The MNL-Bandit Problem: Theory and Applications

Vashist Avadhanula

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

The MNL-Bandit Problem: Theory and Applications

Vashist Avadhanula

One fundamental problem in revenue management that arises in many settings including retail and display-based advertising is assortment planning. Here, the focus is on understanding how consumers select from a large number of substitutable items and identifying the optimal offer set to maximize revenues. Typically, for tractability, we assume a model that captures consumer preferences and focus on computing the optimal offer set. A significant challenge here is the lack of knowledge on consumer preferences. In this thesis, we consider the multinomial logit choice model, the most popular model for this application domain and develop tractable robust algorithms for assortment planning under uncertainty. We also quantify the fundamental performance limits from both computational and information theoretic perspectives for such problems.

The existing methods for the dynamic problem follow "estimate, then optimize" paradigm, which require knowledge of certain parameters that are not readily available, thereby limiting their applicability in practice. We address this gap between theory and practice by developing new theoretical tools which will aid in designing algorithms that judiciously combine exploration and exploitation to maximize revenues. We first present an algorithm based on the principle of "optimism under uncertainty" that is simultaneously robust and adaptive to instance complexity. We then leverage this theory to develop a Thompson Sampling (TS) based framework with theoretical guarantees for the dynamic problem. This is primarily motivated by the growing popularity of TS approaches in practice due to their attractive empirical properties. We also indicate how to generalize the TS framework to design scalable dynamic learning algorithms for high-dimensional data and discuss empirical gains of such approaches from preliminary implementations on Flipkart, e-commerce firm in India.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

I owe a great deal of gratitude to my advisers, Shipra Agrawal, Vineet Goyal and Assaf Zeevi for supporting me in writing this thesis. They have been a constant source of motivation and I consider myself incredibly lucky to be advised by the three of them. Their unique perspectives have not only accelerated my learning process, but also helped me in identifying interesting research problems which eventually resulted in this thesis. I have immensely benefited from my interactions with all my advisers, which helped me become the researcher I am today. I am still trying (albeit unsuccessfully) to emulate Assaf's writing style, who over the years has taught me the importance of having clarity of thought in describing and identifying fundamental research questions. I often identify key contributions of a research paper by asking myself what would Vineet ask me if I describe the paper to him. My interactions with Shipra has instilled in me the belief that the solution is always around the corner if we persist long enough.

I also consider myself lucky to be a graduate student in the DRO division, where I have immensely benefited from the support structures in place. The feedback from Brown bag seminars and regular interactions with faculty and other students have contributed to my over all well being. I would particularly like to thank Omar Besbes, Costis Maglaras and Yash Kanoria for their support over the years. I still remember Omar's words of encouragement and advice after my disastrous practice job talk. I am glad that 5 years ago Yash pushed me to join DRO, which not only led me to spend time in a department with brilliant people, but also gave me an opportunity

to spend time in New York City. I would also like to thank him for his constant encouragement and advice during the last 5 years. I have benefited immensely from my interactions with Costis on academic and non-academic issues.

I am extremely thankful for the great cohort of students I've had a chance to interact with. I have particularly benefitted from my interactions with Chun and Antoine, both of whom I have looked up to at various moments during my graduate studies. I would also like to single out Hamsa for being a great friend, her help in networking and more importantly for all the advice and encouragement regarding my job market and research. A special mention to my room-mate and batch-mate, Gowtham, whom I've known for 12 years, for putting up with my ramblings about research and seemingly irrelevant issues. Another special shout out to Yaarit for the fun banter and the occasional debates that helped me see alternative view points. I am particularly thankful for the advice and encouragement during my stressful times and also for letting me be the cool uncle to the kids. I would also like to thank Francisco, Zhe and Fei whose interactions in the office have made my grad school life fun.

Lastly, I am thankful for the support I received from my friends and family throughout my grad school. I would like to acknowledge efforts of Deeksha and Keerti in patiently teaching me Python and Hive, which has helped me significantly in my research. I would like to thank my friends Ayesha, Kavya, Pradeep and my sister Srivalli for putting up with my idiosyncrasies and helped me stay sane during the grad school.

To my parents - Padmavathi Suvarna & Suresh Avadhanula and my grandmother Suvarna Amba who have sacrificed a great deal to get me to where I stand today.

Introduction

The explosive growth of e-commerce firms has brought renewed attention to the field of revenue management. Many e-commerce firms including Amazon and Flipkart operate in a low margin and high volume environment, where even small percentage increase in revenues can translate to substantial profits. Consequently, a central focus of many business applications is the development of predictive models that capture consumer behavior to maximize revenue growth. A significant challenge here is the lack of knowledge on consumer preferences which is further exacerbated by short selling seasons and evolving demand trends. This necessitates a balanced exploration-exploitation approach, where we not only have to learn demand trends, but also simultaneously exploit the information gain. The unprecedented flexibility in operational decisions associated with modern e-commerce systems not only provides the possibility but also makes it essential to optimize decision making with evolving and uncertain consumer tastes. My dissertation expands the scope of revenue management systems by designing tractable robust algorithms to optimize sequential decision making under uncertainty for assortment planning, which is a key component in many revenue management applications.

The first chapter of this dissertation provides an overview of assortment planning and the multinomial logit model (MNL), which is the most popular predictive model for this application domain. In this chapter, we also present an overview of existing approaches for assortment planning under uncertainty and highlight shortcomings that limit their applicability in practice, thereby motivating a need for more adaptive

approaches. In Chapter 2, we address this gap between theory and practice by developing new theoretical tools to design an algorithm based on the principle that is simultaneously robust and adaptive to instance complexity. In Chapter 3, we leverage the theory developed in Chapter 2 to design a Thompson Sampling (TS) based framework with theoretical guarantees for the dynamic problem. This is primarily motivated by the growing popularity of TS approaches in practice due to their attractive empirical properties. In Chapter 4, we indicate how to generalize the TS framework to design scalable dynamic learning algorithms for high-dimensional data and discuss empirical gains of such approaches from preliminary implementations on Flipkart, a large e-commerce firm in India. In Chapter 5, we present tractable algorithms for static assortment planning with constraints under the MNL and more general Nested Logit choice models. This is studied as a first step to developing dynamic assortment planning approaches for more general predictive models.

MNL-Bandit Problem. One fundamental problem in revenue management that arises in many settings including retail and display-based advertising is assortment planning. Here, the focus is on understanding how consumers select from a large number of substitutable items and identifying the optimal offer set to maximize revenues. Typically, for tractability, we assume a model that captures consumer preferences and focus on computing the optimal offer set. However, model selection and estimating the parameters is a challenging problem. In many e-commerce settings such as fast fashion retail, products have short selling seasons. Therefore, the data on consumer choices is either limited or non existent. The retailer needs to learn consumer preferences by offering different assortments and observing purchase decisions, but short selling seasons limit the extent of experimentation. There is a natural trade-off in these settings, where the retailer need to learn consumer preferences and also maximize cumulative revenues simultaneously. Finding the right balance between exploration and exploitation is a challenge.

In Chapter 1, we consider the MNL model which is the most popular model for this application domain and formulate the dynamic learning problem in the framework of multi-armed bandits (MAB). More specifically, we formulate the dynamic problem as a parametric bandit problem, which we will refer to as the MNL-Bandit problem. Though it is common practice to study dynamic problems under the MAB framework, the combinatorial complexity involved with identifying the ideal subset (assortment) presents many theoretical and computational challenges. We discuss these challenges in detail along with a review of the existing methods for the MNL-Bandit problem which typically make restrictive assumptions, severely limiting their applicability in practice.

UCB Approach for the MNL-Bandit. Motivated by the apparent need for a tractable policy, in Chapter 2, we develop an efficient algorithm that judiciously combines exploration of the combinatorial option space and exploitation of that information to maximize revenues. The key idea in our work is a novel estimation technique using sampling, where the samples directly give us unbiased estimates of the model parameters. We use these estimates to leverage the structure of the MNL model, and to adapt the upper confidence bounds (UCB) policy, a popular bandit technique, to our problem. Our sampling technique plays an essential role in avoiding the shortcomings of standard estimation approaches like maximum likelihood, where the estimates are obtained by optimizing a loss function. The convergence bounds for estimates resulting from such approaches typically depend on true parameters, which becomes an impediment in real time implementation. In contrast, our approach which obtains estimates through sampling is completely independent of model parameters. Furthermore, we show that our algorithm's performance is near-optimal as well as adaptive to the complexity of the instance.

Thompson Sampling for the MNL-Bandit. The UCB based approach developed in Chapter 2, focuses on robustness and tends to experiment more than necessary.

To that end, several stream of recent papers observed that Thompson Sampling (TS) significantly outperform more traditional approaches such as UCB policies. For standard MAB problems, despite being easy to implement, TS-based algorithms are hard to analyze and theoretical work on TS is limited. Furthermore, the selection of prior, efficient posterior computation and theoretical analysis remains particularly challenging for parametric bandit settings, where arms are related through a small number of parameters. Motivated by the growing popularity of TS in practice, In Chapter 3, we leverage the sampling technique to present an approach to adapt Thompson Sampling to this problem and show that it achieves near-optimal regret as well as attractive numerical performance. A key ingredient in our approach is a two moment approximation of the posterior and the ability to judicially correlate samples, which is done by embedding this approximation in a normal family.

Thompson Sampling in Practice: Evidence from Flipkart.com. In Chapter 4, we present evidence of empirical gains from employing dynamic assortment planning in optimizing product recommendations on Flipkart, an Indian ecommerce firm. First, we show that choice models like MNL which capture consumer preferences over an assortment have higher predictive power than traditional models which consider each item independently. We will then present empirical evidence to show that firms stand to gain by implementing dynamic learning algorithms instead of the traditional "estimate, then optimize" approaches. In settings like Flipkart, we have a large number of alternatives that are effectively described by a small number of attributes; via what is often referred to as a factor model. The possibility of different items being related to each other only through their attributes raises the possibility that one can design algorithms whose performance is independent of the number of items, which is a major source of complexity. Using the analysis developed in Chapter 3 as a foundation, we present a framework that indicates how to extend our aforementioned TS-based policy to the problem of learning in the attribute space. Specifically, how

to leverage the relation between different items through attributes and achieves a regret bound which is independent of the number of items, and only depends on the number of attributes, thereby accelerating the learning.

Static Assortment Planning. Noting that an important ingredient for dynamic learning is a computationally efficient policy for static optimization, i.e., computing the optimal set of items to offer when the model parameters are known. In Chapter 5, we consider the MNL model and it's generalized version, the nested logit model (NL) and present polynomial time algorithms for computing the optimal assortment under a large class of constraints.

Chapter 1

The MNL-Bandit Problem

1.1 Assortment Optimization

In many settings, a decision maker is faced with the problem of identifying the optimal mix of items, often from a large feasible set to offer to users. For example, a retailer needs to select a subset (assortment) of products and due to substitution effects, the demand for an individual product is influenced by the assortment of products presented to the consumer. In display-based online advertising, a publisher needs to select a set of advertisements to display to users and due to competition between ads, the click rates for an individual ad depends on the overall subset of ads displayed. A recommender system like the one used by Netflix or Amazon, must determine a subset of items to suggest to users from a large pool of similar alternatives. In all these settings, items may be valued differently from the decision maker's perspective and therefore the assortment of items offered to users have significant impact on revenues. To identify the ideal offer set, the decision maker must understand the substitution patterns of users.

Choice models capture these substitution effects among items by specifying the probability that a user selects an item given the offered set. More specifically, let $\mathcal{N} = \{1, \cdots, N\}$ be the set of available items for the decision maker to offer for consumers. For any subset $S \subset \mathcal{N}$ and any item $i \in S$, a choice model describes the probability of a random consumer preferring item i in the set S as,

$$\pi(i, S) = \Pr(\text{customer selects item } i \text{ from offer set } S).$$

We refer to $\pi(i, S)$ as choice probabilities. Using these choice probabilities, one can compute the expected revenue associated with an offer set as the weighted sum of revenues of items in the offer set and the choice probabilities. More specifically, if value (revenue) corresponding to every item $i \in \mathcal{N}$ is r_i, then the expected revenue $R(S)$ of any assortment $S \subset \mathcal{N}$ can be written as

$$R(S) = \sum_{i \in S} r_i \cdot \pi(i, S).$$

Then the decision maker can identify the optimal set by computing the set with highest revenues, an optimization problem commonly referred to as assortment optimization problem and formulated as

$$\max_{S \subseteq \mathcal{N}} R(S). \tag{1.1}$$

Assortment optimization problems also allow for constraints that arise in practice, e.g. budget for inventory, product purchasing, display capacity, etc.

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 [27] for a detailed review. In this dissertation, 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 thesis.

1.1.1 Multinomial Logit Choice Model (MNL)

A fundamental problem in assortment planning is model selection. There is a trade-off between working with models that have greater predictive power and simple models that allow greater tractability. Estimating choice probabilities involving large number of alternatives from transactional data is a highly non-trivial task. Furthermore, note that the assortment optimization problem is a combinatorial optimization problem for which trying all 2^n possible assortment is not a scalable solution. Though theoretically choice models with higher predictive power could result in better assortment solutions, lack of tractable optimization approaches for these problems could make them less interesting for decision makers. The trade-offs between the predictability and the tractability of a choice model is an important consideration for the decision maker in its deployment, particularly in settings where one needs to constantly estimate and optimize the model. In this dissertation, we consider the multinomial logit choice model (MNL) for which the tractability is well understood and develop efficient approaches that learns consumer preferences while simultaneously maximizing revenues. The dynamic learning algorithms developed in this thesis for the MNL model should be viewed as a first step towards efficient algorithms for more general choice models.

MNL was introduced independently by Luce [31] and Plackett [38]. In his seminal work, McFadden [33] showed that the multinomial logit model is part of a larger class of models that can be modeled within the random utility frame work. In the random utility framework, it is assumed that consumers have inherent (random) utility associated with every item and upon presenting an offer set consumers select the item with the highest utility. In the MNL model, the consumer's random utilities are modeled as independent Gumbel random variables. In particular, the utility of item i is given by:

$$U_i = \mu_i + \xi_i,$$

where $\mu_i \in R$ denotes the mean utility that a 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 and represent the idiosyncrasies in consumer population. The choice probabilities for the MNL can be computed in closed form as $\pi_{\mathsf{MNL}}(i, S) = \frac{e^{\mu_i}}{\sum_{j \in S} e^{\mu_j}}$,

In assortment planning problems, consumers always have the option of not choosing any item from the offered set. Such scenarios are modeled by augmenting the available set of items with a further index, 0 that indicates an "outside" option. Consumers purchase some thing from the offered set if the random utility of one of the offered items is more than the random utility corresponding to the outside option. Therefore, the choice probabilities can be written as $\pi_{\mathsf{MNL}}(i, S) = \frac{v_i}{v_0 + \sum_{j \in S} v_j}$ where we denote e^{μ_i} by v_i denotes for notational brevity. We can also without loss of generality assume that $v_0 = 1$ by scaling every other parameter. Hence, the choice probabilities for the MNL model can be reformulated as:

$$\pi_{\mathsf{MNL}}(i, S) = \frac{v_i}{1 + \sum_{j \in S} v_j}, \tag{1.2}$$

and the expected revenue for any assortment S is given by

$$\mathbb{R}(S, \mathbf{v}) = \sum_{i \in S} r_i \frac{v_i}{1 + \sum_{j \in S} v_j}. \tag{1.3}$$

From the choice probabilities we can see that the ratio of choice probabilities of two items, $\pi_{\mathsf{MNL}}(i, S)$ and $\pi_{\mathsf{MNL}}(j, S)$ is independent of the offer set S. This property is known as the independent of irrelevant attributes (IIA) property [8] and is a limitation of the MNL model. Other random utility based choice models like Nested Logit (NL) [47] and Mixed Logit model (mMNL) [34] generalize the MNL model and are not restricted by the IIA property. However, estimation of these models and the corresponding assortment planning problems involved are often intractable highlighting the challenges involved in model selection. See [20] for further discussion on tractability of choice models. The closed form expression of the choice probabilities

make this model extremely tractable from estimation and optimization point of view (see [44].) The tractability of the model in decision making is the primary reason MNL has been extensively used in practice ([25, 8, 46]).

1.2 Dynamic Learning in Assortment Selection

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 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.

Motivated by aforementioned applications, 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. The objective is to design a policy that selects a sequence of history dependent assortments $(S_1, S_2, \ldots, S_T)$ so as to maximize the cumulative expected revenue,

$$\mathbb{E}_\pi \left(\sum_{t=1}^{T} R(S_t, \mathbf{v}) \right), \tag{1.4}$$

where $R(S, \mathbf{v})$ is the revenue corresponding to assortment S as defined as in (1.3). Direct analysis of (1.4) is not tractable given that the parameters $\{v_i, i = 1, \ldots, N\}$ are not known to the seller a priori. Instead we propose to measure the performance of a policy via its *regret*. The objective then is to design a policy that approximately minimizes the *regret* defined as

$$\mathsf{Reg}(T, \mathbf{v}) = \sum_{t=1}^{T} R(S^*, \mathbf{v}) - \mathbb{E}_\pi[R(S_t, \mathbf{v})], \qquad \text{(MNL-Bandit)}$$

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

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 $S \subseteq \{1, \ldots, N\}$, i.e., $x_i(S) = 1$ if product $i \in S$ and 0 otherwise. We consider constraints of the form

$$\mathcal{S} = \{S \subseteq \{1, \ldots, N\} \mid A\, \mathbf{x}(S) \leq \mathbf{b},\ \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}\}, \qquad (1.5)$$

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 [17] for a detailed discussion on assortment and pricing optimization problems that can be formulated under the TU constraints.

1.2.1 Existing Approaches for the MNL-Bandit

The problem of dynamic learning under the MNL choice model has been studied in the literature. [40] and [43] 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," [43] show an asymptotic $O(N \log T)$ regret bound, while [40] 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.

1.3 Summary of contributions of Chapters 2, 3, 4 and 5

We summarize the main contributions of Chapters 2, 3, 4 and 5. The primary contribution of this dissertation is to develop a systematic approach, and supporting theory, for the MNL-Bandit problem. In Chapter 2, we present an efficient learning algorithm that does not require any parameter information and has near-optimal performance. The algorithm is predicated on the upper bound (UCB) type logic, originally developed to balance the aforementioned trade-off in the multi armed bandit

(MAB) problem (c.f. [28]). The UCB based algorithm is easy to analyze theoretically and helps in developing theoretical tools that will aid in designing more efficient learning based algorithms for the MNL-Bandit problem. In Chapter 3, we leverage the theory developed in Chapter 2 to design a Thompson Sampling (TS) based framework with theoretical guarantees for the dynamic problem. This is primarily motivated by the growing popularity of TS approaches in practice due to their attractive empirical properties.

In Chapter 4, we indicate how to generalize the TS framework to design scalable dynamic learning algorithms for high-dimensional data and discuss empirical gains of such approaches from preliminary implementations on Flipkart, a large e-commerce firm in India. In Chapter 5, we present tractable algorithms for static assortment planning with constraints under the MNL and more general Nested Logit choice models. This is studied as a first step to develop dynamic assortment planning approaches for more general predictive models.

1.3.1 UCB Approach for the MNL-Bandit

In this chapter we present 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.

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 [43] and [40] 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 [17]). 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. [28]).

We also 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 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 [15] 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.

1.3.2 Thompson Sampling for the MNL-Bandit

In this chapter, relying on structural properties of the MNL model and theoretical tools developed in Chapter 2, we design a TS approach that is computationally efficient and yet achieves parameter independent (optimal in order) regret bounds. Specifically, we present a computationally efficient TS algorithm for the MNL-Bandit which uses a prior distribution on the parameters of the MNL model such that the posterior update under the MNL-bandit feedback is tractable. A key ingredient in our approach is a two moment approximation of the posterior and the ability to judicially correlate samples, which is done by embedding the two-moment approximation in a

normal family. We show that our algorithm achieves a worst-case (prior-free) regret bound of $O(\sqrt{NT}\log TK)$ under a mild assumption that $v_0 \geq v_i$ for all i (more on the practicality of this assumption later in the text); the bound is non-asymptotic, the "big oh" notation is used for brevity. This regret bound is independent of the parameters of the MNL choice model and hence holds uniformly over all problem instances. The regret is comparable to the existing upper bound of $O(\sqrt{NT})$ achieved by the UCB approach in Chapter 2, yet the numerical results demonstrate that our Thompson Sampling based approach significantly outperforms the UCB-based approach. The methods developed in this paper highlight some of the key challenges involved in adapting the TS approach to the MNL-Bandit, and present a blueprint to address these issues that we hope will be more broadly applicable, and form the basis for further work in the intersection of combinatorial optimization and machine learning.

1.3.3 Empirical Evaluation of Thompson Sampling

In Chapter 4, we present evidence of empirical gains from employing dynamic assortment planning in optimizing product recommendations on Flipkart, an Indian ecommerce firm. First, we show that choice models like MNL which capture consumer preferences over an assortment have higher predictive power than traditional models which consider each item independently. In particular, we consider a structured MNL model, where every item is described by a set of attributes and the mean utility of a product is linear in the values of attributes and show that the fit of this stylized MNL model is better than a simple logistic regression with the same set of attributes, which is the current model used at Flipkart. We will then present empirical evidence to show that firms stand to gain by implementing dynamic learning algorithms instead of the traditional "estimate, then optimize" approaches.

1.3.4 Static Assortment Optimization

In Chapter 5, we consider settings when the model parameters are known and focus on developing tractable optimization algorithms for the MNL and the NL model under totally unimodular constraint structures. The totally unimodular constraints model a rich class of practical assortment planning problems including cardinality constraints, partition matroid constraints and joint display and assortment constraints. We refer the reader to [17] for a detailed discussion on assortment and pricing optimization problems that can be formulated under the TU constraints.

First we consider the assortment planning problem under the MNL model and show that a natural linear programming (LP) relaxation is tight. The LP based approach provides robustness to handle capacity constraints in addition to the existing TU constraints. In particular, we consider an arbitrary additional constraint to the set of TU constraints such that the resulting set of constraints are not TU. We present a Polynomial Time Approximation Scheme (PTAS) for the assortment optimization problem under this more general set of constraints where for any $0 < \epsilon < 1$, we obtain a solution with objective value at least $(1 - \epsilon)$ times the optimal in running time polynomial in the input size for a fixed ϵ. As a consequence of this problem, we obtain PTAS for joint display and assortment optimization problem with an additional capacity constraint.

We then consider the assortment optimization problem under NL model with TU constraints and provide a Fully Polynomial Time Approximation Scheme (FPTAS) for this problem, where for any $0 < \epsilon < 1$, we obtain a solution with objective value at least $(1 - \epsilon)$ times the optimal in running time polynomial in the input size and $1/\epsilon$. We also show that the exact assortment optimization under NL model with TU constraints is NP-hard. For the joint display and assortment optimization problem, we show that under special settings the problem allows for an exact solution in polynomial time.

Summary. In Chapters 2, 3 and 4, we focus on designing efficient algorithms for assortment planning under the most popular choice model. In the final chapter, we work on developing tractable optimization approaches for general choice models with the hope that these approaches are a first step in designing dynamic learning approaches for these choice models in the future.

Chapter 2

A UCB Approach for the MNL-Bandit

In this Chapter, we describe our proposed policy for the MNL-Bandit problem. 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. [28]). A key idea in our algorithm is a novel estimation technique using sampling, where the samples directly give us unbiased estimates of the model parameters. We use these estimates to leverage the structure of the MNL model, and to adapt the UCB policy to our problem. The estimation technique also plays a key role in designing a tractable Thompson Sampling algorithm in Chapter 3.

We first present in Section 2.1, an efficient online algorithm that judiciously balances the exploration and exploitation trade-off intrinsic to our problem. Subsequently, in Section 2.2 show that this algorithm 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. In Section 2.5, we relax the assumption on "no-purchase" and give a learning algorithm that is independent of problem parameters and bound its regret.

In Section 2.3, we show that for "well separated" instances, the regret of our policy is bounded by 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 [43] and [40] respectively, even though we do not require any prior information on Δ unlike the aforementioned work.

In Section 2.4, 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 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 [15] 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.

Finally in Section 2.6, 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 [43] "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 [43]. We observe that the performance of our algorithm is sub-linear, while the performance of [43] is linear, which further emphasizes the limitations of "explore first and exploit later" approaches and highlights the universal applicability of our approach.

2.1 UCB Algorithm

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. Before introducing our algorithm, we present a quick background of the UCB family of algorithms [6] for the classic multi-armed bandit (MAB) problem.

2.1.1 Revisiting UCB for MAB

In the classical MAB problem, there are n arms and a finite time horizon T. The reward obtained upon playing arm i at time t is r_{it}, generated i.i.d (across time) from a distribution $\mathcal{F}_i$ with fixed but unknown mean, μ_i. The objective here is to play arms in an online fashion in order to maximize the cumulative reward or equivalently minimize the regret which is defined as

$$\mathsf{Reg}_{\mathsf{MAB}}(T) = \sum_{t=1}^{T}(\mu_{i_*} - r_t),$$

where $i_* = \arg\max_i \mu_i$ and r_t is the reward corresponding to the arm played at time t. Maximizing cumulative rewards, as with any bandit problems involves experimenting with various arms to learn these unknown means while simultaneously trying to play the "best arm" as many times as possible. UCB algorithm provide a structured framework to judiciously balance the friction between exploration and exploitation for the MAB problems. As the name suggests, the basic idea of the UCB framework is to use the observations from the past plays of each arm to construct estimates $\mathsf{UCB}_{t,i}$ that are "upper confidence bounds" of the true rewards. In particular, the estimates $\mathsf{UCB}_{t,i}$ satisfy the following two key properties.

1. $\mathsf{UCB}_{t.i}$ for every arm is larger than its mean reward with high probability,

$$\mathsf{UCB}_{t,i} \geq \mu_i, \ \forall \ i.t$$

2. As the arm i is played more and more, the estimate $\mathsf{UCB}_{t,i}$ converges to the true mean with high probability,

$$|\mathsf{UCB}_{t,i} - r_{i,t}| \lessapprox \frac{1}{\sqrt{T_i(t)}},$$

where $T_i(t)$ is the number of times arm i is played till time t. The UCB algorithm plays the best arm according to the estimates $\mathsf{UCB}_{t,i}$ and by virtue of first property, we always have the estimate of the current arm higher than the optimal mean, i.e. $\mathsf{UCB}_{t,i_t} \geq \mu_{i_*}$, where i_t is the arm played at time t. Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T} \mu_{i_*} - r_t \leq \sum_{t=1}^{T} \mathsf{UCB}_{t,i_t} - r_t &= \sum_{t=1}^{T}\sum_{i=1}^{n} (\mathsf{UCB}_{t,i} - r_{i,t}) \cdot \mathbb{1}(i_t = i) \\ &\lessapprox \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{1}{\sqrt{T_i(t)}}.\end{aligned}$$

We have $\sum_{t=1}^{T}\sum_{i=1}^{n} \frac{1}{\sqrt{T_i(t)}} \leq \sum_{i=1}^{n} \sqrt{T_i}$, where T_i is the total number of times arm i is played. Noting that $\sum_{i=1}^{n} T_i = T$, by Cauchy Schwartz we have $\sum_i \sqrt{T_i} \leq \sqrt{nT}$.

In our UCB based algorithm, we use this same basic idea for algorithm design and regret analysis. However, the combinatorial nature of the problem brings in new challenges which we elaborate in the following section.

2.1.2 Challenges and overview

A key difficulty in applying standard multi-armed bandit techniques to the MNL-Bandit problem is that the response observed on offering an item i is *not* independent of other items in assortment S. Therefore, the N items 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 feasible assortments as an arm, leading to a computationally inefficient algorithm with exponential regret. 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 [6]. It uses the past observations to maintain increasingly accurate upper confidence bounds for the 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, our policy picks the assortment S with the highest optimistic revenue. There are two main challenges in implementing this scheme. First, the user response to being offered an assortment S depends on the entire set S, and does not directly provide an unbiased sample of demand for an item $i \in S$. 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 an item i is selected 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.

Remark 2.1. (Related UCB Approaches) Popular extensions of UCB for large scale problems include the linear bandit (e.g., 5, 39) and generalized linear bandit (22) problems. However, these do not apply directly to our problem, since the revenue corresponding to an assortment is nonlinear in problem parameters. Other works (see 14) 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 user choices depend on the assortment of

items offered in a time step.

2.1.3 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, 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,

$$\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\tau\}\,, \quad T_i(\ell) = |\mathcal{T}_i(\ell)|. \tag{2.1}$$

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

$$\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}. \tag{2.2}$$

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 2.1) Using these estimates, we compute the upper confidence

bounds, $v_{i,\ell}^{\mathsf{UCB}}$ for v_i as,

$$v_{i,\ell}^{\mathsf{UCB}} := \bar{v}_{i,\ell} + \sqrt{\bar{v}_{i,\ell}\frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}. \tag{2.3}$$

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 2.2). 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,

$$S_{\ell+1} := \operatorname*{argmax}_{S\in\mathcal{S}} \max\left\{R(S,\hat{\mathbf{v}}) : \hat{v}_i \leq v_{i,\ell}^{\mathsf{UCB}}\right\}, \tag{2.4}$$

where $R(S,\hat{\mathbf{v}})$ is as defined in (1.3). We later show that the above optimization problem is equivalent to the following optimization problem (see Lemma 2.3).

$$S_{\ell+1} := \operatorname*{argmax}_{S\in\mathcal{S}} \tilde{R}_{\ell+1}(S), \tag{2.5}$$

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

$$\tilde{R}_{\ell+1}(S) := \frac{\sum_{i\in S} r_i v_{i,\ell}^{\mathsf{UCB}}}{1+\sum_{j\in S} v_{j,\ell}^{\mathsf{UCB}}}. \tag{2.6}$$

We summarize the steps in our policy in Algorithm 1. Finally, we may remark on the computational complexity of implementing (2.4). The optimization problem (2.4) is formulated as a static assortment optimization problem under the MNL model with TU constraints, with model parameters being $v_{i,\ell}^{\mathsf{UCB}}, i = 1,\ldots,N$ (see (2.5)). There are efficient polynomial time algorithms to solve the static assortment optimization problem under MNL model with known parameters (see [17, 40]). We will now briefly comment on how Algorithm 1 is different from the existing approaches of [43] and [40] and also why other standard "bandit techniques" are not applicable to the MNL-Bandit problem.

Algorithm 1 Exploration-Exploitation algorithm for MNL-Bandit

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

while $t < T$ **do**

  Compute $S_\ell = \underset{S \in \mathcal{S}}{\operatorname{argmax}}\, \tilde{R}_\ell(S) = \dfrac{\sum_{i \in S} r_i v_{i,\ell-1}^{\mathsf{UCB}}}{1 + \sum_{j \in S} v_{j,\ell-1}^{\mathsf{UCB}}}$

  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 $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until ℓ that offered product i

    update $\bar{v}_{i,\ell} = \dfrac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}$, 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

Remark 2.2. (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 [43] and [40], 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 2.3. (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 [11]), 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 (2.4), (2.5) and Step-4 of Algorithm 1), which is less accessible for the MLE estimate.

2.2 Worst Case Regret Bounds

In what follows, we make the following assumptions.

Assumption 2.1.

1. *The MNL parameter corresponding to any product* $i \in \{1, \ldots, N\}$ *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 2.5, 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* $\mathbf{v} = (v_0, \ldots, v_N)$ *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,*

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

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

2.2.1 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 (2.3) 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 [6]. We provide the precise statements for the above mentioned properties in Lemma 2.2. 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 (2.6) 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 2.4 provides the precise statement. Lemma 2.4 is akin to the first property in Section 2.1.1.

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 2.5 provides the precise statements of above properties.

In the rest of this section, we make the above notions precise. Note that Lemma 2.4 and Lemma 2.5 are similar in spirit to first and second properties of the UCB estimates $\mathsf{UCB}_{t,i}$ discussed in Section 2.1.1. Therefore, the proof of Theorem 1 follows a similar analysis. However, the combinatorial aspects of the assortment optimization problem brings in additional challenges in completing the proof. In the interest of continuity, we defer the proof of Theorem 1 to Appendix B.1.

2.2.2 Unbiased estimates

Here, we prove that $\hat{v}_{i,\ell}$ and $\bar{v}_{i,\ell}$ are unbiased estimates of the MNL parameter v_i. We show that from the moment generating function of the estimate $\hat{v}_{i,\ell}$

Lemma 2.1 (Moment Generating Function)**.** *The moment generating function of estimate conditioned on S_ℓ, $\hat{v}_i$, is given by,*

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

Proof. From (1.2), we have that probability of no purchase event when assortment S_ℓ is offered is given by

$$\pi(0, S_\ell) = \frac{1}{1 + \sum_{j \in S_\ell} v_j}.$$

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 $\pi(0, S_\ell)$. 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 $\pi(0, S_\ell)$ and $q_i(S_\ell)$. Hence, we have

$$\mathbb{E}\left(e^{\theta \hat{v}_{i,\ell}}\right) = E_{n_\ell}\left\{\mathbb{E}\left(e^{\theta \hat{v}_{i,\ell}} \,\middle|\, n_\ell\right)\right\}. \tag{2.7}$$

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

$$\mathbb{E}\left(e^{\theta \hat{v}_{i,\ell}} \,\middle|\, n_\ell\right) = \mathbb{E}_{n_\ell}\left\{\left(q_i e^\theta + 1 - q_i\right)^{n_\ell}\right\}. \tag{2.8}$$

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,

$$\mathbb{E}_{n_\ell}\left\{\left(q_i e^\theta + 1 - q_i\right)^{n_\ell}\right\} = \frac{p_0}{1 - \left(q_i e^\theta + 1 - q_i\right)(1 - p_0)} = \frac{1}{1 - v_i(e^\theta - 1)}. \tag{2.9}$$

The result follows from (2.7), (2.8) and (2.9). □

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 2.1, we have the following result.

Corollary 2.1 (Unbiased Estimates)**.** *We have the following results.*

1. *$\hat{v}_{i,\ell},\ \ell \leq L$ are i.i.d geometrical random variables with parameter $\frac{1}{1+v_i}$, i .e. for any ℓ, i*
$$\mathsf{Pr}\left(\hat{v}_{i,\ell} = m\right) = \left(\frac{v_i}{1+v_i}\right)^m \left(\frac{1}{1+v_i}\right) \quad \forall\ m = \{0, 1, 2, \cdots\}.$$
2. *$\hat{v}_{i,\ell},\ \ell \leq L$ are unbiased i.i.d estimates of v_i, i .e. $\mathbb{E}\left(\hat{v}_{i,\ell}\right) = v_i\ \forall\ \ell, i$.*

2.2.3 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 2.2. *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*
$$v_{i,\ell}^{\mathsf{UCB}} - v_i \leq C_1\sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}} + C_2\frac{\log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)},$$
with probability at least $1 - \frac{7}{N\ell}$.

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 2.1, 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 [36] and [7]) to geometric random variables. The proof is provided in Appendix A.2

2.2.4 Optimistic estimate and convergence rates

In this section, we show that the estimated revenue converges to the optimal expected revenue from above and akin to the upper confidence bounds described in Section 2.1.1. First we leverage the structural properties of the MNL model to establish two key properties of the optimal expected revenue. In the first property, which we refer to as restricted monotonicity, we note that the optimal expected revenue is monotone in the MNL parameters. In the second property, we present a Lipschitz property of the expected revenue function. In particular, we have the following result.

Lemma 2.3 (Properties of the Optimal Revenue)**.** *Fix $\mathbf{v} \in \mathcal{R}_+^n$, let S^* be an optimal assortment when the MNL are parameters are given by $\mathbf{v}$, i.e. $S^* = \underset{S:|S|\leq K}{arg\ max}\, R(S, \mathbf{v})$. For any $\mathbf{w} \in \mathcal{R}_+^n$, we have:*

1. *(Restricted Monotonicity) If $v_i \leq w_i$ for all $i = 1, \cdots, N$. Then,*

$$R(S^*, \mathbf{w}) \geq R(S^*, \mathbf{v}).$$

2. *(Lipschitz) $|R(S^*, \mathbf{v}) - R(S^*, \mathbf{w})| \leq \dfrac{\sum_{i\in S^*} |v_i - w_i|}{1 + \sum_{j\in S^*} v_j}$.*

Proof. We will first prove the restricted monotonicity property and extend the analysis to prove the Lipschitz property.

Restricted Monotonicity. We prove the result by first showing that for any $j \in S^*$, we have $R(S^*, \mathbf{w}^j) \geq R(S^*, \mathbf{v})$, where $\mathbf{w}^j$ is vector $\mathbf{v}$ with the j^{th} component increased to w_j, i.e. $w_i^j = v_i$ for all $i \neq j$ and $w_j^j = w_j$. 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 $(v_j - w_j)(\sum_{i \in S/j} w_i + 1)$ on both sides of the above inequality and re-arranging terms, we can show that $R(S^*, \mathbf{w}^j) \geq R(S^*, \mathbf{v})$.

Lipschitz. Following the above analysis, we define sets $\mathcal{I}(S^*)$ and $\mathcal{D}(S^*)$ as

$$\mathcal{I}(S^*) = \{i | i \in S^* \text{ and } v_i \geq w_i\}$$

$$\mathcal{D}(S^*) = \{i | i \in S^* \text{ and } v_i < w_i\},$$

and vector $\mathbf{u}$ as,

$$u_i = \begin{cases} w_i & \text{if } \ i \in \mathcal{D}(S^*), \\ v_i & \text{otherwise.} \end{cases}$$

By construction of $\mathbf{u}$, we have $u_i \geq v_i$ and $u_i \geq w_i$ for all i. Therefore from the restricted monotonicity property, we have

$$\begin{aligned} R(S^*, \mathbf{v}) - R(S^*, \mathbf{w}) &\leq R(S^*, \mathbf{u}) - R(S^*, \mathbf{w}) \\ &\leq \frac{\sum\limits_{i \in S^*} r_i u_i}{1 + \sum\limits_{j \in S^*} u_j} - \frac{\sum\limits_{i \in S^*} r_i w_i}{1 + \sum\limits_{j \in S^*} u_j}, \\ &\leq \frac{\sum\limits_{i \in S^*} (u_i - w_i)}{1 + \sum\limits_{j \in S^*} u_j}. \end{aligned}$$

Lipschitz property in Lemma 2.3 follows from the definition of u_i. □

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.

Now, we show that the estimated revenue is an upper confidence bound on the optimal revenue. In particular, we have the following result.

Lemma 2.4. *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*

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

Proof. From Lemma 2.2 and union bound, we have with at least $1-\frac{6}{\ell}$ probability that $v_{i,\ell}^{\mathsf{UCB}} > v_i$ for all $i \in S^*$. In Lemma 2.3, we show that the optimal expected revenue is monotone in the MNL parameters. Therefore, the probability that the estimated revenue is greater than the optimal revenue is at least as large as the probability of $v_{i,\ell}^{\mathsf{UCB}} > v_i$ for all $i \in S^*$. □

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

Lemma 2.5. *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 $1 - \frac{7}{\ell}$.

Proof. Using a union bound, we can argue that the second statement of Lemma 2.2 holds true for all products in the optimal set with at least a probability of $1 - \frac{7}{\ell}$. The result then follows from the Lipschitz property established in Lemma 2.3. □

2.3 Improved Regret Bounds for "well separated" Instances

In this section, we consider the problem instances that are "well separated" and present an improved logarithmic regret bound. More specifically, we derive an $O(\log T)$ regret bound for Algorithm 1 for instances that are "well separated." In Section 2.2,

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 2.2 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.,

$$\Delta(\mathbf{v}) \ = \min_{\{S \in \mathcal{S} | R(S,\mathbf{v}) \neq R(S^*,\mathbf{v})\}} \{R(S^*, \mathbf{v}) - R(S)\}. \tag{2.10}$$

We have the following result.

Theorem 2 (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,*

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

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,

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

and we call it a "bad" epoch otherwise, where C_1 and C_2 are constants as defined in Lemma 2.2. 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 B.3). 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

$$\tau = \frac{4NC \log NT}{\Delta^2(\mathbf{v})}, \tag{2.11}$$

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

Lemma 2.6. *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 $R(S_\ell, \mathbf{v}) = R(S^*, \mathbf{v})$.*

Proof. Let $V(S_\ell) = \sum_{i \in S_\ell} v_i$. From Lemma 2.5, and definition of τ (see (2.11)), we have,

$$\begin{aligned} R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v}) &\leq \frac{1}{V(S_\ell) + 1} \sum_{i \in S_\ell} \left(C_1 \sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}} + C_2 \frac{\log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)} \right), \\ &\leq \Delta(\mathbf{v}) \left(\frac{C_1 \sum_{i \in S_\ell} \sqrt{v_i}}{2\sqrt{NC}\left(V(S_\ell) + 1\right)} + \frac{C_2}{4C} \right). \end{aligned} \tag{2.12}$$

From Cauchy-Schwartz inequality, we have

$$\sum_{i \in S_\ell} \sqrt{v_i} \leq \sqrt{|S_\ell| \sum_{i \in S_\ell} v_i} \leq \sqrt{NV(S_\ell)} \leq \sqrt{N}\left(V(S_\ell) + 1\right).$$

Substituting the above inequality in (2.12) and using the fact that $C = \max\{C_1^2, C_2\}$, we obtain $R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v}) \leq \frac{3\Delta(\mathbf{v})}{4}$. The result follows from the definition of $\Delta(\mathbf{v})$. □

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 2.6.

Lemma 2.7. *Algorithm 1 cannot offer sub-optimal assortments in more than $N\tau$ "good" epochs.*

Proof. We complete the proof using an inductive argument on N.

Lemma 2.7 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 $N \leq M$, we have that the number of "good epochs" offering sub-optimal products is bounded by $N\tau$, where τ is as defined in (2.11). 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.,

$$\mathcal{E}_{\mathcal{G}} = \left\{ \ell \;\middle|\; v_{i,\ell}^{\mathsf{UCB}} \geq v_i \text{ or } v_{i,\ell}^{\mathsf{UCB}} \leq v_i + C_1 \sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}} + C_2 \frac{\log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)} \; \forall i \right\}, \tag{2.13}$$

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

$$\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}} = \left\{ \ell \in \mathcal{E}_{\mathcal{G}} \mid R(S_\ell) < R(S^*) \right\}. \tag{2.14}$$

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,

$$\ell_i \stackrel{\Delta}{=} \min \left\{ \ell \mid T_i(\ell) = \tau \right\}.$$

Without loss of generality, assume that, $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_N$. 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},

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

where $\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}$ is as defined as in (2.14). 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 $|\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt(N)}}| \leq M\tau$. 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,

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

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,

$$|\hat{\mathcal{E}}_{\mathcal{G}}^{\mathsf{sub\_opt}}| \leq M\tau + T_N(\ell_{N-1}). \tag{2.15}$$

Note that from Lemma 2.6, 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,

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

From definition of ℓ_N, we have that $T_N(\ell_N) = \tau$ and substituting (2.15) in the above inequality, we obtain

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

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. □

The proof for Theorem 2 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 2.2 to bound the regret by $O(N^2 \log T)$. We provide a rigorous proof in Appendix B.3 for the sake of completeness. Note that for the special case of cardinality constraints, we have $|S_\ell| \leq K$ for every epoch ℓ. By modifying the definition of τ in (2.11) 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 2.2 (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* $v_0 \geq v_i$ *for all* i*, the regret of the policy given by Algorithm 1 at any time* T *is bounded as,*

$$\mathsf{Reg}(T, \mathbf{v}) \leq B_1 \frac{NK \log NT}{\Delta(\mathbf{v})} + B_2,$$

where, B_1 *and* B_2 *are absolute constants and* $\Delta(\mathbf{v})$ *is defined in* (2.10)*.*

It should be noted that the bound obtained in Corollary 2.2 is similar in magnitude to the regret bounds obtained by [43], when K is assumed to be fixed, and is strictly better than the regret bound $O(N^2 \log^2 T)$ established by [40]. 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.

2.4 Lower Bound for the MNL-Bandit

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 3 (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,*

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

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

Remark 2.4. (Optimality) Theorem 3 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, [15] 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 B.2) and the optimality gap for the unconstrained setting is $\sqrt{|S^*|}$.

2.4.1 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 B.5.1. To that end, we will assume that $K < N$ for the rest of this section. We prove Theorem 2.4 by a reduction to a parametric multi-armed bandit (MAB) problem, for which a lower bound is known.

Definition 2.1 (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,*

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

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 for the MAB problem.

Lemma 2.8. *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,*

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

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 2.8 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 13), and has been provided in Appendix B.5 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 2.2 (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 2.8. This will prove Theorem 2.4 by contradiction.

