Multinomial Logit Contextual Bandits

Min-hwan Oh [1] **Garud Iyengar** [1]

Abstract

We consider a dynamic assortment selection problem where the goal is to offer an assortment with cardinality constraint K from a set of N possible items. The sequence of assortments can be chosen as a function of the contextual information of items, and possibly users, and the goal is to maximize the expected cumulative rewards, or alternatively, minimize the expected regret. The distinguishing feature in our work is that feedback, i.e. the item chosen by the user, has a multinomial logistic distribution. We propose upper confidence interval based algorithms for this multinomial logit contextual bandit. The first algorithm is a simple and computationally more efficient method which achieves an $\tilde{O}(d\sqrt{T})$ regret over T rounds with d dimensional feature vectors. The second algorithm inspired by the work of (Li et al., 2017) achieves an $\tilde{O}(\sqrt{dT})$ with logarithmic dependence on N and increased computational complexity because of pruning processes.

1. Introduction

In many human-algorithm interactions, a learning agent (algorithm) makes sequential decisions and receives user (human) feedback for the choices it makes. The multi-armed bandit (Lattimore & Szepesvári, 2019) is a model for this sequential decision making with partial feedback. It is a classic reinforcement learning problem that exemplifies the exploration-exploitation tradeoff dilemma. This multi-armed bandit model has found applications to a very diverse set of problems such as sequential design of experiments including learning click-through rates in search engines, product recommendations in online retailing, movie suggestions on streaming services, etc. Often, feature information about options that the agent has or contextual information about the user is available. The contextual bandit extends the multi-armed bandit by making the decision conditional on this contextual/feature information. In many real world problems including the aforementioned examples, the agent offers multiple options to the user, rather than a single option as in traditional bandit action selection. Then the user chooses one of the offered options (or chooses none) and the agent receives a reward associated with the user feedback.

In this paper, we consider a dynamic assortment selection which is a combinatorial variant of the bandit problem. The goal is to offer a sequence of assortments of at most K items from a set of N possible items. The sequence can be chosen as a function of the contextual information of items, and possibly users in order to minimize the expected regret. The d-dimensional contextual information, or a set of feature vectors, is revealed at each round t, allowing the feature/contextual information of products to change over time. The feedback here is the particular item chosen by the user from the offered assortment. We assume that the item choice follows a multinomial logistic (MNL) distribution. This is a widely used model in dynamic assortment optimization literature (Caro & Gallien, 2007; Rusmevichientong et al., 2010; Sauré & Zeevi, 2013; Agrawal et al., 2017a;b; Chen & Wang, 2017).

The multinomial logit (MNL) contextual bandit is a multinomial generalization of (generalized) linear contextual bandits (Filippi et al., 2010; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Chu et al., 2011; Li et al., 2017) — if an assortment contains a single item, then the problem reduces to (generalized) linear bandits. This generalization is non-trivial since the MNL model cannot be expressed in the form of generalized linear model (Chen et al., 2018); hence, the results of Li et al. (2017) do not apply. Furthermore, in contrast to the standard contextual bandit problems, in the MNL contextual bandit, the item choice (feedback) is a function of the offered assortment. Thus the regret analysis is more complicated.

We propose the first UCB based algorithms for this MNL contextual bandit. We propose two different algorithms that trade-off computational complexity with achieved regret. The first algorithm is computationally efficient with a $\tilde{O}(d\sqrt{T})$ regret bound over T rounds. The second algorithm, which utilizes the technique adapted from the works

[1]Columbia University, New York, New York, USA. Correspondence to: Min-hwan Oh <m.oh@columbia.edu>.

Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36^{th} International Conference on Machine Learning, Long Beach, California, USA, 2019.

of Auer (2002) and Li et al. (2017), achieves the improved regret bound of $\tilde{O}(\sqrt{dT})$. However, this second algorithm now has logarithmic dependence on N and increased computational complexity because of assortment pruning. We also generalize a finite sample normality type confidence bound for the maximum likelihood estimator (MLE) of GLM (Li et al., 2017) to the MNL model.

2. Related Work

The multinomial logit model (MNL) (Plackett, 1975; McFadden, 1978; Luce, 2012) is one of the most widely used choice models for assortment selection problems. The problem of computing the optimal assortment (*static* assortment optimization problem), when the MNL parameters, i.e. user preferences, are known a priori, is well studied (Talluri & Van Ryzin, 2004; Davis et al., 2014; Désir et al., 2014). Our work belongs to the literature on *dynamic* assortment optimization. Caro & Gallien (2007) consider the setting where the demand for items in an assortment is independent. Rusmevichientong et al. (2010) and Sauré & Zeevi (2013) consider the problem of minimizing regret under the MNL choice model and present an "explore first then exploit later" approach. Rusmevichientong et al. (2010) showed $O(N^2 \log^2 T)$ regret bound, where N is the number of total candidate items. Sauré & Zeevi (2013) later improved the bound to $O(N \log T)$. However, these methods require a priori knowledge of "separability" between the true optimal assortment and the other sub-optimal alternatives.

