A Near-Optimal Exploration-Exploitation Approach for Assortment Selection

SHIPRA AGRAWAL, VASHIST AVADHANULA, VINEET GOYAL, and ASSAF ZEEVI,
Columbia University

We consider an online assortment optimization problem, where in every round, the retailer offers a K-cardinality subset (assortment) of N substitutable products to a consumer, and observes the response. We model consumer choice behavior using the widely used *multinomial logit* (MNL) model, and consider the retailer's problem of dynamically learning the model parameters, while optimizing cumulative revenues over the selling horizon T. Formulating this as a variant of a multi-armed bandit problem, we present an algorithm based on the principle of "optimism in the face of uncertainty." A naive MAB formulation would treat each of the $\binom{N}{K}$ possible assortments as a distinct "arm," leading to regret bounds that are exponential in K. We show that by exploiting the specific characteristics of the MNL model it is possible to design an algorithm with $\tilde{O}(\sqrt{NT})$ regret, under a mild assumption. We also establish a lower bound, by showing that any algorithm must incur a regret of $\Omega(\sqrt{NT/K})$ for $K < N$. This establishes that the performance of our algorithm is tight for constant K.

General Terms: Exploration-Exploitation, Upper Confidence Bound, Optimal regret

Additional Key Words and Phrases: revenue optimization, multi-armed bandit, regret bounds, assortment optimization, multinomial logit model

1. INTRODUCTION AND PROBLEM FORMULATION

Consider an online planning problem over a discrete option space containing N distinct elements each ascribed with a certain value. At each time step the decision maker needs to select a subset $S \subset N$, with cardinality $|S| \leq K$, after which s/he observes a response that is dependent on the nature of the elements contained in S. Thinking of the N primitive elements as *products*, the subset S as an *assortment*, K as a *display constraint*, and assuming a model that governs how consumers respond and *substitute* among their choice of products (a so-called choice model), the set up is referred to in the literature as an (dynamic) assortment optimization problem. Such problems have their origin in retail, but have since been used in a variety of other application areas. Roughly speaking, the typical objective in such problems is to determine the assortment that maximizes a yield-related objective, involving the likelihood of an item in the assortment being selected by a consumer and the value it creates for the retailer. In settings where the consumer response and substitution patterns are not known a priori and need to be inferred over the course of repeated (say, T) interactions, the problem involves a trade off between exploration (learning consumer preferences) and exploitation (selecting the optimal assortment), and this variant of the problem is the subject of the present paper. In particular, foreshadowing what is to come later, our interest focuses on the complexity of the problem as measured primarily by the interaction between N and K (governing the static combinatorial nature of the problem) and T (the problem horizon over which the aforementioned exploration and exploitation objectives need to be suitably balanced).

To formally state the online assortment optimization problem, let us index the N products described above by $1, 2, \cdots, N$ and their values will be referred to henceforth

V. Goyal is supported by the NSF grants CMMI 1201116 and CMMI 1351838. Author's addresses: S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi, Columbia University, New York 10027. Email: {sa3305, va2297, vg2277, ajz2001}@columbia.edu.

as revenues, and denoted as $r_1, \cdots, r_N$, respectively. Since the consumer need not select any product in a given assortment, we model this "no purchase option" as an additional product denoted "0" which augments the product index set. Let $p_i(S)$ be the probability, specified by the underlying choice model that a consumer purchases product i when assortment S is offered. Then the expected revenue corresponding to the assortment S, $R(S)$ is given by

$$R(S) = \sum_{i \in S} r_i p_i(S), \tag{1}$$

and the corresponding *static* assortment optimization problem is

$$\max_{S \in \mathcal{S}} R(S), \tag{2}$$

where $\mathcal{S}$ is the set of feasible assortments, with the constraint

$$\mathcal{S} = \{S \subset \{1, \cdots, N\} \mid |S| \leq K\}.$$

To complete the description of this problem, a choice model needs to be specified. 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 [Luce 1959] and Plackett [Plackett 1975], see also [Ben-Akiva and Lerman 1985; McFadden 1978; Train 2003; Wierenga 2008] for further discussion and survey of other commonly used choice models.) Under this model the probability that a consumer purchases product i when offered an assortment $S \subset \{1, \cdots, N\}$ is given by,

$$p_i(S) = \begin{cases} \dfrac{v_i}{v_0 + \sum_{j \in S} v_j}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where v_i is a parameter of the MNL model corresponding to product i. Without loss of generality, we can assume that $v_0 = 1$. It is also assumed that the MNL parameter corresponding to any product is less than or equal to one, i.e. $v_i \leq 1$. This is assumption is equivalent to claiming that the no purchase option is preferred to any other product (an observation which holds in most realistic retail setting and certainly in online display advertising). From (1) and (3), the expected revenue when assortment S is offered is given by

$$R(S) = \sum_{i \in S} r_i p_i(S) = \sum_{i \in S} \frac{r_i v_i}{1 + \sum_{j \in S} v_j}. \tag{4}$$

As alluded to above, many instances of assortment optimization problems commence with very limited or even no a priori information about consumer preferences. Traditionally, due to production considerations, retailers used to forecast the uncertain demand before the selling season starts and decide on an optimal assortment to be held throughout. There are a growing number of industries like fast fashion and online display advertising where demand trends change constantly and new products (or advertisements) can be introduced (or removed) from offered assortments in a fairly frictionless manner. In such situations, it is possible (and in fact essential) to experiment by offering different assortments and observing resulting purchases. Of course, gathering more information on consumer choice in this manner reduces the time remaining to exploit said information. Balancing this exploration-exploitation tradeoff is essential for maximizing expected revenues over the planning horizon. To formalize this, consider a time horizon T, where assortments can be offered at time periods $t = 1, \cdots, T$. If S^* is the optimal assortment for (2), when the values of $p_i(S)$, as given by (3), are

known a priori, and the decision maker has chosen to offer $S_1, \cdots, S_T$ at times $1, \cdots, T$ respectively, then his/her objective would be to select a (non-anticipating) sequence of assortments in a path-dependent manner (namely, based on observed responses) to maximize cumulative expected revenues over said horizon, or alternatively, minimize the *regret* defined as

$$Reg(T) = \sum_{t=1}^{T} R(S^*) - E[R(S_t)], \tag{5}$$

where $R(S)$ is the expected revenue when assortment S is offered as defined in (1). This exploration-exploitation problem, which we refer to as **bandit-MNL**, is the focus of this paper.

Further discussion on the MNL choice model. McFadden [McFadden 1973] showed that the multinomial logit model is based on a random utility model, where consumer's utilities for different products are independent Gumbel random variables and the consumers prefer the product that maximizes their utility. In particular, the utility of a product i is given by: $U_i = \mu_i + \xi_i$, where $\mu_i \in R$ denotes the mean utility that the consumer assigns to product i. $\xi_0, \cdots, \xi_N$ are independent and identically distributed random variables having a Gumbel distribution with location parameter 0 and scale parameter 1. If we let $\mu_i = \log v_i$, then the choice probabilities are given by the equation (3). Note that from equation (3), the probability of a consumer choosing product i decreases if there is a product in the offer set with high mean utility and increases if the products in the offer set with low mean utilities. Although, MNL is restricted by the independence of irrelevant attributes ($p_i(S)/p_j(S)$ is independent of S), the structure of choice probabilities (3) offers tractability in finding the optimal assortment and estimating the parameters v_i.

1.1. Our Contributions

Our main contributions are the following.

Parameter Independent Regret Bounds. We propose an online algorithm that judiciously balances the exploration and exploitation trade-off intrinsic to our problem. Under a mild assumption that no purchase outcome is the most frequent outcome, our dynamic policy achieves a regret bound of $O(\sqrt{NT}\log T + N\log^3 T)$; the bound is non-asymptotic, the "big oh" notation is used for brevity. Subject to the aforementioned mild assumption, this regret bound is independent of the parameters of the MNL choice model and hence holds uniformly over all problem instances. To the best of our knowledge, this is the first policy to have a parameter independent regret bound for the MNL choice model. It is also interesting to note that there is no dependence on the cardinality constraint K, despite the combinatorial complexity that is dictated by the relationship between N and K. Our algorithm is predicated on upper confidence bound (UCB) type logic, originally developed in the context of the multi-armed bandit (MAB) problem (cf. [Auer et al. 2002]); 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 and Optimality. We establish a non-asymptotic lower bound for the online assortment optimization problem under the MNL model. In particular, we show that any algorithm must incur a regret of $\Omega(\sqrt{NT})$. The bound is derived via a reduction of the online problem with the MNL model to a parametric multi-armed bandit problem, for which such lower bounds are constructed by means of standard information theoretic arguments. In particular, the lower bound constructs a "hard" instance

of the problem by considering arms with Bernoulli distributions that are barely distinguishable (from a hypothesis testing perspective), yet incur "high" regret for any algorithm. The online algorithm discussed above matches this lower bound up to a logarithmic (in T) term, establishing the near optimality of our proposed algorithm.

Intuitively, a large K implies combinatorially more possibilities of assortments, but it also allows the algorithm to learn more in every round since the algorithm observes consumer's response on K products (though the response for one product is not independent of other products in the offered assortment). Our upper and lower bounds demonstrate that the two factors balance each other out, so that the optimal algorithm can achieve regret bounds independent of the value of K.