2.4.2 Construction of the MAB algorithm using the MNL algorithm

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 $v_i / (v_0 + \sum_{i \in S_\ell} v_i)$. 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.

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

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

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 $\mathcal{A}_{\mathsf{MNL}}$.

2.4.3 Proof of Theorem 2.4

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.

$$\mathsf{Reg}_{\mathcal{A}_{\mathsf{MAB}}}(T, \boldsymbol{\mu}) \leq \frac{1}{(1+\alpha)}\mathbb{E}\left(\sum_{\ell=1}^{L}\frac{1}{K}(V(S^*) - V(S_\ell))\right), \tag{2.16}$$

where the expectation in equation (2.16) 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,

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

Therefore, it follows that,

$$\mathsf{Reg}_{\mathcal{A}_{\mathsf{MNL}}}(L,\mathbf{v}) \geq \frac{1}{(1+\alpha)}\mathbb{E}\left[\sum_{\ell=1}^{L}\frac{1}{K}(V(S^*)-V(S_\ell))-\frac{\epsilon v^* L}{(1+\alpha)^2}\right], \tag{2.17}$$

where $v^*=\alpha+\epsilon$ and the expectation in equation (2.17) 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. The following result makes this precise.

Lemma 2.9. *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 $\mathbb{E}(L) \leq 2T+1 \leq 3T$. $\square$

Now we are ready to prove Theorem 3. From (2.16) and (2.17), we have

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

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

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

From lemma B.4, we have that $\mathbb{E}(L) \leq 3T$. Therefore, we have, $c\sqrt{\frac{\hat{N}\mathbb{E}(L)}{K}} = c\sqrt{N\mathbb{E}(L)} \leq c\sqrt{3NT} = c\epsilon T\sqrt{\frac{3}{\alpha}} < \frac{\epsilon T}{12}$ on setting $c < \frac{1}{12}\sqrt{\frac{\alpha}{3}}$. Also, using $v^* = \alpha + \epsilon \leq 2\alpha$, and 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

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

thus arriving at a contradiction.

2.5 Relaxing the "no-purchase" assumption

In this section, we extend our approach (Algorithm 1) to the setting where the assumption that $v_i \leq v_0$ 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 [40] and [43], 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.

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 $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until ℓ that offered product i
        update $\bar{v}_{i,\ell} = \dfrac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}$, 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

We can extend the analysis in Section 2.2 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,*

$$\mathsf{Reg}(T, \mathbf{v}) \leq C_1 \sqrt{BNT \log NT} + C_2 N \log^2 NT + C_3 NB \log NT,$$

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 2.2 to establish the correctness of the confidence intervals, the optimistic assortment, and finally deriving the convergence rates and regret bounds. We will now make the above notions precise.

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

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

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* $v_i \leq B v_0$ *for all* $i = 1, \ldots, N$*, for some* $B \geq 1$*, then we have that,*

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

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 $v_i \leq Bv_0$, 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 $p \geq 1/(B|S_\ell|+1)$. Therefore, we have the conditional expectation of $|\mathcal{E}_\ell|$ given that assortment S_ℓ is offered is bounded as,

$$\mathbb{E}\left(|\mathcal{E}_\ell| \;\middle|\; S_\ell\right) \leq B|S_\ell| + 1. \tag{2.18}$$

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

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

Substituting the above inequality in (2.18), we obtain

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

Confidence Intervals. We will now show a result analogous to Lemma 2.2, 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 2.11. *For every epoch ℓ, if $T_i(\ell) \geq 48 \log\left(\sqrt{N}\ell + 1\right)$ for all $i \in S_\ell$, then we have,*

1. *$v_{i,\ell}^{\mathsf{UCB2}} \geq v_i$ 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*

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

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

The proof is very similar to the proof of Lemma 2.2, where we first establish the following concentration inequality for the estimates $\hat{v}_{i,\ell}$, when $T_i(\ell) \geq 48 \log(\sqrt{N}\ell + 1)$ from which the above result follows. The proof of Lemma 2.11 along with the proof of Lemma 2.2 is deferred to Appendix A.2.

Convergence Rates of the Revenue Estimate: Using a union bound, we can argue that the second statement of Lemma 2.11 holds true for all products in the optimal set with at least a probability of $1 - \frac{7}{\ell}$. The following result which specifies the convergence rate of the revenue estimate follows from the Lipschitz property established in Lemma 2.3.

Lemma 2.12. *For every epoch ℓ, if $r_i \in [0,1]$ and $T_i(\ell) \geq 48 \log(\sqrt{N}\ell + 1)$ for all $i \in S_\ell$, then there exists constants C_1 and C_2 such that for every ℓ, we have*

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

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

Note that Lemma 2.4 and Lemma 2.12 are similar in spirit to first and second properties of the UCB estimates $\mathsf{UCB}_{t,i}$ discussed in Section 2.1.1. Therefore, the proof of Theorem 4 follows a similar analysis. However, the combinatorial aspects of the assortment optimization problem brings in additional challenges in completing the proof. In the interest of continuity, we defer the proof of Theorem 1 to Appendix B.4.

2.6 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 [43] 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 [43] 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.

2.6.1 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 (2.19)), 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 $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else ,} \end{cases} \tag{2.19}$$

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.

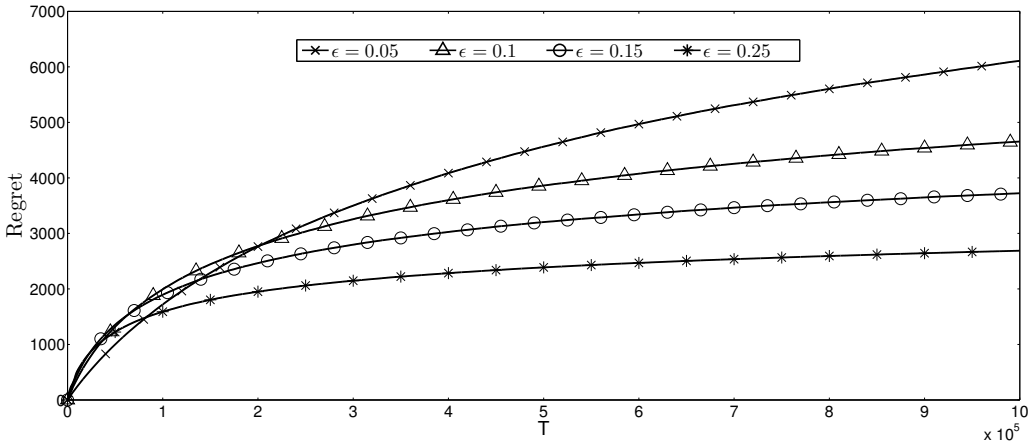

Figure 2.1: Performance of Algorithm 1 measured as the regret on the parametric instance (2.19). The graphs illustrate the dependence of the regret on T for "separation gaps" $\epsilon = 0.05, 0.1, 0.15$ and 0.25 respectively.

Results. Figure 2.1 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 2.1, the regret behavior is fundamentally different in the case of "small" ϵ and "large" ϵ. To see this, in Figure 2.2 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 2.3, where we showed that the regret grows logarithmically with time, if the optimal assortment and next best assortment are "well separated," while the worst-case regret scales as $\sqrt{T}$.

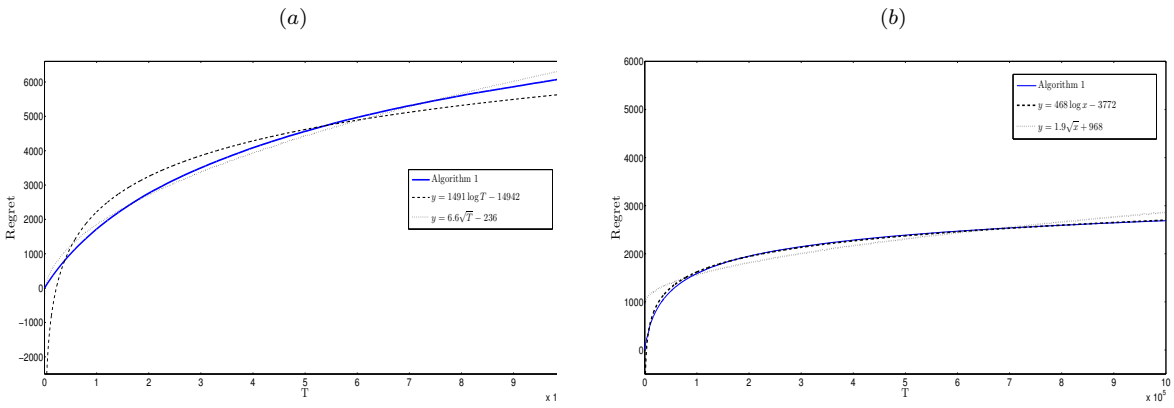

Figure 2.2: Best fit for the regret of Algorithm 1 on the parametric instance (2.19). The graphs (a), (b) illustrate the dependence of the regret on T for "separation gaps" $\epsilon = 0.05,$ and 0.25 respectively. The best $y = \beta_1 \log T + \beta_0$ fit and best $y = \beta_1\sqrt{T} + \beta_0$ fit are superimposed on the regret curve.

2.6.2 Comparison with existing approaches

In this section, we present a computational study comparing the performance of our algorithm to that of [43]. (To the best of our knowledge, [43] 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 [43] 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 [43] "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 (2.19) and for each value of $\epsilon \in \{0.05, 0.1, 0.15, 0.25\}$. Since the implementation of the policy in [43] 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 [43] and the selling horizon $T = 10^6$. Figure 2.3 compares the regret of Algorithm 1 with that of [43]. The results are based on running 100 independent simulations with standard error of 2%. We observe that the regret for [43] 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 [43] 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.

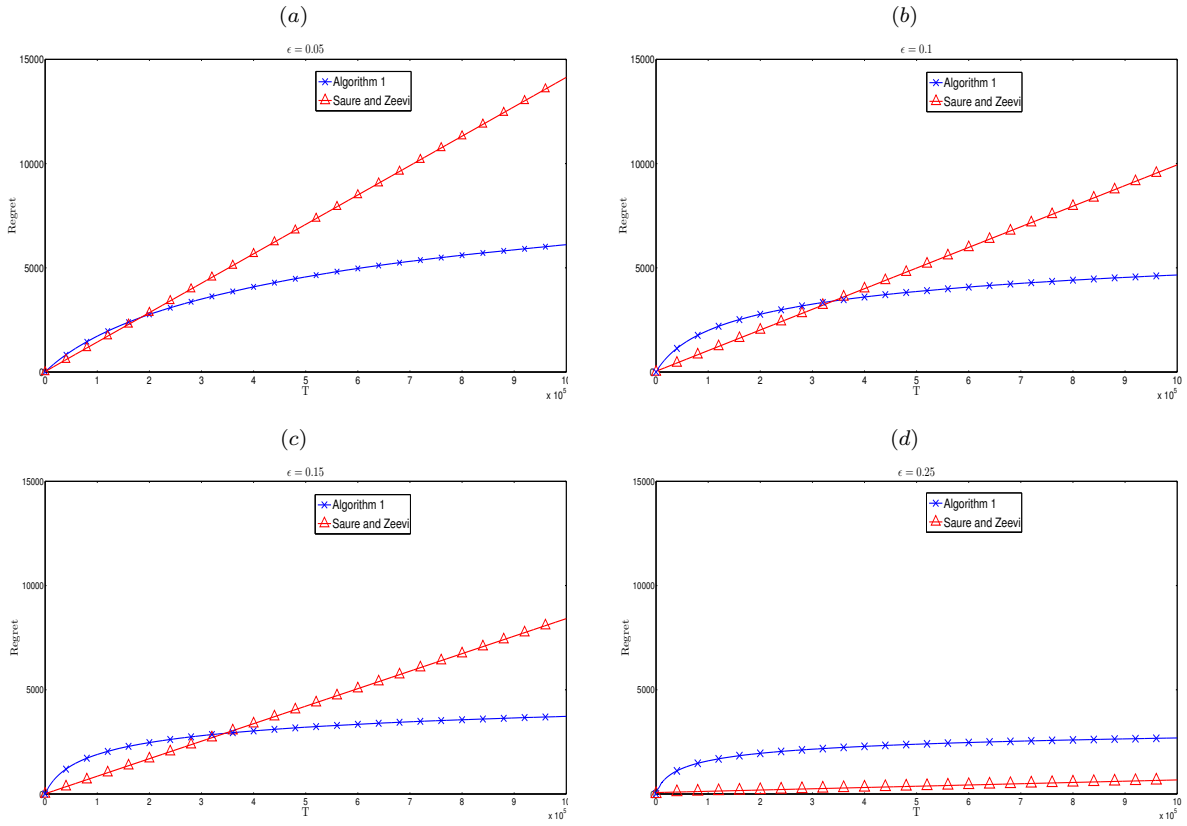

Figure 2.3: Comparison with the algorithm of [43]. The graphs (a), (b), (c) and (d) compares the performance of Algorithm 1 to that of [43] on problem instance (2.19), for $\epsilon = 0.05, 0.1, 0.15$ and 0.25 respectively.

2.6.3 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 [43].

Data description. We consider the "UCI Car Evaluation Database" (see [29]) which contains attributes for $N = 1728$ cars and consumer ratings for each car. The exact details of the attributes are provided in Table 2.1. 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.

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 2.1: Attribute information of cars in the database

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 2.1 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,

$$p_i(S) = \begin{cases} \dfrac{e^{\theta \cdot m_i}}{1 + \sum_{j \in S} e^{\theta \cdot m_j}}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{2.20}$$

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

$$Reg_\pi(T, \mathbf{m}) = \sum_{t=1}^{T} \left(\sum_{i \in S^*} p_i(S) - \sum_{i \in S_t} p_i(S_t) \right), \tag{2.21}$$

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 (2.21) is a special case formulation of the regret defined in (MNL-Bandit) 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

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

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

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

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 [43].

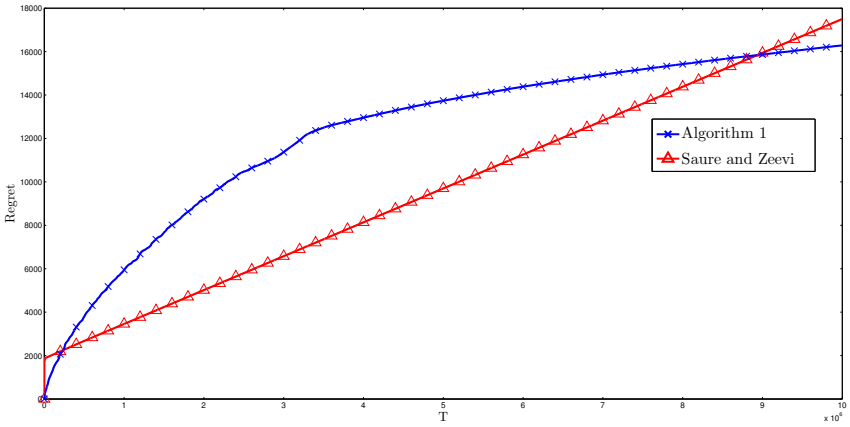

Figure 2.4: Comparison with the algorithm of [43] on real data. The graph compares the performance of Algorithm 1 to that of [43] on the "UCI Car Evaluation Databse' for $T = 10^7$.

We compare the performance of Algorithm 1 with that of [43], in terms of regret as defined in (2.21) with $\theta = \theta_{\mathsf{MLE}}$ and at each time index, the retailer can only show at most $k = 100$ cars. We implement [43]'s approach with their suggested mandotary exploration period, which explores every product for at least $20 \log T$ periods. Figure 2.4 plots the regret of Algorithm 1 and the [43] 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 [43] 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.

Summary and main insights. In this Chapter, 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."

Chapter 3

Thompson Sampling for the MNL-Bandit

It is widely recognized that UCB-type algorithms that optimize the worst case regret typically tend to spend "too much time" in the exploration phase, resulting in poor performance in practice (regret-optimality bounds notwithstanding). To that end, several studies (see [37], [24], [32]) have demonstrated that TS significantly outperforms the state of the art methods in practice. Motivated by the attractive empirical properties, in this chapter, we focus on a Thompson Sampling (TS) approach to the MNL-Bandit problem.

In Section 3.1 we give an overview of TS approach for the classical multi-armed bandit problem and highlight challenges associated with designing TS policies. In Section 3.2, we present our adaptations of the Thompson Sampling algorithm for the MNL-Bandit. In particular, we describe how to leverage the sampling technique introduced in Chapter 2 and design a prior distribution on the parameters of the MNL model such that the posterior update under the MNL-bandit feedback is tractable.

In Section 2.2, we prove our main result that our algorithm achieves a regret bounded as $\tilde{O}(\sqrt{NT}\log TK)$. Here, we also highlight the key ingredient of our approach, a two moment approximation of the posterior and the ability to judicially correlate samples, which is done by embedding the two-moment approximation in a normal family. In Section 3.4 demonstrates the empirical efficiency of our algorithm design.

3.1 Overview of Thompson Sampling

Thompson Sampling, introduced by Thompson [45] in 1933 is one of the oldest algorithms for bandit problems. For the classical MAB problem, where there are n arms with unknown means, $\{\mu_i\}_{i=1,\cdots,n}$, a TS based policy proceeds in the following manner

Algorithm 4 Basic Structure of TS policy for the classical MAB Problem

1: Assume a prior distribution $\mathsf{Pr}_0(\mu)$ on the parameters $\{\mu\}_{i=1,\cdots,n}$.
2: **for** $t = 1, 2, \ldots$ **do**
3: Sample parameters $\tilde{\mu}$ from the prior/posterior distribution $\mathsf{Pr}_{t-1}(\mu)$
4: Play the arm with highest sampled parameters, i.e. $i_t = \arg\max_i \tilde{\mu}_i$
5: Observe reward r_t which is generated from the distribution $\mathsf{Pr}(r_t|\mu)$
6: Update the posterior $\mathsf{Pr}_t(\mu) = \mathsf{Pr}(\mu|r_t) \propto \mathsf{Pr}(r_t|\mu)\mathsf{Pr}(\mu)$
7: **end for**

For ease of exposition we will consider the special case of two-armed bandits to highlight the intuition for why TS works in practice. In the TS algorithm, we generate samples $\tilde{\mu}_{1,t}$ and $\tilde{\mu}_{2,t}$ and play the arm with the larger sample. In the worst case that the sub-optimal arm has been played large number of times, the posterior distribution of the sub-optimal arm (say arm 2) will be concentrated around its true mean, μ_2, $(\mu_2 < \mu_1)$. If the optimal arm is not played often, then the posterior distribution of the optimal arm will have larger variance, which will frequently result in the sampled values being larger than the true mean, $\tilde{\mu}_1 > \mu_1$, which further ensures that the optimal arm is played more often. Typically, a worst case regret analysis of TS proceeds by showing that the best arm is optimistic (in the sense that the sampled parameter is larger than the true parameter) at least once every few steps.

Despite being intuitive, there are a number of challenges involved in designing a TS based approach. The primary concern is the the choice of prior, which not only has to ensure the posterior update is tractable but also guarantee that the posterior distribution has sufficient variance to explore the optimal arm. The tractability of the posterior update also impede the analysis of such an algorithm. For example, in all

existing work ([3], [2]) 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 (see [41]) have been proven, which are much weaker than the worst case regret notion we consider in this dissertation. There are also a number of heuristics or posterior approximation (see [42, 37]) algorithm that indicate how to approximate the TS algorithm. However, it is not immediately clear if these approximate TS based approaches facilitate theoretical analysis.

3.2 Algorithm.

In this section, we describe our posterior sampling (aka Thompson Sampling) based algorithm for the MNL-Bandit problem. The basic structure of Thompson Sampling involves maintaining a posterior on the unknown problem parameters, which is updated every time new feedback is obtained. In the beginning of every round, a sample set of parameters is generated from the current posterior distribution, and the algorithm chooses the best option according to these sample parameters. In the MNL-Bandit problem, there is one unknown parameter v_i associated with each item. To adapt the TS algorithm for this problem, we would need to maintain a joint posterior for $(v_1, \ldots, v_N)$. However, updating such a joint posterior is non-trivial since the feedback observed in every round is a sample from multinomial choice probability, $v_i/(1 + \sum_{j \in S} v_j)$, which clearly depends on the subset S offered in that round. In particular, even if we initialize with an independent prior from a popular analytical family such as multivariate Gaussian, the posterior distribution after observing the MNL choice feedback can have a complex description. As a first step in addressing this challenge, we attempt to design a Thompson Sampling algorithm with independent priors. In particular, we leverage a sampling technique introduced in in Chapter 2 that allows us to decouple individual parameters from the MNL choice feedback

and provides unbiased estimates of these parameters. We can utilize these unbiased estimates to efficiently maintain independent conjugate Beta priors for the parameters v_i for each i. We present the details in Algorithm 1 below.

3.2.1 A TS algorithm with independent conjugate Beta priors

Here, we present the first version of our Thompson sampling algorithm, which will serve as an important building block for our main algorithm in Section 3.2.3. In this version, we maintain a Beta posterior distribution for each item $i = 1, \ldots, N$, which is updated as we observe users' choice of items from the offered subsets. A key challenge here is to choose priors that can be efficiently updated on observing user choice feedback, in order to obtain increasingly accurate estimates of parameters $\{v_i\}$. To address this, we use the sampling technique introduced in Chapter 2 to decouple estimates of individual parameters from the complex MNL feedback. The idea is to offer a set S multiple times; in particular, a chosen set S is offered repeatedly until the "outside option" is repeatedly picked (in the motivating application discussed earlier, this corresponds displaying the same subset of ads until we observe a user who does not click on any of the displayed ads). Proceeding in this manner, due to the structure of the MNL model, the average number of times an item i is selected provides an unbiased estimate of parameter v_i. Moreover, the number of times an item i is selected is independent of the displayed set and is a geometric distribution with success probability $1/(1 + v_i)$ and mean v_i (see Lemma 2.1 in Chapter 2). This observation is used as the basis for our epoch based algorithmic structure and our choice of prior/posterior, as a conjugate to this geometric distribution.

Epoch based offerings: Our algorithm proceeds in epochs $\ell = 1, 2, \ldots$ similar to Algorithm 1 in Chapter 2. An epoch is a group of consecutive time steps, where a set

S_ℓ is offered repeatedly until the outside option is picked in response to offering S_ℓ. The set S_ℓ to be offered in an epoch ℓ is picked in the beginning of the epoch based on the sampled parameters from the current posterior distribution; the construction of these posteriors and choice of S_ℓ is described in the next paragraph. We denote the group of time steps in an epoch as $\mathcal{E}_\ell$, which includes the time step at which an outside option was preferred.

The following lemma which establishes the existence of a conjugate prior to our estimates play a key role in algorithmic construction.

Lemma 3.1 (Conjugate Priors)**.** *For any $\alpha > 3, \beta > 0$, let $X_{\alpha,\beta} = \frac{1}{\mathsf{Beta}(\alpha,\beta)} - 1$ and $f_{\alpha,\beta}$ be a probability distribution of the random variable $X_{\alpha,\beta}$. If v_i is distributed as $f_{\alpha,\beta}$ and $\tilde{v}_{i,\ell}$ is a geometric random variable with success probability $\frac{1}{v_i+1}$, then we have,*

$$\mathbb{P}\left(v_i \middle| \tilde{v}_{i,\ell} = m\right) = f_{\alpha+1,\beta+m}(v_i).$$

Proof. The proof of the lemma follows from the following result on the probability density function of the random variable $X_{\alpha,\beta}$. Specifically, we have for any $x > 0$

$$f_{\alpha,\beta}(x) = \frac{1}{B(\alpha,\beta)} \left(\frac{1}{1+x}\right)^{\alpha+1} \left(\frac{x}{x+1}\right)^{\beta-1}, \tag{3.1}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a)$ is the gamma function. Since we assume that the parameter v_i's prior distribution is same as that of $X_{\alpha,\beta}$, we have from (3.1) and Lemma 2.1,

$$\mathbb{P}\left(v_i \middle| \tilde{v}_{i,\ell} = m\right) \propto \left(\frac{1}{1+v_i}\right)^{\alpha+2} \left(\frac{v_i}{v_i+1}\right)^{\beta+m-1}.$$

□

Construction of conjugate prior/posterior: From Lemma 2.1, we have that for any epoch ℓ and for any item $i \in S_\ell$, the estimate $\tilde{v}_{i,\ell}$, the number of picks of item i in epoch ℓ is geometrically distributed with success probability $1/(1+v_i)$. Suppose

that the prior distribution for parameter v_i in the beginning of an epoch ℓ is same as that of

$$X_i = \frac{1}{\mathsf{Beta}(n_i, V_i)} - 1,$$

where $\mathsf{Beta}(n_i, V_i)$ is the Beta random variable with parameters n_i and V_i. In Lemma 3.1, we show that after observing the geometric variable $\tilde{v}_{i,\ell} = m$, the posterior distribution of v_i is same as that of,

$$X_i' = \frac{1}{\mathsf{Beta}(n_i + 1, V_i + m)} - 1.$$

Therefore, we use the distribution of $\frac{1}{\mathsf{Beta}(1,1)} - 1$ as the starting prior for v_i, and then, in the beginning of epoch ℓ, the posterior is distributed as $\frac{1}{\mathsf{Beta}(n_i(\ell), V_i(\ell))} - 1$, with $n_i(\ell)$ being the number of epochs the item i has been offered before epoch ℓ (as part of an assortment), and $V_i(\ell)$ being the number of times it was picked by the user.

Selection of subset to be offered: To choose the subset to be offered in epoch ℓ, the algorithm samples a set of parameters $\mu_1(\ell), \ldots, \mu_N(\ell)$ independently from the current posteriors and finds the set that maximizes the expected revenue as per the sampled parameters. In particular, the set S_ℓ to be offered in epoch ℓ is chosen as:

$$S_\ell := \arg\max_{|S| \leq K} R(S, \boldsymbol{\mu}(\ell)) \tag{3.2}$$

There are efficient polynomial time algorithms available to solve this optimization problem (e.g., refer to [17] and [40]).

The details of our procedure are provided in Algorithm 5.

3.2.2 Challenges and key ideas.

Posterior approximation and Correlated sampling. Algorithm 5 presents some unique challenges in theoretical analysis. A worst case regret analysis of Thompson Sampling based algorithms for MAB typically relies on showing that the best arm is optimistic at least once every few steps, in the sense that the parameter sampled

Algorithm 5 A TS algorithm for MNL-Bandit with Independent Beta priors

Initialization: For each item $i = 1, \cdots, N$, $V_i = 1$, $n_i = 1$.

$t = 1$, keeps track of the time steps

$\ell = 1$, keeps count of total number of epochs

while $t \leq T$ **do**

(a) (*Posterior Sampling)* For each item $i = 1, \cdots, N$, sample $\theta_i(\ell)$ from the $\mathsf{Beta}(n_i, V_i)$ and compute $\mu_i(\ell) = \frac{1}{\theta_i(\ell)} - 1$

(b) (*Subset Selection)* Compute $S_\ell = \underset{|S| \leq K}{\arg\max}\, R(S, \boldsymbol{\mu}(\ell)) = \frac{\sum_{i \in S} r_i \mu_i(\ell)}{1 + \sum_{j \in S} \mu_j(\ell)}$

(c) (*Epoch-based offering)*

repeat

Offer the set S_ℓ, and observe the user choice c_t;

Update $\mathcal{E}_\ell = \mathcal{E}_\ell \cup t$, time indices corresponding to epoch ℓ; $t = t + 1$

until $c_t = 0$

(d) (*Posterior update)*

For each item $i \in S_\ell$, compute $\tilde{v}_{i,\ell} = \sum_{t \in \mathcal{E}_\ell} \mathbb{I}(c_t = i)$, no. of picks of item i in epoch ℓ.

Update $V_i = V_i + \tilde{v}_{i,\ell}$, $n_i = n_i + 1$, $\ell = \ell + 1$.

end while

from the posterior is better than the true parameter. Due to the combinatorial nature of our problem, such a proof approach requires showing that every few steps, all the K items in the optimal offer set have sampled parameters that are better than their true counterparts. However, Algorithm 1 samples the posterior distribution for each parameter *independently* in each round. This makes the probability of being optimistic exponentially small in K.

We address this challenge by employing *correlated sampling* across items. To implement correlated sampling, we find it useful to approximate the Beta posterior distribution by a Gaussian distribution with approximately the same mean and variance as the former; to obtain what was referred to in the introduction as a two-moment approximation. This allows us to generate correlated samples from the N Gaussian distributions as linear transforms of a single standard Gaussian random

variable. Under such correlated sampling, the probability of all K optimal items to be simultaneously optimistic is a constant, as opposed to being exponentially small (in K) in the case of independent samples. However, such correlated sampling reduces the overall variance of the maximum of N samples severely, thus reducing exploration. We boost the variance by taking K samples instead of a single sample of the standard Gaussian. The resulting variant of Thompson Sampling algorithm is presented in Algorithm 6 in Section 3.2.3. We prove a near-optimal regret bound for this algorithm in Section 3.3.

3.2.3 A TS algorithm with posterior approximation and correlated sampling

Motivated by the challenges in theoretical analysis of Algorithm 5 described earlier, in this section we design a variant, Algorithm 6. There are three main changes in this version of the algorithm; posterior approximation by means of a Gaussian distribution, correlated sampling, and taking multiple samples (for "variance boosting"). We describe each of these changes below. First, we present the following result that helps us in approximating the posterior.

Lemma 3.2 (Moments of the Posterior Distribution)**.** *If X is a random variable distributed as* $\mathsf{Beta}(\alpha, \beta)$*, then*

$$\mathbb{E}\left(\tfrac{1}{X}-1\right) = \tfrac{\beta}{\alpha-1}, \quad \textit{and } \mathsf{Var}\left(\tfrac{1}{X}-1\right) \;= \tfrac{\frac{\beta}{\alpha-1}\left(\frac{\beta}{\alpha-1}+1\right)}{\alpha-2}.$$

Proof. We prove the result by relating the mean of the posterior to the mean of the Beta distribution. Let $\hat{X} = \frac{1}{X} - 1$. From (3.1), we have

$$\mathbb{E}(\hat{X}) = \frac{1}{B(\alpha,\beta)} \int_0^\infty x \left(\frac{1}{1+x}\right)^{\alpha+1} \left(\frac{x}{x+1}\right)^{\beta-1} dx,$$

Substituting $y = \frac{1}{1+x}$, we have

$$\mathbb{E}(\hat{X}) = \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha-2}(1-y)^{\beta} dx = \frac{B(\alpha-1,\beta+1)}{B(\alpha,\beta)} = \frac{\beta}{\alpha-1}.$$

Similarly, we can derive the expression for the $\mathsf{Var}(\hat{X})$. $\square$

Posterior approximation: We approximate the posterior distributions used in Algorithm 5 for the MNL parameters v_i, by Gaussian distributions with approximately the same mean and variance (refer to Lemma 3.2). In particular, let

$$\bar{v}_{i,\ell} := \frac{V_i(\ell)}{n_i(\ell)}, \quad \hat{\sigma}_i(\ell) := \sqrt{\frac{50\bar{v}_{i,\ell}(\bar{v}_{i,\ell}+1)}{n_i(\ell)}} + 75\frac{\sqrt{\log TK}}{n_i(\ell)}, \tag{3.3}$$

where $n_i(\ell)$ is the number of epochs the item i has been offered before epoch ℓ (as part of an assortment), and $V_i(\ell)$ being the number of times it was picked by the user. We will use $\mathcal{N}\left(\bar{v}_{i,\ell}, \hat{\sigma}_i^2(\ell)\right)$ as the posterior distribution for item i in the beginning of epoch ℓ. The Gaussian approximation of the posterior will facilitate efficient correlated sampling from posteriors. The correlated sampling will play a key role in avoiding some theoretical challenges in analyzing Algorithm 5.

Correlated sampling: Given the posterior approximation by Gaussian distributions, we correlate the samples by using a common standard normal variable and constructing our posterior samples as an appropriate transform of this common standard normal. More specifically, in the beginning of an epoch ℓ, we generate a sample from the standard normal distribution, $\theta \sim \mathcal{N}(0,1)$ and the posterior sample for item i, is generated as $\bar{v}_{i,\ell} + \theta\hat{\sigma}_i(\ell)$. This allows us to generate sample parameters for $i = 1, \ldots, N$ that are either simultaneously high or simultaneously low, thereby, boosting the probability that the sample parameters for *all* the K items in the best assortment are optimistic (the sampled parameter values are higher than the true parameter values).

Multiple (K) samples: The correlated sampling decreases the joint variance of the sample set. More specifically, if θ_i were sampled independently from the standard normal distribution for every i, then for any epoch ℓ, we have that

$$\mathsf{Var}\left(\max_{i=1,\cdots,N}\{\hat{v}_i(\ell) + \theta\hat{\sigma}_i(\ell)\}\right) < \mathsf{Var}\left(\max_{i=1,\cdots,N}\{\hat{v}_i(\ell) + \theta_i\hat{\sigma}_i(\ell)\}\right).$$

In order to boost this joint variance and ensure sufficient exploration, we modify the procedure to generate multiple sets of samples. In particular, in the beginning of an epoch ℓ, we now generate K independent samples from the standard normal distribution, $\theta^{(j)} \sim \mathcal{N}(0,1), j = 1, \ldots, K$. And then for each j, a sample parameter set is generated as:

$$\mu_i^{(j)}(\ell) := \hat{v}_i(\ell) + \theta^{(j)}\hat{\sigma}_i(\ell), \qquad i = 1, \ldots, N,$$

Then, we use the highest valued samples

$$\mu_i(\ell) := \max_{j=1,\cdots,K} \mu_i^{(j)}(\ell), \forall i,$$

to decide the assortment to offer in epoch ℓ,

$$S_\ell := \arg\max_{S \in \mathcal{S}} R(S, \boldsymbol{\mu}(\ell))$$

We summarize the steps in Algorithm 6. Here, we also have an "initial exploration period," where for every item i, we offer a set containing only i until the user selects the outside option.

Intuitively, the second moment approximation by Gaussian distribution and multiple samples in Algorithm 6 may make posterior converge slower and increase exploration. However, the correlated sampling may compensate for these effects by reducing the variance of the maximum of N samples, and therefore reducing the overall exploration. In Section 3.4, we illustrate some of these insights through some preliminary numerical simulations. Here, correlated sampling is observed to provide significant improvements when compared to independent sampling, and posterior approximation by Gaussian distribution has little impact.

3.3 Regret Analysis

We prove an upper bound on the regret of Algorithm 6 for the MNL-Bandit problem, under the following assumption.

Algorithm 6 A TS algorithm with Gaussian approximation and correlated sampling

Input Parameters $\alpha = 50$ and $\beta = 75$
Initialization: $t = 0$, $\ell = 0$, $n_i = 0$ for all $i = 1, \cdots, N$.
for each item, $i = 1, \cdots, N$ **do**

Display item i to users until the user selects the "outside option". Let $\tilde{v}_{i,1}$ be the number of times item i was offered. Update: $V_i = \tilde{v}_{i,1} - 1$, $t = t + \tilde{v}_{i,1}$, $\ell = \ell + 1$ and $n_i = n_i + 1$.

end for
while $t \leq T$ **do**

(a) (*Correlated Sampling)* **for** $j = 1, \cdots, K$

Sample $\theta^{(j)}(\ell)$ from the distribution $\mathcal{N}(0, 1)$; update $\bar{v}_{i,\ell} = \frac{V_i}{n_i}$.
For each item $i \leq N$, compute $\mu_i^{(j)}(\ell) = \bar{v}_{i,\ell} + \theta^{(j)}(\ell) \cdot \left(\sqrt{\frac{\alpha \bar{v}_{i,\ell}(\bar{v}_{i,\ell}+1)}{n_i}} + \frac{\beta \sqrt{\log TK}}{n_i} \right)$.
end

For each item $i \leq N$, compute $\mu_i(\ell) = \max_{j=1,\cdots,K} \mu_i^{(j)}(\ell)$

(b) *(Subset selection)* Same as step (b) of Algorithm 5.

(c) *(Epoch-based offering)* Same as step (c) of Algorithm 5.

(d) *(Posterior update)* Same as step (d) of Algorithm 5.

end while

Assumption 3.1. For every item $i \in \{1, \ldots, N\}$, the MNL parameter v_i satisfies $v_i \leq v_0 = 1$.

This assumption is equivalent to the outside option being more preferable to any other item. This assumption holds for many applications like display advertising, where users do not click on any of the displayed ads more often than not. Our main theoretical result is the following upper bound on the regret of Algorithm 6.

Theorem 3.1. *For any instance* $\mathbf{v} = (v_0, \cdots, v_N)$ *of the MNL-Bandit problem with* N *products,* $r_i \in [0, 1]$*, and satisfying Assumption 4.1, the regret of Algorithm 6 in time* T *is bounded as,*

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

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

3.3.1 Proof Sketch

We break down the expression for total regret

$$\mathsf{Reg}(T, \mathbf{v}) := \mathbb{E}\left[\sum_{t=1}^{T} R(S^*, \mathbf{v}) - R(S_t, \mathbf{v})\right],$$

into regret per epoch, and rewrite it as follows:

$$\begin{aligned}\mathsf{Reg}(T, \mathbf{v}) = &\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} |\mathcal{E}_\ell| \left(R(S^*, \mathbf{v}) - R(S_\ell, \boldsymbol{\mu}(\ell))\right)\right]}_{\mathsf{Reg}_1(T,\mathbf{v})} \\ &+ \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} |\mathcal{E}_\ell| \left(R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v})\right)\right]}_{\mathsf{Reg}_2(T,\mathbf{v})},\end{aligned} \tag{3.4}$$

where $|\mathcal{E}_\ell|$ is the number of time steps in epoch ℓ, and S_ℓ is the set repeatedly offered by our algorithm in epoch ℓ. Then, we bound the two terms: $\mathsf{Reg}_1(T, \mathbf{v})$ and $\mathsf{Reg}_2(T, \mathbf{v})$ separately.

Since S_ℓ was chosen as optimal set for MNL instance with parameters $\boldsymbol{\mu}(\ell)$, the first term $\mathsf{Reg}_1(T, \mathbf{v})$ is essentially the difference between the optimal revenue of the true instance and the optimal revenue of the sampled instance. This term contributes no regret if the revenues corresponding to the sampled instances are optimistic, i.e. if $R(S_\ell, \boldsymbol{\mu}(\ell)) > R(S^*, \mathbf{v})$. Unlike optimism under uncertainty approaches like UCB, this property is not directly ensured by our Thompson Sampling based algorithm. To bound this term, we utilize anti-concentration properties of the posterior, as well as the dependence between samples for different items, in order to prove that at least one of the K sampled instances is optimistic often enough.

The second term $\mathsf{Reg}_2(T, \mathbf{v})$ captures the difference in the revenue of the offered set S_ℓ when evaluated on sampled parameters in comparison to the true parameters. We bound this by utilizing the concentration properties of the posterior distributions.

It involves showing that for the sets that are played often, the posterior will converge quickly, so that revenue on the sampled parameters will be close to that on the true parameters.

In what follows, we will first highlight three key results involved in proving Theorem 1. In Section C.2 we will put together these properties and follow the above outline to prove Theorem 1.

Structural properties of the optimal revenue.

The first step in our regret analysis is to leverage the structural properties of the MNL revenue function established in Lemma 2.3. Re-collect that in the first property, which we refer to as restricted monotonicity, we have that the optimal expected revenue is monotone in the MNL parameters. In the second property, we have a Lipschitz property for the expected revenue function. In particular, the difference between the expected revenue corresponding to two different MNL parameters is bounded in terms of the difference in individual parameters. These properties project the non-linear reward function of the MNL choice into its parameter space and help us focus on analyzing the posterior distribution of the parameters.

Concentration of the posterior distribution.

The next step in the regret analysis is to show that as more observations are made, the posterior distributions concentrate around their means, which in turn concentrate around the true parameters. More specifically, we have the following two results.

Lemma 3.3. *For any $\ell \leq T$ and $i \in \{1, \cdots, N\}$, we have for any $r > 0$,*

$$\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right) \leq \frac{1}{r^4 K^3},$$

where $\hat{\sigma}_i(\ell)$ is as defined in (3.3).

Lemma 3.4. *If $v_i \leq 1$ for all $i = 1, \cdots, N$, then for any $m, \rho > 0$, $\ell \in \{1, 2, \cdots\}$ and $i \in \{1, \cdots, N\}$ we have,*

1. $$\mathcal{P}\left(|\hat{v}_i(\ell) - v_i| > 4\sqrt{\frac{\hat{v}_i(\ell)(\hat{v}_i(\ell)+1)m\log(\rho+1)}{n_i(\ell)}} + \frac{24m\log(\rho+1)}{n_i(\ell)}\right) \leq \frac{5}{\rho^m}.$$

2. $$\mathcal{P}\left(|\hat{v}_i(\ell) - v_i| \geq \sqrt{\frac{12v_i m\log(\rho+1)}{n_i(\ell)}} + \frac{24m\log(\rho+1)}{n_i(\ell)}\right) \leq \frac{4}{\rho^m}.$$

The above results indicate that for any item i and at the beginning of any epoch ℓ, the difference between the sample from the posterior distribution $\mu_i(\ell)$ and the true parameter v_i is bounded and is decreasing over time. Lemma 3.3 follows from the large deviation properties of Gaussian distribution and Lemma A.1 is similar to Chernoff bounds. For the sake of continuity, we defer the proof of these concentration results to Appendix C.1. Leveraging the Lipschitz property of the optimal revenue, this concentration of sample parameter around its true value will help us prove that the difference between the expected revenue of the offer set S_ℓ corresponding to the sampled parameters, $\boldsymbol{\mu}(\ell)$, and the true parameters, $\mathbf{v}$ also becomes smaller with time. In particular, we have the following result.

Lemma 3.5 (v)**.** *For any epoch ℓ, if $S_\ell = \underset{S:|S|\leq K}{arg\ max}\, R(S, \boldsymbol{\mu}(\ell))$*

$$\mathbb{E}\left\{(1 + \sum_{j\in S_\ell} v_j)\left[R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v})\right]\right\} \leq \mathbb{E}\left[C_1 \sum_{i\in S_\ell}\sqrt{\frac{v_i \log TK}{n_i(\ell)}} + C_2\frac{\log TK}{n_i(\ell)}\right],$$

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

The concentration property of the posterior distribution allows us to bound the second term, $\mathsf{Reg}_2(T, \mathbf{v})$ in (3.4). Therefore to bound the regret, it suffices to bound the difference between the the optimal revenue $R(S^*, \mathbf{v})$ and the expected revenue of the offer set corresponding to sampled parameters $R(S_\ell, \boldsymbol{\mu}(\ell))$.

Anti-Concentration of the posterior distribution.

We refer to an epoch ℓ as optimistic if expected revenue of the optimal set corresponding to the sampled parameters is higher than the expected revenue of the optimal set corresponding to true parameters, i.e., $R(S^*, \boldsymbol{\mu}(\ell)) \geq R(S^*, \mathbf{v})$. Any epoch that is not optimistic is referred to as non-optimistic epoch. Since S_ℓ is an optimal set for the sampled parameters, we have $R(S_\ell, \boldsymbol{\mu}(\ell)) \geq R(S^*, \boldsymbol{\mu}(\ell))$. Hence, for any optimistic epoch ℓ, the difference between the optimal revenue $R(S^*, \mathbf{v})$ and the expected revenue of the offer set corresponding to sampled parameters $R(S_\ell, \boldsymbol{\mu}(\ell))$ is bounded by zero. This suggests that as the number of optimistic epochs increases, the term $\mathsf{Reg}_1(T, \mathbf{v})$ decreases. The final and important technical component of our analysis is showing that the regret over non-optimistic epochs is "small". More specifically, we prove that there are only a "small" number of non-optimistic epochs. From the restricted monotonicity property of the optimal revenue (see Lemma 2.3), we have that an epoch ℓ is optimistic if every sampled parameter, $\mu_i(\ell)$ is at least as high as the true parameter v_i for every item i in the optimal set S^*. Recall that each posterior sample $\mu_i^{(j)}(\ell)$, is generated from a Gaussian distribution, whose mean concentrates around the true parameter v_i. We can use this observation to conclude that any sampled parameter will be greater than the true parameter with constant probability, i.e. $\mu_i^{(j)}(\ell) \geq v_i$. However, to show that an epoch is optimistic, we need to show that sampled parameters for *all* the items in S^* are larger than the true parameters. This is where the correlated sampling feature of our algorithm plays a key role. We use the dependence structure between samples for different items in the optimal set, and variance boosting provided by the sampling of K independence sample sets to prove an upper bound of roughly $1/K$ on the number of consecutive epochs between two optimistic epochs. More specifically, we have the following result.

