SUBMITTED TO OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

MNL-Bandit: A Dynamic Learning Approach to Assortment Selection

Shipra Agrawal
Industrial Engineering and Operations Research, Columbia University, New York, NY. sa3305@columbia.edu

Vashist Avadhanula
Decision Risk and Operations, Columbia Business School, New York, NY. vavadhanula18@gsb.columbia.edu

Vineet Goyal
Industrial Engineering and Operations Research, Columbia University, New York, NY. vg2277@columbia.edu

Assaf Zeevi
Decision Risk and Operations, Columbia Business School, New York, NY. assaf@gsb.columbia.edu

We consider a dynamic assortment selection problem, where in every round the retailer offers a subset (assortment) of N substitutable products to a consumer, who selects one of these products according to a multinomial logit (MNL) choice model. The retailer observes this choice and the objective is to dynamically learn the model parameters, while optimizing cumulative revenues over a selling horizon of length T. We refer to this exploration-exploitation formulation as the *MNL-Bandit problem*. Existing methods for this problem follow an *explore-then-exploit* approach, which estimate parameters to a desired accuracy and then, treating these estimates as if they are the correct parameter values, offers the optimal assortment based on these estimates. These approaches require certain a priori knowledge of "separability," determined by the true parameters of the underlying MNL model, and this in turn is critical in determining the length of the exploration period. (Separability refers to the distinguishability of the true optimal assortment from the other sub-optimal alternatives.) In this paper, we give an efficient algorithm that *simultaneously* explores and exploits, without a priori knowledge of any problem parameters. Furthermore, the algorithm is adaptive in the sense that its performance is near-optimal in both the "well separated" case, as well as the general parameter setting where this separation need not hold.

Key words: Exploration-Exploitation, assortment optimization, upper confidence bound, multinomial logit

1. Introduction

1.1. Overview of the problem

Assortment optimization problems arise widely in many industries including retailing and online advertising where the seller needs to select a subset from a universe of substitutable items with the objective of maximizing expected revenue. Choice models capture substitution effects among products by specifying the probability that a consumer selects a product from the offered set.

Traditionally, assortment decisions are made at the start of the selling period based on a choice model that has been estimated from historical data; see Kök and Fisher (2007) for a detailed review.

In this work, we focus on the dynamic version of the problem where the retailer needs to simultaneously learn consumer preferences and maximize revenue. In many business applications such as fast fashion and online retail, new products can be introduced or removed from the offered assortments in a fairly frictionless manner and the selling horizon for a particular product can be short. Therefore, the traditional approach of first estimating the choice model and then using a static assortment based on the estimates, is not practical in such settings. Rather, it is essential to experiment with different assortments to learn consumer preferences, while simultaneously attempting to maximize immediate revenues. Suitable balancing of this exploration-exploitation tradeoff is the focal point of this paper.

We consider a stylized dynamic optimization problem that captures some salient features of this application domain, where our goal is to develop an exploration-exploitation policy that simultaneously learns from current observations and exploits this information gain for future decisions. In particular, we consider a constrained assortment selection problem under the Multinomial logit (MNL) model with N substitutable products and a "no purchase" option. Our goal is to offer a sequence of assortments, $S_1, \ldots, S_T$, where T is the planning horizon, such that the cumulative expected revenues over said horizon is maximized, or alternatively, minimizing the gap between the performance of a proposed policy and that of an oracle that knows instance parameters a priori, a quantity referred to as the *regret*.

Related literature. The Multinomial Logit model (MNL), owing primarily to its tractability, is the most widely used choice model for assortment selection problems. (The model was introduced independently by Luce (1959) and Plackett (1975), see also Train (2009), McFadden (1978), Ben-Akiva and Lerman (1985) for further discussion and survey of other commonly used choice models.) If the consumer preferences (MNL parameters in our setting) are known a priori, then the problem of computing the optimal assortment, which we refer to as the *static assortment optimization problem*, is well studied. Talluri and van Ryzin (2004) consider the unconstrained assortment planning problem under the MNL model and present a greedy approach to obtain the optimal assortment. Recent works of Davis et al. (2013) and Désir and Goyal (2014) consider assortment planning problems under MNL with various constraints. Other choice models such as Nested Logit (Williams 1977, Davis et al. 2014, Gallego and Topaloglu 2014 and Li et al. 2015), Markov Chain (Blanchet et al. 2016 and Désir et al. 2015) and more general models (Farias et al. 2013 and Gallego et al. 2014) are also considered in the literature.

Most closely related to our work are the papers of Caro and Gallien (2007), Rusmevichientong et al. (2010) and Sauré and Zeevi (2013), where information on consumer preferences is not known and needs to be learned over the course of the selling horizon. Caro and Gallien (2007) consider the setting under which demand for products is independent of each other. Rusmevichientong et al. (2010) and Sauré and Zeevi (2013) consider the problem of minimizing regret under the MNL choice model and present an "explore first and exploit later" approach. In particular, a selected set of assortments are explored until parameters can be estimated to a desired accuracy and then the optimal assortment corresponding to the estimated parameters is offered for the remaining selling horizon. The exploration period depends on certain a priori knowledge about instance parameters. Assuming that the optimal and next-best assortment are "well separated," Sauré and Zeevi (2013) show an asymptotic $O(N \log T)$ regret bound, while Rusmevichientong et al. (2010) establish a $O(N^2 \log^2 T)$ regret bound; recall N is the number of products and T is the time horizon. However, their algorithm relies crucially on a priori knowledge of system parameters which is not readily available in practice. As will be illustrated later, absence of this knowledge, these algorithms can perform quite poorly. In this work, we focus on approaches that simultaneously explore and exploit demand information, do not require any a priori knowledge or assumptions, and whose performance is in some sense best possible; thereby, making our approach more universal in its scope.

Our problem is closely related to the multi-armed bandit (MAB) paradigm (cf. Robbins 1952). A naive mapping to that setting would consider every assortment as an arm, and as such, given the combinatorial nature of the problem would lead to exponentially many arms. Popular extensions of MAB for large scale problems include the linear bandit (e.g., Auer 2003, Rusmevichientong and Tsitsiklis 2010) and generalized linear bandit (Filippi et al. 2010) formulations. However, these do not apply directly to our problem, since the revenue corresponding to an assortment is nonlinear in problem parameters. Other works (see Chen et al. 2013) have considered versions of MAB where one can play a subset of arms in each round and the expected reward is a function of rewards for the arms played. However, this approach assumes that the reward for each arm is generated independently of the other arms in the subset. This is not the case typically in retail settings, and in particular, in the MNL choice model where purchase decisions depend on the assortment of products offered in a time step. In this work, we use the structural properties of the MNL model, along with techniques from MAB literature, to optimally explore and exploit in the presence of a large number of alternatives (assortments).

1.2. Contributions

Parameter independent online algorithm and regret bounds. We give an efficient online algorithm that judiciously balances the exploration and exploitation trade-off intrinsic to our problem and achieves a worst-case regret bound of $O(\sqrt{NT \log NT})$ under a mild assumption, namely

that the no-purchase is the most "frequent" outcome. The assumption regarding no-purchase is quite natural and a norm in online retailing for example. To the best of our knowledge, this is the first such policy with provable regret bounds that does not require prior knowledge of instance parameters of the MNL choice model. Moreover, the regret bound we present for this algorithm is non-asymptotic. The "big-oh" notation is used for brevity and only hides absolute constants.

We also show that for "well separated" instances, the regret of our policy is bounded by $O\left(\min\left(N^2 \log NT/\Delta, \sqrt{NT \log NT}\right)\right)$ where Δ is the "separability" parameter. This is comparable to the regret bounds, $O\left(N \log T/\Delta\right)$ and $O\left(N^2 \log^2 T/\Delta\right)$, established in Sauré and Zeevi (2013) and Rusmevichientong et al. (2010) respectively, even though we do not require any prior information on Δ unlike the aforementioned work. It is also interesting to note that the regret bounds hold true for a large class of constraints, e.g., we can handle matroid constraints such as assignment, partition and more general totally unimodular constraints (see Davis et al. 2013). Our algorithm is predicated on upper confidence bound (UCB) type logic, originally developed to balance the aforementioned exploration-exploitation trade-off in the context of the multi-armed bandit (MAB) problem (cf. Lai and Robbins 1985). In this paper the UCB approach, also known as optimism in the face of uncertainty, is customized to the assortment optimization problem under the MNL model.

Lower bounds. We establish a non-asymptotic lower bound for the online assortment optimization problem under the MNL model. In particular, we show that for the cardinality constrained problem under the MNL model, any algorithm must incur a regret of $\Omega(\sqrt{NT/K})$, where K is the bound on the number of products that can be offered in an assortment. This bound is derived via a reduction to a parametric multi-armed bandit problem, for which such lower bounds are constructed by means of information theoretic arguments. This result establishes that our online algorithm is nearly optimal, the upper bound being within a factor of $\sqrt{K}$ of the lower bound. A recent work by Chen and Wang (2017) demonstrates a lower bound of $\Omega(\sqrt{NT})$ for the MNL-Bandit problem, thus suggesting that our algorithm's performance is optimal even with respect to its dependence on K.

Computational study. We present a computational study that highlights several salient features of our algorithm. In particular, we test the performance of our algorithm over instances with varying degrees of separability between optimal and sub-optimal solutions and observe that the performance is bounded irrespective of the "separability parameter." In contrast, the approach of Sauré and Zeevi (2013) "breaks down" and results in linear regret for some values of the "separability parameter." We also present results of a simulated study on a real world data set, where we compare the performance of our algorithm to that of Sauré and Zeevi (2013). We observe that

the performance of our algorithm is sub-linear, while the performance of Sauré and Zeevi (2013) is linear, which further emphasizes the limitations of "explore first and exploit later" approaches and highlights the universal applicability of our approach.

Outline. In Section 2, we give the precise problem formulation. In Section 3, we present our algorithm for the MNL-Bandit problem, and in Section 4, we prove the worst-case regret bound of $\tilde{O}(\sqrt{NT})$ for our policy. In Section 5, we present our non-asymptotic lower bound on regret for any algorithm for MNL-Bandit. In Section 6, we present two extensions including improved logarithmic regret bound for "well-separated" instances and regret bound when the "no purchase" assumption is relaxed. In Section 7, we present results from our computational study.

2. Problem formulation

The basic assortment problem. In our problem, at every time instance t, the seller selects an assortment $S_t \subset \{1,\dots,N\}$ and observes the customer purchase decision $c_t \in S_t \cup \{0\}$, where $\{0\}$ denotes the no-purchase alternative, which is always available for the consumer. As noted earlier, we assume consumer preferences are modeled using a multinomial logit (MNL) model. Under this model, the probability that a consumer purchases product i at time t when offered an assortment $S_t = S \subset \{1,\dots,N\}$ is given by,

$$p_i(S) := \mathbb{P}\left(c_t = i | S_t = S\right) = \begin{cases} \dfrac{v_i}{v_0 + \sum_{j\in S} v_j}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

for all t, where v_i is the *attraction parameter* for product i in the MNL model. The random variables $\{c_t : t = 1,2,\dots\}$ are conditionally independent, namely, c_t conditioned on the event $\{S_t = S\}$ is independent of $c_1,\dots,c_{t-1}$. Without loss of generality, we can assume that $v_0 = 1$. It is important to note that the parameters of the MNL model v_i, are not known to the seller. From (2.1), the expected revenue when assortment S is offered and the MNL parameters are denoted by the vector $\mathbf{v}$ is given by

$$R(S,\mathbf{v}) = \mathbb{E}\left[\sum_{i\in S} r_i \mathbb{1}\{c_t = i | S_t = S\}\right] = \sum_{i\in S} \frac{r_i v_i}{1 + \sum_{j\in S} v_j}, \tag{2.2}$$

where r_i is the revenue obtained when product i is purchased and is known a priori.

We consider several naturally arising constraints over the assortments that the retailer can offer. These include cardinality constraints (where there is an upper bound on the number of products that can be offered in the assortment), partition matroid constraints (where the products are partitioned into segments and the retailer can select at most a specified number of products from each segment) and joint display and assortment constraints (where the retailer needs to decide both the assortment as well as the display segment of each product in the assortment and there is an

upper bound on the number of products in each display segment). More generally, we consider the set of totally unimodular (TU) constraints on the assortments. Let $\mathbf{x}(S) \in \{0,1\}^N$ be the incidence vector for assortment $S \subseteq \{1,\dots,N\}$, i.e., $x_i(S) = 1$ if product $i \in S$ and 0 otherwise. We consider constraints of the form

$$\mathcal{S} = \{S \subseteq \{1,\dots,N\} \mid A\,\mathbf{x}(S) \leq \mathbf{b},\, \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}\}, \tag{2.3}$$

where $\mathbf{A}$ is a totally unimodular matrix and $\mathbf{b}$ is integral (i.e., each component of the vector $\mathbf{b}$ is an integer). The totally unimodular constraints model a rich class of practical assortment planning problems including the examples discussed above. We refer the reader to Davis et al. (2013) for a detailed discussion on assortment and pricing optimization problems that can be formulated under the TU constraints.

Admissible Policies. To define the set of policies that can be used by the seller, let U be a random variable, which encodes any additional sources of randomization and $(\mathbb{U},\mathcal{U},\mathbb{P}_u)$ be the corresponding probability space. We define $\{\pi_t,\, t = 1,2,\dots\}$ to be measurable mappings as follows:

$$\pi_1 : \mathbb{U} \to \mathcal{S}$$
$$\pi_t : \mathbb{U} \times \mathcal{S}^{t-1} \times \{0,\dots,N\}^{t-1} \to \mathcal{S}, \text{ for each } t = 2,3,\dots,$$

where $\mathcal{S}$ is as defined in (2.3). Then the assortment selection for the seller at time t is given by

$$S_t = \begin{cases} \pi_1(U), & t = 1 \\ \pi_t(U, c_1, \dots, c_{t-1}, S_1, \dots, S_{t-1}), & t = 2,3,\dots. \end{cases} \tag{2.4}$$

For further reference, let $\{\mathcal{H}_t : t = 1,2,\dots\}$ denote the filtration associated with the policy $\pi = (\pi_1, \pi_2, \dots, \pi_t, \dots)$. Specifically,

$$\mathcal{H}_1 = \sigma(U)$$
$$\mathcal{H}_t = \sigma(U, c_1, \dots, c_{t-1}, S_1, \dots, S_{t-1}), \text{ for each } t = 2,3,\dots.$$

We denote by $\mathbb{P}_\pi\{.\}$ and $\mathbb{E}_\pi\{.\}$ the probability distribution and expectation value over path space induced by the policy π.

The online assortment optimization problem. The objective is to design a policy $\pi = (\pi_1,\dots,\pi_T)$ that selects a sequence of history dependent assortments $(S_1, S_2, \dots, S_T)$ so as to maximize the cumulative expected revenue,

$$\mathbb{E}_\pi\left(\sum_{t=1}^{T} R(S_t, \mathbf{v})\right), \tag{2.5}$$

where $R(S,\mathbf{v})$ is defined as in (2.2). Direct analysis of (2.5) is not tractable given that the parameters $\{v_i, i = 1,\dots,N\}$ are not known to the seller a priori. Instead we propose to measure the

performance of a policy π via its *regret*. The objective then is to design a policy that approximately minimizes the *regret* defined as

$$Reg_\pi(T,\mathbf{v}) = \sum_{t=1}^{T} R(S^*,\mathbf{v}) - \mathbb{E}_\pi[R(S_t,\mathbf{v})], \tag{2.6}$$

where S^* is the optimal assortment for (2.2), namely, $S^* = \underset{S\in\mathcal{S}}{\operatorname{argmax}}\, R(S,\mathbf{v})$. This exploration-exploitation problem, which we refer to as **MNL-Bandit**, is the focus of this paper.

3. The proposed policy

In this section, we describe our proposed policy for the MNL-Bandit problem. The policy is designed using the characteristics of the MNL model based on the principle of optimism under uncertainty.

3.1. Challenges and overview

A key difficulty in applying standard multi-armed bandit techniques to this problem is that the response observed on offering a product i is *not* independent of other products in assortment S. Therefore, the N products cannot be directly treated as N independent arms. Our policy utilizes the specific properties of the dependence structure in MNL model to obtain an efficient algorithm with order $\sqrt{NT}$ regret.

Our policy is based on a non-trivial extension of the UCB algorithm in Auer et al. (2002), which is predicated on Lai and Robbins (1985). It uses the past observations to maintain increasingly accurate upper confidence bounds for the MNL parameters $\{v_i, i = 1,\dots,N\}$, and uses these to (implicitly) maintain an estimate of expected revenue $R(S,\mathbf{v})$ for every feasible assortment S. In every round, our policy picks the assortment S with the highest optimistic revenue. There are two main challenges in implementing this scheme. First, the customer response to being offered an assortment S depends on the entire set S, and does not directly provide an unbiased sample of demand for a product $i \in S$. In order to obtain unbiased estimates of v_i for all $i \in S$, we offer a set S multiple times: specifically, it is offered repeatedly until a no-purchase occurs. We show that proceeding in this manner, the average number of times a product i is purchased provides an unbiased estimate of the parameter v_i. The second difficulty is the computational complexity of maintaining and optimizing revenue estimates for each of the exponentially many assortments. To this end, we use the structure of the MNL model and define our revenue estimates such that the assortment with maximum estimated revenue can be efficiently found by solving a simple optimization problem. This optimization problem turns out to be a static assortment optimization problem with upper confidence bounds for v_i's as the MNL parameters, for which efficient solution methods are available.

3.2. Details of the policy

We divide the time horizon into epochs, where in each epoch we offer an assortment repeatedly until a no purchase outcome occurs. Specifically, in each epoch ℓ, we offer an assortment S_ℓ repeatedly. Let $\mathcal{E}_\ell$ denote the set of consecutive time steps in epoch ℓ. $\mathcal{E}_\ell$ contains all time steps after the end of epoch $\ell-1$, until a no-purchase happens in response to offering S_ℓ, including the time step at which no-purchase happens. The length of an epoch $|\mathcal{E}_\ell|$ conditioned on S_ℓ is a geometric random variable with success probability defined as the probability of no-purchase in S_ℓ. The total number of epochs L in time T is implicitly defined as the minimum number for which $\sum_{\ell=1}^{L}|\mathcal{E}_\ell| \geq T$.

At the end of every epoch ℓ, we update our estimates for the parameters of MNL, which are used in epoch $\ell+1$ to choose assortment $S_{\ell+1}$. For any time step $t \in \mathcal{E}_\ell$, let c_t denote the consumer's response to S_ℓ, i.e., $c_t = i$ if the consumer purchased product $i \in S_\ell$, and 0 if no-purchase happened. We define $\hat{v}_{i,\ell}$ as the number of times a product i is purchased in epoch ℓ,

$$\hat{v}_{i,\ell} := \sum_{t\in\mathcal{E}_\ell} \mathbb{1}(c_t = i). \tag{3.1}$$

For every product i and epoch $\ell \leq L$, we keep track of the set of epochs before ℓ that offered an assortment containing product i, and the number of such epochs. We denote the set of epochs by $\mathcal{T}_i(\ell)$ and the number of epochs by $T_i(\ell)$. That is,

$$\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\tau\}, \quad T_i(\ell) = |\mathcal{T}_i(\ell)|. \tag{3.2}$$

We compute $\bar{v}_{i,\ell}$ as the number of times product i was purchased per epoch,

$$\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau\in\mathcal{T}_i(\ell)} \hat{v}_{i,\tau}. \tag{3.3}$$

We show that for all $i \in S_\ell$, $\hat{v}_{i,\ell}$ and $\bar{v}_{i,\ell}$ are unbiased estimators of the MNL parameter v_i (see Corollary A.1) Using these estimates, we compute the upper confidence bounds, $v_{i,\ell}^{\mathsf{UCB}}$ for v_i as,

$$v_{i,\ell}^{\mathsf{UCB}} := \bar{v}_{i,\ell} + \sqrt{\bar{v}_{i,\ell}\frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}. \tag{3.4}$$

We establish that $v_{i,\ell}^{\mathsf{UCB}}$ is an upper confidence bound on the true parameter v_i, i.e., $v_{i,\ell}^{\mathsf{UCB}} \geq v_i$, for all i,ℓ with high probability (see Lemma 4.1). The role of the upper confidence bounds is akin to their role in hypothesis testing; they ensure that the likelihood of identifying the parameter value is sufficiently large. We then offer the optimistic assortment in the next epoch, based on the previous updates as follows,

$$S_{\ell+1} := \operatorname*{argmax}_{S\in\mathcal{S}} \max\left\{R(S,\hat{\mathbf{v}}) : \hat{v}_i \leq v_{i,\ell}^{\mathsf{UCB}}\right\}, \tag{3.5}$$

where $R(S,\hat{\mathbf{v}})$ is as defined in (2.2). We later show that the above optimization problem is equivalent to the following optimization problem (see Lemma A.3).

$$S_{\ell+1} := \operatorname*{argmax}_{S\in\mathcal{S}} \tilde{R}_{\ell+1}(S), \tag{3.6}$$

where $\tilde{R}_{\ell+1}(S)$ is defined as,

$$\tilde{R}_{\ell+1}(S) := \frac{\sum_{i\in S} r_i v_{i,\ell}^{\mathsf{UCB}}}{1+\sum_{j\in S} v_{j,\ell}^{\mathsf{UCB}}}. \tag{3.7}$$

We summarize the steps in our policy in Algorithm 1. Finally, we may remark on the computa-

Algorithm 1 Exploration-Exploitation algorithm for MNL-Bandit

Initialization: $v_{i,0}^{\mathsf{UCB}} = 1$ for all $i = 1,\dots,N$
$t = 1$; $\ell = 1$ keeps track of the time steps and total number of epochs respectively
while $t < T$ **do**
  Compute $S_\ell = \operatorname*{argmax}_{S\in\mathcal{S}} \tilde{R}_\ell(S) = \dfrac{\sum_{i\in S} r_i v_{i,\ell-1}^{\mathsf{UCB}}}{1+\sum_{j\in S} v_{j,\ell-1}^{\mathsf{UCB}}}$
  Offer assortment S_ℓ, observe the purchasing decision, c_t of the consumer
  if $c_t = 0$ **then**
    compute $\hat{v}_{i,\ell} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}(c_t = i)$, no. of consumers who preferred i in epoch ℓ, for all $i \in S_\ell$
    update $\mathcal{T}_i(\ell) = \{\tau \le \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until ℓ that offered product i
    update $\bar{v}_{i,\ell} = \dfrac{1}{T_i(\ell)} \sum_{\tau\in\mathcal{T}_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
    update $v_{i,\ell}^{\mathsf{UCB}} = \bar{v}_{i,\ell} + \sqrt{\bar{v}_{i,\ell} \dfrac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \dfrac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}$; $\ell = \ell+1$
  else
    $\mathcal{E}_\ell = \mathcal{E}_\ell \cup t$, time indices corresponding to epoch ℓ
  end if
  $t = t+1$
end while

tional complexity of implementing (3.5). The optimization problem (3.5) is formulated as a static assortment optimization problem under the MNL model with TU constraints, with model parameters being $v_{i,\ell}^{\mathsf{UCB}}, i = 1,\dots,N$ (see (3.6)). There are efficient polynomial time algorithms to solve the static assortment optimization problem under MNL model with known parameters (see Avadhanula et al. 2016, Davis et al. 2013, Rusmevichientong et al. 2010). We will now briefly comment

on how Algorithm 1 is different from the existing approaches of Sauré and Zeevi (2013) and Rusmevichientong et al. (2010) and also why other standard "bandit techniques" are not applicable to the MNL-Bandit problem.