More recent work by Agrawal et al. (2017a;b); Cheung & Simchi-Levi (2017); Chen & Wang (2017) also incorporated MNL models into dynamic assortment optimization and formulated the problem into an online regret minimization problem without requiring a priori knowledge on separability. Agrawal et al. (2017a) proposed UCB-style algorithm which shows $\tilde{O}(\sqrt{NT})$ regret bound (where $\tilde{O}$ suppresses logarithmic dependence on T, N, or K). Agrawal et al. (2017b) achieve the same order of the regret bound $\tilde{O}(\sqrt{NT})$ using Thompson sampling (Thompson, 1933) technique with improved empirical performance. Chen & Wang (2017) show a matching lower bound of $\Omega(\sqrt{NT})$. All of this previous work on MNL bandits assumes each item is associated with a unique parameter, i.e. one cannot learn across items. In our proposed MNL contextual bandits, the utility of item i at time t is of the form $x_{t,i}^\top \theta^*$ some fixed but unknown *utility parameter* θ^*; hence, we can learn across items. When the feature dimension $d \ll \sqrt{N}$, learning across items allows one to reduce the regret bound from $\tilde{O}(\sqrt{NT})$ to $\tilde{O}(d\sqrt{T})$. However, one cannot directly incorporate (time-varying) contextual information into the previous work (see, e.g. Agrawal et al. (2017a;b)) since these methods require that the same assortment be offered repeatedly for a random number of time periods until an outside choice is observed. Chen et al. (2018) establishes $\tilde{O}(d\sqrt{T})$ regret bound for the MNL contextual bandit similar to our first algorithm, UCB-MNL (Algorithm 1) with $\tilde{O}(d\sqrt{T})$ regret. However, our UCB-MNL still has a tighter logarithmic dependence and additive terms (and clearly our second algorithm, CB-MNL (Alogithm 2) has a much tighter bound with $\tilde{O}(\sqrt{dT})$ regret). There is a fundamental difference between Chen et al. (2018) and our UCB-MNL. Chen et al. (2018) enumerates the exponentially many (N choose K) assortments and builds confidence bounds for each of them. In contrast, UCB-MNL only builds confidence bounds for each of the N different items. Chen et al. (2018) do recognize the computational issue and propose an approximate optimization algorithm to somewhat remedy it; however, not completely. Consider the simple case where each item has unit revenue. In this case, assortment selection under UCB-MNL reduces to sorting items based upper-confidence bounds and the run time is independent of K, whereas Chen et al. (2018) still have to consider all the (N choose K) assortments.

Linear contextual bandits have been widely studied (Auer, 2002; Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Chu et al., 2011; Agrawal & Goyal, 2013). Filippi et al. (2010); Li et al. (2017) extend the linear contextual bandit to scalar, monotone, generalized linear bandit. Filippi et al. (2010) established $\tilde{O}(d\sqrt{T})$ regret bound. Li et al. (2017) improved the regret bound to $\tilde{O}(\sqrt{dT})$ by establishing a new finite-sample confidence bound for MLE in generalized linear models (GLM). However, these results do not apply directly to our problem, since the choise probability of an item in an assortment is non-linear and non-monotone in the utility parameter θ^*.

Notations

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

3. Problem Formulation

We formulate the problem of MNL contextual bandits as follows. Consider an option space containing N distinct items, indexed by $i \in [N]$. At time t, feature vectors $x_{t,i} \in$

$\mathbb{R}^d$ for every base item $i \in [N]$ are revealed to the agent. Each feature vector combines the information of the user and the corresponding item i. For example, suppose the user at time t is characterized by a feature vector u_t and the base item i has a feature vector $v_{t,i}$ (note that we allow feature vectors for an item and a user to change over time), then we can use $x_{t,i} = \text{vec}(u_t v_{t,i}^\top)$, the vectorized outer-product of u_t and $v_{t,i}$, as the combined feature vector of item i a at time t. If u_t is not available, we can use product dependent features only $x_{t,i} = v_{t,i}$. Given this contextual information, at every round t, the agent selects an assortment $S_t \subset \mathcal{S}$ and observes the user purchase decision $c_t \in S_t \cup \{0\}$, where $\{0\}$ denotes "outside option" which means the user did not choose any item offered in S_t. This selection is given by a multinomial logit (MNL) choice model. Under this model, the probability that a user chooses item $i \in S_t$ is given by,

$$p_{t,i}(S_t, \theta^*) = \begin{cases} \dfrac{\exp\{x_{t,i}^\top \theta^*\}}{1 + \sum_{j \in S_t} \exp\{x_{t,j}^\top \theta^*\}}, & \text{if } i \in S_t \\ \dfrac{1}{1 + \sum_{j \in S_t} \exp\{x_{t,j}^\top \theta^*\}}, & \text{if } i = 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\theta^* \in \mathbb{R}^d$ is an unknown time-invariant parameter. We first make the following assumptions:

Assumption 1. *z Each feature vector $x_{t,i}$ is drawn i.i.d. from an unknown distribution ν, with $\|x_{t,i}\| \leq 1$ all t, i and there exists a constant $\sigma_0 > 0$ such that $\mathbb{E}[x_{t,i} x_{t,i}^\top] \geq \sigma_0$.*

In fact, the i.i.d. assumption on feature vectors is only required during the initialization phase to ensure the invertibility of V_{T_0}, which we discuss in Section 5.1.

Assumption 2. *For every item $i \in S$ and any $S \subset \mathcal{S}$ and all t, $\kappa := \min_{\|\theta - \theta^*\| \leq 1} p_{t,i}(S, \theta) p_{t,0}(S, \theta) > 0$.*

The asymptotic normality of MLE implies the necessity of this assumption. Note that this assumption is typical and equivalent or more restrictive assumptions appear in (generalized) linear contextual bandit literature (Filippi et al., 2010; Li et al., 2017) to ensure the Fisher information matrix is invertible. We discuss the need for this assumption in more detail in Section 4.