Lemma 3.6 (Spacing of optimistic epochs)**.** *Let $\mathcal{E}^{\mathsf{An}}(\tau)$ denotes the group of consecutive epochs between an optimistic epoch τ and the next optimistic epoch τ'. For any*

$p \in [1, 2]$, *we have,*

$$\mathbb{E}^{1/p}\left[\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p\right] \leq \frac{e^{12}}{K} + 30^{1/p}.$$

Proof. Note that for any non-negative discrete random variable, X, we have $E(X) = \sum_x P(X \geq x)$. Hence, we will first establish a lower bound on the probability $\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p \geq q\right\}$ and use the preceding fact to obtain a bound on the moments of the number of non-optimistic epochs.

For the sake of brevity, let $r = \lfloor q^{1/p} \rfloor$ and $z = \sqrt{\log(rK+1)}$. Hence, we have,

$$\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p \geq q\right\} = \mathbb{P}\left\{|\mathcal{E}(\tau)| \geq r\right\}.$$

By definition, $\mathcal{E}^{\mathsf{An}}(\tau)$ less than r implies that one of the epochs $\tau+1, \cdots, \tau+r$ is optimistic. More specifically we have,

$$\begin{aligned}
\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| > r\right\} &= 1 - \mathbb{P}\left(\left\{\{\mu_i(\ell) \geq v_i \text{ for all } i \in S^*\} \text{ for some } \ell \in (\tau, \tau+r]\right\}\right), \\
&\leq 1 - \mathbb{P}\left(\left\{\{\mu_i(\ell) \geq \hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \geq v_i \text{ for all } i \in S^*\} \text{ for some } \ell \in (\tau, \tau+r]\right\}\right).
\end{aligned}$$

For the sake of brevity, let A_ℓ denote the event that the sampled parameter for every item in the optimal set is larger than z standard deviations away from the mean of the posterior distribution. Furthermore, B_ℓ denote the event that the true parameter of every item in the optimal set is smaller than mean of the posterior distribution plus z times the standard deviation of the posterior distribution. More specifically we have,

$$\begin{aligned}
A_\ell &= \{\mu_i(\ell) \geq \hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \text{ for all } i \in S^*\}, \\
B_\ell &= \{\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \geq v_i \text{ for all } i \in S^*\}. \\
\mathcal{B}_\tau &= \bigcap_{\ell=\tau+1}^{\tau+r} B_\ell.
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| \geq r\right\} &\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c \cup B_\ell^c\right), \\
&\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) + \sum_{\ell=\tau+1}^{\tau+r} \mathbb{P}(B_\ell^c), \\
&\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) + \sum_{i \in S^*} \mathbb{P}\left(\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) < v_i\right).
\end{aligned} \tag{3.5}$$

where the last two inequalities follows from union bound. Note that from the concentration property of the posterior distribution (see Lemma A.1), the probability of every event in the above inequality is small. In particular, substituting $m = 3.1$ and $\rho = rK$ in Lemma A.1 and using the fact that $rK \leq TK$ we obtain,

$$\mathbb{P}\left(\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) < v_i\right) \leq \frac{1}{(rK)^{3.1}}. \tag{3.6}$$

We will now use the tail bounds for Gaussian random variables to bound the probability $\mathbb{P}(A_\ell^c)$. For any Gaussian random variable, Z with mean μ and standard deviation σ, we have,

$$Pr(Z > \mu + x\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-x^2/2}.$$

Note that by design of Algorithm 6, $\mu_i(\ell) = \hat{v}_i(\ell) + \hat{\sigma}_i(\ell) \max_{j \leq K} \theta^{(j)}(\ell)$, where $\theta^{(j)}(\ell)$ are i.i.d standard normal random variables. Therefore, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) &= \mathbb{P}\Big(\theta^{(j)}(\ell) \leq z \text{ for all } \ell \in (\tau, \tau+r] \text{ and for all } j = 1, \cdots, K\Big), \\
&\overset{a}{\leq} \left[1 - \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\log rK}}{\log rK + 1} \cdot \frac{1}{\sqrt{rK}}\right)\right]^{rK}, \\
&\overset{b}{\leq} \exp\left(-\frac{r^{1/2}}{\sqrt{2\pi}} \frac{2\sqrt{\log rK}}{4\log rK + 1}\right), \\
&\overset{c}{\leq} \frac{1}{(rK)^{2.2}} \text{ for any } r \geq \frac{e^{12}}{K},
\end{aligned} \tag{3.7}$$

where inequality (a) follows from the tail bounds for standard normal random variables, inequality (b) follows from the fact that $1 - x \leq e^{-x}$ for all $x \geq 0$ and inequality (c) follows from the fact that $\exp\left(-\sqrt{x/2\pi \log x}\right) \leq 1/x^{2.2}$ for any $x \geq e^{12}$.

Hence from (3.5), (3.6), and (3.7) we have ,

$$\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| \geq r\right\} \leq \frac{1}{(rK)^{2.1}} + \frac{1}{(rK)^{2.2}} \text{ for any } r \geq \frac{e^{12}}{K}.$$

The result follows from the above inequality, definition of r and the fact that $\sum_{x=1}^{\infty} \frac{1}{x^y}$ is constant for any $y > 1$. □

We will now briefly discuss how the above properties are put together to bound $\mathsf{Reg}_1(T, \mathbf{v})$ and $\mathsf{Reg}_2(T, \mathbf{v})$. A complete proof is provided in Appendix C.2.

Bounding the first term $\mathsf{Reg}_1(T, \mathbf{v})$.

Firstly, by our assumption $v_0 \geq v_i, \forall i$, the outside option is picked at least as often as any particular item i. Therefore, it is not difficult to see that the expected value of epoch length $|\mathcal{E}_\ell|$ is bounded by $K + 1$, so that $\mathsf{Reg}_1(T, \mathbf{v})$ is bounded as

$$(K+1)\mathbb{E}\left(\sum_{\ell=1}^{L} R(S^*, \mathbf{v}) - R(S_\ell, \boldsymbol{\mu}(\ell))\right).$$

Recall that for every optimistic epoch, we have that the set S^* has at least as much revenue on the sampled parameters as on the true parameters. Hence, optimistic epochs don't contribute to this term.

To bound the contribution of the remaining epochs, we bound the individual contribution of any "non-optimistic" epoch ℓ by relating it to the closest optimistic epoch τ before it. By definition of an optimistic epoch and by the choice of S_ℓ as the revenue maximizing set for the sampled parameters $\boldsymbol{\mu}(\ell)$, we have

$$R(S^*, \mathbf{v}) - R(S_\ell, \boldsymbol{\mu}(\ell)) \leq R(S_\tau, \boldsymbol{\mu}(\tau)) - R(S_\ell, \boldsymbol{\mu}(\ell)) \leq R(S_\tau, \boldsymbol{\mu}(\tau)) - R(S_\tau, \boldsymbol{\mu}(\ell)).$$

To bound the last term, $R(S_\tau, \boldsymbol{\mu}(\tau)) - R(S_\tau, \boldsymbol{\mu}(\ell))$, the difference in the revenue of the set S_τ corresponding to two different sample parameters: $\boldsymbol{\mu}(\tau)$ and $\boldsymbol{\mu}(\ell)$, we will utilize the concentration property of the posterior and the Lipschitz property of the revenue function. From Lemma 3.5, the difference in the revenues can be bounded

by the sum of sample variances $\hat{\sigma}_i(\tau) + \hat{\sigma}_i(\ell)$ and since the variance at the beginning of epoch τ is larger than the variance at the beginning of epoch ℓ, we have,

$$|R(S_\tau, \boldsymbol{\mu}(\tau)) - R(S_\tau, \boldsymbol{\mu}(\ell))| \lesssim O\left(\sum_{i \in S_\tau} \hat{\sigma}_i(\tau)\right).$$

From the above bound, we have that the regret in non-optimistic epoch is bounded by the sample variance in the closest optimistic epoch before it. Utilizing the fact on an average there are only $1/K$ non-optimistic epochs (see Lemma 3.6) between any two consecutive optimistic epochs, we can bound the term $\mathsf{Reg}_1(T, \mathbf{v})$ as:

$$\mathsf{Reg}_1(T, \mathbf{v}) \leq (K+1) O\left(\mathbb{E}\left[\sum_{\ell \in \mathsf{optimistic}} \frac{1}{K} \sum_{i \in S_\ell} \hat{\sigma}_i(\ell)\right]\right).$$

A bound of $\tilde{O}(\sqrt{NT})$ on the sum of these deviations can be derived, which will also be useful for bounding the second term, as discussed next.

Bounding the second term $\mathsf{Reg}_2(T, \mathbf{v})$.

Noting that the expected epoch length when set S_ℓ is offered is $1 + \sum_{j \in S_\ell} v_i$, $\mathsf{Reg}_2(T, \mathbf{v})$ can be reformulated as

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left[\sum_{\ell=1}^{L} (1 + V(S_\ell)) \left(R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v})\right)\right],$$

Again, as discussed above, using Lipschitz property of revenue function and the concentration properties of the posterior distribution, this can be bounded in terms of posterior standard deviation (see (3.3))

$$\mathsf{Reg}_2(T, \mathbf{v}) \lesssim O\left(\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{i \in S_\ell} \hat{\sigma}_i(\ell)\right]\right).$$

Overall, the above analysis on Reg_1 and Reg_2 implies roughly the following bound on regret

$$O(\sum_{\ell=1}^{L} \sum_{i \in S_\ell} \hat{\sigma}_i(\ell)) = O\left(\sum_{\ell=1}^{L} \sum_{i \in S_\ell} \sqrt{\frac{v_i}{n_i(\ell)}} + \frac{1}{n_i(\ell)}\right) \log TK \leq O(\sum_{i=1}^{N} \log TK \sqrt{v_i n_i}),$$

where n_i is total number of times i was offered in time T. Then, utilizing the bound of T on the expected number of total picks, i.e., $\sum_{i=1}^{N} v_i n_i \leq T$, and doing a worst case scenario analysis, we obtain a bound of $\tilde{O}(\sqrt{NT})$ on $\mathsf{Reg}(T, \mathbf{v})$.

3.4 Empirical study

In this section, we analyze the various design components of our Thompson Sampling approach through numerical simulations. The aim is to isolate and understand the effect of individual features of our algorithm design like Beta posteriors vs. Gaussian approximation, independent sampling vs. correlated sampling, and single sample vs. multiple samples, on the practical performance.

We simulate an instance of MNL-Bandit problem with $N = 1000$, $K = 10$ and $T = 2 \times 10^5$, and the MNL parameters $\{v_i\}_{i=1,\ldots,N}$ generated randomly from $\mathsf{Unif}[0, 1]$. And, we compute the average regret based on 50 independent simulations over the randomly generated instance. In Figure 3.1, we report performance of following different variants of TS:

i) Algorithm 1: Thompson Sampling with independent Beta priors, as described in Algorithm 1.

ii) $\mathsf{TS}_{\text{Indpendent Gaussian Priors}}$: Algorithm 1 with Gaussian posterior approximation and independent sampling. More specifically, for each epoch ℓ and for each item i, we sample a Gaussian random variable independently with the mean and variance equal to the mean and variance of the Beta prior in Algorithm 1 (see Lemma 3.3).

iii) $\mathsf{TS}_{\text{Gaussian Correlated Sampling}}$: Algorithm 1 with Gaussian posterior approximation and correlated sampling. In particular, for every epoch ℓ, we sample a standard normal random variable. Then for each item i, we obtain a corresponding sample by multiplying and adding the preceding sample with the standard deviation and mean of the Beta prior in Algorithm 1 (see Step (a) in Algorithm 6). We use the values $\alpha = \beta = 1$ for this variant of Thompson Sampling.

iv) Algorithm 6: Algorithm 1 with Gaussian posterior approximation with correlated sampling and boosting by using multiple (K) samples. This is essentially the

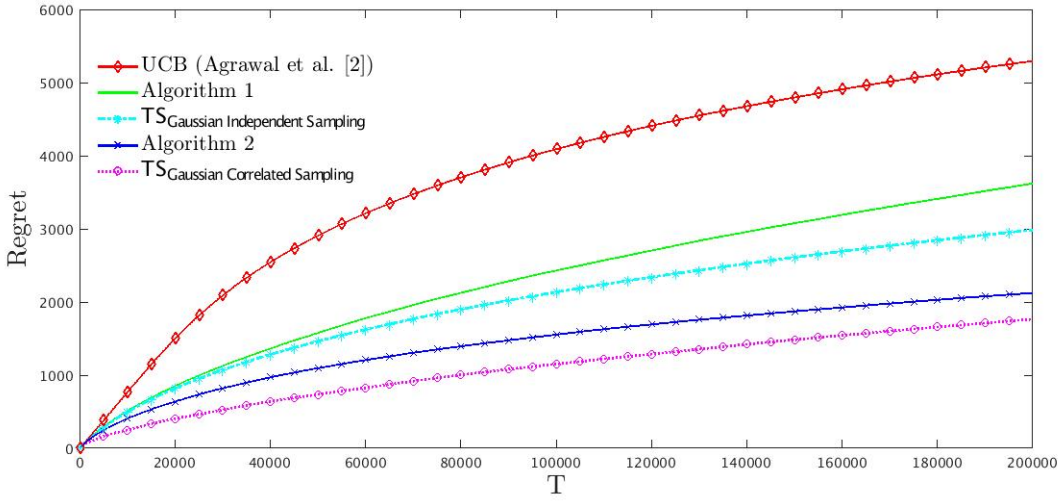

Figure 3.1: Regret growth with T for various heuristics on a randomly generated MNL-Bandit instance with $N = 1000, K = 10$.

version with all the features of Algorithm 6. We use the values $\alpha = \beta = 1$ for this variant of Thompson Sampling.

For comparison, we also present the performance of UCB approach in Chapter 2. We repeated this experiment on several randomly generated instances and a similar performance was observed. The performance of all the variants of TS is observed to be better than the UCB approach in our experiments, which is consistent with the other empirical evidence in the literature.

Among the TS variants, the performance of Algorithm 1, i.e., Thompson Sampling with independent Beta priors is similar to $\mathsf{TS}_{\text{Indpendent Gaussian Priors}}$, the version with independent Gaussian (approximate) posteriors; indicating that the effect of posterior approximation is minor. The performance of $\mathsf{TS}_{\text{Gaussian Correlated Sampling}}$, where we generated correlated samples from the Gaussian distributions, is significantly better than all the other variants of the algorithm. This is consistent with our remark earlier that to adapt the Thompson sampling approach of the classical MAB problem to our setting, ideally we would like to maintain a joint prior over the parameters $\{v_i\}_{i=1,\ldots,N}$ and update it to a joint posterior on observing the bandit feedback. However, since

this can be quite challenging and intractable, we used independent priors over the parameters. The superior performance of $\mathsf{TS}_{\text{Gaussian Correlated Sampling}}$ demonstrates the potential benefits of considering a joint (correlated) prior/posterior in such settings with combinatorial arms. Finally, we observe that the performance of Algorithm 6, where an additional "variance boosting" is provided through K independent samples, is worse than $\mathsf{TS}_{\text{Gaussian Correlated Sampling}}$ as expected, but still significantly better than the independent Beta posterior version Algorithm 1. Hence, significant improvements in performance due to correlated sampling feature of Algorithm 6 compensate for the slight deterioration caused by boosting.

3.5 Conclusion

In this Chapter, relying on structural properties of the MNL model, we develop a TS approach that is computationally efficient and yet achieves parameter independent (optimal in order) regret bounds. Specifically, we present a computationally efficient TS algorithm for the MNL-Bandit which uses a prior distribution on the parameters of the MNL model such that the posterior update under the MNL-bandit feedback is tractable. A key ingredient in our approach is a two moment approximation of the posterior and the ability to judicially correlate samples, which is done by embedding the two-moment approximation in a normal family. We show that our algorithm achieves a worst-case (prior-free) regret bound of $O(\sqrt{NT}\log TK)$ under a mild assumption that $v_0 \geq v_i$ for all i (more on the practicality of this assumption later in the text); the bound is non-asymptotic, the "big oh" notation is used for brevity. This regret bound is independent of the parameters of the MNL choice model and hence holds uniformly over all problem instances. The regret is comparable to the existing upper bound of $O(\sqrt{NT})$ proved in Chapter 2, yet the numerical results demonstrate that our Thompson Sampling based approach significantly outperforms

the UCB-based approach. Furthermore, the regret bound is also comparable to the lower bound of $\Omega(\sqrt{NT})$ established by [15] under the same assumption, suggesting the optimality of our algorithm. The methods developed in this paper highlight some of the key challenges involved in adapting the TS approach to the MNL-Bandit, and present a blueprint to address these issues that we hope will be more broadly applicable, and form the basis for further work in the intersection of combinatorial optimization and machine learning.

Chapter 4

Empirical Evaluation of Thompson Sampling: Evidence from Flipkart

In this chapter, we present evidence of empirical gains from employing dynamic assortment planning in optimizing product recommendations on Flipkart, an Indian ecommerce firm. First, in Section 4.2 we show that choice models like MNL which capture consumer preferences over an assortment have higher predictive power than traditional models which consider each item independently. In particular, we consider a structured MNL model, where every item is described by a set of attributes and the mean utility of a product is linear in the values of attributes. We show that the fit of this stylized MNL model is better than a simple logistic regression with the same set of attributes, which is the current model used at Flipkart. In Section 4.3, we will then present empirical evidence using click data from Flipkart to show that there is much to gain by implementing dynamic learning algorithms instead of the traditional "estimate, then optimize" approaches. In particular, we observe that an online algorithm like Thompson Sampling performs better in comparison to traditional approaches like estimating the model parameters based on initial observations and optimizing the decisions based on these estimates for the rest of the time period.

An important technical contribution of this chapter is the generalization of the learning algorithms from Chapters 2 and Chapters 3, which were designed to learn the model parameters in the product space. The possibility of different items being related to each other only through their attributes raises the possibility that one can design algorithms whose performance is independent of the number of items,

which is a major source of complexity. In Section 4.4, using the analysis developed in Chapter 3 as a foundation, we discuss how to extend the TS policy of Chapter 3 to the problem of learning in the attribute space. Specifically, we study how to leverage the relation between different items through attributes and obtain a regret bound which is independent of the number of items, and only depends on the number of attributes.

In this chapter, we describe our collaboration with Flipkart's homepage optimization team, where we consider the problem of improving product recommendations on the homepage while accounting for substitution patterns and adjusting the recommendations "on the fly." We will now present a brief background on the Flipkart's homepage optimization problem.

4.1 Introduction

Flipkart is an Indian e-commerce firm that has been founded in 2007 and has grown rapidly since to capture 39% of the total Indian e-commerce market [10]. It deals with a diverse range of products, serving more than 15 million active monthly consumers ([30]) who have collectively generated a revenue of US $ 7 billion in 2017 ([9]). Most of Flipkart's consumer base access Flipkart using a mobile app or a browser on the mobile phone, providing the firm with an unprecedented access in tracking consumer behavior on their site and using this information for future decision making.

One fundamental problem that concerns Flipkart is that of identifying the relevant set of products to display to a user. However, the challenges involved in identifying the optimal set of products to display are multi-fold. In settings like Flipkart, where the inventory is regularly updated with new items and demand trends constantly change, one has to constantly learn consumer preferences while concurrently attempting to maximizing revenues. This problem is further compounded by the fact that we can

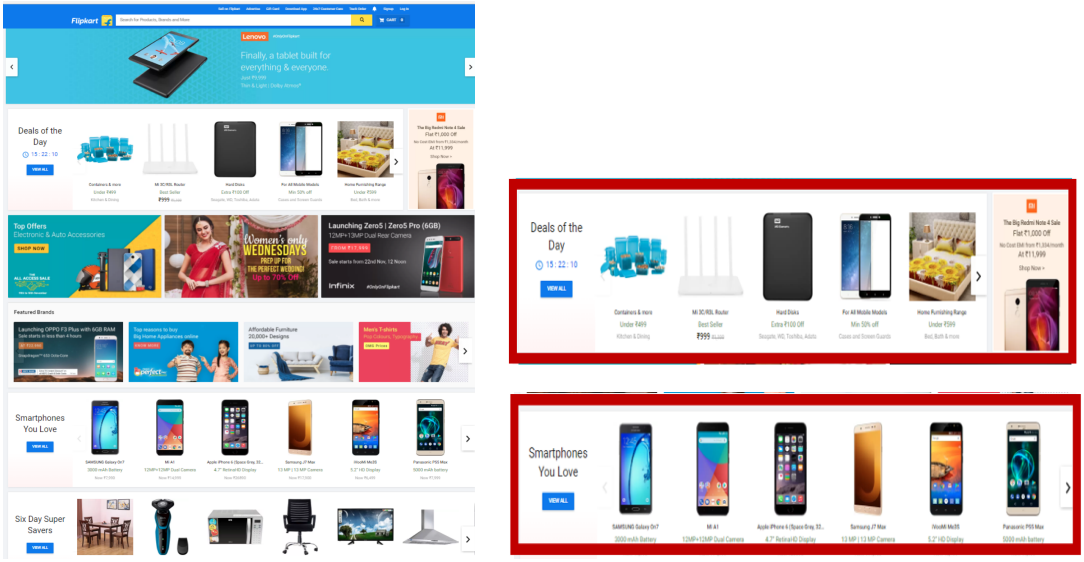

Figure 4.1: (Left) Example of Flipkart's Homepage. (Right) The enlarged widget, containing group of products. articles. The widget on the top has products that is being pushed by the sales team with discounts, while the widget below has smartphones.

only show a small number of products from a large selection of product categories and consumer preferences for a product depend on the overall set displayed (substitution effect). Moreover, apart from selecting the set of items to display we also need to decide how to bundle the items and where to display them. Motivated by this apparent need for a structured framework to recommend relevant set of items to consumers, in this dissertation, we consider the problem of identifying the optimal configuration of products on the homepage while accounting for substitution patterns and uncertainty in consumer preferences.

4.1.1 Background

On Flipkart, when a consumer visits their homepage they are displayed a wide range of products (see Figure 4.1). The standard practice at Flipkart is to group a selected set of products that follow a common theme or serve a common sales purpose as a widget and display an assortment of widgets to the consumer. For example, in

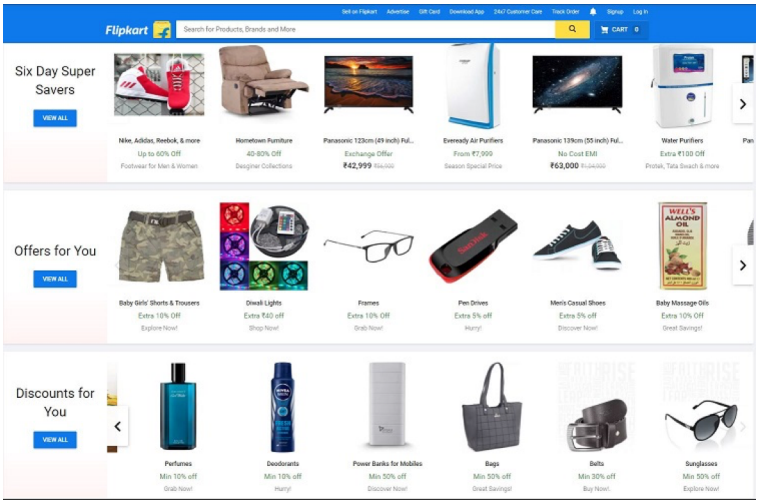

Figure 4.2: Example of homepage displaying widgets of similar theme

Figure 4.1, observe that there are 6 widgets (on the left image) and the products in each widget follow a common theme. More specifically, the widget titled "Deals of the Day" consists of products for which there's an ongoing discount offer, while the widget titled "Smartphones You Love" exclusively contains a pre-decided set of smartphones.

To manage the large number of products that could be displayed on the homepage, Flipkart follows the following mechanism in generating and selecting the widgets to be displayed. There are several units/teams within Flipkart that generate content (widgets) that serve their team's business function. For example, a sales team creates widgets consisting of products that are being offered with discounts. A merchandising team generates widgets consisting of specific brands that they want to advertise on the home page. Similarly, a recommendation team generates widgets consisting of products that the team perceives are ideal fit for the consumer on whom they have collected data before. Anytime a consumer visits the homepage, all the teams (automatically/algorithmically) generate their content and sends the widget requests to a centralized team, referred to as the homepage optimization team, which will then identify the optimal combination of widgets to be displayed for the user.

On an average, the homepage team receives around $30 - 40$ widget requests from various teams, after which it has to decide on the order in which the suggested widgets should be displayed. Typically, consumers only interact with around $3 - 5$ widgets depending on the screen space and hence, it is essential for the homepage team to optimize the rankings of the widgets to display so that the most relevant widgets are displayed in the most visible segments. Moreover widgets generated by different teams have an overlap in the theme of products and leads to substitution among the widgets displayed. For example, in Figure 4.2 we can observe that widgets "Six Day Super Savers," "Offers for You," and "Discounts for You" that are similar in spirit being displayed to a consumer. Consumers who are shopping around for a good deal on items would be equally interested in all the three widgets in contrast to the case where only one of the widget pertains to an offer. If one estimates the popularity of the widgets individually without accounting for substitution patterns, then the estimates will be significantly different in the above mentioned two scenarios. Therefore, to ensure the optimal configuration of the widgets, it is essential to consider a framework that accounts for substitution among the available alternatives.

4.2 Multinomial Logit and Logistic Regression

Here we present empirical evidence of the aforementioned discussion. More specifically, we argue that the MNL model which accounts for presence of similar alternatives has a better predictive power than Logistic model, which is the current model used for estimating the popularity of individual widgets by the homepage optimization team at Flipkart. Before going into the details of the logistic and MNL model, we first briefly describe the data available for the study.

Attribute	Description
Gender	Binary: male/female
Single category customer	Binary: only interested in single category
Is Parent	Binary: true/false
Is Student	Binary: true/false
Monetary	Categorical, indicating spending power: {1,2,3,4,5}
RFM	Categorical, indicating the status: {Platnum, Bronze, Gold, Silver}
Recency	Categorical, indicating the activity: {1,2,3,4,5}

Table 4.1: Description of available user attributes

4.2.1 Data Description

User Attributes. Flipkart's customer base predominantly interact with the firm either via the mobile app or a mobile browser. This makes it easy for Flipkart to track user attributes and personalize the widgets for that specific user. We provide the details of user attributes available in Table 4.1. However, due to an actively growing user base, there are still a considerable number of users for whom the personal attributes are unknown. For these users, Flipkart typically displays widgets assuming the average value for the unknown attributes.

Understandably, we have to account for heterogeneity in user preferences to develop models with higher predictive power. In this Chapter, our focus is on developing a better understanding of the impact of product recommendations that account for substitutions. Therefore, for the purpose of this study to avoid accounting for user heterogeneity, we focus only on a specific category of users for which all the observable attributes are same. Table 4.2 provides the details of the attributes for our considered segment.

Widget Attributes. As discussed earlier, every time a consumer interacts with the Flipkart's app or the homepage, different business units generate new widgets and request the homepage team to display their content to the user. The homepage optimization team, in order to predict which widgets will be more relevant for the user

Attribute	Value
Gender	male
Single category customer	false
Is Parent	Binary: false
Is Student	Binary: true
Monetary	5
RFM	Platinum
Recency	5

Table 4.2: User attributes for the segment under consideration

keeps track of certain widget attributes including the business unit that generated the widget, the content in the widget, the theme of the widget, at what position (rank) and with what layout has it been displayed. Table 4.3 provides the detailed descriptions of the widget attributes.

We convert the categorical attributes described in Table 4.3 to attributes with binary values by adding dummy attributes (for example each of the 13 widget types is considered as different as a different attribute that can take values 1 or 0) resulting in 1564 attributes. Now every widget is associated with an attribute vector $x_i \in \{0,1\}^{1564}$. We focus on consumer click data on a single day, 16th of April in 2018. There were approximately 250,000 unique user requests with the users having attributes described in Table 4.2. The click rate for individual widgets was 10%, while the click rate for the homepage (i.e., at least one of the widget is clicked) is around 35%. Around 8% of the users have clicked on multiple widgets, since random utility choice models do not allow for the possibility of clicking multiple items, we assume that only one these widgets is clicked and randomly select a widget (out of the clicked ones) to be the clicked widget. In what follows, we will discuss the fit of the Logistic Regression and the MNL model on this data set.

Attribute	Description
Widget Type	Categorical (13 types)- indicating the type of widget, for example if it is an advertisement/product card/deal card
Content Type	Categorical (14 types) - indicating the content and generator of the widget, for example personalized recommendation card based on past purchases
Is Pinned	If the widget is forced to be displayed by one of the business unit
View type	Categorical (12 types) Display configuration of the widget.
Rank	Position/Rank of the widget displayed. There were 40 unique rank/positions.
Store Categories	Product categories grouped in the widget. On an average there are 2 product categories for every widget. Over all there are 1483 unique product categories.
Store Null	A dummy feature to indicate product categories information is not available.

Table 4.3: Description of Widget Attributes

4.2.2 Logistic Regression

In the logistic model, every item's demand is estimate independently of the offer set. More specifically, under the logistic model, the probability that a consumer will click on a widget whose attributes are defined by the vector $x \in \{0,1\}^{1564}$ and the attribute weight vector θ^{LogReg} is given by

$$p_{\mathsf{Click}}(\theta^{\mathsf{LogReg}}, x) \triangleq \mathbb{P}\left(\text{Click}\middle|\theta^{\mathsf{LogReg}}\right) = \frac{e^{\theta^{\mathsf{LogReg}}.x}}{1+e^{\theta^{\mathsf{LogReg}}.x}}.$$

We utilize the click information on each widget offered and then leverage the maximum likelihood estimation $\theta^{\mathsf{LogReg}}_{\mathsf{MLE}}$ for θ^{LogReg} to estimate the click through rate of the offered

widgets and study the fit of the logistic model for the estimated θ_{MLE}. In particular, we compute $\theta_{\mathsf{MLE}}^{\mathsf{LogReg}}$ that maximizes the following regularized log-likelihood

$$\theta_{\mathsf{MLE}}^{\mathsf{LogReg}} = \arg\max_{\theta} \sum_{t=1}^{T} \log p_{\mathsf{Click}}(\theta, x_t) - \|\theta\|_2.$$

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, $\theta_{\mathsf{MLE}}^{\mathsf{LogReg}}$ (see [12].) We obtain the estimates using the popular stochastic gradient descent technique.

4.2.3 MNL model

In the MNL choice model, we assume that the mean utility of a product is linear in the values of attributes. More specifically, the mean utility of widget i, with attribute vector x_i is given by the inner product

$$\mu_i = \theta^{\mathsf{MNL}} \cdot x_i \quad \forall\ i$$

where $\theta^{\mathsf{MNL}} \in \mathbb{R}^{1564}$ is some fixed but initially unknown attribute weight vector. Under this model, the probability that a consumer clicks on widget i when offered an assortment of widgets $S \subset \{1, \ldots, N\}$ is assumed to be,

$$p_i(S) = \begin{cases} \dfrac{e^{\theta^{\mathsf{MNL}} \cdot x_i}}{1 + \sum_{j \in S} e^{\theta^{\mathsf{MNL}} \cdot x_j}}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

Here, again, we utilize the click information for each user request and then leverage the maximum likelihood estimation θ_{MLE} for θ to estimate the click through rate of the offered widgets and study the fit of the logistic model for the estimated θ_{MLE}. In particular, we compute θ_{MLE} that maximizes the following regularized log-likelihood

$$\theta_{\mathsf{MLE}}^{\mathsf{MNL}} = \arg\max_{\theta} \sum_{t=1}^{T} \sum_{i \in S_t \cup \{0\}} \mathbb{1}\left(\mathsf{widget\ i\ is\ clicked}\right) \cdot \log p_i(\theta, x_i) - \|\theta\|_2. \tag{4.2}$$

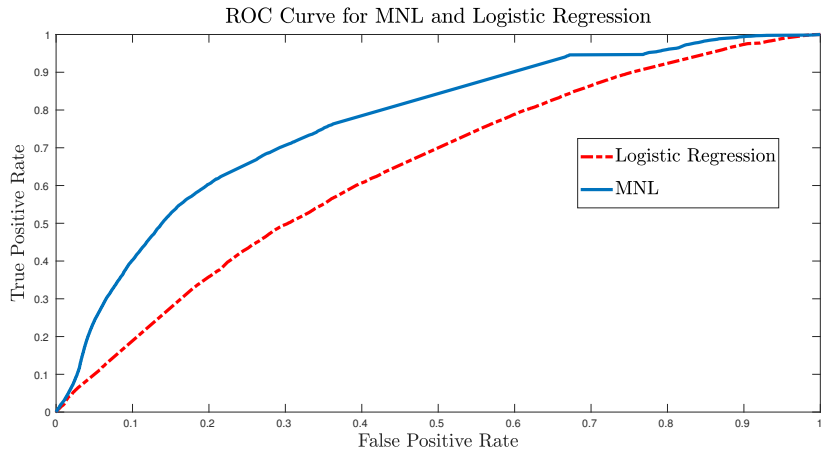

Figure 4.3: Fit of logistic and MNL regression on Flipkart's consumer click data.

The objective function in the preceding optimization problem is also convex and therefore we can use any of the standard convex optimization techniques to obtain the estimate, $\theta^{\mathsf{MNL}}_{\mathsf{MLE}}$ (see [12].) We obtain the estimates using the popular stochastic gradient descent technique.

4.2.4 Results: Fit of Logistic Regression vs Fit of the MNL model

For both logistic regression and MNL regression, we perform a $10-$ fold cross validation with 30-70 % train and test split. In particular, we randomly split the consumer click data into training and testing with 30% of the data in training segment and the remaining 70% of the data in the testing segment. We repeat this 10 times and compute the average of the 10 results for a more robust comparison. In Figure 4.3, we plot the roc curves of the fit on the test data. We can observe that the fit corresponding to the MNL model is significantly larger than the fit corresponding to the Logistic Regression with the corresponding area under the curve (AUC) being 77% and 64% respectively. This suggests that working with models that accounts for substitution patterns will provide better handle on understanding consumer preferences and thereby help in making optimal decisions.

4.3 Thompson Sampling for Optimal Configuration of Widgets

As discussed earlier, in settings like Flipkart, the inventory is regulary updated with new items and the demand trends constantly change. For example, 61 new product categories were introduced on the next day and 428 new product categories were introduced over the period of next one week, for which we have no historical data. The standard approach of estimating the widget preferences over a small amount of historical data and then optimizing the decisions based on this estimates is no longer optimal in such settings. Several studies (see [37], [24], [32]) in the literature have demonstrated that TS significantly improves the decision making under uncertainty. However, designing learning approaches in the attribute space are associated with new challenges as the difficulty not only arises from the fact that there are combinatorial number of assortments that can be offered, but also from the fact that a small number of attributes can lead to significantly large number of products to consider, making the number of possibilities to choose from extremely large.

In this Section, we will first present a common heuristic approach to implement a TS policy for learning in the attribute space. Subsequently in the next section, we will indicate how to generalize some of the theoretical ideas from Chapter 3 to design a provable TS algorithm for the attribute space[1]. More specifically, using the click data on Flipkart, we present empirical evidence of efficacy of our approximate Thompson Sampling approach in identifying the optimal configuration of the widgets, i.e. the optimal ranking of the widgets. Furthermore, we will also show that the TS approach perform better than the " estimate, then optimize" methods, contributing to the growing literature that advocates for moving beyond such standard practices.

[1] In the data, we do not have information regarding what widgets were rejected from being displayed. Therefore, we focus on optimizing the ranking of the widgets that were displayed to study the performance of Thompson Sampling.

Problem Description. Here we assume that the MNL choice model estimated in Section 4.2 as the ground truth model and further assume that we do not know the model parameters. Our objective is to learn these unknown attribute weights to identify the optimal configuration of the widgets (i.e. ranking among widgets), while simultaneously maximizing the over all click rates. With some abuse of notation, whenever we refer an assortment by it's attribute vector, we will assume that the attribute vector x_i does not include the display position/rank information of the widget. We will now describe the rank optimization problem more formally.

Let $S_t = (\mathbf{x}_{1t}, \mathbf{x}_{2t}, \cdots, \mathbf{x}_{kt})$ be the assortment of widgets that has been displayed for the t^{th} user request and r_{it} denote the rank/display positions of the widget with the attribute x_{it}. Our goal is to offer the assortment of widgets $S_t = (\mathbf{x}_{1t}, \mathbf{x}_{2t}, \cdots, \mathbf{x}_{kt})$ at the optimal ranks r^*_{it} such that the cumulative clicks are maximized according to the assumed MNL choice model, i.e.

$$\max_{\mathbf{r_1}, \cdots, \mathbf{r_T}} \sum_{t=1}^{T} p_i(S_t, \mathbf{r_t}) \tag{4.3}$$

where $p_i(S_t, \mathbf{r_t})$ is the choice probability (see (4.1)) of item i being clicked when widgets with attributes $\{x_{it}\}_{i=1,\cdots,k}$ are displayed at positions r_{it} respectively. Note that if $\theta^{\mathsf{MNL}}_{\mathsf{MLE}}$ is known a priori, then one can compute the optimal ranking for the widgets in a straightforward manner. In particular, for any collection of widgets $S_t = (\mathbf{x}_{1t}, \mathbf{x}_{2t}, \cdots, \mathbf{x}_{kt})$, we compute the inner product $x_{it} \cdot \theta^{\mathsf{MLE}}_{\mathsf{MNL}}$. and assign the widgets to the ranks in the decreasing order of the magnitude of the inner product. However, since $\theta^{\mathsf{MLE}}_{\mathsf{MNL}}$ is assumed to be unknown we have to focus on learning these weights while trying to maximize the cumulative click rate as described in (4.3). In what follows, we present a TS based learning approach to identify the optimal configuration of the widgets while maximizing the click through rates.

4.3.1 Laplacian Approximation

As discussed in Section 3.1, a fundamental challenge involved in pursing TS based approaches is the prior selection in order to allow tractable posterior update. One approach to address the challenge of tractable posterior update is approximating the posterior distribution by a multi-variate Gaussian distribution, a technique introduced by [37] and commonly referred to as Laplace Approximation. We refer the reader to [42] for other approximations for posterior sampling.

The likelihood function corresponding to the MNL choice model when assortments $S_1, \cdots, S_\tau$ are offered, is as follows:

$$\mathsf{Pr}(\text{Data observed until time t}) = \prod_{\tau=1}^{t-1} \prod_{i \in S_\tau \cup 0} \left(\frac{e^{\theta^{\mathsf{MNL}} \cdot x_i}}{1 + \sum_{j \in S_\tau} e^{\theta^{\mathsf{MNL}} \cdot x_j}} \right)^{\mathbb{1}(\text{widget i is clicked})} .$$

Therefore, if we assume a prior $f_0(\theta)$ on the attribute weights, by Bayes rules, we have that the posterior density at time $t-1$ for θ satisfies

$$f_{t-1}(\theta) \propto f_0(\theta) \prod_{\tau=1}^{t-1} \prod_{i \in S_\tau \cup 0} \left(\frac{e^{\theta^{\mathsf{MNL}} \cdot x_i}}{1 + \sum_{j \in S_\tau} e^{\theta^{\mathsf{MNL}} \cdot x_j}} \right)^{\mathbb{1}(\text{widget i is clicked})} .$$

For notational brevity, let $g(\theta)$ denote the right hand side of the above equation. Note that if $f_0(\theta)$ is a concave function, then $\log g(\theta)$ is concave and furthermore, $g(\theta)$ is unimodal (say the mode is $\hat{\theta}$.) In Laplacian approximation, we consider a second-order Taylor approximation to the log-density around it's mode and assume that $g(\theta) = e^{\log g(\theta)}$ sharply peaks around $\hat{\theta}$. In particular,

$$\log g(\theta) \approx \log g(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T C (\theta - \hat{\theta}),$$

where $C = -\nabla^2 \log g(\hat{\theta})$ is the Hessian of $\log g(\theta)$ at its mode. Therefore, if we start with a uniform prior, in Laplace Approximation, we compute the mode as the maximum likelihood estimate and approximate the posterior as Gaussian distribution

with mean as the maximum likelihood estimate $\theta^{\mathsf{MNL}}_{\mathsf{MLE}}$ and co-variance matrix as the inverse of the hessian matrix at the MLE estimate. Algorithm 7 provides the details of the approach.

Algorithm 7 TS with Laplacian Approximation for the Rank Optimization problem

Input: Tuning parameter α, warm up period T_0
while $t \leq T_0$ **do**
  Offer widgets in assortment S_t at random positions, $\mathbf{r}_t$
  Observe click information c_t and track data $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(S_t, \mathbf{r}_t, c_t)\}$
  $t = t + 1$
end while
while $t \leq T$ **do**
  Compute θ_t^{MLE} as the MLE from observations $\mathcal{D}_{t-1}$ (see (4.2))
  Compute $\mathcal{H}_t$ as the Hessian of the log-likelihood function at θ_t^{MLE}.
  Sample $\theta_t^{\mathsf{TS}} \sim \mathcal{N}\left(\theta_t^{\mathsf{MLE}}, \alpha H_t^{-1}\right)$
  Offer widgets in the assortment S_t at the optimal ranks assuming θ_t^{TS} as the true parameter
  Observe click information c_t and track data $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(S_t, \mathbf{r}_t, c_t)\}$
  $t = t + 1$.
end while

Note that the Hessian H_t in our problem setting is a matrix of dimension 1564, computing H_t^{-1} at every time step is a computationally expensive process. Therefore, to speed up the learning algorithm, we follow the approach in [37] and further approximate the covariance matrix by a diagonal matrix. We also perform the updates in a batch fashion to further enhance the computational speed of TS algorithm. Algorithm 8 provide the details of our tractable TS algorithm.

4.3.2 Results

Performance Metric. A policy which does not learn attribute weights would configure the widgets randomly. In contrast, an algorithm which has a priori knowledge of the weights offers the widgets in the optimal order there by resulting in an increased click through rates (CTR). Any policy π which does not know the attribute weights but attempts to learn them will perform better than a random policy but worse than

Algorithm 8 TS with Diagonal Approximation of Laplacian

Input: Tuning parameter α, batch size T_0.
$\hat{\theta}_i = 0$, $q_i = \lambda$ for all $i = 1, \cdots, 1564$.
while $t \leq T$ **do**
Sample $\theta_i^{\mathsf{TS}} \sim \mathcal{N}\left(\theta_i, q_i^{-1}\right)$
Offer widgets in the assortment S_t at the optimal ranks assuming θ^{TS} as the true parameter. Observe click information c_t
$t = t + 1$
if t is a multiple of T_0 **then**
Consider a batch of observations $\{(S_\tau, \mathbf{r}_\tau, c_\tau)\}_{\{\tau = t - T_0, \cdots, t\}}$
Compute θ^{MLE} as the regularized MLE from the new observations, i.e. argmax of the following objective function.

$$-\frac{1}{2}\sum_{i=1}^{1564} q_i(\theta_i - \hat{\theta}_i)^2 + \sum_{\tau=t-T_0}^{t}\left[\sum_{j\in S_\tau} \mathbb{1}\left(\text{widget i is clicked}\right)\theta\cdot x_j - \log\left(1 + \sum_{\ell\in S_\tau}\exp\left(\theta\cdot x_\ell\right)\right)\right].$$

Update $\hat{\theta} = \theta^{\mathsf{MLE}}$ and $q_i = \sum_{\tau=t-T_0}^{t}\sum_{\ell\in S_\tau} x_{\ell i}^2 \cdot p_\ell(S_\tau) - \left(\sum_{\ell\in S_\tau} x_{\ell i} \cdot p_\ell(S_\tau)\right)^2$,
where $p_\ell(S_\tau)$ is the choice probability as defined in (4.1).
end if
end while

the policy that knows these weights a priori. Therefore, we evaluate the performance of the policy π by comparing the gain in CTR it obtains over a random policy and the gain in CTR obtained by a policy which knows weights a priori over a random policy. More specifically, let

$$\mathsf{Reg}(T) = \sum_{t=1}^{T}\left(\sum_{i\in S_t} p_i(S_t, \mathbf{r}_t^*) - \sum_{i\in S_t} p_i(S_t, \mathbf{r}_t^{\mathsf{Random}})\right), \tag{4.4}$$

be the gain in CTR obtained by the policy that knows the attribute weights over a policy that configures widgets randomly and et

$$\mathsf{Reg}_\pi(T) = \sum_{t=1}^{T}\left(\sum_{i\in S_t} p_i(S_t, \mathbf{r}_t^{\pi}) - \sum_{i\in S_t} p_i(S_t, \mathbf{r}_t^{\mathsf{Random}})\right), \tag{4.5}$$

be the gain in CTR obtained by policy π over a random policy. We measure the performance of policy π by the ration $\mathsf{Reg}_\pi/\mathsf{Reg}(T)$. Higher value of the ratio suggests that the algorithm learns the weights quickly and is mimicking the policy which has

knowledge of weights a priori. On the contrary, lower value of the ratio suggests that the algorithm has not still figured out attribute weights and is behaving similar to a random policy.