Remark 1 (Universality) Note that Algorithm 1 does not require any prior knowledge/information about the problem parameters $\mathbf{v}$ (other than the assumption $v_i \leq v_0$, which is subsequently relaxed in Algorithm 3). This is in contrast with the approaches of Sauré and Zeevi (2013) and Rusmevichientong et al. (2010), which require the knowledge of the "separation gap," namely, the difference between the expected revenues of the optimal assortment and the second-best assortment. Assuming knowledge of this "separation gap," both these existing approaches explore a pre-determined set of assortments to estimate the MNL parameters within a desired accuracy, such that the optimal assortment corresponding to the estimated parameters is the (true) optimal assortment with high probability. This forced exploration of pre-determined assortments is avoided in Algorithm 1, which offers assortments adaptively, based on the current observed choices. The confidence regions derived for the parameters $\mathbf{v}$ and the subsequent assortment selection, ensure that Algorithm 1 judiciously maintains the balance between exploration and exploitation that is central to the MNL-Bandit problem.

Remark 2 (Estimation Approach) Because the MNL-Bandit problem is parameterized with parameter vector ($\mathbf{v}$), a natural approach is to build on standard estimation approaches like maximum likelihood (MLE), where the estimates are obtained by optimizing a loss function. However, the confidence regions for estimates resulting from such approaches are either:

1. asymptotic and are not necessarily valid for finite time with high probability, or
2. typically depend on true parameters, which are not known a priori. For example, finite time confidence regions associated with maximum likelihood estimates require the knowledge of $\sup_{\mathbf{v}\in\mathcal{V}} I(\mathbf{v})$ (see Borovkov 1984), where I is the Fisher information of the MNL choice model and $\mathcal{V}$ is the set of feasible parameters (that is not known apriori). Note that using $I(\mathbf{v}^{\mathsf{MLE}})$ instead of $\sup_{\mathbf{v}\in\mathcal{V}} I(\mathbf{v})$ for constructing confidence intervals would only lead to asymptotic guarantees and not finite sample guarantees.

In contrast, in Algorithm 1, we solve the estimation problem by a sampling method designed to give us unbiased estimates of the model parameters. The confidence bounds of these estimates and the algorithm do not depend on the underlying model parameters. Moreover, our sampling method allows us to compute the confidence regions by simple and efficient "book keeping" and avoids computational issues that are typically associated with standard estimation schemes such as MLE. Furthermore, the confidence regions associated with the unbiased estimates also facilitate

a tractable way to compute the optimistic assortment (see (3.5), (3.6) and Step-4 of Algorithm 1), which is less accessible for the MLE estimate.

Remark 3 (Alternative Approaches) Recently, Thompson Sampling (TS) has attracted considerable attention and several studies (Oliver and Li 2011, May et al. 2012) have demonstrated that TS significantly outperforms the state of the art bandit policies in practice. Typically, TS approaches proceed by assuming a prior distribution on the underlying parameters ($\mathbf{v}$ in the MNL-Bandit problem) and at every time step the posterior distribution on the parameters is updated based on the observed rewards and an arm (assortment) is selected with its posterior probability of it being the best arm. To implement a TS approach for the MNL-Bandit problem, one would need to specify the choice of prior, address the tractability of posterior sampling, etc. These issues also impede the analysis of such an algorithm. For example, in all existing work (Agrawal and Goyal 2017, Agrawal and Goyal 2013) on worst-case regret analysis for TS, the prior is chosen to allow a conjugate posterior, which permits theoretical analysis. For general posteriors, only Bayesian regret bounds have been proven, which are much weaker than the regret notion we consider in this paper. We return to discuss TS sampling in the concluding remarks of the paper.

4. Main results

In what follows, we make the following assumptions.
Assumption 4.1

1. *The MNL parameter corresponding to any product* $i \in \{1, \dots, N\}$ *satisfies* $v_i \leq v_0 = 1$.
2. *The family of assortments* $\mathcal{S}$ *is such that* $S \in \mathcal{S}$ *and* $Q \subseteq S$ *implies that* $Q \in \mathcal{S}$.

The first assumption is equivalent to the 'no purchase option' being the most likely outcome. We note that this holds in many realistic settings, in particular, in online retailing and online display-based advertising. The second assumption implies that removing a product from a feasible assortment preserves feasibility. This holds for most constraints arising in practice including cardinality, and matroid constraints more generally. We would like to note that the first assumption is made for ease of presentation of the key results and is not central to deriving bounds on the regret. In section 6.2, we relax this assumption and derive regret bounds that hold for any parameter instance.

Our main result is the following upper bound on the regret of the policy stated in Algorithm 1.

Theorem 1 (Performance Bounds for Algorithm 1) *For any instance* $\mathbf{v} = (v_0, \dots, v_N)$ *of the MNL-Bandit problem with* N *products,* $r_i \in [0, 1]$ *and Assumption 4.1, the regret of the policy given by Algorithm 1 at any time* T *is bounded as,*

$$Reg_\pi(T, \mathbf{v}) \leq C_1 \sqrt{NT \log NT} + C_2 N \log^2 NT,$$

where C_1 *and* C_2 *are absolute constants (independent of problem parameters).*

4.1. Proof Outline

In this section, we provide an outline of different steps involved in proving Theorem 1.

Confidence intervals. The first step in our regret analysis is to prove the following two properties of the estimates $v_{i,\ell}^{UCB}$ computed as in (3.4) for each product i. Specifically, that v_i is bounded by $v_{i,\ell}^{\mathsf{UCB}}$ with high probability, and that as a product is offered an increasing number of times, the estimates $v_{i,\ell}^{\mathsf{UCB}}$ converge to the true value with high probability. Intuitively, these properties establish $v_{i,\ell}^{UCB}$ as upper confidence bounds converging to actual parameters v_i, akin to the upper confidence bounds used in the UCB algorithm for MAB in Auer et al. (2002). We provide the precise statements for the above mentioned properties in Lemma 4.1. These properties follow from an observation that is conceptually equivalent to the IIA (Independence of Irrelevant Alternatives) property of MNL, and shows that in each epoch τ, $\hat{v}_{i,\tau}$ (the number of purchases of product i) provides an independent unbiased estimates of v_i. Intuitively, $\hat{v}_{i,\tau}$ is the ratio of probabilities of purchasing product i to preferring product 0 (no-purchase), which is independent of S_τ. This also explains why we choose to offer S_τ repeatedly until no-purchase occurs. Given these unbiased i.i.d. estimates from every epoch τ before ℓ, we apply a multiplicative Chernoff-Hoeffding bound to prove concentration of $\bar{v}_{i,\ell}$.

Validity of the optimistic assortment. The product demand estimates $v_{i,\ell-1}^{\mathsf{UCB}}$ were used in (3.7) to define expected revenue estimates $\tilde{R}_\ell(S)$ for every set S. In the beginning of every epoch ℓ, Algorithm 1 computes the optimistic assortment as $S_\ell = \arg\max_S \tilde{R}_\ell(S)$, and then offers S_ℓ repeatedly until no-purchase happens. The next step in the regret analysis is to leverage the fact that $v_{i,\ell}^{\mathsf{UCB}}$ is an upper confidence bound on v_i to prove similar, though slightly weaker, properties for the estimates $\tilde{R}_\ell(S)$. First, we show that estimated revenue is an upper confidence bound on the optimal revenue, i.e., $R(S^*, \mathbf{v})$ is bounded by $\tilde{R}_\ell(S_\ell)$ with high probability. The proof for these properties involves careful use of the structure of MNL model to show that the value of $\tilde{R}_\ell(S_\ell)$ is equal to the highest expected revenue achievable by any feasible assortment, among all instances of the problem with parameters in the range $[0, v_i^{\mathsf{UCB}}], i = 1, \ldots, n$. Since the actual parameters lie in this range with high probability, we have $\tilde{R}_\ell(S_\ell)$ is at least $R(S^*, \mathbf{v})$ with high probability. Lemma 4.2 provides the precise statement.

Bounding the regret. The final part of our analysis is to bound the regret in each epoch. First, we use the fact that $\tilde{R}_\ell(S_\ell)$ is an upper bound on $R(S^*, \mathbf{v})$ to bound the loss due to offering the assortment S_ℓ. In particular, we show that the loss is bounded by the difference between the "optimistic" revenue estimate, $\tilde{R}_\ell(S_\ell)$, and the actual expected revenue, $R(S_\ell)$. We then prove a Lipschitz property of the expected revenue function to bound the difference between these estimates

in terms of errors in individual product estimates $|v_{i,\ell}^{\mathsf{UCB}} - v_i|$. Finally, we leverage the structure of the MNL model and the properties of $v_{i,\ell}^{UCB}$ to bound the regret in each epoch. Lemma 4.3 provides the precise statements of above properties.

In the rest of this section, we make the above notions precise. Finally, in Appendix A.3, we utilize these properties to complete the proof of Theorem 1.

4.2. Upper confidence bounds

In this section, we will show that the upper confidence bounds $v_{i,\ell}^{\mathsf{UCB}}$ converge to the true parameters v_i from above. Specifically, we have the following result.

Lemma 4.1 *For every $\ell = 1, \cdots, L$, we have:*

1. $v_{i,\ell}^{\mathsf{UCB}} \geq v_i$ *with probability at least* $1 - \frac{6}{N\ell}$ *for all* $i = 1, \ldots, N$.
2. *There exists constants C_1 and C_2 such that*

$$v_{i,\ell}^{\mathsf{UCB}} - v_i \leq C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell + 1)}{T_i(\ell)}} + C_2 \frac{\log(\sqrt{N}\ell + 1)}{T_i(\ell)},$$

with probability at least $1 - \frac{7}{N\ell}$.

We first establish that the estimates $\hat{v}_{i,\ell}$, $\ell \leq L$ are unbiased i.i.d estimates of the true parameter v_i for all products. It is not immediately clear a priori if the estimates $\hat{v}_{i,\ell}$, $\ell \leq L$ are independent. In our setting, it is possible that the distribution of the estimate $\hat{v}_{i,\ell}$ depends on the offered assortment S_ℓ, which in turn depends on the history and therefore, previous estimates, $\{\hat{v}_{i,\tau},\, \tau = 1, \ldots, \ell - 1\}$. In Lemma A.1, we show that the moment generating function of $\hat{v}_{i,\ell}$ conditioned on S_ℓ only depends on the parameter v_i and not on the offered assortment S_ℓ, there by establishing that estimates are independent and identically distributed. Using the moment generating function, we show that $\hat{v}_{i,\ell}$ is a geometric random variable with mean v_i, i.e., $E(\hat{v}_{i,\ell}) = v_i$. We will use this observation and extend the classical multiplicative Chernoff-Hoeffding bounds (see Mitzenmacher and Upfal (2005) and Babaioff et al. (2015)) to geometric random variables. The proof is provided in Appendix A.1

4.3. Optimistic estimate and convergence rates

In this section, we show that the estimated revenue converges to the optimal expected revenue from above. First, we show that the estimated revenue is an upper confidence bound on the optimal revenue. In particular, we have the following result.

Lemma 4.2 *Suppose $S^* \in \mathcal{S}$ is the assortment with highest expected revenue, and Algorithm 1 offers $S_\ell \in \mathcal{S}$ in each epoch ℓ. Then, for every epoch ℓ, we have*

$$\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*) \geq R(S^*, \mathbf{v}) \text{ with probability at least } 1 - \frac{6}{\ell}.$$

In Lemma A.3, we show that the optimal expected revenue is monotone in the MNL parameters. It is important to note that we do not claim that the expected revenue is in general a monotone function, but only the value of the expected revenue corresponding to the optimal assortment increases with increase in the MNL parameters. The result follows from Lemma 4.1, where we established that $v_{i,\ell}^{\mathsf{UCB}} > v_i$ with high probability. We provide the detailed proof in Appendix A.2.

The following result provides the convergence rates of the estimate $\tilde{R}_\ell(S_\ell)$ to the optimal expected revenue.

Lemma 4.3 *If $r_i \in [0,1]$, there exists constants C_1 and C_2 such that for every $\ell = 1, \cdots, L$, we have*

$$(1+\textstyle\sum_{j\in S_\ell} v_j)(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v})) \leq C_1\sqrt{\frac{v_i \log(\sqrt{N}\ell+1)}{|\mathcal{T}_i(\ell)|}} + C_2\frac{\log(\sqrt{N}\ell+1)}{|\mathcal{T}_i(\ell)|},$$

with probability at least $1 - \frac{13}{\ell}$.

In Lemma A.4, we show that the expected revenue function satisfies a certain kind of Lipschitz condition. Specifically, the difference between the estimated, $\tilde{R}_\ell(S_\ell)$, and expected revenues, $R_\ell(S_\ell)$, is bounded by the sum of errors in parameter estimates for the products, $|v_{i,\ell}^{\mathsf{UCB}} - v_i|$. This observation in conjunction with the "optimistic estimates" property will let us bound the regret as an aggregated difference between estimated revenues and expected revenues of the offered assortments. Noting that we have already computed convergence rates between the parameter estimates earlier, we can extend them to show that the estimated revenues converge to the optimal revenue from above. We complete the proof in Appendix A.2.

5. Lower bounds and near-optimality of the proposed policy

In this section, we consider the special case of TU constraints, namely, a cardinality constrained assortment optimization problem, and establish that any policy must incur a regret of $\Omega(\sqrt{NT/K})$. More precisely, we prove the following result.

Theorem 2 (Lower bound on achievable performance) *There exists a (randomized) instance of the MNL-Bandit problem with $v_0 \geq v_i\,, i = 1, \ldots, N$, such that for any N and K, and any policy π that offers assortment $\mathcal{S}_t^\pi$, $|S_t^\pi| \leq K$ at time t, we have for all $T \geq N$ that,*

$$Reg_\pi(T, \mathbf{v}) := \mathbb{E}_\pi\left(\sum_{t=1}^{T} R(S^*, \mathbf{v}) - R(S_t^\pi, \mathbf{v})\right) \geq C\sqrt{\frac{NT}{K}},$$

where S^* is (at-most) K-cardinality assortment with maximum expected revenue, and C is an absolute constant.

Remark 4 (Optimality) Theorem 2 establishes that Algorithm 1 is optimal if we assume K to be fixed. We note that the assumption that K is fixed holds in many realistic settings, in particular, in online retailing, where there are a large number of products, but only fixed number of slots to show these products. Algorithm 1 is nearly optimal if K is also considered to be a problem parameter, with the upper bound being within a factor of $\sqrt{K}$ of the lower bound. In recent work, Chen and Wang (2017) established a lower bound of $\Omega\left(\sqrt{NT}\right)$ for the MNL-Bandit problem, when $K < N/4$, thus suggesting that Algorithm 1 is optimal even with respect to its dependence on K. For the special case of the unconstrained MNL-Bandit problem (i.e., $K = N$), the regret bound of Algorithm 1 can be improved to $O(\sqrt{|S^*|T})$, where $|S^*|$ is the size of the optimal assortment (see Appendix A.4) and the optimality gap for the unconstrained setting is $\sqrt{|S^*|}$.

5.1. Proof overview

For ease of exposition, we focus here on the case where $K < N$, and present the proof for lower bound when $K = N$ in Appendix E.1. To that end, we will assume that $K < N$ for the rest of this section. We prove Theorem 2 by a reduction to a parametric multi-armed bandit (MAB) problem, for which a lower bound is known.

Definition 5.1 (MAB instance I_{MAB}) *Define I_{MAB} as a (randomized) instance of MAB problem with $N \geq 2$ Bernoulli arms (reward is either 0 or 1) and the probability of the reward being 1 for arm i is given by,*

$$\mu_i = \begin{cases} \alpha, & \text{if } i \neq j, \\ \alpha + \epsilon, & \text{if } i = j, \end{cases} \quad \text{for all } i = 1, \ldots, N,$$

where j is set uniformly at random from $\{1, \ldots, N\}$, $\alpha < 1$ and $\epsilon = \frac{1}{100}\sqrt{\frac{N\alpha}{T}}$.

Throughout this section we will use the terms algorithm and policy interchangeably. An algorithm $\mathcal{A}$ is referred to as online if it adaptively selects a history dependent $\mathcal{A}_t \in \{1, \ldots, n\}$ at each time t as in (2.4) for the MAB problem.

Lemma 5.1 *For any $N \geq 2$, $\alpha < 1$, T and any online algorithm $\mathcal{A}$ that plays arm $\mathcal{A}_t$ at time t, the expected regret on instance I_{MAB} is at least $\frac{\epsilon T}{6}$. That is,*

$$Reg_{\mathcal{A}}(T, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{t=1}^{T}(\mu_j - \mu_{\mathcal{A}_t})\right] \geq \frac{\epsilon T}{6},$$

where, the expectation is both over the randomization in generating the instance (value of j), as well as the random outcomes that result from pulled arms.

The proof of Lemma 5.1 is a simple extension of the proof of the $\Omega(\sqrt{NT})$ lower bound for the Bernoulli instance with parameters $\frac{1}{2}$ and $\frac{1}{2}+\epsilon$ (for example, see Bubeck and Cesa-Bianchi 2012), and has been provided in Appendix E for the sake of completeness. We use the above lower bound for the MAB problem to prove that any algorithm must incur at least $\Omega(\sqrt{NT/K})$ regret on the following instance of the MNL-Bandit problem.

Definition 5.2 (MNL-Bandit instance I_{MNL}) *Define I_{MNL} as the following (randomized) instance of MNL-Bandit problem with K-cardinality constraint, $\hat{N}=NK$ products, parameters $v_0=K$ and for $i=1,\ldots,\hat{N}$,*

$$v_i=\begin{cases}\alpha, & \text{if } \lceil \frac{i}{K} \rceil \neq j,\\ \alpha+\epsilon, & \text{if } \lceil \frac{i}{K} \rceil = j,\end{cases}$$

where j is set uniformly at random from $\{1,\ldots,N\}$, $\alpha<1$, and $\epsilon=\frac{1}{100}\sqrt{\frac{N\alpha}{T}}$ and $r_i=1$.

We will show that any MNL-Bandit algorithm has to incur a regret of $\Omega\left(\sqrt{\frac{NT}{K}}\right)$ on instance I_{MNL}. The main idea in our reduction is to show that if there exists an algorithm $\mathcal{A}_{\mathsf{MNL}}$ for MNL-Bandit that achieves $o(\sqrt{\frac{NT}{K}})$ regret on instance I_{MNL}, then we can use $\mathcal{A}_{\mathsf{MNL}}$ as a subroutine to construct an algorithm $\mathcal{A}_{\mathsf{MAB}}$ for the MAB problem that achieves strictly less than $\frac{\epsilon T}{6}$ regret on instance I_{MAB} in time T, thus contradicting the lower bound of Lemma 5.1. This will prove Theorem 2 by contradiction.

5.2. Construction of the MAB algorithm using the MNL algorithm

Algorithm 2 Algorithm $\mathcal{A}_{\mathsf{MAB}}$

Initialization: $t=0$, $\ell=0$
while $t\leq T$ **do**
  Update $\ell=\ell+1$
  Call $\mathcal{A}_{\mathsf{MNL}}$, receive assortment $S_\ell\subset[\hat{N}]$
  Repeat until 'exit loop'
    With probability $\frac{1}{2}$, send **Feedback to** $\mathcal{A}_{\mathsf{MNL}}$ 'no product was purchased', **exit loop**
    Update $t=t+1$
    With probability $\frac{1}{2K}$, **pull** arm $\mathcal{A}_t=\lceil \frac{i}{K} \rceil$, where $i\in S_\ell$
    With probability $\frac{1}{2}-\frac{|S_\ell|}{2K}$, **continue the loop** (go to Step-5)
    If reward is 1, send **Feedback to** $\mathcal{A}_{\mathsf{MNL}}$ 'i was purchased' and **exit loop**
  end loop
end while

Algorithm 2 provides the exact construction of $\mathcal{A}_{\mathsf{MAB}}$, which simulates the $\mathcal{A}_{\mathsf{MNL}}$ algorithm as a "black-box." Note that $\mathcal{A}_{\mathsf{MAB}}$ pulls arms at time steps $t = 1, \ldots, T$. These arm pulls are interleaved by simulations of $\mathcal{A}_{\mathsf{MNL}}$ steps (**Call** $\mathcal{A}_{\mathsf{MNL}}$, **Feedback to** $\mathcal{A}_{\mathsf{MNL}}$). When step ℓ of $\mathcal{A}_{\mathsf{MNL}}$ is simulated, it uses the feedback from $1, \ldots, \ell-1$ to suggest an assortment S_ℓ; and recalls a feedback from $\mathcal{A}_{\mathsf{MAB}}$ on which product (or no product) was purchased out of those offered in S_ℓ, where the probability of purchase of product $i \in S_\ell$ is $v_i \big/ (v_0 + \sum_{i \in S_\ell} v_i)$. Before showing that the $\mathcal{A}_{\mathsf{MAB}}$ indeed provides the right feedback to $\mathcal{A}_{\mathsf{MNL}}$ in the ℓ^{th} step for each ℓ, we introduce some notation.

Let M_ℓ denote the length of the loop at step ℓ, more specifically, the cumulative number of times, $\mathcal{A}_{\mathsf{MNL}}$ was executing steps 6, 8 or 9 in the ℓ^{th} step before exiting the loop. For every $i \in S_\ell \cup 0$, let ζ_ℓ^i denote the event that the feedback to $\mathcal{A}_{\mathsf{MNL}}$ from $\mathcal{A}_{\mathsf{MAB}}$ after step ℓ of $\mathcal{A}_{\mathsf{MNL}}$ is "product i is purchased". We have,

$$\mathcal{P}(M_\ell = m \cap \zeta_\ell^i) = \frac{v_i}{2K} \left(\frac{1}{2K} \sum_{i \in S_\ell} (1 - v_i) + \frac{1}{2} - \frac{|S_\ell|}{2K} \right)^{m-1} \quad \text{for each } i \in S_\ell \cup \{0\}.$$

Hence, the probability that $\mathcal{A}_{\mathsf{MAB}}$'s feedback to $\mathcal{A}_{\mathsf{MNL}}$ is "product i is purchased" is,

$$p_{S_\ell}(i) = \sum_{m=1}^{\infty} \mathcal{P}(M_\ell = m \cap \zeta_\ell^i) = \frac{v_i}{v_0 + \sum_{q \in S_\ell} v_q}.$$

This establish that $\mathcal{A}_{\mathsf{MAB}}$ provides the appropriate feedback to $\mathcal{A}_{\mathsf{MNL}}$.

5.3. Proof of Theorem 2

We prove the result by establishing three key results. First, we upper bound the regret for the MAB algorithm, $\mathcal{A}_{\mathsf{MAB}}$. Then, we prove a lower bound on the regret for the MNL algorithm, $\mathcal{A}_{\mathsf{MNL}}$. Finally, we relate the regret of $\mathcal{A}_{\mathsf{MAB}}$ and $\mathcal{A}_{\mathsf{MNL}}$ and use the established lower and upper bounds to show a contradiction.

For the rest of this proof, assume that L is the total number of calls to $\mathcal{A}_{\mathsf{MNL}}$ in $\mathcal{A}_{\mathsf{MAB}}$. Let S^* be the optimal assortment for I_{MNL}. For any instantiation of I_{MNL}, it is easy to see that the optimal assortment contains K items, all with parameter $\alpha + \epsilon$, i.e., it contains all i such that $\lceil \frac{i}{K} \rceil = j$. Therefore, $V(S^*) = K(\alpha + \epsilon) = K\mu_j$. Note that if an algorithm offers an assortment, S_ℓ, such that $|S_\ell| < K$, then we can improve the regret incurred by this algorithm for the MNL-Bandit instance I_{MNL} by offering an assortment $\hat{S}_\ell = S_\ell \cup \{i\}$ for some $i \notin S_\ell$. Since our focus is on lower bounding the regret, we will assume, without loss of generality, that $|S_\ell| = K$ for the rest of this section.