Outline. We provide a literature review in Section 2. In Section 3, we present our algorithm for the bandit-MNL problem, and in Section 4, we prove our main result that this algorithm achieves an $\tilde{O}(\sqrt{NT})$ regret upper bound. Section 5 demonstrates the optimality of our regret bound by proving a matching lower bound of $\Omega(\sqrt{NT})$.

2. RELATED WORK

Static Assortment Optimization. The static assortment planning literature focuses on finding an optimal assortment assuming that the information on consumer preferences is known a priori and does not change throughout the entire selling period. Static assortment planning under various choice models has been studied extensively; [Kök and Fisher 2007] provides a detailed review, below we cite representative work avoiding an exhaustive survey. [Talluri and Van Ryzin 2004] consider the unconstrained assortment planning problem under the MNL model and establish that the optimal assortment can be obtained by a greedy algorithm, where products are added to the optimal set in order of their revenues. In the constrained case, recent work, following [Rusmevichientong et al. 2010] that treats the cardinality constrained problem, provides polynomial time algorithms to find optimal (or near optimal) assortments under the MNL model under capacity constraints ([Désir and Goyal 2014]) and totally unimodular constraints ([Davis et al. 2013]). As alluded to earlier, there are many extensions and generalization of the MNL that are still tractable, including mixed logit, nested logit and Markov chain based choice models; for some examples of work on these approaches, as well as further references see [Blanchet et al. 2013], [Davis et al. 2011], [Gallego et al. 2015], and [Farias et al. 2012].

Dynamic Assortment Optimization. In most dynamic settings, either the information on consumer preferences is not known, the demand trends (and substitution patterns) evolve over the selling horizon, or there are inventory constraints that are part of the "state" descriptor. The formulation and analysis of these problems tend to differ markedly. The present paper focuses on the case of dynamically learning consumer preferences (while jointly optimizing cumulative revenues), and therefore we restrict attention to relevant literature to this problem. To the best of our knowledge, [Caro and Gallien 2007] were the first to study the dynamic assortment planning under model/parameter uncertainty. Their work focuses on an independent demand model, where the demand for each product is not influenced by the demand for other products (that is, absent substitution), and employ a Bayesian learning formulation to estimate demand rates. Closer to the current paper is the work by [Rusmevichientong et al. 2010] and [Sauré and Zeevi 2013]. They consider a problem where the parameters of an ambient choice model are unknown a priori (the former exclusively focusing on MNL, the latter extending to more general Luce-type models). Both papers design algorithms that separate estimation and optimization into separate batches sequentially in time. Assuming that the optimal assortment and second best assortment are "well

separated," their main results are essentially upper bounds on the regret which are predicated in the observation that one can localize the optimal solution with high probability. In particular, in [Rusmevichientong et al. 2010] it is shown that $O\left(CN^2 \log T\right)$ exploration batches are needed while in [Sauré and Zeevi 2013] $O\left(CN \log T\right)$ explorations are required to compute an optimal solution with probability at least $\Omega\left(1 - \frac{1}{T}\right)$. As indicated, this leads to regret bounds which are $O(CN^2 \log T)$ for [Rusmevichientong et al. 2010], and $O(CN \log T)$ in [Sauré and Zeevi 2013], for a constant C that depends on the parameters of the MNL. The number of exploration batches in their approach specifically depend on the separability assumption and cannot be implemented in practice without an estimate of C.

Relationship to MAB problems. A naive translation of the bandit-MNL problem to an MAB-type setting would create $\binom{N}{K}$ "arms" (one for each assortment of size K). For an "arm" corresponding to subset S, the reward is give by $R(S)$. One can apply a standard UCB-type algorithm to this structure. Of course the non-empty intersection of elements in these "arms" creates dependencies which are not being exploited by any generic MAB algorithm that is predicated on the arms being independent. Perhaps more importantly, this translation would naturally result in a bound that is combinatorial in the leading constant. Our approach in this paper customizes a UCB-type algorithm to the specifics of the assortment problem in a manner that creates a tractable complexity, which is also shown to be best possible in the sense of the achieved regret.

A closely related setting is that of bandit submodular maximization under cardinality constraints, see [Golovin and Krause 2012], where the revenue for set S is given by a submodulnar function $f(S)$. On offering subset S, the marginal benefit $f(S_i) - f(S_{i-1})$ of each item i in S is observed, assuming the items of S were offered in some order. Under K-cardinality constraint, the best available regret bounds for this problem (in non-stochastic setting) are upper and lower bounds of $O(K\sqrt{NT \log(N)})$ and $\Omega(\sqrt{KT \log N})$, respectively [Streeter and Golovin 2009]. Many special cases of the submodular maximization problem have been considered for applications in learning to rank documents in *web search* (e.g., see [Radlinski et al. 2008]).

In comparison, in the bandit-MNL problem considered in this paper, the reward function $R(S)$ for assortment S is not submodular – it only has a *restricted submodularity* property [Aouad et al. 2015], where the submodularity property holds over sets containing less than certain number of elements. We provide an algorithm with regret upper bound of $\tilde{O}(\sqrt{NT})$ for any $K \leq N$, and present a matching lower bound of $\Omega(\sqrt{NT})$, in stochastic setting.

Other related work includes *limited feedback settings* where on offering S, only $f(S)$ is observed by the algorithm, and *not* individual feedback for arms in S. For example, in [Hazan and Kale 2012], $f(S)$ is submodular, and in linear bandit problem [Auer 2003], $f(S)$ is a linear function. There, due to limited feedback, the available regret guarantees are much worse, and depend linearly on (dimension) N.

3. ALGORITHM

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

3.1. Challenges and overview

A key difficulty in applying standard UCB-like multi-armed bandit techniques for 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. As mentioned before, a naive extension of MAB algorithms for this problem would treat each of the $\binom{N}{K}$ possible assortments as an arm, leading to a computationally inefficient algorithm with regret exponential in K. Our algorithm utilizes the specific properties of the dependence structure in MNL model to obtain an efficient algorithm with $\tilde{O}(\sqrt{NT})$ regret.

Our algorithm is based on a non-trivial extension of the UCB algorithm [Auer et al. 2002]. It uses the past observations to maintain increasingly accurate upper confidence bounds for MNL parameters $\{v_i, i = 1, \ldots, N\}$, and uses these to (implicitly) maintain an estimate of expected revenue $R(S)$ for every feasible assortment S. In every round, it picks the assortment S with the highest estimated revenue. There are two main challenges in implementing this scheme. First, the customer response on offering 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: a chosen S is offered repeatedly until a no-purchase happens. We show that on proceeding in this manner, the average number of times a product i is purchased provides an unbiased estimate of 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 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. Algorithmic details

We divide the time horizon into epochs, where in each epoch we offer an assortment repeatedly until a no purchase outcome happens. 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 $|\mathcal{E}_\ell|$ is a geometric random variable with success probability as 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 $i \in S$, 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.

$$\hat{v}_{i,\ell} := \sum_{t \in \mathcal{E}_\ell} \mathbb{I}(c_t = i) \tag{6}$$

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

$$\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\tau\}, \;\; T_i(\ell) = |\mathcal{T}_i(\ell)|. \tag{7}$$

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

$$\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}. \tag{8}$$

In Claim 2, we prove 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. Using these estimates, we compute $v_{i,\ell}^{UCB}$ as,

$$v_{i,\ell}^{UCB} := \bar{v}_{i,\ell} + \sqrt{\frac{12\bar{v}_{i,\ell}}{T_i(\ell)}} \log T + \frac{30 \log^2 T}{T_i(\ell)}. \tag{9}$$

In next section (Lemma 4.2), we prove that $v_{i,\ell}^{UCB}$ is an upper confidence bound on true parameter v_i, i.e., $v_{i,\ell}^{UCB} \geq v_i, \forall i, \ell$ with high probability.

Based on the above estimates, we define an estimate $\tilde{R}_{\ell+1}(S)$ for expected revenue of each assortment S, as

$$\tilde{R}_{\ell+1}(S) := \frac{\sum_{i \in S} r_i v_{i,\ell}^{UCB}}{1 + \sum_{j \in S} v_{j,\ell}^{UCB}}. \tag{10}$$

In epoch $\ell + 1$, the algorithm picks assortment $S_{\ell+1}$, computed as the assortment $S \in \mathcal{S}$ with highest value of $\tilde{R}_{\ell+1}(S)$, i.e.,

$$S_{\ell+1} := \underset{S \in \mathcal{S}}{\text{argmax}}\ \tilde{R}_{\ell+1}(S). \tag{11}$$

