Optimal Dynamic Assortment Planning

Denis Saure*
Columbia University

Assaf Zeevi†
Columbia University

October 2009

Abstract

We study a family of stylized assortment planning problems, where arriving customers make purchase decisions among offered products based on maximizing their utility. Given limited display capacity and no a priori information on consumers' utility, the retailer must select which subset of products to offer. By offering different assortments and observing the resulting purchase behavior, the retailer learns about consumer preferences, but this experimentation should be balanced with the goal of maximizing revenues. We develop a family of dynamic policies that judiciously balance the aforementioned tradeoff between exploration and exploitation, and prove that their performance cannot be improved upon in a precise mathematical sense. One salient feature of these policies is that they "quickly" recognize, and hence limit experimentation on, strictly suboptimal products.

Short Title: Optimal Dynamic Assortment Planning

Keywords: assortment planning, bandit problem, on-line algorithm, demand learning .

1 Introduction

Motivation and main objectives. Product assortment selection is among the most critical decisions facing retailers. Inferring customer preferences and responding accordingly with updated product offerings plays a central role in a growing number of industries, in particular when companies are capable of revisiting product assortment decisions during the selling season as demand information becomes available. From an operations perspective, a retailer is often not capable of simultaneously displaying every possible product to prospective customers due to limited shelf space, stocking restrictions and other capacity related considerations. One of the central decisions is therefore which products to include in the retailer's product assortment. This will be referred to

*Graduate School of Business, e-mail: dsaure05@gsb.columbia.edu
†Graduate School of Business, e-mail: assaf@gsb.columbia.edu

as *assortment planning problem*; see (Kok, Fisher, and Vaidyanathan 2006) for an overview. Our interest lies in *dynamic* instances of the problem where assortment planning decisions can be revisited frequently throughout the selling season (these could correspond to periodic review schedules, for example). This will be referred to as *dynamic assortment planning*. Here are two motivating examples that arise in very different application domains.

Example 1: Fast fashion. In recent years "fast" fashion companies such as Zara, Mango or World co have implemented highly flexible and responsive supply chains that allow them to make and revisit most product design and assortment decisions *during* the selling season. Customers visiting one of their stores will only see a fraction of the potential products that the retailer has to offer, and their purchase decisions will effectively depend on the specific assortment presented at the store. The essence of fashion retail entails offering new products for which no demand information is available, and hence the ability to revisit these decisions at a high frequency is key to the "fast fashion" business model; each season there is a need to learn the current fashion trend by exploring with different styles and colors, and to exploit such knowledge before the season is over.

Example 2: On-line advertising. This emerging area of business is the single most important source of revenues for thousands of web sites. Giants such as Yahoo and Google, depend almost completely on on-line advertisement to subsist. One of the most prevalent business models here builds on the cost-per-click statistic: advertisers pay the web site (a "publisher") only when a user clicks on their ads. Upon each visit, users are presented with a finite set of ads, on which they may or may not click depending on what is being presented. Roughly speaking, the publisher's objective is to learn ad click-through-rates (and their dependence on the set of ads being displayed) and present the set of ads that maximizes revenues within the life span of the contract with the advertiser.

The above motivating applications share common features. For products/ads for which little or no demand information is available a priori, retailers/publishers must learn their desirability/effectiveness by dynamically adjusting their product/ad offering and observing customer behavior. It is natural to think that any *good* assortment strategy should gather some information on consumer preferences before committing to assortments that are thought to be profitable. This is the classical "exploration versus exploitation" trade-off: on the one hand, the longer a retailer/publisher spends learning consumer preferences, the less time remains to exploit that knowledge and optimize profits. On the other hand, less time spent on studying consumer behavior translates into more residual uncertainty, which could hamper revenue maximization objective. Moreover, demand information must be gathered carefully as product/ad profitability depends on the assortments offered: the retailer/publisher may learn consumer preferences more effectively by experimenting with a particular set of assortments.

The purpose of this paper is to study a family of stylized dynamic assortment problems that

consider homogeneous utility maximizing customers: each customer assigns a (random) utility to each offered product, and purchases the product that maximizes his/her utility. The retailer needs to devise an assortment policy to maximize revenues over the relevant time horizon by properly adapting the offered assortment based on observed customer purchase decisions and subject to capacity constraints that limit the size of the assortment.

Our main focus in this work is on the impact of learning consumer behavior via suitable assortment experimentation, and doing this in a manner that guarantees minimal revenue loss over the selling horizon. To shed light on this facet of the problem, we ignore other effects such as inventory considerations, additional costs (such as assortment switching costs), operational constraints (e.g. restrictions on the sequence of offered assortments), and finally, we assume that product prices are fixed throughout the selling season. Returning to the motivating examples we discussed earlier, it is worth noting that such considerations are absent almost altogether from the on line advertisement problem, and are often ignored in the fast fashion setting; see, for example, the work of Caro and Gallien (2007).

Key insights and qualitative results. As indicated above, we consider assortment policies that can only use observed purchase decisions to adjust assortment choices at each point in time (this will be defined more formally later as a class of non-anticipating policies). Performance of such a policy will be measured in terms of the expected revenue loss relative to an oracle that knows in advance the product utility distributions. This is the loss due to the absence of *a priori* knowledge of consumer behavior. Our objective is to characterize the minimum loss attainable by any non-anticipating assortment policy.

The main findings of this paper are summarized below.

i.) We establish fundamental bounds on the performance of any policy. Specifically, we identify the magnitude of the loss, relative to the oracle performance, that *any* policy must incur in terms of its dependence on: the length of the selling horizon; the number of products; and the capacity constraint (see Theorem 1 for a precise statement).

ii.) We propose a family of adaptive policies that achieve the fundamental bound mentioned above. These policies "quickly" identify the optimal assortment of products (the one that maximizes the expected single sale profit) with "high" probability while successfully limiting the extent of exploration. Our performance analysis, in section 5.2, makes these terms rigorous; see Theorem 3.

iii.) We prove that not all products available to the retailer need to be extensively tested: under mild assumptions, some of them can be easily and quickly identified as suboptimal. In particular, a specific subset of said products can be detected after a "small" number of experiments

(independent of the length of the selling horizon); see Theorems 1 and 3. Moreover, we show that our proposed policy successfully limits the extent of such an exploration (see Corollary 1 for a precise statement).

iv.) We highlight salient features of the dynamic assortment problem that distinguish it from similar problems of sequential decision making under model uncertainty, and we show how exploiting these features helps to dramatically decrease the complexity of the assortment problem, relative to using existing non-customized strategies, e.g., from the multi-armed bandit literature.

