* \subseteq [N]$.

Proposition 1 (Relate to PTOP-K)**.** *$S^* = \mathcal{F}([N], K, \boldsymbol{u})$ and $\theta^* = \sum_{i \in S^*} u_i$.*

We defer the proof to Appendix B.1, which uses a classical method in optimization theory called fractional programming (Dinkelbach, 1967; Rusmevichientong et al., 2010). Our proposition indicates that pairwise partial orders and early decision rules in MNL-bandit are the same as those in the PTOP-K problem, which is very similar to the top-K arm identification problem. Since algorithms with fully gap-dependent bounds are well-studied in the top-K arm problem, we can obtain such bounds for the MNL-bandit problem by combining those algorithms with our relation proposition.

However, two issues arise when combining them. First, the gap-dependent bounds for the top-K problems use the gaps of scores, not our suboptimality gap for items. Second, estimating the advantage scores u_i is much more difficult than estimating the means of arms in the top-K arm problem, because the definition of u_i involves the optimal revenue θ^*, which could depend on items *other* than i. In contrast, the mean of each arm only depends on the arm itself. The first issue can be resolved by Lemma B.1, which shows that our gap is always smaller and thus bounds for top-K problems translate to our MNL-bandit setting. The second issue is difficult to resolve. In Lemma 3.2, we will show that a naive solution could lead to two extra K factors in the guarantee.

Algorithmic Framework Let us assume a procedure EST that estimates the advantage score. We introduce our algorithmic framework $\mathsf{SAR\text{-}MNL}$ (Algorithm 1). We summarize below its sample complexity guarantee and defer the proof to Appendix B.2.

Lemma 3.1 ($\mathsf{SAR\text{-}MNL}$)**.** *Assume with probability $1-\delta^{(k)}$, EST (a) returns within $C_{\mathsf{EST}} \cdot \frac{|B^{(k-1)}| \log(N/\delta^{(k)})}{\epsilon_k^2}$ time steps in phase k for a numeric constant C_{EST}, and (b) $u_i \in [\check{\xi}_i, \hat{\xi}_i]$ and $\hat{\xi}_i - \check{\xi}_i \leq \frac{\epsilon_k}{2}$ for every $i \in B^{(k-1)}$. Then $\mathsf{SAR\text{-}MNL}$ with EST is δ-PAC with sample complexity $C_{\mathsf{EST}} \cdot O(\sum_{i \in [N]} \frac{\log N + \log \delta^{-1} + \log\log \Delta_i^{-1}}{\Delta_i^2})$.*

Our framework is similar to the successive accept-reject algorithms used to solve the top-K arm identification problem (Chen et al., 2017; Rejwan and Mansour, 2020; Bubeck et al., 2013). The idea is to alternate in phases between estimate the scores of pending items and accept-reject them. For each phase k, accepted items are stored in $A^{(k)}$ and rejected items are in $[N] \setminus (A^{(k)} \cup B^{(k)})$. In phase k, after building up the confidence intervals of scores u_i at Line 5, the algorithm accepts-rejects items by some rules. Since

$$\begin{aligned} S^* = \mathcal{F}([N], K, \boldsymbol{u}) &= \mathcal{F}(A^{(k-1)} \cup B^{(k-1)}, K, \boldsymbol{u}) \\ &= A^{(k-1)} \sqcup \mathcal{F}(B^{(k-1)}, M, \boldsymbol{u}), \end{aligned}$$

where $M = \min\{K - |A^{(k-1)}|, |B^{(k-1)}|\}$, the rules are to accept items in $B^{(k-1)}$ with positive and top-M scores and reject those with negative or not top-M scores. In the framework, Line 6 handles the sign rule and Lines 7-9 handle the top-M rule.