We summarize the steps in our algorithm as Algorithm 1. Finally, we may remark

ALGORITHM 1: Exploration-Exploitation algorithm for bandit-MNL

Initialization: $v_{i,0}^{UCB} = 1$ for all $i = 1, \cdots, N$.
$t = 1$, keeps track of the time steps
$\ell = 1$, keeps count of total number of epochs
repeat
  Compute $S_\ell = \underset{S \in \mathcal{S}}{\text{argmax}}\ \tilde{R}_\ell(S) = \frac{\sum_{i \in S} r_i v_{i,\ell-1}^{UCB}}{1 + \sum_{j \in S} v_{j,\ell-1}^{UCB}}$
  Offer assortment S_ℓ, observe the purchasing decision, $\mathbf{c_t}$ of the consumer
  if $c_t = 0$ **then**
    compute $\hat{v}_{i,\ell} = \sum_{t \in \mathcal{E}_\ell} \mathbb{I}(c_t = i)$, no. of consumers who preferred i in epoch ℓ, for all $i \in S_\ell$.
    update $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}$, $T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until ℓ that offered product i.
    update $\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
    update $v_{i,\ell}^{UCB} = \bar{v}_{i,\ell} + \sqrt{\frac{12\bar{v}_{i,\ell}}{T_i(\ell)}} \log T + \frac{30 \log^2 T}{T_i(\ell)}$
    $\ell = \ell + 1$
  else
    $\mathcal{E}_\ell = \mathcal{E}_\ell \cup t$, time indices corresponding to epoch ℓ.
  end
  $t = t + 1$
until $t < T$;

on the computational complexity of implementing (11). Since we are only interested in finding the assortment $S \in \mathcal{S}$ with the largest value of $\tilde{R}_\ell(S)$ in epoch ℓ, we can avoid explicitly calculating $\tilde{R}_\ell(S)$ for all S. Instead, we observe that (11) can be formulated

as a static K-cardinality constrained assortment optimization problem under MNL model, with model parameters being $v_{i,\ell}^{UCB}, i = 1, \ldots, N$. There are efficient polynomial time algorithms to solve the static assortment optimization problem under MNL model with known parameters. [Davis et al. 2013] showed a simple linear programming formulation of this problem. [Rusmevichientong et al. 2010] proposed an enumerative method that utilizes an observation that optimal assortment belongs to an efficiently enumerable collection of N^2 assortments.

4. REGRET ANALYSIS

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

THEOREM 4.1. *For any instance of the bandit-MNL problem with N products, $1 \leq K \leq N$, $r_i \in [0,1]$ and $v_0 \geq v_i$ for $i = 1, \ldots, N$, the regret of Algorithm 1 in time T is bounded as,*

$$Reg(T) = O(\sqrt{NT}\log T + N\log^3 T).$$

4.1. Proof Outline

The first step in our regret analysis is to prove the following two properties of the estimates $v_{i,\ell}^{UCB}$ computed as in (9) for each product i. 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 [Auer et al. 2002].

(1a) The estimate $v_{i,\ell}^{UCB}$ for every i, is larger than v_i with high probability, i.e.,

$$v_{i,\ell}^{UCB} \geq v_i, \forall i, \ell$$

(2a) As a product is offered more and more, its estimate approaches the actual parameter v_i, so that in epoch ℓ, with high probability the difference between the estimate and actual parameter can be bounded as

$$v_{i,\ell}^{UCB} - v_i \leq \tilde{O}\left(\sqrt{\tfrac{v_i}{T_i(\ell)}} + \tfrac{1}{T_i(\ell)}\right), \forall i, \ell$$

Lemma 4.2 provides the precise statements of above properties and proves that these hold with probability at least $1 - O\left(\frac{1}{T^2}\right)$. To prove this lemma, we first employ an observation conceptually equivalent to the IIA (Independence of Irrelevant Alternatives) property of MNL model to show 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 chose to offer S_τ repeatedly until no-purchase happened. Given these unbiased i.i.d. estimates from every epoch τ before ℓ, we apply a multiplicative Chernoff-Hoeffding bound to prove concentration of $\bar{v}_{i,\ell}$. Then, above properties follow from definition of $v_{i,\ell}^{UCB}$.

The product demand estimates $v_{i,\ell-1}^{UCB}$ were used in (10) to define expected revenue estimates $\tilde{R}_\ell(S)$ for every set S. In the beginning of every epoch ℓ, Algorithm 1 computes the maximizer $S_\ell = \arg\max_S \tilde{R}_\ell(S)$, and then offers S_ℓ repeatedly until no-purchase happens. The next step in regret analysis is to use above properties of $v_{i,\ell}^{UCB}$ to prove similar, though slightly weaker, properties for estimates $\tilde{R}_\ell(S)$. We prove that the following hold with high probability.

(1b) The estimate $\tilde{R}_\ell(S^*)$ is an upper confidence bound on $R(S^*)$, i.e., $\tilde{R}_\ell(S^*) \geq R(S^*)$. By choice of S_ℓ, it directly follows that

$$\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*) \geq R(S^*)$$

Note that we do not claim that for every S, $\tilde{R}_\ell(S)$ is an upper confidence bound on $R(S)$; infact we observe that this property holds only for S^* and certain other special $S \in \mathcal{S}$. Above weaker guarantee will suffice for our regret analysis, and it allows a more efficient algorithm that does not require to maintain an explicit upper confidence bound for every set S.

(2b) The difference between the estimated revenue and actual expected revenue for the *offered assortment* S_ℓ is bounded as

$$(1 + \sum_{j \in S_\ell} v_j)(\tilde{R}_\ell(S_\ell) - R(S_\ell)) \leq \tilde{O}\left(\sum_{i \in S_\ell} \sqrt{\frac{v_i}{T_i(\ell)}} + \frac{1}{T_i(\ell)}\right), \forall i, \ell$$

Lemma 4.3 and Lemma 4.4 provide the precise statements of above properties, and prove that these hold with probability at least $1 - O\left(\frac{1}{T^2}\right)$. The proof of the property (1b) above involves careful use of the structure of MNL model to show that the value of

$$\tilde{R}_\ell(S_\ell) = \max_{S \in \mathcal{S}} \frac{\sum_{i \in S} r_i v_{i,\ell}^{UCB}}{1 + \sum_{j \in S} v_{j,\ell}^{UCB}}$$

is equal to the highest expected revenue achievable by any assortment (of at most K-cardinality), among *all instances of the problem with parameters in the range* $[0, v_i^{UCB}], i = 1, \ldots, n$. Since the actual parameters lie in this range with high probability, we obtain that $\tilde{R}_\ell(S_\ell)$ is at least $R(S^*)$. For property (2b) above, we prove a Lipschitz property of function $\tilde{R}_\ell(S)$ and bound its error in terms of errors in individual product estimates $|v_{i,\ell}^{UCB} - v_i|$.

Given above properties, the rest of the proof is relatively straightforward. Recall that in epoch ℓ, assortment S_ℓ is offered, for which expected revenue is $R(S_\ell)$. Epoch ℓ ends when a no purchase happens on offering S_ℓ, where the probability of the no-purchase event is $1/(1 + \sum_{j \in S_\ell} v_j)$. Therefore, expected length of an epoch is given by $(1 + \sum_{j \in S_\ell} v_j)$. Using these observations, we show that the total expected regret can be bounded by

$$Reg(T) \leq E[\sum_{\ell=1}^{L} (1 + V(S_\ell))(R(S^*) - R(S_\ell))],$$

where $V(S_\ell) := \sum_{j \in S_\ell} v_j$. Then, using property (1b) and (2b) above, we can further bound this as

$$Reg(T) \leq \sum_\ell (1 + V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell)) \leq \sum_\ell \tilde{O}\left(\sum_{i \in S_\ell} \sqrt{\frac{v_i}{T_i(\ell)}} + \frac{1}{T_i(\ell)}\right) = \tilde{O}(\sum_i \sqrt{v_i T_i}),$$

where T_i denotes the total number of epochs in which product i was offered. Note that $\sum_i T_i \leq TK$, since in each epoch, the set S_ℓ can contain at most K products, and there are at most T epochs. Using this loose bound, we would obtain that in worst case, $T_i = TK/N$, and using $v_i \leq 1$ for each i, we get that regret is bounded by $\tilde{O}(\sqrt{NKT})$. We derive a more careful bound on number of epochs T_i based on the value of corresponding parameter v_i to obtain an $\tilde{O}(\sqrt{NT})$ regret, as stated in Theorem 4.1.

In rest of this section, we follow the above outline to provide a detailed proof of Theorem 4.1. The proof is organized as follows. In Section 4.2, we prove Property (1a) and (2a) for estimates $v_{i,\ell}^{UCB}$. In Section 4.3, we prove Property (1b) and (2b) for estimates $\tilde{R}_\ell(S_\ell)$. Finally, in Section 4.4, we utilize these properties to complete the proof of Theorem 4.1 .

4.2. Properties of estimates $v_{i,\ell}^{UCB}$

First, we focus on the concentration properties of $\hat{v}_{i,\ell}$ and $\bar{v}_{i,\ell}$, and then utilize those to establish the necessary properties of $v_{i,\ell}^{UCB}$.

4.2.1. Unbiased Estimates. It is not clear if the estimates $\hat{v}_{i,\ell}, \ell \leq L$ are independent of each other. In our setting, it is possible that the distribution of estimate $\hat{v}_{i,\ell}$ depends on the offered assortment S_ℓ, which in turn depends on previous estimates. In the following result, we show that the moment generating of $\hat{v}_{i,\ell}$ only depends on the parameter v_i and not on the offered assortment S_ℓ, there by establishing that estimates are identically and independently distributed. Using the moment generating function, we show that $\hat{v}_{i,\ell}$ is an unbiased estimate for v_i, i.e., $E(\hat{v}_{i,\ell}) = v_i$ and bounded with high probability.