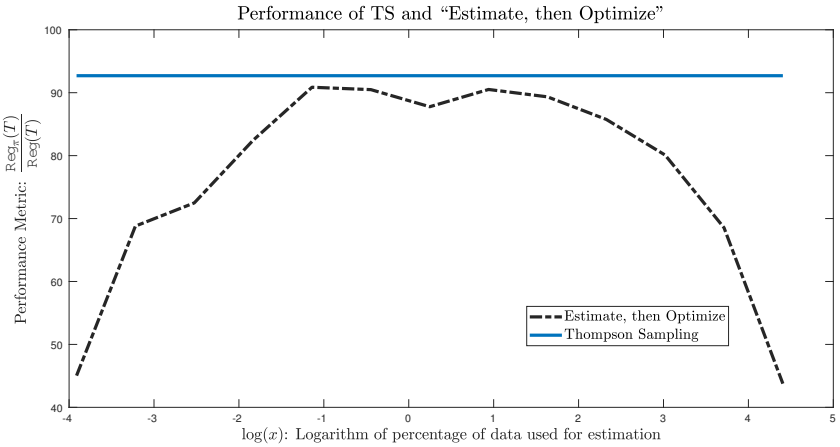

Figure 4.4: Comparing the Performance of Thompson Sampling with "Estimate, then Optimize" approach

We implement Algorithm 8 with $T_0 = 2000$ and $\lambda = 1$. We also implement the common "estimate, then optimize" policy, where the weights are estimated based on randomly selected $x\%$ of the data and then the ranks are optimized for the remaining $100 - x\%$ of the data based on the aforementioned estimates. Figure 4.4 plots the performance of Algorithm 8 and the performance of "estimate, then optimize" policy for various values of x ranging from 0.02% to 81%. The results are based on running 50 independent simulations and have a standard error of 2%. We can observe that the Thompson Sampling's performance is around 93% of the policy that has knowledge of the weights a priori, suggesting that TS based policy is almost comparable to the policy that knows the weights. We also observe that the performance of the "estimate, then optimize" approach depends on how much data is leveraged for estimation. One can note that estimating over small amount of data leads to under exploration, while estimating with large amounts of data leads to over exploration and poor performance, highlighting the common challenge associated with these approaches. Furthermore, we can observe that TS based policy outperforms the "estimate, then optimize"

approach for all possible estimation data sizes. izon and does not require any a priori knowledge on the parameters, making a case for the universal applicability of our approach.

4.4 Theoretical Guarantees

In this Section, we indicate how to generalize the theory developed in Chapter 3 to design an algorithm with provable guarantees. Here, we consider a variant of the MNL-Bandit problem with cardinality constraints and product attributes. More specifically, we assume that the choice probabilities are as described in 4.1, i.e., the probability that a consumer clicks on product i when offered an assortment of products $S \subset \{1, \ldots, N\}$ is assumed to be,

$$p_i(S) = \begin{cases} \dfrac{e^{\theta_* \cdot x_i}}{1 + \sum_{j \in S} e^{\theta_* \cdot x_j}}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta_* \in \mathbb{R}^d$ is some fixed, but initially unknown parameter vector. Let $R(S, \theta_*)$ denote the expected revenue when assortment S is offered and the parameter vector is given by θ_*. Our goal is to offer assortments $S_1, \cdots, S_T$ at times $1, \cdots, T$ such that $|S_t| \leq K$ to minimize regret defined as,

$$\mathsf{Reg}(T, \theta_*) = \sum_{t=1}^{T} R(S^*, \theta_*) - \mathbb{E}\left[R(S_t, \theta_*)\right], \tag{4.6}$$

and more specifically obtain regret bounds that depend on the dimension of attributes, d and not on the number of products.

We will briefly describe how the techniques of TS algorithm for MNL-Bandit developed in Chapter 3 can be used to design an algorithm for the above problem.

Challenges and overview. A key difficulty in our problem arises not only from the fact that there are combinatorial number of assortments that can be offered, but

also from the fact that a small number of attributes can lead to significantly large number of products to consider, making the number of possibilities to choose from extremely large. We adapt some fundamental ideas from Algorithm 5. Since our primary objective is to obtain regret bounds that are not dependent on the number of products, we focus on the attribute space rather than the product space, where it can exploit the past purchase behavior to update the posterior distribution for the attribute weight vector θ. However, there are two main challenges in implementing this scheme.

First, unlike the MNL-Bandit scenario, it is not easy to obtain unbiased estimates for the values of parameter vector θ. To overcome this issue, we use a maximum likelihood estimate and use the Laplacian approximation described in Section 4.3.1 to update the posterior distribution. A key aspect in our analysis is establishing the concentration bounds for the MLE estimate. We use the martingale argument (see [22]) for the MLE estimates to derive such bounds. As discussed earlier in Chapter 3, the worst case analysis of TS typically proceeds by showing that the best arm is optimistic at least once every few steps, in the sense that its sampled parameter is better than the true parameter. To avoid the challenges involved with the combinatorial structure of the MNL Bandit problem, we correlated sampling and variance boosting as in Chapter 3 to facilitate theoretical analysis.

4.4.1 Algorithm.

Algorithm 9 provides the details of our TS algorithm.

We make the following assumptions to facilitate analysis.

Assumption 4.1. *The norm of attribute vectors m_i is bounded for all $i = 1, \cdots, N$, i.e. there exists $c_x < \infty$ such that for all $i \leq N$, we have $\|x_i\|_2 \leq c_x$.*

Algorithm 9 Approximate TS for Assortment Planning in Attribute Space

Input: Tuning parameter α, warm up period t_0

while $t \leq T$ **do**

Compute θ_t^{MLE} as the MLE from observations $\mathcal{D}_{t-1}$ (see (4.2))

Compute $\mathcal{H}_t$ as the Hessian of the log-likelihood function at θ_t^{MLE}.

(a) (*Correlated Sampling)* **for** $j = 1, \cdots, K$

Sample $\theta^{(j)}(t)$ from the distribution $\mathcal{N}(0, 1)$;

For each item $i \leq N$, compute $\mu_i^{(j)}(t) = \theta_t^{\mathsf{MLE}} \cdot x_i + \alpha\theta^{(j)}(t) \cdot \|x_i\|_{H_t^{-1}}$.

end

For each item $i \leq N$, compute $\mu_i(t) = \max_{j=1,\cdots,K} \mu_i^{(j)}(t)$

(b) *(Subset selection)* Compute $S_t = \arg\max_{|S|\leq K} R(S, \boldsymbol{\mu}(t)) = \dfrac{\sum_{i\in S} r_i e^{\mu_i(t)}}{1 + \sum_{j\in S} e^{\mu_j(t)}}$

Observe click information c_t and track data $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(S_t, \mathbf{r}_t, c_t)\}$

$t = t + 1$.

end while

Assumption 4.2. *There exists a constant $c_\mu > 0$ such that $c_\mu = \inf_{\theta\in\Theta,\, i\leq N} \dot{p}_{S_\tau}(\theta \cdot x_i)$, where Θ is the set of all feasible attribute weights.*

Assumption 4.3. $\sup_{\theta\in\Theta,\, i\leq N} exp(\theta' x_i) \leq 1$.

We first establish a concentration inequality for the MLE estimate using the martingale argument of [22]. Observe that the correlated sampling in step (a) of Algorithm 9 is similar to the correlated sampling step of Algorithm 6. Therefore, we can follow the proof technique of Algorithm 6 to leverage the anti-concentration properties of Gaussian distribution to argue that the best arm is optimistic often enough. Noting that the optimistic arm is played often, we can then follow the UCB analysis of [22] to derive the following result. We present a detailed proof in Appendix D.

Theorem 4.1. *If $\alpha = \dfrac{60\sqrt{d \log d}}{c_\mu}$, then under Assumptions 4.1 and 4.2, the regret*

of Algorithm 9 is bounded as

$$Regret \leq O\left(d\frac{\log KTd}{c_\mu}\sqrt{KT\log\left(\frac{c_x^2 T}{\lambda_0}\right)}\right),$$

where λ_0 *is the minimum eigenvalue of* $M_{t_0} \triangleq \sum_{t=1}^{t_0}\sum_{i\in S_t} x_i x_i^T$.

4.5 Conclusion.

In this Chapter, we have demonstrated empirical gains from employing dynamic assortment planning in optimizing product recommendations on Flipkart, an Indian ecommerce firm. We have also argued that choice models like MNL which capture consumer preferences over an assortment have higher predictive power than traditional models which consider each item independently. Using the analysis developed in Chapter 3 as a foundation, we have presented a framework that indicates how to design TS-based policy to the problem of learning in the attribute space. However, the complete development of an algorithm for the attribute-based MNL-Bandit with regret depending only on the number of attributes, and an algorithm independent of any problem parameters remains an interesting open problem.

Chapter 5

Algorithms for Static Assortment Planning

In this Chapter, we consider settings when the model parameters are known and focus on developing tractable optimization algorithms for the MNL and the NL model under totally unimodular constraint structures. The totally unimodular constraints model a rich class of practical assortment planning problems including cardinality constraints, partition matroid constraints and joint display and assortment constraints.

First we consider the assortment planning problem under the MNL model and show that a natural linear programming (LP) relaxation is tight. The LP based approach provides robustness to handle capacity constraints in addition to the existing TU constraints. In particular, we consider an arbitrary additional constraint to the set of TU constraints such that the resulting set of constraints are not TU. We present a Polynomial Time Approximation Scheme (PTAS) for the assortment optimization problem under this more general set of constraints where for any $0 < \epsilon < 1$, we obtain a solution with objective value at least $(1 - \epsilon)$ times the optimal in running time polynomial in the input size for a fixed ϵ. As a consequence of this problem, we obtain PTAS for joint display and assortment optimization problem with an additional capacity constraint.

We then consider the assortment optimization problem under NL model with TU constraints and provide a Fully Polynomial Time Approximation Scheme (FPTAS) for this problem, where for any $0 < \epsilon < 1$, we obtain a solution with objective value at least $(1 - \epsilon)$ times the optimal in running time polynomial in the input size and $1/\epsilon$. We also show that the exact assortment optimization under NL model

Choice model	Assortment optimization problem
Multinomial Logit (MNL)	• Tight LP relaxation for totally unimodular constraint structures. (Theorem 5.1) • Joint assortment and display optimization problem is polynomially solvable. (Section 5.1.3) • PTAS for Joint assortment and display optimization problem with an additional capacity constraint. (Theorem 5.2)
Nested Logit (MNL)	• Hardness result for TU constraint structures (Corollary 5.2) • Hardness result for certain parameter settings ($v_{i0} \neq 0$) (Corollary 5.3) • FPTAS for TU constraint structures ($v_{i0} = 0$). (Section 5.3.2) • Joint assortment and display optimization problem polynomially solvable under a mild assumption. (Section 5.3.3)

Table 5.1: Summary of contributions for static assortment optimization.

with TU constraints is NP-hard. For the joint display and assortment optimization problem, we show that under special settings the problem allows for an exact solution in polynomial time.

5.1 Assortment Optimization Under MNL with TU Constraints

In this section, we consider the assortment optimization problem with TU constraints under the MNL choice model. In particular, we consider the following optimization problem.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} \frac{r_i v_i x_i}{v_0 + \sum_{j=1}^{n} v_j x_j} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \in \{0,1\}^n, \end{aligned} \tag{5.1}$$

First, we show that a natural LP relaxation for the above problem is tight. We will then consider an arbitrary additional constraint to the set of TU constraints such that the resulting set of constraints are not and present a Polynomial Time Approximation Scheme (PTAS) for the assortment optimization problem (5.1) under this more general set of constraints where for any $0 < \epsilon < 1$, we obtain a solution with objective value at least $(1 - \epsilon)$ times the optimal in running time polynomial in the input size for a fixed ϵ.

Remark 5.1. We would like to note that, Davis et al. [17] also use the LP relaxation to show that the assortment optimization under TU constraints can be solved optimally under the MNL choice model. However, they do not explicitly analyze the structure of extreme points of the LP relaxation. Here, we show that an extreme point optimal solution for the LP relaxation is "integral" and therefore, gives an optimal solution for the assortment optimization problem (**??**) under TU constraints for the MNL model. This structural property of the extreme points of the LP relaxation allows us to obtain near-optimal solutions for more general set of constraints that we discuss below.

5.1.1 Assortment Optimization: LP relaxation

In this section, we present a LP relaxation for (5.1) and show that the formulation is tight. Let

$$p_0 = \frac{1}{v_0 + \sum_{j=1}^{n} v_j x_j}, \quad p_i = x_i p_0, \ \forall i = 1, \ldots, n.$$

We can reformulate (5.1) as follows.

$$\begin{aligned} \underset{(\mathbf{p},p_0)}{\text{maximize}} \quad & \sum_{i=0}^{n} r_i v_i p_i \\ \text{subject to} \quad & \mathbf{Ap} \leq p_0 \mathbf{b} \\ & \sum_{i=0}^{n} v_i p_i = 1 \\ & p_i \in \{0, p_0\} \ \forall \ i \in \{1, 2, \cdots, n\} \\ & p_0 \geq 0. \end{aligned} \tag{5.2}$$

Note that (5.2) is an exact reformulation of (5.1). It can be easily reformulated as a mixed integer program using binary variables as follows.

$$\begin{aligned} & 0 \leq p_i \leq x_i p_0 \ \ \forall \ i \in \{1, 2, \cdots, n\} \\ & p_i + (1 - x_i) \geq p_0 \ \ \forall \ i \in \{1, 2, \cdots, n\} \\ & x_i \in \{0, 1\} \ \ \forall \ i \in \{1, 2, \cdots, n\}. \end{aligned} \tag{5.3}$$

5.1.2 Tightness of the LP relaxation

We consider the following LP relaxation for (5.2).

$$\begin{aligned} z_{\mathsf{LP}} \ = \ \max_{(\mathbf{p},p_0)} & \sum_{i=1}^{n} r_i v_i p_i \\ & \mathbf{Ap} \leq p_0 \mathbf{b} \\ & \sum_{j=0}^{n} v_j p_j = 1 \\ & 0 \leq p_i \leq p_0, \ \forall i = 1, \ldots, n, \\ & p_0 \geq 0. \end{aligned} \tag{5.4}$$

where we relax the constraints $p_i \in \{0, p_0\}$ to $0 \leq p_i \leq p_0$ for all $i = 1, \ldots, n$. Let $\mathcal{P}$ be the polytope defined by the constraints in (5.4), i.e.,

$$\mathcal{P} = \left\{ (\mathbf{p}, p_0) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \mathbf{Ap} \leq p_0 \mathbf{b}, \ \mathbf{v}'\mathbf{p} + v_0 p_0 = 1, \quad 0 \leq p_i \leq p_0, \forall i \right\}. \tag{5.5}$$

We show that all extreme points of $\mathcal{P}$ are "integral". We say that an extreme point $(\mathbf{p}, p_0) \in \mathcal{P}$ is *integral* if $p_i \in \{0, 1\}$ for all $i = 1, \ldots, n$ and *fractional* otherwise.

Theorem 5.1. *For any extreme point* $(\mathbf{p}, p_0) \in \mathcal{P}$, $p_i \in \{0, p_0\}$ *for all* $i = 1, \ldots, n$.

We will prove Theorem 5.1 by establishing a correspondence between extreme points of $\mathcal{P}$ and $\mathcal{Q}$, where

$$\mathcal{Q} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b},\ 0 \leq x_i \leq 1 \ \text{ for all } i = 1, 2, \ldots, n\},$$

is the polytope corresponding to relaxed constraints of the optimization problem (5.1).

Lemma 5.1. *If* $(\mathbf{p}, p_0)$ *is an extreme point of* $\mathcal{P}$, *then* $\mathbf{x} = \frac{\mathbf{p}}{p_0}$ *is an extreme point of* $\mathcal{Q}$. *Conversely, if* $\mathbf{x}$ *is an extreme point of* $\mathcal{Q}$, *then* $(\mathbf{p}, p_0)$ *where*

$$p_0 = \frac{1}{(v_0 + \mathbf{v}'\mathbf{x})},\ \mathbf{p} = p_0\mathbf{x}$$

is an extreme point of $\mathcal{P}$.

Theorem 5.1 follows from Lemma 5.1 and the fact that any extreme point $\mathbf{x}$ of $\mathcal{Q}$ is integral, i.e. $x_i \in \{0, 1\}$. Theorem 5.1 proves that any extreme point optimal solution of LP relaxation (5.4) is the same as the MIP reformulation (5.2) and hence it suffices to solve the relaxation. We defer the proof of Lemma 5.1 to Appendix.

5.1.3 Example of TU Constraints

Here, we present an important application of assortment planning under the TU constraints, namely the joint display and assortment optimization problem under the TU model for the MNL choice model. This problem arises in retailing and online advertising where the display slot of the product/ad affects the choice probability. In particular, we consider a model with m display segments and each segment has an upper bound on the number of products that can be displayed.

Let n be the total number of products and m be the number of display segments. There is a bound K_j on the number of products in display segment j for all $j \in [m]$. We assume that every product can only be displayed in at most one display segments. Let $x_{ij} \in \{0,1\}$ denote whether we offer product i in display segment j. For any product i, let r_i denote the revenue and v_{ij} denote the attraction parameter in display segment j. Now, the expected revenue optimization problem can be formulated as:

$$\begin{aligned}
\underset{\mathbf{X}\in\{0,1\}^{n\times m}}{\text{maximize}} \quad & R(\mathbf{X}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} r_i v_{ij} x_{ij}}{v_0 + \sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij} x_{ij}} \\
\text{subject to} \quad & \mathbf{C_i}: \quad \sum_{j=1}^{m} x_{ij} \le 1,\ i = 1,\dots,n \\
& \mathbf{C_j}: \quad \sum_{i=1}^{n} x_{ij} \le K_j,\ j = 1,\dots,m \\
& \quad\ x_{ij} \in \{0,1\},\ \ i = 1,\dots,n,\ j = 1,\dots,m.
\end{aligned} \tag{5.6}$$

Constraints $\{\mathbf{C_i}\}$ enforce that every product can be displayed only in one of the display segments, while constraints $\{\mathbf{C_j}\}$ enforce the cardinality constraints in each segment. The constraints in problem (5.6) are identical to the constraints in a transportation problem and hence are TU.

5.1.4 Extension to More General Constraints

We will now consider a more general variant of the assortment optimization problem (5.1), where constraints are not necessarily TU. In particular, we consider the following problem where we have a set of TU constraints and one additional constraint such that the overall constraints are not TU:

$$\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{n} \frac{v_i r_i x_i}{v_0 + \sum_{j=1}^{n} v_j x_j} \\
\text{subject to} \quad & \mathbf{A}\mathbf{x} \le \mathbf{b} \\
& \boldsymbol{\alpha}^T \mathbf{x} \le \gamma \\
& \mathbf{x} \in \{0,1\}^n,
\end{aligned} \tag{5.7}$$

where $\mathbf{A}$ is a $\{0,1\}^{m\times n}$ TU matrix, $\mathbf{b} \in Z^m$, $v_i \geq 0$ and $\alpha_i \geq 0$ for all i. Let

$$\mathcal{Q} = \{\mathbf{x} \mid A\mathbf{x} \leq \mathbf{b},\ 0 \leq x_i \leq 1 \ \text{ for all } i = 1, 2, \ldots, n\}$$
$$\hat{\mathcal{Q}} = \left\{\mathbf{x} \in \mathcal{Q} \ \middle|\ \boldsymbol{\alpha}^T\mathbf{x} \leq \gamma\right\},$$

be the polytopes corresponding to the relaxations of (5.1) and (5.7) respectively.

Similar to our approach in Section 5.1.1, we consider the following LP relaxation for (5.7),

$$\begin{aligned} \underset{(\mathbf{p},p_0)}{\text{maximize}} \quad & \sum_{i=0}^{n} a_i p_i \\ \text{subject to} \quad & (\mathbf{p}, p_0) \in \mathcal{P} \\ & \boldsymbol{\alpha}^T\mathbf{p} \leq p_0\gamma. \end{aligned} \tag{5.8}$$

where $\mathcal{P}$ is the polytope corresponding to the LP relaxation of the optimization problem (5.1) as defined in (5.5). Let

$$\hat{\mathcal{P}} = \left\{(\mathbf{p}, p_0) \in \mathcal{P} \ \middle|\ \boldsymbol{\alpha}^T\mathbf{p} \leq p_0\gamma \ \right\},$$

be the polytope corresponding to the LP relaxation of (5.7). Since constraints in (5.7) are not TU, the LP relaxation (5.8) may not be tight. In this section, we present a polynomial time approximation scheme (PTAS) for (5.7) under certain assumptions on $\mathcal{Q}$. In other words, for a fixed ϵ, we compute a $(1-\epsilon)$-approximation for (5.7) in running time polynomial in the input but exponential in $1/\epsilon$. Our PTAS is based on the following structure of extreme points of (5.8).

Observe that the polytope $\hat{\mathcal{Q}}$ (respectively $\hat{\mathcal{P}}$) is the intersection of the polytope $\mathcal{Q}$ (respectively $\mathcal{P}$) and the hyperplane $\boldsymbol{\alpha}^T\mathbf{x} \leq \gamma$ (respectively $\boldsymbol{\alpha}^T\mathbf{p} \leq p_0\gamma$). Hence, any extreme point of $\hat{\mathcal{Q}}$ (respectively $\hat{\mathcal{P}}$) is either an extreme point of $\mathcal{Q}$ (respectively $\mathcal{P}$) or a convex combination of two adjacent extreme points of $\mathcal{Q}$ (respectively $\mathcal{P}$). Therefore, if any two adjacent extreme points of $\mathcal{Q}$ "differ" only in a small number of components, then the number of "fractional components" in any extreme point of $\hat{\mathcal{Q}}$ and $\hat{\mathcal{P}}$ is small. Therefore, we can obtain an approximate solution for (5.7) by

ignoring the small number of "fractional components" from the optimal solution of (5.8) after appropriate pruning.

More specifically, for any two extreme points $\mathbf{x}_1, \mathbf{x}_2$ of $\mathcal{Q}$, let

$$d(\mathbf{x}_1, \mathbf{x}_2) = |\{i \mid x_{1i} \neq x_{2i}\}|$$
$$d(\mathcal{Q}) = \max\{d(\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_1, \mathbf{x}_2 \text{ are adjacent extr pts of } \mathcal{Q}\}.$$

Here $d(\mathcal{Q})$ denotes the maximum number of components by which the two adjacent extreme points of $\mathcal{Q}$ can differ. If $d(\mathcal{Q}) \leq \ell$, then the number of fractional components for any extreme point of $\hat{\mathcal{Q}}$ is at most ℓ. From Lemma 5.1, we know that there is a correspondence between extreme points of $\mathcal{P}$ and $\mathcal{Q}$. A similar correspondence also holds for extreme points of $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$. Hence, the number of "fractional components" in any extreme point of $\hat{\mathcal{P}}$ is also bounded by ℓ. In particular, for any extreme point $(\mathbf{p}, p_0)$ of $\hat{\mathcal{P}}$, let

$$\mathcal{F}((\mathbf{p}, p_0)) = \{i \geq 1 \mid 0 < p_i < p_0\},$$

denote the set of fractional components in $(\mathbf{p}, p_0)$. We have the following result,

Corollary 5.1. *If $d(\mathcal{Q}) \leq \ell$, then the number of fractional components for any extreme point $(\mathbf{p}, p_0)$ of $\hat{\mathcal{P}}$ is bounded by ℓ, i.e. $|\mathcal{F}((\mathbf{p}, p_0))| \leq \ell$.*

PTAS when $d(\mathcal{Q})$ is constant. Now we will present a PTAS for the case when $d(\mathcal{Q})$ is a constant (say ℓ). From Lemma 5.1, we know that optimality (feasibility) of $(\mathbf{p}, p_0)$ is equivalent to optimality (feasibility) of $\mathbf{x} = \mathbf{p}/p_0$ for (5.7). From Corollary 5.1, we know that any extreme point to (5.8) has at most ℓ fractional variables as $d(\mathcal{Q}) = \ell$. A simple idea to construct a feasible solution for (5.7) from an optimal solution of (5.8) is to ignore the "fractional variables". In particular, let $(\mathbf{p}, p_0) \in \hat{\mathcal{P}}$ be an optimal extreme point of (5.8). Define $(\hat{\mathbf{p}}, \hat{p}_0)$ as

$$\hat{p}_i = \begin{cases} 0 & \text{if } p_i < p_0 \\ \hat{p}_0 & \text{otherwise} \end{cases}$$

where

$$\hat{p}_0 = \frac{1}{v_0 + \sum\limits_{i:\hat{p}_i \neq 0} v_i}.$$

Observe that we ignore at most ℓ variables of $(\mathbf{p}, p_0)$. If the contribution of these variables to the objective value is small, then the total decrease in objective value is also bounded. Let R^* denote the optimal objective value of (5.7). If

$$a_i p_i \leq \frac{\epsilon}{\ell} R^* \quad \forall\ i:\ p_i < p_0, \tag{5.9}$$

then

$$\sum_{i \in \mathcal{F}(\mathbf{p}, p_0)} a_i p_i \leq \epsilon R^*,$$

which implies

$$(1-\epsilon) R^* \leq \sum_{i \notin \mathcal{F}(\mathbf{p}, p_0)} a_i p_i = \sum_i a_i \hat{p}_i,$$

and $(\hat{\mathbf{p}}, \hat{p}_0)$ is a $(1-\epsilon)$-approximate solution for (5.7). Note that in $(\mathbf{p}, p_0)$ there can be at most $\left\lfloor \frac{\ell}{\epsilon} \right\rfloor$ variables such that $a_i p_i > \frac{\epsilon}{\ell} R^*$. Therefore, to ensure (5.9) we guess the top $\left\lfloor \frac{\ell}{\epsilon} \right\rfloor$ variables contributing to the objective in (5.7), set those variables $p_i = p_0$ and solve the resulting linear program. The running time is exponential in $\left\lfloor \frac{\ell}{\epsilon} \right\rfloor$. We provide the details in Algorithm 10 and Theorem 5.2 establishes its correctness.

Theorem 5.2. *Let $d(\mathcal{Q}) \leq \ell$ and $(\hat{\mathbf{p}}, \hat{p}_0)$ be the solution obtained by Algorithm 10. Then $\sum\limits_{i=0}^{n} a_i \hat{p}_i > (1-\epsilon) R^*$, where R^* is the optimal value of* (5.7).

Examples of $\mathcal{Q}$ with small $d(\mathcal{Q})$: The polytope $\mathcal{Q}$ corresponding to the feasible region of cardinality constrained joint assortment and display optimization problem (5.6) is

$$\mathcal{Q} = \left\{ \mathbf{X} \;\middle|\; \sum_{j=1}^{m} x_{ij} \leq 1\ \forall i,\ \sum_{i=1}^{n} x_{ij} \leq K_j\ \forall j,\ 0 \leq x_{ij} \leq 1\ ,\ \forall\ i,j \right\}.$$

The constraints in problem (5.6) are the same as the transportation problem, the number of variables that are different in two adjacent extreme points of the LP

Algorithm 10 PTAS for (5.7)

Set $\mathcal{S} = \left\{ S_t \subset \{1, 2, \ldots, n\} \;\middle|\; |S_t| \leq \left\lfloor \frac{\ell}{\epsilon} \right\rfloor \right\}$.
for $S_t \in \mathcal{S}$ **do**
  if $|S_t| < \left\lfloor \frac{\ell}{\epsilon} \right\rfloor$ **then**
    Obtain $(\hat{\mathbf{p}}_t, \hat{p}_{t0})$ as follows: $\hat{p}_{t0} = \dfrac{1}{v_0 + \sum_{i \in S_t} v_i}$

$$\hat{p}_{ti} = \begin{cases} \hat{p}_{t0} & \text{if } i \in S_t \\ 0 & \text{otherwise} \end{cases}$$

  end if
  if $(\hat{\mathbf{p}}_t, \hat{p}_{t0})$ is feasible in (5.7) **then** Set $R_t = \sum_{i=0}^{n} a_i \hat{p}_{ti}$
  else Set $Q_t = \{i \in \{1, 2, \ldots, n\} \mid i \notin S_t \text{ and } \exists\ j \in S_t \text{ such that } a_j \leq a_i\}$
    Consider modified (5.8), $z_{\mathsf{LP}}(t)$ with additional constraints

$$p_i = p_0,\ \forall i \in S_t$$
$$p_i = 0,\ \forall i \in Q_t$$

    if $z_{\mathsf{LP}}(t)$ is feasible **then** Set $(\mathbf{p}_t^*, p_{t0}^*)$ as the optimal extreme point of $z_{\mathsf{LP}}(t)$.
      Set $\hat{S}_t = \{i \mid p_{ti}^* = p_{t0}^*\}$
      Obtain $(\hat{\mathbf{p}}_t, p_{t0})$ as follows: $\hat{p}_{t0} = \dfrac{1}{v_0 + \sum_{i \in S_t} v_i}$

$$\hat{p}_{ti} = \begin{cases} \hat{p}_{t0} & \text{if } i \in \hat{S}_t \\ 0 & \text{otherwise} \end{cases}$$

      Set $R_t = \sum_{i=0}^{n} a_i \hat{p}_{ti}$
    end if
  end if
end for
Set $t^* = \arg\max_t R_t$;
Output $(\hat{\mathbf{p}}, \hat{p}_0) = (\hat{\mathbf{p}}_{t^*}, \hat{p}_{t^*0})$

relaxation of problem (5.6) is bounded by the maximum cycle length in the corresponding transportation network. Since the transportation network is a bipartite graph, the maximum cycle length cannot exceed twice the number of nodes in either of the partitions. Hence, we have the following result,

Lemma 5.2. *For $\mathcal{Q}$ corresponding to feasible region of cardinality constrained joint assortment and display optimization problem* (5.6)*, we have $d(\mathcal{Q}) \leq 2m$, where m is the number of display segments.*

Theorem-5.2 and the Lemma 5.2 establishes that there exists a PTAS for the joint assortment and display optimization problem under the MNL choice model in the presence of an additional constraint.

A Computational Study: Here, we study the computational performance of our PTAS algorithm for rational optimization over a TU constraint set with one additional constraint. In particular, we consider the joint assortment and display optimization problem with both cardinality and capacity constraints. Each item has capacity c_i and there is a bound C on the total capacity of items selected. The problem formulation is shown below.

$$\begin{aligned} \underset{\mathbf{X}\in\{0,1\}^{n\times m}}{\text{maximize}} \quad & R(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{C_i}: \sum_{j=1}^{m} x_{ij} \leq 1,\ \forall i\ ;\ \mathbf{C_j}: \sum_{j=1}^{m} x_{ij} \leq K_j; \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} c_i x_{ij} \leq C\ ;\quad x_{ij} \in \{0,1\}\ \forall\ i,\ j \end{aligned} \tag{5.10}$$

We would like to note that (5.10) is NP hard even for $m = 1$ ([20]). Algorithm 10 gives a PTAS for the above problem when the number of display m is a constant.

To evaluate the performance of our PTAS algorithm we perform 5 experiments by varying the number of products ($n \in \{10, 50, 100\}$) and the number of display segments ($m \in \{2, 3\}$). For each experiment, we generate 10 random instances of problem (5.10). The parameters $\mathbf{v}, \mathbf{c}$ and $\mathbf{r}$ are chosen as uniform random numbers

products	segments	Optimality Ratio	Time for PTAS(secs)
10	2	0.9408	0.653
50	2	0.996	261.876
50	3	0.947	3606.466
100	2	0.994	2886.35
100	3	0.869	3648.761

Table 5.2: PTAS Performance for different number of products (n) and display segments (m).

between 0 and 1, as the scale of these parameters does not change the optimal solution. For every instance, we solve the corresponding LP relaxation and implement a slightly modified version of the PTAS algorithm. All implementations have been done using Gurobi libraries in C++. In the modified version of PTAS, we enforce a time limit on the running time of the algorithm. Specifically, we restrict the time spent in guessing the top variables (steps 8-9 in Algorithm 10) to one hour. Although Lemma 5.2 bounds the number of fractional variables to $2m$, based on empirical observations, we relaxed the bound to m in order to decrease the number of computations. Hence, we only considered subsets of size not exceeding $\left\lfloor \frac{m}{\epsilon} \right\rfloor$ instead of the theoretically correct $\left\lfloor \frac{2m}{\epsilon} \right\rfloor$. To avoid trivial cases, the value of the capacity bound C is appropriately chosen to ensure that the additional capacity constraint is tight and the optimal solution of LP relaxation has atleast $\lfloor m/\epsilon \rfloor$ positive components.

Table 5.2 summarizes performance for our PTAS approach. For each experiment, we report two quantities of interest namely i) the average ratio of objective values obtained by the PTAS method and the LP solution (z_{PTAS}/z_{LP}) and ii) the average running time of the PTAS method. It is important to note that the LP solution (i.e. optimal solution to LP relaxation of (5.10)) is clearly an upper bound to the optimal solution to (5.10) itself and hence the ratio z_{PTAS}/z_{LP} is a conservative measure of PTAS performance. Even though we fixed $\epsilon = 0.8$, which theoretically guarantees only a 0.2-approximation, the approximate optimal value is on an average about 85%

of the optimal value. This suggests that one can use a higher value of ϵ to avoid large computations and still obtain a good approximation.

5.2 Nested Logit Model

In this section, we describe the Nested Logit Model (NL). In the NL model the products are assumed to be partitioned into different nests. Each nest has a certain number of products and it is assumed that the consumer first selects a nest and then selects a product within the nest according to a MNL model. More specifically, assume the products are partitioned into m nests and each nest has n products [1]. Here, an assortment S of products typically refers to a $m-$tuple $(S_1, \cdots, S_m)$, where S_i is the assortment or subset of products offered in nest i. In the NL model, every product j in nest i is associated with a parameter v_{i_j}. These parameters are similar to the utility parameters associated with every product in the MNL model. In addition to these parameters, every nest i is associated with two additional parameters, γ_i and v_{i0}. γ_i is the dissimilarity parameter that indicates the strength of correlation for demand of products within a nest and v_{i0} is the parameter corresponding to the outside option after a consumer has selected nest i. Finally, the parameter v_0 represents the outside option to indicate the setting where a consumer does not chose to explore any nest. For brevity, we will assume that for every nest i,

$$V(S_i) = v_{i0} + \sum_{j \in S_\ell} v_{ij}. \tag{5.11}$$

The choice probabilities of a consumer selecting product j from nest i is given as

$$\pi_{\mathsf{NL}}(j, S_1, \cdots, S_m) = \frac{V^{\gamma_i}(S_i)}{v_0 + \sum_{k=1}^{m} V^{\gamma_k}(S_k)} \cdot \frac{v_{i_j}}{v_{i0} + \sum_{\ell \in S_i} v_{i\ell}} \tag{5.12}$$

The first term on the right hand side of (5.12) is the probability that a consumer selects nest i when offered assortments $S_1, \cdots, S_m$ and the second term in (5.12) is

[1] we assume that every nest has same number of products for ease of exposition, one can easily generalize the framework to handle each nest having different number of products

the probability that a consumer selects product j in nest i given that the consumer has already selected nest i.

Williams [47] has showed that when the parameters satisfy the conditions $\gamma_i \leq 1$ for all the nests i, then the Nested Logit Model can be modeled within the random utility framework. In particular, [47] shows that utility of a product j in nest i can be decomposed as,

$$U_{i_j} = \mu_{i_j} + \epsilon_{i_j} + \xi_i, \tag{5.13}$$

where μ_{i_j} is the mean utility for product i in nest j, while ϵ_{i_j} are i.i.d random variables having a Gumbel distribution with location and scale parameters 0 and 1 respectively that represents the idiosyncrasies of consumer with regard to the product j. ξ_i represents the idiosyncrasies of consumer with regard to the nests and is distributed in such a way that $\max_{j \in S_i} \mu_{i_j} + \epsilon_{i_j} + \xi_i$ is a Gumbel distribution with scale γ_i. By substituting $v_{i_j} = e^{\mu_{i_j}}$, we obtain the choice probabilities specified in (5.12). It is also to see that when $m = 1$ and $\gamma_i = 1$, the choice probabilities are same as the MNL model, indicating that Nested Logit model generalizes the MNL model. Recently, [18] showed that the assortment planning problem under the NL model is NP-hard for the settings when the NL model cannot be formulated within the random utility framework, i.e. settings when $v_{i,0} \neq 0$ or $\gamma_i > 1$ for some nest i. [2] Therefore in this chapter, we restrict ourselves to the settings when $v_{i0} = 0$ and $\gamma_i \leq 1$.

5.3 Assortment Optimization Under NL with TU Constraints

In this section, we consider the assortment optimization problem with TU constraints under the NL choice model. In particular, we consider the following optimization

[2]The proof of hardness discussed in [18] for the case when $v_{i,0} \neq 0$ is incomplete, as the authors in their reduction to a partition problem assume square root can be computed exactly in polynomial time. We build on their reduction technique to close the gap in their proof.

problem.

$$\begin{aligned} \text{maximize} \quad & \Pi(\mathbf{x}_1, \cdots, \mathbf{x}_m) = \sum_{i=1}^{m} \frac{V_i(\mathbf{x}_i)^{\gamma_i}}{v_0 + \sum_{j=1}^{m} V_j(\mathbf{x}_j)^{\gamma_j}} R_i(\mathbf{x}_i) \\ \text{subject to} \quad & \mathbf{A}_i \mathbf{x}_i \leq \mathbf{b}_i \ \ \forall \ \ i \in \{1, \cdots, m\} \\ & \mathbf{x}_i \in \{0,1\}^n \ \ \forall \ \ i \in \{1, \cdots, m\}, \end{aligned} \tag{5.14}$$

where $\mathbf{A}_i$ is a Totally Unimodular matrix corresponding to the nest i and x_i is the the incidence vector for assortment $S_i \subseteq \{1, \ldots, N\}$, i.e., $x_j(S_i) = 1$ if product $j \in S_i$ and 0 otherwise. In this section we will interchangeably refer to the assortment by the subset S_i or it's incident vector $\mathbf{x}_i$. In a slight of abuse of notation, at places we will refer to the assortment S_i by its incident vector $\mathbf{x}_i$ and at places by the subset S_i.

Here, we first establish that the assortment optimization problem (5.14) is NP-hard. We will then provide a Fully Polynomial Time Approximation Scheme (FPTAS) for this problem. Finally, we will the specific application of (5.14), namely, the joint display and assortment optimization problem and present a polynomial time algorithm making a mild assumption on the parameters of the NL-model.

5.3.1 Hardness Result

We show that the general version of problem (5.14) is NP-hard.
Let $(S_1^*, \cdots, S_m^*)$ be the optimal solution to problem (5.14). Problem (**??**) is in P if for any $\epsilon > 0$, there is an algorithm that computes a solution $(S_1(\epsilon), \cdots, S_m(\epsilon))$ with

$$\Pi(S_1^*, \cdots, S_m^*) - \Pi(S_1(\epsilon), \cdots, S_m(\epsilon)) \leq \epsilon \tag{5.15}$$

in time polynomial in the input size and $\log(1/\epsilon)$.

Based on the above definition, we show that there is no polynomial time algorithm for the assortment optimization problem under the general nested logit model. As

in Davis et al. [18], we consider the reduction from the partition problem to the assortment optimization problem under a nested logit model. Consider the following instance, $\mathcal{I}$ of the partition problem given by:

$$c_1, \ldots, c_n \in \mathbb{Z}_+, \ \sum_{j=1}^{n} c_j = 2T. \tag{5.16}$$

The instance, $\mathcal{I}'$ for the nested logit problem is constructed as follows: There are two nests. The preference weight for the option of not choosing any of the nests is $v_0 = 0$. The dissimilarity parameters of the two nests are $\gamma_1 = \gamma_2 = \frac{1}{2}$. The first nest N_1 has two products. The revenue and utility parameters associated with the two products are respectively $r_{11} = 0$, $v_{11} = 2$ and $r_{12} = 2(T+1)(T+3)$ and $v_{12} = 2(2T+1)$. The second nest has $n+1$ products. The revenue and utility parameter associated with the first product is $r_{21} = 0$ and $v_{21} = 1$. The revenues of the other products in the second nest are identical and they are given by $r_{2j} = (T+1)(2T+1) \forall j = 2, \cdots, n+1$. The utility parameters for the other products in the second nest are given by $v_{2j} = c_j \ \forall j = 2, \cdots, n+1$. The retailer here is constrained to offer the first product in both the nests, i.e. the constraints are given by $x_{11} = 1$ and $x_{21} = 0$.

We prove the following theorem. We defer the proof of Theorem to Appendix.

Theorem 5.3. *For any*

$$0 < \epsilon < \frac{2T+1}{(6T+3)\,(3T+2)^2},$$

there exist an assortment (S_1, S_2) *such that* $\mathbf{x}_1(S_1) = 1$ *and* $\mathbf{x}_1(S_2) = 1$ *for instance* $\mathcal{I}'$ *with* $\Pi(S_1, S_2) \geq (T+2)(2T+1) - \epsilon$ *if and only if instance* $\mathcal{I}$ *has a partition.*

As a direct consequence of Theorem 5.3, we have the following two results.

Corollary 5.2. *Assortment optimization problem (*(**??**)*) with TU constraints under the NL model is NP-hard.*

Corollary 5.3. *If we allow the utility parameters of the no purchase options within the nests to take on strictly positive values, then the assortment feasibility problem is NP-hard.*

5.3.2 FPTAS for Assortment Optimization with TU constraints

In this section, we focus on the setting where $v_{i0} = 0$ for all the nests i and consider a class of TU constraints. We present fully polynomial time approximation scheme (FPTAS) for the assortment optimization problem for this setting. In particular, if not offering any product is a feasible assortment, then our algorithm computes a $(1-\epsilon)$ approximation of the optimal assortment in time polynomial in the input size and $1/\epsilon$.

[23] presented a linear programming reformulation for the assortment optimization problem (5.14). We build our FPTAS on the LP reformulation.

Theorem 5.4 ([23])**.** *The assortment optimization problem* (5.14) *is equivalent to the following linear program*

$$\begin{aligned} &\underset{(\mathbf{y},z)}{\textit{minimize}} \quad z \\ &\textit{subject to} \quad z \geq \sum_{i=1}^{m} y_i \\ &\qquad\qquad y_i \geq V(\mathbf{x}_i)^{\gamma_i}(R(\mathbf{x}_i) - z) \;\forall\; \mathbf{x}_i \in \{0,1\}^n \textit{ such that } \mathbf{A}_i\mathbf{x}_i \leq \mathbf{b}_i \end{aligned} \tag{5.17}$$

It should be noted that the linear program (5.17) has exponential number of constraints. Consider the following separation problem.

Separation Problem: For a given $(z, y_1, \cdots, y_m)$, for each nest $i \in M$ decide whether

$$y_i \geq V(\mathbf{x}_i)^{\gamma_i}(R(\mathbf{x}_i) - z) \;\forall\; \mathbf{x}_i \in \{0,1\}^n \text{ such th-at } \mathbf{A}_i\mathbf{x}_i \leq \mathbf{b}_i \tag{5.18}$$

or find an $\mathbf{x}_i \in \{0,1\}^n$ such that $\mathbf{A}_i\mathbf{x}_i \leq \mathbf{b}_i$ and $y_i < V(\mathbf{x}_i)^{\gamma_i}(R(\mathbf{x}_i) - z)$. It is well established in linear optimization that if we can solve the separation problem in polynomial time, then the linear program (5.17) can also be solved in polynomial time. Hence, we focus our efforts on the following optimization problem, for a given z

$$\begin{aligned} \underset{\mathbf{x}_i}{\text{maximize}} \quad & g_i(\mathbf{x}_i) \stackrel{\triangle}{=} V(\mathbf{x}_i)^{\gamma_i}(R(\mathbf{x}_i) - z) \\ \text{subject to} \quad & \mathbf{A}_i\mathbf{x}_i \leq \mathbf{b}_i \\ & \mathbf{x}_i \in \{0,1\}^n \end{aligned} \tag{5.19}$$

Given z, the optimization problem (5.19) is sub-problem corresponding to the individual nests. Our FPTAS approach is based on the idea that if we can solve the sub-problem approximately, then the linear program reformulation (5.17) of the assortment optimization (5.14) can be solved approximately.