On a more practical side, our results establish that an oracle with advance knowledge of customer behavior only gains additional revenue on the order of the *logarithm* of the total number of customers visiting the retailer during the selling season. Moreover, we establish that this is a fundamental price that any feasible assortment policy must pay. Regarding the "exploration versus exploitation" trade-off, we establish the precise frequency and extent of assortment experimentation that guarantee this best achievable performance. While in general it is necessary to experiment with "inferior" products at a precise and critical frequency that is increasing with the time horizon, for a certain subset of these products experimentation can be kept to a minimum (a bounded number of trials independent of the time horizon). This result differs markedly from most of the literature on similar sequential decision making problems.

The remainder of the paper. The next section reviews related work. Section 3 formulates the dynamic assortment problem. Section 4 provides a fundamental limit on the performance of any assortment policy, and analyzes its implications for policy design. Section 5 proposes a dynamic assortment algorithm that achieves this performance bound, and Subsection 5.3 customizes our proposed algorithm for the most widely used customer choice model, namely the Logit. Finally, Section 6 presents our concluding remarks. Proofs are relegated to two appendices.

2 Literature Review

Static assortment planning. The static planning literature focuses on finding an optimal assortment that is held unchanged throughout the entire selling season. Customer behavior is assumed to be known a priori, but inventory decisions are considered; see Kok, Fisher, and Vaidyanathan (2006) for a review of the state-of-the-art in static assortment optimization. Within this area, van Ryzin and Mahajan (1999) formulate the assortment planning problem using a Multinomial Logit model (hereafter, MNL) of consumer choice. Assuming that customers do not look for a substitute if their choice is stocked out (known as static substitution), they prove that the optimal assortment is always in the "popular assortment set" and establish structural properties of the optimal

assortment and ordering quantities. Gaur and Honhon (2006) use the locational choice model and characterize properties of the optimal solution under static substitution. In a recent paper Goyal, Levi, and Segev (2009) prove that the assortment problem is NP-hard, in the static setting when stock-out based substitution is allowed, and propose a near-optimal heuristic solution for a particular choice model; see also Mahajan and van Ryzin (2001), Honhon, Gaur, and Seshadri (2009) and Hopp and Xu (2008).

While we will assume perfect replenishment, and hence eliminate stock-out based substitution considerations, it is important to note that even in this setting the *static* one-period profit maximization problem remains NP-hard in general; see Goyal, Levi, and Segev (2009). The work of Rusmevichientong, Shen, and Shmoys (2008) identifies a polynomial-time algorithm for the static optimization problem when consumer preferences are represented using particular choice models; hence at least in certain instances the problem can be solve efficiently.

Dynamic assortment planning. This problem setting allows to revisit assortment decisions at each point in time as more information is collected about initially unknown demand/consumer preferences. To the best of our knowledge Caro and Gallien (2007) were the first to study this type of problem, motivated by an application in fast fashion. In their formulation, customer demand for a product is *exogenous*, and independent of demand and availability for other products. The rate of demand is constant throughout the selling season, and their formulation ignores inventory considerations. Taking a Bayesian approach to demand learning, they study the problem using dynamic programming. They derive bounds on the value function and propose an index-based policy that is shown to be *near* optimal when there is certain prior information on demand. Closer to our paper is the work by Rusmevichientong, Shen, and Shmoys (2008). There, utility maximizing customers make purchase decisions according to the MNL choice model (a special case of the more general setting treated in the present paper), and an adaptive algorithm for joint parameter estimation and assortment optimization is developed, see further discussion below.

Connection to the multi-armed bandit literature. In the canonical multi-armed bandit problem the decision maker can select in each period to pull a single arm out of a set of K possible arms, where each arm delivers a random reward whose distribution is not known a priori, and the objective is to maximize the revenue over a finite horizon. See Lai and Robbins (1985) and Auer, Cesa-Bianchi, and Fisher (2002) for a classical formulation and solution approach to the problem, respectively.

The model of Caro and Gallien (2007) is in fact equivalent to a multi-armed bandit problem with multiple simultaneous plays. The dynamic programming formulation and the Bayesian learning approach aims to solve the "exploration versus exploitation" trade-off optimally. See also Farias and Madan (2009) for a similar bandit-formulation with multiple simultaneous plays under

more restricted type of policies. In the work of Rusmevichientong, Shen, and Shmoys (2008) the connection is less straightforward. Their proposed algorithm works in cycles. Each cycle mixes parameter estimation (exploration) and assortment optimization (exploitation). In the exploration phase $O(N^2)$ assortments are tested, where N is the number of products. Estimators based on the exploration phase are fed into the static optimization problem, which returns $O(N)$ assortments among which the optimal one is found with high probability. From there, a standard multi-armed bandit algorithm is prescribed to find the optimal assortment, and an upper bound on the regret of order $O(N^2 \log^2 T)$ is established, where T is the length of the planning horizon.

There is a thematic connection between multi-armed bandits and assortment planning problems, in the sense that both look to balance exploration and exploitation. However, the fact that product utility does not map directly to retailer revenues in the dynamic assortment problem is essentially what distinguishes these problems. In the bandit setting all products are ex-ante identical, and only individual product exploration allows the decision maker to differentiate them. Nevertheless, there is always the possibility that a poorly explored arm is in fact optimal. This last fact prevents limiting exploration on arms that have been observed to be empirically inferior. (In their seminal work, Lai and Robbins (1985) showed that "good" policies should explore each arm at least $O(\log T)$ times.) In the assortment planning setting, products are not ex-ante identical, and product revenue is capped by its profit margin. In section 4 we show how this observation can be exploited to limit exploration on certain "suboptimal" products (a precise definition will be advanced in what follows). Moreover, the possibility to test several products simultaneously has the potential to further reduce the complexity of the assortment planning problem. Our work builds on some of the ideas present in the multi-armed bandit literature, most notably the lower bound technique developed by Lai and Robbins (1985), but also exploits salient features of the assortment problem in constructing optimal algorithms and highlighting key differences from traditional bandit results.

3 Problem Formulation

Model primitives and basic assumptions. We consider a price-taking retailer that has N different products to sell. For each product $i \in \mathcal{N} := \{1, \ldots, N\}$, let $w_i > 0$ denote the *marginal profit* resulting from selling one unit of the product, and let $w := (w_1, \ldots, w_N)$ denote the vector of product margins. Due to display space constraints, the retailer can offer at most C products simultaneously.

Let T to denote the total number of customers that arrive during the selling season after which sales are discontinued. (The value of T is in general not known to the retailer a priori.) We use t to index customers according to their arrival times, so $t = 1$ corresponds to the first arrival, and

$t = T$ the last. We assume the retailer has both a perfect replenishment policy, and the flexibility to offer a different assortment to every customer without incurring any switching cost. (While these assumptions do not typically hold in practice, they provide sufficient tractability for analysis purposes, and allow us to extract structural insights.)

