Multinomial Logit Contextual Bandits: Provable Optimality and Practicality

Min-hwan Oh[1] **and Garud Iyengar**[2]
[1]Seoul National University, Seoul, South Korea
[2]Columbia University, New York, USA
[1]minoh@snu.ac.kr, [2]garud@ieor.columbia.edu

Abstract

We consider a sequential assortment selection problem where the user choice is given by a multinomial logit (MNL) choice model whose parameters are unknown. In each period, the learning agent observes a d-dimensional contextual information about the user and the N available items, and offers an assortment of size K to the user, and observes the bandit feedback of the item chosen from the assortment. We propose upper confidence bound based algorithms for this MNL contextual bandit. The first algorithm is a simple and practical method which achieves an $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret over T rounds. Next, we propose a second algorithm which achieves a $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. This matches the lower bound for the MNL bandit problem, up to logarithmic terms, and improves on the best known result by a $\sqrt{d}$ factor. To establish this sharper regret bound, we present a non-asymptotic confidence bound for the maximum likelihood estimator of the MNL model that may be of independent interest as its own theoretical contribution. We then revisit the simpler, significantly more practical, first algorithm and show that a simple variant of the algorithm achieves the optimal regret for a broad class of important applications.

Introduction

In many of the human-algorithm interactions today, a learning agent (algorithm) makes sequential decisions and receives user (human) feedback *only* for the chosen decisions. The multi-armed bandit (Lattimore and Szepesvári 2019) is a model for this sequential decision making with partial feedback. It is a classic reinforcement learning problem that exemplifies the dilemma of exploration vs. exploitation. This multi-armed bandit model has found diverse applications, e.g. learning click-through rates in search engines, product recommendations in online retailing, movie suggestions on streaming services, news feeds, etc. Note that in several of the applications, the goal is to maximize an appropriate "clickthrough" rate. Often information about the features of the agent's actions and contextual information about the user are available. The contextual bandit extends the multi-armed bandit by making the decision conditional on this context and feature information. In many real-world problems including the aforementioned examples, the agent offers a menu of options to the user, rather than a single option as in traditional bandit action selection. The user chooses at most one of the offered options, and the agent receives a reward associated with the user choice.

In this paper, we consider a sequential assortment selection problem 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, which is defined as the gap between the expected revenue generated by the algorithm and the offline optimal expected revenue when the true parameter is known. The d-dimensional contextual information, or a set of feature vectors, is revealed at each round t, allowing the feature information of items 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 (McFadden 1978). This is one of the most widely used model in dynamic assortment optimization literature (Caro and Gallien 2007; Rusmevichientong, Shen, and Shmoys 2010; Sauré and Zeevi 2013; Agrawal et al. 2019, 2017; Aouad, Levi, and Segev 2018).

For sequential decision-making with contextual information, (generalized) linear bandits (Abe and Long 1999; Auer 2002; Filippi et al. 2010; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011; Li, Lu, and Zhou 2017) and their variants have been widely studied. However, these methods are only limited to a single item selection which is increasingly rarer in practice as compared to multiple item offering that we consider in this work. There are a line of works in combinatorial variants of contextual bandit problems (Qin, Chen, and Zhu 2014; Wen, Kveton, and Ashkan 2015; Kveton et al. 2015; Zong et al. 2016) mostly with semi-bandit feedback or cascading feedback. However, these methods do not take the user choice into account. Hence, substitution effect is not considered. In contrast to these contextual bandit problems and their combinatorial variants, in the multinomial logit (MNL) contextual bandit, the item choice (feedback) is a function of all items in the offered assortment. The key challenges are how to design an algorithm that offers assortments to simultaneously learn the unknown parameter

	METHOD	CONTEXT	REGRET
AGRAWAL ET AL. (2019)	UCB	NO	$\widetilde{\mathcal{O}}(\sqrt{NT})$, $\Omega(\sqrt{NT/K})$
AGRAWAL ET AL. (2017)	TS	NO	$\widetilde{\mathcal{O}}(\sqrt{NT})$
CHEUNG AND SIMCHI-LEVI (2017)	TS	YES	$\widetilde{\mathcal{O}}(d\sqrt{T})^*$
CHEN AND WANG (2017)	N/A	N/A	$\Omega(\sqrt{NT})$ $(\equiv \Omega(\sqrt{dT}))$
OU ET AL. (2018)	UCB	YES	$\widetilde{\mathcal{O}}(Kd\sqrt{T})$
CHEN, WANG, AND ZHOU (2018)	UCB	YES	$\widetilde{\mathcal{O}}(d\sqrt{T})$, $\Omega(d\sqrt{T}/K)$
OH AND IYENGAR (2019)	TS	YES	$\widetilde{\mathcal{O}}(d\sqrt{T})^*$, $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$
THIS WORK (ALGORITHM 1)	UCB	YES	$\widetilde{\mathcal{O}}(d\sqrt{T})$
THIS WORK (ALGORITHMS 2)	UCB	YES	$\widetilde{\mathcal{O}}(\sqrt{dT})$

Table 1: Comparison of regret bounds in related works on MNL bandits. T is the number of total rounds, K is the assortment size, N is the total number of items, and d is the feature dimension. UCB denotes upper-confidence bound and TS denotes Thompson sampling, and starred (*) regrets denote Bayesian regrets. $\widetilde{\mathcal{O}}$ is a big-$\mathcal{O}$ notation up to logarithmic factors.