The revenue parameter $r_{t,i}$ for each item is also revealed at round t. Without loss of generality, assume $\|r_{t,i}\| \leq 1$. Then, the expected revenue corresponding to the assortment S_t is given by

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

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

$$S_t^* = \arg\max_{S \subset \mathcal{S}} R_t(S, \theta^*). \tag{2}$$

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

$$\mathcal{R}_T = \mathbb{E}\left(\sum_{t=1}^{T} \left(R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \right) \right)$$

where $R_t(S_t^*, \theta^*)$ is the expected revenue corresponding to the offline optimal assortment at time t.

4. MLE for Multinomial Logistic Regression

We consider the use of maximum likelihood to estimate the unknown parameter θ^* of the MNL model. Before we express the likelihood function, we first use a one-hot encoding with a binary vector $y_t \in \{0, 1\}^{|S_t|}$ for the user choice c_t in which $y_{t,i} = 1$ if the i-th item in the assortment S_t is chosen and and $y_{t,j} = 0$ for all $j \in S_t, j \neq i$. Then, the likelihood function is then given by

$$p(\boldsymbol{Y} | \theta^*) = \prod_{t=1}^{n} \prod_{i \in S_t} \left(p_{t,i}(S_t, \theta^*) \right)^{y_{t,i}}$$

where $\boldsymbol{Y}$ is $\{y_t\}_{t=1}^n$. Taking the negative logarithm gives

$$E(\theta^*) = -\log p(\boldsymbol{Y} | \theta^*) = -\sum_{t=1}^{n} \sum_{i \in S_t} y_{t,i} \log p_{t,i}(S_t, \theta^*)$$

which is known as the cross-entropy error function for the multi-class classification problem. Now, taking the gradient of this error function with respect to θ^*, we obtain

$$\nabla_{\theta^*} E(\theta^*) = \sum_{t=1}^{n} \sum_{i \in S_t} (p_{t,i}(S_t, \theta^*) - y_{t,i}) x_{t,i}$$

From the classical likelihood theory (Lehmann & Casella, 2006), as the sample size n goes to infinity, we know the MLE $\hat{\theta}_n$ is asymptotically normal, with $\hat{\theta}_n - \theta^* \to \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1})$ where $\mathcal{I}_{\theta^*}$ is the Fisher information matrix. We show in the proof of Theorem 2 that $\mathcal{I}_{\theta^*}$ is lower

bounded by $\sum_t \sum_{i \in S_t} p_{t,i}(\theta^*) p_{t,0}(\theta^*) x_{t,i} x_{t,i}^\top$. Hence, if $p_{t,i}(\theta^*) p_{t,0}(\theta^*) \geq \kappa > 0$, then we can ensure that $\mathcal{I}_{\theta^*}$ is invertible and prevent asymptotic variance of $x^\top \hat{\theta}$ from going to infinity for any x. Therefore, this justifies the need for Assumption 2 to ensure $p_{t,i}(\theta^*) p_{t,0}(\theta^*)$ is not too small.

5. Algorithms and Main Results

In this section, we present two algorithms for the MNL contextual bandit problem and their regret analyses. While the first algorithm is simpler and computationally more efficient, the second algorithm provides lower expected regret.

5.1. Algorithm UCB-MNL

The basic idea of the algorithm is to maintain a confidence set for the unknown true utility values, or more the precisely parameter of the utility function. The techniques of upper confidence bounds (UCB) have been widely known to be effective in balancing the exploration and exploitation trade-off in many bandit problems, including K-arm bandits (Auer et al., 2002; Lattimore & Szepesvári, 2019), linear bandits (Abbasi-Yadkori et al., 2011; Auer, 2002; Chu et al., 2011; Dani et al., 2008) and generalized linear bandits (Filippi et al., 2010; Li et al., 2017).

For each round t, the confidence set of θ^* is constructed from the feature vectors $\{x_{1,i}\}_{i \in S_1}$, ..., $\{x_{t,i}\}_{i \in S_t}$ and the observed feedback of selected items $y_1, ..., y_{t-1}$ from all previous rounds. Suppose $\hat{\theta}_t$ is the current estimator of the true parameter θ^* after time t. And, our estimator $\hat{\theta}_t$ lies within in the confidence set with radius $\alpha > 0$, which we show later in Lemma 2 holds true with a high probability as a function of α. Intuitively, the larger the radius α is, the more exploration will take place. Now, an exploitation is to offer a set S such that maximizes the estimated average revenue $R_t(S, \hat{\theta}_t)$, whereas an exploration is to choose a set which has a large variance in expected revenue. In the case of linear bandits or generalized linear bandits with an increasing inverse link function, balancing exploitation and exploration can be done simply by taking an action that maximizes the sum of $x_{t,i}^\top \hat{\theta}_t$ and the variance. However, in MNL bandits since the choice probability of an item is non-linear with the parameters of utility function (and the expected revenue is also weighted by the revenue for each item), we need to construct upper confidence bounds more carefully.

First we consider the following upper bound of $x_{t,i}^\top \hat{\theta}_t$,

$$z_{t,i} := x_{t,i}^\top \hat{\theta}_{t-1} + \alpha \|x_{t,i}\|_{V_{t-1}^{-1}} \tag{3}$$

where $V_t = \sum_{\tau=1}^{t} \sum_{i \in S_t} x_{\tau,i} x_{\tau,i}^\top \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix. We show later in Lemma 3 that $z_{t,i}$ is indeed an upper bound of $x_{t,i}^\top \theta^*$ if $\hat{\theta}_{t-1}$ lies within in the confidence set of θ^*. Then, we construct the following optimistic estimate of the expected revenue

$$\tilde{R}_t(S) := \frac{\sum_{i \in S} r_{t,i} \exp(z_{t,i})}{1 + \sum_{j \in S} \exp(z_{t,j})}.$$

Now, we assume an access to an optimization oracle which returns the assortment choice at time t, $S_t = \arg\max_{S \subset \mathcal{S}} \tilde{R}_t(S)$. This leads to Algorithm 1.