CLAIM 1. *The moment generating function of estimate $\hat{v}_i$, $E\left(e^{\theta \hat{v}_{i,\ell}}\right)$ is given by,*

$$E\left(e^{\theta \hat{v}_{i,\ell}}\right) = \frac{1}{1 - v_i(e^\theta - 1)}, \text{ for all } \theta \leq \log 2, \text{ for all } i = 1, \cdots, N.$$

PROOF. From (3), 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

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

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

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

If $\alpha(1-p) < 1$ and n is a geometric random variable with parameter p, we have

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

Note that for all $\theta < \log 2$, we have $\left(q_i e^\theta + (1 - q_i)\right)(1 - p_0) = (1 - p_0) + p_0 v_i(e^\theta - 1) < 1$. Therefore, we have $E\left(e^{\theta \hat{v}_{i,\ell}}\right) = \frac{1}{1 - v_i(e^\theta - 1)}$ for all $\theta < \log 2$. □

We can establish that $\hat{v}_{i,\ell}$ is unbiased estimator of v_i by computing the differential the moment generating function and setting $\theta = 0$. Since $\hat{v}_{i,\ell}$ is an unbiased estimate, it follows by definition (refer to (8)) that $\bar{v}_{i,\ell}$ is also an unbiased estimate for v_i. Therefore, from Claim 1, we have the following result.

CLAIM 2. *We have the following claims.*

(1) $\hat{v}_{i,\ell},\ \ell \leq L$ are unbiased i.i.d estimates of v_i, i .e. $E\left(\hat{v}_{i,\ell}\right) = v_i\ \forall\ \ell, i$.
(2) $E\left(\bar{v}_{i,\ell}\right) = v_i$

(3) $\mathcal{P}\left(\hat{v}_{i,} > 8\log T\right) \leq \frac{2}{T^3}$ *for all* $i = \{1, \cdots, N\}$
(4) $\mathcal{P}\left(\bar{v}_{i,\ell} > 2v_i + 8\log T\right) \leq \frac{2}{T^3}$ *for all* $i = \{1, \cdots, N\}$

PROOF. We establish (1) by differentiating the moment generating function established in Claim 1 and setting $\theta = 0$. (2) directly follows from (1). Evaluating the moment generating function at $\theta = \log 3/2$ and using Chernoff bound, we establish (3). Applying Chernoff bounds on $\sum_{\tau=1}^{\ell} \hat{v}_{i,\ell}$ and using the fact that $\hat{v}_{i,\ell}$ are i.i.d., we show (4). The proof for (4) is non trivial and the details are provided in Claim A.1.

4.2.2. Concentration Bounds. From Claim 2, it follows that $\hat{v}_{i,\tau}, \tau \in \mathcal{T}_i(\ell)$ are i.i.d random variables that are bounded with high probability and $E(\hat{v}_{i,\tau}) = v_i$ for all $\tau \in \mathcal{T}_i(\ell)$. We will combine these two observations and extend multiplicative Chernoff-Hoeffding [Babaioff et al. 2011] inequality to establish the following result.

CLAIM 3. *We have the following inequalities.*

(1) $\mathcal{P}\left(|\bar{v}_{i,\ell} - v_i| \geq \sqrt{12\frac{\bar{v}_{i,\ell}}{T_i(\ell)}}\log T + \frac{30\log^2 T}{T_i(\ell)}\right) \leq O\left(\frac{1}{T}\right)$.
(2) $\mathcal{P}\left(|\bar{v}_{i,\ell} - v_i| \geq \sqrt{6\frac{v_i}{T_i(\ell)}}\log T + \frac{30\log^2 T}{T_i(\ell)}\right) \leq O\left(\frac{1}{T}\right)$.

Note that to apply Chernoff-Hoeffding inequality, we must have the individual sample values bounded by some constant, which is not the case with our estimates $\hat{v}_{i,\tau}$. However, we proved earlier that these estimates are bounded by $\Omega\left(8\log T\right)$ with probability at least $1 - O(\frac{1}{T^3})$ and we use truncation technique to establish Claim 3. We complete the proof of Claim 3 in Appendix A.

The following result follows from Claim 2 and 3, and establishes the necessary properties of $v_{i.\ell}^{UCB}$ alluded to as properties 1(a) and 2(a) in the proof outline.

LEMMA 4.2. *We have the following claims.*

(1) $v_{i,\ell}^{UCB} \geq v_i$ *with probability at least* $1 - O\left(\frac{1}{T}\right)$ *for all* $i = 1, \cdots, N$.
(2) *There exists constants* C_1 *and* C_2 *such that*

$$v_{i,\ell}^{UCB} - v_i \leq C_1\sqrt{\frac{v_i}{T_i(\ell)}}\log T + C_2\frac{\log^2 T}{T_i(\ell)}$$

with probability at least $1 - O\left(\frac{1}{T}\right)$.

4.3. Properties of estimate $\tilde{R}(S)$

In this section we establish properties of upper bound estimate $\tilde{R}_\ell(S)$. First, we establish the following result (property 1(b) in the proof outline).

LEMMA 4.3. *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 any epoch* ℓ*, we have*

$$\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*) \geq R(S^*) \text{ with probability at least } 1 - O\left(\frac{1}{T}\right).$$

Let $R(S, \mathbf{w})$ denote the expected revenue when assortment S is offered and if the parameters of the MNL were given by the vector $\mathbf{w}$, i.e.

$$R(S, \mathbf{w}) := \sum_{i \in S} \frac{w_i r_i}{1 + \sum_{j \in S} w_j},$$

Then, $R(S) = R(S, \mathbf{v})$, and from definition of $\tilde{R}_\ell(S)$ (refer to (10)),

$$\tilde{R}_\ell(S) = R(S, \mathbf{v}_\ell^{UCB}).$$

CLAIM 4. *Assume $0 \le w_i \le v_i^{UCB}$ for all $i = 1, \cdots, n$. Suppose S is an optimal assortment when the MNL are parameters are given by* $\mathbf{w}$. *Then,*

$$R(S, \mathbf{v}^{UCB}) \ge R(S, \mathbf{w}).$$

PROOF. We prove the result by first showing that for any $j \in S$, we have

$$R(S, \mathbf{w}(j)) \ge R(S, \mathbf{w}), \tag{12}$$

where $\mathbf{w}(j)$ is vector $\mathbf{v}$ with the j^{th} component increased to v_j^{UCB},

$$\mathbf{w}(j) = \begin{cases} w_i & \text{if } i \neq j \\ v_j^{UCB} & \text{if } i = j \end{cases}.$$

We first establish that for any $j \in S,\ r_j \ge R(S)$. For the sake of contradiction, suppose for some $j \in S$, we have, $r_j < R(S)$, then by multiplying with w_j on both sides of the inequality, we have,

$$w_j r_j (1 + \sum_{i \in S} w_i) < w_j (\sum_{i \in S} r_i w_i),$$

adding $(\sum_{i \in S} r_i w_i)(\sum_{i \in S} w_i + 1)$ to both sides of the inequality, we get,

$$(\sum_{i \in S} r_i w_i)(\sum_{i \in S} w_i + 1) + w_j r_j (1 + \sum_{i \in S} w_i) < (\sum_{i \in S} r_i w_i)(\sum_{i \in S} w_i + 1) + w_j (\sum_{i \in S} r_i w_i).$$

Rearranging the terms from the above inequality, it follows that,

$$(\sum_{i \in S} r_i w_i)(\sum_{i \in S} w_i + 1) - w_j r_j (1 + \sum_{i \in S} w_i) > (\sum_{i \in S} r_i w_i)(\sum_{i \in S} w_i + 1) - w_j (\sum_{i \in S} r_i w_i).$$

implying,

$$\frac{\sum_{i \in S} r_i w_i - w_j r_j}{1 + \sum_{i \in S} w_i - w_j} > \frac{\sum_{i \in S} r_i w_i}{\sum_{i \in S} w_i + 1},$$

which can be rewritten as,

$$\frac{\sum_{i \in S/j} r_i w_i}{1 + \sum_{i \in S/j} w_i} > \frac{\sum_{i \in S} r_i w_i}{\sum_{i \in S} w_i + 1}$$

contradicting that S is the optimal assortment when the parameters are $\mathbf{w}$. Therefore,

$$r_j \ge \frac{\sum_{i \in S} r_i w_i}{1 + \sum_{i \in S} w_i} \text{ for all } j \in S.$$

Multiplying by $(v_j^{UCB} - w_j)(\sum_{i \in S/j} w_i + 1)$ on both sides of the above inequality, we obtain

$$(v_j^{UCB} - w_j) r_j \left(\sum_{i \in S/j} w_i + 1 \right) \ge (v_j^{UCB} - w_j) \left(\sum_{i \in S/j} w_i r_i \right),$$

from which we have inequality (12). The result follows from (12), which establishes that increasing one parameter of MNL to the highest possible value increases the value of $R(S, \mathbf{w})$. □

Let $\hat{S}, \mathbf{w}^*$ be maximizers of the following optimization problem.

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

Then applying Claim 4 on assortment $\hat{S}$ and parameters $\mathbf{v}^*$ and noting that $v_{i,\ell}^{UCB} > v_i$ with high probability, we have that

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

Now we will establish the connection between the error on the expected revenues and the error on the estimates of MNL parameters. In particular, we have the following result.