and maximize the expected revenue through sequential interactions with users and how to guarantee its performance. There has been an emerging body of literature on MNL bandits in both non-contextual and contextual settings (Agrawal et al. 2017, 2019; Cheung and Simchi-Levi 2017; Ou et al. 2018; Chen, Wang, and Zhou 2018; Oh and Iyengar 2019). However, an open question in the MNL contextual bandit problem is whether one can close the gap between lower and upper bounds of regret. Often, meeting such a criterion comes at the cost of practicality. Hence, designing a practical algorithm that achieves the provable optimality becomes a greater challenge. Our contributions are as follows:

- `UCB-MNL` (Algorithm 1) is an upper confidence bound based algorithm for MNL contextual bandits that, to our knowledge, is the first polynomial time algorithm that achieves an N independent $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret. This result matches the previous best upper bound (up to logarithmic factors).
- We show that $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret is achievable in the MNL contextual bandits (Theorem 3). This improves on the best previous result by $\sqrt{d}$ factor, and matches the lower bound for the MNL bandit problem to within logarithmic factor. However, the resulting algorithm is not practical as with other provably optimal bandit algorithms that rely on a framework proposed in Auer (2002).
- `DBL-MNL` (Algorithms 2), a simple variant of `UCB-MNL`, achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret when revenue is uniform for all items — a setting that arises in a wide range of applications. `DBL-MNL` does *not* rely on the framework in Auer (2002), and has state-of-the-art computational efficiency. Thus, this work is the first one to provide a practical algorithm with provable $\sqrt{d}$ dependence on the dimension of the context.
- To establish a sharper regret bound, we prove a non-asymptotic confidence bound for the maximum likelihood estimator of the MNL model, which may be of independent interest.

Problem Formulation

Notations

For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its ℓ_2-norm. 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 eigenvalues of a symmetric matrix V are written as $\lambda_{\min}(V)$ and $\lambda_{\max}(V)$ 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. For a positive integer n, we define $[n] = \{1, 2, ..., n\}$. Finally, we define $\mathcal{S}$ to be the set of candidate assortments with size constraint at most K, i.e. $\mathcal{S} = \{S \subset [N] : |S| \leq K\}$. Although we treat $\mathcal{S}$ as stationary for ease of exposition, we can allow $\mathcal{S}$ (as well as the item set $[N]$) to change over time.

MNL Contextual Bandits

The MNL contextual bandits problem is defined as follows. The agent has a set of N distinct items. At each round t, the agent observes feature vectors $x_{ti} \in \mathbb{R}^d$ for every item $i \in [N]$. Given this contextual information, at every round t, the agent offers an assortment $S_t = \{i_1, \ldots, i_\ell\} \in \mathcal{S}$, $\ell \leq K$, 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 (McFadden 1978) under which the choice probability for item $i_k \in S_t$ (and the outside option) is defined as

$$p_t(i_k|S_t, \theta^*) = \frac{\exp\{x_{ti_k}^\top \theta^*\}}{1 + \sum_{j \in S_t} \exp\{x_{tj}^\top \theta^*\}},$$
$$p_t(0|S_t, \theta^*) = \frac{1}{1 + \sum_{j \in S_t} \exp\{x_{tj}^\top \theta^*\}}$$

where $\theta^* \in \mathbb{R}^d$ is a time-invariant parameter unknown to the agent. The choice response for each item $i_k \in S_t$ is defined as $y_{ti_k} := \mathbb{1}(c_t = i_k) \in \{0, 1\}$ and $y_{t0} := \mathbb{1}(c_t = 0)$ for the outside option. Hence the choice response variable

$y_t = (y_{t0}, y_{ti_1}, ..., y_{ti_\ell})$ is a sample from this multinomial distribution:

$$y_t \sim \text{multinomial}\left\{1, (p_t(0|S_t, \theta^*), ..., p_t(i_\ell|S_t, \theta^*))\right\}$$

where the parameter 1 indicates that y_t is a single-trial sample, i.e. $y_{t0} + \sum_{k=1}^{\ell} y_{ti_k} = 1$. For each $i \in S_t \cup \{0\}$ and t, we define the noise $\epsilon_{ti} := y_{ti} - p_t(i|S_t, \theta^*)$. Since each ϵ_{ti} is a bounded random variable in $[0, 1]$, ϵ_{ti} is σ^2-sub-Gaussian with $\sigma^2 = 1/4$; however, ϵ_{ti} is *not* independent across $i \in S_t$ due to the substitution effect in the MNL model. The revenue parameter r_{ti} for each item is also given at round t. r_{ti} is the revenue from the sale if item i is sold in round t. Without loss of generality, assume $|r_{ti}| \leq 1$ for all i and t. Then, the expected revenue of the assortment S_t is given by

$$R_t(S_t, \theta^*) = \sum_{i \in S_t} r_{ti} p_t(i|S_t, \theta^*) \tag{1}$$

Note that for a very broad class of MNL applications, including search ranking and media recommendation, the goal is to maximize the click-through rate; therefore, the item revenue is uniform.