Solving the Sub-problem approximately: We present a key property of $g_i(\mathbf{x})$ that would be helpful in building our FPTAS for the assortment optimization problem.

Lemma 5.3. *$g_i(\mathbf{x})$ is a quasi-convex function over $\mathcal{Q}_i = \{\mathbf{x} \in \mathcal{R}^n | g_i(\mathbf{x}) \geq 0\}$.*

Proof. We have $g_i(\mathbf{x}) = \frac{\sum_{j=1}^n v_{ij}(r_{ij}-z)x_j}{\left(\sum_{j=1}^n v_{ij}x_j\right)^{1-\gamma_i}}$. Observe that $\sum_{j=1}^n v_{ij}r_{ij}x_j$ is a linear function, while $\left(\sum_{j=1}^n v_{ij}s_j\right)^{1-\gamma_i}$ is a concave function in $\mathbf{x}$. Since $g_i(\mathbf{x})$ is non-negative over $\mathcal{Q}_i$ and is a ratio of a linear function and a concave function, $g_i(\mathbf{x})$ is quasi-convex over $\mathcal{P}_i$. □

If we assume that not offering any product is a feasible assortment, then we can assume without loss of generality that $g_i(\mathbf{x}_i) \geq 0$. From Lemma 5.3, we have that whenever $g_i(\mathbf{x}_i) \geq 0$, we have $g_i(\mathbf{x}_i)$ as a quasi-convex function. Hence, for the purpose of optimizing (5.19), $g_i(\mathbf{x})$ is a quasi-convex function. [35] developed FPTAS based on the work of [21] for minimizing low rank quasi-concave functions over combinatorial sets. Noting that problem (5.19) is maximizing a quasi-convex function, we develop our FPTAS on the work of [35].

We introduce certain concepts and establish results that would lead us to solving the sub-problem (5.19) approximately.

Let $\mathcal{P}_i$ be the polytope corresponding to the relaxation of (5.19), i.e.

$$\mathcal{P}_i = \{\mathbf{x} \in [0,1]^n \mid \mathbf{A}_i\mathbf{x} \leq \mathbf{b}_i\}. \tag{5.20}$$

Let $g_{i1}(\mathbf{x}) = \sum_{j\in N} v_{ij}(r_{ij} - z)s_j$, and $g_{i2}(\mathbf{x}) = \sum_{j\in N} v_{ij}s_j$ be the linear functions corresponding to the numerator and denominator of $g_i(\mathbf{x})$.

Definition 5.1. *For $\epsilon > 0$, an ϵ-convex Pareto set, denoted by $\mathcal{CP}_\epsilon$, is a set of solutions such that for all $\mathbf{x} \in \mathcal{P}_i$, there is $\mathbf{x}' \in Conv(\mathcal{CP}_\epsilon)$ such that*

$$g_{i1}(\mathbf{x}') \geq \frac{g_{i1}(\mathbf{x})}{1+\epsilon} \;\&\; g_{i2}(\mathbf{x}) \leq (1+\epsilon)g_{i2}(\mathbf{x})$$

By definition of the ϵ-convex Pareto set, there exists a $\mathbf{x} \in Conv(\mathcal{CP}_\epsilon)$ such that

$$\begin{array}{rcll} & & \underset{\mathbf{x}_i}{\text{maximize}} & g_i(\mathbf{x}_i) \\ (1+\epsilon)^{2-\gamma_i} g_i(\mathbf{x}) & \geq & \text{subject to} & \mathbf{A}_i\mathbf{x}_i \leq \mathbf{b}_i \\ & & & \mathbf{x}_i \in \{0,1\}^n \end{array}$$

Since, $g_i(\mathbf{x})$ is a quasi-convex function, the maximum value of $g_i(\mathbf{x})$ over a polytope occurs at an extreme point. Hence for all $\mathbf{x} \in Conv(\mathcal{CP}_\epsilon)$, there exists a $\mathbf{x}' \in \mathcal{CP}_\epsilon$ such that

$$g_i(\mathbf{x}') \geq g_i(\mathbf{x}).$$

Therefore, we have the following result which highlights the relation between $\mathcal{CP}_\epsilon$ and $(1+\epsilon)$ approximate solution to the sub-problem (5.19).

Theorem 5.5. *There exists $\mathbf{x} \in \mathcal{CP}_\epsilon$ such that $\mathbf{x}$ is an $(1+\epsilon)^{2-\gamma_i}$ approximation to the sub-problem (5.19).*

We provide the details on computing the ϵ-convex Pareto set and complete the proof of Theorem 5.5 in the Appendix.

5.3.3 Joint Assortment and Display Optimization Problem

Here we formulate the joint assortment and display optimization problem, where the retailer needs to select the subset of products to offer and also decide on the display segment when the customers choose according to an NL model. We will then present a polynomial time algorithm for this optimization under mild assumptions.

Let m be the total number of nests, n number of products in each segment and ℓ be the number of display segments. There is a bound N_{ik} on the number of products in display segment k for all $k \in [\ell]$. We assume that every product can only be displayed in at most one display segments. We use the matrix

$$\mathbf{X}_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1\ell} \\ \vdots & \vdots & \vdots \\ x_{in1} & \cdots & x_{in\ell} \end{bmatrix} \in \{0,1\}^{n\times\ell},$$

to denote the assortment of products that we offer in nest i and their display positions, where $x_{ijk} = 1$ if we offer product j in nest i at display slot k and $x_{ijk} = 0$ otherwise. Hence, If we offer the assortment $\mathbf{x}_i$ in nest i, the total preference weight associated with nest i as defined in (5.11) will be given by

$$V_i(\mathbf{x}_i) = \sum_{j=1}^{n}\sum_{k=1}^{\ell} v_{ijk}x_{ijk},$$

and the expected revenue, conditioned on the fact that a consumer decided to make a purchase from nest i is given by

$$R_i(\mathbf{x}_i) = \sum_{j=1}^{n}\sum_{k=1}^{\ell} \frac{v_{ijk}x_{ijk}}{V_i(\mathbf{x}_i)} r_{ij}.$$

Hence, if we offer the assortment $(\mathbf{x}_1, \cdots, \mathbf{x}_m)$, then the expected revenue is given by

$$\Pi(\mathbf{x}_1, \cdots, \mathbf{x}_m) = \sum_{i\in M} Q_i(\mathbf{x}_1, \cdots, \mathbf{x}_m) R_i(\mathbf{x}_i) = \sum_{i\in M} \frac{V_i(\mathbf{x}_i)^{\gamma_i}}{v_0 + \sum_{l\in M} V_l(\mathbf{x}_l)^{\gamma_l}} R_i(\mathbf{x}_i).$$

We are interested in finding the optimal assortment of products and their display positions in each nest such that the expected revenue is maximized and the number of products in a display segment is less than a specified bound. Specifically, we are interested in solving the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \Pi(\mathbf{x}_1, \cdots, \mathbf{x}_m) = \sum_{i=1}^{M} \frac{V_i(\mathbf{x}_i)^{\gamma_i}}{v_0 + \sum_{j=1}^{M} V_j(\mathbf{x}_j)^{\gamma_j}} R_i(\mathbf{x}_i) \\ \text{subject to} \quad & \sum_{k} x_{ijk} \leq 1 \ \ \forall\ j \in [n],\ i \in [m] \\ & \sum_{j} x_{ijk} \leq N_{ik} \ \ \forall\ k \in [\ell],\ i \in [m] \\ & \mathbf{x}_{ijk} \in \{0,1\} \ \ \forall\ j \in [n],\ k \in [\ell],\ i \in [m], \end{aligned} \tag{5.21}$$

where N_{ik} is the upper bound on the number of products allowed at display position k in nest i. The first constraint in (5.21) ensures that every product in each nest can be assigned to only one display position, while the second one enforces the cardinality constraints on display segments.

We will make the following assumption and present a polynomial time algorithm for (5.21).

Assumption 5.1. *The utility parameter v_{ijk} corresponding to product j of nest i when displayed at level k takes the following form,*

$$v_{ijk} = v_{ij}\lambda_{ik} \ \forall i \in \{1, \cdots, m\} \ \ j \in \{1, \cdots, n\} \ \ k \in \{1, \cdots, \ell\},$$

for some $\lambda_{ik} > 0$.

Remark 5.2. Note that Assumption 5.1 is common for choice models, where the mean utility of a product is a linear combination of the attributes. In particular, $\mu_{i_j} = \boldsymbol{\beta}_i \cdot \mathbf{f}_{ij}$, where $\mathbf{f}_{ij} \in \mathcal{R}^d$ is a vector of attributes of the product i in nest j, while $\boldsymbol{\beta}_i \in \mathcal{R}^d$ is the weight associated with each of these attributes for products in nest i. One of the product attributes that can potentially influence it's attractiveness

to a consumer is where it is displayed. Therefore, assuming display position as an attribute, Assumption 5.1 is a natural consequence of the linear utility model.

Solution concept: The joint assortment and display optimization problem (5.21) is a non-linear optimization over $\{0,1\}$ variables. We will use the linear programming formulation (see Theorem 5.4) to design a tractable algorithm for our problem. Note that the linear program (5.17) has exponential number of constraints. In the rest of this section, we will show that only polynomial number of constraints are sufficient to describe the exponential number of constraints. Let $\mathcal{P}_i$ be the feasible region of the optimization problem (5.21). In particular,

$$\mathcal{P}_i = \left\{ \mathbf{X}_i \in \{0,1\}^{n\times \ell} \;\middle|\; \sum_k x_{ijk} \leq 1 \;\forall\; j \in [n], \sum_j x_{ijk} \leq N_{ik} \;\; \forall\; k \in [\ell] \right\}$$

Definition 5.2. *A collection of assortments $\mathcal{T}_i$ is* ideal*, if*

- *$\mathcal{T}_i \subset \mathcal{P}_i$ and $|\mathcal{T}_i|$ is polynomial in n,*
- *for every z, there exists a $\hat{\mathbf{S}}_i \in \mathcal{T}_i$ such that*

$$V_i(\hat{\mathbf{S}}_i)^{\gamma_i}(R_i(\hat{\mathbf{S}}_i) - z) \geq V_i(\mathbf{S}_i)^{\gamma_i}(R_i(\mathbf{S}_i) - z) \text{ for all } \mathbf{S}_i \in \mathcal{P}_i.$$

We will prove the existence of *ideal* collection of assortments for each nest. Therefore, the linear program (5.17) is equivalent to the following linear program

$$\begin{aligned} \underset{(\mathbf{y},z)}{\text{minimize}} \quad & z \\ \text{subject to} \quad & z \geq \sum_{i=1}^{m} y_i \\ & y_i \geq V_i(\mathbf{S}_i)^{\gamma_i}(R(\mathbf{s}_i) - z) \;\forall\; \mathbf{S}_i \in \mathcal{T}_i \;\; \forall\; i = \{1, \cdots, m\}. \end{aligned} \tag{5.22}$$

By transforming the LP (5.17) into LP (5.22), we have shown that the exponential number of constraints in (5.17) are redundant and it is sufficient to consider only a polynomial number of constraints to solve the LP (5.17), thus enabling us to solve the linear program (5.17) in polynomial time. Our proof technique relies heavily on the following result established by [23],

Lemma 5.4 ([23]). *Fix i, for every z, there exists a u_i such that*

$$\arg\max_{\mathbf{S}_i \in \mathcal{P}_i} \; V_i(\mathbf{S}_i)^{\gamma_i}(R(\mathbf{s}_i) - z) \;=\; \arg\max_{\mathbf{S}_i \in \mathcal{P}_i} \; V_i(\mathbf{S}_i)(R(\mathbf{s}_i) - u_i).$$

Lemma 5.4 allows for the following interpretation of *ideal* collection of assortments,

Corollary 5.4. *A collection of assortments $\mathcal{T}_i$ is* ideal, *if*

- *$\mathcal{T}_i \subset \mathcal{P}_i$ and $|\mathcal{T}_i|$ is polynomial in n,*
- *for every u_i, there exists a $\hat{\mathbf{S}}_i \in \mathcal{T}_i$ such that*

$$V_i(\hat{\mathbf{S}}_i)(R_i(\hat{\mathbf{S}}_i) - u_i) \geq V_i(\mathbf{S}_i)(R_i(\mathbf{S}_i) - u_i) \text{ for all } \mathbf{S}_i \in \mathcal{P}_i.$$

Corollary 5.4 suggests that to prove the existence of *ideal* collection of assortments, it sufficies to show that the parametric optimization problem,

$$\max_{\mathbf{S}_i \in \mathcal{P}_i} \; V_i(\mathbf{S}_i)(R(\mathbf{s}_i) - u_i), \tag{5.23}$$

has polynomial number of optimal solutions. In the rest of this section, we prove the existence of *ideal* collection of assortments and show how to obtain an *ideal* collection of assortments.

Note that $V_i(\mathbf{S}_i)(R_i(\mathbf{S}_i) - u_i) = \sum_{j=1}^{n} \sum_{k=1}^{\ell} v_{ijk}(r_{ij} - u_i)x_{ijk}$. Therefore the parametric optimization problem (5.23) is equivalent to

$$\begin{aligned}
\text{maximize} \quad & \sum_{j \in [n]} \sum_{k \in [\ell]} v_{ijk}(r_{ij} - u_i)x_{ijk} \\
\text{subject to} \quad & \sum_{k} x_{ijk} \leq 1 \;\; \forall\, j \in [n] \\
& \sum_{j} x_{ijk} \leq N_{ik} \;\; \forall\, k \in [\ell] \\
& \mathbf{x}_{ijk} \in \{0, 1\} \;\; \forall\, j \in [n],\; k \in [\ell].
\end{aligned} \tag{5.24}$$

Consider the following LP relaxation of (5.24),

$$\begin{aligned} \text{maximize} \quad & \sum_{j\in[n]}\sum_{k\in[\ell]} v_{ijk}(r_{ij}-u_i)x_{ijk} \\ \text{subject to} \quad & \sum_{k} x_{ijk} \leq 1 \ \ \forall\ j \in [n] \\ & \sum_{j} x_{ijk} \leq N_{ik} \ \ \forall\ k \in [\ell] \\ & 0 \leq \mathbf{x}_{ijk} \leq 1 \ \ \forall\ j \in [n],\ k \in [\ell]. \end{aligned} \tag{5.25}$$

Since the constraints in the LP relaxation (5.25) are totally unimodular, we have a tight relaxation. Hence, to prove the existence of *ideal* collection of assortments, it suffices to prove that the parametric linear program (5.25) has polynomial number of optimal solutions.

"Ideal" collection of assortments: We show that it suffices to consider a polynomial number of values of u_i to find the set of optimal solutions to the parametric linear program (5.25). Define linear functions,

$$f_{ij}(u) = v_{ij}(r_{ij}-u),$$

where v_{ij} is as defined in Assumption (5.1). We show that just by considering the intersection points of any two linear functions $f_{ij}(u)$ and $f_{ij'}(u)$ and points where $f_{ij}(u)$ vanishes for some j is sufficient to obtain the set of optimal solutions for the parametric linear program (5.25). Algorithm 11 describes how to obtain an *ideal* collection of assortments. Consider the set U_i described in Algorithm 11, let $u_1, \cdots, u_t$ be the elements of U_i indexed in ascending order of values, i.e.

$$u_p < u_{p'}\ \forall\ 1 \leq p < p' \leq t,$$

where t is the number of elements in U_i. We will prove that for any $u_i \in [u_p, u_{p+1}]$, the optimal solution of the parametric linear program (5.25) remains the same. In particular, we have the following result

Algorithm 11 Obtaining *ideal* collection of assortments $\mathcal{T}_i$

Set $U^1 = \{u \mid f_{ij}(u) = 0 \text{ for some } j \in [n]\}$
Set $U^2 = \{u \mid f_{ij}(u) = f_{ij'}(u) \text{ for some } j \neq j' \text{ and } j, j' \in [n]\}$
Set $U_i = U^1 \cup U^2$
Set $\mathcal{T}_i = \phi$
for each $u \in U_i$ **do**
    Set $\mathbf{S}_i$ as the optimal solution of the linear program (5.25) with $u_i = u$, i.e.

$$\begin{aligned} \mathbf{S}_i \leftarrow \arg\max \quad & \sum_{j\in[n]} \sum_{k\in[\ell]} v_{ijk}(r_{ij} - u) s_{ijk} \\ \text{subject to} \quad & \sum_{k} s_{ijk} \leq 1 \quad \forall\, j \in [n] \\ & \sum_{j} s_{ijk} \leq N_{ik} \quad \forall\, k \in [\ell] \\ & 0 \leq \mathbf{s}_{ijk} \leq 1 \quad \forall\, j \in [n],\ k \in [\ell]. \end{aligned}$$

    $\mathcal{T}_i = \mathcal{T}_i \cup \{\mathbf{S}_i\}$
end for
Return $\mathcal{T}_i$

Lemma 5.5. *Fix $u \in [u_p, u_{p+1}]$, if*

$$\mathbf{S}_i^* \in \arg\max_{\mathbf{S}_i \in \mathcal{P}_i} \sum_{j\in[n]} \sum_{k\in[\ell]} v_{ijk}(r_{ij} - u) s_{ijk},$$

*where $\mathcal{P}_i$ is the set of feasible assortments in nest i as defined in (**??**), then for every $u' \in [u_p, u_{p+1}]$, we have*

$$\mathbf{S}_i^* \in \arg\max_{\mathbf{S}_i \in \mathcal{P}_i} \sum_{j\in[n]} \sum_{k\in[\ell]} v_{ijk}(r_{ij} - u') s_{ijk}.$$

Therefore, the collection of assortments obtained by Algorithm 11 are *ideal* and the following theorem is direct consequence of Lemma 5.5.

Theorem 5.6. *There exist a collection of assortments $\mathcal{T}_i$ with $|\mathcal{T}_i| = O(n^2)$ that includes an optimal solution to the linear program (5.25) for any $u_i \in R$.*

We can use Algorithm 11 to find the *ideal* collection of assortments in each nest and then solve the linear program (5.17) to obtain the optimal assortment of products and their display positions. From Theorem 5.6, we have that the linear program (5.17)

has $m+1$ variables and $O(mn^2)$ constraints and hence can be solved in time that is polynomial in the number of products and number of display segments.

5.4 Conclusion.

In this Chapter, we considered variants of the static assortment optimization problem under the Nested Logit and Multinomial Logit model and presented (near) optimal algorithms. Understandably, Multinomial Logit model, owing to its tractability, is a well studied choice model for assortment planning. [44], [40], [20] have considered the unconstrained, cardinality constrained and capacity constrained problems respectively and presented near optimal algorithms. Recently, [17] has presented a linear programming based solution for a large class of TU constraints. In this chapter, we contribute to the growing literature for assortment optimization under the MNL model by presenting a general framework for assortment planning under a large class of constraints. Our framework based on linear programming is robust enough to generalize for additional constraint for which exact approach is a hard problem.

Given the IIA property of the MNL model, NL model is attracting considerable attention. [23] has presented a polynomial time algorithm for cardinality constrained assortment planning under the NL model. [16] has extended the approach of [23] to present a polynomial algorithm to compute the optimal assortment of products and simultaneously compute the prices of the offered products such that the better quality products have higher prices. In this chapter, we have shown that the general problem under TU is NP-hard, presented an FPTAS under mild assumptions for the TU constraint structures and further presented an exact algorithm under specific parameter settings for a special application of the TU constraints, namely the joint assortment and the display optimization problem. Our work add to the literature of assortment planning under the NL model by consider the assortment planning under

the TU constraint structures. However, an FPTAS for the general TU problem or an exact algorithm for the joint assortment and display optimization problem for general parameters are still open questions.

Bibliography

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

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

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

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

[5] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 2003.

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

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

[8] Moshe Ben-Akiva and Steven Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*, volume 9. MIT press, 1985.

[9] Bloomberg. https://www.bloomberg.com/view/articles/2018-05-09/walmart-s-flipkart-deal-is-right-move-despite-investor-qualms.

[10] Bloomberg. https://www.bloombergquint.com/business/2018/03/23/this-is-why-amazon-hasnt-beaten-flipkart-in-india-yet.

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

[12] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

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

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

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

[16] J. Davis, H. Topaloglu, and Williamson David. Quality consistent discrete pricing under the nested logit model. *Working Paper*, 2016.

[17] James Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Technical Report*, 2013.

[18] James M. Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2):250–273, 2014.

[19] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

[20] Antoine Désir, Vineet Goyal, and Jiawei Zhang. Near-optimal algorithms for capacity constrained assortment optimization. 2014. Available at SSRN 2543309.

[21] Ilias Diakonikolas and Mihalis Yannakakis. Succinct approximate convex pareto curves. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 74–83, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

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

[23] Guillermo Gallego, Richard Ratliff, and Sergey Shebalov. A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research*, 63(1):212–232, 2015.

[24] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 13–20, 2010.

[25] William H Greene. The econometric approach to efficiency analysis. *The measurement of productive efficiency and productivity growth*, 1(1):92–250, 2008.

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

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

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

[29] M. Lichman. UCI machine learning repository, 2013.

[30] Livemint. https://www.livemint.com/industry/a8wttqtrj8dualthltkubi/flipkart-to-look-beyond-gross-sales-numbers-kalyan-krishnam.html.

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

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

[33] D. McFadden. Conditional logit analysis of qualitative choice behavior. *in P. Zarembka, ed., Frontiers in Econometrics*, 1973.

[34] Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

[35] Shashi Mittal and AndreasS. Schulz. An fptas for optimizing a class of low-rank functions over a polytope. *Mathematical Programming*, 141(1-2):103–120, 2013.

[36] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

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

[38] R.L. Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975.

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

[40] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

[41] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[42] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

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

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

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

[46] Kenneth Train. *Discrete choice methods with simulation*. Cambridge University Press, New York, NY, 2003.

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

[48] J.M.Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *Cambridge University Press*, 2015.

Appendices

Appendix A

Concentration Inequalities for Sum of Geometric Random Variables

Here, we prove concentration inequalities for sum of geometric random variables. Note that the estimates obtained from the epoch approach are distributed geometrically. The tail bound established in this section will help us in understanding how fast our estimate $\bar{v}_{i,\ell}$ converges to its true mean, v_i. The concentration bounds we prove in this section are similar to Chernoff bounds discussed in [36] (originally discussed in [4]), but for the fact that in bandit applications the number of arms over which we estimate the mean is a random variable. Hence, we use a self-normalized martingale technique to derive concentration bounds.

A.1 Exponential Inequalities for self-normalized martingales with Geometric distribution

Theorem A.1. *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$. Let $\mathcal{F}_\ell = \sigma(X_1, \cdots, X_{\ell-1})$ be the filtration corresponding to the random variables $\{X_i\}_{i=1,\cdots,n}$ and $\mathbb{1}_\ell$ be a $0-1$ random variable that is $\mathcal{F}_\ell$ measurable. Further, let*

$$\bar{X}_\ell \triangleq \frac{\sum_{i=1}^{\ell} X_i \cdot \mathbb{1}_i}{\sum_{i=1}^{\ell} \mathbb{1}_i},\ n_\ell \triangleq \sum_{i=1}^{\ell} \mathbb{1}_i \text{ and } \mu \triangleq \mathbb{E}(X_i) = \frac{1-p}{p}.$$

Then for any random variable δ, we have

1. $\mathsf{Pr}\left(\bar{X}_\ell > (1+\delta)\mu\right) \leq \begin{cases} \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \mu \delta^2}{2(1+\delta)(1+\mu)^2}\right)\right] & \text{if } \mu \leq 1, \\ \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \delta^2 \mu^2}{6(1+\mu)^2}\left(3 - \frac{2\delta\mu}{1+\mu}\right)\right)\right] & \text{if } \mu \geq 1. \end{cases}$

2. $\mathsf{Pr}\left(\bar{X}_\ell < (1-\delta)\mu\right) \leq \begin{cases} \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \delta^2 \mu}{6(1+\mu)^2}\left(3 - \frac{2\delta\mu}{1+\mu}\right)\right)\right] & \text{if } \mu \leq 1, \\ \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \delta^2 \mu^2}{2(1+\mu)^2}\right)\right] & \text{if } \mu \geq 1. \end{cases}$

Proof. We have

$$\bar{X}_i(\ell) = \frac{1}{n_\ell} \sum_{i=1}^{\ell} X_i \mathbb{1}_\ell.$$

Therefore, bounding $\mathsf{Pr}\left(\bar{X}_\ell > (1+\delta)\mu\right)$ and $\mathsf{Pr}\left(\bar{X}_\ell < (1-\delta)\mu\right)$ is equivalent to bounding $\mathsf{Pr}\left(\sum_{i=1}^{\ell} X_i \mathbb{1}_i > (1+\delta)\mu n_\ell\right)$ and $\mathsf{Pr}\left(\sum_{i=1}^{\ell} X_i \mathbb{1}_i < (1-\delta)\mu n_\ell\right)$. We will bound the first term and then follow a similar approach for bounding the second term to complete the proof.

Bounding $\mathsf{Pr}\left(\bar{X}_\ell > (1+\delta)\mu\right)$:

We have for any $\lambda > 0$,

$$\begin{aligned} \mathsf{Pr}\left(\sum_{i=1}^{\ell} X_i \mathbb{1}_i > (1+\delta)\mu n_\ell\right) &= \mathsf{Pr}\left\{\exp\left(\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i\right) > \exp\left(\lambda(1+\delta_i) v_i n_i(\ell)\right)\right\}, \\ &= \mathsf{Pr}\left\{\exp\left(\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i - \lambda(1+\delta)\mu n_\ell\right) > 1\right\}, \qquad \text{(A.1)} \\ &\leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i - \lambda(1+\delta)\mu n_\ell\right)\right], \end{aligned}$$

where the last inequality follows from Markov inequality. For notational brevity, denote $f(\lambda, \mu)$ by the function,

$$f(\lambda, \mu) = -\frac{\log\left(1 - \mu(e^{2\lambda} - 1)\right)}{2}.$$

We have,

$$\begin{aligned}
&\mathbb{E}\left[e^{\left(\lambda\sum_{i=1}^{\ell}X_i\mathbb{1}_i-\lambda(1+\delta)\mu n_\ell\right)}\right]\\
&=\mathbb{E}\left[e^{\left(\sum_{i=1}^{\ell}(\lambda X_i-f(\lambda,\mu))\cdot\mathbb{1}_i\right)}\cdot e^{\left(-\lambda(1+\delta)\mu(1-f(\lambda,\mu))n_\ell\right)}\right], \qquad (A.2)\\
&\leq\mathbb{E}^{\frac{1}{2}}\left[e^{\left(\sum_{\tau=1}^{\ell}(2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_i\right)}\right]\cdot\mathbb{E}^{\frac{1}{2}}[e^{\left(-2\lambda(1+\delta)\mu(1-f(\lambda,\mu))n_\ell\right)}],
\end{aligned}$$

where the above inequality follows from Cauchy-Schwartz inequality. Note that for any i, $\mathbb{1}_i$ conditioned on F_i is a constant and $\{X_i|\mathcal{F}_i\}$ is a geometric random variable. From the proof of Lemma 2.1, for all $i\geq 1$ and for any $0<\lambda<\frac{1}{2}\log\frac{1+\mu}{\mu}$, we have,

$$\mathbb{E}\left(e^{2\lambda X_i\mathbb{1}_i}\middle|\mathcal{F}_i\right)=\left(\frac{1}{1-\mu(e^{2\lambda}-1)}\right)^{\mathbb{1}_i}.$$

Therefore, it follows that

$$\mathbb{E}\left(e^{(2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_i}\middle|\mathcal{F}_i\right)\leq 1, \qquad (A.3)$$

and

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{i=1}^{\ell}(2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\right]&=\mathbb{E}\left[\mathbb{E}\left\{\exp\left((2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_i\right)|\mathcal{F}_\ell\right\}\right]\\
&=\mathbb{E}\left[\prod_{i=1}^{\ell-1}\exp\left((2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_\ell\right)\cdot\mathbb{E}\left(e^{(2\lambda X_\ell-2f(\lambda,\mu))\cdot\mathbb{1}_i}\middle|\mathcal{F}_\ell\right)\right]\\
&\leq\mathbb{E}\left[\prod_{i=1}^{\ell-1}\exp\left((2\lambda X_\ell-2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\right],
\end{aligned}$$

where the inequality follows from (A.3). Similarly by conditioning with $\mathcal{F}_{\ell-1},\cdots,\mathcal{F}_1$, we obtain,

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{\ell}(2\lambda X_i-2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\right]\leq 1.$$

From (A.1) and (A.2), we have

$$\mathsf{Pr}\left(\sum_{i=1}^{\ell}X_i\mathbb{1}_i>(1+\delta)\mu n_\ell\right)\leq\mathbb{E}^{\frac{1}{2}}\left[\exp\Big(-2\lambda(1+\delta)\mu(1-f(\lambda,\mu))n_\ell\Big)\right].$$

Therefore, we have

$$\mathsf{Pr}\left(\sum_{i=1}^{\ell}X_i\mathbb{1}_i>(1+\delta)\mu n_\ell\right)\leq\mathbb{E}^{\frac{1}{2}}\left[\min_{\lambda\in\Omega}\exp\Big(-2\lambda(1+\delta)\mu(1-f(\lambda,\mu))n_\ell\Big)\right], \qquad (A.4)$$

where $\Omega = \{\lambda | 0 < \lambda < \frac{1}{2}\log\frac{1+\mu}{\mu}\}$ is the range of λ for which the moment generating function in (A.3) is well definred. Taking logarithm of the objective in (A.4), we have,

$$\underset{\lambda\in\Omega}{\operatorname{argmin}}\ e^{-2\lambda(1+\delta)\mu(1-f(\lambda,\mu))\cdot n_\ell} = \underset{\lambda\in\Omega}{\operatorname{argmin}} -2(1+\delta)\lambda n_\ell\mu - n_\ell\log\left(1-\mu(e^{2\lambda}-1)\right) \quad \text{(A.5)}$$

Noting that the right hand side in the above equation is a convex function in λ, we obtain the optimal λ by solving for the zero of the derivative. Specifically, at optimal t, we have

$$e^{2\lambda} = \frac{(1+\delta)(1+\mu)}{1+\mu(1+\delta)}.$$

Substituting the above expression in (A.4), we obtain the following bound.

$$\mathsf{Pr}\left(\bar{X}_\ell > (1+\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\left(1-\frac{\delta}{(1+\delta)(1+\mu)}\right)^{n_\ell\mu(1+\delta)}\left(1+\frac{\delta\mu}{1+\mu}\right)^{n_\ell}\right]. \quad \text{(A.6)}$$

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

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

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

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

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 (A.6), we have

$$Pr\left(\bar{X}_\ell > (1+\delta)\mu\right) \leq \exp\left(-\frac{n_\ell\mu\delta^2}{2(1+\delta)(1+\mu)^2}\right). \quad \text{(A.7)}$$

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

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

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

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

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,

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

Therefore, substituting preceding results in (A.6), we have

$$Pr\left(\bar{X}_\ell > (1+\delta)\mu\right) \leq \begin{cases} \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \delta^2 \mu^2}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right)\right)\right] & \text{if } \mu \geq 1 \text{ and } \delta \in (0,1), \\ \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell \delta \mu^2}{6(1+\mu)^2}\left(3-\frac{2\mu}{1+\mu}\right)\right)\right] & \text{if } \mu \geq 1 \text{ and } \delta \geq 1. \end{cases} \tag{A.8}$$

Bounding $\mathsf{Pr}\left(\bar{X}_\ell < (1-\delta)\mu\right)$:

Now to bound the other one sided inequality, we use the fact that for any $\lambda > 0$,

$$\mathbb{E}\left(e^{-\lambda X_i \mathbb{1}_i} \middle| \mathcal{F}_i\right) = \left(\frac{1}{1-\mu(e^{-\lambda}-1)}\right)^{\mathbb{1}_i}.$$

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

$$\begin{aligned} \mathsf{Pr}\left(\sum_{i=1}^{\ell} X_i \mathbb{1}_i < (1-\delta) v_i n_\ell\right) &= \mathsf{Pr}\left\{\exp\left(-\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i\right) > \exp\left(-\lambda(1-\delta)\mu n_\ell\right)\right\}, \\ &= \mathsf{Pr}\left\{\exp\left(-\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i + \lambda(1-\delta)\mu n_\ell\right) > 1\right\}, \quad \text{(A.9)} \\ &\leq \mathbb{E}\left[\exp\left(-\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i + \lambda(1-\delta)\mu n_\ell\right)\right]. \end{aligned}$$

For notational brevity, denote $f(\lambda,\mu)$ by the function,

$$f(\lambda,\mu) = -\frac{\log\left(1-\mu(e^{-2\lambda}-1)\right)}{2}.$$

We have,

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\lambda \sum_{i=1}^{\ell} X_i \mathbb{1}_i + \lambda(1-\delta)\mu n_\ell\right)\right] \\ &= \mathbb{E}\left[\exp\left(\sum_{i=1}^{\ell}(-\lambda X_i - f(\lambda,\mu)) \cdot \mathbb{1}_i\right) \cdot \exp\left(\lambda(1-\delta)\mu(1+f(\lambda,\mu))n_\ell\right)\right], \quad \text{(A.10)} \\ &\leq \mathbb{E}^{\frac{1}{2}}\left[\exp\left(\sum_{i=1}^{\ell}(-2\lambda X_i - 2f(\lambda,\mu)) \cdot \mathbb{1}_i\right)\right] \cdot \mathbb{E}^{\frac{1}{2}}\left[\exp\left(2\lambda(1-\delta)\mu(1+f(\lambda,\mu))n_\ell\right)\right], \end{aligned}$$

where the above inequality follows from Cauchy-Schwartz inequality. Noting that for any τ, $\mathbb{1}_i$ conditioned on F_i is a constant and $\{X_i|\mathcal{F}_i\}$ is a geometric random variable. Therefore, for all $i \geq 1$ and for any $\lambda > 0$, we have,

$$\mathbb{E}\left(e^{-2\lambda X_i \mathbb{1}_i}\middle|\mathcal{F}_i\right) = \left(\frac{1}{1-\mu(e^{-2\lambda}-1)}\right)^{\mathbb{1}_i}.$$

Therefore, it follows that

$$\mathbb{E}\left(e^{(-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i}\middle|\mathcal{F}_i\right) \leq 1, \tag{A.11}$$

and

$$\begin{aligned}
\mathbb{E}\left[e^{\left(\sum_{i=1}^{\ell}(-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i\right)}\right] &= \mathbb{E}\left[\mathbb{E}\left\{e^{\left(\sum_{i=1}^{\ell}(-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i\right)}\middle|\mathcal{F}_\ell\right\}\right], \\
&= \mathbb{E}\left[\prod_{i=1}^{\ell-1}\exp\left((-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\cdot\mathbb{E}\left(e^{(-2\lambda X_\ell - 2f(\lambda,\mu))\cdot\mathbb{1}_i}\middle|\mathcal{F}_\ell\right)\right], \\
&= \mathbb{E}\left[\prod_{i=1}^{\ell-1}\exp\left((-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\right],
\end{aligned}$$

where the inequality follows from (A.11). Similarly by conditioning with $\mathcal{F}_{\ell-1}, \cdots, \mathcal{F}_1$, we obtain,

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{\ell}(-2\lambda X_i - 2f(\lambda,\mu))\cdot\mathbb{1}_i\right)\right] \leq 1.$$

From (A.9) and (A.10), we have

$$\mathsf{Pr}\left(\sum_{i=1}^{\ell} X_i\mathbb{1}_i < (1-\delta)\mu n_\ell\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\exp\left(2\lambda(1-\delta)\mu(1+f(\lambda,\mu))n_\ell\right)\right].$$

Therefore, we have

$$\mathsf{Pr}\left(\bar{X}_\ell < (1-\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\min_{\lambda>0}\exp\left(2\lambda(1-\delta)\mu(1+f(\lambda,\mu))n_\ell\right)\right].$$

Following similar approach as in optimizing the previous bound (see (A.4)) to establish the following result.

$$\mathsf{Pr}\left(\bar{X}_\ell < (1-\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\left(1+\frac{\delta}{(1-\delta)(1+\mu)}\right)^{n_\ell\mu(1-\delta)}\left(1-\frac{\delta\mu}{1+\mu}\right)^{n_\ell}\right].$$

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

$$\mathsf{Pr}\left(\bar{X}_\ell) < (1-\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\left(1+\frac{\delta}{1+\mu}\right)^{n_\ell\mu}\left(1-\frac{\delta\mu}{1+\mu}\right)^{n_\ell}\right]. \tag{A.12}$$

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

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

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

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

Therefore, substituting the preceding equations in (A.12), we have,

$$Pr\left(\bar{X}_\ell < (1-\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell\delta^2\mu}{6(1+\mu)^2}\left(3-\frac{2\delta\mu}{1+\mu}\right)\right).\right] \tag{A.13}$$

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

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

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

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

Therefore, substituting the preceding equations in (A.12), we have,

$$Pr\left(\bar{X}_\ell < (1-\delta)\mu\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n_\ell\delta^2\mu^2}{2(1+\mu)^2}\right)\right]. \tag{A.14}$$

The result follows from (A.7), (A.8), (A.13) and (A.14). □

Now, we will adapt a non-standard corollary from [7] and [26] to our estimates to obtain sharper bounds.

Lemma A.1. *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$. Let $\mathcal{F}_\ell = \sigma(X_1, \cdots, X_{\ell-1})$ be the filtration corresponding to the random variables $\{X_i\}_{i=1,\cdots,n}$ and $\mathbb{1}_\ell$ be a $0-1$ random variable that is $\mathcal{F}_\ell$ measurable. Further, let*

$$\bar{X}_\ell \triangleq \frac{\sum_{i=1}^{\ell} X_i \cdot \mathbb{1}_i}{\sum_{i=1}^{\ell} \mathbb{1}_i},\ n_\ell \triangleq \sum_{i=1}^{\ell} \mathbb{1}_i \text{ and } \mu \triangleq \mathbb{E}(X_i) = \frac{1-p}{p}.$$

If for any $m > 0$, $n_\ell > 48 \log(m+1)$, then we have for any ℓ,

1. $\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| > \max\left\{\sqrt{\bar{X}_\ell}, \bar{X}_\ell\right\}\sqrt{\frac{48\log(m+1)}{n_\ell}} + \frac{48\log(m+1)}{n_\ell}\right) \leq \frac{6}{m^2}$.
2. $\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| \geq \max\{\sqrt{\mu}, \mu\}\sqrt{\frac{24\log(m+1)}{n_\ell}} + \frac{48\log(m+1)}{n_\ell}\right) \leq \frac{4}{m^2}$,
3. $\mathcal{P}\left(\bar{X}_\ell \geq \frac{3\mu}{2} + \frac{48\log(m+1)}{n}\right) \leq \frac{3}{m^2}$.

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

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

$$\begin{aligned} \mathcal{P}\left(\bar{X}_\ell - \mu > \delta\mu\right) &\leq \frac{1}{m^2}, \\ \mathcal{P}\left(\bar{X}_\ell - \mu < -\delta\mu\right) &\leq \frac{1}{m^2}, \\ \mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| < (\mu+1)\sqrt{\frac{6\mu\log(m+1)}{n_\ell}}\right) &\geq 1 - \frac{2}{m^2}. \end{aligned} \tag{A.15}$$

Since $\delta \leq \frac{1}{2}$, we have $\mathcal{P}\left(\bar{X}_\ell - \mu \leq -\frac{\mu}{2}\right) \leq \mathcal{P}\left(\bar{X}_\ell - \mu \leq -\delta\mu\right)$. Hence, from (A.15), we have,

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

and hence, it follows that,

$$\mathcal{P}\left(2\bar{X} \geq \mu\right) \geq 1 - \frac{1}{N\ell^2}. \tag{A.16}$$

From (A.15) and (A.16), we have,

$$\mathcal{P}\left(\left|\bar{X}-\mu\right|<\sqrt{\frac{48\bar{X}\log\left(m+1\right)}{n}}\right)\geq\mathcal{P}\left(\left|\bar{X}-\mu\right|<\sqrt{\frac{24\mu\log\left(m+1\right)}{n}}\right)\geq1-\frac{3}{m^{2}}. \tag{A.17}$$

Since $\delta\leq\frac{1}{2}$, we have, $\mathcal{P}\left(\bar{X}_{\ell}\leq\frac{3\mu}{2}\right)\geq\mathcal{P}\left(\bar{X}_{\ell}<(1+\delta)\mu\right)$. Hence, from (A.15), we have

$$\mathcal{P}\left(\bar{X}_{\ell}\leq\frac{3\mu}{2}\right)\geq1-\frac{1}{m^{2}}. \tag{A.18}$$

Since, $\mu\leq1$, we have $\mathcal{P}\left(\bar{X}_{\ell}\leq\frac{3}{2}\right)\geq1-\frac{1}{m^{2}}$ and

$$\mathcal{P}\left(\bar{X}_{\ell}\leq\sqrt{\frac{3\bar{X}}{2}}\right)\geq1-\frac{1}{m^{2}}.$$

Therefore, substituting above result in (C.2), the first inequality in Lemma A.1 follows.

$$\mathcal{P}\left(\left|\bar{X}_{\ell}-\mu\right|>\max\left\{\sqrt{\bar{X}_{\ell}},\sqrt{\frac{2}{3}}\bar{X}_{\ell}\right\}\sqrt{\frac{48\log\left(m+1\right)}{n_{\ell}}}\right)\leq\frac{4}{m^{2}}. \tag{A.19}$$

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

$$\delta_{1}\triangleq\frac{12(\mu+1)^{2}\log\left(m+1\right)}{\mu n_{\ell}}\geq\frac{1}{2},$$

which implies,

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

Therefore, substituting the value of δ_1 in Theorem A.1, we have

$$\mathcal{P}\left(\left|\bar{X}_{\ell}-\mu\right|>\frac{48\log\left(m+1\right)}{n_{\ell}}\right)\leq\frac{2}{m^{2}}. \tag{A.20}$$

Hence, from (A.20) and (A.19), it follows that,

$$\mathcal{P}\left(\left|\bar{X}_{\ell}-\mu\right|>\max\left\{\sqrt{\bar{X}_{\ell}},\sqrt{\frac{2}{3}}\bar{X}_{\ell}\right\}\sqrt{\frac{48\log\left(m+1\right)}{n_{\ell}}+\frac{48\log\left(m+1\right)}{n_{\ell}}}\right)\leq\frac{6}{m^{2}}. \tag{A.21}$$

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

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

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

$$\mathcal{P}\left(2\bar{X}_\ell \geq \mu\right) \geq 1 - \frac{1}{m^2}.$$

Hence we have,

$$\begin{aligned}\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| < \bar{X}_\ell\sqrt{\frac{48\log(m+1)}{n_\ell}}\right) &\geq \mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| < \mu\sqrt{\frac{12\log(m+1)}{n_\ell}}\right) \\ &\geq 1 - \frac{3}{ml^2}.\end{aligned} \tag{A.22}$$

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

$$\mathcal{P}\left(\bar{X}_\ell \geq \sqrt{\frac{\bar{X}_\ell}{2}}\right) \geq 1 - \frac{1}{m^2}. \tag{A.23}$$

Therefore, from (A.22) and (A.23), we have

$$\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| > \max\left\{\bar{X}_\ell, \sqrt{\frac{\bar{X}_\ell}{2}}\right\}\sqrt{\frac{48\log(m+1)}{n_\ell}}\right) \leq \frac{4}{m^2}. \tag{A.24}$$

We complete the proof by stating that first inequality follows from (A.21) and (A.24), while second inequality follows from (C.2) and (A.22) and third inequality follows from (A.18) and (A.20). □

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

Corollary A.1. *Consider n i.i.d geometric random variables $X_1, \cdots, X_n$. Let $\mathcal{F}_\ell = \sigma(X_1, \cdots, X_{\ell-1})$ be the filtration corresponding to the random variables $\{X_i\}_{i=1,\cdots,n}$ and $\mathbb{1}_\ell$ be a $0-1$ random variable that is $\mathcal{F}_\ell$ measurable. Further, let*

$$\bar{X}_\ell \triangleq \frac{\sum_{i=1}^{\ell} X_i \cdot \mathbb{1}_i}{\sum_{i=1}^{\ell} \mathbb{1}_i},\ n_\ell \triangleq \sum_{i=1}^{\ell} \mathbb{1}_i \text{ and } \mu \triangleq \mathbb{E}(X_i) = \frac{1-p}{p}.$$

If $\mu \leq 1$, then we have for any $m > 0$

1. $\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| > \sqrt{\frac{48\bar{X}_\ell \log(m+1)}{n_\ell}} + \frac{48\log(m+1)}{n_\ell}\right) \leq \frac{6}{m^2}$.