Algorithm 1 UCB-MNL

Input: total rounds T, initialization rounds T_0 and confidence radius α
Initialization: **for** $t \in [T_0]$
  Randomly choose S_t with $|S_t| = K$
  $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{t,i} x_{t,i}^\top$
for all $t = T_0 + 1$ to T **do**
  Compute $S_t = \arg\max_{S \subset \mathcal{S}} \tilde{R}_t(S)$
  Offer S_t and observe y_t (user choice at time t)
  Update $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{t,i} x_{t,i}^\top$
  Compute MLE $\hat{\theta}_t$ by solving the equation

$$\sum_{\tau=1}^{t} \sum_{i \in S_\tau} \left(p_{\tau,i}(S_\tau, \hat{\theta}_t) - y_{\tau,i} \right) x_{\tau,i} = 0 \tag{4}$$

  $t \leftarrow t + 1$
end for

In Algorithm 1, during the initialization phase, we first randomly choose an assortment S_t with exactly K items (note that after initialization, S_t can be smaller than K) to ensure a unique solution of (4). The initialization duration T_0 is specified later in Theorem 1, which is chosen to ensure that $\lambda_{\min}(V_{T_0})$ is large enough so that V_{T_0} to be invertible. The following proposition allows us to find such T_0. Its proof is deferred to the appendix.

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

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

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

The proposition implies that we can have $\lambda_{\min}(V_{T_0}) \geq K$ with a high probability if we run the initialization for $O(\sigma_0^{-2}(d + \log T))$ rounds. Similar to Filippi et al. (2010) and Li et al. (2017), the i.i.d. assumption (in Assumption 1) on the context $x_{t,i}$ is only needed to ensure that V_{T_0} is invertible at the end of the initialization phase. In the rest of the regret analysis, we do not require this stochastic

assumption. Hence, after the initialization, $x_{t,i}$ can even be chosen adversarially as long as $\|x_{t,i}\|$ is bounded.

The following lemma shows that the optimistic expected revenue $\tilde{R}_t(S_t)$ is an upper bound of the true expected revenue of the optimal assortment $R_t(S_t^*, \theta^*)$. The lemma is an adaptation of Lemma 4.2 in (Agrawal et al., 2017a) which is shown for non-contextual setting.

Lemma 1. *Suppose S_t^* is the offline optimal assortment as defined in (2), and suppose $S_t = \arg\max_{S \subset \mathcal{S}} \tilde{R}_t(S)$. If for every item $i \in S_t^*$, $z_{t,i} \geq x_i^\top \theta^*$, then the revenues satisfy the following inequalities for all round t:*

$$R_t(S_t^*, \theta^*) \leq \tilde{R}_t(S_t^*) \leq \tilde{R}_t(S_t).$$

It is important to note that Lemma 1 does not claim that the expected revenue is generally a monotone function, but only the value of the expected revenue corresponding to the optimal assortment increases with an increase in the MNL parameters (Agrawal et al., 2017a).

Regret Analysis of UCB-MNL

We present the following upper bound on the regret of the policy stated in Algorithm 1.

Theorem 1. *There exists a universal constant $C_0 > 0$, such that if we run UCB-MNL with $\alpha = \frac{\sigma}{\kappa}\sqrt{2d\log(T)}$ for total of T rounds with $T_0 = O(\sigma_0^{-2}(d + \log T))$ assortment size constraint K, then the expected regret of the algorithm with is upper-bounded by*

$$\begin{aligned}\mathcal{R}_T &\leq T_0 + O(1) + \frac{2\sigma d}{\kappa}\sqrt{2T\log\left(\frac{T}{d}\right)\log T} \\ &\leq O\left(d\sqrt{T\log(T/d)\log T}\right)\end{aligned}$$

The theorem demonstrates an $\tilde{O}(d\sqrt{T})$ regret bound for UCB-MNL which is independent of N; hence, applicable to the case of infinite items. Chen et al. (2018) established the lower bound result $\Omega(d\sqrt{T}/K)$ for MNL bandits. When K is small, which is typically true in most applications, the regret upper-bound in Theorem 1 demonstrates that our policy is almost optimal.

Proof. Proof of Theorem 1

The proof of the regert bound involves bounding $\|\hat{\theta}_t - \theta^*\|_{V_t}$ and $\sum_{t=1}^T \sum_{i \in S_t} \|x_{t,i}\|^2_{V_{t-1}^{-1}}$ as well as an immediate regret bound. We present the following lemmas whose proofs are deferred to the appendix.

The first lemma below shows that the true parameter θ^* lies within an ellipsoid centered at $\hat{\theta}_t$ with confidence radius α under V_t norm. Recall that Proposition 1 ensures that we have $\lambda_{\min}(V_{T_0}) \geq K$ at the end of the initialization phase if we run the initialization with size K assortments; hence the algorithm satisfies the condition of the following lemma.

Lemma 2. *Define $\alpha_t = \frac{\sigma}{\kappa}\sqrt{2d\log\left(1 + \frac{t}{d}\right) + \log t}$. If $\lambda_{\min}(V_{T_0}) \geq K$, then it follows that*

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t \tag{5}$$

holds for all $t > T_0$ with a probability $1 - O(t^{-1})$.

We emphasize that this finite-sample estimation error bound is new for MNL model.