We define S_t^* to be the offline optimal assortment at time t when θ^* is known apriori, i.e. when the true MNL probabilities $p_t(i|S, \theta^*)$ are known a priori:

$$S_t^* = \operatorname*{argmax}_{S \subset \mathcal{S}} R_t(S, \theta^*). \tag{2}$$

The learning agent does not know the value of θ^*, and therefore, can only choose the assortment S_t in period t based on the choices S_τ for periods $\tau < t$, and the observed responses. We measure the performance of the agent by the regret $\mathcal{R}_T$ for the time horizon of T periods, which is the gap between the expected revenue generated by the assortment chosen by the agent and that of the offline optimal assortment, i.e.,

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

where $R_t(S_t^*, \theta^*)$ is the expected revenue corresponding to the offline optimal assortment in period t, i.e., the highest revenue which can be obtained with the knowledge of θ^*. Hence, maximizing the cumulative expected revenue is equivalent to minimizing the cumulative expected regret.

MLE for Multinomial Logistic Regression

We briefly discuss the maximum likelihood estimation of the unknown parameter θ^* for the MNL model. First, recall that $y_t \in \{0, 1\}^{|S_t|+1}$ is the user choice response variable where y_{ti} is the i-th component of y_t. Then, the negative log-likelihood function under parameter θ is then given by $\ell_n(\theta) := -\sum_{t=1}^{n} \sum_{i \in S_t \cup \{0\}} y_{ti} \log p_t(i|S_t, \theta)$ which is also known as the cross-entropy error function for the multi-class classification problem. Taking the gradient of this negative log-likelihood with respect to θ, we obtain

$$\nabla_\theta \ell(\theta) = \sum_{t=1}^{n} \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti}) x_{ti}$$

As the sample size n goes to infinity, the MLE $\hat{\theta}_n$ is asymptotically according to the classical likelihood theory (Lehmann and Casella 2006), 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 $\sum_t \sum_{i \in S_t} p_t(i|\theta^*) p_t(0|\theta^*) x_{ti} x_{ti}^\top$. 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.

Algorithms and Main Results

In this section, we present algorithms for the MNL contextual bandit problem and their regret bounds.

Algorithm: `UCB-MNL`

The basic idea of our first algorithm is to maintain a confidence set for the parameter θ^*. 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, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2019), linear bandits (Auer 2002; Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011) and generalized linear bandits (Filippi et al. 2010; Li, Lu, and Zhou 2017).

For each round t, the confidence set $\mathcal{C}_t$ for θ^* is constructed from the feature vectors $\{x_{t'i}, i \in S_{t'}\}_{t' \leq t}$ and the observed feedback of selected items $y_1, ..., y_{t-1}$ from all previous rounds. Let $\hat{\theta}_t$ denote the estimate of the unknown parameter θ^* after t periods, and suppose we are guaranteed that θ^* lies within the confidence set $\mathcal{C}_t$ centered at MLE $\hat{\theta}_t$ with radius $\alpha_t > 0$ with a high probability. The radius α_t has to be chosen carefully: larger α_t induces more exploration; however, too large α_t can cause regret to increase. In the MNL setting, exploitation is to offer $\operatorname{argmax}_{S \in \mathcal{S}} R_t(S, \hat{\theta}_t)$, whereas exploration is to choose a set S that has the potential for high expected revenue $R_t(S, \theta)$ as θ varies over $\mathcal{C}_t$. Thus, a direct way to introduce optimism, and induce exploration, is to define an optimistic revenue for each $\binom{N}{K}$ assortments. This is the approach taken in Chen, Wang, and Zhou (2018); however, this enumeration has exponential complexity when N is large and K is relatively small. We show that one can induce sufficient exploration by defining an optimistic utility z_{ti} for each item, and defining the optimistic revenue for any assortment S using the optimistic utility.

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

where $V_t = \sum_{t'=1}^{t} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix. The optimistic utility z_{ti} consists of two components: mean utility estimate $x_{ti}^\top \hat{\theta}_{t-1}$ and standard deviation $\alpha_t \|x_{ti}\|_{V_{t-1}^{-1}}$. In the proof of the regret bound of the algorithm, we show that z_{ti} is, indeed, an upper bound of $x_{ti}^\top \theta^*$ if θ^* lies within in the confidence ellipsoid centered at $\hat{\theta}_{t-1}$. Based on z_{ti}, we construct the following optimistic

Algorithm 1 UCB-MNL

Input: initialization T_0, confidence radius α_t
Initialization: **for** $t \in [T_0]$
  Randomly choose S_t with $|S_t| = K$
  $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{ti} x_{ti}^\top$
for all $t = T_0 + 1$ to T **do**
  Compute $z_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \alpha_t \|x_{ti}\|_{V_{t-1}^{-1}}$ for all i
  Offer $S_t = \operatorname{argmax}_{S \subset \mathcal{S}} \widetilde{R}_t(S)$ and observe y_t
  Update $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{ti} x_{ti}^\top$
  Compute MLE $\hat{\theta}_t$ by solving $\sum_{t'=1}^{t} \sum_{i \in S_{t'}} \left(p_{t'}(i|S_{t'}, \hat{\theta}_t) - y_{t'i}\right) x_{t'i} = \mathbf{0}$
end for

estimate of the expected revenue

$$\widetilde{R}_t(S) := \frac{\sum_{i \in S} r_{ti} \exp(z_{ti})}{1 + \sum_{j \in S} \exp(z_{tj})}. \tag{4}$$

We assume an access to an assortment optimization method which returns the assortment at time t for a given parameter estimate, $S_t = \arg\max_{S \subset \mathcal{S}} \widetilde{R}_t(S)$. There are efficient polynomial-time algorithms available to solve this optimization problem (Rusmevichientong, Shen, and Shmoys 2010; Davis, Gallego, and Topaloglu 2014). We now have all the ingredients for our algorithm, UCB-MNL (see Algorithm 1).