2. $\mathcal{P}\left(\left|\bar{X}_\ell - \mu\right| \geq \sqrt{\frac{24\mu \log(m+1)}{n_\ell}} + \frac{48\log(m+1)}{n_\ell}\right) \leq \frac{4}{m^2}$.

3. $\mathcal{P}\left(\bar{X}_\ell \geq \frac{3\mu}{2} + \frac{48\log(m+1)}{n_\ell}\right) \leq \frac{3}{m^2}$.

A.2 Proof of Lemma 2.2 and Lemma 2.11

From Corollary 2.1, it follows that $\hat{v}_{i,\ell}$ are i.i.d geometric random variables with mean v_i. Furthermore, we have $\bar{v}_{i,\ell} = \frac{\sum_{\tau=1}^{\ell} \hat{v}_{i,\tau} \mathbb{1}\{i \in S_\tau\}}{\sum_{\tau=1}^{\ell} \mathbb{1}\{i \in S_\tau\}}$. Therefore, in the rest of this proof whenever we refer to Theorem A.1 or Lemma A.1 or Corollary A.1, it is assumed that $\mu = v_i$ and $\bar{X}_\ell = \bar{v}_{i,\ell}$.

Proof of Lemma 2.2: By design of Algorithm 1, we have,

$$v_{i,\ell}^{\mathsf{UCB}} = \bar{v}_{i,\ell} + \sqrt{48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}. \tag{A.25}$$

Therefore from Corollary A.1, we have

$$\mathcal{P}_\pi\left(v_{i,\ell}^{\mathsf{UCB}} < v_i\right) \leq \frac{6}{N\ell}. \tag{A.26}$$

The first inequality in Lemma 2.2 follows from (A.26). From triangle inequality and (A.25), we have,

$$\begin{aligned}\left|v_{i,\ell}^{\mathsf{UCB}} - v_i\right| &\leq \left|v_{i,\ell}^{\mathsf{UCB}} - \bar{v}_{i,\ell}\right| + \left|\bar{v}_{i,\ell} - v_i\right| \\ &= \sqrt{48\bar{v}_{i,\ell}\frac{\log(\sqrt{N}\ell+1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)} + \left|\bar{v}_{i,\ell} - v_i\right|.\end{aligned} \tag{A.27}$$

From Corollary A.1, we have

$$\mathsf{Pr}\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log(\sqrt{N}\ell+1)}{T_i(\ell)}\right) \leq \frac{3}{N\ell},$$

which implies

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

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

$$\mathsf{Pr}\left(\sqrt{\frac{\bar{v}_{i,\ell}\log\left(\sqrt{N}\ell+1\right)}{48T_i(\ell)}} + \frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)} > \sqrt{\frac{v_i\log\left(\sqrt{N}\ell+1\right)}{32T_i(\ell)}} + \frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{3}{N\ell}, \tag{A.28}$$

From Corollary A.1, we have,

$$\mathsf{Pr}\left(|\bar{v}_{i,\ell}-v_i| > \sqrt{24v_i\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{4}{N\ell}. \tag{A.29}$$

From (A.27) and applying union bound on (A.28) and (A.29), we obtain,

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

Lemma 2.2 follows from the above inequality and (A.26).

Proof of Lemma 2.11 By design of Algorithm 3, we have,

$$v_{i,\ell}^{\mathsf{UCB2}} = \bar{v}_{i,\ell} + \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{\frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}. \tag{A.30}$$

Therefore from Lemma A.1, we have

$$\mathsf{Pr}\left(v_{i,\ell}^{\mathsf{UCB2}} < v_i\right) \leq \frac{6}{N\ell}. \tag{A.31}$$

The first inequality in Lemma 2.2 follows from (A.31). From (A.30), we have,

$$\begin{aligned}\left|v_{i,\ell}^{\mathsf{UCB2}}-v_i\right| &\leq \left|v_{i,\ell}^{\mathsf{UCB}}-\bar{v}_{i,\ell}\right| + |\bar{v}_{i,\ell}-v_i| \\ &= \max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}\sqrt{48\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)} + |\bar{v}_{i,\ell}-v_i|.\end{aligned} \tag{A.32}$$

From Lemma A.1, we have

$$\mathsf{Pr}\left(\bar{v}_{i,\ell} > \frac{3v_i}{2} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{3}{N\ell},$$

which implies

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

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

$$\mathsf{Pr}\left(\frac{\max\left\{\sqrt{\bar{v}_{i,\ell}}, \bar{v}_{i,\ell}\right\}}{\max\left\{\sqrt{v_i}, v_i\right\}}\sqrt{\bar{v}_{i,\ell}\frac{\log\left(\sqrt{N}\ell+1\right)}{48T_i(\ell)}} > \sqrt{\frac{\log\left(\sqrt{N}\ell+1\right)}{32T_i(\ell)}} + \frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{3}{N\ell}, \tag{A.33}$$

From Lemma A.1, we have,

$$\mathsf{Pr}\left(|\bar{v}_{i,\ell} - v_i| > \max\left\{\sqrt{v_i}, v_i\right\}\sqrt{24\frac{\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}} + \frac{48\log\left(\sqrt{N}\ell+1\right)}{T_i(\ell)}\right) \leq \frac{4}{N\ell}. \tag{A.34}$$

From (A.32) and applying union bound on (A.33) and (A.34), we obtain,

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

Lemma 2.11 follows from the above inequality and (A.31). □

Appendix B

UCB Approach for the MNL-Bandit

B.1 Proof of Theorem 1

In this section, we utilize the results established in Section 2.2 and complete the proof of Theorem 1.

Proof. Let S^* denote the optimal assortment, our objective is to minimize the *Regret* defined in (MNL-Bandit), which is same as

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

Note that L, $\mathcal{E}_\ell$ and S_ℓ are all random variables and the expectation in equation (B.1) is over these random variables. Let $\mathcal{H}_\ell$ be the filtration (history) associated with the policy upto 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.

$$\pi(0, S_\ell) = \frac{1}{1 + \sum_{j \in S_\ell} v_j}.$$

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

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

and hence the Regret can be reformulated as

$$\mathsf{Reg}(T, \mathbf{v}) = \mathbb{E}\left\{\sum_{\ell=1}^{L} \left(1 + V(S_\ell)\right) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}, \tag{B.2}$$

the expectation in equation (B.2) is over the random variables L and S_ℓ. For the sake of brevity, for each $\ell \in 1, \cdots, L$, let

$$\Delta R_\ell = (1 + V(S_\ell)) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v}) \right). \tag{B.3}$$

Now the Regret can be reformulated as

$$\mathsf{Reg}(T, \mathbf{v}) = \mathbb{E}\left\{ \sum_{\ell=1}^{L} \Delta R_\ell \right\}. \tag{B.4}$$

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

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

Therefore, from Lemma 2.2, we have,

$$\mathsf{Pr}(\mathcal{A}_\ell) \leq \frac{13}{\ell}. \tag{B.5}$$

Since $\mathcal{A}_\ell$ is a "low probability" event (see (B.5)), 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:

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

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 (B.3), we have $\Delta R_\ell \leq N + 1$. Substituting the preceding inequality in the above equation, we obtain,

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

Whenever $\mathbb{1}(\mathcal{A}_{\ell-1}^c) = 1$, from the restricted monotonicity property of Lemma 2.3, 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

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

From the definition of the event, $\mathcal{A}_\ell$ and the Lipschitz property of Lemma 2.3, it follows that,

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

Therefore, we have

$$\mathbb{E}\left\{\Delta R_\ell\right\} \leq (N+1)\mathsf{Pr}\left(\mathcal{A}_{\ell-1}\right) + C \sum_{i \in S_\ell} \mathbb{E}\left(\sqrt{\frac{v_i \log \sqrt{NT}}{T_i(\ell)}} + \frac{\log \sqrt{NT}}{T_i(\ell)} \right), \tag{B.6}$$

where $C = \max\{C_1, C_2\}$. Combining equations (B.2) and (B.6), we have

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

Therefore, from Lemma 2.2, we have

$$\begin{aligned}
\mathsf{Reg}(T, \mathbf{v}) &\leq C\mathbb{E}\left\{ \sum_{\ell=1}^{L} \frac{N+1}{\ell} + \sum_{i \in S_\ell} \sqrt{\frac{v_i \log \sqrt{NT}}{T_i(\ell)}} + \sum_{i \in S_\ell} \frac{\log \sqrt{NT}}{T_i(\ell)} \right\}, \\
&\overset{(a)}{\leq} CN \log T + CN \log^2 \sqrt{NT} + C\mathbb{E}\left(\sum_{i=1}^{n} \sqrt{v_i T_i \log \sqrt{NT}} \right), \qquad \text{(B.7)} \\
&\overset{(b)}{\leq} CN \log T + CN \log^2 NT + C \sum_{i=1}^{N} \sqrt{v_i \log (NT) \mathbb{E}(T_i)}.
\end{aligned}$$

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

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

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

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

Therefore, it follows that

$$\sum v_i \mathbb{E}(T_i) \leq T. \tag{B.8}$$

To obtain the worst case upper bound, we maximize the bound in equation (B.7) subject to the condition (B.8) and hence, we have $\mathsf{Reg}(T, \mathbf{v}) = O(\sqrt{NT \log NT} + N \log^2 NT)$. □

B.2 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 B.1 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 B.1. *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,*

$$\mathsf{Pr}\left(S_\ell \subset S^*\right) \geq 1 - \frac{6}{\ell}.$$

Proof. If there exists a product i, such that $r_i \geq R(S^*, \mathbf{v})$, then following the proof of Lemma 2.3, 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,

$$i \in S^* \text{ if and only if } r_i \geq R(S^*, \mathbf{v}). \tag{B.9}$$

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

$$i \in S_\ell \text{ if and only if } r_i \geq R(S_\ell, \mathbf{v}_\ell^{\mathsf{UCB}}). \tag{B.10}$$

From Lemma 2.4, we have ,

$$\mathsf{Pr}\left(R(S_\ell, \mathbf{v}_\ell^{\mathsf{UCB}}) \geq R(S^*, \mathbf{v})\right) \geq 1 - \frac{6}{\ell}. \tag{B.11}$$

Lemma B.1 follows from (B.9), (B.10) and (B.11). □

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

Corollary B.1 (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* Reg*ret of the policy defined in Algorithm 1 at any time* T *is bounded as,*

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

B.3 Proof of Theorem 2

First we state the following auxiliary result that is helpful in proving Theorem 2. Following the proof of Lemma 2.7, we can establish the following result.

Corollary B.2. *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,*

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

We will re-use the ideas from proof of Theorem 1 to prove Theorem 2. Briefly, we breakdown the Regret into Regret over "good epochs" and "bad epochs." First we argue using Lemma 2.2, 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 2.7 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 B.1. We will now make these notions rigorous and complete the proof of Theorem 2.

Proof. Following the analysis in Appendix B.1, we reformulate the Regret as

$$\mathsf{Reg}(T, \mathbf{v}) = \mathbb{E}\left\{ \sum_{\ell=1}^{L} \left(1 + V(S_\ell)\right) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right) \right\}, \tag{B.12}$$

where S^* is the optimal assortment, $V(S_\ell) = \sum_{j \in S_\ell} v_j$ and the expectation in equation (B.12) is over the random variables L and S_ℓ. Similar to the analysis in Appendix B.1, for the sake of brevity, we define,

$$\Delta R_\ell = \left(1 + V(S_\ell)\right) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right). \tag{B.13}$$

Now the Regret can be reformulated as

$$\mathsf{Reg}(T, \mathbf{v}) = \mathbb{E}\left\{ \sum_{\ell=1}^{L} \Delta R_\ell \right\}. \tag{B.14}$$

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.

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

Using the fact that $R(S^*, \mathbf{v})$ and $R(S_\ell, \mathbf{v})$ are both bounded by one and $V(S_\ell) \leq N$ in (B.13), we have $\Delta R_\ell \leq N + 1$. Substituting the preceding inequality in the above equation, we obtain,

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

From the analysis in Appendix **??** (see (B.5)), we have $\mathcal{P}(\mathcal{A}_\ell) \leq \frac{13}{\ell}$. Therefore, it follows that,

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

Substituting the above inequality in (B.14), we obtain

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

From Corollary B.2, we have that $\sum_{\ell=1}^{L} \mathbb{1}(\ell \in \xi) \leq N \log NT$. Hence, we have,

$$\mathsf{Reg}(T, \mathbf{v}) \leq 14N \log T + N(N+1) \log NT + \mathbb{E}\left[\sum_{\ell=1}^{L} \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \mathbb{1}(\ell \in \xi^c) \right]. \tag{B.15}$$

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,

$$\mathbb{E}\left[\sum_{\ell=1}^{L} \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \mathbb{1}(\ell \in \xi^c) \right] = \mathbb{E}\left[\sum_{\ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}} \Delta R_\ell \cdot \mathbb{1}(\ell \in \xi^c) \right]. \tag{B.16}$$

Whenever $\mathbb{1}(\mathcal{A}_{\ell-1}^c) = 1$, from Lemma 2.3, 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

$$\begin{aligned}
\mathbb{E}\left\{\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_\ell^c)\right\} &\leq \mathbb{E}\left\{\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v}))\right] \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \mathbb{1}(\ell \in \xi^c)\right\}, \\
&\leq \sum_{i \in S_\ell} \left(C_1 \sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}} + \frac{C_2 \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)} \right) \cdot \mathbb{1}(\ell \in \xi^c), \\
&\leq C \sum_{i \in S_\ell} \sqrt{\frac{v_i \log\left(\sqrt{N}\ell + 1\right)}{T_i(\ell)}}.
\end{aligned} \tag{B.17}$$

where $C = C_1 + C_2$, the second inequality in (B.17) follows from the definition of the event, $\mathcal{A}_\ell$ and the last inequality follows from the definition of set ξ. From equations (B.15), (B.16), and (B.17) , we have,

$$\mathsf{Reg}(T, \mathbf{v}) \leq 14N^2 \log NT + C\mathbb{E}\left\{ \sum_{\ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}} \sum_{i \in S_\ell} \sqrt{\frac{\log NT}{T_i(\ell)}} \right\}, \tag{B.18}$$

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

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

Substituting the inequality $\sum_{\ell \in \mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}} \frac{1}{\sqrt{T_i(\ell)}} \leq \sqrt{T_i}$ in (B.18) and noting that $T_i \leq T$, we obtain,

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

From Jenson's inequality, we have $\mathbb{E}_\pi\left(\sqrt{T}_i\right) \leq \sqrt{\mathbb{E}_\pi\left(T_i\right)}$ and therefore, it follows that,

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

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

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

For any epoch ℓ, we have $|S_\ell| \leq N$. Hence, we have $\sum_{i=1}^{N} T_i \leq N|\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}|$. From Lemma 2.7, we have $|\mathcal{E}_{\mathcal{G}}^{\mathsf{sub\_opt}}| \leq N\tau$. Therefore, we have $\sum_{i=1}^{N} \mathbb{E}_\pi(T_i) \leq N^2\tau$ and hence, it follows that,

$$\begin{aligned}\mathsf{Reg}(T, \mathbf{v}) &\leq 14N^2 \log NT + CN\sqrt{N\tau \log NT}, \\ &\leq 14N^2 \log NT + C\frac{N^2 \log NT}{\Delta(\mathbf{v})}.\end{aligned} \tag{B.19}$$

□

B.4 Proof of Theorem 4

Proof of Theorem 4 is very similar to the proof of Theorem 1. 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:

$$\mathsf{Reg}(T, \mathbf{v}) = \underbrace{\mathbb{E}\left\{\sum_{\ell \in E_L} |\mathcal{E}_\ell| \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}}_{\mathsf{Reg}_1(T,\mathbf{v})} + \underbrace{\mathbb{E}\left\{\sum_{\ell \notin E_L} |\mathcal{E}_\ell| \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}}_{\mathsf{Reg}_2(T,\mathbf{v})}.$$

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

$$\mathsf{Reg}(T, \mathbf{v}) \leq \mathbb{E}\left\{\sum_{\ell \in E_L} |\mathcal{E}_\ell|\right\} + \mathsf{Reg}_2(T, \mathbf{v}).$$

From Lemma 2.10, it follows that,

$$\mathsf{Reg}(T, \mathbf{v}) \leq 49NB \log NT + \mathsf{Reg}_2(T, \mathbf{v}). \tag{B.20}$$

We will focus on the second term in the above equation, $\mathsf{Reg}_2(T, \mathbf{v})$. Following the analysis in Appendix B.1, we can show that,

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left\{\sum_{\ell \notin E_L} (1 + V(S_\ell)) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right)\right\}. \tag{B.21}$$

Similar to the analysis in Appendix B.1, for the sake of brevity, we define,

$$\Delta R_\ell = (1 + V(S_\ell)) \left(R(S^*, \mathbf{v}) - R(S_\ell, \mathbf{v})\right). \tag{B.22}$$

Now, $\mathsf{Reg}_2(T, \mathbf{v})$ can be reformulated as

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left\{\sum_{\ell \notin E_L} \Delta R_\ell\right\}. \tag{B.23}$$

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

$$\mathcal{B}_\ell = \bigcup_{i=1}^{N} \{\mathcal{C}_\ell \cup \mathcal{D}_\ell\},$$

where $\mathcal{C}_\ell = \left\{v_{i,\ell}^{\mathsf{UCB2}} < v_i\right\}$ and

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

From union bound, it follows that

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

Therefore, from Lemma 2.11, we have,

$$\mathsf{Pr}(\mathcal{B}_\ell) \leq \frac{13}{\ell}. \tag{B.24}$$

Since $\mathcal{B}_\ell$ is a "low probability" event (see (B.24)), 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.

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

Using the fact that $R(S^*, \mathbf{v})$ and $R(S_\ell, \mathbf{v})$ are both bounded by one and $V(S_\ell) \leq BN$ in (B.22), we have $\Delta R_\ell \leq N + 1$. Substituting the preceding inequality in the above equation, we obtain,

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

Whenever $\mathbb{1}(\mathcal{B}_{\ell-1}^c) = 1$, from Lemma 2.3, 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

$$\mathbb{E}\left\{\Delta R_\ell\right\} \leq B(N+1)\mathsf{Pr}(\mathcal{B}_{\ell-1}) + \mathbb{E}\left\{\left[(1+V(S_\ell))(\tilde{R}_\ell(S_\ell) - R(S_\ell, \mathbf{v}))\right] \cdot \mathbb{1}(\mathcal{B}_{\ell-1}^c)\right\}. \tag{B.25}$$

From the definition of the event, $\mathcal{B}_\ell$ and Lemma 2.12, we have,

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

and therefore, substituting above inequality in (B.25), we have

$$\mathbb{E}\left\{\Delta R_\ell\right\} \leq B(N+1)\mathsf{Pr}\left(\mathcal{B}_{\ell-1}\right) + C \sum_{i \in S_\ell} \mathbb{E}\left(\max\{v_i, \sqrt{v_i}\} \sqrt{\frac{\log \sqrt{N}T}{T_i(\ell)}} + \frac{\log \sqrt{N}T}{T_i(\ell)} \right), \tag{B.26}$$

where $C = \max\{C_1, C_2\}$. Combining equations (B.20), (B.23) and (B.26), we have

$$\mathsf{Reg}(T, \mathbf{v}) \leq 49BN \log NT + \mathbb{E}\left\{ \sum_{\ell=1}^{L} B(N+1)\mathsf{Pr}\left(\mathcal{A}_{\ell-1}\right) \right\}$$
$$+ \sum_{\ell=1}^{L} \mathbb{E}\left[C \max\{v_i, \sqrt{v_i}\} \sum_{i \in S_\ell} \left(\sqrt{\frac{\log \sqrt{N}T}{T_i(\ell)}} + \frac{\log \sqrt{N}T}{T_i(\ell)} \right) \right].$$

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

$$\begin{aligned}
\mathsf{Reg}(T, \mathbf{v}) &\leq 98NB \log NT + C\mathbb{E}\left\{ \sum_{\ell=1}^{L} \sum_{i \in S_\ell} \left(\max\left\{\sqrt{v_i}, v_i\right\} \sqrt{\frac{\log \sqrt{N}T}{T_i(\ell)}} + \frac{\log \sqrt{N}T}{T_i(\ell)} \right) \right\}, \\
&\overset{(a)}{\leq} 98NB \log NT + CN \log^2 NT + C\mathbb{E}\left(\sum_{i \in \mathcal{D}} \sqrt{v_i T_i \log NT} + \sum_{i \in \mathcal{I}} v_i \sqrt{T_i \log NT} \right), \\
&\overset{(b)}{\leq} 98NB \log NT + CN \log^2 NT + C \sum_{i \in \mathcal{D}} \sqrt{v_i \mathbb{E}(T_i) \log NT} + \sum_{i \in \mathcal{I}} v_i \sqrt{\mathbb{E}(T_i) \log NT},
\end{aligned} \tag{B.27}$$

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

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

while inequality (b) follows from Jensen's inequality. From (B.8), we have that,

$$\sum v_i \mathbb{E}(T_i) \leq T.$$

To obtain the worst case upper bound, we maximize the bound in equation (B.27) subject to the above constraint. Noting that the objective in (B.27) is concave, we use the KKT conditions to derive the worst case bound as $\mathsf{Reg}(T, \mathbf{v}) = O(\sqrt{BNT \log NT} + N \log^2 NT + BN \log NT)$. □

B.5 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.2. *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. 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 $x = \frac{\epsilon}{1-\alpha}$, 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.3. *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. 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,

$$J_1 = \left\{ i \,\middle|\, E_{P_0}(N_i) \leq \frac{3t}{N} \right\}, J_2 = \left\{ i \,\middle|\, \mathcal{P}_0(a_t = i) \leq \frac{3}{N} \right\} \text{ and } J = J_1 \cap J_2. \tag{B.28}$$

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

$$|\mathcal{P}_j(a_t = j) - \mathcal{P}_0(a_t = j)| \leq \frac{1}{2}\sqrt{2 \log 2 \cdot KL(P_0 \| P_j)}, \tag{B.29}$$

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.2 and (B.28), we have that Therefore,

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

Therefore,

$$\begin{aligned} \mathcal{P}_j(a_t = j) &\leq \mathcal{P}_0(a_t = j) + \frac{1}{2}\sqrt{2\log 2 \cdot KL(P_0 \| P_j)}, \\ &\leq \frac{3}{N} + \frac{1}{2}\sqrt{(2\log 2)\frac{4\epsilon^2}{\alpha} E_{P_0}(N_j)}, \\ &\leq \frac{3}{N} + \sqrt{2\log 2}\sqrt{\frac{3t\epsilon^2}{N\alpha}} \leq \frac{1}{2}. \end{aligned} \tag{B.30}$$

The second inequality follows from (B.28), while the last inequality follows from the fact that $N > 12$ and $t \leq \frac{N\alpha}{60\epsilon^2}$ $\square$.

Proof of Lemma 2.8. 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,

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

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

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 $\mathsf{Reg}ret \geq \frac{T\epsilon}{6}$.

Lemma B.4. *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 $\mathbb{E}(L) \leq 2T + 1 \leq 3T$. □

B.5.1 Lower Bound for the unconstrained MNL-Bandit problem ($K = N$)

We will complete proof of Theorem 2.4 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 B.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 B.1, we have that,

$$i \in S^* \text{ if and only if } r_i \geq R(S^*, \mathbf{v}), \tag{B.31}$$

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 (B.31), we have that optimal assortment is either $\{1\}$ or $\{1, 2\}$, specifically, we have that

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

Therefore, we have,

$$S^* = \begin{cases} \{1\}, & \text{if } v_1 = \frac{1}{2} + \epsilon, \\ \{1, 2\}, & \text{if } v_1 = \frac{1}{2}. \end{cases} \tag{B.32}$$

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

$$R(\{1\}, \mathbf{v}) - R(S, \mathbf{v}) \geq R(\{1\}, \mathbf{v}) - R(\{1, 2\}, \mathbf{v}) \geq \frac{\epsilon}{20}, \tag{B.33}$$

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

$$R(\{1\}, \mathbf{v}) - R(S, \mathbf{v}) \geq R(\{1, 2\}, \mathbf{v}) - R(\{1\}, \mathbf{v}) = \frac{\epsilon}{12} \geq \frac{\epsilon}{20}. \tag{B.34}$$

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 2.8. First, we present two auxillary results required to prove Lemma 2.4.

Lemma B.5. *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,*

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

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

Then for any S,

$$\mathsf{KL}\left(\mathcal{P}_0^S \middle\| \mathcal{P}_1^S\right) \leq 4\epsilon^2, \tag{B.35}$$

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 $1 \in S$.

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

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

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

□

Lemma B.6. *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,

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = \sum_{\mathbf{c} \in \{0,1,\cdots,N\}^T} \mathcal{P}(\mathbf{c}) \log\left(\frac{\mathcal{P}(\mathbf{c})}{\mathcal{P}_1(\mathbf{c})}\right), \tag{B.36}$$

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,

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = \sum_{t=1}^{T} \sum_{\{c_1,\cdots,c_{t-1}\} \in \{0,1,\cdots,N\}^{t-1}} \mathbb{P}_0(\mathbf{c^t}) \mathsf{KL}\left(\mathbb{P}_0(c_t) \| \mathbb{P}_1(c_t) | c_1, \cdots, c_{t-1}\right), \tag{B.37}$$

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,

$$\mathsf{KL}\left(\mathbb{P}_0(c_t) \| \mathbb{P}_1(c_t) | c_1, \cdots, c_{t-1}\right) = \mathsf{KL}\left(\mathcal{P}_0^{S_t}(c_t) \middle\| \mathcal{P}_1^{S_t}(c_t)\right), \tag{B.38}$$

where $\mathcal{P}_0^{S_t}$ and $\mathcal{P}_1^{S_t}$ are defined as in Lemma B.5. Therefore from (B.37), (B.38) and Lemma B.5, we have,

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

□

Proof of Theorem 2.4: 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 $\mathcal{A}$.

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}_0}(T_2) - \mathbb{E}_{\mathbb{P}_1}(T_2)| &\leq \left| \sum_{t=1}^{T} \mathcal{P}_0(2 \in S_t) - \mathbb{P}_1(2 \in S_t) \right|, \\
&\leq \sum_{t=1}^{T} \left| \mathbb{P}_0(2 \in S_t) - \mathbb{P}_1(2 \in S_t) \right|, \qquad \text{(B.39)} \\
&\leq \sum_{t=1}^{T} \frac{1}{2} \sqrt{2 \log 2 \cdot \mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right)} = \frac{T}{2} \sqrt{2 \log 2 \cdot \mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right)},
\end{aligned}$$

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 (B.36) and the last inequality follows from Pinsker's inequality. From Lemma B.6, we have that,

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

Substituting the value of ϵ, we obtain $\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) \leq \frac{1}{2}$ and from (B.39), we have

$$|\mathbb{E}_{\mathbb{P}_0}(T_2) - \mathbb{E}_{\mathbb{P}_1}(T_2)| \leq \frac{T}{4}. \qquad \text{(B.40)}$$

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

$$\mathsf{Reg}_{\mathcal{A}_{\mathsf{MNL}}}(T, \mathbf{v}) = \frac{1}{2} \mathsf{Reg}_{\mathcal{A}_{\mathsf{MNL}}}\left(T, \mathbf{v}, \left| v_1 = \frac{1}{2}\right.\right) + \frac{1}{2} \mathsf{Reg}_{\mathcal{A}_{\mathsf{MNL}}}\left(T, \mathbf{v}, \left| v_1 = \frac{1}{2} + \epsilon\right.\right). \qquad \text{(B.41)}$$

From (B.34) 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,

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

and similarly from (B.33) 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,

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

Therefore, from (B.41) and (B.40), it follows that,

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

□

Appendix C

Thompson Sampling for the MNL-Bandit

C.1 Bounds on the deviation of MNL Expected Revenue

Here, we bound the difference between the expected revenues of the offer set S_ℓ corresponding to the sampled parameters, $\boldsymbol{\mu}(\ell)$ and the true parameters, $\mathbf{v}$. In order to establish this bound, we will first present two concentration results. In the first result, utilizing the large deviation properties of Gaussian distribution, we show that over time, the posterior distributions concentrate around their means. The second result proves a Chernoff-like bound which suggests that the means of the posterior distribution concentrates around the true parameters. The second result is similar to the Corollary A.1 which is a consequence of the exponential inequalities for Geometric random variables that were derived in Theorem A.1.

Proof of Lemma A.1.

Let $\delta_i = \sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}}$. We analyze the cases $\delta_i \leq \frac{1}{2}$ and $\delta_i \geq \frac{1}{2}$ separately.

Case 1: $\delta_i \leq \frac{1}{2}$: For any $v_i \leq 1$ and $\delta_i \leq 1/2$, we have,

$$\frac{v_i\delta_i^2 n_i(\ell)}{2(1+\delta_i)(1+v_i)^2} \geq \frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)} \geq m\log(\rho+1),$$

and

$$\frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)^2}\left(3-\frac{2\delta_i v_i}{1+v_i}\right) \geq \frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)} \geq m\log(\rho+1).$$

Therefore, substituting $\delta_i = \sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}}$ in Lemma A.1 with δ_i, we have,

$$\begin{aligned} \mathcal{P}\left(2\hat{v}_i(\ell) \geq v_i\right) &\geq 1 - \frac{1}{\rho^m}, \\ \mathcal{P}\left(|\hat{v}_i(\ell) - v_i| < \sqrt{\frac{4v_i(v_i+2)m\log(\rho+1)}{n_i(\ell)}}\right) &\geq 1 - \frac{2}{\rho^m}. \end{aligned} \tag{C.1}$$

From the above three results, we have,

$$\begin{aligned} &\mathcal{P}\left(|\hat{v}_i(\ell) - v_i| < \sqrt{\frac{16\hat{v}_i(\ell)\left(\hat{v}_i(\ell)+1\right)\log(\rho+1)}{n_i(\ell)}}\right) \\ &\geq \mathcal{P}\left(|\hat{v}_i(\ell) - v_i| < \sqrt{\frac{4v_i(v_i+2)\log(\rho+1)}{n_i(\ell)}}\right) \geq 1 - \frac{3}{\rho^m}. \end{aligned} \tag{C.2}$$

By assumption, $v_i \leq 1$. Therefore, we have $v_i(v_i+2) \leq 3v_i$ and,

$$\mathcal{P}\left(|\hat{v}_i(\ell) - v_i| < \sqrt{\frac{12v_i\log(\rho+1)}{n_i(\ell)}}\right) \geq 1 - \frac{3}{\rho^m}.$$

Case 2: $\delta_i > \frac{1}{2}$: Now consider the scenario, when $\sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}} > \frac{1}{2}$. Then, we have,

$$\bar{\delta}_i \triangleq \frac{8(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)} \geq \frac{1}{2},$$

which implies for any $v_i \leq 1$,

$$\begin{aligned} \frac{nv_i\bar{\delta}_i^2}{2(1+\bar{\delta}_i)(1+v_i)^2} &\geq \frac{nv_i\bar{\delta}_i}{12(1+v_i)}, \\ \frac{n\bar{\delta}_i^2 v_i}{6(1+v_i)^2}\left(3 - \frac{2\bar{\delta}_i v_i}{1+v_i}\right) &\geq \frac{nv_i\bar{\delta}_i}{12(1+v_i)}. \end{aligned}$$

Therefore, substituting the value of $\bar{\delta}_i$ in Lemma A.1, we have

$$\mathcal{P}\left(|\hat{v}_i(\ell) - v_i| > \frac{24m\log(\rho+1)}{n}\right) \leq \frac{2}{\rho^m}.$$

Proof of Lemma 3.3: Note that we have $\mu_i(\ell) = \hat{v}_i(\ell) + \hat{\sigma}_i(\ell) \cdot \max_{j=1,\cdots,K}\{\theta^{(j)}(\ell)\}$. Therefore, from union bound, we have,

$$\begin{aligned} \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK} \;\middle|\; \hat{v}_i(\ell)\right\} &= \mathbb{P}\left(\bigcup_{j=1}^{K}\left\{\theta^j(\ell) > 4\sqrt{\log rK}\right\}\right) \\ &\leq \sum_{j=1}^{K}\mathbb{P}\left(\theta^j(\ell) > 4\sqrt{\log rK}\right) \end{aligned}$$

The result follows from the above inequality and the following anti-concentration bound for the normal random variable $\theta^{(j)}(\ell)$ (see formula 7.1.13 in [1]).

$$\frac{1}{4\sqrt{\pi}} \cdot e^{-7z^2/2} < \mathsf{Pr}\left(|\theta^{(j)}(\ell)| > z\right) \leq \frac{1}{2} e^{-z^2/2}.$$

Corollary C.1. *For any item i and any epoch ℓ, we have*

$$\mathbb{E}\left(|\mu_i(\ell) - \hat{v}_i(\ell)|\right) \leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK}.$$

Proof. In Lemma 3.3, we show that for any $r > 0$ and $i = 1, \cdots, N$, we have,

$$\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right) \leq \frac{1}{r^4 K^3},$$

where $\hat{\sigma}_i(\ell) = \sqrt{\frac{50\hat{v}_i(\hat{v}_i+1)}{n_i}} + \frac{75\sqrt{\log TK}}{n_i}$. Since $S_\ell \subset \{1, \cdots, N\}$, we have for any $i \in S_\ell$ and $r > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK} \text{ for any } i \in S_\ell\right) \\
&\leq \mathbb{P}\left(\bigcup_{i=1}^{N} |\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right), \qquad \text{(C.3)} \\
&\leq \frac{N}{r^4 K^3}.
\end{aligned}$$

Since $|\mu_i(\ell) - \hat{v}_i(\ell)|$ is a non-negative random variable, we have

$$\begin{aligned}
\mathbb{E}(|\mu_i(\ell) - \hat{v}_i(\ell)|) &= \int_0^\infty \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx, \\
&= \int_0^{4\hat{\sigma}_i(\ell)\sqrt{\log TK}} \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx + \int_{4\hat{\sigma}_i(\ell)\sqrt{\log TK}}^\infty \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx, \\
&\leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK} + \sum_{r=T}^{\infty} \int_{4\hat{\sigma}_i(\ell)\sqrt{\log rK}}^{4\hat{\sigma}_i(\ell)\sqrt{\log (r+1)K}} \mathbb{P}\left\{Y \geq x\right\} dx, \qquad \text{(C.4)} \\
&\overset{a}{\leq} 4\hat{\sigma}_i(\ell)\sqrt{\log TK} + \sum_{r=T}^{\infty} \frac{N\sqrt{\log\left(rK+1\right)} - N\sqrt{\log rK}}{r^4 K^3}, \\
&\leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK} \text{ for any } T \geq N,
\end{aligned}$$

where the inequality (a) follows from (C.3). □

C.2 Proof of Theorem 3.1

In this section, we will utilize the above properties and follow the outline discussed in Section 3.3.1 to complete the proof of Theorem 1. For the sake of brevity we will use the following notation for the rest of this section.

- For any assortment S, $V(S) \triangleq \sum_{i \in S} v_i$
- For any $\ell, \tau \leq L$, define ΔR_ℓ and $\Delta R_{\ell,\tau}$ in the following manner

$$\Delta R_\ell \triangleq (1 + V(S_\ell)) \left[R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v}) \right]$$
$$\Delta R_{\ell,\tau} \triangleq (1 + V(S_\tau)) \left[R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \boldsymbol{\mu}(\tau)) \right]$$

- Let $\mathcal{A}_0$ denote the complete set Ω and for all $\ell = 1, \ldots, L$, define events $\mathcal{A}_\ell$ as

$$\mathcal{A}_\ell = \left\{ |\hat{v}_i(\ell) - v_i| \geq \sqrt{\frac{24 v_i \log(\ell+1)}{n_i(\ell)}} + \frac{48 \log(\ell+1)}{n_i(\ell)} \text{ for some } i = 1, \cdots, N \right\}$$

We will bound the regret by bounding both the terms in (3.4). We will first focus on bounding the second term, $\mathsf{Reg}_2(T, \mathbf{v})$ and then extend this analysis to bound the first term, $\mathsf{Reg}_1(T, \mathbf{v})$.

Bounding $\mathsf{Reg}_2(T, \mathbf{v})$: Note that conditioned on event S_ℓ, the length of the ℓ^{th} epoch, $|\mathcal{E}^{\mathsf{AI}}|$ is a geometric random variable with probability of success $p_0(S_\ell) = 1/(1 + V(S_\ell))$. Therefore using conditional expectations, we can reformulate $\mathsf{Reg}_2(T, \mathbf{v})$ as,

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left\{ \sum_{\ell=1}^{L} \Delta R_\ell \right\}. \tag{C.5}$$

Noting that $\mathcal{A}_\ell$ is a "low probability" event, we analyze the regret in two scenarios, one when $\mathcal{A}_\ell$ is true and another when $\mathcal{A}_\ell^c$ is true. More specifically,

$$\begin{aligned} \mathbb{E}(\Delta R_\ell) &= \mathbb{E}\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \right], \\ &\leq \frac{K+1}{\ell^2} + \mathbb{E}\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c) \right], \end{aligned} \tag{C.6}$$

where the last inequality follows from Lemma A.1 and the fact that both $R(S_\ell, \boldsymbol{\mu}(\ell))$ and $R(S_\ell, \mathbf{v})$ are both bounded by one and $V(S_\ell) \leq K$. Therefore from Lemma 2.3 it follows that,

$$\begin{aligned}\mathbb{E}\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right] &\leq \mathbb{E}\left[\sum_{i\in S_\ell} |\mu_i(\ell) - v_i| \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right]. \\ &\leq \mathbb{E}\left[\sum_{i\in S_\ell} |\mu_i(\ell) - \hat{v}_i(\ell)|\right] + \mathbb{E}\left[\sqrt{\frac{24 v_i \log(\ell+1)}{n_i(\ell)}} + \frac{48\log(\ell+1)}{n_i(\ell)}\right],\end{aligned} \tag{C.7}$$

where the last inequality follows from the definition of event $\mathcal{A}_\ell$ and triangle inequality.

In Corollary C.1, we use Lemma 3.3 to show that the first term in above inequality, which is difference between the sampled parameter and the mean of the posterior distribution is bounded. Therefore, from (D.23), (D.24), (C.7), Corollary C.1 and Lemma A.1, we have,

$$Reg_2(T, \mathbf{v}) \leq C_1 \mathbb{E}\left(\sum_{\ell=1}^{L}\sum_{i\in S_\ell}\sqrt{\frac{v_i \log TK}{n_i(\ell)}}\right) + C_2\mathbb{E}\left(\sum_{\ell=1}^{L}\sum_{i\in S_\ell}\frac{\log TK}{n_i(\ell)}\right), \tag{C.8}$$

where C_1 and C_2 are absolute constants. If T_i denote the total number of epochs product i is offered, then we have,

$$\begin{aligned}Reg_2(T, \mathbf{v}) &\overset{(a)}{\leq} C_2 N \log^2 TK + C_1 \mathbb{E}\left(\sum_{i=1}^{n}\sqrt{v_i T_i \log TK}\right), \\ &\overset{(b)}{\leq} C_2 N \log^2 TK + C_1 \sum_{i=1}^{N}\sqrt{v_i \log(TK)\mathbb{E}(T_i)}.\end{aligned} \tag{C.9}$$

Inequality (a) follows from the observation that $L \leq T$, $T_i \leq T$, $\sum_{n_i(\ell)=1}^{T_i} \frac{1}{\sqrt{n_i(\ell)}} \leq \sqrt{T_i}$ and $\sum_{n_i(\ell)=1}^{T_i} \frac{1}{n_i(\ell)} \leq \log T_i$, while Inequality (b) follows from Jensen's inequality.

Since that expected epoch length condition on the event $S = S_\ell$ is $1 + V(S_\ell)$, we have, $\sum v_i \mathbb{E}(T_i) \leq T$. To obtain the worst case upper bound, we maximize the bound in equation (C.9) subject to the above condition and hence, we have

$$Reg_2(T, \mathbf{v}) \leq C_1\sqrt{NT\log TK} + C_2 N \log^2 TK). \tag{C.10}$$

We will now focus on the first term in (3.4), $\mathsf{Reg}_1(T, \mathbf{v}$.

Bounding $\mathsf{Reg}_1(T, \mathbf{v})$: Let $\mathcal{T}$ denote the set of optimistic epochs. Recall that $\mathcal{E}^{\mathsf{An}}(\ell)$ is the set of non-optimistic epochs between ℓ^{th} epoch and the subsequent optimistic epoch. Therefore, we can reformulate $Reg_1(T, \mathbf{v})$ as,

$$\mathsf{Reg}_1(T, \mathbf{v}) = \mathbb{E}[\sum_{\ell=1}^{L} \mathbb{1}(\ell \in \mathcal{T}) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} |\mathcal{E}_\tau|(R(S^*, \mathbf{v}) - R(S_\tau, \boldsymbol{\mu}(\tau)))]$$

Note that for any ℓ, by algorithm design we have that S_ℓ is the optimal set for the sampled parameters, i.e., $R(S_\ell, \boldsymbol{\mu}(\ell)) \geq R(S^*, \boldsymbol{\mu}(\ell))$. From the restricted monotonicity property, for any $\ell \in \mathcal{T}$, we have $R(S^*, \boldsymbol{\mu}(\ell)) \geq R(S^*, \mathbf{v})$. Therefore, it follows that,

$$\begin{aligned}
\mathsf{Reg}_1(T, \mathbf{v}) &\leq \mathbb{E}[\sum_{\ell=1}^{L} \mathbb{1}(\ell \in \mathcal{T}) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} |\mathcal{E}_\tau|(R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\tau, \boldsymbol{\mu}(\tau)))], \\
&\overset{(a)}{\leq} \mathbb{E}[\sum_{\ell=1}^{L} \mathbb{1}(\ell \in \mathcal{T}) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} |\mathcal{E}_\tau|(R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \boldsymbol{\mu}(\tau)))], \qquad \text{(C.11)} \\
&\overset{(b)}{\leq} \mathbb{E}[\sum_{\ell=1}^{L} \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} \Delta R_{\ell,\tau}]
\end{aligned}$$

where inequality (a) follows from the fact S_τ is the optimal assortment for the sampled parameters $\boldsymbol{\mu}(\tau)$ and inequality (b) follows from the observation that the expected length of the τ^{th} epoch conditioned on event $S = S_\tau$ is $1 + V(S_\tau)$. Following the approach of bounding $\mathsf{Reg}_2(T, \mathbf{v})$, we analyze the first term, $\mathsf{Reg}_1(T, \mathbf{v})$ in two scenarios, one when $\mathcal{A}_\ell$ is true and another when $\mathcal{A}_\ell^c$ is true. More specifically,

$$\begin{aligned}
&\frac{\mathbb{E}\left(\sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} \Delta R_{\ell,\tau}\right)}{K+1} = \frac{\mathbb{E}[\sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} \Delta R_{\ell,\tau} \cdot \mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_{\ell,\tau} \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)]}{K+1}, \\
&\overset{(a)}{\leq} \mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)| \cdot \mathbb{1}(\mathcal{A}_{\ell-1})] + \Delta R_{\ell,\tau} \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)], \\
&\overset{(b)}{\leq} \mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)| \cdot \mathbb{1}(\mathcal{A}_{\ell-1})] + \mathbb{E}[\mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} \sum_{i \in S_\ell} |\mu_i(\ell) - \mu_i(\tau)|], \qquad \text{(C.12)} \\
&\overset{(c)}{\leq} \mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)| \cdot \mathbb{1}(\mathcal{A}_{\ell-1})] + \mathbb{E}\left[\mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(\ell)} \sum_{i \in S_\ell} |\mu_i(\ell) - v_i| + |\mu_i(\tau) - v_i|\right],
\end{aligned}$$

where inequality (a) follows from the fact that $R(S_\ell, \boldsymbol{\mu}(\ell))$ and $R(S_\ell, \boldsymbol{\mu}(\tau))$ are both bounded by one and $V(S_\tau) \leq K$, inequality (b) follows from Lemma 2.3 and inequality (c) follows from the triangle inequality.