With regard to demand, we will adopt a random utility approach to model customer preferences over products: customer t assigns a utility U_i^t to product i, for $i \in \mathcal{N} \cup \{0\}$, with

$$U_i^t := \mu_i + \zeta_i^t,$$

were $\mu_i \in \mathbb{R}_+$ denotes the mean utility assigned to product i, $\zeta_i^1, \ldots, \zeta_i^T$ are independent random variables drawn from a distribution F common to all customers, and product 0 represents a no-purchase alternative. Let $\mu := (\mu_1, \ldots, \mu_N)$ denote the vector of mean utilities. We assume all customers assign μ_0 to a no-purchase alternative; when offered an assortment, customers select the product with the highest utility if that utility is greater than the one provided by the no-purchase alternative. For convenience, and without loss of generality, we set $\mu_0 := 0$.

The static assortment optimization problem. Let $\mathcal{S}$ denote the set of possible assortments, i.e., $\mathcal{S} := \{S \subseteq \mathcal{N} : |S| \leq C\}$, where $|S|$ denotes the cardinality of the set $S \subset \mathcal{N}$. For a given assortment $S \in \mathcal{S}$ and a given vector of mean utilities μ, the probability $p_i(S, \mu)$ that a customer chooses product $i \in S$ is

$$p_i(S, \mu) = \int_{-\infty}^{\infty} \prod_{j \in S \cup \{0\} \setminus \{i\}} F(x - \mu_j) \quad dF(x - \mu_i), \tag{1}$$

and $p_i(S, \mu) = 0$ for $i \notin S$. The expected profit $f(S, \mu)$ associated with an assortment S and mean utility vector μ is given by

$$f(S, \mu) = \sum_{i \in S} w_i p_i(S, \mu).$$

If the retailer knows the value of the vector μ, then it is optimal to offer $S^*(\mu)$, the solution to the static optimization problem, to every customer:

$$S^*(\mu) \in \operatorname*{argmax}_{S \in \mathcal{S}} f(S, \mu). \tag{2}$$

In what follows we will assume that the solution to the static problem is unique (this assumption simplifies our construction of fundamental performance bounds, and can be relaxed as long as there is at least one product that is suboptimal). Note that for a given product $i \in \mathcal{N}$, $p_i(S, \mu)$ in (1) depends in a non-trivial manner on the assortment S. Efficiently solving problem (2) is beyond the scope of this paper: we will assume that the retailer can compute $S^*(\mu)$ for any vector μ.

Remark 1 (**Complexity of the static problem).** We note that for specific utility distributions there exist efficient algorithms for solving the static problem. For example, Rusmevichientong,

Shen, and Shmoys (2008) present a $O(N^2)$ algorithm to solve the static problem when an MNL choice model is assumed, i.e., when F is assumed to be a standard Gumbel distribution (location parameter 0 and scale parameter 1) for all $i \in \mathcal{N}$. This is an important contribution given that the MNL is by far the most commonly used choice model. The algorithm, based on a more general solution concept developed by Megiddo (1979), can in fact be used to solve the static problem efficiently for any Luce-type choice model[1] (see, for example, Anderson, De Palma, and Thisse (1992)).

The dynamic optimization problem. We assume that the retailer knows F, the distribution that generate the idiosyncracies of customer utilities, but *does not know* the mean vector μ. That is, the retailer has some understanding of how customers differ in their valuations for any given product, but has no prior information on how customers rank different products on average.

The retailer is able to observe purchase/no-purchase decisions made by each customer. S/he needs to decide what assortment to offer to each customer, taking into account all information gathered up to that point in time, in order to maximize expected cumulative profits. More formally, let $(S_t \in \mathcal{S} \,:\, 1 \leq t \leq T)$ denote an *assortment process*, which is comprised of a sequence of admissible assortments over $\{1, \ldots, T\}$. Let

$$X_i^t := \mathbf{1}\left\{i \in S_t \,,\, U_i^t > U_j^t \,,\, j \in S_t \setminus \{i\} \cup \{0\}\right\},$$

denote the purchase decision of customer t regarding product $i \in S_t$, where $X_i^t = 1$ indicates that customer t decided to purchase product i, and $X_i^t = 0$ otherwise. Also, let $X_0^t := \mathbf{1}\left\{U_0 > U_j \,,\, j \in S_t\right\}$ denote the overall purchase decision of customer t, where $X_0^t = 1$ if customer t opted not to purchase any product, and $X_0^t = 0$ otherwise. Here, and in what follows, $\mathbf{1}\{A\}$ denotes the indicator function of a set A. We denote by $X^t := (X_0^t, X_1^t, \ldots, X_N^t)$ the vector of purchase decisions of customer t. Let $\mathcal{F}_t = \sigma((S_u, X^u), 1 \leq u \leq t)$ $t = 1, \ldots, T$, denote the filtration or history associated with the assortment process and purchase decisions up to (including) time t, with $\mathcal{F}_0 = \emptyset$. An assortment process is said to be *non-anticipating* if S_t is determined only on the basis of previous assortment choices and observed purchase decisions, i.e., is $\mathcal{F}_{t-1}$-measurable, for all t. An admissible *assortment policy* π is a mapping from past history to $\mathcal{S}^T$ such that $(S_t(\mathcal{F}_{t-1}) \in \mathcal{S} \,:\, 1 \leq t \leq T)$ is non-anticipating. We will restrict attention to the set of such policies and denote it by $\mathcal{P}$. We will use $\mathbb{E}_\pi$ and $\mathbb{P}_\pi$ to denote expectations and probabilities of random variables, when the assortment policy $\pi \in \mathcal{P}$ is used.

The retailer's objective is to choose a policy $\pi \in \mathcal{P}$ to maximize the expected cumulative revenues over the selling season

$$J^\pi(T, \mu) := \mathbb{E}_\pi\left[\sum_{t=1}^{T} \sum_{i \in \mathcal{N}} w_i \, X_i^t\right].$$

[1] These are choice models for which $p_i(S) = v_i/(\sum_{j \in S} v_j)$ for a vector $v \in \mathbb{R}_+^N$, and any $S \subseteq \mathcal{N}$.

If the mean utility vector μ were known to the retailer at the start of the selling season, the optimal policy would clearly be to choose the assortment that maximizes the one-sale expected value, namely $S^*(\mu)$, and offer it to every customer. The corresponding performance, denoted by $J^*(T,\mu)$, is given by

$$J^*(T,\mu) := Tf(S^*(\mu),\mu). \tag{3}$$

This quantity provides an upper bound on expected revenues generated by any admissible policy, i.e., $J^*(T,\mu) \geq J^\pi(T,\mu)$ for all $\pi \in \mathcal{P}$. With this in mind we define the *regret* $\mathcal{R}^\pi(T,\mu)$ associated with a policy π, to be

$$\mathcal{R}^\pi(T,\mu) := T - \frac{J^\pi(T,\mu)}{f(S^*(\mu),\mu)}.$$

The regret measures to the number of customers to whom non-optimal assortments are offered by π over $\{1,\ldots,T\}$. One may also view this as a normalized measure of revenue loss due to the lack of a priori knowledge of consumer behavior.