In Algorithm 1, during the initialization phase, we first randomly choose an assortment S_t with exactly K items (after initialization, S_t can be smaller than K) to ensure a unique MLE solution. The initialization T_0, specified in Theorem 1, is chosen to ensure that $\lambda_{\min}(V_{T_0})$ is large enough.

Regret Bound for UCB-MNL Algorithm

We present the regret upper-bound of UCB-MNL under the following assumptions on the context process and the MNL model, both standard in the literature.

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

The boundedness is used to make the regret bounds scale-free. The i.i.d. assumption is also made in generalized linear bandit (Li, Lu, and Zhou 2017) and MNL contextual bandit (Chen, Wang, and Zhou 2018; Oh and Iyengar 2019) literature.

Assumption 2. *There exists $\kappa > 0$ such that for every item $i \in S$ and any $S \in \mathcal{S}$ and all round t, $\min_{\|\theta - \theta^*\| \leq 1} p_t(i|S, \theta) p_t(0|S, \theta) \geq \kappa$.*

The asymptotic normality of MLE implies the necessity of this assumption. This is a standard assumption in MNL contextual bandits (Cheung and Simchi-Levi 2017; Chen, Wang, and Zhou 2018; Oh and Iyengar 2019), which is also equivalent to the standard assumption for the link function in generalized linear contextual bandits (Filippi et al. 2010; Li, Lu, and Zhou 2017) to ensure the Fisher information matrix is invertible.

Theorem 1 (Regret of UCB-MNL)**.** *Suppose Assumptions 1 and 2 hold and we run UCB-MNL with confidence width $\alpha_t = \frac{1}{2\kappa}\sqrt{2d \log\left(1 + \frac{t}{d}\right) + 2 \log t}$ and $T_0 = \mathcal{O}(\max\{\kappa^{-2}(d \log(T/d) + 4 \log T), K/\sigma^2\})$. Then the expected regret of UCB-MNL is upper-bounded by*

$$\mathcal{R}_T = \mathcal{O}\left(d\sqrt{T \log(1 + T/d) \log(T/d)}\right).$$

Discussion of Theorem 1. In terms of key problem primitives, Theorem 1 demonstrates $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound for UCB-MNL which is independent of N; hence, it is applicable to the case with a very large number of candidate items. Chen, Wang, and Zhou (2018) established the lower bound result $\Omega(d\sqrt{T}/K)$ for MNL bandits. When K is small, which is typically true in many applications, the regret upper-bound in Theorem 1 demonstrates that UCB-MNL is almost optimal. The established regret of UCB-MNL improves the previous worst-case regret bound of Oh and Iyengar (2019) by $\sqrt{d}$ factor and that of Chen, Wang, and Zhou (2018) in both logarithmic and additive factors. Moreover, although having the same rate of $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret up to logarithmic factors, the UCB method in Chen, Wang, and Zhou (2018) has exponential computational cost, since it needs to enumerate all of the possible (N choose K) assortments. Therefore, UCB-MNL is the first polynomial-time algorithm that achieves $\widetilde{\mathcal{O}}(d\sqrt{T})$ worst-case regret.

Extension to online parameter update. UCB-MNL is simple to implement and works very well in practice. We further improve both the time and space complexities of the algorithm by using an online parameter update version (Algorithm 3 in the appendix). Exploiting the fact that the loss for the MNL model is strongly convex over bounded domain, we apply a variant of the online Newton step inspired by Hazan, Koren, and Levy (2014); Zhang et al. (2016) to find an approximate solution rather than computing the exact MLE. We show that the modified algorithm still enjoys the same order of the statistical efficiency with $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret even with the online update.

Corollary 1. *UCB-MNL with online parameter update still has $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret.*

Non-asymptotic Normality of the MLE for the MNL Model

We have shown that UCB-MNL is both statistically and computationally efficient. The algorithm also shows state-of-the-art practical performances as we report later in the numerical experiments. However, the regret bound in Theorem 1 has a linear dependence on feature dimension d and, therefore, is not very attractive when the feature vectors are high dimensional. We next investigate whether a sublinear dependence on d is possible. In the regret analysis for UCB-MNL, we upper-bound the prediction error $x^\top(\theta^* - \hat{\theta}_t)$ using Hölder's inequality, $|x^\top \hat{\theta}_t - x^\top \theta^*| \leq \|x\|_{V_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{V_t}$, where we show each of the terms on the right hand side is bounded by $\widetilde{\mathcal{O}}(\sqrt{d})$, hence resulting in a linear dependence on d when combined. A potential solution to circumvent this challenge

is to control the prediction error directly without bounding two terms separately.

In Theorem 2, we propose a non-asymptotic normality bound for the MLE for the MNL model in order to establish a sharper concentration result for $|x^\top(\hat{\theta}_t - \theta^*)|$. This is a generalization of Theorem 1 in Li, Lu, and Zhou (2017) to the MNL model. To the best of our knowledge, there was no existing finite-sample normality results for the prediction error of the utility for the MNL model. This concentration result can be of independent interest beyond the bandit problem we address in this work.