Estimation Procedure We present a naive estimation procedure $\mathsf{EST\text{-}NAIVE}$ (Algorithm 2). The procedure estimates the score $u_i = v_i(r_i - \theta^*)$ by estimating both v_i and θ^*. Line 3 is because the optimal revenue is

Algorithm 1: $\mathsf{SAR\text{-}MNL}(\delta)$: Successive Accept-Reject Framework for MNL-bandit

```
1  $A^{(0)} = \emptyset, B^{(0)} = [N]$;                      ▷ $A^{(k)}, B^{(k)}$ store accepted, pending items
2  for $k \leftarrow 1, 2, \ldots$ do          ▷ maintain $A^{(k)} \subseteq S^* \subseteq A^{(k)} \sqcup B^{(k)}$ for each phase $k$
3      $\epsilon_k = 2^{-k}, \delta^{(k)} = \frac{\delta}{3k^2}, M = M^{(k-1)} = \min\{K - A^{(k-1)}, |B^{(k-1)}|\}$;
4      if $M = 0$ then return $A^{(k-1)}$;
5      $\{\check{\xi}_i, \hat{\xi}_i\}_{i \in B^{(k-1)}} \leftarrow \mathsf{EST}(A^{(k-1)}, B^{(k-1)}, \delta^{(k)}, \frac{\epsilon_k}{2})$;          ▷ estimate the scores
6      $B_{\text{acc}} \leftarrow \{b \in B^{(k-1)} : \check{\xi}_b > 0\}, B_{\text{rej}} \leftarrow \{b \in B^{(k-1)} : \hat{\xi}_b < 0\}$;
7      if $|B^{(k-1)}| > M$ then          ▷ if $|B^{(k-1)}| \leq M$ then all items have top-$M$ scores
8          $\alpha \leftarrow M$-th largest value of $\{\check{\xi}_i\}_{i \in B^{(k-1)}}, \beta \leftarrow (M+1)$-th largest value of $\{\hat{\xi}_i\}_{i \in B^{(k-1)}}$;
9          $B_{\text{acc}} \leftarrow B_{\text{acc}} \cap \{b \in B^{(k-1)} : \check{\xi}_b > \beta\}, B_{\text{rej}} \leftarrow B_{\text{rej}} \cup \{b \in B^{(k-1)} : \hat{\xi}_b < \alpha\}$;
10     $A^{(k)} \leftarrow A^{(k-1)} \cup B_{\text{acc}}, B^{(k)} \leftarrow B^{(k-1)} \setminus (B_{\text{acc}} \cup B_{\text{rej}})$;          ▷ accepts-rejects the items
```

Algorithm 2: $\mathsf{EST\text{-}NAIVE}(A, B, \delta, \epsilon)$: Naive Estimation of u_i for $i \in B$

```
1  for $i \in A \sqcup B$ do Keep offering $\{i\}$ until "no purchase" occurs for $K\tau$ times;          ▷ $\tau = \widetilde{O}(\frac{1}{\epsilon^2})$
2  $\forall i \in A \sqcup B$: Compute the confidence intervals $v_i \in [\check{v}_i, \hat{v}_i]$;          ▷ formulas of $\check{v}_i, \hat{v}_i$ in Appendix B.3
3  Compute $\check{\theta} = \max_{S \subseteq A \sqcup B, |S| \leq K} R(S, \check{v}), \hat{\theta} = \max_{S \subseteq A \sqcup B, |S| \leq K} R(S, \hat{v})$;
4  Compute $\check{\xi}_i = (\check{v}_i(r_i - \hat{\theta})) \wedge (\hat{v}_i(r_i - \hat{\theta})), \hat{\xi}_i = (\check{v}_i(r_i - \check{\theta})) \vee (\hat{v}_i(r_i - \check{\theta}))$, return $\{\check{\xi}_i, \hat{\xi}_i\}$;
```

a monotonic function of the model parameters (Agrawal et al., 2016, 2019). (We emphasize that the revenue is not monotonic in general.) The maximization step at Line 3 can be solve efficiently (Rusmevichientong et al., 2010). Line 4 is based on $0 \leq v_i \leq 1$ and $|r_i - \theta^*| \leq 1$. The procedure leads to the following guarantee.