LEMMA 4.4. *There exists constants C_1 and C_2 such that*

$(1+\sum_{j \in S_\ell} v_j)(\tilde{R}_\ell(S_\ell) - R(S_\ell)) \le C_1 \sqrt{\frac{v_i}{|\mathcal{T}_i(\ell)|}} \log T + C_2 \frac{\log^2 T}{|\mathcal{T}_i(\ell)|}$, *with probability at least* $1 - O\left(\frac{1}{T}\right)$

The above result follows directly from the following result and Lemma 4.2.

CLAIM 5. *If $0 \le v_i \le v_{i,\ell}^{UCB}$ for all $i \in S_\ell$, then*

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

PROOF.

$$\tilde{R}_\ell(S_\ell) - R(S_\ell) = \frac{\sum_{i \in S_\ell} r_i v_{i,\ell}^{UCB}}{1 + \sum_{j \in S_\ell} v_{j,\ell}^{UCB}} - \frac{\sum_{i \in S_\ell} r_i v_i}{1 + \sum_{j \in S_\ell} v_j}.$$

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

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

□

4.4. Putting it all together: Proof of Theorem 4.1

In this section, we formalize the intuition developed in the previous sections and complete the proof of Theorem 4.1.

Let S^* denote the optimal assortment and $r_t(S_\ell)$ be the expected revenue generated by offering the assortment S_ℓ at time t. Our objective is to minimize the *regret* defined in (5), which is same as

$$Reg(T) = E\left(\sum_{\ell=1}^{L} \sum_{t \in \mathcal{E}_\ell} \left(R(S^*) - r_t(S_\ell) \right) \right). \tag{13}$$

For every epoch ℓ, let t_ℓ denote the time index when the no purchase happened, after which the algorithm progressed to the next epoch. Observe Algorithm 1 by design,

offers an assortment until a no purchase happens. Hence, the conditional expectation of $r_t(S_\ell)$ given S_ℓ, $E\left(r_t(S_\ell) \,|\, S_\ell\right)$ is not the same as $R(S_\ell)$, but is given by

$$E\left(r_t(S_\ell) \,|\, S_\ell\right) = \begin{cases} E\left(r_t(S_\ell) \,|\, S_\ell, \{r_t(S_\ell) \neq 0\}\right) & \text{if } t \neq t_\ell \\ E\left(r_t(S_\ell) \,|\, S_\ell, \{r_t(S_\ell) = 0\}\right) & \text{if } t = t_\ell \end{cases}.$$

Hence, we have

$$E\left(r_t(S_\ell) \,|\, S_\ell\right) = \begin{cases} \dfrac{1 + \sum_{j \in S_\ell} v_j}{\sum_{i \in S_\ell} v_i} R(S_\ell) & \text{if } t < t_\ell \\ 0 & \text{if } t = t_\ell \end{cases}.$$

Note that L, $\mathcal{E}_\ell$, S_ℓ and $r_t(S_\ell)$ are all random variables and the expectation in equation (13) is over these random variables. Therefore, the regret can be reformulated as

$$Reg(T) = E\left\{ \sum_{\ell=1}^{L} (1 + \sum_{j \in S_\ell} v_j) \left[R(S^*) - R(S_\ell)\right] \right\}, \tag{14}$$

the expectation in equation (14) is over the random variables L and S_ℓ. We now provide the proof for Theorem 4.1.

PROOF. of Theorem 4.1 Let $V(S_\ell) = \sum_{j \in S_\ell} v_j$, from equation (14), we have that

$$Reg(T) = E\left\{ \sum_{\ell=1}^{L} \left(1 + V(S_\ell)\right) \left(R(S^*) - R(S_\ell)\right) \right\}$$

For sake of brevity, let ΔR_ℓ=$(1 + V(S_\ell))\left(R(S^*) - R(S_\ell)\right)$, for all $\ell = 1, \cdots, L$. Now the regret can be reformulated as

$$Reg(T) = E\left\{ \sum_{\ell=1}^{L} \Delta R_\ell \right\} \tag{15}$$

Let T_i denote the total number of epochs that offered an assortment containing product i. Let $\mathcal{A}_0$ denote the complete set Ω and for all $\ell = 1, \ldots, L$, event $\mathcal{A}_\ell$ is given by

$$\mathcal{A}_\ell = \left\{ v_{i,\ell}^{UCB} < v_i \text{ or } v_{i,\ell}^{UCB} > v_i + C_1 \sqrt{\frac{v_i}{T_i(\ell)}} \log T + C_2 \frac{\log^2 T}{T_i(\ell)} \text{ for some } i \in S_\ell \cup S^* \right\}.$$

Noting that $\mathcal{A}_\ell$ is a rare event and our earlier results on the bounds are true whenever event $\mathcal{A}_\ell^c$ happens, we try to analyze the regret in two scenarios, one when $\mathcal{A}_\ell$ is true and another when $\mathcal{A}_\ell^c$ is true. For any event A, let $\mathbb{I}(A)$ denote the indicator random variable for the event A. Hence, we have

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

Using the fact that $R(S^*)$ and $R(S_\ell)$ are both bounded by one and $V(S_\ell) \leq K$, we have

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

Whenever $\mathbb{I}(\mathcal{A}_{\ell-1}^c) = 1$, from Lemma 4, we have $\tilde{R}_\ell(S^*) \geq R(S^*)$ and by our algorithm design, we have $\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*)$ for all $\ell \geq 2$. Therefore, it follows that

$$E\left\{\Delta R_\ell\right\} \leq (K+1)\mathcal{P}(\mathcal{A}_{\ell-1}) + E\left\{ \left[(1 + V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell))\right] \cdot \mathbb{I}(\mathcal{A}_{\ell-1}^c) \right\}$$

From Lemma 4.4, it follows that

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

Therefore, we have

$$E\left\{\Delta R_\ell\right\}\leq(K+1)\mathcal{P}\left(\mathcal{A}_{\ell-1}\right)+CE\left(\log T\sum_{j\in S_\ell}\left(\sqrt{\frac{v_i}{T_i(\ell)}}+\frac{\log T}{T_i(\ell)}\right)\right)\tag{16}$$

where $C=\max\{C_1,C_2\}$. Combining equations (15) and (16), we have

$$Reg(T)\leq E\left\{\sum_{\ell=1}^{L}\left[(K+1)\mathcal{P}\left(\mathcal{A}_{\ell-1}\right)+C\log T\sum_{j\in S_\ell}\left(\sqrt{\frac{v_i}{T_i(\ell)}}+\frac{\log T}{T_i(\ell)}\right)\right]\right\}.$$

Therefore, from Lemma 4.2, we have

$$\begin{aligned}Reg(T)&\leq CE\left\{\sum_{\ell=1}^{L}\frac{K+1}{T}+\sum_{j\in S_\ell}\sqrt{\frac{v_i}{T_i(\ell)}}\log T+\sum_{j\in S_\ell}\frac{\log^2 T}{T_i(\ell)}\right\},\\&\overset{(a)}{\leq}CK+CN\log^3 T+(C\log T)E\left(\sum_{i=1}^{n}\sqrt{v_iT_i}\right)\\&\overset{(b)}{\leq}CK+CN\log^3 T+(C\log T)\sum_{i=1}^{N}\sqrt{v_iE(T_i)}\end{aligned}\tag{17}$$

Inequality (a) follows from the observation that $L\leq T$, $T_i\leq T$, $\sum_{T_i(\ell)=1}^{T_i}\frac{1}{\sqrt{T_i(\ell)}}\leq\sqrt{T_i}$ and $\sum_{T_i(\ell)=1}^{T_i}\frac{1}{T_i(\ell)}\leq\log T_i$, 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 $\sum_{\ell=1}^{L}n_\ell\leq T$. Hence, we have $E\left(\sum_{\ell=1}^{L}n_\ell\right)\leq T$. Let $\mathcal{S}$ denote the filtration corresponding to the offered assortments $S_1,\cdots,S_L$, then by law of total expectation, we have,

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

Therefore, it follows that

$$\sum v_iE(T_i)\leq T.\tag{18}$$

To obtain the worst case upper bound, we maximize the bound in equation (17) subject to the condition (18) and hence, we have $Reg(T)=O(\sqrt{NT}\log T+N\log^3 T)$ $\square$

5. LOWER BOUNDS

In this section, we establish that any algorithm must incur a regret of $\Omega(\sqrt{NT/K})$. More precisely, we prove the following result.

THEOREM 5.1. *There exists a (randomized) instance of the bandit-MNL problem with $v_0 \geq v_i$, $i = 1, \cdots, N$, such that for any N, $K < N$, $T \geq N$, and any algorithm $\mathcal{A}$ that offers assortment $\mathcal{S}_t^{\mathcal{A}}$, $|S_t^{\mathcal{A}}| \leq K$ at time t, we have*

$$E\left[Reg(T)\right] := E\left(\sum_{t=1}^{T} R(S^*) - R(S_t^{\mathcal{A}})\right) \geq C\sqrt{\frac{NT}{K}}$$

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

5.1. Proof Overview

We prove Theorem 5.1 by a reduction to a parametric multi-armed bandit (MAB) problem, for which a lower bound is known.

Definition 5.2 (*MAB instance I_{MAB}*). Define I_{MAB} as a (randomized) instance of MAB problem with $N \geq 2$ Bernoulli arms, and following parameters (probability of reward 1)

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

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

LEMMA 5.3. *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 of algorithm $\mathcal{A}$ on instance I_{MAB} is at least $\frac{\epsilon T}{6}$. That is,*

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

where expectation is both over the randomization in generating the instance (value of j), and the random outcomes of pulled arms during execution of the algorithm on an instance.