Theorem 2 (Non-asymptotic normality of MLE)**.** *Suppose we have independent responses $y_1, ..., y_n$ conditioned on feature vectors $\{x_{ti}\}_{t=1,i=1}^{n,K}$. Define $V_n = \sum_{t=1}^{n}\sum_{i\in S_t} x_{ti}x_{ti}^\top$, and let $\delta > 0$ be given. Furthermore, assume that $\lambda_{\min}(V_n) \geq \max\left\{\frac{9\mathcal{D}^4}{\kappa^4 \log(1/\delta)}, \frac{144\mathcal{D}^2}{\kappa^4}\right\}$ where $\mathcal{D} := \min\left\{4\sqrt{2d + \log\frac{1}{\delta}}, \sqrt{d\log(n/d) + 2\log\frac{1}{\delta}}\right\}$. Then, for any $x \in \mathbb{R}^d$, the maximum likelihood estimator $\hat{\theta}_n$ of the MNL model satisfies with probability at least $1 - 3\delta$ that*

$$|x^\top\hat{\theta}_n - x^\top\theta^*| \leq \frac{5}{\kappa}\sqrt{\log\frac{1}{\delta}}\|x\|_{V_n^{-1}}.$$

Hence, the prediction error can be bounded by $\widetilde{\mathcal{O}}(\sqrt{d})$ with high probability as long as the conditions on independence of samples and the minimum eigenvalue are satisfied. Note that although the statement of Theorem 2 is similar to that of the generalized linear model version in Li, Lu, and Zhou (2017), the extension to the MNL model is non-trivial because choice probability for any given item $i \in S_t$ is function of the all the items in the assortment S_t, and hence the analysis is much more involved. Theorem 2 implies that we can control the behavior of the MLE in every direction allowing us to handle the prediction error in a tighter fashion.

Provably Optimal but Impractical

Unfortunately, we cannot directly apply the tight bound for the MLE shown in Theorem 2 to `UCB-MNL` since Theorem 2 requires independent samples (as well as the minimum eigenvalue being large enough, but this condition can be satisfied by initial exploration). `UCB-MNL` is not guaranteed to produce independent samples since the algorithm chooses assortments based on previous observations, causing dependence between collected samples. This issue can be handled by generating independent samples using a framework in Auer (2002), which we denote as "Auer-framework." This Auer-framework has been previously used in several variants of (generalized) linear bandits (Chu et al. 2011; Li, Lu, and Zhou 2017; Zhou, Xu, and Blanchet 2019). We show that the adaptation of the Auer-framework to the MNL contextual bandit problem is possible[1] and establish the following regret bound.

Theorem 3 (Provably optimal regret)**.** *Suppose Assumptions 1 and 2 hold. There exists an algorithm which establishes $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret for the MNL contextual bandits.*

[1]We defer the details of the algorithm to the appendix since this is not the focus of the paper.

Algorithm 2 `DBL-MNL`

Input: sampling parameter q_k, confidence radius β_k
Set $\tau_1 \leftarrow d$, $t \leftarrow 1$, $V_0 \leftarrow \mathbf{0}_{d\times d}$
Initialization: **for** $t \in [d]$
  Randomly choose $S_t \in \mathcal{S}$ with $|S_t| = K$
  $V_t \leftarrow V_{t-1} + \sum_{i\in S_t} x_{ti}x_{ti}^\top$
for each episode $k = 2, 3, ...$ **do**
  Set the last round of k-th episode: $\tau_k \leftarrow 2^{k-1}$
  Compute MLE $\hat{\theta}_k$ by solving $\sum_{t=\tau_{k-2}+1}^{\tau_{k-1}}\sum_{i\in S_t}\big(p_t(i|S_t, \hat{\theta}_k) - y_{ti}\big)x_{ti} = \mathbf{0}$
  Update $W_{k-1} \leftarrow V_{\tau_{k-1}+1}$; Reset $V_{\tau_{k-1}+1} \leftarrow 0_{d\times d}$
  for each round $t = \tau_{k-1}+1, ..., \tau_k$ **do**
    if $\tau_k - t \leq q_k$ and $\lambda_{\min}(V_t) \leq \frac{Kq_k\sigma_0}{2}$ **then**
      Randomly choose $S_t \in \mathcal{S}$ with $|S_t| = K$
    else
      Offer $S_t = \text{argmax}_{S\in\mathcal{S}}\widetilde{R}_t(S)$
    end if
    Update $V_{t+1} \leftarrow V_t + \sum_{i\in S_t} x_{ti}x_{ti}^\top$
  end for
end for

$\Omega(\sqrt{NT})$ lower bound was shown in Chen and Wang (2017) for the non-contextual MNL bandits. This lower bound can be translated to $\Omega(\sqrt{dT})$ if each item is represented as one-hot encoding. Hence the regret bound in Theorem 3 matches the lower bound for the MNL bandit problem with finite items. To our knowledge, this is the first result that achieves the rate of $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret and establishes the provable optimality in the MNL contextual bandit problem. However, this comes at a cost. The algorithm based on the Auer-framework, although provably optimal, is not practical (see the numerical experiments)! In fact, this is true for *all* optimal methods (Chu et al. 2011; Li, Lu, and Zhou 2017; Zhou, Xu, and Blanchet 2019) that rely on the Auer-framework (Auer 2002) because the framework wastes too many samples with random exploration.[2] Next, we investigate whether $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret can be achieved in a practical manner for the class of applications where the revenue for each item is uniform. As discussed earlier that this class includes web search and media recommendations.