Maximizing expected cumulative revenues is equivalent to minimizing the regret over the selling season, and to this end, the retailer must balance suboptimal demand exploration (which adds directly to the regret) with exploitation of the gathered information. On the one hand the retailer has incentives to explore demand extensively in order to "guess" the optimal assortment $S^*(\mu)$ with high probability. On the other hand the longer the exploration lasts the less consumers will be offered the "optimal" assortment, and therefore the retailer has incentives to shorten the length of the exploration phase in favor to the exploitation phase.

4 Fundamental Limits on Achievable Performance

4.1 A lower bound on the performance of any admissible policy

Let us begin by narrowing down the set of "interesting" policies worthy of consideration. We say that an admissible policy is *consistent* if for all $\mu \in \mathbb{R}_+^N$

$$\frac{\mathcal{R}^\pi(T,\mu)}{T^a} \to 0, \tag{4}$$

as $T \to \infty$, for every $a > 0$. In other words, the per-consumer normalized revenue of consistent policies converges to 1. The restriction in (4) ensures that this convergence occurs at least at a polynomial rate in T. Let $\mathcal{P}' \subseteq \mathcal{P}$ denote the set of non-anticipating, consistent assortment policies.

Let $\underline{\mathcal{N}}$ denote the set of products that cannot be made to enter the optimal assortment by increasing/decreasing, *ceteris paribus*, their mean utilities. Namely,

$$\underline{\mathcal{N}} := \left\{ i \in \mathcal{N} \,:\, i \notin S^*(\nu)\,,\, \nu := (\mu_1,\ldots,\mu_{i-1},v,\mu_{i+1},\ldots,\mu_N)\,,\, \forall\, v \in \mathbb{R}_+ \right\}.$$

We will refer to a product in $\underline{\mathcal{N}}$ as *strictly-suboptimal*. Similarly, define $\overline{\mathcal{N}}$ as the set of non-optimal products that *can* be made to enter the optimal assortment by increasing/decreasing their mean utility,

$$\overline{\mathcal{N}} := \mathcal{N} \setminus (S^*(\mu) \cup \underline{\mathcal{N}}).$$

We will assume the common distribution function F is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$, and that its density function is positive everywhere. This assumption is quite standard and satisfied by many commonly used distributions. Roughly speaking, this implies that one cannot infer products' mean utilities solely from observing the associated utility realizations. The result below establishes a fundamental limit on what can be achieved by any consistent assortment policy.

Theorem 1. *For any $\pi \in \mathcal{P}'$, and any $\mu \in \mathbb{R}_+^N$,*

$$\mathcal{R}^\pi(T, \mu) \geq K_1(\mu) \frac{\left|\overline{\mathcal{N}}\right|}{C} \log T + K_2(\mu),$$

for finite positive constants $K_1(\mu)$ and $K_2(\mu)$, and all T.

This result asserts that *any* consistent assortment policy *must* offer non-optimal assortments to *at least* order $(\left|\overline{\mathcal{N}}\right|/C) \log T$ customers (in expectation), and that this holds for all values of μ. (Explicit expressions for the constants $K_1(\mu)$ and $K_2(\mu)$ are given in the proof.). When *all* non-optimal products are strictly-suboptimal, the result suggest that a finite regret may be attainable. This last observation highlights the importance of strictly-suboptimal product detection, and hence the inefficiency of a naive multi-armed bandit approach to assortment planning: treating each possible assortment as a different arm in the bandit setting will result in the regret scaling linearly with the combinatorial term $\binom{N}{C}$, instead of the much smaller constant $(\left|\overline{\mathcal{N}}\right|/C)$.

Remark 2 (**Implications for design of "good" policies.).** The proof of Theorem 1, which can be found in the e-companion to this paper and is outlined below, suggests certain desirable properties for "optimal" policies: (i.) non-optimal products that can be made to be part of the optimal assortment are to be tested on order-$(\log T)$ customers; (ii) this type of non-optimal product experimentation is to be conducted in batches of size C; and (iii.) *strictly-suboptimal* products (the ones that cannot be made to be part of the optimal assortment) need only be tested on a *finite* number of customers (in expectation), independent of T.

4.2 Proof outline and intuition behind Theorem 1

Intuition and main underlying ideas. For the purpose of proving Theorem 1 we will exploit the connection between the regret and testing of suboptimal assortments. In particular, we will bound

the regret by computing lower bounds on the expected number of tests involving non-optimal products (those in $\mathcal{N} \setminus S^*(\mu)$): each time a non-optimal product is offered, the corresponding assortment must be sub-optimal, contributing directly to the policy's regret.

To bound the number of tests involving non-optimal products we will use a change-of-measure argument introduced by (Lai and Robbins 1985) for proving an analogous result for a multi-armed bandit problem, hence our proof establishes a direct connection between the two areas. To adapt this idea we consider the fact that underlying realizations of the random variables (product utilities) are non-observable in the assortment setting, which differs from the multi-armed bandit setting where reward realizations are observed directly. The argument can be roughly described as follows. Any non-optimal product $i \in \overline{\mathcal{N}}$ is in the optimal assortment for at least one suitable choice of mean utility vector μ^i. When such a configuration is considered, any consistent policy π must offer this non-optimal product to all but a sub-polynomial (in T) number of customers. If this configuration does not differ in a significant manner from the original (made precise in the e-companion to this paper), then one would expect such a product to be offered to a "large" number of customers under the μ-configuration. In particular, we prove that for any policy π

$$\mathbb{P}_\pi \left\{ T_i(T) \leq \log T / K_i \right\} \to 0 \tag{5}$$

as $T \to \infty$, where $T_i(t)$ is the number of customers product i has been offered to up until customer $t-1$, and K_i is a finite positive constant. The relation in (5) says that, asymptotically, any non-optimal product that can be "made" optimal must be offered to at least order-$(\log T / K_i)$ customers. Note that this asymptotic minimum-testing requirement is inversely proportional to K_i, which turns out to be a measure of "closeness" of a product to "optimality"(how close the vector μ is to a configuration that makes product i be part of the optimal assortment). This also has immediate consequences on the expected number of times a non-optimal product is tested: using Markov's inequality we have that for any $i \in \overline{\mathcal{N}}$,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\pi \left\{ T_i(T) \right\}}{\log T} \geq \frac{1}{K_i}.$$

The result in Theorem 1 follows directly from the equation above and the connection between the regret and testing of suboptimal assortments mentioned at the beginning of this section.

5 Dynamic Assortment Planning Policies

This section introduces an assortment policy whose structure is guided by the key ideas gleaned from Theorem 1. We introduce the following natural assumption.