The proof of Lemma 5.3 is a simple extension of the proof of $\Omega(\sqrt{NT})$ 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 for the sake of completeness. We use the above lower bound for MAB to prove that any algorithm must incur at least $\Omega(\sqrt{NT/K})$ regret on the following instance of the bandit-MNL problem.

Definition 5.4 (*bandit-MNL instance I_{MNL}*). Define I_{MNL} as the following (randomized) instance of bandit-MNL 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 $\epsilon = \frac{1}{100}\sqrt{\frac{N\alpha}{T}}$.

We will show that any bandit-MNL 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 $\mathcal{A}_{MNL}$ for bandit-MNL that achieves $o(\sqrt{\frac{NT}{K}})$ regret on instance I_{MNL}, then we can use $\mathcal{A}_{MNL}$ as a subroutine to construct an algorithm $\mathcal{A}_{MAB}$ for MAB that achieves strictly less than $\frac{\epsilon T}{6}$ regret on instance I_{MAB} in time T, thus contradicting the lower bound of Lemma 5.3. This will prove Theorem 5.1 by contradiction.

5.2. Construction of algorithm $\mathcal{A}_{MAB}$ using $\mathcal{A}_{MNL}$

ALGORITHM 2: Algorithm $\mathcal{A}_{MAB}$

Initialization: $t = 0$, $\ell = 0$.
$\mathcal{A}_{MNL}$ suggests to offer assortment $S_0^{\mathcal{A}_{MNL}} \subset [\hat{N}]$ such that $|S_0^{\mathcal{A}_{MNL}}| \leq K$
repeat
  Update $\ell = \ell + 1$;
  Call $\mathcal{A}_{MNL}$, receive assortment $S_\ell \subset [\hat{N}]$;
  Repeat until 'exit loop'
  With probability $\frac{1}{2}$, send **Feedback to** $\mathcal{A}_{MNL}$ 'no product was purchased', **exit loop**;
    provide feedback to $\mathcal{A}_{MNL}$ no product was purchased, and **exit**;
  Update $t = t + 1$;
  Pull arm $\mathcal{A}_t = \lceil \frac{i}{K} \rceil$,where $i \in S_\ell$ is chosen uniformly at random.
  If reward is 1, send **Feedback to** $\mathcal{A}_{MNL}$ 'i was purchased' and **exit loop**;
until $t \leq T$;

Algorithm 2 provides the exact construction of $\mathcal{A}_{MAB}$. $\mathcal{A}_{MAB}$ simulates $\mathcal{A}_{MNL}$ algorithm as a blackbox. Note that $\mathcal{A}_{MAB}$ pulls arms at time steps $t = 1, \ldots, T$. These arm pulls are interleaved by simulations of $\mathcal{A}_{MNL}$ steps (**Call** $\mathcal{A}_{MNL}$, **Feedback to** $\mathcal{A}_{MNL}$). When step ℓ of $\mathcal{A}_{MNL}$ is simulated, it uses the feedback from $1, \ldots, \ell - 1$ to suggest an assortment S_ℓ; and expects a feedback from $\mathcal{A}_{MAB}$ about which product (or no product) was purchased out of those offered in S_ℓ, where the probability of purchase of product $i \in S_\ell$ must be $\frac{v_i}{v_0 + \sum_{i \in S_\ell} v_i}$. Following claim shows that $\mathcal{A}_{MAB}$ provides the right feedback to $\mathcal{A}_{MNL}$. Proof is omitted due to space constraints.

CLAIM 6. *For any assortment S_ℓ suggested by $\mathcal{A}_{MNL}$, for each $i \in S_\ell$, let $\mathcal{P}_{S_\ell}(i)$ denote the probability that $\mathcal{A}_{MAB}$ gives the feedback that product i is purchased. And, let $\mathcal{P}_{S_\ell}(0)$ denote the probability that $\mathcal{A}_{MAB}$ gives the feedback that no product is purchased. Then, for each $i \in S_\ell \cup \{0\}$,*

$$\mathcal{P}_{S_\ell}(i) = \frac{v_i}{v_0 + \sum_{j \in S_\ell} v_j}.$$

Let L be the total number of calls to $\mathcal{A}_{MNL}$ in $\mathcal{A}_{MAB}$. Intuitvely, after any call to $\mathcal{A}_{MNL}$ ("**Call** $\mathcal{A}_{MNL}$ " 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. Following claim makes this intuition precise.

CLAIM 7. *Let L be the total number of calls to $\mathcal{A}_{MNL}$ when $\mathcal{A}_{MAB}$ is executed for T time steps. Then,*

$$\mathcal{P}(T \geq \frac{L}{3}) \geq \Omega(1 - \frac{1}{T}).$$

Next, we use the properties of this construction to relate regret of $\mathcal{A}_{MAB}$ in T time steps to regret of $\mathcal{A}_{MNL}$ in L steps.

5.3. Relating regret of $\mathcal{A}_{MNL}$ and $\mathcal{A}_{MAB}$ to prove Theorem 5.1

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, $V(S^*) = K(\alpha + \epsilon) = K\mu_j$. The following lemmas relate regret of $\mathcal{A}_{MNL}$ to regret of $\mathcal{A}_{MAB}$ by bounding both in terms of $(\sum_\ell V(S^*) - V(S_\ell))$.

LEMMA 5.5. *Total expected regret of $\mathcal{A}_{MAB}$ on instance I_{MAB} in T time steps is upper bounded as*

$$Reg(\mathcal{A}_{MAB}\ , T) \leq \frac{1}{(1+\alpha)} \sum_{\ell=1}^{L} \frac{1}{K}(V(S^*) - V(S_\ell)).$$

LEMMA 5.6. *Total expected regret of $\mathcal{A}_{MNL}$ on instance I_{MNL} in L time steps is lower bounded as*

$$Reg(\mathcal{A}_{MNL}\ , L) \geq \frac{1}{(1+\alpha)} \sum_{\ell=1}^{L} \frac{1}{K}(V(S^*) - V(S_\ell)) - \frac{\epsilon v^* L}{(1+\alpha)^2}$$

REFERENCES

A. Aouad, R. Levi, and D. Segev. 2015. A Constant-Factor Approximation for Dynamic Assortment Planning Under the Multinomial Logit Model. *Available at SSRN* (2015).

P. Auer. 2003. Using Confidence Bounds for Exploitation-exploration Trade-offs. *J. Mach. Learn. Res.* (2003).

P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* (2002).

M. Babaioff, S. Dughmi, R. Kleinberg, and A. Slivkins. 2011. Dynamic Pricing with Limited Supply. *CoRR* (2011).

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

J.H. Blanchet, G. Gallego, and V. Goyal. 2013. A markov chain approximation to choice modeling. In *ACM Conference on Electronic Commerce (EC '13)*.

S. Bubeck and N. Cesa-Bianchi. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* (2012).

F. Caro and J. Gallien. 2007. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *Management Science* (2007).

J.M. Davis, G. Gallego, and H. Topaloglu. 2011. *Assortment optimization under variants of the nested logit model*. Technical Report. Technical report, Cornell University, School of Operations Research and Information Engineering.

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

A. Désir and V. Goyal. 2014. Near-Optimal Algorithms for Capacity Constrained Assortment Optimization. *Available at SSRN* (2014).

V.F. Farias, S. Jagabathula, and D. Shah. 2012. A Nonparametric Approach to Modeling Choice with Limited Data. *Management Science (To Appear)* (2012).

G. Gallego, R. Ratliff, and S. Shebalov. 2015. A General Attraction Model and Sales-Based Linear Program for Network Revenue Management Under Customer Choice. *Operations Research* (2015).

D. Golovin and A. Krause. 2012. Submodular Function Maximization. (2012).

E. Hazan and S. Kale. 2012. Online Submodular Minimization. *J. Mach. Learn. Res.* (2012).

R. Kleinberg, A. Slivkins, and E. Upfal. 2008. Multi-armed Bandits in Metric Spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing (STOC '08)*.

A.G. Kök and M.L. Fisher. 2007. Demand Estimation and Assortment Optimization Under Substitution: Methodology and Application. *Operations Research* (2007).
R.D. Luce. 1959. *Individual choice behavior: A theoretical analysis*. Wiley.
D. McFadden. 1973. Conditional logit analysis of qualitative choice behavior. *in P. Zarembka, ed., Frontiers in Econometrics* (1973).
D. McFadden. 1978. *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.
M. Mitzenmacher and E. Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*.
R. L. Plackett. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* (1975).
F. Radlinski, R. Kleinberg, and T. Joachims. 2008. Learning Diverse Rankings with Multi-armed Bandits. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*.
P. Rusmevichientong, Z. M. Shen, and D.B. Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* (2010).
D. Sauré and A. Zeevi. 2013. Optimal Dynamic Assortment Planning with Demand Learning. *Manufacturing & Service Operations Management* (2013).
M. Streeter and D. Golovin. 2009. An Online Algorithm for Maximizing Submodular Functions. In *Advances in Neural Information Processing Systems 21*.
K. Talluri and G. Van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* (2004).
K. Train. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.
B. Wierenga. 2008. Handbook of Marketing Decision Models, vol. 121 of International Series in Operations Research and Management Science. (2008).