Lemma 3.2. *SAR-MNL with EST-NAIVE is δ-PAC with sample complexity $\widetilde{O}(\sum_{i \in [N]} \frac{K^2}{\Delta_i^2})$.*

We sketch the proof here and complete it in Appendix B.3. Note that the procedure offers each item in the set $A \cup B$ for $\widetilde{O}(\frac{K}{\epsilon^2})$ time steps, so it achieves $C_{\mathsf{EST}} \asymp K \cdot \frac{|A \cup B|}{|B|}$ for Lemma 3.1. In the worst case, we have $|A \cup B| \asymp K|B|$, so we have $C_{\mathsf{EST}} \leq K^2$, which implies Lemma 3.2.

We inspect the sources of two K factors in Lemma 3.2. The first is because we use that $\frac{|A \cup B|}{|B|} \leq K$, which is ultimately because the naive procedure needs to estimate v_i for $i \in A$. The second is because the procedure needs to estimate each v_i to a fixed accuracy $\frac{\epsilon}{K}$ in order to estimate θ^*. One may ask why the procedure only offers singletons at Line 1 and why the accuracy needs to be $\frac{\epsilon}{K}$ instead of $O(\epsilon)$. Interestingly, we show in Appendix B.3 that both could be optimal for some instance.

3.2 Reduced Revenue Function and Generalized Epoch-based Offering

To eliminate the first K factor in $\mathsf{EST\text{-}NAIVE}$, we introduce a reduced revenue function and a generalized epoch-based offering procedure in this subsection. The reduced revenue function enables us to estimate θ^* without estimating the parameters v_i for $i \in A$. The generalized procedure is used to estimate the parameters in the reduced revenue function.

Reduced Revenue Function We note that $\mathsf{SAR\text{-}MNL}$ invokes EST with $A \subseteq S^* \subseteq A \sqcup B$. However, $\mathsf{EST\text{-}NAIVE}$ only uses $S^* \subseteq A \sqcup B$. Now we exploit $A \subseteq S^*$. Let $M = \min\{K - |A|, |B|\}$. For an assortment S satisfying $A \subseteq S$, we rewrite its revenue as

$$
\begin{aligned}
R(S, \boldsymbol{v}) &= \frac{\sum_{i \in S} v_i r_i}{1 + \sum_{i \in S} v_i} \\
&= \frac{\zeta + \sum_{i \in S \setminus A} \nu_i r_i}{1 + \sum_{i \in S \setminus A} \nu_i} = R(S \setminus A, \nu, \zeta),
\end{aligned}
$$

where we define $\zeta = \frac{\sum_{j \in A} v_j r_j}{1 + \sum_{j \in A} v_j} = R(A, \boldsymbol{v})$ and $\nu_i = \frac{v_i}{1 + \sum_{j \in A} v_j}$ for $i \notin A$. Note that if we use $R(S, \boldsymbol{v})$ to compute the revenue of S, we need $|S|$ parameters (v_i for each $i \in S$). In contrast, if we use $R(S \setminus A, \nu, \zeta)$, we only need $(|S \setminus A| + 1)$ parameters (ν_i for $i \in S \setminus A$ and ζ). Thus we refer to the function $R(S \setminus A, \nu, \zeta)$ as the *reduced revenue function*, since it reduces the number of required parameters. We note that $\theta^* = \max_{S_0 \subseteq B : |S_0| \leq M} R(S_0, \nu, \zeta)$ and Lemma C.1 further shows that the maximization used by θ^* is still monotonic in the parameters ν and ζ. Therefore, given the confidence intervals $\zeta \in [\check{\zeta}, \hat{\zeta}]$ and $\nu_i \in [\check{\nu}_i, \hat{\nu}_i]$, we have the confidence interval $\theta^* \in [\check{\theta}, \hat{\theta}]$, where

$$\begin{aligned} \check{\theta} &= \max_{S \subseteq B : |S| \leq M} R(S, \check{\nu}, \check{\zeta}), \\ \hat{\theta} &= \max_{S \subseteq B : |S| \leq M} R(S, \hat{\nu}, \hat{\zeta}). \end{aligned} \tag{2}$$