Assumption 1 (Identifiability). *For any assortment $S \in \mathcal{S}$, and any vector $\rho \in \mathbb{R}_+^N$ such that $\sum_{i \in S} \rho_i < 1$, the system of equations $\{p_i(S, \eta) = \rho_i \,,\, i \in S\}$ has a unique solution $\mathcal{T}(S, \rho)$ in $\eta \in \mathbb{R}_+^N$*

such that $\eta_i = 0$ *for all* $i \notin S$. *In addition* $p(S,\cdot)$ *is Lipschitz continuous, and* $\mathcal{T}(S,\cdot)$ *is locally Lipschitz continuous in the neighborhood of* ρ, *for all* $S \in \mathcal{S}$.

Under this assumption one can compute mean utilities for products in a given assortment based solely on the associated purchase probabilities. This characteristic enables our approach to parameter estimation: we will estimate purchase probabilities by observing consumer decisions during an exploration phase and we will use those probabilities to reconstruct a mean utility vector that rationalizes such observed behavior. Note that the Logit model, for which F is a standard Gumbel, satisfies this assumption.

5.1 Intuition and a simple "separation-based" policy

To build some intuition towards the construction of our ultimate dynamic assortment policy (given in §5.2) it is helpful to first consider a policy that *separates* exploration from exploitation. The idea is to insolate the effect of imposing the "right" order of exploration (suggested by Theorem 2) on the regret. Assuming prior knowledge of T, such a policy first engages in an exploration phase over $\lceil N/C \rceil$ assortments encompassing all products, each offered sequentially to order-$(\log T)$ customers. Then, in light of Assumption 1, an estimator for μ is computed. Based on this estimator a proxy for the optimal assortment is computed, and offered to the remaining customers. For this purpose consider the set of test-assortments $\mathcal{A} := \left\{A_1, \ldots, A_{\lceil N/C \rceil}\right\}$, where

$$A_j = \{(j-1)\,C + 1, \ldots, \min\{j\,C, N\}\},$$

Fix $j \leq |\mathcal{A}|$. Suppose $t-1$ customers have arrived to that point. We will use $\hat{p}_{i,t}$ to estimate $p_i(A_j, \mu)$ when customer t arrives, where

$$\hat{p}_{i,t}(A_j) := \frac{\sum_{u=1}^{t-1} X_i^u \, \mathbf{1}\{S_u = A_j\}}{\sum_{u=1}^{t-1} \mathbf{1}\{S_u = A_j\}},$$

for $i \in A_j$. Define $\hat{p}_t(A_j) := (\hat{p}_{1,t}(A_j), \ldots, \hat{p}(A_j)_{N,t})$ to be the vector of product selection probabilities. For any $i \in A_j$ we will use $\hat{\mu}_{t,i}(A_j)$ to estimate μ_i when customer t arrives, where

$$\hat{\mu}_{i,t}(A_j) := \left(\mathcal{T}(A_j, \hat{p}_t(A_j))\right)_i,$$

and $(a)_i$ denotes the i-th component of vector a. Let $\hat{\mu}_t := (\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{N,t})$ denote the vector of mean utilities estimates. (In all of the above we are suppressing the dependence on A_j and in particular the index j, to avoid cluttering the notation.)

The underlying idea is the following: when an assortment $A_j \in \mathcal{A}$ has been offered to a "large" number of customers one expects $\hat{p}_{t,i}$ to be "close" to $p_i(A_j, \mu)$ for all $i \in A_j$. If this is the case for all assortments in $\mathcal{A}$, by Assumption 1 we also expect $\hat{\mu}_t$ to be close to μ. With this in mind,

Algorithm 1 : $\pi_1 = \pi(\mathcal{C}, T, w)$

STEP 1. Exploration:

for j =1 to $|\mathcal{A}|$ **do**

Offer A_j to $\mathcal{C}\log T$ customers (if possible). [Exploration]

end for

STEP 2. Exploitation:

for t =$(\mathcal{C}\log T)\,|\mathcal{A}|+1$ to T **do**

for j =1 to $|\mathcal{A}|$ **do**

Set $\hat{p}_{i,t}(A_j) := \dfrac{\sum_{u=1}^{t-1} X_i^u\,\mathbf{1}\{S_u = A_j\}}{\sum_{u=1}^{t-1}\mathbf{1}\{S_u = A_j\}}$ for $i \in A_j$. [Probability estimates]

Set $\hat{\mu}_{i,t}(A_j) := \eta_i$ for $i \in A_j$, where $\eta = \mathcal{T}(A_j, \hat{p}_t)$. [Mean utility estimates]

end for

Offer $S_t = S^*(\hat{\mu}_t)$ to customer t. [Exploitation]

end for

we propose a separation-based policy defined through a positive constant $\mathcal{C}$ that serves as a tuning parameter. The policy is summarized for convenience in Algorithm 1.

Performance analysis. This policy is constructed to guarantee that the probability of not choosing the optimal assortment decays polynomially (in T) when using the estimator $\hat{\mu}$ to compute product choice probabilities. This, in turn, translates into a $O(\lceil N/C\rceil \log T)$ regret. The next result, whose proof can be found in the e-companion to this paper, formalizes this.

Theorem 2. *Let $\pi_1 := \pi(\mathcal{C}, T, w)$ be defined by Algorithm 1 and let Assumption 1 hold. Then, for some finite constants $K_1, K_2 > 0$, the regret associated with π_1 is bounded for all T as follows*

$$\mathcal{R}^{\pi}(T,\mu) \leq \mathcal{C}\lceil N/C\rceil \log T + K_1,$$

provided that $\mathcal{C} > K_2$.

Constants K_1 and K_2 depend on instance specific quantities (e.g., the minimum optimality gap), but not on the number of products, N, or the length of the selling horizon, T. The proof of Theorem 2 elucidates that K_1 is the expected cumulative loss during the exploitation phase for an infinite horizon setting, while K_2 represents the minimum length of the exploration phase that makes K_1 finite. This trade off is balanced by the construction of the policy π_1. The quantity on the right hand side of 2 is essentially the one in Theorem 1, where $\left|\overline{N}\right|$ is replaced by N. This indicates that: (i.) imposing the right order (in T) of exploration is enough to get the right dependence (in T) of the regret; and (ii.) to reach the fundamental limit one needs to limit exploration on strictly-suboptimal products.