Algorithm: `DBL-MNL`

We propose a new algorithm, `DBL-MNL` (Algorithm 2) that is *both* provably optimal and practical. `DBL-MNL` operates in an episodic manner. At the beginning of each episode, the MLE is computed using the samples from a previous episode. Within an episode, the parameter is not updated, but the algorithm takes an UCB action based on the parameter computed at the beginning of the episode. In particular, for round t in the k-th episode, the upper-bound of an utility

[2]These previous methods (Chu et al. 2011; Li, Lu, and Zhou 2017; Zhou, Xu, and Blanchet 2019) that use techniques in (Auer 2002) do not provide numerical evaluations.

estimate is computed as

$$\widetilde{z}_{ti} = x_{ti}^\top \hat{\theta}_k + \alpha_k \|x_{ti}\|_{W_{k-1}^{-1}}$$

$$\text{where } W_{k-1} = \sum_{t'=\tau_{k-1}+1}^{\tau_{k-1}} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top$$

and τ_{k-1} is the last round of the $k-1$-th episode. Note that the Gram matrix resets every episode. Under this action selection, samples within each episode are independent of each other. Episode lengths are doubled over time such that the length of the k-th episode is twice as large as the $k-1$-th episode. This doubling technique is inspired by Jaksch, Ortner, and Auer (2010); Javanmard and Nazerzadeh (2019). Towards the end of each episode, the algorithm checks whether $\lambda_{\min}(V_t)$ is suitably large. If not, it performs random exploration. Since episode lengths are growing exponentially and the threshold for $\lambda_{\min}(V_t)$ is only logarithmic in t, even in the worst case, the algorithm draws $\mathcal{O}(\log T)$ random samples. Note that the algorithm may not even take these exploratory actions since $\lambda_{\min}(V_t)$ may already surpass the threshold for large enough episodes (this is clearly observed in numerical evaluations). This makes DBL-MNL much more practical since it would perform minimal random exploration. Furthermore, the algorithm is computationally efficient with only logarithmic number of parameter updates instead of updating in every period.

Regret Bound of DBL-MNL

We analyze the regret of DBL-MNL for which we aim to establish $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. For our analysis, we add the following mild assumption which encompasses many canonical distributions.

Assumption 3 (Relaxed symmetry)**.** *For a joint distribution p_X, there exists $\rho_0 < \infty$ such that $\frac{p_X(-x)}{p_X(x)} \leq \rho_0$ for all x.*

This assumption is also used in the analysis of sparse bandits Oh, Iyengar, and Zeevi (2020). Assumption 3 states that the joint distribution $p_{\mathcal{X}}$ can be skewed but this skewness is bounded. For symmetrical distributions, $\rho_0 = 1$. One can see that a large class of continuous and discrete distributions satisfy Assumption 3, e.g., Gaussian, truncated Gaussian, uniform distribution, and Rademacher distribution, and many more. Under this suitable regularity, we establish the following regret bound for DBL-MNL.

Theorem 4 (Regret bound of DBL-MNL)**.** *Suppose Assumptions 1-3 hold and the revenue $r_i \equiv r$ is uniform. Then the expected regret of DBL-MNL over horizon T is $\mathcal{R}_T = \mathcal{O}\big(\sqrt{dT \log(T/d) \log(TN) \log(T)}\big)$.*

Discussion of Theorem 4. DBL-MNL achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret when the revenue for each item is uniform. This encompasses all applications where the goal is to maximize an appropriate "click-through rate" from offering the assortment. Theorem 4 provides insights beyond the MNL contextual bandits: it shows that under the suitable regularity condition, it is possible for a practical algorithm to attain $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. We expect this technique to yield practical provably optimal algorithms for other variants of contextual bandit problems. The regret bound of UCB-MNL is N independent; in contrast, DBL-MNL has a logarithmic dependence on N (as is common for $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret algorithms). In fact, the numerical experiments suggest that performance does have at least logarithmic dependence on N for all methods (as indicated by Theorem 4 for DBL-MNL).

Proof Outline of Theorem 4

Since the length of an episode grows exponentially, the number of episodes up to round T is logarithmic in T. In particular, the T-th round belongs to the L-th episode with $L = \lfloor \log_2 T \rfloor + 1$. Let $\mathcal{T}_k := \{\tau_{k-1}+1, ..., \tau_k\}$ denote an index set of rounds that belong to the k-th episode. Note that the length of the k-th episode is $|\mathcal{T}_k| = \tau_k/2$. Then, we let $\texttt{Reg}(k\text{-th episode})$ denote the cumulative regret of the k-th episode, i.e.,

$$\texttt{Reg}(k\text{-th episode}) := \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \Big(R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*)\Big)\right]$$

so that the cumulative expected regret over T rounds is $\mathcal{R}(T) = \sum_{k=1}^{L} \texttt{Reg}(k\text{-th episode})$. Therefore, it suffices to bound each $\texttt{Reg}(k\text{-th episode})$. Now, for each episode $k \in [L]$, we consider the following two cases.