Generalized Epoch-based Offering With Eq. (2) in hand, it remains how to estimate ζ and ν_i. Note that ν_i is a *ratio* of two *unknown* quantities v_i and $(1 + \sum_{j \in A} v_j)$, so it is virtually impossible to estimate ν_i by separately estimating the two quantities. The generalized epoch-based offering procedure (Algorithm 3) allows us to *directly* estimate the ratio ν_i. It generalizes those used in (Agrawal et al., 2016, 2017, 2019) by introducing a stopping set Z, which is fixed as $Z = \emptyset$ in the original version. When we set $Z = A$, we can use the procedure to estimate parameters ν_i and also ζ.

Proposition 2 (Generalized Epoch-based Offering)**.** *After Explore(S), we have*

(a) $z \in [0, 1]$ is an independent bounded random variable with mean ζ;

(b) x_i is an independent geometric random variable with mean ν_i for every item $i \in S$;

(c) $(E_\ell - 1)$ is an independent geometric random variable with mean $\sum_{i \in S} \nu_i$.

We defer the proof to Appendix C.1. Statement (c) can give the sample complexity bound when using the procedure, as in Lemma C.2. Combined with corresponding concentration inequalities in Appendix A, statements (a)(b) can give the confidence intervals $\zeta \in [\check{\zeta}, \hat{\zeta}]$ and $\nu_i \in [\check{\nu}_i, \hat{\nu}_i]$, where

$$\begin{aligned} \check{\zeta} &= 0 \vee (\bar{\zeta} - \sqrt{\frac{\log(2/\delta)}{2T_Z}}), \\ \hat{\zeta} &= 1 \wedge (\bar{\zeta} + \sqrt{\frac{\log(2/\delta)}{2T_Z}}), \end{aligned} \tag{3}$$

and

$$\begin{aligned} \check{\nu}_i = 0 \vee (\bar{\nu}_i - \sigma(\nu_i)), \hat{\nu}_i = 1 \wedge (\bar{\nu}_i + \sigma(\nu_i)), \\ \sigma(\nu_i) = \sqrt{\frac{48\bar{\nu}_i \log(2/\delta)}{T_i}} + \frac{48 \log(2/\delta)}{T_i}. \end{aligned} \tag{4}$$

Finally, we define the reduced score $\xi_i = \nu_i(r_i - \theta^*)$ and its confidence interval $\xi_i \in [\check{\xi}_i, \hat{\xi}_i]$, where

$$\begin{aligned} \check{\xi}_i &= (\check{\nu}_i(r_i - \hat{\theta})) \wedge (\hat{\nu}_i(r_i - \hat{\theta})), \\ \hat{\xi}_i &= (\check{\nu}_i(r_i - \check{\theta})) \vee (\hat{\nu}_i(r_i - \check{\theta})). \end{aligned} \tag{5}$$

Now we assume the procedure EST used by SAR-MNL estimates ξ_i instead of u_i. In Appendix C.2, we show that the same sample complexity bound as Lemma 3.1 still hold.

To demonstrate the technique in this subsection, we show in Appendix C.3 that we can achieve $C_{\mathsf{EST}} = O(K)$ using the generalized epoch-based offering, which implies an $\widetilde{O}(\sum_{i=1}^N K \Delta_i^{-2})$ sample complexity bound and eliminates an extra K factor in EST-NAIVE.