The following lemma shows our optimistic utility estimate $z_{t,i}$ is an upper confidence bound for the expected utility $x_{t,i}^\top \theta^*$ if the true θ^* is contained in the confidence set of $\hat{\theta}_t$.

Lemma 3. *Let $z_{t,i}$ be defined as (3). If event $\mathcal{E}^\theta$ holds, then we have*

$$0 \leq z_{t,i} - x_{t,i}^\top \theta^* \leq 2\alpha_t \|x_{t,i}\|_{V_{t-1}^{-1}}.$$

Then we show that the expected revenue has some Lipschitz property and bound the immediate regret with the maximum variance over the assortment.

Lemma 4. *Suppose that $z_{t,i} - x_{t,i}^\top \theta^* \leq 2\alpha_t \|x_{t,i}\|_{V_{t-1}^{-1}}$ holds for $i \in S_t$ where S_t is the chosen assortment in round t. Then, we have*

$$\tilde{R}_t(S_t) - R_t(S_t, \theta^*) \leq 2\alpha_t \max_{i \in S_t} \|x_{t,i}\|_{V_{t-1}^{-1}}$$

The next technical lemma upper bounds the sum of squared norms.

Lemma 5. *Define $V_{T_0} = \sum_{t=1}^{T_0} \sum_{i \in S_\tau} x_{t,i} x_{t,i}^\top$ and $V_T = V_{T_0} + \sum_{t=V_0+1}^{T} \sum_{i \in S_t} x_{t,i} x_{t,i}^\top$. If $\lambda_{\min}(V_{T_0}) \geq K$, then we have*

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{t,i}\|^2_{V_{t-1}^{-1}} \leq 2d\log\left(\frac{T}{d}\right)$$

Now we can combine the results to show the cumulative regret bound. First we define the high probability events for the concentration of parameter and the random initialization.

Definition 1. *Define the following events:*

$$\begin{aligned}\mathcal{E}^\lambda &= \{\lambda_{\min}(V_{T_0}) \geq K\} \\ \mathcal{E}^\theta &= \{\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t, \forall t \leq T\}\end{aligned}$$

Then we break the regret into the initialization phase and the learning phase:

$$\begin{aligned}\mathcal{R}_T &\leq T_0 + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\left(R(\tilde{S}_t^*, \theta^*) - R(S_t, \theta^*)\right)\right] \\ &\leq T_0 + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\left(\tilde{R}_t(S_t) - R(S_t, \theta^*)\right)\right]\end{aligned}$$

where the last inequality comes from optimistic revenue estimation by Lemma 1. Now, we break the regret of the learning phase further into two components – when both event $\mathcal{E}^\theta$ in Lemma 2 and event $\mathcal{E}^\lambda$ in Proposition 1 are true (i.e. $\mathcal{E}^\theta \cap \mathcal{E}^\lambda$) and when either of the events is not true, (i.e. $\bar{\mathcal{E}}^\theta \cup \bar{\mathcal{E}}^\lambda$).

$$\begin{aligned}\mathcal{R}_T &\le T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\tilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}\left(\mathcal{E}^\theta \cap \mathcal{E}^\lambda\right)\right] \\ &\quad + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\tilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}\left(\bar{\mathcal{E}}^\theta \cup \bar{\mathcal{E}}^\lambda\right)\right] \\ &\le T_0 + \sum_{t=1}^{T} 2\alpha \max_{i\in S_t}\|x_{t,i}\|_{V_{t-1}^{-1}} + O(1)\end{aligned}$$

where the last inequality is from Lemma 4 and $\alpha = \alpha_T$. Applying Cauchy-Schwarz inequality in the second term, it follows that

$$\mathcal{R}_T \le T_0 + 2\alpha\sqrt{\sum_{t=1}^{T}\max_{i\in S_t}\|x_{t,i}\|^2_{V_{t-1}^{-1}}} + O(1).$$

Applying Lemma 5 for $\sum_{t=1}^{T}\max_{i\in S_t}\|x_{t,i}\|^2_{V_{t-1}^{-1}}$,

$$\mathcal{R}_T \le T_0 + 2\alpha\sqrt{2d\log\left(\frac{T}{d}\right)} + O(1).$$

Finally, choosing $\alpha = \frac{\sigma}{\kappa}\sqrt{2d\log\left(1+\frac{T}{d}\right) + \log T}$, we have

$$\mathcal{R}_T \le T_0 + \frac{2\sigma d}{\kappa}\sqrt{2T\log\left(\frac{T}{d}\right)\log T} + O(1)$$

□

5.2. Non-asymptotic Normality of the MLE for MNL

Before we present our next algorithm CB-MNL, we first present Theorem 2, which is a crucial component in the regret analysis of CB-MNL which is proposed in Section 5.3. The following theorem is a generalization of Theorem 1 in Li et al. (2017), which presents a finite-sample version of the classical asymptotic normality of the MLE for generalized linear model (GLM). Our version is a generalization to a multinomial setting, i.e. if S_t contains only a single item, then it is equivalent to the GLM version presented in Li et al. (2017).