Upper bound for the regret of the MAB algorithm. The first step in our analysis is to prove an upper bound on the regret of the MAB algorithm, $\mathcal{A}_{\mathsf{MAB}}$ on the instance I_{MAB}. Let us label the loop following the ℓth call to $\mathcal{A}_{\mathsf{MNL}}$ in Algorithm 2 as ℓth loop. Note that the probability

of exiting the loop is $p = E[\frac{1}{2} + \frac{1}{2}\mu_{\mathcal{A}_t}] = \frac{1}{2} + \frac{1}{2K}V(S_\ell)$, where $V(S_\ell) \triangleq \sum_{i\in S_\ell} v_i$. In every step of the loop until exited, an arm is pulled with probability 1/2. The optimal strategy would pull the best arm so that the total expected optimal reward in the loop is $\sum_{r=1}^{\infty}(1-p)^{r-1}\frac{1}{2}\mu_j = \frac{\mu_j}{2p} = \frac{1}{2Kp}V(S^*)$. Algorithm $\mathcal{A}_{\mathsf{MAB}}$ pulls a random arm from S_ℓ, so total expected algorithm's reward in the loop is $\sum_{r=1}^{\infty}(1-p)^{r-1}\frac{1}{2K}V(S_\ell) = \frac{1}{2Kp}V(S_\ell)$. Subtracting the algorithm's reward from the optimal reward, and substituting p, we obtain that the total expected regret of $\mathcal{A}_{\mathsf{MAB}}$ over the arm pulls in loop ℓ is

$$\frac{V(S^*) - V(S_\ell)}{(K + V(S_\ell))}.$$

Noting that $V(S_\ell) \geq K\alpha$, we have the following upper bound on the regret for the MAB algorithm.

$$Reg_{\mathcal{A}_{\mathsf{MAB}}}(T, \boldsymbol{\mu}) \leq \frac{1}{(1+\alpha)}\mathbb{E}\left(\sum_{\ell=1}^{L}\frac{1}{K}(V(S^*) - V(S_\ell))\right), \tag{5.1}$$

where the expectation in equation (5.1) is over the random variables L and S_ℓ.

Lower bound for the regret of the MNL algorithm. Here, we derive a lower bound on the regret of the MNL algorithm, $\mathcal{A}_{\mathsf{MNL}}$ on the instance I_{MNL}. Specifically,

$$\begin{aligned} Reg_{\mathcal{A}_{\mathsf{MNL}}}(L, \mathbf{v}) &= \mathbb{E}\left[\sum_{\ell=1}^{L}\frac{V(S^*)}{v_0 + V(S^*)} - \frac{V(S_\ell)}{v_0 + V(S_\ell)}\right] \\ &\geq \frac{1}{K(1+\alpha)}\mathbb{E}\left[\sum_{\ell=1}^{L}\left(\frac{V(S^*)}{1+\frac{\epsilon}{1+\alpha}} - V(S_\ell)\right)\right]. \end{aligned}$$

Therefore, it follows that,

$$Reg_{\mathcal{A}_{\mathsf{MNL}}}(L, \mathbf{v}) \geq \frac{1}{(1+\alpha)}\mathbb{E}\left[\sum_{\ell=1}^{L}\frac{1}{K}(V(S^*) - V(S_\ell)) - \frac{\epsilon v^* L}{(1+\alpha)^2}\right], \tag{5.2}$$

where $v^* = \alpha + \epsilon$ and the expectation in equation (5.2) is over the random variables L and S_ℓ.

Relating the regret of the MNL algorithm and the MAB algorithm. Finally, we relate the regret of the MNL algorithm $\mathcal{A}_{\mathsf{MNL}}$ and MAB algorithm $\mathcal{A}_{\mathsf{MAB}}$ to derive a contradiction. The first step in relating the regret involves relating the length of the horizons of $\mathcal{A}_{\mathsf{MNL}}$ and $\mathcal{A}_{\mathsf{MAB}}$, L and T respectively. Note that, after every call to $\mathcal{A}_{\mathsf{MNL}}$ ("**Call** $\mathcal{A}_{\mathsf{MNL}}$" in Algorithm 2), many iterations of the following loop may be executed; in roughly 1/2 of those iterations, an arm is pulled and t is advanced (with probability 1/2, the loop is exited without advancing t). Therefore, T should be at least a constant fraction of L. Lemma E.3 in Appendix E makes this precise by showing that $\mathbb{E}(L) \leq 3T$.

Now we are ready to prove Theorem 2. From (5.1) and (5.2), we have

$$Reg_{\mathcal{A}_{\mathsf{MAB}}}(T, \boldsymbol{\mu}) \leq \mathbb{E}\left(Reg_{\mathcal{A}_{\mathsf{MNL}}}(L, \mathbf{v}) + \frac{\epsilon v^* L}{(1+\alpha)^2}\right).$$

For the sake of contradiction, suppose that the regret of the $\mathcal{A}_{\mathsf{MNL}}$, $Reg_{\mathcal{A}_{\mathsf{MNL}}}(L,\mathbf{v}) \le c\sqrt{\frac{\hat{N}L}{K}}$ for a constant c to be prescribed below. Then, from Jensen's inequality, it follows that,

$$Reg_{\mathcal{A}_{\mathsf{MAB}}}(T,\boldsymbol{\mu}) \le c\sqrt{\frac{\hat{N}\mathbb{E}(L)}{K}} + \frac{\epsilon v^* \mathbb{E}(L)}{(1+\alpha)^2}.$$

From lemma E.3, we have that $\mathbb{E}(L) \le 3T$. Therefore, we have, $c\sqrt{\frac{\hat{N}\mathbb{E}(L)}{K}} = c\sqrt{N\mathbb{E}(L)} \le c\sqrt{3NT} = c\epsilon T\sqrt{\frac{3}{\alpha}} < \frac{\epsilon T}{12}$ on setting $c < \frac{1}{12}\sqrt{\frac{\alpha}{3}}$. Also, using $v^* = \alpha + \epsilon \le 2\alpha$, and setting α to be a small enough constant, we can get that the second term above is also strictly less than $\frac{\epsilon T}{12}$. Combining these observations, we have

$$Reg_{\mathcal{A}_{\mathsf{MAB}}}(T,\boldsymbol{\mu}) < \tfrac{\epsilon T}{12} + \tfrac{\epsilon T}{12} = \tfrac{\epsilon T}{6},$$

thus arriving at a contradiction. □

6. Extensions

In this section, we consider two extensions of the MNL-Bandit problem. In the first extension, we consider problem instances that are "well separated" and present an improved logarithmic regret bound. We will then consider a setting where the "no purchase" assumption ($v_i \le v_0$ for all i) is relaxed and present a modified algorithm that works for more general class of MNL parameters and establish $\tilde{O}(\sqrt{BNT})$ regret bounds.

6.1. Improved regret bounds for "well-separated" instances

In this section, we derive an $O(\log T)$ regret bound for Algorithm 1 for instances that are "well separated." In Section 4, we established worst case regret bounds for Algorithm 1 that hold for all problem instances satisfying Assumption 4.1. Although our algorithm ensures that the exploration-exploitation tradeoff is balanced at all times, for problem instances that are "well separated," our algorithm quickly converges to the optimal solution leading to better regret bounds. More specifically, we consider problem instances where the optimal assortment and "second best" assortment are sufficiently "separated" and derive a $O(\log T)$ regret bound that depends on the parameters of the instance. Note that, unlike the regret bound derived in Section 4 that holds for all problem instances satisfying Assumption 4.1, the bound we derive here only holds for instances having certain separation between the revenues corresponding to optimal and second best assortments. In particular, let $\Delta(\mathbf{v})$ denote the difference between the expected revenues of the optimal and second-best assortment, i.e.,

$$\Delta(\mathbf{v}) = \min_{\{S\in\mathcal{S} | R(S,\mathbf{v}) \neq R(S^*,\mathbf{v})\}} \{R(S^*,\mathbf{v}) - R(S)\}. \tag{6.1}$$

We have the following result.

Theorem 3 (Performance Bounds for Algorithm 1 in "well separated" case) *For any instance* $\mathbf{v} = (v_0, \ldots, v_N)$ *of the MNL-Bandit problem with* N *products,* $r_i \in [0,1]$ *and Assumption 4.1, the regret of the policy given by Algorithm 1 at any time* T *is bounded as,*

$$Reg_\pi(T, \mathbf{v}) \leq B_1 \left(\frac{N^2 \log T}{\Delta(\mathbf{v})} \right) + B_2,$$

where B_1 *and* B_2 *are absolute constants.*

Proof outline. In this setting, we analyze the regret by separately considering the epochs that satisfy certain desirable properties and the ones that do not. Specifically, we denote epoch ℓ as a "good" epoch if the parameters $v_{i,\ell}^{\mathsf{UCB}}$ satisfy the following property,

$$0 \leq v_{i,\ell}^{\mathsf{UCB}} - v_i \leq C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell + 1)}{T_i(\ell)}} + C_2 \frac{\log(\sqrt{N}\ell + 1)}{T_i(\ell)},$$

and we call it a "bad" epoch otherwise, where C_1 and C_2 are constants as defined in Lemma 4.1. Note that every epoch ℓ is a good epoch with high probability $(1 - \frac{13}{\ell})$ and we show that regret due to "bad" epochs is bounded by a constant (see Appendix C). Therefore, we focus on "good" epochs and show that there exists a constant τ, such that after each product has been offered in at least τ "good" epochs, Algorithm 1 finds the optimal assortment. Based on this result, we can then bound the total number of "good" epochs in which a sub-optimal assortment can be offered by our algorithm. Specifically, let

$$\tau = \frac{4NC \log NT}{\Delta^2(\mathbf{v})}, \tag{6.2}$$

where $C = \max\{C_1^2, C_2\}$. Then we have the following result.

Lemma 6.1 *Let* ℓ *be a "good" epoch and* S_ℓ *be the assortment offered by Algorithm 1 in epoch* ℓ*. If every product in assortment* S_ℓ *is offered in at least* τ *"good epochs," i.e.* $T_i(\ell) \geq \tau$ *for all* i*, then we have* $R(S_\ell, \mathbf{v}) = R(S^*, \mathbf{v})$ *.*

We prove Lemma 6.1 in Appendix C. The next step in the analysis is to show that Algorithm 1 will offer a small number of sub-optimal assortments in "good" epochs. We make this precise in the following observation whose proof amounts to a simple counting exercise using Lemma 6.1 (see full proof in Appendix C.)

Lemma 6.2 *Algorithm 1 cannot offer sub-optimal assortments in more than* $N\tau$ *"good" epochs.*

The proof for Theorem 3 follows from the above result. In particular, noting that the number of epochs in which sub-optimal assortment is offered is small, we re-use the regret analysis of Section 4 to bound the regret by $O(N^2 \log T)$. We provide a rigorous proof in Appendix C for the sake of

completeness. Note that for the special case of cardinality constraints, we have $|S_\ell| \le K$ for every epoch ℓ. By modifying the definition of τ in (6.2) to $\tau = 4KC \log NT/\Delta^2(\mathbf{v})$ and following the above analysis, we can improve the regret bound to $O(NK \log T)$ for this case. Specifically, we have the following.

Corollary 6.1 (Performance bounds in well separated case under cardinality constraints) *For any instance $\mathbf{v} = (v_0, \ldots, v_N)$ of the MNL-Bandit problem with N products and cardinality constraint K, $r_i \in [0,1]$ and $v_0 \ge v_i$ for all i, the regret of the policy given by Algorithm 1 at any time T is bounded as,*

$$Reg_\pi(T, \mathbf{v}) \le B_1 \frac{NK \log NT}{\Delta(\mathbf{v})} + B_2,$$

where, B_1 and B_2 are absolute constants and $\Delta(\mathbf{v})$ is defined in (6.1).

It should be noted that the bound obtained in Corollary 6.1 is similar in magnitude to the regret bounds obtained by Sauré and Zeevi (2013), when K is assumed to be fixed, and is strictly better than the regret bound $O(N^2 \log^2 T)$ established by Rusmevichientong et al. (2010). Moreover, our algorithm does not require the knowledge of $\Delta(\mathbf{v})$, unlike the aforementioned papers which build on a conservative estimate of $\Delta(\mathbf{v})$ to implement their proposed policies.

6.2. Relaxing the "no purchase" assumption

In this section, we extend our approach (Algorithm 1) to the setting where the assumption that $v_i \le v_0$ for all i is relaxed. The essential ideas in the extension remain the same as our earlier approach, specifically optimism under uncertainty, and our policy is structurally similar to Algorithm 1. The modified policy requires a small but mandatory initial exploration period. However, unlike the works of Rusmevichientong et al. (2010) and Sauré and Zeevi (2013), the exploratory period does not depend on the specific instance parameters and is constant for all problem instances. Therefore, our algorithm is parameter independent and remains relevant for practical applications. Moreover, our approach continues to simultaneously explore and exploit after the initial exploratory phase. In particular, the initial exploratory phase is to ensure that the estimates converge to the true parameters from above particularly in cases when the attraction parameter v_i (frequency of purchase), is large for certain products. We describe our approach in Algorithm 3.

We can extend the analysis in Section 4 to bound the regret of Algorithm 3 as follows.

Theorem 4 (Performance Bounds for Algorithm 3) *For any instance $\mathbf{v} = (v_0, \ldots, v_N)$, of the MNL-Bandit problem with N products, $r_i \in [0,1]$ for all $i = 1, \ldots, N$, the regret of the policy corresponding to Algorithm 3 at any time T is bounded as,*

$$Reg_\pi(T, \mathbf{v}) \le C_1 \sqrt{BNT \log NT} + C_2 N \log^2 NT + C_3 NB \log NT,$$

Algorithm 3 Exploration-Exploitation algorithm for MNL-Bandit general parameters

Initialization: $v_{i,0}^{\mathsf{UCB}} = 1$ for all $i = 1, \ldots, N$
$t = 1$; $\ell = 1$ keeps track of the time steps and total number of epochs respectively
$T_i(1) = 0$ for all $i = 1, \ldots, N$
while $t < T$ **do**
  Compute $S_\ell = \underset{S \in \mathcal{S}}{\operatorname{argmax}}\, \tilde{R}_\ell(S) = \dfrac{\sum_{i \in S} r_i v_{i,\ell-1}^{\mathsf{UCB}}}{1 + \sum_{j \in S} v_{j,\ell-1}^{\mathsf{UCB}}}$
  if $T_i(\ell) < 48 \log(\sqrt{N}\ell + 1)$ for some $i \in S_\ell$ **then**
    Consider $\hat{S} = \{i | T_i(\ell) < 48 \log(\sqrt{N}\ell + 1)\}$
    Choose $S_\ell \in \mathcal{S}$ such that $S_\ell \subset \hat{S}$
  end if
  Offer assortment S_ℓ, observe the purchasing decision, c_t of the consumer
  if $c_t = 0$ **then**
    compute $\hat{v}_{i,\ell} = \sum_{t \in \mathcal{E}_\ell} \mathbb{1}(c_t = i)$, no. of consumers who preferred i in epoch ℓ, for all $i \in S_\ell$
    update $\mathcal{T}_i(\ell) = \{\tau \le \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until ℓ that offered product i
    update $\bar{v}_{i,\ell} = \dfrac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
    update $v_{i,\ell}^{\mathsf{UCB2}} = \bar{v}_{i,\ell} + \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\} \sqrt{\frac{48 \log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48 \log(\sqrt{N}\ell+1)}{T_i(\ell)}$
    $\ell = \ell + 1$
  else
    $\mathcal{E}_\ell = \mathcal{E}_\ell \cup t$, time indices corresponding to epoch ℓ
  end if
  $t = t + 1$
end while

where C_1, C_2 and C_3 are absolute constants and $B = \max\{\max_i \frac{v_i}{v_0}, 1\}$.

Proof outline. Note that Algorithm 3 is very similar to Algorithm 1 except for the initial exploratory phase. Hence, to bound the regret we first prove that the initial exploratory phase is indeed bounded and then follow the approach discussed in Section 4 to establish the correctness of the confidence intervals, the optimistic assortment, and finally deriving the convergence rates and regret bounds. We make the above notions precise and provide the complete proof in Appendix B.

7. Computational study

In this section, we present insights from numerical experiments that test the empirical performance of our policy and highlight some of its salient features. We study the performance of Algorithm 1

from the perspective of robustness with respect to the "separability parameter" of the underlying instance. In particular, we consider varying levels of separation between the revenues corresponding to the optimal assortment and the second best assortment and perform a regret analysis numerically. We contrast the performance of Algorithm 1 with the approach in Sauré and Zeevi (2013) for different levels of separation. We observe that when the separation between the revenues corresponding to optimal assortment and second best assortment is sufficiently small, the approach in Sauré and Zeevi (2013) breaks down, i.e., incurs linear regret, while the regret of Algorithm 1 only grows sub-linearly with respect to the selling horizon. We also present results from a simulated study on a real world data set.

7.1. Robustness of Algorithm 1

Here, we present a study that examines the robustness of Algorithm 1 with respect to the instance separability. We consider a parametric instance (see (7.1)), where the separation between the revenues of the optimal assortment and next best assortment is specified by the parameter ϵ and compare the performance of Algorithm 1 for different values of ϵ.

Experimental setup. We consider the parametric MNL setting with $N = 10$, $K = 4$, $r_i = 1$ for all i and utility parameters $v_0 = 1$ and for $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else} , \end{cases} \tag{7.1}$$

where $0 < \epsilon < 0.25$, specifies the difference between revenues corresponding to the optimal assortment and the next best assortment. Note that this problem has a unique optimal assortment, $\{1, 2, 9, 10\}$ with an expected revenue of $1 + 4\epsilon/2 + 4\epsilon$ and next best assortment has revenue of $1 + 3\epsilon/2 + 3\epsilon$. We consider four different values for ϵ, $\epsilon = \{0.05, 0.1, 0.15, 0.25\}$, where higher value of ϵ corresponds to larger separation, and hence an "easier" problem instance.

Results. Figure 1 summarizes the performance of Algorithm 1 for different values of ϵ. The results are based on running 100 independent simulations, the standard errors are within 2%. Note that the performance of Algorithm 1 is consistent across different values of ϵ; with a regret that exhibits sub linear growth. Observe that as the value of ϵ increases the regret of Algorithm 1 decreases. While not immediately obvious from Figure 1, the regret behavior is fundamentally different in the case of "small" ϵ and "large" ϵ. To see this, in Figure 2 we focus on the regret for $\epsilon = 0.05$ and $\epsilon = 0.25$ and fit to $\log T$ and $\sqrt{T}$ respectively. (The parameters of these functions are obtained via simple linear regression of the regret vs $\log T$ and $\sqrt{T}$ respectively). It can be observed that the regret is roughly logarithmic when $\epsilon = 0.25$, and in contrast roughly behaves like $\sqrt{T}$ when $\epsilon = 0.05$. This illustrates the theory developed in Section 6.1, where we showed that the regret

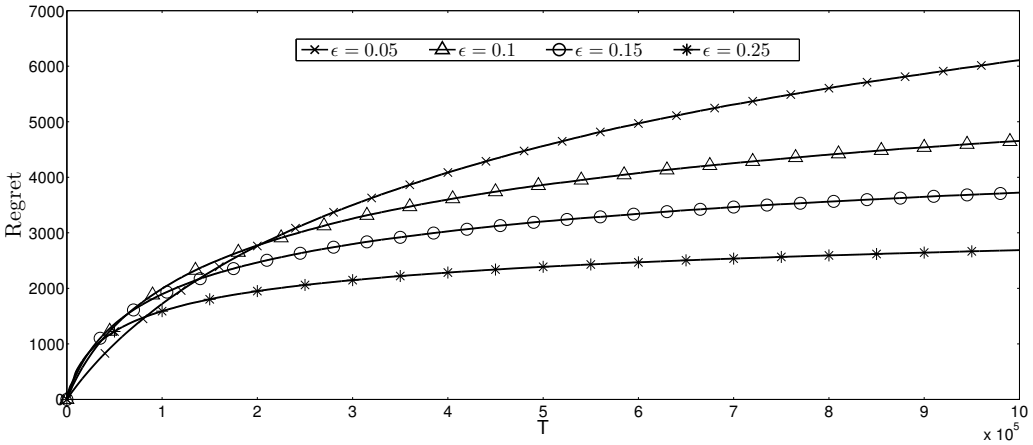

Figure 1 Performance of Algorithm 1 measured as the regret on the parametric instance (7.1). The graphs illustrate the dependence of the regret on T for "separation gaps" $\epsilon = 0.05, 0.1, 0.15$ and 0.25 respectively.

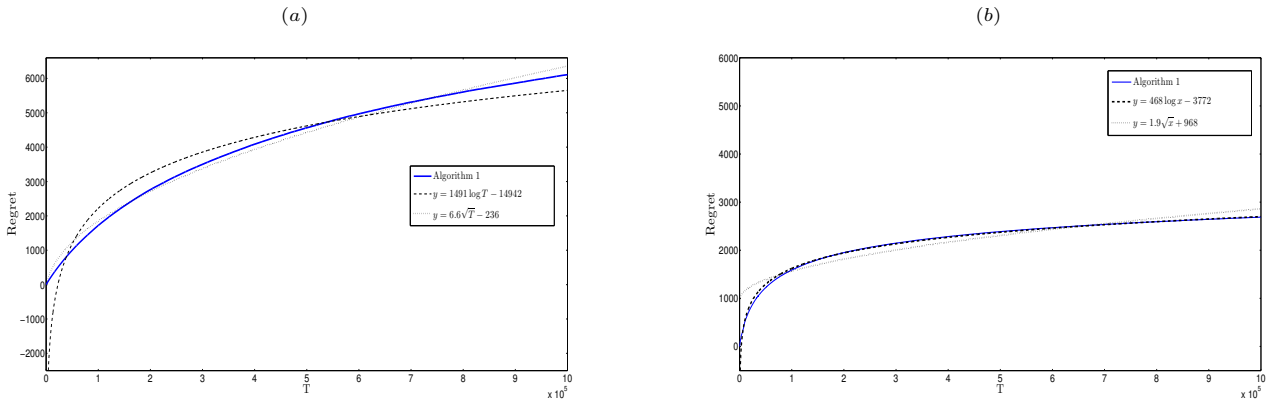

Figure 2 Best fit for the regret of Algorithm 1 on the parametric instance (7.1). The graphs (a), (b) illustrate the dependence of the regret on T for "separation gaps" $\epsilon = 0.05$, and 0.25 respectively. The best $y = \beta_1 \log T + \beta_0$ fit and best $y = \beta_1 \sqrt{T} + \beta_0$ fit are superimposed on the regret curve.

grows logarithmically with time, if the optimal assortment and next best assortment are "well separated," while the worst-case regret scales as $\sqrt{T}$.

7.2. Comparison with existing approaches

In this section, we present a computational study comparing the performance of our algorithm to that of Sauré and Zeevi (2013). (To the best of our knowledge, Sauré and Zeevi (2013) is currently the best existing approach for our problem setting.) To be implemented, their approach requires certain a priori information of a "separability parameter"; roughly speaking, measuring the degree to which the optimal and next-best assortments are distinct from a revenue standpoint. More specifically, their algorithm follows an *explore-then-exploit* approach, where every product is offered

for a minimum duration of time that is determined by an estimate of said "separability parameter." After this mandatory exploration phase, the parameters of the choice model are estimated based on the past observations and the optimal assortment corresponding to the estimated parameters is offered for the subsequent consumers. If the optimal assortment and the next best assortment are "well separated," then the offered assortment is optimal with high probability, otherwise, the algorithm could potentially incur linear regret. Therefore, the knowledge of this "separability parameter" is crucial. For our comparison, we consider the exploration period suggested by Sauré and Zeevi (2013) and compare it with the performance of Algorithm 1 for different values of separation (ϵ). We will see that for any given exploration period, there is an instance where the approach in Sauré and Zeevi (2013) "breaks down" or in other words incurs linear regret, while the regret of Algorithm 1 grows sub-linearly ($O(\sqrt{T})$, more precisely) for all values of ϵ as asserted in Theorem 1.