A. MULTIPLICATIVE CHERNOFF BOUNDS

LEMMA A.1. *Let $\hat{v}_1, \cdots, \hat{v}_m$ be m be i.i.d random variables such that the moment generating function is given by*

$$E\left(e^{\theta \hat{v}_\ell}\right) = \frac{1}{1 - v(e^\theta - 1)}, \textit{for all } \theta < \log 2,$$

where $v \leq 1$. Let $\bar{v}_m = \dfrac{\sum_{\ell=1}^m \hat{v}_\ell}{m}$. Then, it follows that

$$\mathcal{P}\left(\bar{v}_m > 2v + a\right) \leq \exp\left(\frac{-m \cdot a}{3}\right).$$

PROOF.

$$\mathcal{P}\left(\bar{v}_m > 2v + a\right) = \mathcal{P}\left(\sum_{\ell=1}^m \hat{v}_\ell > 2m \cdot v + m \cdot a\right),$$

$$\mathcal{P}\left(\bar{v}_m > 2v + a\right) \leq \frac{E\left\{\exp\left(\theta \sum_{\ell=1}^m \hat{v}_\ell\right)\right\}}{e^{\theta(m \cdot v + m \cdot a)}},$$

$$= e^{-\theta m \cdot a}\left(\frac{E\left\{\exp\left(\theta \hat{v}_\ell\right)\right\}}{e^{2\theta \cdot v}}\right)^m.$$

The last equality follows the fact that $\hat{v}_\ell$ are i.i.d. Therefore,

$$\mathcal{P}\left(\bar{v}_m > 2v + a\right) \leq e^{-\theta m \cdot a} \frac{1}{\left[(1 - v(e^\theta - 1))e^{2\theta v}\right]^m}.$$

Let

$$f(\theta, v) = \log\left[(1 - v(e^\theta - 1))e^{2\theta v}\right].$$

Note that $f(\theta, v)$ is a concave function in $v \in [0,1]$ for all $\theta < \log 2$ and hence the minimum value of $f(\theta, v)$ occurs at a boundary point for all θ. In particular, we have

$$f(\theta, v) \geq \min\{0, \log(2e^{2\theta} - e^{3\theta})\}$$

Substituting $\theta = \log 3/2$, we get $f(\theta, v) \geq 0$. Therefore, it follows that

$$\mathcal{P}\left(\bar{v}_m > 2v + a\right) \leq e^{-(\log 3/2)m \cdot a} \leq \exp\left(\frac{-m \cdot a}{3}\right). \quad \square$$

We will use the following concentration inequality from [Mitzenmacher and Upfal 2005].

THEOREM A.2. *Consider n i.i.d random variables $X_1, \cdots, X_n$ with values in [0,1] and $EX_1 = \mu$. Then:*

$$Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right\} < 2e^{-\mu n\delta^2/3} \textit{ for any } \delta \in (0,1).$$

Theorem A.2 requires that the random variables be bounded, which is not the case with our estimate, $\hat{v}_{i,\tau}$. However, Corollary 2 established that our estimate is bounded by $8 \log T$ with high probability. Therefore, we can use a truncation technique to derive Chernoff bounds for our estimate. Define truncated random variables, $X_{i,\tau}$, $\tau \in \mathcal{T}_i(\ell)$,

$$X_{i,\tau} = \hat{v}_{i,\tau} \mathbb{I}\left(\hat{v}_{i,\tau} \leq 8 \log T\right) \text{ for all } \tau \in \mathcal{T}_i(\ell),$$

and let $\bar{X}_{i,\ell}$ be the sample mean of $X_{i,\tau}$, $\tau \in \mathcal{T}_i(\ell)$,

$$\bar{X}_{i,\ell} = \frac{1}{|\mathcal{T}_i(\ell)|} \sum_{\tau \in \mathcal{T}_i(\ell)}^{|\mathcal{T}_i(\ell)|} X_{i,\tau}$$

We have from Lemma 1 that the random variables $X_{i,\tau}$, $\tau \in \mathcal{T}_i(\ell)$ are independent and indentical in distribution. Now, we will adapt a non-standard corollary from [Babaioff et al. 2011] and [Kleinberg et al. 2008] to our estimates to obtain sharper bounds.

LEMMA A.3. $v_i - E(X_{i,\tau}) \leq \dfrac{6 \log^2 T}{T}$, *if* $T > 10$

PROOF. Define $Y_i = \hat{v}_{i,1} - X_{i,\tau}$. Note that $Y_i = \hat{v}_{i,1} \mathbb{I}(\hat{v}_{i,1} > 8 \log T)$ and hence

$$\begin{aligned} E(Y_i) &= \sum_{y=8\log T}^{\infty} yP(Y_i = y) \\ &\leq \sum_{y=8\log T}^{\infty} yP(Y_i \geq y) \\ &= \sum_{y=8\log T}^{\infty} yP(\hat{v}_{i,1} \geq y). \end{aligned}$$

Using Lemma 1 we can prove that for all $m \geq 1$,

$$\mathcal{P}\left(\hat{v}_{i,1} > 2^{m+2} \log T\right) \leq \frac{1}{T^{1+m}},$$

using Chernoff bound techniques as we did in Corollary 2. Bounding each term in the summation in interval

$$[2^m \cdot 8 \log T,\ 2^{m+1} \cdot 8 \log T]$$

by $2^{m+1} \cdot 8 \log T$, we have

$$E(Y_i) \leq 32 \frac{\log^2 T}{T^2} \sum_{m=1}^{\infty} \left(\frac{4}{T}\right)^m \leq 64 \frac{\log^2 T}{T^2} \leq 6 \frac{\log^2 T}{T}, \text{ if } T > 10. \quad \square$$

We will prove equivalent of Lemma 3 for the truncated variables.

LEMMA A.4. *Let $E(X_{i,\tau}) = \mu_i$. Then:*

(1) $\mathcal{P}\left(\left|\bar{X}_{i,\ell} - v_i\right| \geq \sqrt{\frac{12\bar{v}_{i,\ell}}{|\mathcal{T}_i(\ell)|}} \log T + \frac{30 \log^2 T}{|\mathcal{T}_i(\ell)|}\right) \leq \frac{4}{T^2}$ *for all* $i = 1, \cdots, n$.

(2) $\mathcal{P}\left(\left|\bar{X}_{i,\ell} - v_i\right| \geq \sqrt{\frac{6v_i}{|\mathcal{T}_i(\ell)|}} \log T + \frac{30 \log^2 T}{|\mathcal{T}_i(\ell)|}\right) \leq \frac{4}{T^2}$ *for all* $i = 1, \cdots, n$.

PROOF. Fix i, First assume $\mu_i \leq \frac{24 \log^2 T}{|\mathcal{T}_i(\ell)|}$. From Lemma A.3, we have

$$v_i \leq \mu_i + \frac{6 \log^2 T}{T} \leq \frac{30 \log^2 T}{|\mathcal{T}_i(\ell)|}$$

and hence, we have $\bar{X}_{i,\ell} - v_i \geq -30 \frac{\log^2 T}{|\mathcal{T}_i(\ell)|}$. Since $\bar{v}_{i,\ell} \geq \bar{X}_{i,\ell}$, we have,

$$\mathcal{P}\left(\bar{X}_{i,\ell} > v_i + \frac{30 \log^2 T}{T} + \frac{6 \log T}{|\mathcal{T}_i(\ell)|}\right) \leq \mathcal{P}\left(\bar{v}_{i,\ell} > 2v_i + \frac{6 \log T}{|\mathcal{T}_i(\ell)|}\right).$$

From Lemma A.1, we have $\mathcal{P}\left(\bar{v}_{i,\ell} > 2v_i + \frac{\log T}{|\mathcal{T}_i(\ell)|}\right) \leq \frac{1}{T^2}$. Hence, trivially we have $\mathcal{P}\left(\bar{v}_{i,\ell} > 2v_i + \frac{30 \log^2 T}{|\mathcal{T}_i(\ell)|}\right) \leq \frac{1}{T^2}$. Therefore it follows that,

$$\mathcal{P}\left(\left|\bar{X}_{i,\ell} - v_i\right| > \frac{30 \log^2 T}{|\mathcal{T}_i(\ell)|}\right) \leq \frac{1}{T^2}. \tag{19}$$

Now suppose $\mu_i \geq \frac{24 \log^2 T}{|\mathcal{T}_i(\ell)|}$, using Lemma A.2 with $\delta = \frac{1}{2}\sqrt{\frac{24 \log^2 T}{\mu_i |\mathcal{T}_i(\ell)|}}$, we have

$$\mathcal{P}\left(\left|\frac{\bar{X}_{i,\ell}}{\log T} - \frac{\mu_i}{\log T}\right| < \delta \frac{\mu_i}{\log T}\right) \geq 1 - 2\exp\left(\frac{-\mu_i |\mathcal{T}_i(\ell)| \delta^2}{3 \log T}\right) = 1 - \frac{2}{T^2}.$$

Substituting the value of δ, and noting that $v_i \geq \mu_i$, we have

$$\mathcal{P}\left(\left|\bar{X}_{i,\ell} - \mu_i\right| < \sqrt{\frac{6v_i \log^2 T}{|\mathcal{T}_i(\ell)|}}\right) \geq \mathcal{P}\left(\left|\bar{X}_{i,\ell} - \mu_i\right| < \sqrt{\frac{6\mu_i \log^2 T}{|\mathcal{T}_i(\ell)|}}\right) \geq 1 - \frac{2}{T^2}.$$