Theorem 2. *Define $V_t = \sum_{u=1}^{t}\sum_{i\in S_t} x_{\tau,i}x_{\tau,i}^\top$, and let $\delta > 0$ be given. Furthermore, assume that*

$$\lambda_{\min}(V_t) \ge \frac{4096\sigma^2}{\kappa^2}\left(d^2 + \log\frac{1}{\delta}\right) \tag{6}$$

Then, with probability at least $1 - 3\delta$, the maximum likelihood estimator satisfies, for any $x \in \mathbb{R}^d$, that

$$|x^\top(\hat{\theta}_t - \theta^*)| \le \frac{5\sigma}{\kappa}\sqrt{\log\frac{1}{\delta}}\|x\|_{V_t^{-1}}$$

The proof of Theorem 2 is presented in the appendix. It is important to note that although the statement of the theorem is similar to the GLM version in Li et al. (2017). The GLM version is not directly applicable to the MNL model, due to the dependency of the choice probability over different items $i, j \in S_t$ and their outer product of the contexts $x_{t,i}x_{t,j}^\top$ in the Fisher information matrix (see the proof of Theorem 2 in the appendix). This theorem characterizes the behavior of MLE on every direction. The theorem implies that $x^\top(\hat{\theta}_t - \theta^*)$ has a sub-Gaussian tail bound for any $x \in \mathbb{R}^d$, which enables us to improve the regret bound by the factor of $\sqrt{d}$ compared to Theorem 1 for UCB-MNL.

5.3. Algorithm CB-MNL

Inspired by Li et al. (2017), we propose another algorithm CB-MNL (Algorithm 3) which uses MLE-MNL (Algorithm 2) as a sub-routine. This algorithm operates on the radius of the confidence bound, independent of expected mean utility, to perform exploration. At round t, the algorithm screens the candidate actions based on the value of $\max_{i\in S} w_{t,i}^{(\ell)}$ through L epochs until an assortment S_t is chosen.

The algorithm maintains $\{\Psi_\ell\}_{\ell=0}^{L}$, the sets of time indices which are the partitions of the entire time horizon $\{1, 2, ..., T\}$. The purpose of this partitioning is to ensure that the choice responses y_t in each index set Ψ_ℓ are independent, so that we can apply the normality result in Theorem 2 to each of Ψ_ℓ individually (then combine them together to get the total regret). The idea was first introduced by Auer (2002) and further developed by Li et al. (2017).

Algorithm 2 MLE-MNL

1: **Input**: parameter α, index set $\Psi(t)$, candidate set A
2: Compute MLE $\hat{\theta}_t$ by solving the equation

$$\sum_{\tau\in\Psi(t)}\sum_{i\in S_\tau}\left(p_{\tau,i}(S_\tau,\theta) - y_{\tau,i}\right)x_{\tau,i} = 0$$

3: Update $V_t = \sum_{\tau\in\Psi(t)}\sum_{i\in S_\tau} x_{\tau,i}x_{\tau,i}^\top$
4: Compute the following:

$$w_{t,i} = \alpha\|x_{t,i}\|_{V_t^{-1}} \text{ for all } i \in [N]$$
$$\mathcal{W}_t = 2\max_{i\in\mathcal{I}} w_{t,i}$$

where $\mathcal{I} = \{i \in S : S \in A\}$

At each round t, the algorithm goes though epochs ℓ up to L until S_t is chosen.

- *Sub-routine:* in step (a), we run MLE-MNL (Algorithm 2) which uses the normality result to compute $w_{t,i}^{(\ell)}$ for all i, $\mathcal{W}_t^{(\ell)}$, and $\hat{\theta}_t^{(\ell)}$. We can utilize the normality result here since $\{y_t, t \in \Psi_\ell\}$'s are independent given the feature vectors in each Ψ_ℓ (see Lemma 6).
- *Exploitation:* in step (b), if the maximal confidence interval of an assortment is very small, smaller than $\frac{1}{K\sqrt{T}}$, for all possible candidate sets, then we perform pure exploitation. This step's contribution to the total regret will be small.
- *Exploration:* in step (c), if there is a set that has large confidence interval (larger than $2^{-\ell}$), then we choose that set as S_t. Then we update the index set Ψ_ℓ to include the timestamp t.
- *Pruning:* finally, step (d) is a pruning step, where we remove clearly sub-optimal sets and keep the sets which are possibly optimal.

Algorithm 3 CB-MNL

Input: Lengths of trials T and pilot τ, parameter α
Initialization: **for** $t \in [\tau]$
  randomly choose S_t with $|S_t| = K$
  set $L = \lfloor \frac{1}{2} \log_2 T \rfloor$, and $\Psi_0 = \cdots = \Psi_L = \varnothing$.
for all $t = \tau + 1$ to T **do**
  Initialize $A_1 = \mathcal{S}$ and $\ell = 1$
  while S_t is empty **do**
    (a). Run MLE-MNL with A_ℓ, α and $\Psi_\ell \cup [\tau]$ to compute $\hat{\theta}_t^{(\ell)}$, $w_{t,i}^{(\ell)}$, $\mathcal{W}_t^{(\ell)}$
    (b). **If** $\mathcal{W}_t^{(\ell)} \le \frac{1}{\sqrt{T}}$,
      set $S_t = \arg\max_{S \in A_\ell} R_t(S, \hat{\theta}_t^{(\ell)})$;
      update $\Psi_0 = \Psi_0 \cup \{t\}$
    (c). **Else if** $\mathcal{W}_t^{(\ell)} > 2^{-\ell}$,
      set $S_t = \arg\max_{S \in A_\ell} \sum_{i \in S} w_{t,i}^{(\ell)}$;
      update $\Psi_\ell = \Psi_\ell \cup \{t\}$
    (d). **Else if** $\mathcal{W}_t^{(\ell)} \le 2^{-\ell}$,
      compute $\mathcal{M}_t^{(\ell)} = \max_{S \in A_\ell} R_t(S, \hat{\theta}_t^{(\ell)})$
      $A_{\ell+1} = \left\{ S \in A_\ell : R_t(S, \hat{\theta}_t^{(\ell)}) \ge \mathcal{M}_t^{(\ell)} - 2 \cdot 2^{-\ell} \right\}$
      $\ell \leftarrow \ell + 1$
  end while
end for

If the algorithm does not choose S_t in epoch ℓ, then it moves on to the next epoch $\ell + 1$ and repeat the process until S_t is chosen either through exploitation action in (b) or exploration action in (c). Note that when maximizing the expected revenue $R_t(S, \hat{\theta})$ in step (b) or in step (d), it uses the expected revenue defined in (1) replacing θ^* with the current estimator $\hat{\theta}_t^{(\ell)}$ — it is not the optimistic expected revenue $\tilde{R}_t(S)$ which is used in Algorithm 1.