Experimental setup and results. We consider the parametric MNL setting as described in (7.1) and for each value of $\epsilon \in \{0.05, 0.1, 0.15, 0.25\}$. Since the implementation of the policy in Sauré and Zeevi (2013) requires knowledge of the selling horizon and minimum exploration period a priori, we take the exploration period to be $20 \log T$ as suggested in Sauré and Zeevi (2013) and the selling horizon $T = 10^6$. Figure 3 compares the regret of Algorithm 1 with that of Sauré and Zeevi (2013). The results are based on running 100 independent simulations with standard error of 2%. We observe that the regret for Sauré and Zeevi (2013) is better than the regret of Algorithm 1 when $\epsilon = 0.25$ but is worse for other values of ϵ. This can be attributed to the fact that for the assumed exploration period, their algorithm fails to identify the optimal assortment within the exploration phase with sufficient probability and hence incurs a linear regret for $\epsilon = 0.05, 0.1$ and 0.15. Specifically, among the 100 simulations we tested, the algorithm of Sauré and Zeevi (2013) identified the optimal assortment for only $7\%, 40\%, 61\%$ and 97% cases, when $\epsilon = 0.05, 0.1, 0.15$, and 0.25, respectively. This highlights the sensitivity to the "separability parameter" and the importance of having a reasonable estimate for the exploration period. Needless to say, such information is typically not available in practice. In contrast, the performance of Algorithm 1 is consistent across different values of ϵ, insofar as the regret grows in a sub-linear fashion in all cases.

7.3. Performance of Algorithm 1 on a simulation of real data

Here, we present the results of a simulated study on a real data set and compare the performance of Algorithm 1 to that of Sauré and Zeevi (2013).

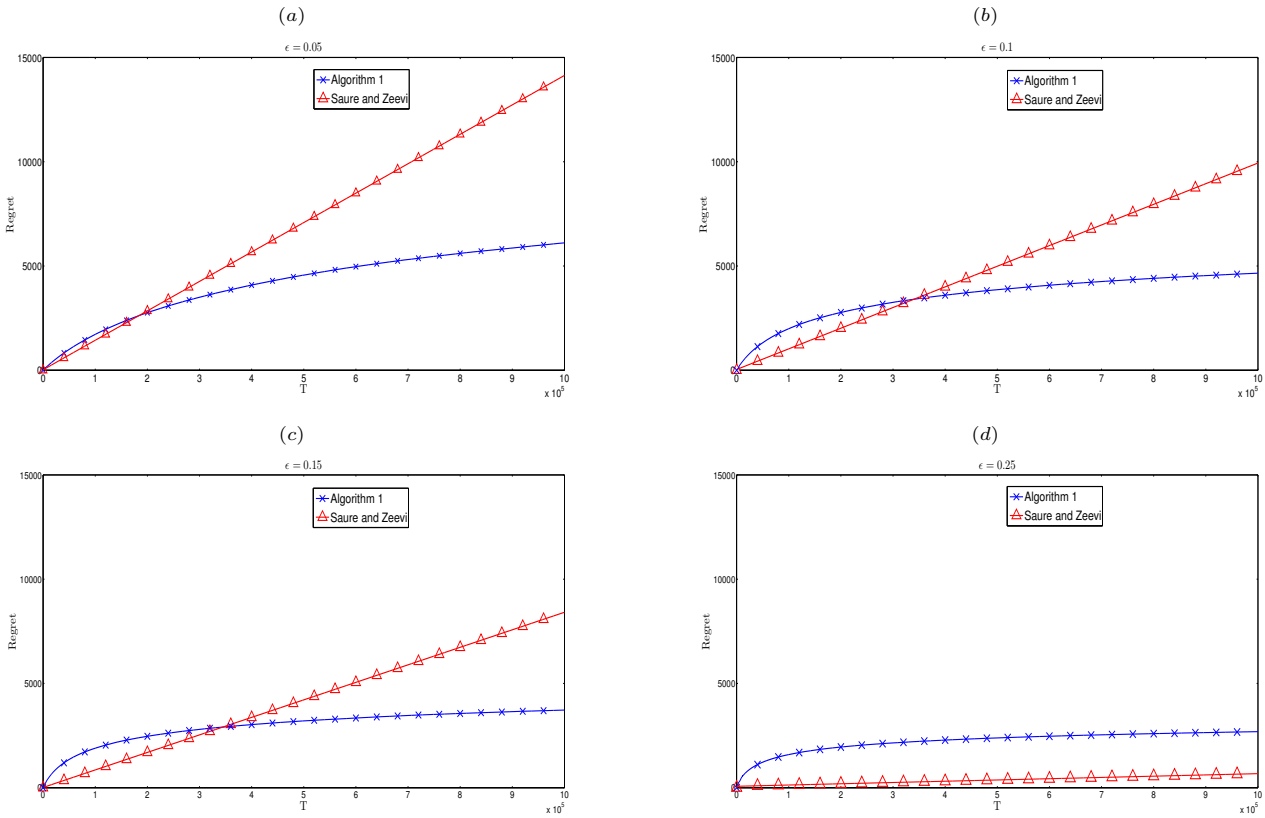

Figure 3 Comparison with the algorithm of Sauré and Zeevi (2013). The graphs (a), (b), (c) and (d) compares the performance of Algorithm 1 to that of Sauré and Zeevi (2013) on problem instance (7.1), for $\epsilon = 0.05, 0.1, 0.15$ and 0.25 respectively.

Attribute	Attribute Values
price	Very-high, high, medium, low
maintenance costs	Very-high, high, medium, low
# doors	2, 3, 4, 5 or more
passenger capacity	2, 4, more than 4
luggage capacity	small, medium and big
safety perception	low, medium, high

Table 1 Attribute information of cars in the database

Data description. We consider the "UCI Car Evaluation Database" (see Lichman (2013)) which contains attributes for $N = 1728$ cars and consumer ratings for each car. The exact details of the attributes are provided in Table 1. Rating for each car is also available. In particular, every car is associated with one of the following four ratings, unacceptable, acceptable, good and very good.

Assortment optimization framework. We assume that the consumer choice is modeled by the MNL model, where the mean utility of a product is linear in the values of attributes. More specifically, we convert the categorical attributes described in Table 1 to attributes with binary

values by adding dummy attributes (for example "price very high", "price low" are considered as two different attributes that can take values 1 or 0). Now every car is associated with an attribute vector $m_i \in \{0,1\}^{22}$, which is known a priori and the mean utility of product i is given by the inner product

$$\mu_i = \theta \cdot m_i \quad i = 1, \ldots, N,$$

where $\theta \in \mathbb{R}^{22}$ is some fixed but initially unknown attribute weight vector. Under this model, the probability that a consumer purchases product i when offered an assortment $S \subset \{1, \ldots, N\}$ is assumed to be,

$$p_i(S) = \begin{cases} \dfrac{e^{\theta \cdot m_i}}{1 + \sum_{j \in S} e^{\theta \cdot m_j}}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{7.2}$$

Let $\mathbf{m} = (m_1, \ldots, m_N)$. Our goal is to offer assortments $S_1, \ldots, S_T$ at times $1, \ldots, T$ respectively such that the cumulative sales are maximized or alternatively, minimize the regret defined as

$$Reg_\pi(T, \mathbf{m}) = \sum_{t=1}^{T} \left(\sum_{i \in S^*} p_i(S) - \sum_{i \in S_t} p_i(S_t) \right), \tag{7.3}$$

where

$$S^* = \arg\max_S \sum_{i \in S} \frac{e^{\theta \cdot m_i}}{1 + \sum_{j \in S} e^{\theta \cdot m_j}}.$$

Note that regret defined in (7.3) is a special case formulation of the regret defined in (2.6) with $r_i = 1$ and $v_i = e^{\theta \cdot m_i}$ for all $i = 1, \ldots, N$.

Experimental setup and results. We first estimate a ground truth MNL model as follows. Using the available attribute level data and consumer rating for each car, we estimate a logistic model assuming every car's rating is independent of the ratings of other cars to estimate the attribute weight vector θ. Specifically, under the logistic model, the probability that a consumer will purchase a car whose attributes are defined by the vector $m \in \{0,1\}^{22}$ and the attribute weight vector θ is given by

$$p_{\mathsf{buy}}(\theta, m) \triangleq \mathbb{P}(\text{buy}|\theta) = \frac{e^{\theta \cdot m}}{1 + e^{\theta \cdot m}}.$$

For the purpose of training the logistic model on the available data, we consider the consumer ratings of "acceptable," "good," and "very good" as success or intention to buy and the consumer rating of "unacceptable" as a failure or no intention to buy. We then use the maximum likelihood estimate θ_{MLE} for θ to run simulations and study the performance of Algorithm 1 for the realized θ_{MLE}. In particular, we compute θ_{MLE} that maximizes the following regularized log-likelihood

$$\theta_{\mathsf{MLE}} = \arg\max_\theta \sum_{i=1}^{N} \log p_{\mathsf{buy}}(\theta, m_i) - \|\theta\|_2.$$

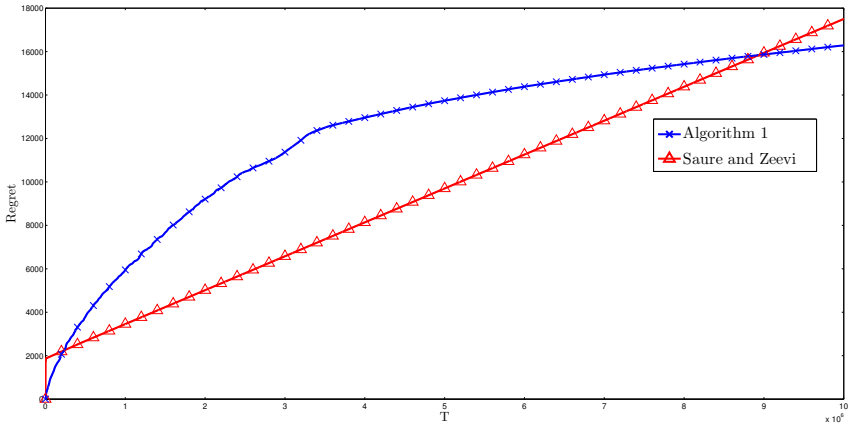

Figure 4 Comparison with the algorithm of Sauré and Zeevi (2013) on real data. The graph compares the performance of Algorithm 1 to that of Sauré and Zeevi (2013) on the "UCI Car Evaluation Databse' for $T = 10^7$.

The objective function in the preceding optimization problem is convex and therefore we can use any of the standard convex optimization techniques to obtain the estimate, θ_{MLE}. It is important to note that the logistic model is only employed to obtain an estimate for θ, θ_{MLE}. The estimate θ_{MLE} is assumed to be the ground truth MNL model and is used to simulate the feedback of consumer choices for our learning Algorithm 1 and the learning algorithm proposed by Sauré and Zeevi (2013).

We compare the performance of Algorithm 1 with that of Sauré and Zeevi (2013), in terms of regret as defined in (7.3) with $\theta = \theta_{\mathsf{MLE}}$ and at each time index, the retailer can only show at most $k = 100$ cars. We implement Sauré and Zeevi (2013)'s approach with their suggested mandotary exploration period, which explores every product for at least $20 \log T$ periods. Figure 4 plots the regret of Algorithm 1 and the Sauré and Zeevi (2013) policy, when the selling horizon is $T = 10^7$. The results are based on running 100 independent simulations and have a standard error of 2%. We can observe that while the initial regret of Sauré and Zeevi (2013) is smaller, the regret grows linearly with time, suggesting that the exploration period was too small. This further illustrates the shortcomings of an explore-then-exploit approach which requires knowledge of underlying parameters. In contrast, the regret of Algorithm 1 grows in a sublinear fashion with respect to the selling horizon and does not require any a priori knowledge on the parameters, making a case for the universal applicability of our approach.

8. Conclusions and future work

Summary and main insights. In this paper, we have studied the dynamic assortment selection problem under the widely used multinomial logit choice model. Formulating the problem as a

parametric multi-arm bandit problem, we present a policy that learns the parameters of the choice model while simultaneously maximizing the cumulative revenue. Focusing on a policy that would be universally applicable, we highlight the limitations of existing approaches and present a novel computationally efficient algorithm, whose performance (as measured by the regret) is nearly-optimal. Furthermore, our policy is adaptive to the complexity of the problem instance, as measured by "separability" of items. The adaptive nature of the algorithm is manifest in its "rate of learning" the unknown instance parameters, which is more rapid if the problem instance is "less complex."

Limitations and future research. In this work we primarily focused on developing an algorithm that would be broadly applicable. In so doing, we only consider the setting where every product has its own utility parameter and has to be estimated separately. However, in many settings a large number of products are effectively described by a small number of product features, via what is often referred to as factor model. An important extension of our problem would be to consider a policy that leverages the relation between products as measured via their features, and achieves a regret bound that is independent of the number of products and only depends on the dimensionality of feature space.

Another interesting direction is to consider the settings where the consumers are heterogeneous. If the consumer type is known a priori, then we can easily generalize our algorithm to learn only model parameters of that type. In a recent work, Kallus and Udell (2016) consider the setting of heterogeneous consumers where each consumer segment follows a separate MNL model, but the underlying structure of these parameters over all the segments has low dimension. Assuming the consumer type is observable a priori, they present an explore first exploit later approach to dynamically learn the preferences of heterogeneous consumer population. Their work also demonstrates significant improvements in performance in comparison to trivially extending the existing dynamic learning approaches (Sauré and Zeevi 2013, Rusmevichientong et al. 2010) to learn a different MNL model for each consumer type. Generalizing our work to design a parameter independent algorithm to learn the preferences of heterogeneous consumers with an underlying low rank structure would be an important extension with significant practical implications.

As discussed earlier, Thompson Sampling is a natural algorithm for the MNL-Bandit problem. Despite being empirically superior to other bandit policies, TS-based algorithms remain challenging to analyze and theoretical work on TS is limited. An interesting direction is to consider a TS-based approach for the MNL-Bandit problem and derive similar regret bounds to the ones obtained in this paper. Due to its combinatorial nature, selecting a suitable prior and efficiently updating the posterior present a significant challenge in designing a good TS-based algorithm for the MNL-Bandit problem. Some preliminary results in this direction are reported in Agrawal et al. (2017).

Acknowledgments

V. Goyal is supported in part by NSF Grants CMMI-1351838 (CAREER) and CMMI-1636046. A. Zeevi is supported in part by NSF Grants NetSE-0964170 and BSF-2010466.

References

Agrawal, S., V. Avadhanula, V. Goyal, A. Zeevi. 2017. Thompson sampling for the mnl-bandit. *Proceedings of Machine Learning Research* **(65)** 76–78.

Agrawal, S., N. Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on International Conference on Machine Learning* **28**.

Agrawal, S., N. Goyal. 2017. Near-optimal regret bounds for thompson sampling. *J. ACM* **64**(5).

Angluin, D., L. G. Valiant. 1977. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Proceedings of the Ninth Annual ACM Symposium on Theory of Computing*. STOC '77, 30–41.

Auer, P. 2003. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research.*

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

Avadhanula, V., J. Bhandari, V. Goyal, A. Zeevi. 2016. On the tightness of an lp relaxation for rational optimization and its applications. *Operations Research Letters* **44**(5) 612–617.

Babaioff, M., S. Dughmi, R. Kleinberg, A. Slivkins. 2015. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation* **3**(1) 4.

Ben-Akiva, M., S. Lerman. 1985. *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press.

Blanchet, J., G. Gallego, V. Goyal. 2016. A markov chain approximation to choice modeling. *Operations Research* **64**(4) 886–905.

Borovkov, AA. 1984. Mathematical statistics. estimation of parameters, testing of hypotheses .

Bubeck, S., N. Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* .

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

Chen, W., Y. Wang, Y. Yuan. 2013. Combinatorial multi-armed bandit: General framework, results and applications. *Proceedings of the 30th international conference on machine learning*. 151–159.

Chen, X., Y. Wang. 2017. A note on tight lower bound for mnl-bandit assortment selection models. *ArXiv e-prints* .

Davis, J., G. Gallego, H. Topaloglu. 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Technical Report, Cornell University.* .

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

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

Désir, A., V. Goyal, D. Segev, C. Ye. 2015. Capacity constrained assortment optimization under the markov chain based choice model. *Working Paper, Columbia University* .

Farias, V., S. Jagabathula, D. Shah. 2013. A nonparametric approach to modeling choice with limited data. *Management Science* **59**(2) 305–322.

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

Gallego, G., R. Ratliff, S. Shebalov. 2014. A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research* **63**(1) 212–232.

Gallego, G., H. Topaloglu. 2014. Constrained assortment optimization for the nested logit model. *Management Science* **60**(10) 2583–2601.

Kallus, N., M. Udell. 2016. Dynamic assortment personalization in high dimensions. *ArXiv e-prints* .

Kleinberg, R., A. Slivkins, E. Upfal. 2008. Multi-armed bandits in metric spaces. *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. STOC '08, 681–690.

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

Lai, T.L., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**(1) 4–22.

Li, G., P. Rusmevichientong, H. Topaloglu. 2015. The d-level nested logit model: Assortment and price optimization problems. *Operations Research* **63**(2) 325–342.

Lichman, M. 2013. UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Luce, R.D. 1959. *Individual choice behavior: A theoretical analysis*. Wiley.

May, B. C., N. Korda, A. Lee, D. S. Leslie. 2012. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* **(13)** 2069–2106.

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

Mitzenmacher, M., E. Upfal. 2005. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press.

Oliver, C., L. Li. 2011. An empirical evaluation of thompson sampling. *In Advances in Neural Information Processing Systems (NIPS)* **24** 2249–2257.

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

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58**(5) 527–535.

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

Rusmevichientong, P., J. N. Tsitsiklis. 2010. Linearly parameterized bandits. *Math. Oper. Res.* **35**(2) 395–411.

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

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

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

Williams, H.C.W.L. 1977. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A* **9**(3) 285–344.

A. Proof of Theorem 1

In this section, we provide a detailed proof of Theorem 1 following the outline discussed in Section 4.1. The proof is organized as follows. In Section A.1, we complete the proof of Lemma 4.1 and in Section A.2, we prove Lemma 4.2 and Lemma 4.3. Finally, in Section A.3, we utilize these results to complete the proof of Theorem 1.

A.1. Properties of estimates $v_{i,\ell}^{\mathsf{UCB}}$: Proof of Lemma 4.1

First, we prove Lemma 4.1. To complete the proof, we establish certain properties of the estimates $v_{i,\ell}^{\mathsf{UCB}}$, and then extend these properties to establish the necessary properties of $\hat{v}_{i,\ell}$ and $\bar{v}_{i,\ell}$.

Lemma A.1 (Moment Generating Function) *The moment generating function of estimate conditioned on S_ℓ, $\hat{v}_i$, is given by,*

$$\mathbb{E}_\pi\left(e^{\theta\hat{v}_{i,\ell}}\middle|S_\ell\right) = \frac{1}{1-v_i(e^\theta-1)}, \text{ for all } \theta \le \log\frac{1+v_i}{v_i}, \text{ for all } i=1,\cdots,N.$$

Proof. From (2.1), we have that probability of no purchase event when assortment S_ℓ is offered is given by

$$p_0(S_\ell) = \frac{1}{1+\sum_{j\in S_\ell} v_j}.$$

Let n_ℓ be the total number of offerings in epoch ℓ before a no purchased occurred, i.e., $n_\ell = |\mathcal{E}_\ell| - 1$. Therefore, n_ℓ is a geometric random variable with probability of success $p_0(S_\ell)$. And, given any

fixed value of n_ℓ, $\hat{v}_{i,\ell}$ is a binomial random variable with n_ℓ trials and probability of success given by

$$q_i(S_\ell) = \frac{v_i}{\sum_{j\in S_\ell} v_j}.$$

In the calculations below, for brevity we use p_0 and q_i respectively to denote $p_0(S_\ell)$ and $q_i(S_\ell)$. Hence, we have

$$\mathbb{E}_\pi\left(e^{\theta\hat{v}_{i,\ell}}\right) = E_{n_\ell}\left\{\mathbb{E}_\pi\left(e^{\theta\hat{v}_{i,\ell}} \,\middle|\, n_\ell\right)\right\}. \tag{A.1}$$

Since the moment generating function for a binomial random variable with parameters n, p is $(pe^\theta + 1 - p)^n$, we have

$$\mathbb{E}_\pi\left(e^{\theta\hat{v}_{i,\ell}} \,\middle|\, n_\ell\right) = \mathbb{E}_{n_\ell}\left\{\left(q_i e^\theta + 1 - q_i\right)^{n_\ell}\right\}. \tag{A.2}$$

For any α, such that $\alpha(1-p) < 1$, if n is a geometric random variable with parameter p, then we have

$$\mathbb{E}(\alpha^n) = \frac{p}{1-\alpha(1-p)}.$$

Since n_ℓ is a geometric random variable with parameter p_0 and by definition of q_i and p_0, we have, $q_i(1-p_0) = v_i p_0$, it follows that for any $\theta < \log\frac{1+v_i}{v_i}$, we have,

$$\mathbb{E}_{n_\ell}\left\{\left(q_i e^\theta + 1 - q_i\right)^{n_\ell}\right\} = \frac{p_0}{1-\left(q_i e^\theta + 1 - q_i\right)(1-p_0)} = \frac{1}{1-v_i(e^\theta - 1)}. \tag{A.3}$$

The result follows from (A.1), (A.2) and (A.3). □

From the moment generating function, we can establish that $\hat{v}_{i,\ell}$ is a geometric random variable with parameter $\frac{1}{1+v_i}$. Thereby also establishing that $\hat{v}_{i,\ell}$ and $\bar{v}_{i,\ell}$ are unbiased estimators of v_i. More specifically, from Lemma A.1, we have the following result.

Corollary A.1 (Unbiased Estimates) *We have the following results.*

1. *$\hat{v}_{i,\ell}$, $\ell \le L$ are i.i.d geometrical random variables with parameter $\frac{1}{1+v_i}$, i .e. for any ℓ, i*

$$\mathbb{P}_\pi\left(\hat{v}_{i,\ell} = m\right) = \left(\frac{v_i}{1+v_i}\right)^m \left(\frac{1}{1+v_i}\right) \quad \forall\, m = \{0,1,2,\cdots\}.$$

2. *$\hat{v}_{i,\ell}$, $\ell \le L$ are unbiased i.i.d estimates of v_i, i .e. $\mathbb{E}_\pi\left(\hat{v}_{i,\ell}\right) = v_i \,\forall\, \ell, i$.*

From Corollary A.1, it follows that $\hat{v}_{i,\tau}, \tau \in \mathcal{T}_i(\ell)$ are i.i.d geometric random variables with mean v_i. We will use this observation and extend the multiplicative Chernoff-Hoeffding bounds discussed in Mitzenmacher and Upfal (2005) and Babaioff et al. (2015) to geometric random variables and derive the following result.