From Lemma A.3, we have,

$$\mathcal{P}\left(\left|\bar{X}_{i,\ell} - v_i\right| < \sqrt{\frac{6v_i \log^2 T}{|\mathcal{T}_i(\ell)|}} + 6\frac{\log T}{T}\right) \geq 1 - \frac{4}{T^2}. \tag{20}$$

By assumption, we have $\delta \leq \frac{1}{2}$ and hence $\mathcal{P}\left(2\bar{X}_{i,\ell} \geq \mu_i\right) \geq 1 - \frac{2}{T^2}$. Since $\bar{v}_{i,\ell} > \bar{X}_{i,\ell}$, we have,

$$\mathcal{P}\left(\left|\bar{X}_{i,\ell} - \mu_i\right| < \sqrt{\frac{12\bar{v}_{i,\ell}\log^2 T}{|\mathcal{T}_i(\ell)|}}\right) \geq \mathcal{P}\left(\left|\bar{X}_{i,\ell} - \mu_i\right| < \sqrt{\frac{12\bar{X}_{i,\ell}\log^2 T}{|\mathcal{T}_i(\ell)|}}\right) \geq 1 - \frac{4}{T^2}.$$

From Lemma A.3, we have,

$$\mathcal{P}\left(\left|\bar{X}_{i,\ell} - v_i\right| < \sqrt{\frac{12\bar{v}_{i,\ell}\log^2 T}{|\mathcal{T}_i(\ell)|}} + 6\frac{\log T}{T}\right) \geq 1 - \frac{4}{T^2}. \tag{21}$$

From (19), (20) and (21), we have the required result. □

We will break up the error on the estimate into two scenarios, one where $\hat{v}_{i,\tau}$ is bounded by $8\log T$ and other wise. In the first scenario, we will use Lemma A.4 to bound the error estimates and since the second scenario is a rare event, we have bounded the errors with high probability.

Proof of Lemma 3 Fix i. Define the events,

$$\mathcal{A}_{i,\ell} = \left\{|\bar{v}_{i,\ell} - v_i| > 4\sqrt{\frac{\bar{v}_{i,\ell}}{|\mathcal{T}_i(\ell)|}}\log T + \frac{4\log^2 T}{|\mathcal{T}_i(\ell)|}\right\}.$$

We will prove the result by showing $\mathcal{P}\left(\mathcal{A}_{i,\ell}\right)$is bounded by $\frac{4}{T^2}$.

Let $\mathcal{N}_{i,\ell}$ denote the event,

$$\mathcal{N}_{i,\ell} = \left\{\hat{v}_{i,\tau} > 8\log T \text{ for some } \tau = \{1, \cdots, |\mathcal{T}_i(\ell)|\}\right\}.$$

Note that the event $\mathcal{N}_{i,\ell}$ is an extremely low probability event. Whenever $\mathcal{N}^c_{i,\ell}$ is true, we have the estimate $\hat{v}_{i,\tau}$ bounded and can use multiplicative Chernoff Bounds to bound the difference between sample mean of the estimates $\hat{v}_{i,\tau}$ and v_i. Our proof will follow a similar approach, where we first show the probability of event $\mathcal{N}_{i,\ell}$ is $\mathcal{O}(\frac{1}{T^2})$ and then derive concentration bounds assuming $\mathcal{N}^c_{i,\ell}$ is true.

$$\begin{aligned}
\mathcal{P}\left(\mathcal{A}_{i,\ell}\right) &= \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}_{i,\ell}\right) + \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right), \\
&\leq \mathcal{P}\left(\mathcal{N}_{i,\ell}\right) + \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right), \\
&\leq \mathcal{P}\left(\bigcup_{\tau \in \mathcal{T}_i(\ell)} \{\hat{v}_{i,\tau} > 8\log T\}\right) + \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right) \qquad (22) \\
&\leq \sum_{\tau \in \mathcal{T}_i(\ell)} \frac{2}{T^3} + \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right) \leq \frac{2}{T^2} + \mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right).
\end{aligned}$$

The second inequality in (22) follows from the union bound and last inequality follows from Lemma 2. Observe that,

$$\mathcal{P}\left(\mathcal{A}_{i,\ell} \cap \mathcal{N}^c_{i,\ell}\right) \leq \mathcal{P}\left(\left|\frac{1}{|\mathcal{T}_i(\ell)|}\sum_{\ell=1}^{|\mathcal{T}_i(\ell)|} \hat{v}_{i,\tau}\mathbb{I}\left(\hat{v}_{i,\tau} \leq 8\log T\right) - v_i\right| > \sqrt{\frac{12\bar{v}_{i,\ell}}{|\mathcal{T}_i(\ell)|}}\log T + \frac{30\log^2 T}{|\mathcal{T}_i(\ell)|}\right), \tag{23}$$

where (23) follows from Lemma A.4. We can establish the second inequality in a similar manner. □

B. 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 B.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 expansion for $-\log 1-x$ by $x+2*x^2$ for $x=\frac{\epsilon}{1-\alpha}$, 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 B.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,

$$J_1 = \left\{i \,\middle|\, E_{P_0}(N_i) \leq \frac{3t}{N}\right\}, J_2 = \left\{i \,\middle|\, \mathcal{P}_0(a_t=i) \leq \frac{3}{N}\right\} \text{ and } J = J_1 \cap J_2. \tag{24}$$

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

$$|\mathcal{P}_j(a_t = j) - \mathcal{P}_0(a_t = j)| \leq \frac{1}{2}\sqrt{2\log 2 \cdot KL(P_0 \| P_j)}, \tag{25}$$

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 B.1 and (24), we have that Therefore,

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

Therefore,

$$\begin{aligned}
\mathcal{P}_j(a_t = j) &\leq \mathcal{P}_0(a_t = j) + \frac{1}{2}\sqrt{2\log 2 \cdot KL(P_0\|P_j)} \\
&\leq \frac{3}{N} + \frac{1}{2}\sqrt{(2\log 2)\frac{4\epsilon^2}{\alpha}E_{P_0}(N_j)} \\
&\leq \frac{3}{N} + \sqrt{2\log 2}\sqrt{\frac{3t\epsilon^2}{N\alpha}} \leq \frac{1}{2}.
\end{aligned} \tag{26}$$

The second inequality follows from (24), while the last inequality follows from the fact that $N > 12$ and $t \leq \frac{N\alpha}{60\epsilon^2}$ □.

Proof of Lemma 5.3 . 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,

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

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

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 $Regret \geq \frac{T\epsilon}{6}$. □

Proof of Lemma 5.5 Let us label the loop following the ℓth call to $\mathcal{A}_{MNL}$ in Algorithm 2 as ℓth loop. Then, we show that the total expected regret of $\mathcal{A}_{MAB}$ over the arm pulls in loop ℓ is

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

The lemma statement will then follow from substituting $V(S_\ell) \geq K\alpha$ and summing over $\ell = 1, \ldots, L$.

To see above, note that the probability of exiting the loop is $p = E[\frac{1}{2} + \frac{1}{2}\mu_{\mathcal{A}_t}] = \frac{1}{2} + \frac{1}{2K}V(S_\ell)$. 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 $\mathcal{A}_{MAB}$ 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 optimal reward, and substituting p, we obtain the above expression for expected regret over the arm pulls in a loop.

Proof of Lemma 5.6 Now we are ready to prove Theorem 5.1. From the previous two lemmas, we have

$$Reg(\mathcal{A}_{MAB}\ , T) \leq Reg(\mathcal{A}_{MNL}\ , L) + \frac{\epsilon v^* L}{(1+\alpha)^2}$$

Now for contradiction suppose that the regret of the $\mathcal{A}_{MNL}$, $Reg(\mathcal{A}_{MNL}\ , L) \leq c\sqrt{\frac{\hat{N}L}{K}}$ for a constant c to be prescribed in the following. Then, from above,

$$Reg(\mathcal{A}_{MAB}\ , T)\ \leq\ c\sqrt{\frac{\hat{N}L}{K}} + \frac{\epsilon v^* L}{(1+\alpha)^2}$$

From Claim 7, we have that $L \leq 3T$ with high probability. In what follows, for an easy understanding of the proof, we assume that the above event occur with probability 1. It is easy to derive a rigorous proof without this assumption, but we omit that here due to space constraints. Now, $c\sqrt{\frac{\hat{N}L}{K}} = c\sqrt{NL} \leq c\sqrt{3NT} = c\epsilon T\sqrt{\frac{3}{\alpha}} < \frac{\epsilon T}{12}$ on setting $c < \frac{1}{12}\sqrt{\frac{\alpha}{3}}$. Also, using $v^* = \alpha + \epsilon \leq 2\alpha$, and $L \leq 3T$, 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

$$Reg(\mathcal{A}_{MAB}\ , T) < \tfrac{\epsilon T}{12} + \tfrac{\epsilon T}{12} = \tfrac{\epsilon T}{6},$$

thus arriving at a contradiction. This proves that $Reg(\mathcal{A}_{MNL}\ , L) > c\sqrt{\frac{\hat{N}L}{K}}$ for a constant c.