Adapted from Lemma 14 of Auer (2002) and Lemma 4 of Li et al. (2017), the following result shows that the samples collected from Algorithm 3 in each index set Ψ_ℓ are independent. The proof is presented in the appendix.

Lemma 6. *For all $\ell \in [L]$ and $t \in [T]$, given the set of feature vectors in index set Ψ_ℓ, $\{\{x_{t,i}\}_{i \in S_t}, t \in \Psi_\ell\}$, the corresponding choice responses $\{y_t, t \in \Psi_\ell\}$ are independent random variables.*

Regret Analysis of CB-MNL

Independent samples enable us to apply the non-asymptotic normality result in Theorem 2. We present the following regret bound of CB-MNL (Algorithm 3).

Theorem 3. *There exists a universal constants C and C_0, such that if we run CB-MNL algorithm with $T_0 = \frac{C_0}{\sigma_0^2 K} \sqrt{dT} \log T$ and $\alpha = \frac{5\sigma}{\kappa} \sqrt{\log(TN \log T)}$ for $T \ge T'$ rounds, where*

$$T' = \Omega\left(\frac{\sigma^2}{\kappa^4} \max\left\{ \frac{\log TN}{d}, d^3 \right\} \right), \tag{7}$$

the expected regret of the algorithm is upper-bounded as

$$\begin{aligned} R_T &\le \frac{C\sigma}{\kappa} \sqrt{dT \log(T/d) \log(TN \log T) \log T} \\ &\quad + T_0 + O(1) + 2\sqrt{T} \\ &\le O\left(\sqrt{dT \log(T/d) \log(TN \log T) \log T} \right). \end{aligned}$$

The theorem establishes $\tilde{O}(\sqrt{dT})$ regret bound for CB-MNL algorithm. Chu et al. (2011) showed the minimax lower bound of the expected regret of $\Omega(\sqrt{dT})$ in finite-armed linear bandits, a special of the GLM bandits, which is again a special case (when the assortment size is exactly 1) of MNL contextual bandits we consider in the work. To the best of our knowledge, this is the first algorithm which achieves the rate of $\tilde{O}(\sqrt{dT})$ regret in MNL contextual bandits. Comparing with Theorem 1 for UCB-MNL (Algorithm 1), the improvement of $\sqrt{d}$ factor comes from avoiding separately bounding $\|\theta^* - \hat{\theta}_t\|_{V_t}$ and $\|x\|_{V_t^{-1}}$, each of which contains $\sqrt{d}$ term; hence resulting in extra $\sqrt{d}$ when combined. Note that the regret bound in Theorem 3 has logarithmic dependence on N, therefore CB-MNL is not applicable to an infinite number of total items. However, when N is not exponentially large, the rate of CB-MNL is faster.

Proof of Theorem 3. We first present two lemmas to help bound the cumulative expected regret. The first lemma ensures that normality results (Theorem 2) holds with given confidence radius α for all items.

Lemma 7. *Let $T_0 = \frac{C_0}{\sigma_0^2 K}\sqrt{dT}\log T$. Suppose $T \geq T_0$ where T_0 is defined as (7). Fix $\alpha > 0$. Define the following event:*

$$\mathcal{E}_t^u := \left\{ |x_{t,i}^\top \hat{\theta}_{t-1} - x_{t,i}^\top \theta^*| \leq \alpha \|x_{t,i}\|_{V_t^{-1}}, \forall i \in [N] \right\} \tag{8}$$

Then, event $\mathcal{E}_t^u$ holds for all $t > T_0$ (and for all epochs ℓ within each round t) with probability at least $1 - O(t^{-1})$

Proof sketch. The proof follows Lemma 6 and Theorem 2. Lemma 6 ensure independent samples. Using Proposition 2 we lower bound $\lambda_{\min}(V_{T_0})$ at the end of initialization to successfully apply . Denote this event by $\mathcal{E}_\tau^\lambda$. Then our choice of τ indeed ensures $\mathcal{E}_\tau^\lambda$ holds with high probability. This allows us to use Theorem 2. Then we apply union bound for all items i. Then solving for δ gives the results. □

The next lemma bounds the immediate regret of CB-MNL, breaking down to two choice scenarios — when an assortment is chosen for exploitation (step (b)) or for exploration (step (c)) in Algorithm 2. The proof is deferred to the appendix.