Lemma A.2 (Concentration Bounds) *If $v_i \le v_0$ for all i, for every epoch ℓ, in Algorithm 1, we have the following concentration bounds.*

1. $\mathbb{P}_\pi\left(|\bar{v}_{i,\ell} - v_i| > \sqrt{48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\ell+1)}{T_i(\ell)}\right) \le \frac{6}{N\ell}.$
2. $\mathbb{P}_\pi\left(|\bar{v}_{i,\ell} - v_i| > \sqrt{24v_i\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right) \le \frac{4}{N\ell}.$
3. $\mathbb{P}_\pi\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right) \le \frac{3}{N\ell}.$

Note that to apply standard Chernoff-Hoeffding inequality (see p.66 in Mitzenmacher and Upfal 2005), we must have the individual sample values bounded by some constant, which is not the case with our estimates $\hat{v}_{i,\tau}$. However, these estimates are geometric random variables and therefore have extremely small tails. Hence, we can extend the Chernoff-Hoeffding bounds discussed in Mitzenmacher and Upfal (2005) and Babaioff et al. (2015) to geometric variables and prove the above result. Lemma 4.1 follows directly from Lemma A.2 (see below.) The proof of Lemma A.2 is long and tedious and in the interest of continuity, we complete the proof in Appendix D. Following the proof of Lemma A.2, we obtain a very similar result that is useful to establish concentration bounds for the general parameter setting.

Proof of Lemma 4.1: By design of Algorithm 1, we have,

$$v_{i,\ell}^{\mathsf{UCB}} = \bar{v}_{i,\ell} + \sqrt{48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}. \tag{A.4}$$

Therefore from Lemma A.2, we have

$$\mathcal{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB}} < v_i\right) \le \frac{6}{N\ell}. \tag{A.5}$$

The first inequality in Lemma 4.1 follows from (A.5). From triangle inequality and (A.4), we have,

$$\begin{aligned}\left|v_{i,\ell}^{\mathsf{UCB}} - v_i\right| &\le \left|v_{i,\ell}^{\mathsf{UCB}} - \bar{v}_{i,\ell}\right| + |\bar{v}_{i,\ell} - v_i| \\ &= \sqrt{48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)} + |\bar{v}_{i,\ell} - v_i|.\end{aligned} \tag{A.6}$$

From Lemma A.2, we have

$$\mathbb{P}_\pi\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right) \le \frac{3}{N\ell},$$

which implies

$$\mathbb{P}_\pi\left(48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)} > 72v_i\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)} + \left(\frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right)^2\right) \le \frac{3}{N\ell},$$

Using the fact that $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$, for any positive numbers a, b, we have,

$$\mathbb{P}_\pi \left(\sqrt{48\bar{v}_{i,\ell} \frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)} > \sqrt{72 v_i \frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{96\log(\sqrt{N}\ell+1)}{T_i(\ell)} \right) \leq \frac{3}{N\ell}, \tag{A.7}$$

From Lemma A.2, we have,

$$\mathbb{P}_\pi \left(|\bar{v}_{i,\ell} - v_i| > \sqrt{24 v_i \frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)} \right) \leq \frac{4}{N\ell}. \tag{A.8}$$

From (A.6) and applying union bound on (A.7) and (A.8), we obtain,

$$\mathcal{P} \left(\left| v_{i,\ell}^{\mathsf{UCB}} - v_i \right| > (\sqrt{72} + \sqrt{24}) \sqrt{\frac{v_i \log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{144\log(\sqrt{N}\ell+1)}{T_i(\ell)} \right) \leq \frac{7}{N\ell}.$$

Lemma 4.1 follows from the above inequality and (A.5). □

A.2. Properties of estimate $\tilde{R}(S)$: Proof of Lemma 4.2 and Lemma 4.3

In this section, we prove Lemma 4.2 and Lemma 4.3. To complete the proofs, we will establish two auxiliary results, in the first result (see Lemma A.3) we show that the expected revenue corresponding to the optimal assortment is monotone in the MNL parameters $\mathbf{v}$ and in the second result (see Lemma A.4) we bound the difference between the estimate of the optimal revenue and the true optimal revenue.

Lemma A.3 (Optimistic Estimates) *Assume $0 \leq w_i \leq v_i^{\mathsf{UCB}}$ for all $i = 1, \cdots, n$. Suppose S is an optimal assortment when the MNL are parameters are given by $\mathbf{w}$. Then, $R(S, \mathbf{v}^{UCB}) \geq R(S, \mathbf{w})$.*

Proof. We prove the result by first showing that for any $j \in S$, we have $R(S, \mathbf{w}^j) \geq R(S, \mathbf{w})$, where $\mathbf{w}^j$ is vector $\mathbf{w}$ with the j^{th} component increased to v_j^{UCB}, i.e. $w_i^j = w_i$ for all $i \neq j$ and $w_j^j = v_j^{\mathsf{UCB}}$. We can use this result iteratively to argue that increasing each parameter of MNL to the highest possible value increases the value of $R(S, \mathbf{w})$ to complete the proof.

If there exists $j \in S$ such that $r_j < R(S)$, then removing the product j from assortment S yields higher expected revenue contradicting the optimality of S. Therefore, we have

$$r_j \geq R(S). \ \forall j \in S.$$

Multiplying by $(v_j^{UCB} - w_j)(\sum_{i \in S/j} w_i + 1)$ on both sides of the above inequality and re-arranging terms, we can show that $R(S, \mathbf{w}^j) \geq R(S, \mathbf{w})$. □

We would like to remind the readers that Lemma A.3 does not claim that the expected revenue is in general a monotone function, but only that the value of the expected revenue corresponding to the optimal assortment is monotone in the MNL parameters.

Proof of Lemma 4.2: Let $\hat{S}, \mathbf{w}^*$ be maximizers of the optimization problem,

$$\max_{S\in\mathcal{S}} \max_{0\le w_i\le v_{i,\ell}^{\mathsf{UCB}}} R(S,\mathbf{w}).$$

Assume $v_{i,\ell}^{\mathsf{UCB}} > v_i$ for all i. Then from Lemma A.3 it follows that,

$$\tilde{R}_\ell(S_\ell) = \max_{S\in\mathcal{S}} R(S,\mathbf{v}_\ell^{UCB}) \ge \max_{S\in\mathcal{S}} \max_{0\le w_i\le v_{i,\ell}^{\mathsf{UCB}}} R(S,\mathbf{w}) \ge R(S^*,\mathbf{v}). \tag{A.9}$$

From Lemma 4.1, for each ℓ and $i\in\{1,\cdots,N\}$, we have that,

$$\mathcal{P}\left(v_{i,\ell}^{\mathsf{UCB}} < v_i\right) \le \frac{6}{N\ell}.$$

Hence, from union bound, it follows that,

$$\mathcal{P}\left(\bigcap_{i=1}^{N}\left\{v_{i,\ell}^{\mathsf{UCB}} < v_i\right\}\right) \ge 1-\frac{6}{\ell}. \tag{A.10}$$

Lemma 4.2 follows from (A.9) and (A.10). □ □

Lemma A.4 (Bounding Regret) *If $r_i\in[0,1]$ and $0\le v_i\le v_{i,\ell}^{\mathsf{UCB}}$ for all $i\in S_\ell$, then*

$$\tilde{R}_\ell(S_\ell) - R(S_\ell,\mathbf{v}) \le \frac{\sum_{j\in S_\ell}\left(v_{j,\ell}^{\mathsf{UCB}} - v_j\right)}{1+\sum_{j\in S_\ell} v_j}.$$

Proof. Since $1+\sum_{i\in S_\ell} v_{i,\ell}^{\mathsf{UCB}} \ge 1+\sum_{i\in S_\ell} v_{i,\ell}$, we have

$$\begin{aligned}\tilde{R}_\ell(S_\ell) - R(S_\ell,\mathbf{v}) &\le \frac{\sum_{i\in S_\ell} r_i v_{i,\ell}^{\mathsf{UCB}}}{1+\sum_{j\in S_\ell} v_{j,\ell}^{\mathsf{UCB}}} - \frac{\sum_{i\in S_\ell} r_i v_i}{1+\sum_{j\in S_\ell} v_{j,\ell}^{\mathsf{UCB}}},\\ &\le \frac{\sum_{i\in S_\ell}\left(v_{i,\ell}^{\mathsf{UCB}} - v_i\right)}{1+\sum_{j\in S_\ell} v_{j,\ell}^{\mathsf{UCB}}} \le \frac{\sum_{i\in S_\ell}\left(v_{i,\ell}^{\mathsf{UCB}} - v_i\right)}{1+\sum_{j\in S_\ell} v_j}.\end{aligned}$$

Proof of Lemma 4.3: From Lemma A.4, we have,

$$\left(1+\sum_{j\in S_\ell} v_j\right)\left(\tilde{R}_\ell(S_\ell) - R(S_\ell,\mathbf{v})\right) \le \sum_{j\in S_\ell}\left(v_{j,\ell}^{\mathsf{UCB}} - v_j\right). \tag{A.11}$$

From Lemma 4.1, we have that for each $i=1,\cdots,N$ and ℓ,

$$\mathcal{P}\left(v_{i,\ell}^{\mathsf{UCB}} - v_i > C_1\sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \le \frac{7}{N\ell}.$$

Therefore, from union bound, it follows that,

$$\mathcal{P}\left(\bigcap_{i=1}^{N}\left\{v_{i,\ell}^{\mathsf{UCB}} - v_i < C_1\sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right\}\right) \ge 1-\frac{7}{\ell}. \tag{A.12}$$

Lemma 4.3 follows from (A.11) and (A.12).

A.3. Putting it all together: Proof of Theorem 1

In this section, we utilize the results established in the previous sections and complete the proof of Theorem 1.

Let S^* denote the optimal assortment, our objective is to minimize the *regret* defined in (2.6), which is same as

$$Reg_\pi(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L} |\mathcal{E}_\ell| \left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right)\right\}, \tag{A.13}$$

Note that L, $\mathcal{E}_\ell$ and S_ℓ are all random variables and the expectation in equation (A.13) is over these random variables. Let $\mathcal{H}_\ell$ be the filtration (history) associated with the policy upto epoch ℓ. In particular,

$$\mathcal{H}_\ell = \sigma(U, C_1, \cdots, C_{t(\ell)}, S_1, \cdots, S_{t(\ell)}),$$

where $t(\ell)$ is the time index corresponding to the end of epoch ℓ. The length of the ℓ^{th} epoch, $|\mathcal{E}_\ell|$ conditioned on S_ℓ is a geometric random variable with success probability defined as the probability of no-purchase in S_ℓ, i.e.

$$p_0(S_\ell) = \frac{1}{1+\sum_{j\in S_\ell} v_j}.$$

Let $V(S_\ell) = \sum_{j\in S_\ell} v_j$, then we have $\mathbb{E}_\pi\left(|\mathcal{E}_\ell| \;\middle|\; S_\ell\right) = 1+V(S_\ell)$. Noting that S_ℓ in our policy is determined by $\mathcal{H}_{\ell-1}$, we have $\mathbb{E}_\pi\left(|\mathcal{E}_\ell| \middle| \mathcal{H}_{\ell-1}\right) = 1+V(S_\ell)$. Therefore, by law of conditional expectations, we have

$$Reg_\pi(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L} \mathbb{E}_\pi\left[|\mathcal{E}_\ell| \left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right) \;\middle|\; \mathcal{H}_{\ell-1}\right]\right\},$$

and hence the regret can be reformulated as

$$Reg_\pi(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L} \left(1+V(S_\ell)\right)\left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right)\right\}, \tag{A.14}$$

the expectation in equation (A.14) is over the random variables L and S_ℓ. For the sake of brevity, for each $\ell \in 1,\cdots,L$, let

$$\Delta R_\ell = \left(1+V(S_\ell)\right)\left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right). \tag{A.15}$$

Now the regret can be reformulated as

$$Reg_\pi(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L} \Delta R_\ell\right\}. \tag{A.16}$$

Let T_i denote the total number of epochs that offered an assortment containing product i. For all $\ell = 1,\ldots,L$, define events $\mathcal{A}_\ell$ as,

$$\mathcal{A}_\ell = \bigcup_{i=1}^{N}\left\{v_{i,\ell}^{\mathsf{UCB}} < v_i \text{ or } v_{i,\ell}^{\mathsf{UCB}} > v_i + C_1\sqrt{\frac{v_i \log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right\}.$$

From union bound, it follows that

$$\begin{aligned}\mathbb{P}_\pi\left(\mathcal{A}_\ell\right) &\leq \sum_{i=1}^{N} \mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB}} < v_i \text{ or } v_{i,\ell}^{\mathsf{UCB}} > v_i + C_1\sqrt{\frac{v_i \log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right),\\ &\leq \sum_{i=1}^{N} \mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB}} < v_i\right) + \mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB}} > v_i + C_1\sqrt{\frac{v_i \log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right).\end{aligned}$$

Therefore, from Lemma 4.1, we have,

$$\mathbb{P}_\pi(\mathcal{A}_\ell) \leq \frac{13}{\ell}. \tag{A.17}$$

Since $\mathcal{A}_\ell$ is a "low probability" event (see (A.17)), we analyze the regret in two scenarios, one when $\mathcal{A}_\ell$ is true and another when $\mathcal{A}_\ell^c$ is true. We break down the regret in an epoch into the following two terms:

$$\mathbb{E}_\pi\left(\Delta R_\ell\right) = E\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right].$$

Using the fact that $R(S^*, \mathbf{v})$ and $R(S_\ell, \mathbf{v})$ are both bounded by one and $V(S_\ell) \leq N$ in (A.15), we have $\Delta R_\ell \leq N+1$. Substituting the preceding inequality in the above equation, we obtain,

$$\mathbb{E}_\pi\left(\Delta R_\ell\right) \leq (N+1)\mathbb{P}_\pi(\mathcal{A}_{\ell-1}) + \mathbb{E}_\pi\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right].$$

Whenever $\mathbb{1}(\mathcal{A}_{\ell-1}^c) = 1$, from Lemma A.3, we have $\tilde{R}_\ell(S^*) \geq R(S^*, \mathbf{v})$ and by our algorithm design, we have $\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*)$ for all $\ell \geq 1$. Therefore, it follows that

$$\mathbb{E}_\pi\left\{\Delta R_\ell\right\} \leq (N+1)\mathbb{P}_\pi(\mathcal{A}_{\ell-1}) + \mathbb{E}_\pi\left\{\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v}))\right] \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right\}.$$

From the definition of the event, $\mathcal{A}_\ell$ and Lemma A.4, it follows that,

$$\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v}))\right] \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \leq \sum_{i \in S_\ell}\left(C_1\sqrt{\frac{v_i \log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{C_2 \log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right).$$

Therefore, we have

$$\mathbb{E}_\pi\left\{\Delta R_\ell\right\} \leq (N+1)\mathbb{P}_\pi\left(\mathcal{A}_{\ell-1}\right) + C\sum_{i \in S_\ell}\mathbb{E}_\pi\left(\sqrt{\frac{v_i \log\sqrt{N}T}{T_i(\ell)}} + \frac{\log\sqrt{N}T}{T_i(\ell)}\right), \tag{A.18}$$

where $C = \max\{C_1, C_2\}$. Combining equations (A.14) and (A.18), we have

$$Reg_\pi(T, \mathbf{v}) \leq \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L}\left[(N+1)\mathbb{P}_\pi\left(\mathcal{A}_{\ell-1}\right) + C\sum_{i \in S_\ell}\left(\sqrt{\frac{v_i \log\sqrt{N}T}{T_i(\ell)}} + \frac{\log\sqrt{N}T}{T_i(\ell)}\right)\right]\right\}.$$

Therefore, from Lemma 4.1, we have

$$\begin{aligned} Reg_\pi(T,\mathbf{v}) &\le C\mathbb{E}_\pi\left\{\sum_{\ell=1}^{L}\frac{N+1}{\ell}+\sum_{i\in S_\ell}\sqrt{\frac{v_i\log\sqrt{N}T}{T_i(\ell)}}+\sum_{i\in S_\ell}\frac{\log\sqrt{N}T}{T_i(\ell)}\right\},\\ &\overset{(a)}{\le} CN\log T+CN\log^2\sqrt{N}T+C\mathbb{E}_\pi\left(\sum_{i=1}^{n}\sqrt{v_iT_i\log\sqrt{N}T}\right), \qquad \text{(A.19)}\\ &\overset{(b)}{\le} CN\log T+CN\log^2 NT+C\sum_{i=1}^{N}\sqrt{v_i\log(NT)\mathbb{E}_\pi(T_i)}. \end{aligned}$$

Inequality (a) follows from the observation that $L\le T$, $T_i\le T$,

$$\sum_{T_i(\ell)=1}^{T_i}\frac{1}{\sqrt{T_i(\ell)}}\le\sqrt{T_i},\text{ and }\sum_{T_i(\ell)=1}^{T_i}\frac{1}{T_i(\ell)}\le\log T_i,$$

while Inequality (b) follows from Jensen's inequality.

For any realization of L, $\mathcal{E}_\ell$, T_i, and S_ℓ in Algorithm 1, we have the following relation

$$\sum_{\ell=1}^{L}n_\ell\le T.$$

Hence, we have $\mathbb{E}_\pi\left(\sum_{\ell=1}^{L}n_\ell\right)\le T$. Let $\mathcal{F}$ denote the filtration corresponding to the offered assortments $S_1,\cdots,S_L$, then by law of total expectation, we have,

$$\begin{aligned}\mathbb{E}_\pi\left(\sum_{\ell=1}^{L}n_\ell\right)&=\mathbb{E}_\pi\left\{\sum_{\ell=1}^{L}E_{\mathcal{F}}\left(n_\ell\right)\right\}=\mathbb{E}_\pi\left\{\sum_{\ell=1}^{L}1+\sum_{i\in S_\ell}v_i\right\},\\ &=\mathbb{E}_\pi\left\{L+\sum_{i=1}^{n}v_iT_i\right\}=\mathbb{E}_\pi\{L\}+\sum_{i=1}^{n}v_i\mathbb{E}_\pi(T_i).\end{aligned}$$

Therefore, it follows that

$$\sum v_i\mathbb{E}_\pi(T_i)\le T. \qquad \text{(A.20)}$$

To obtain the worst case upper bound, we maximize the bound in equation (A.19) subject to the condition (A.20) and hence, we have $Reg_\pi(T,\mathbf{v})=O(\sqrt{NT\log NT}+N\log^2 NT)$. □

A.4. Improved regret bounds for the unconstrained MNL-Bandit

Here, we focus on the special case of the unconstrained MNL-Bandit problem and use the analysis of Appendix A.3 to establish a tighter bound on the regret for Algorithm 1. First, we note that, in the case of the unconstrained problem, for any epoch ℓ, with high probability, the assortment, S_ℓ suggested by Algorithm 1 is a subset of the optimal assortment, S^*. More specifically, the following holds.

Lemma A.5 *Let $S^* = \underset{S\in\{1,\cdots,N\}}{argmax}\ R(S,\mathbf{v})$ and S_ℓ be the assortment suggested by Algorithm 1. Then for any $\ell = 1,\cdots,L$, we have,*

$$\mathbb{P}_\pi\left(S_\ell \subset S^*\right) \geq 1 - \frac{6}{\ell}.$$

Proof. If there exists a product i, such that $r_i \geq R(S^*,\mathbf{v})$, then following the proof of Lemma A.3, we can show that $R(S^*\cup i,\mathbf{v}) \geq R(S^*,\mathbf{v})$ and similarly, if there exists a product i, such that $r_i < R(S^*,\mathbf{v})$, we can show that $R(S^*\backslash\{i\},\mathbf{v}) \geq R(S^*,\mathbf{v})$. Since there are no constraints on the set of feasible assortment, we can add and remove products that will improve the expected revenue. Therefore, we have,

$$i \in S^* \text{ if and only if } r_i \geq R(S^*,\mathbf{v}). \tag{A.21}$$

Fix an epoch ℓ, let S_ℓ be the assortment suggested by Algorithm 1. Using similar arguments as above, we can show that,

$$i \in S_\ell \text{ if and only if } r_i \geq R(S_\ell,\mathbf{v}_\ell^{\mathsf{UCB}}). \tag{A.22}$$

From Lemma 4.2, we have ,

$$\mathbb{P}_\pi\left(R(S_\ell,\mathbf{v}_\ell^{\mathsf{UCB}}) \geq R(S^*,\mathbf{v})\right) \geq 1 - \frac{6}{\ell}. \tag{A.23}$$

Lemma A.5 follows from (A.21), (A.22) and (A.23). □

From Lemma A.5, it follows that Algorithm 1 only considers products from the set S^* with high probability, and hence, we can follow the proof in Appendix A.3 (by replacing N with $|S^*|$) to derive sharper regret bounds. In particular, we have the following result,

Corollary A.2 (Performance Bounds for unconstrained case) *For any instance, $\mathbf{v} = (v_0,\ldots,v_N)$ of the MNL-Bandit problem with N products and no constraints, $r_i \in [0,1]$ and $v_0 \geq v_i$ for $i = 1,\ldots,N$, there exists finite constants C_1 and C_2, such that the regret of the policy defined in Algorithm 1 at any time T is bounded as,*

$$Reg_\pi(T,\mathbf{v}) \leq C_1\sqrt{|S^*|T\log NT} + C_2 N\log NT.$$

B. Proof of Theorem 4

The proof for Theorem 4 is very similar to the proof of Theorem 1. Specifically, we first prove that the initial exploratory phase is indeed bounded and then follow the proof of Theorem 1 to establish the correctness of confidence intervals, optimistic assortment and finally deriving the convergence rates and regret bounds.

Bounding Exploratory Epochs. We would denote an epoch ℓ as an "exploratory epoch" if the assortment offered in the epoch contains a product that has been offered in less than

$48\log(\sqrt{N}\ell+1)$ epochs. It is easy to see that the number of exploratory epochs is bounded by $48N\log NT$, where T is the selling horizon under consideration. We then use the observation that the length of any epoch is a geometric random variable to bound the total expected duration of the exploration phase. Hence, we bound the expected regret due to explorations.

Lemma B.1 *Let L be the total number of epochs in Algorithm 3 and let $\mathcal{E}_L$ denote the set of "exploratory epochs," i.e.*

$$E_L = \left\{ \ell \;\middle|\; \exists\, i \in S_\ell \text{ such that } T_i(\ell) < 48\log(\sqrt{N}\ell+1) \right\},$$

where $T_i(\ell)$ is the number of epochs product i has been offered before epoch ℓ. If $\mathcal{E}_\ell$ denote the time indices corresponding to epoch ℓ and $v_i \le Bv_0$ for all $i=1,\dots,N$, for some $B \ge 1$, then we have that,

$$\mathbb{E}_\pi\left(\sum_{\ell\in E_L} |\mathcal{E}_\ell|\right) < 49NB\log NT,$$

where the expectation is over all possible outcomes of Algorithm 3.

Proof. Consider an $\ell \in E_L$, note that $|\mathcal{E}_\ell|$ is a geometric random variable with parameter $1/V(S_\ell)+1$. Since $v_i \le Bv_0$, for all i and we can assume without loss of generality $v_0 = 1$, we have $|\mathcal{E}_\ell|$ as a geometric random variable with parameter p, where $p \ge 1/(B|S_\ell|+1)$. Therefore, we have the conditional expectation of $|\mathcal{E}_\ell|$ given that assortment S_ℓ is offered is bounded as,

$$\mathbb{E}_\pi\left(|\mathcal{E}_\ell| \mid S_\ell\right) \le B|S_\ell|+1. \tag{B.1}$$

Note that after every product has been offered in at least $48\log NT$ epochs, then we do not have any exploratory epochs. Therefore, we have that

$$\sum_{\ell\in E_L} |S_\ell| \le 48N\log NT.$$

Substituting the above inequality in (B.1), we obtain

$$\mathbb{E}_\pi\left(\sum_{\ell\in E_L} |\mathcal{E}_\ell|\right) \le 48BN\log NT + 48N\log NT. \quad \square$$

Confidence Intervals. We will now show a result analogous to Lemma 4.1, that establish the updates in Algorithm 3, $v_{i,\ell}^{\mathsf{UCB2}}$, as upper confidence bounds converging to actual parameters v_i. Specifically, we have the following result.