3.3 Layer-based Adaptive Estimation

To eliminate another extra K factor in EST-NAIVE, we present a layer-based adaptive estimation procedure based on a detailed error analysis of the reduced revenue function and the tail bounds in Eqs. (3)(4). The error analysis in Appendix D.1 suggest we offer each item $b \in B$ for $T_b \gtrsim T'_b \tau$ epochs, where we define $T'_b = (\frac{1}{\nu_b} \wedge M)$ and $\tau = \widetilde{O}(\epsilon^{-2})$. Next we show how to accomplish this offering task in $O(|B|\tau)$ time steps, which gives $C_{\mathsf{EST}} = O(1)$ in Lemma 3.1 and eliminates the extra K factor. To better convey our idea, we first consider an ideal but unrealistic case where the exact values of ν_b are given. We divide the set B into $m = \lceil \log_2 M \rceil$ layers:

$$\begin{aligned} B_i = \{b \in B : \nu_b \in (2^{-(i+1)}, 2^{-i}]\} \quad (0 \leq i < m), \\ B_m = \{b \in B : \nu_b \in [0, 2^{-i}]\}. \end{aligned} \tag{6}$$

Let $d_i = 2^i$ for $i < m$ and $d_m = M$. The key observation is that items form the same layer have similar ν_b and need to be explored for a similar number of epochs (up to a factor $\kappa = 2$): we have $T'_b \leq \kappa d_i$ and $\nu_b \leq \frac{\kappa}{d_i}$ for $b \in B_i$. We note that $d_i \leq M$. Therefore, we can divide each layer B_i into groups of size d_i. Since we have $\nu_b \lesssim \frac{1}{d_i}$ for $b \in B_i$, by Proposition 2, the expected epoch length of explore a group is $d_i \cdot \frac{1}{d_i} = O(1)$. So if we explore each group for $d_i \tau$ epochs, in expectation it costs us $d_i \tau$ time steps, which is τ time steps per item. Since we have $|B|$ items, we can accomplish the offering task within $O(|B|\tau)$ time steps, which gives the desired $C_{\mathsf{EST}} = O(1)$.

We note that the exact values of ν_b are not necessary, since we only need the layer of each item. In fact, we can surrogate by using a *rough* estimation of ν_b satisfying $\tilde{\nu}_b \in [\frac{\nu_b}{2}, 2\nu_b \vee \frac{1}{M}]$ in place of ν_b in Eq. (6) to divide the layers. Our key observation is still satisfied, though perhaps with a different factor $\kappa = 4$. This fact is utilized by our layer-based adaptive estimation procedure EST-ADAPTIVE (Algorithm 5). It first uses the results from the procedure EST-ROUGH (Algorithm 4) to build up $\tilde{\nu}_b$ at Line 1, based on which it divides the layers at Line 3. We summarize the guarantees of our two procedures in the following two lemmas and defer their proofs to Appendices D.2 and D.3.

Lemma 3.3 (EST-ROUGH)**.** *With probability* $1 - \delta_0$, *(a) EST-ROUGH ends in $O(NK \log \frac{N}{\delta_0})$ time steps and*

Algorithm 3: Explore(S): Generalized Epoch-based Offering with Stopping Set $Z(Z \cap S = \emptyset)$

```
Initialize: $z \leftarrow 0, \ell \leftarrow \ell + 1, E_\ell = 0, \forall i \in S : x_i \leftarrow 0$;
while TRUE do                                          ▷ Epoch: time steps used in the while-loop
    $t \leftarrow t + 1, E_\ell \leftarrow E_\ell + 1$;                              ▷ $E_\ell$ is the length of epoch $\ell$
    Offer assortment $S_t = Z \cup S$, observe purchase decision $c_t$;
    if $c_t \in Z \cup \{0\}$ then $z \leftarrow r_{c_t}$, break; else $x_{c_t} \leftarrow x_{c_t} + 1$;
$n_Z \leftarrow n_Z + z, T_Z \leftarrow T_Z + 1, \bar{\zeta} \leftarrow \frac{n_Z}{T_Z}, \forall i \in S : n_i \leftarrow n_i + x_i, T_i \leftarrow T_i + 1, \bar{\nu}_i \leftarrow \frac{n_i}{T_i}$;
```