Example 1: Performance of the Separation-based policy π_1. Consider $N = 10$ and $C = 4$, with

$$\begin{aligned} w &= (0.98, 0.88, 0.82, 0.77, 0.71, 0.60, 0.57, 0.16, 0.04, 0.02), \\ \mu &= (0.36, 0.84, 0.62, 0.64, 0.80, 0.31, 0.84, 0.78, 0.38, 0.34), \end{aligned}$$

and assume $\{\zeta_i^t\}$ have a standard Gumbel distribution, for all $i \in \mathcal{N}$ and all $t \geq 1$, i.e., we consider the MNL choice model. In this case it is easily verified that $S^*(\mu) = \{1, 2, 3, 4\}$ and $f(S^*(\mu), \mu) = 0.76$. Also, $\underline{\mathcal{N}} = \{5, 6, 7, 8, 9, 10\}$. One can use the test assortments $A_1 = \{1, 2, 3, 4\}$, $A_2 = \{5, 6, 7, 8\}$ and $A_3 = \{9, 10\}$ to conduct the exploration phase in the algorithm described above. Figure 1 depicts the average performance of policy π_1 over 500 replications, using $\mathcal{C} = 20$, and considering selling horizons ranging from $T = 500$ to $T = 10000$. Two important points are worth noting: from panel (a) we observe that the regret is indeed of order-$(\log T)$, as predicted by Theorem 2; from panel (b) we observe that policy π_1 makes suboptimal decisions on a very small fraction of customers, ranging from around 10% when the horizon is 2000 sales attempts, and diminishing to around 2.5% for a horizon of 10,000. (Recall that the regret is measuring the number of suboptimal sales.)

From the setting of this example we observe that A_2 and A_3 are tested on order-$(\log T)$ customers, despite being composed exclusively of *strictly-suboptimal* products. That is, the separation algorithm does not attempt to limit testing efforts over suboptimal products. Moreover, it assumes a priori knowledge of the total number of customers, T. The next section proposes a policy that addresses these two issues.

5.2 A refined dynamic assortment policy

To account for strictly-suboptimal product detection it is necessary to be able to "identify" them a priori, even under partial knowledge of the mean utility vector. For that purpose we introduce the following assumption

Assumption 2 (Revenue preferences). *For any two vectors $\nu, \mu \in \mathbb{R}_+^N$ such that $\nu \leq \mu$ (component-wise),*

$$f(S^*(\nu), \nu) \leq f(S^*(\mu), \mu).$$

This revenue preferences assumption states that the retailer always prefers to sell *better* products (i.e., those with a higher mean utility). We should note that this assumption may not hold in general since, for example, increasing the mean utility of an optimal low-margin product may reduce the optimal single-sale expected profit. However, this property does hold for Luce-type choice models (the MNL being a special case). Under Assumption 2, we have that

$$\underline{\mathcal{N}} = \{i \in \mathcal{N} \,:\, w_i < f(S^*(\mu), \mu)\}.$$

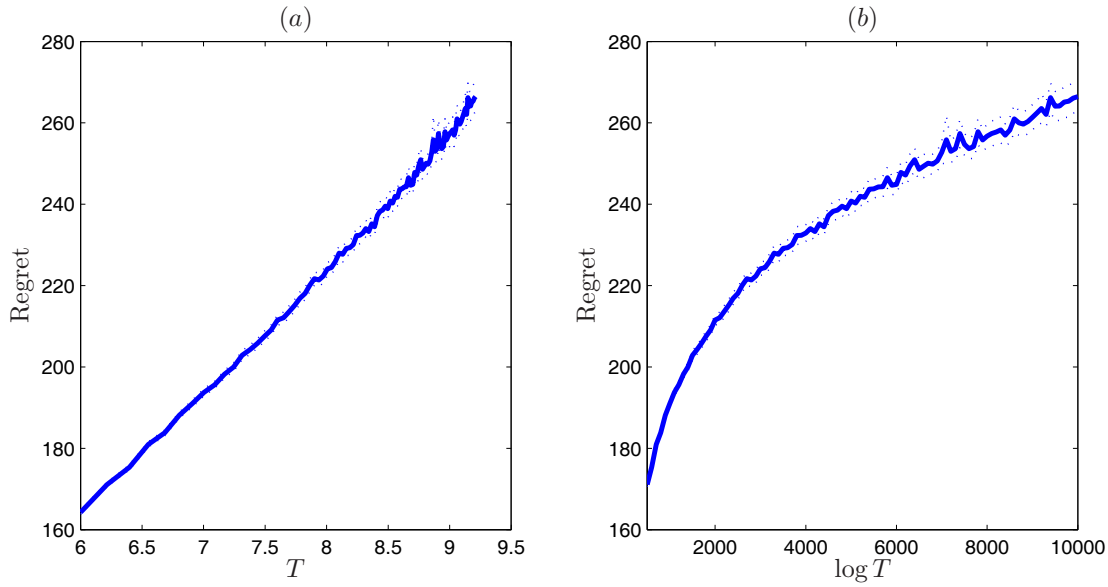

Figure 1: **Performance of the separation-based policy π_1.** The graphs (a) and (b) illustrates the dependence of the regret on T and $\log T$, respectively. The dotted lines represent 95% confidence intervals for the simulation results.

That is, any product with margin less than the optimal single sale profit is *strictly-suboptimal* and vice versa. The implications for strictly-suboptimal product detection are as follows: the value of $f(S^*(\mu), \mu)$ acts as a threshold, differentiating potentially optimal products from strictly-suboptimal ones. Designing test assortments based on product margins translates this to a threshold over assortments. Consider the set of valid assortments $\mathcal{A} := \{A_1, \ldots, A_{\lceil N/C \rceil}\}$, where

$$A_j = \left\{ i_{((j-1)\,C+1)}, \ldots, i_{(\min\{j\,C, N\})} \right\},$$

and $i_{(k)}$ corresponds to the product with the k-th highest margin in w. Suppose one has a proxy for $f(S^*(\mu), \mu)$. One can then use this value to identify assortments containing at least one potentially optimal product and to force the "right" order of exploration on such assortments. If successful, such a scheme will limit exploration on assortment containing only strictly-suboptimal products.

We propose an assortment policy, that, for each customer executes the following logic: using the current estimate of μ at time t, the static problem is solved and S_t, the estimate-based optimal assortment, and f_t, the estimate of the optimal value, are obtained. If all assortments in $\mathcal{A}$ containing products with margins greater than or equal to f_t have been tested on a minimum number of customers, then assortment S_t is offered to the t-th customer. Otherwise, we select, arbitrarily, an "under-tested" assortment in $\mathcal{A}$ containing at least one product with margin greater than or equal to f_t, and offer it to the current customer. The term "under-tested" means tested on less

than order-$(\log t)$ customers prior arrival of customer t. Note that this logic will enforce the correct order of exploration for any value of T.