Following the approach of Bounding $\mathsf{Reg}_2(T, \mathbf{v})$, specifically along the lines of (C.7) and Corollary C.1, we can show that

$$\mathbb{1}(\mathcal{A}_{\ell-1}^c) \cdot |\mu_i(\ell) - v_i| \leq C_1 \sqrt{\frac{v_i \log TK}{n_i(\ell)}} + \frac{\log TK}{n_i(\ell)}.$$

Since $\tau \geq \ell$ we have $n_i(\ell) \leq n_i(\tau)$. Therefore, from (C.11), (C.12) and Lemma A.1 we obtain the following inequality.

$$\mathsf{Reg}_1(T, \mathbf{v}) \leq \mathbb{E}\left[\sum_{\ell \in \mathcal{T}} |\mathcal{E}^{\mathsf{An}}(\ell)| \sum_{i \in S_\ell} \left(C_1 \sqrt{\frac{v_i \log TK}{n_i(\ell)}} + C_2 \frac{\log TK}{n_i(\ell)} \right)\right], \tag{C.13}$$

for some constants C_1 and C_2. If $|\mathcal{E}^{\mathsf{An}}(.)|$ is not a random variable and constant, then bounding the above inequality is similar to bounding $\mathsf{Reg}_1(T, \mathbf{v})$ (see (C.8)). In the remainder of this section, we will show how to utilize Lemma 3.6 to bound $\mathsf{Reg}_1(T, \mathbf{v})$. From Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\sum_{\ell \in \mathcal{T}} \sum_{i \in S_\ell} |\mathcal{E}^{\mathsf{An}}(\ell)| C_1 \sqrt{\frac{v_i \log TK}{n_i(\ell)}}\right] \leq C_1 \sum_{\ell} \sum_{i \in S_\ell} \mathbb{E}^{1/2}\left[|\mathcal{E}^{\mathsf{An}}(\ell)|^2\right] \cdot \mathbb{E}^{1/2}\left[\frac{v_i \log TK}{n_i(\ell)}\right],$$

$$\mathbb{E}\left[\sum_{\ell \in \mathcal{T}} \sum_{i \in S_\ell} |\mathcal{E}^{\mathsf{An}}(\ell)| C_2 \frac{\log TK}{n_i(\ell)}\right] \leq C_2 \sum_{\ell} \sum_{i \in S_\ell} \mathbb{E}^{1/2}\left[|\mathcal{E}^{\mathsf{An}}(\ell)|^2\right] \mathbb{E}^{1/2}\left[\frac{\log^2 TK}{n_i^2(\ell)}\right].$$

Therefore from Lemma 3.6 for some absolute constant C, we have,

$$\begin{aligned}
\mathsf{Reg}_1(T, \mathbf{v}) \leq & \frac{C}{K}\left(\sum_{\ell} \sum_{i \in S_\ell} \mathbb{E}^{1/2}\left[\frac{v_i \log TK}{n_i(\ell)}\right] + \sum_{\ell} \sum_{i \in S_\ell} \mathbb{E}^{1/2}\left[\frac{\log^2 TK}{n_i^2(\ell)}\right]\right), \\
\leq & \frac{C}{K}\left(\sqrt{TK\mathbb{E}\left[\sum_{\ell} \sum_{i \in S_\ell} \frac{v_i \log TK}{n_i(\ell)}\right]} + \sqrt{TK\mathbb{E}\left[\sum_{\ell} \sum_{i \in S_\ell} \frac{\log^2 TK}{n_i^2(\ell)}\right]}\right),
\end{aligned} \tag{C.14}$$

where the last inequality follows Cauchy-Schwarz inequality. Since $v_i \leq 1$ for all i, we have,

$$\sum_{\ell} \sum_{i \in S_\ell} \frac{v_i \log TK}{n_i(\ell)} \leq \sum_{i=1}^{N} \sum_{n_i(\ell)=1}^{T_i} \frac{\log TK}{n_i(\ell)} \leq N \log TK \cdot \log T,$$

and

$$\sum_{\ell}\sum_{i\in S_\ell}\frac{\log^2 TK}{n_i^2(\ell)} = \sum_{i=1}^{N}\sum_{n_i(\ell)=1}^{T_i}\frac{\log^2 TK}{n_i(\ell)} \leq 4N\log^2 TK,$$

Therefore by substituting preceeding two inequalities in (C.14), we obtain that

$$\mathsf{Reg}_1(T,\mathbf{v}) \leq C\sqrt{\frac{NT}{K}},$$

for some constant C. The result follows from this inequality and (C.10).

Appendix D

Thompson Sampling Approach for Attribute Based Learning

First, we introduce some notation which we will use throughout this section and establish some strutural results that will play a key role in proving

D.1 Notation and Key Structural Results

For the rest of this section, we will use the following notations.

1.
$$M_t \triangleq \sum_{\tau=1}^{t-1} \sum_{i \in S_\tau} x_i x_i' \tag{D.1}$$

2. For any $\mathbf{v} \in \mathbb{R}^n$ and $i \in S$, let
$$\begin{aligned} p_S(v_i) &\triangleq \frac{e^{v_i}}{1 + \sum_{j \in S} e^{v_j}}, \\ g_t(\theta) &\triangleq \sum_{\tau=1}^{t-1} \sum_{i \in S_\tau} p_{S_\tau}(\theta \cdot x_i) \cdot x_i. \end{aligned} \tag{D.2}$$

3. Let $\mathbb{1}_i(t)$ be the indicator random variable corresponding to the event that item i has been clicked at time t.

4. Let θ^t_{MLE} be the MLE estimate of θ_* at time t. From the first order conditions we have,
$$\sum_{\tau=1}^{t-1} \sum_{i \in S_\tau} p_{S_\tau}(\theta^t_{\mathsf{MLE}} \cdot x_i) x_i = \sum_{\tau=1}^{t-1} \sum_{i \in S_\tau} \mathbb{1}_i(\tau) \tag{D.3}$$

5.
$$\begin{aligned}\eta_\tau &\triangleq \sum_{i\in S_\tau} \left(p_{S_\tau}(\theta_* \cdot x_i) - \mathbb{1}(\tau)\right) \cdot x_i,\\ \xi_t &\triangleq g_t(\theta_*) - g_t(\theta^t_{\mathsf{MLE}}).\end{aligned} \tag{D.4}$$

6. Let $\mathcal{F}_\tau$ be the filtration (history) associated with the policy upto epoch τ. Note that by definition of choice model, η_τ is a martingale adapted to $(\mathcal{F}_\tau)$.

7. From (D.3), we have,
$$\xi_t = \sum_{\tau=1}^{t-1} \eta_\tau \tag{D.5}$$

8.
$$\begin{aligned}\tilde{v}^t_i &= \exp\left(\theta^t_{\mathsf{MLE}} \cdot x_i + \alpha \cdot \theta^t_{\max} \|x_i\|_{H_t^{-1}}\right),\\ \hat{v}^t_i &= \exp\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right),\\ v_i &= \exp\left(\theta_* \cdot x_i\right).\end{aligned} \tag{D.6}$$

9. Following (**??**), we define the analysis epoch.
$$\begin{aligned}\mathcal{T} &= \left\{t : \tilde{v}^t_i(t) \geq v_i \text{ for all } i \in S^*\right\},\\ \mathsf{succ}(\tau) &= \min\{\bar{\tau} \in \mathcal{T} : \bar{\tau} > \tau\}\\ \mathcal{E}^{\mathsf{An}}(t) &= \{\tau : \tau \in (t, \mathsf{succ}(t))\} \text{ for all } t \in \mathcal{T}.\end{aligned} \tag{D.7}$$

D.1.1 Key Technical Lemmas

In the following result, we show that the inner product of any real valued vector ρ and the martingale η_τ is bounded.

Lemma D.1. *For any* $\rho \in \mathbb{R}^d$, $|\rho \cdot \eta_\tau| < \sqrt{2\sum_{i\in S_\tau} (\rho \cdot x_i)^2}$.

Proof. By definition of η_τ, we have

$$|\rho \cdot \eta_\tau|^2 = |\sum_{i\in S_\tau} \rho \cdot x_i \left(p_{S_\tau}(\theta_* \cdot x_i) - \mathbb{1}\left(\text{ i is clicked at time } \tau\right)\right)|^2.$$

From Cauchy-Schwartz inequality, it follows that,

$$|\rho \cdot \eta_\tau|^2 \leq \sum_{i\in S_\tau} (\rho \cdot x_i)^2 \sum_{i\in S_\tau} \left(p_{S_\tau}(\theta_* \cdot x_i) - \mathbb{1}_i(\tau)\right)^2.$$

We have for any non-negative numbers a and b, we have $(a-b)^2 \leq a^2 + b^2$. Noting that $p_{S_\tau}(.)$ and $\mathbb{1}(.)$ are non-negative numbers, we have,

$$|\rho \cdot \eta_\tau|^2 \leq \sum_{i \in S_\tau} (\rho \cdot x_i)^2 \left(\sum_{i \in S_\tau} p_{S_\tau}^2 (\theta_* \cdot x_i) + \sum_{i \in S_\tau} \mathbb{1}_i(\tau) \right).$$

Since at most one item can be clicked, we have $\sum_{i \in S_\tau} \mathbb{1}\left(\text{i is clicked at time } \tau\right) \leq 1$ and therefore we have,

$$|\rho \cdot \eta_\tau|^2 \leq \sum_{i \in S_\tau} (\rho \cdot x_i)^2 \left(\sum_{i \in S_\tau} p_{S_\tau}^2 (\theta_* \cdot x_i) + 1 \right).$$

Noting that $p_{S_\tau}(.)$ is non-negative, we have $\sum_{i \in S_\tau} p_{S_\tau}^2 (\theta \cdot x_i) \leq \left(\sum_{i \in S_\tau} p_{S_\tau}(\theta \cdot x_i) \right)^2 \leq 1$ and therefore it follows that,

$$|\rho \cdot \eta_\tau|^2 \leq 2 \sum_{i \in S_\tau} (\rho \cdot x_i)^2 .$$

□

We follow the approach of [22] and use the following facts to derive concentration properties of the MLE etimate. We refer the reader to Exercise 2.4 of [48] and Corollary 2.2 of [19] for the proof of these facts.

Fact D.1. *For any filtration $(\mathcal{F}_k; k \geq 0)$ and real valued martingale η_k adapted to $(\mathcal{F}_k)$. If $|\eta_k| \leq B$, then we have $\mathbb{E}\left[e^{\gamma \eta_k} | \mathcal{F}_{k-1}\right] \leq e^{\frac{\gamma^2 B^2}{2}}$.*

Fact D.2. *If A and B are random variables such that*

$$\mathbb{E}\left[\exp\left\{ \gamma A - \frac{\gamma^2}{2} B^2 \right\} \right] \leq 1 \ \text{ for all } \ \gamma \in R.$$

Then, for any $\delta \geq \sqrt{2}$ and $y > 0$,

$$\mathbb{P}\left(|A| \geq \delta \sqrt{(B^2 + y)\left(1 + \frac{1}{2} \log\left(\frac{B^2}{y} + 1 \right)\right)} \right) \leq \exp\left(-\frac{\delta^2}{2} \right).$$

In the following result, we establish a sub-gaussian property for the inner product of any real valued vector ρ and the martingale η_τ.

Corollary D.1. *For any $\rho \in \mathbb{R}^d$, we have for any τ,*

$$\mathbb{E}\left[e^{\gamma\rho\cdot\eta_\tau}|\mathcal{F}_{\tau-1}\right] \le e^{\gamma^2\rho\cdot x_i}. \tag{D.8}$$

$$\mathbb{E}\left[\exp\left(\gamma\rho\cdot\xi_\tau - \gamma^2\|\rho\|^2_{M_t}\right)\right] \le 1. \tag{D.9}$$

Proof. From Lemma D.1, we have that $\rho\cdot\eta_\tau$ is a martingale bounded by $\sqrt{2\sum_{i\in S_\tau}(\rho\cdot x_i)^2}$ and therefore from Fact D.2 we have the first result.

Now we will prove the second result. From the law of conditional expectations, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\gamma\rho\cdot\xi_\tau - \gamma^2\|\rho\|^2_{M_t}\right)\right] &= \mathbb{E}\left(\mathbb{E}\left[\exp\left(\gamma\rho\cdot\xi_\tau - \gamma^2\|\rho\|^2_{M_t}\right)\middle|\mathcal{F}_{t-1}\right]\right)\\
&= \mathbb{E}\left(\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{t-1}\gamma\left(\rho\cdot\eta_\tau - \sum_{i\in S_\tau}(\rho\cdot x_i)^2\right)\right)\middle|\mathcal{F}_{t-1}\right]\right)\\
= \mathbb{E}\left(\exp\left[\sum_{\tau=1}^{t-2}\gamma\left(\rho\cdot\eta_\tau - \sum_{i\in S_\tau}(\rho\cdot x_i)^2\right)\right]\cdot\mathbb{E}\left[\exp\left(\rho\cdot\eta_{t-1} - \sum_{i\in S_{t-1}}(\rho\cdot x_i)^2\right)\middle|\mathcal{F}_{t-1}\right]\right).
\end{aligned}$$

From (D.8), we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\gamma\rho\cdot\xi_\tau - \gamma^2\|\rho\|^2_{M_t}\right)\right] &\le \mathbb{E}\left(\exp\left[\sum_{\tau=1}^{t-2}\gamma\left(\rho\cdot\eta_\tau - \sum_{i\in S_\tau}(\rho\cdot x_i)^2\right)\right]\right).\\
&\vdots\\
&\le \mathbb{E}\left(\exp\left[\gamma\left(\rho\cdot\eta_1 - \sum_{i\in S_1}(\rho\cdot x_i)^2\right)\right]\right) \le 1.
\end{aligned}$$

□

Claim D.1. *For any θ and $\hat{\theta}$ and any assortment S, we have,*

$$\left|p_S\left(\theta\cdot x_i\right) - p_S\left(\hat{\theta}\cdot x_i\right)\right| \le 5\sum_{i\in S}\left|(\theta-\hat{\theta})\cdot x_i\right|.$$

Proof. From triangle inequality we have,

$$\begin{aligned}
&\left|p_S\left(\theta_*\cdot x_i\right) - p_S\left(\theta^t_{\mathsf{MLE}}\cdot x_i\right)\right|\\
&\qquad\le \sum_{i\in S}\left|\frac{\left(e^{\theta\cdot x_i} - e^{\hat{\theta}\cdot x_i}\right) + \sum_{j\in S}e^{\left(\theta\cdot x_j+\hat{\theta}\cdot x_i\right)} - e^{\left(\hat{\theta}\cdot x_j+\theta\cdot x_i\right)}}{(1+\sum_{j\in S}e^{\theta\cdot x_j})(1+\sum_{j\in S}e^{\hat{\theta}\cdot x_j})}\right|,\\
&\qquad\le \frac{\sum_{i\in S}\left|e^{\theta\cdot x_i} - e^{\hat{\theta}\cdot x_i}\right| + \sum_{i\in S}\sum_{j\in S}\left|e^{\left(\theta\cdot x_j+\hat{\theta}\cdot x_i\right)} - e^{\left(\hat{\theta}\cdot x_j+\theta\cdot x_i\right)}\right|}{(1+\sum_{j\in S}e^{\theta\cdot x_j})(1+\sum_{j\in S}e^{\hat{\theta}\cdot x_j})}.
\end{aligned}$$

For any $x < 0$, we have $1 - e^x \leq x$. Therefore, for any x, y, it follows that $|e^x - e^y| \leq \max\{e^x, e^y\}|x - y|$. Using this fact,

$$\left|p_S\left(\theta_* \cdot x_i\right) - p_S\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right)\right| \leq \sum_{i \in S} \frac{\max\{e^{\theta \cdot x_i}, e^{\hat{\theta} \cdot x_i}\} \left|(\theta - \hat{\theta}) \cdot x_i\right|}{(1 + \sum_{j \in S} e^{\theta \cdot x_j})(1 + \sum_{j \in S} e^{\hat{\theta} \cdot x_j})},$$
$$+ \frac{\sum_{i \in S} \sum_{j \in S} \max\left\{e^{\left(\hat{\theta} \cdot x_j + \theta \cdot x_i\right)}, e^{\left(\hat{\theta} \cdot x_i + \theta \cdot x_j\right)}\right\} \left(\left|(\theta - \hat{\theta}) \cdot x_i\right| + \left|(\theta - \hat{\theta}) \cdot x_j\right|\right)}{(1 + \sum_{j \in S} e^{\theta \cdot x_j})(1 + \sum_{j \in S} e^{\hat{\theta} \cdot x_j})}.$$

Using the fact that

$$\max\{e^{\theta \cdot x_i}, e^{\hat{\theta} \cdot x_i}\} \leq (1 + \sum_{j \in S} e^{\theta \cdot x_j})(1 + \sum_{j \in S} e^{\hat{\theta} \cdot x_j})$$
$$\sum_{j \in S} \max\{e^{\theta \cdot x_i + \hat{\theta} \cdot x_j}, e^{\hat{\theta} \cdot x_i + \theta \cdot x_j}\} \leq 2(1 + \sum_{j \in S} e^{\theta \cdot x_j})(1 + \sum_{j \in S} e^{\hat{\theta} \cdot x_j}),$$

we will have,

$$\left|p_S\left(\theta \cdot x_i\right) - p_S\left(\hat{\theta} \cdot x_i\right)\right| \leq 5 \sum_{i \in S} \left|(\theta - \hat{\theta}) \cdot x_i\right|.$$

□

We adapt the following result from [22].

Lemma D.2. *Let $t_0 \geq d + 1$. Then,*

$$\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|^2_{M_t^{-1}}, 1\} \leq 2d \log\left(\frac{c_x^2 KT}{\lambda_0}\right)$$

Proof. Enumerate the vectors $\{x_i\}_{i \in S_t, t=1,\cdots,T}$ as $y_1, \cdots, y_p$, where $p = \sum_{t=1}^{T} |S_t|$. Let $\hat{M}_\ell = \sum_{\tau=1}^{\ell-1} y_\tau y_\tau^T$. Furthermore, let $\hat{t}_0$ be such that

$$\mathsf{det}\left(\hat{M}_{\hat{t}_0}\right) = \mathsf{det}(M_{t_0}).$$

Since $y_\tau y_\tau^T$ is a positive semi-definite matrix for any τ we have,

$$\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|^2_{M_t^{-1}}, 1\} \leq \sum_{\ell=\hat{t}_0}^{p} \min\{\|y\|^2_{\hat{M}_t^{-1}}, 1\} \tag{D.10}$$

We will now focus on bounding the right hand side of (D.10). By definition of $\hat{M}_{\ell+1}$, we have

$$\begin{aligned}\mathsf{det}(\hat{M}_{\ell+1}) &= \mathsf{det}(\hat{M}_\ell + y_\ell y_{\ell+1}^T) = \mathsf{det}(\hat{M}_\ell)\left(I + \hat{M}_\ell^{-1/2} y_\ell \cdot \left(\hat{M}_\ell^{-1/2} y_\ell\right)'\right) \\ &= \mathsf{det}(\hat{M}_\ell)\left(1 + \|y_\ell\|_{\hat{M}_\ell^{-1}}\right) = \mathsf{det}(\hat{M}_{\hat{t}_0}) \prod_{\ell=\hat{t}_0}^{p} \left(1 + \|y\|_{\hat{M}_\ell^{-1}}\right),\end{aligned}$$

where the last line follows from the fact that $1 + \|y_\ell\|^2_{\hat{M}_\ell}$ is an eigenvalue of the matrix $I + \hat{M}_\ell^{-1/2} y_\ell \cdot \left(\hat{M}_\ell^{-1/2} y_\ell\right)'$ and that all other eigenvalues are equal to one. Thus, using the fact that $x \le 2\log(1+x)$ which holds for any $0 \le x \le 1$, we have

$$\begin{aligned}\sum_{\ell=\hat{t}_0} \min\{\|y\|^2_{\hat{M}_t^{-1}}, 1\} &\le 2\sum_{\ell=\hat{t}_0}^{p} \log\left(\|y\|^2_{\hat{M}_t^{-1}} + 1\right) \\ &= 2\log \prod_{\ell=\hat{t}_0}^{p} \left(\|y\|^2_{\hat{M}_t^{-1}} + 1\right) \\ &= 2\log\left(\frac{\mathsf{det}(\hat{M}_p)}{\mathsf{det}(\hat{M}_{t_0})}\right) = 2\log\left(\frac{\mathsf{det}(M_T)}{\mathsf{det}(M_{t_0})}\right).\end{aligned}$$

Note that the trace of M_t is upper bounded by Ktc_x^2. Then, since the trace of the positive definite matrix M_t is equal to the sum of its eigenvalues and $\mathsf{det}(M_t)$ is the product of its eigen values, we have $\mathsf{det}(M_t) \le (ktc_x^2)^d$. In addition, $\mathsf{det}(M_{t_0}) \ge \lambda_0^d$ since $t_0 \ge d+1$. Thus,

$$\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|^2_{M_t^{-1}}, 1\} \le 2d \log\left(\frac{Kc_x^2 T}{\lambda_0}\right).$$

□

Corollary D.2. *Let $t_0 \ge d+1$. Then,*

$$\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|_{M_t^{-1}}, 1\} \le 2d \log\left(\frac{Kc_x^2 T}{\lambda_0}\right)$$

Proof. Using the Cauchy-Schwarz inequality and Lemma D.2, we have

$$\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|_{M_t^{-1}}, 1\} \le \sqrt{KT}\sqrt{\sum_{t=t_0}^{T} \sum_{i \in S_t} \min\{\|x_i\|^2_{M_t^{-1}}, 1\}} \le \sqrt{2dKT \log\left(\frac{Kc_x^2 T}{\lambda_0}\right)}$$

□

D.2 Concentration Bounds

Lemma D.3. *For any $\delta > 0$, we have*

$$\mathbb{P}\left(\|\xi_t\|_{M_t^{-1}} \geq 12\sqrt{d\log\delta d}\right) \leq \frac{1}{d^8\delta^8}.$$

Proof. Let $\delta = 4\sqrt{d\log td}$, S_t be such that $S_t^2 = M_t$ and $\mathbf{e}_i$ be the i^{th} unit vector (i.e., for all $j \neq i$, $e_{ij} = 0$ and $e_{ii} = 1$). Noting that $\|\xi_t\|^2_{M_t^{-1}} = \sum_{i=1}^d \xi^T S_t^{-1}\mathbf{e}_i\mathbf{e}_i^T S_t^{-1}\xi_t$, we have,

$$\begin{aligned}
\mathbb{P}\left(\|\xi_t\|^2_{M_t^{-1}} \geq 3d\delta^2\right) &= \mathbb{P}\left[\sum_{i=1}^d \xi_t^T S_t^{-1} e_i e_i^T S_t^{-1}\xi_t \geq 9d\delta^2\right], \\
&\leq \mathbb{P}\left(\bigcup_{i=1}^d \left\{\xi_t^T S_t^{-1} e_i e_i^T S_t^{-1}\xi_t \geq 9\delta^2\right\}\right), \\
&\leq \sum_{i=1}^d \mathbb{P}\left(\left\{\xi_t^T S_t^{-1} e_i e_i^T S_t^{-1}\xi_t \geq 9\delta^2\right\}\right), \\
&= \sum_{i=1}^d \mathbb{P}\left(\left|\xi_t^T S_t^{-1} e_i\right| \geq 3\delta\right).
\end{aligned} \tag{D.11}$$

Let $\rho_i = S_t^{-1}\mathbf{e}_i$. From Corollary D.1 it follows that random variables $A = \rho_i \cdot \xi_t$ and $B = \|\rho_i\|^2_{M_t^{-1}}$ satisfy the conditions of Fact D.2. Using the fact $\|\rho_i\|^2_{M_t^{-1}} = 1$ and substituting $y = 2$ in Fact D.2, we have,

$$\mathbb{P}\left(\left|\xi_t^T S_t^{-1} e_i\right| \geq 3\delta\right) \leq \frac{1}{d^7 t^8}.$$

The result follows from the above inequality and (D.11). □

Now we will prove the finite time concentration bounds for the MLE estimate, θ^t_{MLE} of Algorithm 9. However, these bounds depend on the problem parameters. Specifically, we prove that,

Lemma D.4. *For any $\delta > 0$, we have,*

$$\mathbb{P}\left(\left|p_{S_\tau}\left(\theta_* \cdot x_i\right) - p_{S_\tau}\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right)\right| > \frac{60\sqrt{d\log\delta Kd}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}\right) \leq \frac{1}{K^7 d^7 \delta^8}.$$

Proof. We will follow the proof steps of [22]. From Fundamental Theorem of Calculus, we have

$$\xi_t = g_t(\theta_*) - g_t(\theta^t_{\mathsf{MLE}}) = G_t(\theta_* - \theta^t_{\mathsf{MLE}}),$$

where

$$G_t = \int_0^1 \nabla(s\theta_* + (1-s)\theta^t_{\mathsf{MLE}})ds.$$

From Claim D.1, we have

$$\begin{aligned}
\left|p_{S_\tau}(\theta_* \cdot x_i) - p_{S_\tau}\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right)\right| &\leq \min\left\{5\sum_{i\in S}|(\theta_* - \theta^t_{\mathsf{MLE}}) \cdot x_i|, 1\right\}, \\
&\leq \min\left\{5\sum_{i\in S}|G_t^{-1}\xi_t \cdot x_i|, 1\right\} \\
&\leq \min\left\{5\sum_{i\in S}\|\xi_t\|_{G_t^{-1}}\|x_i\|_{G_t^{-1}}, 1\right\}.
\end{aligned}$$

We have $\nabla g_t(\theta) = \sum_{\tau=1}^{t-1}\sum_{i\in S_\tau} x_i x_i' \dot{p}_{S_\tau}(\theta \cdot x_i)$. By Assumption 4.2, for any $\theta \in \Theta$ and i, we have $\dot{p}_{S_\tau}(\theta \cdot x_i) \geq c_\mu$. Therefore, we have $G_t \geq c_\mu M_t$ and it follows that,

$$\left|p_{S_\tau}(\theta_* \cdot x_i) - p_{S_\tau}\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right)\right| \leq \min\left\{\frac{5}{c_\mu}\sum_{i\in S_\tau}\|\xi_t\|_{M_t^{-1}}\|x_i\|_{M_t^{-1}}, 1\right\}.$$

If $\frac{5}{c_\mu}\sum_{i\in S_\tau}\|\xi_t\|_{M_t^{-1}}\|x_i\|_{M_t^{-1}} \geq 1$, then the result is trivial since the probability of the event under consideration is zero. Therefore, we focus on the case when

$$\frac{5}{c_\mu}\sum_{i\in S_\tau}\|\xi_t\|_{M_t^{-1}}\|x_i\|_{M_t^{-1}} < 1.$$

Therefore, we have,

$$\begin{aligned}
&\mathbb{P}\left(\left|p_{S_\tau}(\theta_* \cdot x_i) - p_{S_\tau}\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right)\right| > \frac{60}{c_\mu}\sum_{i\in S_\tau}\sqrt{d\log\delta Kd}\cdot\|x_i\|_{M_t^{-1}}\right) \\
&\leq \mathbb{P}\left(\sum_{i\in S_\tau}\|\xi_t\|_{M_t^{-1}}\|x_i\|_{M_t^{-1}} > \frac{12}{c_\mu}\sum_{i\in S_\tau}\sqrt{d\log\delta Kd}\cdot\|x_i\|_{M_t^{-1}},\right) \\
&\leq \sum_{i\in S_\tau}\mathbb{P}\left(\|\xi_t\|_{M_t^{-1}}\|x_i\|_{M_t^{-1}} > \frac{12}{c_\mu}\sqrt{d\log\delta Kd}\cdot\|x_i\|_{M_t^{-1}}\right).
\end{aligned}$$

The result follows from the above result and Lemma D.3 □

Lemma D.5. *For any $t \leq T$, we have for any $r > 0$,*

$$\mathsf{Pr}\left(|\theta^t_{\max}| > 4\sqrt{\log rK}\right) \leq \frac{1}{r^4K^3}.$$

Proof. By definition we have $\theta^t_{\max} = \max_{j=1,\cdots,K} \{\theta^{(j)}(t)\}$. Therefore, from union bound, we have,

$$\mathsf{Pr}\left(\bigcup_{j=1}^{K}\left\{\theta^j(t) > 2\sqrt{\log mk}\right\}\right) \leq \sum_{j=1}^{K}\mathsf{Pr}\left(\theta^j(t) > 4\sqrt{\log rK}\right).$$

The result follows from the above inequality and the following anti-concentration bound for the normal random variable $\theta^{(j)}(t)$ (see formula 7.1.13 in [1]).

$$\frac{1}{4\sqrt{\pi}} \cdot e^{-7z^2/2} < \mathsf{Pr}\left(|\theta^{(j)}(t)| > z\right) \leq \frac{1}{2}e^{-z^2/2}.$$

□

Corollary D.3. *For any assortment S and time t,*

$$\mathbb{E}\left(\left|p_S\left(\tilde{v}^t_i\right) - p_S\left(\hat{v}^t_i\right)\right|\right) \leq \frac{5\sqrt{\log TK}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}.$$

Proof. From Claim D.1, we have

$$\begin{aligned}
\left|p_S\left(\tilde{v}^t_i\right) - p_S\left(\hat{v}^t_i\right)\right| &\leq \min\left\{5\sum_{i\in S}\left|\log(\tilde{v}^t_i) - \log(\hat{v}^t_i)\right|, 1\right\}, \\
&= \min\left\{5\sum_{i\in S}|\theta^t_{\max}| \cdot \|x_i\|_{H_t^{-1}}, 1\right\}, \qquad \text{(D.12)} \\
&\leq \min\left\{\frac{5}{c_\mu}\sum_{i\in S}|\theta^t_{\max}| \cdot \|x_i\|_{M_t^{-1}}, 1\right\},
\end{aligned}$$

where the last inequality follows from Assumption 4.2. In Lemma D.5, we show that for any $r > 0, \mathsf{Pr}\left(|\theta^t_{\max}| > 4\sqrt{\log rK}\right) \leq \frac{1}{r^4K^3}$. Therefore it follows that,

$$\mathsf{Pr}\left(\left|p_S\left(\tilde{v}^t_i\right) - p_S\left(\hat{v}^t_i\right)\right| > \min\left\{\frac{5}{c_\mu}\sqrt{\log rK}\sum_{i\in S}\|x_i\|_{M_t^{-1}}, 1\right\}\right) \leq \frac{1}{r^4K^3}. \qquad \text{(D.13)}$$

If $\frac{5}{c_\mu}\sum_{i\in S_\tau}\sqrt{\log rK}\|x_i\|_{M_t^{-1}} \geq 1$, then the result is trivial since the probability of the event under consideration is zero. Therefore, we focus on the case when $\frac{5}{c_\mu}\sum_{i\in S_\tau}\sqrt{\log rK}\|x_i\|_{M_t^{-1}} < 1$.

Let $X = |p_S(\tilde{v}_i^t) - p_S(\hat{v}_i^t)|$ and $y = \frac{5}{c_\mu}\sum_{i\in S}\|x_i\|_{M_t^{-1}}$ Since X is a non-negative random variable, we have

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^\infty \mathsf{Pr}\{X \geq x\}\, dx, \\
&= \int_0^{y\sqrt{\log rK}} \mathsf{Pr}\{X \geq x\}\, dx + \int_{y\sqrt{\log rK}}^\infty \mathsf{Pr}\{X \geq x\}\, dx, \\
&\leq y\sqrt{\log rK} + \sum_{r=T}^\infty \int_{y(t)\sqrt{\log rK}}^{y\sqrt{\log(r+1)K}} \mathsf{Pr}\{X \geq x\}\, dx,
\end{aligned} \tag{D.14}$$

$$\begin{aligned}
&\overset{a}{\leq} y\sqrt{\log TK} + \sum_{r=T}^\infty \frac{N\sqrt{\log(rK+1)} - N\sqrt{\log rK}}{r^4K^3}, \\
&\leq \frac{5}{c_\mu}\sum_{i\in S_t}\|x_i\|_{M_t^{-1}}\sqrt{\log TK}
\end{aligned}$$

where the inequality (a) follows from (D.13). □

D.3 Anti-Concentration Property: Bounding the Length of the Analysis Epoch

Here, we prove that the expected length (and higher moments) of the analysis epoch (see D.7) is bounded by a constant. Specifically, we have the following result.

Lemma D.6. *Let $\mathcal{E}^{\mathsf{An}}$ be the group of consecutive epochs between an optimistic epoch t and the next optimistic epoch $\hat{t}$, excluding the epochs t and $\hat{t}$. Then, for any $p \in [1, 2]$, we have,*

$$\mathbb{E}^{1/p}\left[|\mathcal{E}^{\mathsf{An}}(t)|\right] \leq \frac{e^{12}}{K} + 30^{1/p}.$$

Proof. For notational brevity, we introduce some notation.

Notation.

- $r = \lfloor (q+1)^{1/p} \rfloor$, $z = \sqrt{\log(rk+1)}$, and for each $i = 1, \cdots, N$,

$$\hat{\sigma}_i(t) = \frac{60\sqrt{d\log d}}{c_\mu}\|x_i\|_{H_t^{-1}}$$

- Define events,

$$\begin{aligned} A_t &= \left\{ p_{S_\tau}\left(\tilde{v}_i^t\right) \geq p_{S_\tau}\left(\hat{v}_i^t\right) + z\hat{\sigma}_i(t) \ \text{ for all } i \in S^* \right\}, \\ B_t &= \left\{ p_{S_\tau}\left(\hat{v}_i^t\right) + z\hat{\sigma}_i(t) \geq p_{S_\tau}\left(v_i\right) \ \text{ for all } i \in S^* \right\}, \\ \mathcal{B}_t &= \bigcap_{\tau=t+1}^{t+r} B_\tau. \end{aligned} \tag{D.15}$$

We have,

$$\mathsf{Pr}\left\{ \left|\mathcal{E}^{\mathsf{An}}(t)\right|^p < q+1 \right\} = \mathsf{Pr}\left\{ |\mathcal{E}(t)| \leq r \right\}.$$

By definition, length of the analysis epoch, $\mathcal{E}^{\mathsf{An}}(t)$ less than r, implies that one of the algorithm epochs from $t+1, \cdots, t+r$ is optimistic. Hence we have,

$$\begin{aligned} \mathsf{Pr}\left\{ \left|\mathcal{E}^{\mathsf{An}}(t)\right| < r \right\} &= \mathsf{Pr}\left(\left\{ \{\tilde{v}_i^\tau \geq v_i \text{ for all } i \in S^*\} \text{ for some } \tau \in (t, t+r] \right\} \right), \\ &\geq \mathsf{Pr}\left(\left\{ \{\tilde{v}_i^\tau \geq \hat{v}_i^\tau + z\hat{\sigma}_i(t) \geq v_i \text{ for all } i \in S^*\} \text{ for some } \tau \in (t, t+r] \right\} \right). \end{aligned}$$

From (D.15), we have,

$$\begin{aligned} \mathsf{Pr}\left\{ \left|\mathcal{E}^{\mathsf{An}}(t)\right| < r \right\} &\geq \mathsf{Pr}\left(\bigcup_{\tau=t+1}^{t+r} A_\tau \cap B_\tau \right), \\ &= 1 - \mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c \cup B_\tau^c \right). \end{aligned} \tag{D.16}$$

We will now focus on the term, $\mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c \cup B_\tau^c \right)$,

$$\begin{aligned} \mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c \cup B_\tau^c \right) &= \mathsf{Pr}\left(\left\{ \bigcap_{\tau=t+1}^{t+r} A_\tau^c \cup B_\tau^c \right\} \cap \mathcal{B}_\tau \right) + \mathsf{Pr}\left(\left\{ \bigcap_{\tau=t+1}^{t+r} A_\tau^c \cup B_\tau^c \right\} \cap \mathcal{B}_\tau^c \right), \\ &\leq \mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c \right) + \mathsf{Pr}(\mathcal{B}_\tau^c), \\ &\leq \mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c \right) + \sum_{t=\tau+1}^{\tau+r} \mathsf{Pr}(B_\tau^c), \end{aligned} \tag{D.17}$$

where the inequality follows from union bound. Note that,

$$\begin{aligned} \mathsf{Pr}(B_\tau^c) &= \mathsf{Pr}\left(\bigcup_{i \in S^*} \left\{ \hat{v}_i^t + z\hat{\sigma}_i(\tau) < v_i \right\} \right), \\ &\leq \sum_{i \in S^*} \mathsf{Pr}\left(\hat{v}_i^t + z\hat{\sigma}_i(\tau) < v_i \right). \end{aligned} \tag{D.18}$$

Noting that $\sqrt{d \log d \cdot \log rK} \geq \log(drK)$. Noting that $\|x_i\|_{H_t^{-1}} \geq \|x_i\|_{M_t^{-1}}$, we substitute $\delta = rK$ in Lemma D.4 to obtain,

$$\mathsf{Pr}\left(\hat{v}_i^t + z\hat{\sigma}_i(t) < v_i\right) \leq \frac{1}{(dK)^7 r^8}. \tag{D.19}$$

From (D.18) and (D.19), we obtain,

$$\begin{aligned} \mathsf{Pr}(B_\tau^c) &\leq \frac{1}{(dK)^7 r^8}, \text{ and} \\ \sum_{t=\tau+1}^{\tau+r} \mathsf{Pr}(B_t^c) &\leq \frac{1}{(drK)^7}. \end{aligned} \tag{D.20}$$

We will now use the tail bounds for Gaussian random variables to bound the probability $\mathsf{Pr}(A_t^c)$. For any Gaussian random variable, Z with mean μ and standard deviation σ, we have,

$$Pr(Z > \mu + x\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-x^2/2}.$$

Noting that e^x is a monotonic function, by construction of $\mu_i^{(j)}(t)$ in Algorithm 9. We have,

$$\mathsf{Pr}\left(\bigcap_{\tau=t+1}^{t+r} A_\tau^c\right) = \mathsf{Pr}\Big(\theta^{(j)}(t) \leq z \text{ for all } t \in (\tau, \tau+r] \text{ and for all } j = 1, \cdots, K\Big).$$

Since $\theta^{(j)}(t),\ j = 1, \cdots, K,\ t = \tau+1, \cdots, \tau+r$ are independently sampled from the distribution, $\mathcal{N}(0,1)$, we have,

$$\begin{aligned} \mathsf{Pr}\left\{\bigcap_{t=\tau+1}^{\tau+r} A_t^c\right\} &\leq \left[1 - \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\log(rK+1)}}{\log(rK+1)+1} \cdot \frac{1}{\sqrt{rK+1}}\right)\right]^{rK} \\ &\leq \exp\left(-\frac{r^{1/2}}{\sqrt{2\pi}} \frac{2\sqrt{\log(rK+1)}}{4\log(rK+1)+1}\right) \\ &\leq \frac{1}{(rK)^{2.2}} \text{ for any } r \geq \frac{e^{12}}{K}. \end{aligned} \tag{D.21}$$

From (D.16), (D.17), (D.20) and (D.21), we have that,

$$\mathsf{Pr}\left\{\left|\mathcal{E}^{\mathsf{An}}(t)\right| < r\right\} \geq 1 - \frac{1}{(rK)^{2.1}} - \frac{1}{(rK)^{2.2}} \text{ for any } r \geq \frac{e^{12}}{K}.$$

From definition $r \geq (q+1)^{1/p} - 1$, we obtain

$$\mathsf{Pr}\left\{\left|\mathcal{E}^{\mathsf{An}}(t)\right|^p < q+1\right\} \geq 1 - \frac{1}{(q+1)^{2.1/p}-1} - \frac{1}{(q+1)^{2.2/p}-1} \text{ for any } q \geq \left(\frac{e^{12}}{K}+1\right)^p.$$

Therefore, we have,

$$\begin{aligned}
\mathbb{E}\left[\left|\mathcal{E}^{\mathsf{An}}(t)\right|^p\right] &= \sum_{q=0}^{\infty} \mathsf{Pr}\left\{|\mathcal{E}(\tau)|^p \geq t\right\}, \\
&\leq \left(\frac{e^{12}}{K}+1\right)^p + \sum_{q=\frac{e^{12p}}{K^p}}^{\infty} \mathsf{Pr}\left\{|\mathcal{E}(t)|^p \geq t\right\}, \\
&\leq e^{12p} + \sum_{q=\frac{e^{12p}}{K^p}}^{\infty} \frac{1}{t^{2.1/p}} + \frac{1}{t^{2.2/p}} \leq \left(\frac{e^{12}}{K}+1\right)^p + 30.
\end{aligned}$$

The result follows from the above inequality. □

D.4 Proof of Theorem 4.1

Notations. For the sake of brevity, we introduce some notations.

- For any $t, \tau \leq T$, define ΔR_t and $\Delta R_{t,\tau}$ in the following manner

$$\Delta R_t \triangleq \left[R(S_t, \tilde{\boldsymbol{v}}^t) - R(S_t, \mathbf{v})\right]$$
$$\Delta R_{t,\tau} \triangleq \left[R(S_t, \tilde{\boldsymbol{v}}^{\boldsymbol{t}}) - R(S_t, \tilde{\boldsymbol{v}}^{\tau})\right]$$

- Let $\mathcal{A}_0$ denote the complete set Ω and for all $t = 1, \ldots, T$, define events $\mathcal{A}_t$ as

$$\mathcal{A}_t = \left\{ \left| p_{S_t}\left(\theta^t_{\mathsf{MLE}} \cdot x_i\right) - p_{S_t}\left(\theta_* \cdot x_i\right)\right| > \frac{60\sqrt{d \log tKd}}{c_\mu} \sum_{i \in S_t} \|x_i\|_{M_t^{-1}} \text{ for some } i \right\}$$

- $\hat{\mathcal{A}} = \bigcup_{\tau=1}^{t} \mathcal{A}_\tau$

$$\begin{aligned}
\mathsf{Reg}(T, \theta_*) :&= \mathbb{E}\left[\sum_{t=1}^{T} \left(R(S^*, \mathbf{v}) - R(S_t, \mathbf{v})\right)\right] \\
&= \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[R(S^*, \mathbf{v}) - R(S_t, \tilde{\mathbf{v}}^t)\right]}_{\mathsf{Reg}_1(T, \mathbf{v})} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[R(S_t, \tilde{\mathbf{v}}^t) - R(S_t, \mathbf{v})\right]}_{\mathsf{Reg}_2(T, \mathbf{v})}.
\end{aligned} \tag{D.22}$$

We will complete the proof by bounding the two terms in (D.22). We first focus on bounding $\mathsf{Reg}_2(T, \mathbf{v})$.

Bounding $\mathsf{Reg}_2(T, \mathbf{v})$: We have the second term in (D.22) reformulated as

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left\{\sum_{t=1}^{T} \Delta R_t\right\}. \tag{D.23}$$

Noting that $\mathcal{A}_t$ is a "low probability" event, we analyze the Regret in two scenarios, one when $\mathcal{A}_t$ is true and another when $\mathcal{A}_t^c$ is true. More specifically,

$$\mathbb{E}\left(\Delta R_t\right) = \mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}) + \Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right],$$

Using the fact that $R(S_t, \tilde{\mathbf{v}}^t)$ and $R(S_t, \mathbf{v})$ are both bounded by one, we have

$$\mathbb{E}\left(\Delta R_t\right) \leq \mathsf{Pr}(\mathcal{A}_{t-1}) + \mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right].$$

Substituting $\delta = t$ in Lemma D.4, we obtain that $\mathsf{Pr}(\mathcal{A}_{t-1}) \leq \frac{1}{K^7 d^7 t^8}$. Therefore, it follows that,

$$\mathbb{E}\left\{\Delta R_t\right\} \leq \frac{1}{t^2} + \mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right]. \tag{D.24}$$

From triangle inequality, we have,

$$\left|R(S_t, \tilde{\mathbf{v}}^{\mathbf{t}}) - R(S_t, \mathbf{v})\right| \leq \sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right|$$

Hence, it follows that,

$$\mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right] \leq \mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right| \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right].$$

From triangle inequality, we have

$$\begin{aligned}\mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right] \leq\ & \mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(\hat{v}_i^t)\right| \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right] \\ & + \mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\hat{v}_i^t) - p_{S_t}(v_i)\right| \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right],\end{aligned}$$

and from the definition of the event $\mathcal{A}_{t-1}^c$, it follows that,

$$\begin{aligned}\mathbb{E}\left[\Delta R_t \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right] \leq \mathbb{E}\left[\sum_{i\in S_t}\left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(\hat{v}_i^t)\right|\right] \\ + \frac{60\sqrt{d\log tKd}}{c_\mu}\mathbb{E}\left[\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}\right].\end{aligned} \tag{D.25}$$

From Corollary D.1, we have

$$\mathbb{E}\left[\sum_{i\in S_t}\left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(\hat{v}_i^t)\right|\right] \leq \frac{5\sqrt{\log TK}}{c_\mu}\mathbb{E}\left[\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}\right].$$

From (D.23), (D.24) and (D.25), we have,

$$\mathsf{Reg}_2(T, \mathbf{v}) \leq \frac{65\sqrt{d\log dTK}}{c_\mu}\mathbb{E}\left(\sum_{t=1}^{T}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}\right).$$

From Corollary D.2, we have

$$\mathsf{Reg}_2(T, \mathbf{v}) \leq 130d\frac{1}{c_\mu}\sqrt{\log\left(\frac{c_x^2 KT}{\lambda_0}\right)TK\log d\log TK} \tag{D.26}$$

We will now focus on the first term in (D.22).