Algorithm 4: EST-ROUGH(δ_0): Rough Estimation of v_i for $i \in [N]$

```
$C_0 = 196, \delta = \frac{\delta_0}{17N}, \tau = 4KC_0 \log(2/\delta), Z \leftarrow \emptyset, \forall i \in [N] : n_i = T_i = 0$;
for $i \in [N]$ do Explore($\{i\}$) for $\tau$ epochs;
$\forall i \in [N]$ : compute $\hat{\nu}_i$ by Eq. (4), let $\tilde{v}_i \leftarrow \hat{\nu}_i$, return $\{\tilde{v}_i\}_{i \in [N]}$;
```

$\tilde{v}_i \in [v_i, 2v_i \vee \frac{1}{K}]$ for every $i \in [N]$; (b) In this case, for every set $S \subseteq [N]$ with $|S| \le K$, we have $\tilde{\nu}_i \in [\frac{\nu_i}{2}, 2\nu_i \vee \frac{1}{K}]$, where $\nu_i = \frac{v_i}{\sum_{i \in S} v_i}$ and $\tilde{\nu}_i = \frac{\tilde{v}_i}{1 + \sum_{i \in S} \tilde{v}_i}$.

Lemma 3.4 (EST-ADAPTIVE)**.** *Assume $A \subseteq S^* \subseteq A \sqcup B$. With probability $1 - \delta_0$, (a) EST-ADAPTIVE returns in $|B|\tau$ time steps; (b) $\xi_i = \frac{u_i}{1 + \sum_{j \in A} v_j} \in [\check{\xi}_i, \hat{\xi}_i]$ and $\hat{\xi}_i - \check{\xi}_i \le \epsilon$ for $i \in B$.*

3.4 Putting Everything Together

We combine all our techniques to design a δ-PAC algorithm: we first invoke EST-ROUGH($\frac{\delta}{2}$), then invoke SAR-MNL($\frac{\delta}{2}$) with EST-ADAPTIVE. We highlight that EST-ROUGH is invoked only *once* while it can help divide layers for *every* phase, as shown by the statement (b) in Lemma 3.3.

Our δ-PAC algorithm becomes (δ, ε)-PAC if we terminate it at the phase k satisfying $\epsilon_k \lesssim \varepsilon$ and let it return the assortment corresponding to $\hat{\theta}$. We summarize the results in the below theorems.

Theorem 1. *There is a δ-PAC algorithm with sample complexity*

$$O(NK \log \frac{N}{\delta} + \sum_{i \in [N]} \Delta_i^{-2} (\log \frac{N}{\delta} + \log \log \Delta_i^{-1})).$$

Theorem 2. *There is a (δ, ε)-PAC algorithm with sample complexity*

$$O(NK \log \frac{N}{\delta} + \sum_{i \in [N]} (\Delta_i^\varepsilon)^{-2} (\log \frac{N}{\delta} + \log \log (\Delta_i^\varepsilon)^{-1})),$$

where $\Delta_i^\varepsilon = \Delta_i \vee \varepsilon$.

The proofs are deferred to Appendix D.4. Both bounds have a gap-independent $O(NK \log \frac{N}{\delta})$ term due to EST-ROUGH, which is arguably much smaller than the gap-dependent $\widetilde{O}(\sum_{i=1}^N \Delta_i^{-2})$ term. Theorem 2 translates to an $\widetilde{O}(N\varepsilon^{-2})$ gap-independent sample complexity bound, which matches a corollary of previous $\widetilde{\Theta}(\sqrt{NT})$ minimax regret bound in (Agrawal et al., 2016, 2019, 2017; Chen and Wang, 2018) as follows. Suppose we run an algorithm with $\widetilde{O}(\sqrt{NT})$ regret bound for $T = O(N\varepsilon^{-2})$ time steps and uniformly choose an assortment S from $\{S_1, \ldots, S_T\}$. In expectation, we have $\mathbb{E}[\theta^* - R(S, \boldsymbol{v})] = \widetilde{O}(\sqrt{NT})/T = \varepsilon$ and thus we get an algorithm with $O(T) = O(N\varepsilon^{-2})$ sample complexity.