Lemma 8. *Suppose that event $\mathcal{E}_t^u$ in (8) holds, and that in round t, the assortment S_t is chosen at stage ℓ_t. Then $S_t^* \in A_\ell$ for all $\ell \leq \ell_t$. Furthermore, we have*

$$R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \leq \begin{cases} \frac{2}{\sqrt{T}}, & \text{if } S_t \text{ chosen in step}(b) \\ \frac{8}{2^{\ell_t}}, & \text{if } S_t \text{ chosen in step}(c) \end{cases}$$

Then, we follow the similar arguments of Li et al. (2017) to show the cumulative expected regret bound. First, define $V_{\ell,t} = \sum_{t\in\Psi_\ell}\sum_{i\in S_t} x_{t,i}x_{t,i}^\top$, then by Lemma 5 and CauchySchwarz inequality, we have

$$\begin{aligned}\sum_{t\in\Psi_\ell} \max_{i\in S_t} w_{t,i}^{(\ell)} &= \sum_{t\in\Psi_\ell} \max_{i\in S_t} \alpha \|x_{t,i}\|_{V_{\ell,t}^{-1}} \\ &\leq \alpha\sqrt{2|\Psi_\ell| d\log(T/d)}.\end{aligned}$$

However, from the choices made at exploration steps (step (c)) of Algorithm 3, we know

$$2^{-\ell}|\Psi_\ell| \leq 2\sum_{t\in\Psi_\ell} \max_{i\in S_t} w_{t,i}^{(\ell)}$$

for $\ell \in \{1, ..., L\}$. Now, we combine the two inequalities above. Then it follows that

$$|\Psi_\ell| \leq 2^{\ell+1}\alpha\sqrt{2|\Psi_\ell| d\log(T/d)}. \tag{9}$$

Note that each index set Ψ_ℓ is a disjoint set with $\cup_{\ell=0}^L \Psi_\ell = \{t+1, ..., T\}$. Then, similar to the first steps in the proof of Theorem 1, we break the regret into three components – when both event $\mathcal{E}_t^u$ in (8) and event $\mathcal{E}^\lambda$, which is to ensure the minimum eigenvalue of V_{T_0} is large enough, are true ($\mathcal{E}_t^u \cap \mathcal{E}^\lambda$) and when either of the events is not true ($\bar{\mathcal{E}}_t^u \cup \bar{\mathcal{E}}^\lambda$), and the random initialization phase with length T_0. Note that we need the minimum eigenvalue of V_{T_0} to be larger than the case in Definition 1 but we can still use Proposition 1 to ensure such case with probability $1 - O(T^{-1})$.

$$\begin{aligned}\mathcal{R}_T \leq T_0 &+ \mathbb{E}\left[\sum_{t=T_0+1}^{T} (R(S^*,\theta^*) - R(S_t,\theta^*))\,\mathbb{1}\left(\mathcal{E}_t^u \cap \mathcal{E}^\lambda\right)\right] \\ &+ \mathbb{E}\left[\sum_{t=T_0+1}^{T} (R(S^*,\theta^*) - R(S_t,\theta^*))\,\mathbb{1}\left(\bar{\mathcal{E}}_t^u \cup \bar{\mathcal{E}}^\lambda\right)\right]\end{aligned}$$

We can decompose the regret into the disjoint stages recorded by Ψ_ℓ.

$$\begin{aligned}\mathcal{R}_T &\leq T_0 + \mathbb{E}\left[\sum_{t\in\Psi_0} (R(S^*,\theta^*) - R(S_t,\theta^*))\,\mathbb{1}\left(\mathcal{E}_t^u \cap \mathcal{E}^\lambda\right)\right] \\ &\quad + \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{t\in\Psi_\ell} (R(S^*,\theta^*) - R(S_t,\theta^*))\,\mathbb{1}\left(\mathcal{E}_t^u \cap \mathcal{E}^\lambda\right)\right] \\ &\quad + O(1) \\ &\leq T_0 + \frac{2}{\sqrt{T}}|\Psi_0| + \sum_{\ell=1}^{L}\frac{8}{2^\ell}|\Psi_\ell| + O(1) \\ &\leq T_0 + 2\sqrt{T} + \sum_{\ell=1}^{L} 16\alpha\sqrt{2|\Psi_\ell| d\log(T/d)} + O(1) \\ &\leq T_0 + 2\sqrt{T} + 16\alpha\sqrt{2dLT\log\frac{T}{d}} + O(1)\end{aligned}$$

where the third inequality uses (9) and the last inequality is by Cauchy-Schwartz inequality. Now, we choose $\alpha = 5\sigma\sqrt{\log(TN\log T)}$, $T_0 = \frac{C_0}{\sigma_0^2}\sqrt{dT}\log T$ and $L = \frac{1}{2}\log T$, then we have the regret bound in Theorem 3.

□

6. Discussions

In this paper, we study the dynamic assortment selection problem under an MNL contextual model. We propose two algorithms for MNL contextual bandits which learn the parameters of the underlying choice model while simultaneously maximizing the cumulative revenue. While the first algorithm UCB-MNL achieves the optimal rate for the case of infinite number of items, the second algorithm CB-MNL achieves a faster rate for the case of finite number of items at each round.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: a dynamic learning approach to assortment selection. *arXiv preprint arXiv:1706.03880*, 2017a.

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

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

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

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

Chen, K., Hu, I., Ying, Z., et al. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Chen, X. and Wang, Y. A note on tight lower bound for mnl-bandit assortment selection models. *arXiv preprint arXiv:1709.06109*, 2017.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment optimization with changing contextual information. *arXiv preprint arXiv:1810.13069*, 2018.

Cheung, W. C. and Simchi-Levi, D. Assortment optimization under unknown multinomial logit choice models. *arXiv preprint arXiv:1704.00108*, 2017.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 355366, 2008.

Davis, J. M., Gallego, G., and Topaloglu, H. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

Désir, A., Goyal, V., and Zhang, J. Near-optimal algorithms for capacity constrained assortment optimization. *Available at SSRN 2543309*, 2014.

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

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080, 2017.

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

McFadden, D. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.

Plackett, R. L. The analysis of permutations. *Applied Statistics*, pp. 193–202, 1975.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

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

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

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

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

Van der Vaart, A. W. and Wellner, J. A. Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608, 1997.