Algorithm 2 : $\pi_2 = \pi(\mathcal{K}, w)$

STEP 1. Initialization:

for t =1 to $|\mathcal{A}|$ **do**

Offer $A_t \in \mathcal{A}$ to customer t and set $n_t = 1$. [Initial test]

end for

STEP 2. Joint exploration and assortment optimization:

for $t = |\mathcal{A}| + 1$ to T **do**

for j =1 to $|\mathcal{A}|$ **do**

Set $\hat{p}_{i,t}(A_j) := \frac{\sum_{u=1}^{t-1} X_i^u \mathbf{1}\{S_u = A_j\}}{\sum_{u=1}^{t-1} \mathbf{1}\{S_u = A_j\}}$ for $i \in A_j$. [Probability estimates]

Set $\hat{\mu}_{i,t} := \eta_i$ for $i \in A_j$, where $\eta = \mathcal{T}(A_j, \hat{p}_t(A_j))$. [Mean utility estimates]

end for

Set $S_t = S^*(\hat{\mu}_t)$ and $f_t = f(S^*(\hat{\mu}_t), \hat{\mu}_t)$. [Static optimization]

Set $\overline{\mathcal{N}}_t = \{i \in \mathcal{N} : w_i \geq f_t\}$. [Candidate optimal products]

if $(n_j \geq \mathcal{K} \log t)$ for all j=1 to $|\mathcal{A}|$ such that $A_j \cap \overline{\mathcal{N}}_t \neq \emptyset$ **then**

Offer S_t to customer t. [Exploitation]

else

Select j such that $A_j \cap \widehat{\overline{\mathcal{N}}}_t \neq \emptyset$ and $n_j < \mathcal{K} \log T$.

Offer A_j to customer t. [Exploration]

$n_j \leftarrow n_j + 1$

end if

end for

This policy, denoted π_2 and summarized for convenience in Algorithm 2, monitors the quality of the estimates for potentially optimal products by imposing *minimum exploration* on assortments containing such products. The specific structure of $\mathcal{A}$ ensures that test assortments do not "mix" high-margin products with low-margin products, thus successfully limiting exploration on strictly-suboptimal products. The policy uses a tuning parameter $\mathcal{K}$ to balance non-optimal assortment testing (which contributes directly to the regret), and the probability of choosing the optimal assortment in the exploitation phase.

Performance analysis. The next result, whose proof can be found in the e-companion to this paper, characterizes the performance of the proposed assortment policy.

Theorem 3. *Let $\pi_2 = \pi(\mathcal{K}, w)$ be defined by Algorithm 2 and let Assumptions 1 and 2 hold. Then, for some finite constants $K_1, K_2 > 0$, the regret associated with π_2 is bounded for all T as follows*

$$\mathcal{R}^{\pi}(T, \mu) \leq \mathcal{K} \lceil \left| \overline{\mathcal{N}} \cup S^*(\mu) \right| / C \rceil \log T + K_1,$$

provided that $\mathcal{K} \geq K_2$.

Theorem 3 implicitly states that assortments containing only strictly-suboptimal products will be tested on a finite number of customers (in expectation); see Corollary 1 below. Note that this policy attains the correct dependence on both T and $\left|\overline{\mathcal{N}}\right|$, as prescribed in Theorem 1 (up to constant values), so it is essentially optimal. Unlike Theorem 2 we see that the proposed policy successfully limits exploration on strictly-suboptimal products. The following corollary, whose proof can be found in the e-companion to this paper, formalizes this statement. Recall $T_i(t)$ denotes the number of customers product i has been offered to up to arrival of customer t.

Corollary 1. *Let Assumptions 1 and 2 hold. Then, for any assortment* $A_j \in \mathcal{A}$ *such that* $A_j \subseteq \underline{\mathcal{N}}$*, and for any selling horizon* T

$$\mathbb{E}_\pi[T_i(T)] \leq K_3,$$

for all $i \in A_j$*, where* K_3 *is a finite positive constant independent of* T*.*

Example 2: Performance of the policy π_2. Consider the setting in Example 1 in section 5.1, for which $\underline{\mathcal{N}} = A_2 \cup A_3$ and $S^*(\mu) = A_1$. Given that the set of test assortments separates products in $\overline{\mathcal{N}}$ from the rest, one would expect Algorithm 2 to effectively limit exploration on all strictly-suboptimal products. Figure 2 depicts the average performance of policies π_1 and π_2 over 500 replications, using $\mathcal{C} = \mathcal{K} = 20$, and considering selling horizons ranging from $T = 500$ to $T = 10000$. The main point to note is that policy π_2 outperforms substantially the separation-based policy π_1. In particular, the operation of π_1 results in lost sales in the range of 2.5-10% (200-260 customers are offered non-optimal choices), depending on the length of selling horizon, while for π_2 we observe sub-optimal decisions being made only about 10-20 times (!) independent of the horizon. This constitutes more than a 10-fold improvement over the performance of π_1. In essence, π_2 adaptively identifies both A_2 and A_3 as suboptimal with increasing probability as t grows large. Since, in this case, the regret is due exclusively to exploration of strictly-suboptimal assortments and incorrect choices in the exploitation phase, we expect the regret to be finite, and this indeed is supported by the numerical results displayed in Figure 2.

Remark 3 (Relationship to bandit problems). The result in Corollary 1 stands in contrast to typical multi-armed bandit results, where *all* suboptimal arms/actions need to be tried at least order-$(\log t)$ times (in expectation). In the assortment problem, product rewards are random variables bounded above by their corresponding margins, therefore, under Assumption 2, the contribution of a product to the overall profit is bounded, independent of its mean utility. More importantly, this features makes some products a priori *better* than others. Such characteristic is not present in the typical bandit problem.

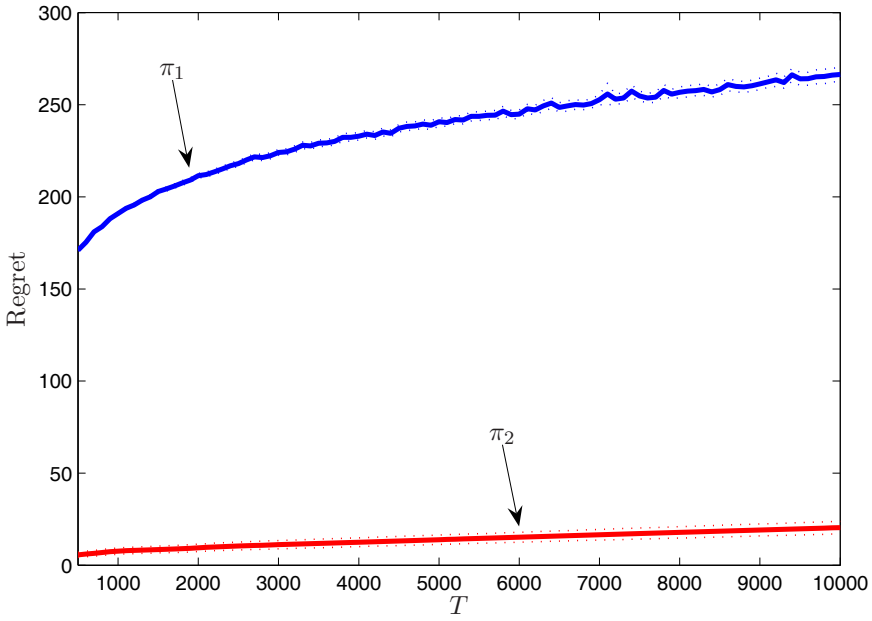

Figure 2: **Performance of the refined policy π_2**: The graph compares the separation-based policy π_1, given by Algorithm 1, and the proposed policy π_2, in terms of regret dependence on T. The dotted lines represent 95% confidence intervals for the simulation result.