(i) $|\mathcal{T}_k| \leq q_k$: In this case, the length of an episode is not large enough to have the concentration of the prediction error due to the failure of ensuring the lower bound on $\lambda_{\min}(V_t)$. Therefore, we cannot control the regret in this case. However, the total number of such rounds is only logarithmic in T, hence the regret corresponding to this case contributes minimally to the total regret.

(ii) $|\mathcal{T}_k| > q_k$: We can apply the fast convergence result in Theorem 2 as long as the lower bound on $\lambda_{\min}(V_t)$ is guaranteed — note that the independence condition is already satisfied since samples in each episode are independent of each other. We show that $\lambda_{\min}(V_t)$ grows linearly as t increases in each episode with high probability. In case of $\lambda_{\min}(V_t)$ not growing as fast as the rate we require, we perform random sampling to satisfy this criterion towards the end of each episode. Therefore, with high probability, the lower bound on $\lambda_{\min}(V_t)$ is satisfied.

For case (i), clearly $q_k \leq q_L$ for any $k \in \{1, ..., L\}$. $|\mathcal{T}_k|$ eventually grows to be larger than q_L for some k since q_L is logarithmic in T. Let k' be the first episode such that $|\mathcal{T}_{k'}| \geq q_L$. Hence, $|\mathcal{T}_{k'}| \leq 2q_L$. Thus, the cumulative regret prior to the k'-th episode is $\mathcal{O}\big(\log d + d^2 + \log^2(TN)\big)$. Then, letting k'' be the first episode such that $|\mathcal{T}_{k''}| \geq q_{k''}$ and noting that $k'' \leq k'$ gives

$$\sum_{k=1}^{k''-1} \texttt{Reg}(k\text{-th episode}) \leq \sum_{k=1}^{k'-1} \texttt{Reg}(k\text{-th episode}).$$

Hence, the cumulative regret corresponding to case (i) is at most poly-logarithic in T.

For case (ii), it suffices to show random sampling ensures the growth of $\lambda_{\min}(V_t)$. We show that random sampling with duration q_k specified in Theorem 4 ensures the minimum eigenvalue condition for the Gram matrix, i.e., $\lambda_{\min}(V_{\tau_k}) \geq$

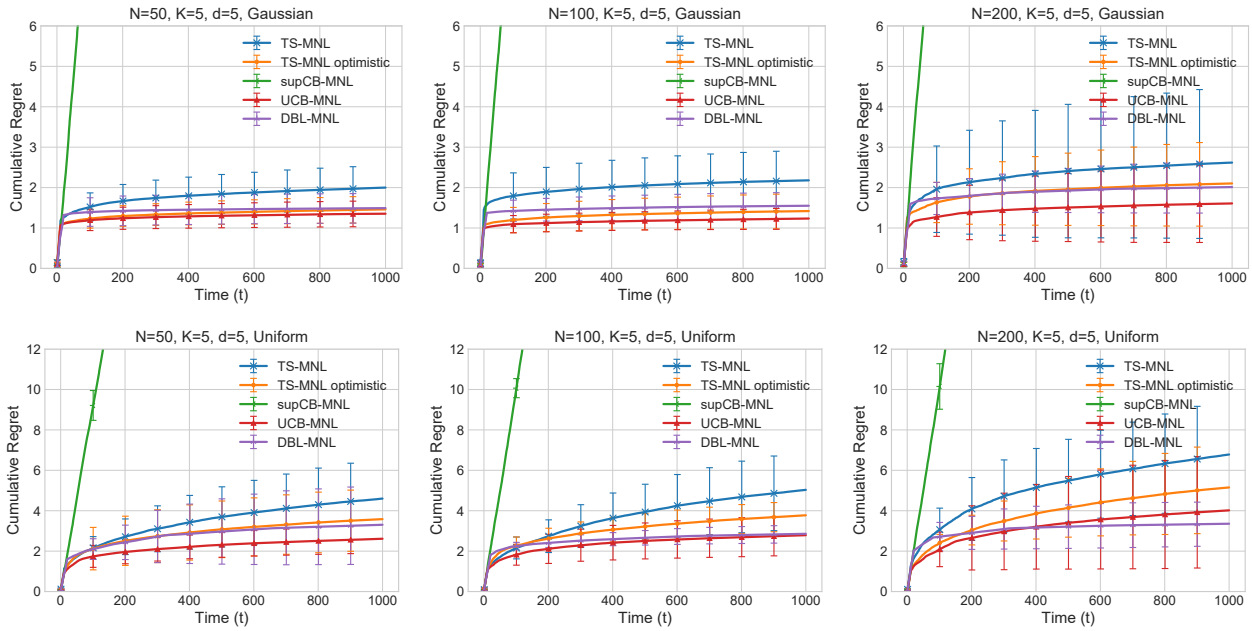

Figure 1: The regret plots show that UCB-MNL and DBL-MNL perform at start-of-the-art levels across different problem instances. Evaluations are for features drawn from a multivariate Gaussian (first row) and uniform (second row) distributions.