Lemma B.2 *For every epoch ℓ, if $T_i(\ell) \ge 48\log(\sqrt{N}\ell+1)$ for all $i \in S_\ell$, then we have,*

1. *$v_{i,\ell}^{\mathsf{UCB2}} \ge v_i$ with probability at least $1-\frac{6}{N\ell}$ for all $i=1,\cdots,N$.*

2. *There exists constants C_1 and C_2 such that*

$$v_{i,\ell}^{\mathsf{UCB2}} - v_i \le C_1 \max\left\{\sqrt{v_i}, v_i\right\}\sqrt{\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)},$$

with probability at least $1-\frac{7}{N\ell}$.

The proof is very similar to the proof of Lemma 4.1, where we first establish the following concentration inequality for the estimates $\hat{v}_{i,\ell}$, when $T_i(\ell) \ge 48\log\left(\sqrt{N}\ell+1\right)$ from which the above result follows. The proof of Lemma B.3 is provided in Appendix D.

Lemma B.3 *If in epoch ℓ, $T_i(\ell) \ge 48\log\left(\sqrt{N}\ell+1\right)$ for all $i \in S_\ell$, then we have the following concentration bounds*

1. $\mathbb{P}_\pi\left(|\bar{v}_{i,\ell} - v_i| \ge \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{n}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \le \frac{6}{N\ell}.$
2. $\mathbb{P}_\pi\left(|\bar{v}_{i,\ell} - v_i| \ge \max\left\{\sqrt{v_i}, v_i\right\}\sqrt{\frac{24\log\left(\sqrt{N}\ell+1\right)}{n}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \le \frac{4}{N\ell}.$
3. $\mathbb{P}_\pi\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \le \frac{3}{N\ell}$.

Proof of Lemma B.2: By design of Algorithm 3, we have,

$$v_{i,\ell}^{\mathsf{UCB2}} = \bar{v}_{i,\ell} + \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}. \tag{B.2}$$

Therefore from Lemma B.3, we have

$$\mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB2}} < v_i\right) \le \frac{6}{N\ell}. \tag{B.3}$$

The first inequality in Lemma 4.1 follows from (B.3). From (B.2), we have,

$$\begin{aligned}\left|v_{i,\ell}^{\mathsf{UCB2}} - v_i\right| &\le \left|v_{i,\ell}^{\mathsf{UCB}} - \bar{v}_{i,\ell}\right| + \left|\bar{v}_{i,\ell} - v_i\right| \\ &= \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{48\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)} + \left|\bar{v}_{i,\ell} - v_i\right|.\end{aligned} \tag{B.4}$$

From Lemma B.3, we have

$$\mathbb{P}_\pi\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \le \frac{3}{N\ell},$$

which implies

$$\mathbb{P}_\pi\left(48\bar{v}_{i,\ell}\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)} > 72v_i\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)} + \left(\frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right)^2\right) \le \frac{3}{N\ell},$$

Using the fact that $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$, for any positive numbers a, b, we have,

$$\mathbb{P}_\pi\left(\max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{48\bar{v}_{i,\ell}\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} > \max\left\{\sqrt{v_i}, v_i\right\}\sqrt{72\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{3}{N\ell}, \tag{B.5}$$

From Lemma B.3, we have,

$$\mathbb{P}_\pi\left(|\bar{v}_{i,\ell} - v_i| > \max\left\{\sqrt{v_i}, v_i\right\}\sqrt{24\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{4}{N\ell}. \tag{B.6}$$

From (B.4) and applying union bound on (B.5) and (B.6), we obtain,

$$\mathbb{P}_\pi\left(\left|v_{i,\ell}^{\mathsf{UCB2}} - v_i\right| > (\sqrt{72}+\sqrt{24})\max\left\{\sqrt{v_i}, v_i\right\}\sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{144\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{7}{N\ell}.$$

Lemma B.2 follows from the above inequality and (B.3). □

Optimistic Estimate and Convergence Rates: We will now establish two results analogous to Lemma 4.2 and 4.3, that show that the estimated revenue converges to the optimal expected revenue from above and also specify the convergence rate. In particular, we have the following two results. The proofs of Lemma B.4 and B.5 follow similar arguments to the proofs of Lemma 4.2 and 4.3 respectively and we skip the proofs in interest of avoiding redundancy.

Lemma B.4 *Suppose $S^* \in \mathcal{S}$ is the assortment with highest expected revenue, and Algorithm 3 offers $S_\ell \in \mathcal{S}$ in each epoch ℓ. Further, if $T_i(\ell) \geq 48\log\left(\sqrt{N}\ell+1\right)$ for all $i \in S_\ell$, then we have,*

$$\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*) \geq R(S^*, \mathbf{v}) \text{ with probability at least } 1 - \frac{6}{N\ell}.$$

Lemma B.5 *For every epoch ℓ, if $r_i \in [0,1]$ and $T_i(\ell) \geq 48\log\left(\sqrt{N}\ell+1\right)$ for all $i \in S_\ell$, then there exists constants C_1 and C_2 such that for every ℓ, we have*

$$(1+\textstyle\sum_{j\in S_\ell} v_j)(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v})) \leq C_1\max\left\{\sqrt{v_i}, v_i\right\}\sqrt{\frac{\log\left(\sqrt{N}\ell+1\right)}{|\mathcal{T}_i(\ell)|}} + C_2\frac{\log\left(\sqrt{N}\ell+1\right)}{|\mathcal{T}_i(\ell)|},$$

with probability at least $1 - \frac{13}{N\ell}$.

B.1. Putting it all together: Proof of Theorem 4

Proof of Theorem 4 is very similar to the proof of Theorem 1. We use the key results discussed above instead of similar results in Section 4 to complete the proof. Note that E_ℓ is the set of "exploratory epochs," i.e. epochs in which at least one of the offered product is offered less than the required number of times. We breakdown the regret as follows:

$$Reg_\pi(T, \mathbf{v}) = \underbrace{\mathbb{E}_\pi\left\{\sum_{\ell\in E_L}|\mathcal{E}_\ell|\left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}}_{Reg_1(T,\mathbf{v})} + \underbrace{\mathbb{E}_\pi\left\{\sum_{\ell\notin E_L}|\mathcal{E}_\ell|\left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}}_{Reg_2(T,\mathbf{v})}.$$

Since for any S, we have, $R(S,\mathbf{v}) \le R(S^*,\mathbf{v}) \le 1$, it follows that,

$$Reg_\pi(T,\mathbf{v}) \le \mathbb{E}_\pi\left\{\sum_{\ell\in E_L}|\mathcal{E}_\ell|\right\} + Reg_2(T,\mathbf{v}).$$

From Lemma B.1, it follows that,

$$Reg_\pi(T,\mathbf{v}) \le 49NB\log NT + Reg_2(T,\mathbf{v}). \tag{B.7}$$

We will focus on the second term in the above equation, $Reg_2(T,\mathbf{v})$. Following the analysis in Appendix A.3, we can show that,

$$Reg_2(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell\notin E_L}(1+V(S_\ell))\left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right)\right\}. \tag{B.8}$$

Similar to the analysis in Appendix A.3, for the sake of brevity, we define,

$$\Delta R_\ell = (1+V(S_\ell))\left(R(S^*,\mathbf{v}) - R(S_\ell,\mathbf{v})\right). \tag{B.9}$$

Now, $Reg_2(T,\mathbf{v})$ can be reformulated as

$$Reg_2(T,\mathbf{v}) = \mathbb{E}_\pi\left\{\sum_{\ell\notin E_L}\Delta R_\ell\right\}. \tag{B.10}$$

Let T_i denote the total number of epochs that offered an assortment containing product i. For all $\ell = 1,\dots,L$, define events $\mathcal{B}_\ell$ as,

$$\mathcal{B}_\ell = \bigcup_{i=1}^{N}\left\{v_{i,\ell}^{\mathsf{UCB2}} < v_i \text{ or } v_{i,\ell}^{\mathsf{UCB2}} > v_i + C_1\max\{\sqrt{v_i}, v_i\}\sqrt{\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + C_2\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right\}.$$

From union bound, it follows that

$$\begin{aligned}\mathbb{P}_\pi(\mathcal{B}_\ell) &\le \sum_{i=1}^{N}\mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB2}} < v_i \text{ or } v_{i,\ell}^{\mathsf{UCB2}} > v_i + C_1\max\{\sqrt{v_i}, v_i\}\sqrt{\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + C_2\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right),\\ &\le \sum_{i=1}^{N}\mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB2}} < v_i\right) + \mathbb{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB2}} > v_i + C_1\max\{\sqrt{v_i}, v_i\}\sqrt{\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + C_2\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right).\end{aligned}$$

Therefore, from Lemma B.2, we have,

$$\mathbb{P}_\pi(\mathcal{B}_\ell) \le \frac{13}{\ell}. \tag{B.11}$$

Since $\mathcal{B}_\ell$ is a "low probability" event (see (B.11)), we analyze the regret in two scenarios: one when $\mathcal{B}_\ell$ is true and another when $\mathcal{B}_\ell^c$ is true. We break down the regret in an epoch into the following two terms.

$$\mathbb{E}_\pi(\Delta R_\ell) = E\left[\Delta R_\ell\cdot\mathbb{1}(\mathcal{B}_{\ell-1}) + \Delta R_\ell\cdot\mathbb{1}(\mathcal{B}_{\ell-1}^c).\right]$$

Using the fact that $R(S^*,\mathbf{v})$ and $R(S_\ell,\mathbf{v})$ are both bounded by one and $V(S_\ell)\le BN$ in (B.9), we have $\Delta R_\ell \le N+1$. Substituting the preceding inequality in the above equation, we obtain,

$$\mathbb{E}_\pi\left(\Delta R_\ell\right) \le B(N+1)\mathbb{P}_\pi(\mathcal{B}_{\ell-1}) + \mathbb{E}_\pi\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{B}^c_{\ell-1})\right].$$

Whenever $\mathbb{1}(\mathcal{B}^c_{\ell-1})=1$, from Lemma A.3, we have $\tilde{R}_\ell(S^*)\ge R(S^*,\mathbf{v})$ and by our algorithm design, we have $\tilde{R}_\ell(S_\ell)\ge \tilde{R}_\ell(S^*)$ for all $\ell\ge 1$. Therefore, it follows that

$$\mathbb{E}_\pi\left\{\Delta R_\ell\right\} \le B(N+1)\mathbb{P}_\pi(\mathcal{B}_{\ell-1}) + \mathbb{E}_\pi\left\{\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell)-R(S_\ell,\mathbf{v}))\right]\cdot\mathbb{1}(\mathcal{B}^c_{\ell-1})\right\}. \tag{B.12}$$

From the definition of the event, $\mathcal{B}_\ell$ and Lemma B.5, we have,

$$\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell)-R(S_\ell,\mathbf{v}))\right]\cdot\mathbb{1}(\mathcal{B}^c_{\ell-1}) \le \sum_{i\in S_\ell}\left(C_1\max\{v_i,\sqrt{v_i}\}\sqrt{\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{C_2\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right),$$

and therefore, substituting above inequality in (B.12), we have

$$\mathbb{E}_\pi\left\{\Delta R_\ell\right\} \le B(N+1)\mathbb{P}_\pi\left(\mathcal{B}_{\ell-1}\right) + C\sum_{i\in S_\ell}\mathbb{E}_\pi\left(\max\{v_i,\sqrt{v_i}\}\sqrt{\frac{\log\sqrt{N}T}{T_i(\ell)}} + \frac{\log\sqrt{N}T}{T_i(\ell)}\right), \tag{B.13}$$

where $C=\max\{C_1,C_2\}$. Combining equations (B.7), (B.10) and (B.13), we have

$$\begin{aligned} Reg_\pi(T,\mathbf{v}) \le\ & 49BN\log NT + \mathbb{E}_\pi\left\{\sum_{\ell=1}^{L} B(N+1)\mathbb{P}_\pi\left(\mathcal{A}_{\ell-1}\right)\right\} \\ & + \sum_{\ell=1}^{L}\mathbb{E}_\pi\left[C\max\{v_i,\sqrt{v_i}\}\sum_{i\in S_\ell}\left(\sqrt{\frac{\log\sqrt{N}T}{T_i(\ell)}} + \frac{\log\sqrt{N}T}{T_i(\ell)}\right)\right]. \end{aligned}$$

Define sets $\mathcal{I}=\{i|v_i\ge 1\}$ and $\mathcal{D}=\{i|v_i<1\}$. Therefore, we have,

$$\begin{aligned} Reg_\pi(T,\mathbf{v}) &\le 98NB\log NT + C\mathbb{E}_\pi\left\{\sum_{\ell=1}^{L}\sum_{i\in S_\ell}\left(\max\left\{\sqrt{v_i},v_i\right\}\sqrt{\frac{\log\sqrt{N}T}{T_i(\ell)}} + \frac{\log\sqrt{N}T}{T_i(\ell)}\right)\right\}, \\ &\overset{(a)}{\le} 98NB\log NT + CN\log^2 NT + C\mathbb{E}_\pi\left(\sum_{i\in\mathcal{D}}\sqrt{v_iT_i\log NT} + \sum_{i\in\mathcal{I}} v_i\sqrt{T_i\log NT}\right), \\ &\overset{(b)}{\le} 98NB\log NT + CN\log^2 NT + C\sum_{i\in\mathcal{D}}\sqrt{v_i\mathbb{E}_\pi(T_i)\log NT} + \sum_{i\in\mathcal{I}} v_i\sqrt{\mathbb{E}_\pi(T_i)\log NT}, \end{aligned} \tag{B.14}$$

inequality (a) follows from the observation that $\sqrt{N}\le N$,$L\le T$, $T_i\le T$,

$$\sum_{T_i(\ell)=1}^{T_i}\frac{1}{\sqrt{T_i(\ell)}} \le \sqrt{T_i} \text{ and } \sum_{T_i(\ell)=1}^{T_i}\frac{1}{T_i(\ell)} \le \log T_i,$$

while inequality (b) follows from Jensen's inequality. From (A.20), we have that,

$$\sum v_i\mathbb{E}_\pi(T_i) \le T.$$

To obtain the worst case upper bound, we maximize the bound in equation (B.14) subject to the above constraint. Noting that the objective in (B.14) is concave, we use the KKT conditions to derive the worst case bound as $Reg_\pi(T,\mathbf{v}) = O(\sqrt{BNT\log NT} + N\log^2 NT + BN\log NT)$. □

C. Improved regret bounds for "well separated" instances

Proof of Lemma 6.1: Let $V(S_\ell) = \sum_{i\in S_\ell} v_i$. From Lemma 4.3, and definition of τ (see (6.2)), we have,

$$\begin{aligned} R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v}) &\le \frac{1}{V(S_\ell)+1} \sum_{i\in S_\ell} \left(C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell+1)}{T_i(\ell)}} + C_2 \frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)} \right), \\ &\le \Delta(\mathbf{v}) \left(\frac{C_1 \sum_{i\in S_\ell} \sqrt{v_i}}{2\sqrt{NC}\,(V(S_\ell)+1)} + \frac{C_2}{4C} \right). \end{aligned} \tag{C.1}$$

From Cauchy-Schwartz inequality, we have

$$\sum_{i\in S_\ell} \sqrt{v_i} \le \sqrt{|S_\ell| \sum_{i\in S_\ell} v_i} \le \sqrt{N V(S_\ell)} \le \sqrt{N}\,(V(S_\ell)+1).$$

Substituting the above inequality in (C.1) and using the fact that $C = \max\{C_1^2, C_2\}$, we obtain $R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v}) \le \frac{3\Delta(\mathbf{v})}{4}$. The result follows from the definition of $\Delta(\mathbf{v})$. □

Proof of Lemma 6.2: We complete the proof using an inductive argument on N.

Lemma 6.2 trivially holds for $N = 1$, since when there is only one product, every epoch offers the optimal product and the number of epochs offering sub-optimal assortment is 0, which is less than τ. Now assume that for any $N \le M$, we have that the number of "good epochs" offering sub-optimal products is bounded by $N\tau$, where τ is as defined in (6.2).

Now consider the setting, $N = M+1$. We will now show that the number of "good epochs" offering sub-optimal products cannot be more than $(M+1)\tau$ to complete the induction argument. We introduce some notation, let $\hat{N}$ be the number of products that are offered in more than τ epochs by Algorithm 1, $\mathcal{E}_\mathcal{G}$ denote the set of "good epochs", i.e.,

$$\mathcal{E}_\mathcal{G} = \left\{ \ell \;\middle|\; v_{i,\ell}^{\mathsf{UCB}} \ge v_i \text{ or } v_{i,\ell}^{\mathsf{UCB}} \le v_i + C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell+1)}{T_i(\ell)}} + C_2 \frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)} \text{ for all } i \right\}, \tag{C.2}$$

and $\mathcal{E}_\mathcal{G}^{\mathsf{sub\_opt}}$ be the set of "good epochs" that offer sub-optimal assortments,

$$\mathcal{E}_\mathcal{G}^{\mathsf{sub\_opt}} = \{\ell \in \mathcal{E}_\mathcal{G} \mid R(S_\ell) < R(S^*)\}. \tag{C.3}$$

Case 1: $\hat{N} = N$: Let L be the total number of epochs and $S_1, \cdots, S_L$ be the assortments offered by Algorithm 1 in epochs $1, \cdots, L$ respectively. Let ℓ_i be the epoch that offers product i for the τ^{th} time, specifically,

$$\ell_i \triangleq \min\{\ell \mid T_i(\ell) = \tau\}.$$

Without loss of generality, assume that, $\ell_1 \le \ell_2 \le \cdots \le \ell_N$. Let $\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}$ be the set of "good epochs" that offered sub-optimal assortments before epoch ℓ_{N-1},

$$\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}} = \left\{ \ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} \,\middle|\, \ell \le \ell_{N-1} \right\},$$

where $\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}$ is as defined as in (C.3). Finally, let $\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}}$ be the set of "good epochs" that offered sub-optimal assortments not containing product N before epoch ℓ_{N-1},

$$\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}} = \left\{ \ell \in \hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}} \,\middle|\, N \notin S_\ell \right\}.$$

Every assortment S_ℓ offered in epoch $\ell \in \hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}}$ can contain at most $N-1 = M$ products, therefore by the inductive hypothesis, we have $|\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}}| \le M\tau$. We can partition $\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}$ as,

$$\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}} = \hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}} \cup \left\{ \ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} \,\middle|\, N \in S_\ell \right\},$$

and hence it follows that,

$$|\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}| \le M\tau + \left| \left\{ \ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} \,\middle|\, N \in S_\ell \right\} \right|.$$

Note that $T_N(\ell_{N-1})$ is the number of epochs until epoch ℓ_{N-1}, in which product N has been offered. Hence, it is higher than the number of "good epochs" before epoch ℓ_{N-1} that offered a sub-optimal assortment containing product N and it follows that,

$$|\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}| \le M\tau + T_N(\ell_{N-1}). \tag{C.4}$$

Note that from Lemma 6.1, we have that any "good epoch" offering sub-optimal assortment must offer product N, since all the the other products have been offered in at least τ epochs. Therefore, we have, for any $\ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} \backslash \hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}$, $N \in S_\ell$ and thereby,

$$T_N(\ell_N) - T_N(\ell_{N-1}) \ge |\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}| - |\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}|.$$

From definition of ℓ_N, we have that $T_N(\ell_N) = \tau$ and substituting (C.4) in the above inequality, we obtain

$$|\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}| \le (M+1)\tau.$$

Case 2: $\hat{N} < N$: The proof for the case when $\hat{N} < N$ is similar along the lines of the previous case (we will make the same arguments using $\hat{N} - 1$ instead of $N - 1$.) and is skipped in the interest of avoiding redundancy. □

Following the proof of Lemma 6.2, we can establish the following result.

Corollary C.1 *The number of epochs that offer a product that does not satisfy the condition, $T_i(\ell) \geq \log NT$, is bounded by $N \log NT$. In particular,*

$$\left|\left\{\ell \;\middle|\; T_i(\ell) < \log NT \text{ for some } i \in S_\ell\right\}\right| \leq N \log NT.$$

Proof of Theorem 3: We will re-use the ideas from proof of Theorem 1 to prove Theorem 3. Briefly, we breakdown the regret into regret over "good epochs" and "bad epochs." First we argue using Lemma 4.1, that the probability of an epoch being "bad epoch" is "small," and hence the expected cumulative regret over the bad epochs is "small." We will then use Lemma 6.2 to argue that there are only "small" number of "good epochs" that offer sub-optimal assortments. Since, Algorithm 1 do not incur regret in epochs that offer the optimal assortment, we can replace the length of the horizon T with the cumulative length of the time horizon that offers sub-optimal assortments (which is "small") and re-use analysis from Appendix A.3. We will now make these notions rigorous and complete the proof of Theorem 3.