4 Regret Minimization

In this section, we present fully gap-dependent regret bounds for Problem 2. Our algorithm is to invoke SAR-MNL($\frac{1}{T}$) with our low-regret estimation procedure EST-REG (Algorithm 6). This algorithm satisfies the following theorem, whose proof is deferred to Appendix E.

Theorem 3 (Regret)**.** *There is an algorithm that achieves regret bound $O(\sum_{i \in [N] \setminus S^*} \frac{K \log(NT)}{\Delta_i})$.*

Our procedure achieves low regret by fixing the accepted set A and offering full assortments. The first fixing idea has been exploited in (Rejwan and Mansour, 2020). The second idea is our novel technique, without which we could have a regret bound $O(\sum_{i=1}^N \frac{K \log(NT)}{\Delta_i})$ that depends on S^*.

5 Discussion on Lower Bounds

An interesting question is whether we can prove fully gap-dependent lower bounds in MNL-bandit for Problems 1 and 2. To begin with, we prove an $\Omega(\sum_{i \notin S^*} \frac{\log T}{K\Delta_i})$ regret lower bound in Appendix F, which matches our regret upper bound when $K = 1$. Besides, for Problem 1, our sample complexity upper bound matches the gap-independent sample complexity

Algorithm 5: EST-ADAPTIVE$(A, B, \delta_0, \epsilon)$: Layer-based Adaptive Estimation of ξ_i for $i \in B$

```
$C_0 = 196, C_2 = 1024, \delta = \frac{\delta_0}{15N}, M = \min\{K - |A|, |B|\}, \tau = \frac{C_2 C_0 \log(2/\delta)}{\epsilon^2}, Z \leftarrow A, n_Z = T_Z = 0,$
$\forall i \in B : n_i = T_i = 0, \tilde{\nu}_i = \frac{\tilde{v}_i}{1+\sum_{j \in Z} \tilde{v}_j}$;        ▷ assume $\tilde{v}_i$ has been computed by EST-ROUGH
for $i \leftarrow 0, 1, \ldots, m = \lceil \log_2 M \rceil$ do
    Compute $B_i$ by Eq. (6) using $\tilde{\nu}_b$ in place of $\nu_b$, let $d_i = 2^i \vee M$;
    Divide $B_i = B_{i,1} \sqcup \cdots \sqcup B_{i,c_i}$ so that $|B_{i,1}| = \cdots = |B_{i,c_i-1}| = d_i$ and $|B_{i,c_i}| \leq d_i$;
    $\forall j \in [c_i]$ : Explore$(B_{i,j})$ for $d_i \tau$ epochs;
Compute $\check{\zeta}, \hat{\zeta}, \check{\nu}_i, \hat{\nu}_i, \check{\theta}, \hat{\theta}, \check{\xi}_i, \hat{\xi}_i$ by Eqs. (3) (4) (2) (5) for $i \in B$, return $\{\check{\xi}_i, \hat{\xi}_i\}_{i \in B}$;
```