Bounding $\mathsf{Reg}_1(T, \mathbf{v})$: Recall, $\mathcal{T}$ is the set of optimistic epoch and the sanalysis epoch $\mathcal{E}^{\mathsf{An}}(t)$ is the set of non-optimistic epochs between t^{th} epoch and the subsequent optimistic epoch. Therefore, we can reformulate $\mathsf{Reg}_1(T, \mathbf{v})$ as,

$$\mathsf{Reg}_1(T, \mathbf{v}) = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}(t\in\mathcal{T})\cdot\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(t)}\left(R(S^*, \mathbf{v}) - R(S_\tau, \tilde{\boldsymbol{v}}^\tau)\right)\right].$$

Note that for any τ, by algorithm design we have that S_τ is the optimal set when the MNL parameters are given by $\tilde{\mathbf{v}}^\tau$, i.e.,

$$R(S_\tau, \tilde{\mathbf{v}}^{\tau)} \geq R(S_t, \tilde{\mathbf{v}}^\tau).$$

Similarly, we have $R(S_t, \tilde{\mathbf{v}}^t) \geq R(S^*, \tilde{\mathbf{v}}^t)$. Furthermore, since t is an optimistic epoch, from the restricted monotonicity property (see Lemma 2.3), we have $R(S^*, \tilde{\mathbf{v}}^t) \geq R(S^*, \mathbf{v})$. Hence, for any $t \in \mathcal{T}$, we have

$$R(S_t, \tilde{\mathbf{v}}^t) \geq R(S^*, \mathbf{v}).$$

Therefore, it follows that,

$$\mathsf{Reg}_1(T, \mathbf{v}) \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(t \in \mathcal{T}) \cdot \sum_{\tau \in \mathcal{E}^{\mathsf{An}}(t)} \Delta R_{t,\tau}\right]. \tag{D.27}$$

Following the approach of bounding $\mathsf{Reg}_2(T, \mathbf{v})$, we analyze the first term, $\mathsf{Reg}_1(T, \mathbf{v})$ in two scenarios, one when $\hat{\mathcal{A}}_t$ is true and another when $\hat{\mathcal{A}}_t^c$ is true. More specifically,

$$\mathbb{E}\left(\sum_{\tau \in \mathcal{E}^{\mathsf{An}}(t)} \Delta R_{t,\tau}\right) = \mathbb{E}\left[\sum_{\tau \in \mathcal{E}^{\mathsf{An}}(t)} \Delta R_{t,\tau} \cdot \mathbb{1}(\mathcal{A}_{t-1}) + \Delta R_{t,\tau} \cdot \mathbb{1}(\mathcal{A}_{t-1}^c)\right].$$

From triangle inequality, we have,

$$\begin{aligned}\left|R(S_t, \tilde{\mathbf{v}}^{\mathbf{t}}) - R(S_t, \mathbf{v}^{\tau})\right| &= \left|R(S_t, \tilde{\mathbf{v}}^{\mathbf{t}}) - R(S_t, \mathbf{v}) + R(S_t, \mathbf{v}) - R(S_t, \tilde{\mathbf{v}}^{\tau})\right| \\ &\leq \sum_{i \in S_t} \left(\left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right| + \left|p_{S_t}(\tilde{v}_i^{\tau}) - p_{S_t}(v_i)\right|\right)\end{aligned}$$

Hence, it follows that,

$$\begin{aligned}\mathbb{E}\left[\Delta R_{t,\tau} \cdot \mathbb{1}(\hat{\mathcal{A}}_{t-1}^c)\right] \leq\ & \mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right| \cdot \mathbb{1}(\hat{\mathcal{A}}_{t-1}^c)\right] \\ &+ \mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\hat{v}_i^{\tau}) - p_{S_t}(v_i)\right| \cdot \mathbb{1}(\hat{\mathcal{A}}_{t-1}^c)\right].\end{aligned}$$

Using the fact that $R(S_t, \tilde{\mathbf{v}}^t)$ and $R(S_t, \tilde{\boldsymbol{v}}^{\tau})$ are bounded by one, we have

$$\mathbb{E}\left(\Delta R_{t,\tau}\right) \leq \mathbb{E}\left[\mathbb{1}(\hat{\mathcal{A}}_{t-1}) + \sum_{i \in S_t} \left(\left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right| + \left|p_{S_t}(\hat{v}_i^{\tau}) - p_{S_t}(v_i)\right|\right) \cdot \mathbb{1}(\hat{\mathcal{A}}_{t-1}^c)\right] \tag{D.28}$$

From Corollary D.1, we have

$$\mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(\hat{v}_i^t)\right|\right] \leq \frac{5\sqrt{\log TK}}{c_\mu} \sum_{i \in S_t} \min\left\{\|x_i\|_{M_t^{-1}}, 1\right\},$$

and by definition of the set $\mathbb{1}\left(\hat{\mathcal{A}}_{t-1}\right)$, we have

$$\mathbb{E}\left[\sum_{i \in S_t} \left|p_{S_t}(\tilde{v}_i^t) - p_{S_t}(v_i)\right|\right] \leq \frac{60\sqrt{d \log tKd}}{c_\mu} \sum_{i \in S_t} \min\left\{\|x_i\|_{M_t^{-1}}, 1\right\}.$$

Therefore from the above two inequalities, it follows that,

$$\mathbb{E}\left[\sum_{i\in S_t}\left|p_{S_t}(\tilde{v}_i^t)-p_{S_t}(v_i)\right|\right]\leq\frac{65\sqrt{d\log TKd}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\}.$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in S_t}\left|p_{S_t}(\tilde{v}_i^\tau)-p_{S_t}(v_i)\right|\right]&\leq\frac{65\sqrt{d\log TKd}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\\&\leq\frac{65\sqrt{d\log TKd}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\},\end{aligned}\tag{D.29}$$

where the last inequality follows form the fact that $\tau>t$ (which implies $M_t\leq M_\tau$). Substituting (D.29) in (D.28), we have,

$$\begin{aligned}\mathbb{E}\left(\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(t)}\Delta R_{t,\tau}\right)&\leq\mathbb{E}\left[|\mathcal{E}^{\mathsf{An}}(t)|\cdot\mathbb{1}(\hat{\mathcal{A}}_{t-1})\right]\\&\quad+\mathbb{E}\left[\frac{130|\mathcal{E}^{\mathsf{An}}(t)|\sqrt{d\log TKd}}{c_\mu}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right].\end{aligned}\tag{D.30}$$

Now we will focus on the first term in the above inequality. From Cauchy-Schwarz inequality, we have,

$$\mathbb{E}\left[|\mathcal{E}^{\mathsf{An}}(t)|\cdot\mathbb{1}(\hat{\mathcal{A}}_{t-1})\right]\leq\mathbb{E}^{1/2}\left(|\mathcal{E}^{\mathsf{An}}(t)|^2\right)\cdot\mathsf{Pr}^{1/2}\left(\hat{\mathcal{A}}_{t-1}\right),$$

Substituting $\delta=t$ in Lemma D.4 and using union bound, we obtain that

$$\mathsf{Pr}(\hat{\mathcal{A}}_{t-1})\leq\frac{1}{K^7d^7t^7}.$$

In Lemma D.6, we show that

$$\mathbb{E}^{1/2}\left[\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^2\right]\leq\frac{e^{12}}{K}+30^{1/2}.$$

Therefore, from the above three inequalities, it follows that,

$$\mathbb{E}\left[\sum_{t=1}^{T}|\mathcal{E}^{\mathsf{An}}(t)|\cdot I(\mathcal{A}_{t-1})\right]\leq\frac{e^{13}}{K}.\tag{D.31}$$

Now focusing on the second term in (D.30). We have,

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}}|\mathcal{E}^{\mathsf{An}}(t)|\cdot\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right]\leq\mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right]\tag{D.32}$$

From Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right)\\
&\leq \mathbb{E}\left(\sqrt{\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|^2\cdot\sum_{t\in\mathcal{T}}\sum_{i\in S_t}\min\left\{\|x_i\|_{M_t^{-1}},1\right\}^2}\right) \qquad \text{(D.33)}\\
&\leq \mathbb{E}\left(\sqrt{\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|^2\cdot\sum_{t\in\mathcal{T}}\sum_{i\in S_t}\min\left\{\|x_i\|^2_{M_t^{-1}},1\right\}}\right)
\end{aligned}$$

From Jensen's inequality, for any two random variables X,Y, we have

$$\mathbb{E}(X\cdot Y)\leq\mathbb{E}^{1/2}(X^2)\mathbb{E}^{1/2}(Y^2).$$

Therefore substituting

$$\begin{aligned}
X&=\sqrt{\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|^2},\\
Y&=\sqrt{\sum_{t\in\mathcal{T}}\sum_{i\in S_t}\left(\|x_i\|^2_{M_t^{-1}}+1\right)},
\end{aligned}$$

we have,

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right)\\
&\leq\sqrt{K}\mathbb{E}^{1/2}\left(\sum_{t\in\mathcal{T}}|\mathcal{E}^{\mathsf{An}}(t)|^2\right)\cdot\mathbb{E}^{1/2}\left(\sum_{t\in\mathcal{T}}\sum_{i\in S_t}\min\left\{\|x_i\|^2_{M_t^{-1}},1\right\}\right).
\end{aligned}$$

From Lemma D.6, it follows that for some constant C,

$$\mathbb{E}\left(\sum_{t\in\mathcal{T}}\sum_{i\in S_t}|\mathcal{E}^{\mathsf{An}}(t)|\min\left\{\|x_i\|_{M_t^{-1}},1\right\}\right)\leq C\sqrt{KT}\mathbb{E}^{1/2}\left(\sum_{t\in\mathcal{T}}\sum_{i\in S_t}\min\left\{\|x_i\|^2_{M_t^{-1}},1\right\}\right).$$

From Lemma D.2, we have

$$\mathsf{Reg}_1(T,\mathbf{v})\leq C\sqrt{d\log\left(\frac{c_x^2KT}{\lambda_0}\right)TK} \qquad \text{(D.34)}$$

The result follows from (D.34) and (D.26).

Appendix E

Static Assortment Optimization

E.1 Proof of Lemma 5.1

Proof. Consider any extreme point, $(\mathbf{p}, p_0)$ of $\mathcal{P}$. Note that there must be $n+1$ linearly independent and active constraints. Let

$$\begin{aligned} S_0 &= \{i \mid p_i = 0\},\ S_1 = \{i \mid p_i = p_0\}, \\ T &= \{i \mid \sum_{j=1}^{n} a_{ij} p_j = b_i p_0\} \\ k &= |S_0| + |S_1| + |T|. \end{aligned} \tag{E.1}$$

We claim that $k \geq n$. This follows as we have $|S_0| + |S_1|$ linearly independent and active constraints from the constraint set $S_0 \cup S_1$, $|T|$ active constraints from the constraint set T and one active constraint from the constraint $\sum_{i=0}^{n} v_i p_i = 1$. Hence the total number of linearly independent and active constraints at $(\mathbf{p}, p_0)$ is at most $k+1$.

Without loss of generality we can assume that $k = n$; since $k > n$ implies that $|S_0| + |S_1| + |T| + 1 > n+1$, making some constraints in T redundant. Let

$$\mathbf{B}_p = \begin{bmatrix} \mathbf{A}(T) & -\mathbf{b}(T) \\ \mathbf{I}(S_0) & \mathbf{0} \\ \mathbf{I}(S_1) & -\mathbf{e} \\ \mathbf{v}' & v_0 \end{bmatrix}, \quad \mathbf{B}_x = \begin{bmatrix} \mathbf{A}(T) \\ \mathbf{I}(S_0) \\ \mathbf{I}(S_1) \end{bmatrix}, \quad \mathbf{b}_x = \begin{bmatrix} \mathbf{b}(T) \\ \mathbf{0} \\ \mathbf{e} \end{bmatrix}, \tag{E.2}$$

Note that $\mathbf{B}_p$ is the basis matrix corresponding to the extreme point $(\mathbf{p}, p_0)$. Hence, $\mathbf{B}_p$ is full rank. For the sake of contradiction, assume that $\mathbf{B}_x$ is not full rank. There

there exists $\boldsymbol{\lambda} \in R^n, \boldsymbol{\lambda} \neq \mathbf{0}$ such that $\boldsymbol{\lambda}'\mathbf{B}_x = \mathbf{0}$, then we have

$$\begin{bmatrix} \boldsymbol{\lambda}' & 0 \end{bmatrix} \mathbf{B}_p = \begin{bmatrix} \boldsymbol{\lambda}'\mathbf{B}_x & -\boldsymbol{\lambda}'\mathbf{b}_x \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\boldsymbol{\lambda}'\mathbf{b}_x \end{bmatrix},$$

which implies

$$\begin{bmatrix} \boldsymbol{\lambda}' & 0 \end{bmatrix} \mathbf{B}_p \begin{bmatrix} \mathbf{p} \\ p_0 \end{bmatrix} = -p_0 \boldsymbol{\lambda}'\mathbf{b}_x,$$

Since $\mathbf{B}_p$ is a full rank, we have $\boldsymbol{\lambda}'\mathbf{b}_x \neq 0$ and $p_0 \neq 0$, contradicting that,

$$\mathbf{B}_p \begin{bmatrix} \mathbf{p} \\ p_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}.$$

Hence, $\mathbf{B}_x$ is a full rank and $\mathbf{x} = \mathbf{p}/p_0$ is a basic feasible solution in $\mathcal{Q}$ corresponding to the basis matrix $\mathbf{B}_x$.

Conversely, consider $\mathbf{x}$, any extreme point of $\mathcal{Q}$. Let

$$p_0 = \frac{1}{v_0 + \mathbf{v}'\mathbf{x}}, \quad \mathbf{p} = p_0\mathbf{x}.$$

Clearly $(\mathbf{p}, p_0) \in \mathcal{P}$. We define the quantities S_0, S_1, T, k as in (E.1) and $\mathbf{B}_p, \mathbf{B}_x, \mathbf{b}_x$ as in (E.2). Using similar arguments, we can assume $k = n$ without loss of generality. Since $\mathbf{x}$ is a basic feasible solution corresponding to the basis $\mathbf{B}_x$, $\mathbf{B}_x$ is full rank.

For the sake of contradiction, suppose $\mathbf{B}_p$ is not full rank. Then there exists $\boldsymbol{\lambda} \in R^{n+1}, \boldsymbol{\lambda} \neq \mathbf{0}$ such that $\boldsymbol{\lambda}'\mathbf{B}_p = \mathbf{0}$. Therefore,

$$\boldsymbol{\lambda}'\mathbf{B}_p \begin{bmatrix} \mathbf{p} \\ p_0 \end{bmatrix} = 0$$

which implies

$$(\boldsymbol{\lambda}([n]))'(\mathbf{B}_x\mathbf{p} + p_0\mathbf{b}_x) + \lambda_{n+1}(\mathbf{v}'\mathbf{p} + v_0p_0) = 0.$$

Since $\mathbf{B}_x\mathbf{x} = \mathbf{b}_x$, we have $\mathbf{B}_x\mathbf{p} + p_0\mathbf{b}_x = \mathbf{0}$ and $\lambda_{n+1} = 0$. Note that, $\boldsymbol{\lambda}'\mathbf{B}_p = \mathbf{0}$ and

$$\boldsymbol{\lambda}'\mathbf{B}_p = \begin{bmatrix} \boldsymbol{\lambda}([n])'\mathbf{B}_x + \lambda_{n+1}\mathbf{v}' & \boldsymbol{\lambda}([n])'\mathbf{b}_x + \lambda_{n+1}v_0 \end{bmatrix}$$

Therefore $\boldsymbol{\lambda}([n])'\mathbf{B}_x = \mathbf{0}$, contradicting the fact that $\mathbf{B}_x$ is full rank. Hence, $\mathbf{B}_p$ is a full rank matrix and $(\mathbf{p}, p_0)$ is the basic feasible solution corresponding to the basis matrix $\mathbf{B}_p$. This completes the proof. □

E.2 Proof of Theorem 5.2

Proof. Let $(\mathbf{p}^*, p_0^*)$ be an optimal solution to (5.7). Let $S = \{i \geq 1 \mid p_i^* > 0\}$. In Steps 3-6 of the algorithm we consider all the solutions that have strictly less than $\left\lfloor \frac{\ell}{\epsilon} \right\rfloor$. Hence, without loss of generality assume that $|S| \geq \left\lfloor \frac{\ell}{\epsilon} \right\rfloor$ and $S = \{1, 2, \ldots, k\}$ for some $k \geq \left\lfloor \frac{\ell}{\epsilon} \right\rfloor$ and $a_k \leq a_{k-1} \leq \cdots \leq a_1$. Also, let $S_1 = \{1, 2, \ldots, k^*\}$ where $k^* = \left\lfloor \frac{\ell}{\epsilon} \right\rfloor$.

Note that $p_1^* = p_2^* = \cdots = p_k^* = p_0^*$. Therefore,

$$a_k p_k^* \leq a_{k-1} p_{k-1}^* \leq \cdots \leq a_1 p_1^*,$$

which implies

$$a_{k^*+1} p_{k^*+1}^* < \frac{\epsilon}{\ell} R^*.$$

Now consider a feasible point of (5.7), $(\mathbf{p}_1, p_{10})$ defined as

$$p_{10} = \frac{1}{v_0 + \sum_{i \in S_1} v_i}, \quad p_{1i} = \begin{cases} p_{10} & \text{if } i \in S_1 \\ 0 & \text{otherwise.} \end{cases}$$

implying $p_{1i}^* < p_{1i}$ for all $i \in S_1$. and since $(\mathbf{p}_1, p_{10})$ is a feasible point to (5.7), it follows that

$$\sum_{i \in S_1} a_i p_{1i} = \sum_{i=1}^{n} a_i p_{1i} < R^* \text{ which implies } \sum_{i \in S_1} a_i p_{1i}^* \leq R^*$$

By construction of $z_{\mathsf{LP}}(1)$, we must also have $p_{1i}^* = 0$ for every $i > k^*$ and $a_{k^*} \leq a_i$, implying

$$a_i p_{1i}^* < a_{k^*} p_{k^*}^* < \frac{\epsilon}{\ell} R^* \quad \text{for all } i > k^*.$$

Observe that $z_{\mathsf{LP}}(1) \geq R^*$ and the variables i in the extreme point $(\mathbf{p}_1^*, p_{10}^*)$ that can be fractional are $i > k^*$. Therefore, $a_i p_{1i}^* < (\epsilon \cdot R^*)/\ell$ for all $i \in \mathcal{F}(\mathbf{p}_1^*, p_{10}^*)$. Thus by Lemma 5.1, it follows that $\sum_{i \in \mathcal{F}(p^*)} a_i p_i^*(1) < \epsilon R^*$, which implies

$$(1 - \epsilon) R^* \leq z_{\mathsf{LP}}(1) - \epsilon R^* < \sum_{i=0}^{n} a_i \hat{p}_i(1).$$

□

E.3 Proof of Theorem 5.3

Proof. Davis et al. [18] show that if instance $\mathcal{I}$ has a partition (U_1, U_2) such that

$$\sum_{j \in U_1} c_j = \sum_{j \in U_2} c_j = T,$$

then, the assortment $(S_1 = \{1\}, S_2 = U_1)$ has expected revenue of $(T+2) \cdot (2T+1)$. Furthermore, they show if there is an assortment (S_1, S_2) such that $\Pi(S_1, S_2) \geq (T+2)(2T+1)$, then instance $\mathcal{I}$ has a partition $(S_2, [n] \setminus S_2)$. We complete the proof by establishing that for any

$$\epsilon < \frac{2T+1}{(6T+3)\,(3T+2)^2},$$

if there exists an assortment (S_1, S_2) such that $\Pi(S_1, S_2) \geq (T+2)(2T+1) - \epsilon$, then we have that $\sum_{i \in S_2} c_i = T$. Davis et al. [18] show that $S_1 = \{1\}$ for maximizing $\Pi(S_1, S_2)$. Suppose there exists (S_1, S_2) such that $\Pi(S_1, S_2) \geq (T+2)(2T+1) - \epsilon$. Then

$$\frac{2(T+3)(2T+1)\sqrt{T+1} + (T+1)(2T+1)\frac{\sum_{j \in S_2} c_j}{\sqrt{1+\sum_{j \in S_2} c_j}}}{\sqrt{1+\sum_{j \in S_2} c_j} + \sqrt{2+2(2T+1)}} \geq (T+2)(2T+1) - \epsilon$$

dividing by $(2T+1)$ on both sides, we have

$$\frac{2(T+3)\sqrt{T+1} + (T+1)\frac{\sum_{j \in S_2} c_j}{\sqrt{1+\sum_{j \in S_2} c_j}}}{\sqrt{1+\sum_{j \in S_2} c_j} + \sqrt{2+2(2T+1)}} \geq (T+2) - \frac{\epsilon}{2T+1}$$

which implies,

$$\begin{aligned}
&2(T+3)\sqrt{T+1}\sqrt{1+\sum_{j \in S_2} c_j} + (T+1)\left(\sum_{j \in S_2} c_j\right) \\
&\geq 2(T+2)\sqrt{T+1}\sqrt{1+\sum_{j \in S_2} c_j} + (T+2)\left(1+\sum_{j \in S_2} c_j\right) \\
&- \frac{\epsilon}{2T+1} \cdot \left(1+\sum_{j \in S_2} c_j + 2\sqrt{T+1}\sqrt{1+\sum_{j \in S_2} c_j}\right),
\end{aligned}$$

Hence we have,

$$\frac{\epsilon}{2T+1} \cdot \left(1 + \sum_{j \in S_2} c_j + 2\sqrt{T+1}\sqrt{1 + \sum_{j \in S_2} c_j}\right)$$
$$\geq -2\sqrt{T+1}\sqrt{1 + \sum_{j \in S_2} c_j} + (T+1) + \left(1 + \sum_{j \in S_2} c_j\right),$$

which implies

$$\epsilon \geq \frac{(2T+1) \cdot \left(\sqrt{T+1} - \sqrt{1 + \sum_{j \in S_2} c_j}\right)^2}{1 + \sum_{j \in S_2} c_j + 2\sqrt{T+1}\sqrt{1 + \sum_{j \in S_2} c_j}}$$
$$> \frac{(2T+1) \cdot \left(\sqrt{T+1} - \sqrt{1 + \sum_{j \in S_2} c_j}\right)^2}{6T+3}$$

The inequality follows from the fact that $1 + \sum_{j \in S_2} c_j \leq 2T+1$ and $\sqrt{T+1} < \sqrt{2T+1}$. Now multiplying with $\left(\sqrt{T+1} + \sqrt{1 + \sum_{j \in S_2} c_j}\right)^2$ in the numerator and denominator of right hand side of the above expression, we have the following inequality

$$\epsilon > \frac{(2T+1) \cdot \left(T - \sum_{j \in S_2} c_j\right)^2}{(6T+3)\left(\sqrt{T+1} + \sqrt{1 + \sum_{j \in S_2} c_j}\right)^2}$$
$$> \frac{(2T+1) \cdot \left(T - \sum_{j \in S_2} c_j\right)^2}{(6T+3)\left(T+1+2T+1\right)^2}$$

Note that if $\sum_{j \in S_2} c_j \neq T$, then since c_j are integers, we have that

$$\epsilon > \frac{2T+1}{(6T+3)\left(3T+2\right)^2},$$

contradicting the hypothesis that $\epsilon < \frac{2T+1}{(6T+3)(3T+2)^2}$. Hence, we have for any $\epsilon < \frac{2T+1}{(6T+3)(3T+2)^2}$, $\Pi(S_1, S_2) \geq (T+2)(2T+1) - \epsilon$ implies that $\sum_{i \in S_2} c_i = T$. □

E.4 Computing the ϵ-convex Pareto set

We now describe a polynomial (in n and $\frac{1}{\epsilon}$) time algorithm to compute the $\mathcal{CP}_\epsilon$. Before proceeding further, let us try and understand the Pareto set $\mathcal{P}(\pi)$.

Definition E.1. *A Pareto set, denoted by $\mathcal{P}_i(\pi)$, is a set of solutions $\mathbf{x} \in \mathcal{P}_i$ such that for all $\mathbf{x} \in \mathcal{P}_i$, there is no $\mathbf{x}' \in \mathcal{P}_i$ such that*

$$g_{i1}(\mathbf{x}') \geq g_{i1}(\mathbf{x}) \ \& \ g_{i2}(\mathbf{x}') \leq g_{i2}(\mathbf{x})$$

The Pareto set corresponds to the optimal solutions of the weighted linear program $\max w_1 g_{i1}(\mathbf{x}) - w_2 g_{i2}(\mathbf{x})$ over the polytope $\mathcal{P}_i$, for all weight vectors $w_1, w_2 \in \mathcal{R}_+$. Since, we are maximizing a linear function over the polytope $\mathcal{P}_i$, the optimal solutions will be extreme points of $\mathcal{P}_i$, which in our case will be feasible to our combinatorial optimization problem (5.19). It is easy to see that the Pareto set $\mathcal{P}_i(\pi)$ contains the optimal solution to the sub-problem (5.19). However, computing a Pareto set may be computationally infeasible. Therefore, we compute the ϵ-convex Pareto set in hope of finding an approximate optimal solution.

The idea behind the algorithm for finding an ϵ-convex Pareto set $\mathcal{CP}_\epsilon$ is to choose a polynomial number of such weight vectors ($\{w_1, w_2\}$) and obtain the corresponding extreme point solutions for the weighted linear programs. We present the algorithm for evaluating the ϵ-convex Pareto set below and later establish the correctness of the algorithm. In steps 3-4 of the algorithm, we fix the weight set (possible choices for w_1, w_2) by enforcing the $\max\{w_1, w_2\} = U$, for some pre-decided U. Let M be such that, $\frac{1}{M} \leq g_{i1}(\mathbf{x}) \leq M$ and $\frac{1}{M} \leq g_{i2}(\mathbf{x}) \leq M$. In steps 5-7, we choose another weight set $R(M)$, which scales the linear functions g_{i1}, g_{i2} appropriately so that we can compute the ϵ-convex Pareto set $\mathcal{CP}_\epsilon$. (see the proof of correctness for better understanding)

The following theorem due to [21] establishes the correctness of the above algorithm. We present the proof specifically to our context for the sake of completeness.

Theorem E.1 (**Diakonikolas and Yannakakis (2008)**). *The above algorithm*

Algorithm 12 Computing ϵ-convex Pareto sets for the sub-problem (5.19)

Choose $\epsilon_1 > 0$ such that $1 - \epsilon_1 = \dfrac{1}{1+\epsilon}$
$U \leftarrow \left\lceil \dfrac{2}{\epsilon_1} \right\rceil$, let $[U] = \{1, 2, \cdots, U\}$
$W_1(U) \leftarrow \{U\} \times [U]$, $W_2(U) \leftarrow [U] \times \{U\}$
$W(U) = W_1(U) \cup W_2(U)$
$S(M) \leftarrow \{2^0, 2^1, \cdots, 2^{2\lceil \log_2 M \rceil}\}$
$R_1(M) \leftarrow \{1\} \times S(M)$, $R_2(M) \leftarrow S(M) \times \{1\}$
$R(M) = R_1(M) \cup R_2(M)$
$\mathcal{CP}_\epsilon \leftarrow \phi$
for each $r \in R(M)$ **do** do
    for each $w \in W(U)$ **do** do
        $\mathbf{x}^* \leftarrow$ optimal extreme point for $\{\max r_1 w_1 g_{i1}(\mathbf{x}) - r_2 w_2 g_{i2}(\mathbf{x}) : \ \mathbf{A}_i \mathbf{x} \leq b_i, \ 0 \leq \mathbf{x} \leq 1\}$
    end for
end for
$\mathcal{CP}_\epsilon \leftarrow \mathcal{CP}_\epsilon \cup \{\mathbf{x}^*\}$
Return $\mathcal{CP}_\epsilon$

yields an ϵ-convex Pareto set $\mathcal{CP}_\epsilon$, i.e. $\forall$ $\mathbf{x}$ $\in \mathcal{P}_i$, $\exists$ $\mathbf{x}' \in Conv(\mathcal{CP}_\epsilon)$ such that

$$g_{i1}(\mathbf{x}') \geq \frac{g_{i1}(\mathbf{x})}{1+\epsilon}$$

$$g_{i2}(\mathbf{x}') \leq (1+\epsilon) g_{i2}(\mathbf{x})$$

Proof. Proof Let $Q(U)$ denote the set of extreme points generated by the above algorithm and also let the set of optimal extreme points generated by the above algorithm be $\mathbf{x}_1, \cdots, \mathbf{x}_l$, where l is the total number of unique solutions obtained during the above algorithm.

Lemma E.1 (**Diakonikolas and Yannakakis(2008)**). *Suppose that $\mathbf{x}$ is in the Pareto-set $\mathcal{P}_i(\pi)$ such that $g_{i1}(\mathbf{x}) \leq 2 g_{i2}(\mathbf{x})$ and $g_{i2}(\mathbf{x}) \leq 2 g_{i1}(\mathbf{x})$ and $\mathbf{x} \notin Q(U)$, then there exists a $\mathbf{x}' \in Conv(Q(U))$, such that*

$$g_{i1}(\mathbf{x}') \geq (1-\epsilon_1) g_{i1}(\mathbf{x})$$

$$g_{i2}(\mathbf{x}') \leq (1+\epsilon_1) g_{i2}(\mathbf{x})$$

Proof. PROOF Suppose there is no $\mathbf{x}' \in Conv(Q(U))$ such that

$$g_{i1}(\mathbf{x}') \geq (1-\epsilon_1)g_{i1}(\mathbf{x}) \;\&\; g_{i2}(\mathbf{x}') \leq (1+\epsilon_1)g_{i2}(\mathbf{x})$$

then the following linear program is infeasible:

$$\sum_{j=1}^{l} \lambda_j g_{i1}(\mathbf{x}_j) \geq (1-\epsilon_1)g_{i1}(\mathbf{x})$$
$$\sum_{j=1}^{l} \lambda_j g_{i2}(\mathbf{x}_j) \leq (1+\epsilon_1)g_{i2}(\mathbf{x})$$
$$\sum_{j=1}^{l} \lambda_j = 1$$
$$\lambda_1, \lambda_2, \cdots, \lambda_l \geq 0$$

By Farka's lemma, there exists w_1, w_2 and v which satisfy the following inequalities,

$$w_1 g_{i1}(\mathbf{x}_j) - w_2 g_{i2}(\mathbf{x}_j) + v \leq 0 \quad \forall \quad j = 1, \cdots, l$$
$$w_1(1-\epsilon_1)g_{i1}(\mathbf{x}) - w_2(1+\epsilon_1)g_{i2}(\mathbf{x}) + v > 0$$
$$w_1, w_2 \geq 0$$

which implies that $w_1, w_2 \geq 0$ and

$$w_1 g_{i1}(\mathbf{x}_j) - w_2 g_{i2}(\mathbf{x}_j) < w_1(1-\epsilon_1)g_{i1}(\mathbf{x}) - w_2(1+\epsilon_1)g_{i2}(\mathbf{x}) \quad \forall \quad j = 1, \cdots, l$$

To establish contradiction to the assumption that there is no such $\mathbf{x}'$, it suffices to show that there exists a j such that $w_1 g_{i1}(\mathbf{x}_j) - w_2 g_{i2}(\mathbf{x}_j) \geq w_1(1-\epsilon_1)g_{i1}(\mathbf{x}) - w_2(1+\epsilon_1)g_{i2}(\mathbf{x})$.

Consider arbitrary $w_1, w_2 \geq 0$, without loss of generality, it can be assumed that the maximum value of $\{w_1, w_2\}$ is U. Let $w_1^* = \lfloor w_1 \rfloor$ and $w_2^* = \lceil w_2 \rceil$, we clearly have $\{w_1^*, w_2^*\} \in W(U)$ and let $\mathbf{x}^*$ be the optimal extreme point for the objective

$\max w_1^* g_{i1}(\mathbf{x}) - w_2^* g_{i2}(\mathbf{x})$, then $\mathbf{x}^* \in Q(U)$ and hence without loss of generality assume that for some $j \leq l, \mathbf{x}^* = \mathbf{x}_j$. We will now show that

$$w_1 g_{i1}(\mathbf{x}_j) - w_2 g_{i2}(\mathbf{x}_j) \geq w_1(1-\epsilon_1) g_{i1}(\mathbf{x}) - w_2(1+\epsilon_1) g_{i2}(\mathbf{x})$$

Note that $w_1 - w_1^* \leq 1$ and $w_2^* - w_2 \leq 1$ and remember that we have enforced by scaling $\max\{w_1, w_2\} = U$, let $t \in \{1,2\}$ be such that $w_t = U$, we have $w_t^* = w_t$ and

$$\begin{aligned}(w_1 - w_1^*) g_{i1}(\mathbf{x}) + (w_2^* - w_2) g_{i2}(\mathbf{x}) &\leq \sum_{k=\{1,2\}/t} g_{ik}(\mathbf{x}_j) \leq 2 g_{it}(\mathbf{x}_j) \leq \epsilon_1 U g_{it}(\mathbf{x}_j) \\ &\leq \epsilon_1 (w_1 g_{i1}(\mathbf{x}) + w_2 g_{i2}(\mathbf{x}))\end{aligned}$$

where the second inequality follows from the fact that the assumption that $g_{i1}(\mathbf{x}) \leq 2 g_{i2}(\mathbf{x})$ and $g_{i2}(\mathbf{x}) \leq 2 g_{i1}(\mathbf{x})$, the third inequality follows from our choice of $U = \lceil 2/\epsilon_1 \rceil$ and the last inequality follows from the fact that $U g_{it}(\mathbf{x}_j) \leq w_1 g_{i1}(\mathbf{x}) + w_2 g_{i2}(\mathbf{x})$. Therefore, from this chain of inequalities, we get

$$w_1(1-\epsilon_1) g_{i1}(\mathbf{x}) - w_2(1+\epsilon_1) g_{i2}(\mathbf{x}) \leq w_1^* g_{i1}(\mathbf{x}) - w_2^* g_{i2}(\mathbf{x})$$

Since $\mathbf{x}_j$ is the optimal solution for the objective $\max w_1^* g_{i1}(\mathbf{x}) - w_2^* g_{i2}(\mathbf{x})$, we have

$$w_1^* g_{i1}(\mathbf{x}) - w_2^* g_{i2}(\mathbf{x}) \leq w_1^* g_{i1}(\mathbf{x}_j) - w_2^* g_{i2}(\mathbf{x}_j)$$

Noting that $w_1^* \leq w_1$ and $w_2^* \geq w_2$, we have

$$w_1^* g_{i1}(\mathbf{x}_j) - w_2^* g_{i2}(\mathbf{x}_j) \leq w_1 g_{i1}(\mathbf{x}_j) - w_2 g_{i2}(\mathbf{x}_j)$$

Combing the above three inequalities, we have the required contradiction.

If the above lemma was true for any $\mathbf{x} \in \mathcal{P}_i(\pi)$, instead of, for only $\mathbf{x} \in \mathcal{P}_i(\pi)$ such that $g_{i1}(\mathbf{x}) \leq 2 g_{i2}(\mathbf{x})$ and $g_{i2}(\mathbf{x}) \leq 2 g_{i1}(\mathbf{x})$, then the Theorem would have followed from the Lemma, since $\mathcal{P}_i(\pi)$ contains at least one optimal solution. Consider any $\mathbf{x} \in \mathcal{P}$, the ratios $g_{i1}(\mathbf{x})/g_{i2}(\mathbf{x})$ and $g_{i2}(\mathbf{x})/g_{i1}(\mathbf{x})$ are both bounded by M^2, hence there exists $\{r_1, r_2\} \in R(M)$, such that $r_1 g_{i1}(\mathbf{x})$ and $r_2 g_{i2}(\mathbf{x})$ are within a factor of

2 of each other. Now $\mathbf{x}$ belongs to the Pareto-set of the weighted objectives $r_1 g_{i1}(\mathbf{x})$ and $r_2 g_{i2}(\mathbf{x})$ and by the above lemma, there exists a $S' \in Conv(Q(U))$ such that

$$\begin{aligned} & r_1 g_{i1}(\mathbf{x}') \geq (1-\epsilon_1) r_1 g_{i1}(\mathbf{x}) \ \& \ r_1 g_{i2}(\mathbf{x}') \leq (1+\epsilon_1) r_2 g_{i2}(\mathbf{x}) \\ \implies & g_{i1}(\mathbf{x}') \geq (1-\epsilon_1) g_{i1}(\mathbf{x}) \ \& \ g_{i2}(\mathbf{x}') \leq (1+\epsilon_1) g_{i2}(\mathbf{x}) \end{aligned}$$

By definition of ϵ_1, we have $1-\epsilon_1 = \dfrac{1}{1+\epsilon}$, which implies that $\epsilon > \epsilon_1$ and hence we have

$$g_{i1}(\mathbf{x}') \geq \frac{g_{i1}(\mathbf{x})}{1+\epsilon} \ \& \ g_{i2}(\mathbf{x}') \leq (1+\epsilon) g_{i2}(\mathbf{x})$$

□

E.5 Proof of Lemma 5.5

Proof. For the sake of contradiction, assume that

$$\mathbf{S}_i^* \notin \ \underset{\mathbf{S}_i \in \mathcal{P}_i}{\arg\max} \sum_{j \in [n]} \sum_{k \in [\ell]} v_{ijk} (r_{ij} - u') s_{ijk}.$$

and let

$$\hat{\mathbf{S}}_i \in \ \underset{\mathbf{S}_i \in \mathcal{P}_i}{\arg\max} \sum_{j \in [n]} \sum_{k \in [\ell]} v_{ijk} (r_{ij} - u') s_{ijk}.$$

We claim that the number of products offered in assortment $\mathbf{S}_i^*$ and $\hat{\mathbf{S}}_i$ is same. This follows by observing that the linear functions $f_{ij}(u)$ and $f_{ij}(u')$ have the same sign in the interval $[u_p, u_{p+1}]$ for all $j \in [n]$. Without loss of generality assume that the number of products offered in $\hat{\mathbf{S}}_i$ is less than the number of products offered in $\mathbf{S^*}_i$. Therefore, there must exist a product j that is offered in $\mathbf{S^*}_i$ and not offered in $\hat{\mathbf{S}}_i$. Hence, we have $r_{ij} < u'$ and $r_{ij} < u$ implying that not offering product j in $\mathbf{S^*}_i$ would increase the value of

$$\sum_{j \in [n]} \sum_{k \in [\ell]} v_{ijk} (r_{ij} - u) s^*_{ijk},$$

contradicting the optimality of $\mathbf{S}_i^*$.

Since $\mathbf{S}_i^*$ and $\hat{\mathbf{S}}_i$ offer the same number of products, we have two cases

1. There exists two products j, j' such that j is offered in $\mathbf{S}_i^*$ and not offered in $\hat{\mathbf{S}}_i$, while j' is offered in $\hat{\mathbf{S}}_i$ and not offered in $\mathbf{S}_i^*$.

2. $\hat{\mathbf{S}}_i$ and $\mathbf{S}_i^*$ offer the same set of products, but at different display positions.

Consider the first scenario, where there are two products j, j' such that j is offered in $\mathbf{S}_i^*$ and not offered in $\hat{\mathbf{S}}_i$, while j' is offered in $\hat{\mathbf{S}}_i$ and not offered in $\mathbf{S}_i^*$. Let k, k' be the display slots of products j, j' in the assortments $\mathbf{S}_i^*$ and $\hat{\mathbf{S}}_i$ respectively. Therefore, we must have

$$v_{ijk}(r_{ij} - u) \geq v_{ij'k}(r_{ij'} - u)$$

$$v_{ij'k'}(r_{ij'} - u') \geq v_{ijk'}(r_{ij} - u')$$

The first inequality follows from the hypothesis that product j is included in $\mathbf{S}_i^*$, while product j' is not. Similarly the second inequality folows from the hypothesis that product j' is included in $\hat{\mathbf{S}}_i$, while product j is not. From (**??**) we have that $v_{ijk} = v_{ij}\lambda_{ik}$. Hence, we have

$$v_{ij}(r_{ij} - u) \geq v_{ij'}(r_{ij'} - u)$$

$$v_{ij'}(r_{ij'} - u') \geq v_{ij}(r_{ij} - u'),$$

contradicting the fact that $f_{ij}(u) - f_{ij'}(u)$ and $f_{ij}(u') - f_{ij'}(u')$ have the same sign in the interval $[u_p, u_{p+1}]$.

Consider the second scenario, where the same products are offered in different display slots in $\mathbf{S}_i^*$ and $\hat{\mathbf{S}}_i$. For such a scenario to occur, there must be a set of products whose display positions in the assortment $\mathbf{S}_i^*$ are shifted in a cyclical fashion from the display positions in the assortment $\hat{\mathbf{S}}_i$. Without loss of generality, let those products be indexed $1, \cdots, q$. Let the display positions of these products in the assortment $\mathbf{S}_i^*$ be $k_1, \cdots, k_q$, then the display positions of these products in the assortment $\hat{\mathbf{S}}_i$ will be $k_2, \cdots, k_q, k_1$ respectively. We have one of the three possibilities,

- There exists $j \in \{1, \cdots, q-2\}$ such that $\lambda_{ik_j} < \lambda_{ik_{j+1}}$ and $\lambda_{ik_{j+1}} > \lambda_{ik_{j+2}}$
- $\lambda_{ik_j} < \lambda_{ik_{j+1}}$ for all $j \in \{1, \cdots, q-1\}$
- $\lambda_{ik_j} > \lambda_{ik_{j+1}}$ for all $j \in \{1, \cdots, q-1\}$,

For the first case, consider the following inequalities

$$v_{ijk_j}(r_{ij} - u) + v_{i(j+1)k_{j+1}}(r_{i(j+1)} - u) \geq v_{ijk_{j+1}}(r_{ij} - u) + v_{i(j+1)k_j}(r_{i(j+1)} - u)$$

$$v_{ijk_{j+1}}(r_{ij} - u') + v_{i(j+1)k_{j+2}}(r_{i(j+1)} - u') \geq v_{ijk_{j+2}}(r_{ij} - u') + v_{i(j+1)k_{j+1}}(r_{i(j+1)} - u')$$

The first inequality follows from the hypothesis that in assortment $\mathbf{S}_i^*$ product j is displayed in slot k_j and product $j+1$ is displayed in slot k_{j+1} and not vice versa. Similarly, the second inequality follows from the hypothesis that in assortment $\hat{\mathbf{S}}_i$ product j is displayed in slot k_{j+1} and product $j+1$ is displayed in slot k_{j+2} and not vice versa. We have that $v_{ijk} = v_{ij}\lambda_{ik}$. Hence, we have

$$(\lambda_{ik_j} - \lambda_{ik_{j+1}})v_{ij}(r_{ij} - u) \geq (\lambda_{ik_j} - \lambda_{ik_{j+1}})v_{i(j+1)}(r_{i(j+1)} - u)$$

$$(\lambda_{ik_{j+1}} - \lambda_{ik_{j+2}})v_{ij}(r_{ij} - u') \geq (\lambda_{ik_{j+1}} - \lambda_{ik_{j+2}})v_{i(j+1)}(r_{i(j+1)} - u'),$$

which implies

$$v_{ij}(r_{ij} - u) \geq v_{i(j+1)}(r_{i(j+1)} - u)$$

$$v_{i(j+1)}(r_{i(j+1)} - u') \geq v_{ij}(r_{ij} - u'),$$

contradicting the fact that $f_{ij}(u) - f_{ij'}(u)$ and $f_{ij}(u') - f_{ij'}(u')$ have the same sign in the interval $[u_p, u_{p+1}]$. We can prove a similar contradiction by consider the products indexed 1, q and considering inequalities corresponding to swapping the display positions of these products. Hence, we have

$$\mathbf{S}_i^* \in \underset{\mathbf{S}_i \in \mathcal{P}_i}{\arg\max} \sum_{j \in [n]} \sum_{k \in [\ell]} v_{ijk}(r_{ij} - u')s_{ijk} \; \forall \; u' \in [u_p, u_{p+1}].$$

□