Following the analysis in Appendix A.3, we reformulate the regret as

$$Reg_\pi(T, \mathbf{v}) = \mathbb{E}_\pi \left\{ \sum_{\ell=1}^{L} \left(1 + V(S_\ell)\right) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right) \right\}, \tag{C.5}$$

where S^* is the optimal assortment, $V(S_\ell) = \sum_{j \in S_\ell} v_j$ and the expectation in equation (C.5) is over the random variables L and S_ℓ. Similar to the analysis in Appendix A.3, for the sake of brevity, we define,

$$\Delta R_\ell = \left(1 + V(S_\ell)\right) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right). \tag{C.6}$$

Now the regret can be reformulated as

$$Reg_\pi(T, \mathbf{v}) = \mathbb{E}_\pi \left\{ \sum_{\ell=1}^{L} \Delta R_\ell \right\}. \tag{C.7}$$

For all $\ell = 1, \ldots, L$, define events $\mathcal{A}_\ell$ as,

$$\mathcal{A}_\ell = \bigcup_{i=1}^{N} \left\{ v_{i,\ell}^{\mathsf{UCB}} < v_i \text{ or } v_{i,\ell}^{\mathsf{UCB}} > v_i + C_1 \sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}} + C_2 \frac{\log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)} \right\}.$$

Let $\xi = \left\{ \ell \;\middle|\; T_i(\ell) < \log NT \text{ for some } i \in S_\ell \right\}$. We breakdown the regret in an epoch into the following terms.

$$\mathbb{E}_\pi \left(\Delta R_\ell\right) = \mathbb{E}_\pi \left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \mathbb{1}(\ell \in \xi) + \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \mathbb{1}(\ell \in \xi^c) \right].$$

Using the fact that $R(S^*,\mathbf{v})$ and $R(S_\ell,\mathbf{v})$ are both bounded by one and $V(S_\ell)\le N$ in (C.6), we have $\Delta R_\ell \le N+1$. Substituting the preceding inequality in the above equation, we obtain,

$$\mathbb{E}_\pi\left(\Delta R_\ell\right) \le (N+1)\mathbb{P}_\pi(\mathcal{A}_{\ell-1}) + (N+1)\mathbb{P}_\pi\left(\ell\in\xi\right) + \mathbb{E}\left[\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right].$$

From the analysis in Appendix A.3 (see (A.17)), we have $\mathcal{P}(\mathcal{A}_\ell)\le\frac{13}{\ell}$. Therefore, it follows that,

$$\mathbb{E}_\pi\left(\Delta R_\ell\right) \le \frac{13(N+1)}{\ell} + (N+1)\mathbb{P}_\pi\left(\ell\in\xi\right) + \mathbb{E}\left[\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right].$$

Substituting the above inequality in (C.7), we obtain

$$Reg_\pi(T,\mathbf{v}) \le 14N\log T + (N+1)\sum_{\ell=1}^{L}\mathbb{P}_\pi\left(\ell\in\xi\right) + \mathbb{E}_\pi\left[\sum_{\ell=1}^{L}\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right].$$

From Corollary C.1, we have that $\sum_{\ell=1}^{L}\mathbb{1}(\ell\in\xi)\le N\log NT$. Hence, we have,

$$Reg_\pi(T,\mathbf{v}) \le 14N\log T + N(N+1)\log NT + \mathbb{E}_\pi\left[\sum_{\ell=1}^{L}\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right]. \tag{C.8}$$

Let $\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}$ be the set of "good epochs" offering sub-optimal products, more specifically,

$$\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} \triangleq \left\{\ell \mid \mathbb{1}(\mathcal{A}_\ell^c)=1 \text{ and } R(S_\ell,\mathbf{v}) < R(S^*,\mathbf{v})\right\}.$$

If $R(S_\ell,\mathbf{v}) = R(S^*,\mathbf{v})$, then by definition, we have $\Delta R_\ell = 0$. Therefore, it follows that,

$$\mathbb{E}_\pi\left[\sum_{\ell=1}^{L}\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right] = \mathbb{E}_\pi\left[\sum_{\ell\in\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}}\Delta R_\ell\cdot\mathbb{1}(\ell\in\xi^c)\right]. \tag{C.9}$$

Whenever $\mathbb{1}(\mathcal{A}_{\ell-1}^c)=1$, from Lemma A.3, we have, $\tilde{R}_\ell(S^*)\ge R(S^*,\mathbf{v})$ and by our algorithm design, we have $\tilde{R}_\ell(S_\ell)\ge\tilde{R}_\ell(S^*)$ for all $\ell\ge 1$. Therefore, it follows that

$$\begin{aligned}
\mathbb{E}_\pi\left\{\Delta R_\ell\cdot\mathbb{1}(\mathcal{A}_\ell^c)\right\} &\le \mathbb{E}_\pi\left\{\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell)-R(S_\ell,\mathbf{v}))\right]\cdot\mathbb{1}(\mathcal{A}_{\ell-1}^c)\cdot\mathbb{1}(\ell\in\xi^c)\right\},\\
&\le \sum_{i\in S_\ell}\left(C_1\sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{C_2\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right)\cdot\mathbb{1}(\ell\in\xi^c),\\
&\le C\sum_{i\in S_\ell}\sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}}.
\end{aligned} \tag{C.10}$$

where $C=C_1+C_2$, the second inequality in (C.10) follows from the definition of the event, $\mathcal{A}_\ell$ and the last inequality follows from the definition of set ξ. From equations (C.8), (C.9), and (C.10) , we have,

$$Reg_\pi(T,\mathbf{v}) \le 14N^2\log NT + C\mathbb{E}_\pi\left\{\sum_{\ell\in\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}}\sum_{i\in S_\ell}\sqrt{\frac{\log NT}{T_i(\ell)}}\right\}, \tag{C.11}$$

Let T_i be the number of "good epochs" that offered sub-optimal assortments containing product i, specifically,

$$T_i = \left|\left\{\ell \in \mathcal{E}_{\mathcal{G}}^{\text{sub\_opt}} \;\middle|\; i \in S_\ell\right\}\right|.$$

Substituting the inequality $\sum_{\ell \in \mathcal{E}_{\mathcal{G}}^{\text{sub\_opt}}} \frac{1}{\sqrt{T_i(\ell)}} \leq \sqrt{T_i}$ in (C.11) and noting that $T_i \leq T$, we obtain,

$$Reg_\pi(T, \mathbf{v}) \leq 14N^2 \log NT + C \sum_{i=1}^{N} \mathbb{E}_\pi \left(\sqrt{T_i \log T}\right).$$

From Jenson's inequality, we have $\mathbb{E}_\pi \left(\sqrt{T_i}\right) \leq \sqrt{\mathbb{E}_\pi (T_i)}$ and therefore, it follows that,

$$Reg_\pi(T, \mathbf{v}) \leq 14N \log T + NC \log NT + C \sum_{i=1}^{N} \sqrt{\mathbb{E}_\pi (T_i) \log NT}.$$

From Cauchy-Schwartz inequality, we have, $\sum_{i=1}^{N} \sqrt{\mathbb{E}_\pi (T_i)} \leq \sqrt{N \sum_{i=1}^{N} \mathbb{E}_\pi (T_i)}$. Therefore, it follows that,

$$Reg_\pi(T, \mathbf{v}) \leq 14N^2 \log NT + C \sqrt{N \sum_{i=1}^{N} \mathbb{E}_\pi (T_i) \log NT}.$$

For any epoch ℓ, we have $|S_\ell| \leq N$. Hence, we have $\sum_{i=1}^{N} T_i \leq N |\mathcal{E}_{\mathcal{G}}^{\text{sub\_opt}}|$. From Lemma 6.2, we have $|\mathcal{E}_{\mathcal{G}}^{\text{sub\_opt}}| \leq N\tau$. Therefore, we have $\sum_{i=1}^{N} \mathbb{E}_\pi (T_i) \leq N^2 \tau$ and hence, it follows that,

$$\begin{aligned} Reg_\pi(T, \mathbf{v}) &\leq 14N^2 \log NT + CN \sqrt{N\tau \log NT}, \\ &\leq 14N^2 \log NT + C \frac{N^2 \log NT}{\Delta(\mathbf{v})}. \end{aligned} \tag{C.12}$$

□

D. Multiplicative Chernoff Bounds

We will extend the Chernoff bounds as discussed in Mitzenmacher and Upfal (2005) [1] to geometric random variables and establish the following concentration inequality.

Theorem 5 *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$ with parameter p, i.e. for any i*

$$Pr(X_i = m) = (1-p)^m p \;\; \forall m = \{0, 1, 2, \cdots\},$$

and let $\mu = \mathbb{E}(X_i) = \frac{1-p}{p}$. We have,

1.
$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) \leq \begin{cases} \exp\left(-\frac{n\mu\delta^2}{2(1+\delta)(1+\mu)^2}\right) & \text{if } \mu \leq 1, \\ \exp\left(-\frac{n\delta^2\mu^2}{6(1+\mu)^2}\left(3 - \frac{2\delta\mu}{1+\mu}\right)\right) & \text{if } \mu \geq 1 \text{ and } \delta \in (0,1). \end{cases}$$

and

[1] (originally discussed in Angluin and Valiant (1977))

2.
$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \begin{cases} \exp\left(-\frac{n\delta^2\mu}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right)\right) & \text{if } \mu \leq 1, \\ \exp\left(-\frac{n\delta^2\mu^2}{2(1+\mu)^2}\right) & \text{if } \mu \geq 1. \end{cases}$$

Proof. We will first bound $Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right)$ and then follow a similar approach for bounding $Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right)$ to complete the proof.

Bounding $Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right)$:

For all i and for any $0 < t < \log\frac{1+\mu}{\mu}$, we have,
$$\mathbb{E}(e^{tX_i}) = \frac{1}{1-\mu(e^t-1)}.$$

Therefore, from Markov Inequality, we have
$$\begin{aligned} Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) &= Pr\left(e^{t\sum_{i=1}^{n} X_i} > e^{(1+\delta)n\mu t}\right), \\ &\leq e^{-(1+\delta)n\mu t}\prod_{i=1}^{n}\mathbb{E}(e^{tX_i}), \\ &= e^{-(1+\delta)n\mu t}\left(\frac{1}{1-\mu(e^t-1)}\right)^n. \end{aligned}$$

Therefore, we have,
$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) \leq \min_{0<t<\log\frac{1+\mu}{\mu}} e^{-(1+\delta)n\mu t}\left(\frac{1}{1-\mu(e^t-1)}\right)^n. \tag{D.1}$$

We have,
$$\operatorname*{argmin}_{0<t<\log\frac{1+\mu}{\mu}} e^{-(1+\delta)n\mu t}\left(\frac{1}{1-\mu(e^t-1)}\right)^n = \operatorname*{argmin}_{0<t<\log\frac{1+\mu}{\mu}} -(1+\delta)n\mu t - n\log\left(1-\mu(e^t-1)\right), \tag{D.2}$$

Noting that the right hand side in the above equation is a convex function in t, we obtain the optimal t by solving for the zero of the derivative. Specifically, at optimal t, we have
$$e^t = \frac{(1+\delta)(1+\mu)}{1+\mu(1+\delta)}.$$

Substituting the above expression in (D.1), we obtain the following bound.
$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) \leq \left(1-\frac{\delta}{(1+\delta)(1+\mu)}\right)^{n\mu(1+\delta)}\left(1+\frac{\delta\mu}{1+\mu}\right)^n. \tag{D.3}$$

First consider the setting where $\mu \in (0,1)$.

Case 1a: If $\mu \in (0,1)$: From Taylor series of $\log(1-x)$, we have that
$$n\mu(1+\delta)\log\left(1-\frac{\delta}{(1+\delta)(1+\mu)}\right) \leq -\frac{n\delta\mu}{1+\mu} - \frac{n\delta^2\mu}{2(1+\delta)(1+\mu)^2},$$

From Taylor series for $\log(1+x)$, we have

$$n\log\left(1+\frac{\delta\mu}{1+\mu}\right) \leq \frac{n\delta\mu}{(1+\mu)},$$

Note that if $\delta > 1$, we can use the fact that $\log(1+\delta x) \leq \delta\log(1+x)$ to arrive at the preceding result. Substituting the preceding two equations in (D.3), we have

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) \leq \exp\left(-\frac{n\mu\delta^2}{2(1+\delta)(1+\mu)^2}\right), \tag{D.4}$$

Case 1b: If $\mu \geq 1$: From Taylor series of $\log(1-x)$, we have that

$$n\mu(1+\delta)\log\left(1-\frac{\delta}{(1+\delta)(1+\mu)}\right) \leq -\frac{n\delta\mu}{1+\mu},$$

If $\delta < 1$, from Taylor series for $\log(1+x)$, we have

$$n\log\left(1+\frac{\delta\mu}{1+\mu}\right) \leq \frac{n\delta\mu}{(1+\mu)} - \frac{n\delta^2\mu^2}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right).$$

If $\delta \geq 1$, we have $\log(1+\delta x) \leq \delta\log(1+x)$ and from Taylor series for $\log(1+x)$, it follows that,

$$n\log\left(1+\frac{\delta\mu}{1+\mu}\right) \leq \frac{n\delta\mu}{(1+\mu)} - \frac{n\delta\mu^2}{6(1+\mu)^2}\left(3-\frac{2\mu}{1+\mu}\right).$$

Therefore, substituting preceding results in (D.3), we have

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > (1+\delta)\mu\right) \leq \begin{cases} \exp\left(-\frac{n\delta^2\mu^2}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right)\right) & \text{if } \mu \geq 1 \text{ and } \delta \in (0,1), \\ \exp\left(-\frac{n\delta\mu^2}{6(1+\mu)^2}\left(3-\frac{2\mu}{1+\mu}\right)\right) & \text{if } \mu \geq 1 \text{ and } \delta \geq 1. \end{cases} \tag{D.5}$$

Bounding $Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right)$:

Now to bound the other one sided inequality, we use the fact that

$$\mathbb{E}(e^{-tX_i}) = \frac{1}{1-\mu(e^{-t}-1)},$$

and follow a similar approach. More specifically, from Markov Inequality, for any $t > 0$ and $0 < \delta < 1$, we have

$$\begin{aligned} Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) &= Pr\left(e^{-t\sum_{i=1}^{n} X_i} > e^{-(1-\delta)n\mu t}\right), \\ &\leq e^{(1-\delta)n\mu t}\prod_{i=1}^{n}\mathbb{E}(e^{-tX_i}), \\ &= e^{(1-\delta)n\mu t}\left(\frac{1}{1-\mu(e^{-t}-1)}\right)^n. \end{aligned}$$

Therefore, we have

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \min_{t>0} e^{-(1+\delta)n\mu t}\left(\frac{1}{1-\mu(e^{-t}-1)}\right)^n,$$

Following similar approach as in optimizing the previous bound (see (D.1)) to establish the following result.

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \left(1+\frac{\delta}{(1-\delta)(1+\mu)}\right)^{n\mu(1-\delta)}\left(1-\frac{\delta\mu}{1+\mu}\right)^{n}.$$

Now we will use Taylor series for $\log(1+x)$ and $\log(1-x)$ in a similar manner as described for the other bound to obtain the required result. In particular, since $1-\delta \leq 1$, we have for any $x>0$ it follows that $(1+\frac{x}{1-\delta})^{(1-\delta)} \leq (1+x)$. Therefore, we have

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \left(1+\frac{\delta}{1+\mu}\right)^{n\mu}\left(1-\frac{\delta\mu}{1+\mu}\right)^{n}. \tag{D.6}$$

Case 2a. If $\mu \in (0,1)$: Note that since $X_i \geq 0$ for all i, we have a zero probability event if $\delta > 1$. Therefore, we assume $\delta < 1$ and from Taylor series for $\log(1-x)$, we have

$$n\log\left(1-\frac{\delta\mu}{1+\mu}\right) \leq -\frac{n\delta\mu}{1+\mu},$$

and from Taylor series for $\log(1+x)$, we have

$$n\mu\log\left(1+\frac{\delta}{1+\mu}\right) \leq \frac{n\delta\mu}{(1+\mu)} - \frac{n\delta^2\mu}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right).$$

Therefore, substituting the preceding equations in (D.6), we have,

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \exp\left(-\frac{n\delta^2\mu}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right)\right). \tag{D.7}$$

Case 2b. If $\mu \geq 1$: For similar reasons as discussed above, we assume $\delta < 1$ and from Taylor series for $\log(1-x)$, we have

$$n\log\left(1-\frac{\delta\mu}{1+\mu}\right) \leq -\frac{n\delta\mu}{1+\mu} - \frac{n\delta^2\mu^2}{2(1+\mu)^2},$$

and from Taylor series for $\log(1+x)$, we have

$$n\log\left(1+\frac{\delta\mu}{1+\mu}\right) \leq \frac{n\delta}{(1+\mu)}.$$

Therefore, substituting the preceding equations in (D.6), we have,

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i < (1-\delta)\mu\right) \leq \exp\left(-\frac{n\delta^2\mu^2}{2(1+\mu)^2}\right). \tag{D.8}$$

The result follows from (D.4), (D.5), (D.7) and (D.8). □

Now, we will adapt a non-standard corollary from Babaioff et al. (2015) and Kleinberg et al. (2008) to our estimates to obtain sharper bounds.

Lemma D.1 *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$ with parameter p, i.e. for any i, $P(X_i = m) = (1-p)^m p \ \forall m = \{0,1,2,\cdots\}$. Let $\mu = \mathbb{E}_\pi(X_i) = \frac{1-p}{p}$ and $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$. If $n > 48\log(\sqrt{N}\ell+1)$, then we have for all $n = 1,2,\cdots$,*

1.
$$\mathcal{P}\left(\left|\bar{X}-\mu\right| > \max\left\{\sqrt{\bar{X}}, \bar{X}\right\}\sqrt{\frac{48\log(\sqrt{N}\ell+1)}{n}} + \frac{48\log(\sqrt{N}\ell+1)}{n}\right) \le \frac{6}{\ell^2}. \tag{D.9}$$

2.
$$\mathcal{P}\left(\left|\bar{X}-\mu\right| \ge \max\{\sqrt{\mu}, \mu\}\sqrt{\frac{24\log(\sqrt{N}\ell+1)}{n}} + \frac{48\log(\sqrt{N}\ell+1)}{n}\right) \le \frac{4}{\ell^2}, \tag{D.10}$$

3.
$$\mathcal{P}\left(\bar{X} \ge \frac{3\mu}{2} + \frac{48\log(\sqrt{N}\ell+1)}{n}\right) \le \frac{3}{\ell^2}. \tag{D.11}$$

Proof. We will analyze the cases $\mu < 1$ and $\mu \ge 1$ separately.

Case-1: $\mu \le 1$. Let $\delta = (\mu+1)\sqrt{\frac{6\log(\sqrt{N}\ell+1)}{\mu n}}$. First assume that $\delta \le \frac{1}{2}$. Substituting the value of δ in Theorem 5, we obtain,

$$\begin{aligned}
\mathcal{P}\left(\bar{X}-\mu > \delta\mu\right) &\le \frac{1}{N\ell^2}, \\
\mathcal{P}\left(\bar{X}-\mu < -\delta\mu\right) &\le \frac{1}{N\ell^2}, \\
\mathcal{P}\left(\left|\bar{X}-\mu\right| < (\mu+1)\sqrt{\frac{6\mu\log(\sqrt{N}\ell+1)}{n}}\right) &\ge 1 - \frac{2}{N\ell^2}.
\end{aligned} \tag{D.12}$$

Since $\delta \le \frac{1}{2}$, we have $\mathcal{P}\left(\bar{X}-\mu \le -\frac{\mu}{2}\right) \le \mathcal{P}\left(\bar{X}-\mu \le -\delta\mu\right)$. Hence, from (D.12), we have,

$$\mathcal{P}\left(\bar{X}-\mu \le -\frac{\mu}{2}\right) \le \frac{1}{N\ell^2},$$

and hence, it follows that,

$$\mathcal{P}\left(2\bar{X} \ge \mu\right) \ge 1 - \frac{1}{N\ell^2}. \tag{D.13}$$

From (D.12) and (D.13), we have,

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| < \sqrt{\frac{48\bar{X}\log(\sqrt{N}\ell+1)}{n}}\right) \ge \mathcal{P}\left(\left|\bar{X}-\mu\right| < \sqrt{\frac{24\mu\log(\sqrt{N}\ell+1)}{n}}\right) \ge 1 - \frac{3}{N\ell^2}. \tag{D.14}$$

Since $\delta \le \frac{1}{2}$, we have, $\mathcal{P}\left(\bar{X} \le \frac{3\mu}{2}\right) \ge \mathcal{P}\left(\bar{X} < (1+\delta)\mu\right)$. Hence, from (D.12), we have

$$\mathcal{P}\left(\bar{X} \le \frac{3\mu}{2}\right) \ge 1 - \frac{1}{N\ell^2}. \tag{D.15}$$

Since, $\mu \le 1$, we have $\mathcal{P}\left(\bar{X} \le \frac{3}{2}\right) \ge 1 - \frac{1}{N\ell^2}$ and

$$\mathcal{P}\left(\bar{X} \le \sqrt{\frac{3\bar{X}}{2}}\right) \ge 1 - \frac{1}{N\ell^2}.$$

Therefore, substituting above result in (D.14), the inequality (D.9) follows.

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| > \max\left\{\sqrt{\bar{X}}, \sqrt{\frac{2}{3}}\bar{X}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{n}}\right) \leq \frac{4}{N\ell^2}. \tag{D.16}$$

Now consider the scenario, when $(\mu+1)\sqrt{\frac{6\log(\sqrt{N}\ell+1)}{\mu n}} > \frac{1}{2}$. Then, we have,

$$\delta_1 \triangleq \frac{12(\mu+1)^2\log\left(\sqrt{N}\ell+1\right)}{\mu n} \geq \frac{1}{2},$$

which implies,

$$\exp\left(-\frac{n\mu\delta_1^2}{2(1+\delta_1)(1+\mu)^2}\right) \leq \exp\left(-\frac{n\mu\delta_1}{6(1+\mu)^2}\right),$$
$$\exp\left(-\frac{n\delta_1^2\mu}{6(1+\mu)^2}\left(3-\frac{2\delta_1\mu}{1+\mu}\right)\right) \leq \exp\left(-\frac{n\mu\delta_1}{6(1+\mu)^2}\right).$$

Therefore, substituting the value of δ_1 in Theorem 5, we have

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| > \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \leq \frac{2}{N\ell^2}. \tag{D.17}$$

Hence, from (D.17) and (D.16), it follows that,

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| > \max\left\{\sqrt{\bar{X}}, \sqrt{\frac{2}{3}}\bar{X}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{n}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \leq \frac{6}{N\ell^2}. \tag{D.18}$$

Case 2: $\boldsymbol{\mu \geq 1}$

Let $\delta = \sqrt{\frac{12\log(\sqrt{N}\ell+1)}{n}}$, then by our assumption, we have $\delta \leq \frac{1}{2}$. Substituting the value of δ in Theorem 5, we obtain,

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| < \mu\sqrt{\frac{12\log\left(\sqrt{N}\ell+1\right)}{n}}\right) \geq 1-\frac{2}{N\ell^2},$$
$$\mathcal{P}\left(2\bar{X} \geq \mu\right) \geq 1-\frac{1}{N\ell^2}.$$

Hence we have,

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| < \bar{X}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{n}}\right) \geq \mathcal{P}\left(\left|\bar{X}-\mu\right| < \mu\sqrt{\frac{12\log\left(\sqrt{N}\ell+1\right)}{n}}\right) \geq 1-\frac{3}{N\ell^2}. \tag{D.19}$$

By assumption $\mu \geq 1$. Therefore, we have $\mathcal{P}\left(\bar{X} \geq \frac{1}{2}\right) \geq 1-\frac{1}{N\ell^2}$ and,

$$\mathcal{P}\left(\bar{X} \geq \sqrt{\frac{\bar{X}}{2}}\right) \geq 1-\frac{1}{N\ell^2}. \tag{D.20}$$

Therefore, from (D.19) and (D.20), we have

$$\mathcal{P}\left(\left|\bar{X}-\mu\right| > \max\left\{\bar{X}, \sqrt{\frac{\bar{X}}{2}}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{n}}\right) \leq \frac{4}{N\ell^2}. \tag{D.21}$$

We complete the proof by stating that (D.9) follows from (D.18) and (D.21), while (D.10) follows from (D.14) and (D.19) and (D.11) follows from (D.15) and (D.17). □

From the proof of Lemma D.1, the following result follows.

Corollary D.1 *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$ with parameter p, i.e. for any i, $P(X_i = m) = (1-p)^m p \ \forall m = \{0,1,2,\cdots\}$. Let $\mu = \mathbb{E}_\pi(X_i) = \frac{1-p}{p}$ and $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$. If $\mu \leq 1$, then we have,*

1. $\mathcal{P}\left(\left|\bar{X}-\mu\right| > \sqrt{\frac{48\bar{X}\log\left(\sqrt{N}\ell+1\right)}{n}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \leq \frac{6}{N\ell^2}$. *for all* $n = 1,2,\cdots$.
2. $\mathcal{P}\left(\left|\bar{X}-\mu\right| \geq \sqrt{\frac{24\mu\log\left(\sqrt{N}\ell+1\right)}{n}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \leq \frac{4}{N\ell^2}$ *for all* $n = 1,2,\cdots$.
3. $\mathcal{P}\left(\bar{X} \geq \frac{3\mu}{2} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{n}\right) \leq \frac{3}{N\ell^2}$.