$\max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$ with high probability for each episode $k \in [L]$. We then apply the confidence bound in Theorem 2 to the k-th episode which requires samples in the $(k-1)$-th episode are independent and $\lambda_{\min}(V_{\tau_{k-1}})$ at the end of the $(k-1)$-th episode is large enough. That is, with a lower bound guarantee on $\lambda_{\min}(V_{\tau_{k-1}})$ and the fact that samples are independent of each other in each episode, we have with high probability

$$|x_{ti}^\top(\hat{\theta}_k - \theta^*)| \leq \beta_k \|x_{ti}\|_{W_{k-1}^{-1}}, \quad \forall i \in [N], \forall t \in \mathcal{T}_k$$

with suitable confidence width β_k specified in Theorem 4. Therefore, the expected regret in the k-th episode can be bounded by $\widetilde{\mathcal{O}}(\sqrt{d\tau_k})$. Then we combine the results over all episodes to establish $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret.

Numerical Experiments

In this section, we evaluate the performances of our proposed algorithms: UCB-MNL (Algorithm 1) and DBL-MNL (Algorithm 2) in numerical experiments. In our evaluations, we report the cumulative regret for each round $t \in \{1, ..., T\}$. For each experimental configuration, we evaluate the algorithms on 20 independent instances and report average performances. In each instance, the underlying parameter θ^* is sampled from the d-dimensional uniform distribution, with each element of θ^* uniformly distributed in $[0, 1]$. The underlying parameters are fixed during each problem instance but not known to the algorithms. For efficient evaluations, we consider uniform revenues, i.e., $r_{ti} = 1$ for all i and t. Therefore, the combinatorial optimization step to solve for the optimal assortment reduces to sorting items according to their utility estimate. Also, recall that the regret bound for DBL-MNL (Theorem 4) is derived under the uniform revenue assumption, therefore, the uniform revenue setting provides a suitable test bed for all methods considered in this section.

Method	Horizon (T) 1000	5000
TS-MNL (Oh and Iyengar 2019)	6.65	73.99
TS-MNL Opt. (Oh and Iyengar 2019)	6.81	77.18
UCB-MNL (Algorithm 1)	6.62	74.28
DBL-MNL (Algorithm 2)	**1.20**	**5.92**

Table 2: Runtime evaluation (sec), $N = 100, K = 5, d = 5$

We compare the performances of the proposed algorithms with those of the state-of-the-art Thompson sampling based algorithms, TS-MNL and "optimistic" TS-MNL, proposed in Oh and Iyengar (2019). Additionally, we evaluate the performance of the provably optimal but impractical algorithm, supCB-MNL (see Algorithm 5 in the appendix), that is based on the Auer-framework. Figure 1 shows that the performances of UCB-MNL and DBL-MNL are superior to or comparable to the state-of-the-art Thompson sampling methods. Moreover, the runtime evaluation shows that DBL-MNL is significantly faster than the other methods due to its logarithmic number of parameter updates.

Ethical Statement

We conform that our work meets the standards listed in the ethics and malpractice statement of the AAAI.

References

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

Abe, N.; and Long, P. M. 1999. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, 3–11.

Agrawal, S.; Avadhanula, V.; Goyal, V.; and Zeevi, A. 2017. Thompson Sampling for the MNL-Bandit. In *Conference on Learning Theory*, 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.

Aouad, A.; Levi, R.; and Segev, D. 2018. Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences. *Operations Research* 66(5): 1321–1345.

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

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

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

Chen, X.; and Wang, Y. 2017. A Note on Tight Lower Bound for MNL-Bandit Assortment Selection Models. *arXiv preprint arXiv:1709.06109* .

Chen, X.; Wang, Y.; and Zhou, Y. 2018. Dynamic Assortment Optimization with Changing Contextual Information. *arXiv preprint arXiv:1810.13069* .

Cheung, W. C.; and Simchi-Levi, D. 2017. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658* .

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

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, 355–366.

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

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

Hazan, E.; Koren, T.; and Levy, K. Y. 2014. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, 197–209.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr): 1563–1600.

Javanmard, A.; and Nazerzadeh, H. 2019. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research* 20(1): 315–363.

Kveton, B.; Szepesvari, C.; Wen, Z.; and Ashkan, A. 2015. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, 767–776.

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

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

Li, L.; Lu, Y.; and Zhou, D. 2017. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *International Conference on Machine Learning*, 2071–2080.

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

Oh, M.-h.; and Iyengar, G. 2019. Thompson Sampling for Multinomial Logit Contextual Bandits. In *Advances in Neural Information Processing Systems*, 3145–3155.

Oh, M.-h.; Iyengar, G.; and Zeevi, A. 2020. Sparsity-agnostic lasso bandit. *arXiv preprint arXiv:2007.08477* .

Ou, M.; Li, N.; Zhu, S.; and Jin, R. 2018. Multinomial Logit Bandit with Linear Utility Functions. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 2602–2608. AAAI Press.

Qin, L.; Chen, S.; and Zhu, X. 2014. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, 461–469. SIAM.

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.

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

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

Wen, Z.; Kveton, B.; and Ashkan, A. 2015. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, 1113–1122.

Zhang, L.; Yang, T.; Jin, R.; Xiao, Y.; and Zhou, Z.-H. 2016. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, 392–401.

Zhou, Z.; Xu, R.; and Blanchet, J. 2019. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, 5198–5209.

Zong, S.; Ni, H.; Sung, K.; Ke, N. R.; Wen, Z.; and Kveton, B. 2016. Cascading Bandits for Large-scale Recommendation Problems. In *Proceedings of the Thirty-Second Confer-*

ence on Uncertainty in Artificial Intelligence, UAI'16, 835–844.