Algorithm 6: EST-REG$(A, B, \delta_0, \epsilon)$: Low-Regret Estimation of u_i

```
$C_0 = 196, C_2 = 1024, \delta = \frac{\delta_0}{13N}, M = \min\{K - |A|, |B|\}, \tau = \frac{C_2 C_0 \log(2/\delta)}{\epsilon_k^2}$, $Z = \emptyset, \forall i \in A \cup B : n_i = T_i = 0$;
Create $m = \left\lceil \frac{|B|}{M} \right\rceil$ sets $B'_1, \ldots, B'_m \subseteq B$ so that $B \subseteq B'_1 \cup \cdots \cup B'_m$ and $|B'_i| = M$;
$\forall i \in [m]$: Explore$(A \cup B'_i)$ for $K \cdot \tau$ epochs;        ▷ we offer full assortments with size $(M + |A|)$
Let $\check{\zeta} = \hat{\zeta} = 0$, Compute $\check{\nu}_i, \hat{\nu}_i, \check{\theta}, \hat{\theta}, \check{\xi}_i, \hat{\xi}_i$ by Eqs. (4) (2) (5) for $i \in A \cup B$, return $\{\check{\xi}_i, \hat{\xi}_i\}_{i \in B}$;
```

bound translated from the previous minimax-optimal regret bound, as specified earlier in Section 3.4.

Furthermore, we discuss about the difficulties in studying gap-dependent lower bounds in MNL-bandit. Given an arbitrary gap sequence $\{\Delta_i\}$, it is not trivial to realize the gaps, where by realizing we mean to find an MNL-bandit instance with such gaps. The fractional revenue function made it hard to determine if a given gap sequence could correspond to an instance and construct such instance when exists. The hardness in constructing instance from gap makes it difficult to prove lower bounds by following the canonical change-one-arm lower bound argument for multi-armed bandits. We believe that perhaps for these reasons, previous work also did not prove gap-dependent lower bounds (even in the term of the global gap Δ, which is much weaker than our gap definition Δ_i, as discussed in Sections 1 and 2) when studying gap-dependent bounds in MNL-bandit.

6 Conclusion

In this paper, we develop multiple techniques to prove the fully gap-dependent sample complexity and regret bounds for the MNL-bandit problems. We leave it a further direction to prove tighter lower bounds for the problems. For the upper bound, a significant question is whether we can remove the K factor in the regret bound. It would be worthwhile to prove a fully gap-dependent regret bound for the original upper confidence bound algorithm in (Agrawal et al., 2016, 2019). An interesting direction is whether the gap-independent term in our sample complexity bound can be reduced or even totally be removed.

Acknowledgments

Jiaqi Yang would like to thank Yuan Zhou and the anonymous reviewers for the comments and suggestions.

References

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2016). A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 599–600.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2017). Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pages 76–78.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019). Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Bubeck, S., Wang, T., and Viswanathan, N. (2013). Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265.

Chen, J., Chen, X., Zhang, Q., and Zhou, Y. (2017). Adaptive multiple-arm identification. In *Proceedings* *of the 34th International Conference on Machine Learning-Volume 70*, pages 722–730. JMLR.org.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016a). Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1659–1667.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016b). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778.

Chen, X., Li, Y., and Mao, J. (2018). 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*, pages 2504–2522. SIAM.

Chen, X. and Wang, Y. (2018). A note on a tight lower bound for capacitated mnl-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management science*, 13(7):492–498.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Jamieson, K. and Nowak, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Janson, S. (2018). Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135:1–6.

Jin, Y., Li, Y., Wang, Y., and Zhou, Y. (2019). On asymptotically tight tail bounds for sums of geometric and exponential random variables. *arXiv preprint arXiv:1902.02852*.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246.

Kök, A. G. and Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6):1001–1021.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

Rejwan, I. and Mansour, Y. (2020). Top-k combinatorial bandits with full-bandit feedback. In *Algorithmic Learning Theory*, pages 752–776.

Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680.

Saha, A. and Gopalan, A. (2019). Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, pages 983–993.

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

Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2013). Preference elicitation for general random utility models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 596–605.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Wang, Y., Chen, X., and Zhou, Y. (2018). Near-optimal policies for dynamic multinomial logit assortment selection models. In *Advances in Neural Information Processing Systems*, pages 3101–3110.