Proof of Lemma A.2 Fix i and ℓ, define the events,

$$\mathcal{A}_{i,\ell} = \left\{|\bar{v}_{i,\ell} - v_i| > \sqrt{48\bar{v}_{i,\ell}\frac{\log\left(\sqrt{N}\ell+1\right)}{|\mathcal{T}_i(\ell)|}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{|\mathcal{T}_i(\ell)|}\right\}.$$

Let $\bar{v}_{i,m} = \dfrac{\sum_{\tau=1}^m \hat{v}_{i,\tau}}{m}$. Then, we have,

$$\begin{aligned}
\mathbb{P}_\pi\left(\mathcal{A}_{i,\ell}\right) &\leq \mathbb{P}_\pi\left\{\max_{m\leq\ell}\left(|\bar{v}_{i,m} - v_i| - \sqrt{48\bar{v}_{i,m}\frac{\log\left(\sqrt{N}\ell+1\right)}{m}} - \frac{48\log\left(\sqrt{N}\ell+1\right)}{m}\right) > 0\right\},\\
&= \mathbb{P}_\pi\left(\bigcup_{m=1}^{\ell}\left\{|\bar{v}_{i,m} - v_i| - \sqrt{48\bar{v}_{i,m}\frac{\log\left(\sqrt{N}\ell+1\right)}{m}} - \frac{48\log\left(\sqrt{N}\ell+1\right)}{m} > 0\right\}\right),\\
&\leq \sum_{m=1}^{\ell}\mathbb{P}_\pi\left(|\bar{v}_{i,m} - v_i| > \sqrt{48\bar{v}_{i,m}\frac{\log\left(\sqrt{N}\ell+1\right)}{m}} - \frac{48\log\left(\sqrt{N}\ell+1\right)}{m}\right),\\
&\overset{(a)}{\leq} \sum_{m=1}^{\ell}\frac{6}{N\ell^2} \leq \frac{6}{N\ell}.
\end{aligned} \tag{D.22}$$

where inequality (a) in (D.22) follows from Corollary D.1. The first inequality in Lemma A.2 follows from definition of $v_{i,\ell}^{\mathsf{UCB}}$, Corollary D.1 and (D.22). The second and third inequality in Lemma A.2 also can be derived in a similar fashion by appropriately modifying the definition of set $\mathcal{A}_{i,\ell}$. □

Proof of Lemma B.3 is similar to the proof of Lemma A.2.

E. Lower Bound

We follow the proof of $\Omega(\sqrt{NT})$ lower bound for the Bernoulli instance with parameters $\frac{1}{2}$. We first establish a bound on KL divergence, which will be useful for us later.

Lemma E.1 *Let p and q denote two Bernoulli distributions with parameters $\alpha+\epsilon$ and α respectively. Then, the KL divergence between the distributions p and q is bounded by $4K\epsilon^2$,*

$$KL(p\|q) \leq \frac{4}{\alpha}\epsilon^2.$$

Proof.

$$\begin{aligned} KL(p\|q) &= \alpha \cdot \log\frac{\alpha}{\alpha+\epsilon} + (1-\alpha)\log\frac{1-\alpha}{1-\alpha-\epsilon} \\ &= \alpha\left[\log\frac{1-\frac{\epsilon}{1-\alpha}}{1+\frac{\epsilon}{\alpha}}\right] - \log\left(1-\frac{\epsilon}{1-\alpha}\right), \\ &= \alpha\log\left(1-\frac{\epsilon}{(1-\alpha)(\alpha+\epsilon)}\right) - \log\left(1-\frac{\epsilon}{1-\alpha}\right), \end{aligned}$$

using $1-x \leq e^{-x}$ and bounding the Taylor series for $-\log 1-x$ by $x+2*x^2$ for $x=\dfrac{\epsilon}{1-\alpha}$, we have

$$\begin{aligned} KL(p\|q) &\leq \frac{-\alpha\epsilon}{(1-\alpha)(\alpha+\epsilon)} + \frac{\epsilon}{1-\alpha} + 4\epsilon^2, \\ &= (\frac{2}{\alpha}+4)\epsilon^2 \leq \frac{4}{\alpha}\epsilon^2. \end{aligned}$$

□.

Fix a guessing algorithm, which at time t sees the output of a coin a_t. Let $P_1,\cdots,P_n$ denote the distributions for the view of the algorithm from time 1 to T, when the biased coin is hidden in the i^{th} position. The following result establishes for any guessing algorithm, there are at least $\frac{N}{3}$ positions that a biased coin could be and will not be played by the guessing algorithm with probability at least $\frac{1}{2}$. Specifically,

Lemma E.2 *Let $\mathcal{A}$ be any guessing algorithm operating as specified above and let $t \leq \frac{N\alpha}{60\epsilon^2}$, for $\epsilon \leq \frac{1}{4}$ and $N \geq 12$. Then, there exists $J \subset \{1,\cdots,N\}$ with $|J| \geq \frac{N}{3}$ such that*

$$\forall j \in J,\ \mathcal{P}_j(a_t = j) \leq \frac{1}{2}.$$

Proof. Let N_i to be the number of times the algorithm plays coin i up to time t. Let P_0 be the hypothetical distribution for the view of the algorithm when none of the N coins are biased. We shall define the set J by considering the behavior of the algorithm if tosses it saw were according to the distribution P_0. We define,

$$J_1 = \left\{ i \,\middle|\, E_{P_0}(N_i) \leq \frac{3t}{N} \right\}, J_2 = \left\{ i \,\middle|\, \mathcal{P}_0(a_t = i) \leq \frac{3}{N} \right\} \text{ and } J = J_1 \cap J_2. \tag{E.1}$$

Since $\sum_i E_{P_0}(N_i) = t$ and $\sum_i \mathcal{P}_0(a_t = i) = 1$, a counting argument would give us $|J_1| \geq \frac{2N}{3}$ and $|J_2| \geq \frac{2n}{3}$ and hence $|J| \geq \frac{N}{3}$. Consider any $j \in J$, we will now prove that if the biased coin is at position j, then the probability of algorithm guessing the biased coin will not be significantly different from the P_0 scenario. By Pinsker's inequality, we have

$$|\mathcal{P}_j(a_t = j) - \mathcal{P}_0(a_t = j)| \leq \frac{1}{2}\sqrt{2\log 2 \cdot KL(P_0 \| P_j)}, \tag{E.2}$$

where $KL(P_0\|P_j)$ is the KL divergence of probability distributions P_0 and P_j over the algorithm. Using the chain rule for KL-divergence, we have

$$KL(P_0\|P_j) = E_{P_0}(N_j) KL(p||q),$$

where p is a Bernoulli distribution with parameter α and q is a Bernoulli distribution with parameter $\alpha + \epsilon$. From Lemma E.1 and (E.1), we have that Therefore,

$$KL(P_0\|P_j) \leq \frac{4\epsilon^2}{\alpha},$$

Therefore,

$$\begin{aligned}
\mathcal{P}_j(a_t = j) &\leq \mathcal{P}_0(a_t = j) + \frac{1}{2}\sqrt{2\log 2 \cdot KL(P_0\|P_j)}, \\
&\leq \frac{3}{N} + \frac{1}{2}\sqrt{(2\log 2)\frac{4\epsilon^2}{\alpha} E_{P_0}(N_j)}, \\
&\leq \frac{3}{N} + \sqrt{2\log 2}\sqrt{\frac{3t\epsilon^2}{N\alpha}} \leq \frac{1}{2}.
\end{aligned} \tag{E.3}$$

The second inequality follows from (E.1), while the last inequality follows from the fact that $N > 12$ and $t \leq \frac{N\alpha}{60\epsilon^2}$ □.

Proof of Lemma 5.1 . Let $\epsilon = \sqrt{\frac{N}{60\alpha T}}$. Suppose algorithm $\mathcal{A}$ plays coin a_t at time t for each $t = 1, \cdots, T$. Since $T \leq \frac{N\alpha}{60\epsilon^2}$, for all $t \in \{1, \cdots, T-1\}$ there exists a set $J_t \subset \{1, \cdots, N\}$ with $|J_t| \geq \frac{N}{3}$ such that

$$\forall\, j \in J_t, P_j(j \in S_t) \leq \frac{1}{2}.$$

Let i^* denote the position of the biased coin. Then,

$$\mathbb{E}_\pi\left(\mu_{a_t} \,|\, i^* \in J_t\right) \leq \frac{1}{2}\cdot(\alpha+\epsilon) + \frac{1}{2}\cdot\alpha = \alpha + \frac{\epsilon}{2},$$

$$\mathbb{E}_\pi\left(\mu_{a_t} \,|\, i^* \notin J_t\right) \leq \alpha + \epsilon.$$

Since $|J_t| \geq \frac{N}{3}$ and i^* is chosen randomly, we have $P(i^* \in J_t) \geq \frac{1}{3}$. Therefore, we have

$$\mu_{a_t} \leq \frac{1}{3}\cdot\left(\alpha + \frac{\epsilon}{2}\right) + \frac{2}{3}\cdot(\alpha+\epsilon) = \alpha + \frac{5\epsilon}{6}$$

We have $\mu^* = \alpha + \epsilon$ and hence the $Regret \geq \frac{T\epsilon}{6}$. □

Lemma E.3 *Let L be the total number of calls to $\mathcal{A}_{\mathsf{MNL}}$ when $\mathcal{A}_{\mathsf{MAB}}$ is executed for T time steps. Then,*

$$\mathbb{E}(L) \leq 3T.$$

Proof. Let η_ℓ be the random variable that denote the duration, assortment S_ℓ has been considered by $\mathcal{A}_{MAB}$, i.e. $\eta_\ell = 0$, if we no arm is pulled when $\mathcal{A}_{MNL}$ suggested assortment S_ℓ and $\eta_\ell \geq 1$, otherwise. We have

$$\sum_{\ell=1}^{L-1} \eta_\ell \leq T.$$

Therefore, we have $\mathbb{E}\left(\sum_{\ell=1}^{L-1} \eta_\ell\right) \leq T$. Note that $\mathbb{E}(\eta_\ell) \geq \frac{1}{2}$. Hence, we have $\mathbb{E}(L) \leq 2T + 1 \leq 3T$. □

E.1. Lower Bound for the unconstrained MNL-Bandit problem ($K = N$)

We will complete proof of Theorem 2 by showing that the lower bound holds true for the case when $K = N$. We will show this by reduction to a parametric multi armed bandit problem with 2 arms.

Definition E.1 (MNL-Bandit instance $\hat{I}_{\mathsf{MNL}}$) *Define $\hat{I}_{\mathsf{MNL}}$ as the following (randomized) instance of unconstrained MNL-Bandit problem, N products, with revenues, $r_1 = 1$, $r_2 = \frac{1+\epsilon}{3+2\epsilon}$ and $r_i = 0.01$ for all $i = 3, \cdots, N$, and MNL parameters $v_0 = 1$, $v_i = \frac{1}{2}$ for all $i = 2, \cdots, N$, while v_1 is randomly set at $\{\frac{1}{2}, \frac{1}{2} + \epsilon\}$, where $\epsilon = \sqrt{\frac{1}{32T}}$.*

Preliminaries on the MNL-Bandit instance $\hat{I}_{\mathsf{MNL}}$: Note that unlike the MNL-Bandit instance, I_{MNL}, where any product can have the biased (higher) MNL parameter, in the MNL-Bandit instance $\hat{I}_{\mathsf{MNL}}$, only one product (product 1) can be biased. From the proof of Lemma A.5, we have that,

$$i \in S^* \text{ if and only if } r_i \geq R(S^*, \mathbf{v}), \tag{E.4}$$

where S^* is the optimal assortment for $\hat{I}_{\mathsf{MNL}}$.
Note that the revenue corresponding to assortment $\{1\}$ is

$$R(\{1\}, \mathbf{v}) = \begin{cases} \dfrac{1+2\epsilon}{3+2\epsilon}, & \text{if } v_1 = \frac{1}{2} + \epsilon \\ \dfrac{1}{3}, & \text{if } v_1 = \frac{1}{2}. \end{cases}$$

Note that $\frac{1+2\epsilon}{3+2\epsilon} > r_2 = \frac{1+\epsilon}{3+2\epsilon} > \frac{1}{3} > r_3 = 0.01$ and since $R(S^*, \mathbf{v}) \geq R(\{1\}, \mathbf{v})$, from (E.4), we have that optimal assortment is either $\{1\}$ or $\{1, 2\}$, specifically, we have that

$$S^* \in \{\{1\}, \{1, 2\}\}.$$

Therefore, we have,

$$S^* = \begin{cases} \{1\}, & \text{if } v_1 = \frac{1}{2} + \epsilon, \\ \{1,2\}, & \text{if } v_1 = \frac{1}{2}. \end{cases} \tag{E.5}$$

Note that since $r_3 < \frac{1}{3}$, for any S and i, such that $i \geq 3$ and $i \notin S$, we have

$$R(S, \mathbf{v}) > R(S \cup \{i\}, \mathbf{v}).$$

Therefore, if $v_i = \frac{1}{2} + \epsilon$, for any $S \neq \{1\}$, we have

$$R(\{1\}, \mathbf{v}) - R(S, \mathbf{v}) \geq R(\{1\}, \mathbf{v}) - R(\{1,2\}, \mathbf{v}) \geq \frac{\epsilon}{20}, \tag{E.6}$$

and similarly if $v_i = \frac{1}{2}$, for any $S \neq \{1,2\}$, we have,

$$R(\{1\}, \mathbf{v}) - R(S, \mathbf{v}) \geq R(\{1,2\}, \mathbf{v}) - R(\{1\}, \mathbf{v}) = \frac{\epsilon}{12} \geq \frac{\epsilon}{20}, \tag{E.7}$$

Before providing the formal proof, we first present the intuition behind the result. Any algorithm that does not offer product 2 when $v_1 = 1/2$ will incur a regret and similarly any algorithm that offers product 2 when $v_1 = 1/2 + \epsilon$. Hence, any algorithm that attempts to minimize regret on instance $\hat{I}_{\mathsf{MNL}}$ has to quickly learn if $v_1 = 1/2 + \epsilon$ or $v_1 = 1/2$. From Chernoff bounds, we know that we need approximately $1/\epsilon^2$ observations to conclude with high probability if $v_1 = 1/2 + \epsilon$ or $1/2$. Therefore in each of these $1/\epsilon^2$ time steps, we are likely to incur a regret of ϵ, leading to a cumulative regret of $1/\epsilon \approx \sqrt{T}$. In what follows, we will formalize this intuition on similar lines to the proof of Lemma 5.1. First, we present two auxillary results required to prove Lemma 2.

Lemma E.4 *Let S be an arbitrary subset of $\{1, \cdots, N\}$ and $\mathcal{P}_0^S, \mathcal{P}_1^S$ denote the probability distributions over the discrete space $\{0, 1, \ldots, N\}$ governed by the MNL feedback on instance $\hat{I}_{MNL}$ when the offer set is S and $v_1 = 1/2$ and $v_1 = 1/2 + \epsilon$ respectively. In particular, we assume,*

$$\mathcal{P}_0^S(i) = \frac{1}{2 + |S|} \times \begin{cases} 0, & \text{if } i \notin S \cup \{0\}, \\ 2, & \text{if } i = 0, \\ 1 & \text{if } i \in S. \end{cases}, \mathcal{P}_1^S(i) = \frac{1}{2 + |S| + 2\epsilon \mathbb{1}(1 \in S)} \times \begin{cases} 0, & \text{if } i \notin S \cup \{0\}, \\ 2, & \text{if } i = 0, \\ 1 & \text{if } i \in S \backslash \{1\} \\ 1 + 2\epsilon & \text{if } i = 1. \end{cases}$$

Then for any S,

$$\mathsf{KL}\left(\mathcal{P}_0^S \middle\| \mathcal{P}_1^S\right) \leq 4\epsilon^2, \tag{E.8}$$

where KL is the Kullback-Leibler divergence.

Proof. If $1 \notin S$, we have $\mathcal{P}_0^S$ and $\mathcal{P}_1^S$ to be the same distributions and the Kullback-Leibler divergence between them is 0. Therefore without loss of generality, assume that $1 \in S$.

$$\begin{aligned}
\mathsf{KL}\left(\mathcal{P}_0^S \middle\| \mathcal{P}_1^S\right) &= \sum_{j=0}^{N} \mathcal{P}_0^S(j) \log\left(\frac{\mathcal{P}_0^S(j)}{\mathcal{P}_1^S(j)}\right), \\
&= \mathcal{P}_0^S(0) \log\left(\frac{\mathcal{P}_0^S(0)}{\mathcal{P}_1^S(0)}\right) + \sum_{j \in \{S\}\setminus 1} \mathcal{P}_0^S(j) \log\left(\frac{\mathcal{P}_0^S(j)}{\mathcal{P}_1^S(j)}\right) + \mathcal{P}_0^S(1) \log\left(\frac{\mathcal{P}_0^S r(1)}{\mathcal{P}_1^S(1)}\right), \\
&= \frac{|S|+1}{|S|+2} \log\left(1 + \frac{2\epsilon}{2+|S|}\right) + \frac{1}{|S|+2} \log\left(1 - \frac{2\epsilon(|S|+1)}{(2+|S|)(1+2\epsilon)}\right), \\
&\leq \frac{2(|S|+1)\epsilon}{(|S|+2)^2}\left(1 - \frac{1}{(1+2\epsilon)}\right) \leq 4\epsilon^2,
\end{aligned}$$

where the first inequality follows from the fact that for any $x \in (0,1)$,

$$\log(1+x) \leq x \text{ and } \log(1-x) \leq -x.$$

□.

Lemma E.5 *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the probability distribution over consumer choices (throughout the planning horizon T) when assortments are offered according to algorithm $\mathcal{A}_{\mathsf{MNL}}$ and feedback to the algorithm is provided via the MNL-Bandit instances $\hat{I}_{MNL}$, when $v_1 = 1/2$ and $v_1 = 1/2 + \epsilon$ respectively. Then, we have,*

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) \leq 4T\epsilon^2,$$

where $\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right)$ is the Kullback-Leibler divergence between the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$. Specifically,

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = \sum_{\mathbf{c} \in \{0,1,\cdots,N\}^T} \mathcal{P}(\mathbf{c}) \log\left(\frac{\mathcal{P}(\mathbf{c})}{\mathcal{P}_1(\mathbf{c})}\right), \tag{E.9}$$

where $\mathbf{c} \in \{0,1,\cdots,N\}^T$ is the observed set of choices by the algorithm $\mathcal{A}_{\mathsf{MNL}}$.

Proof. From the chain rule for Kullback-Liebler divergence, it follows that,

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = \sum_{t=1}^{T} \sum_{\{c_1,\cdots,c_{t-1}\} \in \{0,1,\cdots,N\}^{t-1}} \mathbb{P}_0(\mathbf{c^t}) \mathsf{KL}\left(\mathbb{P}_0(c_t) \| \mathbb{P}_1(c_t) | c_1, \cdots, c_{t-1}\right), \tag{E.10}$$

where,

$$\mathsf{KL}\left(\mathbb{P}_0(c_t) \| \mathbb{P}_1(c_t) | c_1, \cdots, c_{t-1}\right) = \sum_{c_t} \mathbb{P}_0\left\{c_t | c_1, \cdots, c_{t-1}\right\} \log\left(\frac{\mathbb{P}_0\left\{c_t | c_1, \cdots, c_{t-1}\right\}}{\mathbb{P}_1\left\{c_t | c_1, \cdots, c_{t-1}\right\}}\right).$$

Note that assortment offered by $\mathcal{A}_{\mathsf{MNL}}$ at time t, S_t is completely determined by the reward history $c_1, \ldots, c_{t-1}$ and conditioned on S_t, the reward at time t, c_t is independent of the reward history $c_1, \cdots, c_{t-1}$. Therefore, it follows that,

$$\mathbb{P}_0\left(c_t | c_1, \cdots, c_{t-1}\right) = \mathcal{P}_0^{S_t}(c_t) \text{ and } \mathbb{P}_1\left(c_t | c_1, \cdots, c_{t-1}\right) = \mathcal{P}_1^{S_t}(c_t),$$

and hence, we have,

$$\mathsf{KL}\left(\mathbb{P}_0(c_t)\|\mathbb{P}_1(c_t)|c_1,\cdots,c_{t-1}\right)=\mathsf{KL}\left(\mathcal{P}_0^{S_t}(c_t)\middle\|\mathcal{P}_1^{S_t}(c_t)\right), \tag{E.11}$$

where $\mathcal{P}_0^{S_t}$ and $\mathcal{P}_1^{S_t}$ are defined as in Lemma E.4. Therefore from (E.10), (E.11) and Lemma E.4, we have,

$$\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)=\sum_{t=1}^{T}\mathsf{KL}\left(\mathcal{P}_0^{S_t}\middle\|\mathcal{P}_1^{S_t}\right)\le 4T\epsilon^2.$$

□

Proof of Theorem 2: Fix a guessing algorithm $\mathcal{A}_{\mathsf{MNL}}$, which at time t sees the consumer choice based on the offer assortment S_t. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the distributions for the view of the algorithm from time 1 to T, when $v_1=\frac{1}{2}$ and $v_1=\frac{1}{2}+\epsilon$ respectively. Let T_2 be the number of times $\mathcal{A}$ offers product 2 and let $\mathbb{E}_{\mathbb{P}_0}(T_2)$ and $\mathbb{E}_{\mathbb{P}_1}(T_2)$ be the expected number of times product 2 is offered by $\mathcal{A}$.

$$\begin{aligned}
\left|\mathbb{E}_{\mathbb{P}_0}(T_2)-\mathbb{E}_{\mathbb{P}_1}(T_2)\right| &\le \left|\sum_{t=1}^{T}\mathcal{P}_0(2\in S_t)-\mathbb{P}_1(2\in S_t)\right|,\\
&\le \sum_{t=1}^{T}\left|\mathbb{P}_0(2\in S_t)-\mathbb{P}_1(2\in S_t)\right|,\\
&\le \sum_{t=1}^{T}\frac{1}{2}\sqrt{2\log 2\cdot\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)}=\frac{T}{2}\sqrt{2\log 2\cdot\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)},
\end{aligned} \tag{E.12}$$

where $\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)$ as the Kullback-Leibler divergence between the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ as defined in (E.9) and the last inequality follows from Pinsker's inequality. From Lemma E.5, we have that,

$$\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)\le 4T\epsilon^2.$$

Substituting the value of ϵ, we obtain $\mathsf{KL}\left(\mathbb{P}_0\|\mathbb{P}_1\right)\le\frac{1}{2}$ and from (E.12), we have

$$\left|\mathbb{E}_{\mathbb{P}_0}(T_2)-\mathbb{E}_{\mathbb{P}_1}(T_2)\right|\le\frac{T}{4}. \tag{E.13}$$

Since v_1 can be $\frac{1}{2}$ and $\frac{1}{2}+\epsilon$ with equal probability, we have

$$Reg_{\mathcal{A}_{\mathsf{MNL}}}(T,\mathbf{v})=\frac{1}{2}Reg_{\mathcal{A}_{\mathsf{MNL}}}\left(T,\mathbf{v},\middle|v_1=\frac{1}{2}\right)+\frac{1}{2}Reg_{\mathcal{A}_{\mathsf{MNL}}}\left(T,\mathbf{v},\middle|v_1=\frac{1}{2}+\epsilon\right). \tag{E.14}$$

From (E.7) we have that, in every time step we don't offer product $\{2\}$, we incur a regret of at least $\frac{\epsilon}{20}$ and hence it follows that,

$$Reg_{\mathcal{A}_{\mathsf{MNL}}}\left(T,\mathbf{v},\middle|v_1=\frac{1}{2}\right)\ge\frac{\epsilon}{20}(T-\mathbb{E}_{\mathbb{P}_0}(T_2)),$$

and similarly from (E.6) we have that, in every time step we offer product $\{2\}$, we incur a regret of at least $\frac{\epsilon}{20}$ and hence it follows that,

$$Reg_{\mathcal{A}_{\mathsf{MNL}}}\left(T, \mathbf{v}, \middle| v_1 = \frac{1}{2} + \epsilon\right) \geq \frac{\epsilon}{20}\mathbb{E}_{\mathbb{P}_1}(T_2).$$

Therefore, from (E.14) and (E.13), it follows that,

$$Reg_{\mathcal{A}_{\mathsf{MNL}}}(T, \mathbf{v}) \geq \frac{\epsilon}{20}\left[T - (\mathbb{E}_{\mathbb{P}_1}(T_2) - \mathbb{E}_{\mathbb{P}_0}(T_2))\right] \geq \frac{3T\epsilon}{80}.$$

□