Remark 4 (**Selection of the tuning parameter $\mathcal{K}$).** We have established that the lower bound in Theorem 1 can be achieved up to constant terms, for proper choice of $\mathcal{K}$. However, our prescription for $\mathcal{K}$ depends on values that are not known a priori. In particular, setting $\mathcal{K}$ below the specified threshold may compromise the validity of the result. To avoid the risk of miss-specifying $\mathcal{K}$, one can increase the order of the minimum amount of exploration to, say, $\mathcal{K}\log^{1+\alpha} t$, for any $\alpha > 0$. With this, the upper bound above would read

$$\mathcal{R}^{\pi}(T,\mu) \leq \lceil \left|\overline{\mathcal{N}} \cup S^*(\mu)\right| / C \rceil \mathcal{K} \log^{1+\alpha} T + K_1,$$

and the policy becomes optimal up to a $\log^{\alpha} T$-term.

Remark 5 (**Performance of π_2 in absence of Assumption 2).** In absence of Assumption 2 it seems impossible to identify strictly-suboptimal products a priori. Instead, one can modify Algorithm 2 to simply *ignore* strictly-suboptimal product detection. It can be then seen from the proof of Theorem 3 that the upper bound remains valid, with N replacing $\left|\overline{\mathcal{N}} \cup S^*(\mu)\right|$.

5.3 A policy customized to the multinomial logit choice model

For general utility distributions, choice probabilities depend on the offered assortment in a non-trivial way, and hence it is unclear how to combine information originating from different assortments and allow for more efficient use of data gathered on the exploitation phase. We illustrate how to modify parameter estimation to include exploitation-based product information in the case of an MNL choice model (we note that all results in this section extend directly to Luce-type choice models). As indicated earlier, for this model an efficient algorithm for solving the static optimization problem has been developed by Rusmevichientong, Shen, and Shmoys (2008).

MNL choice model properties. Taking F to have a standard Gumbel distribution, then (see, for example, Anderson, De Palma, and Thisse (1992))

$$p_i(S,\nu) = \frac{\nu_i}{1+\sum_{j\in S}\nu_j} \qquad i \in S\,, \text{ for any } S \in \mathcal{S}, \tag{6}$$

where $\nu_i := \exp(\mu_i)$, $i \in \mathcal{N}$, and $v := (v_1, \ldots, v_N)$. In what follows, we will use both ν and μ interchangeably. Given an assortment $S \in \mathcal{S}$ and a vector $\rho \in \mathbb{R}^N_+$ such that $\sum_{i\in S}\rho_i \leq 1$, we have that $\mathcal{T}(S,\rho)$, the unique solution to $\{\rho_i = p_i(S,\nu)$ for $i \in S\,,\ \nu_i = 0$ for $i \in \mathcal{N}\setminus S\}$ is given by

$$\mathcal{T}_i(S,\rho) = \frac{\rho_i}{1-\sum_{j=1}^N \rho_j} \qquad i \in S. \tag{7}$$

From (6) one can see that solving the static optimization problem is equivalent to finding the *largest* value of λ such that

$$\sum_{i\in S} v_i(w_i - \lambda) \geq \lambda, \tag{8}$$

for some $S \in \mathcal{S}$. One can check that (7) and (8) implies that Assumptions 1 and 2 holds, respectively.

A product-exploration-based assortment policy. We propose a customized version of the policy given by Algorithm 2, which we refer to as π_3, defined through a positive constant $\mathcal{M}$ that serves as a tuning parameter. The policy, which is summarized below in algorithmic form, maintains the general structure of Algorithm 2, however parameter estimation, product testing and suboptimal product detection are conducted at the *product-level*. In what follows, the following estimators are used. Suppose $t-1$ customers have shown up so far. We will use $\hat{\nu}_{i,t}$ to estimate ν_i when customer t arrives, where

$$\hat{\nu}_{i,t} := \frac{\sum_{u=1}^{t-1} X_i^u \mathbf{1}\{i \in S_u\}}{\sum_{u=1}^{t-1} X_0^u \mathbf{1}\{i \in S_u\}} \qquad i \in \mathcal{N}.$$

Performance analysis. The tuning parameter $\mathcal{M}$ plays the same role as $\mathcal{K}$ plays in Algorithm 2. The next result, whose proof can be found in the e-companion to this paper, characterizes the performance of the proposed assortment policy.

Algorithm 3 : $\pi_3 = \pi(\mathcal{M}, w)$

STEP 1. Initialization:

Set $n_j = 0$ for all $j \in \mathcal{N}$.

Offer $S_1 = \text{argmax}\{w_j : j \in \mathcal{N}\}$ to customer $t = 1$. Set $n_i = 1$.

STEP 2. Joint exploration and assortment optimization:

for t =2 to T **do**

for i =1 to N **do**

Set $\hat{\nu}_{i,t} := \left(\sum_{u=1}^{t-1} X_i^u \mathbf{1}\{i \in S_u\}\right) / \left(\sum_{u=1}^{t-1} X_0^u \mathbf{1}\{i \in S_u\}\right)$. [Mean utility estimates]

end for

Set $S_t = S^*(\hat{\nu}_t)$ and $f_t = f(S^*(\hat{\nu}_t), \hat{\nu}_t)$. [Static optimization]

Set $O_t = \{i \in \mathcal{N} : w_i \geq f_t\,,\, n_i < \mathcal{M} \log t\}$. [Candidate optimal products]

if $O_t = \emptyset$ **then**

Offer S_t to customer t. [Exploitation]

else

Offer $S_t \in \{S \in \mathcal{S} : S \subseteq O_t\}$. [Exploration]

end if

$n_i \leftarrow n_i + 1$ for all $i \in S_t$.

end for

Theorem 4. *Let $\pi_3 = \pi(\mathcal{M}, w)$ be defined by Algorithm 3. Then, for some finite constants $K_1, K_2 > 0$, the regret associated for π_3 is bounded as follows*

$$\mathcal{R}^{\pi}(T, \mu) \leq \mathcal{M} \left|\overline{\mathcal{N}}\right| \log T + K_1,$$

provided that $\mathcal{M} > K_2$.

Theorem 4 is essentially the equivalent of Theorem 3 for the Logit case, with the exception of the dependence on the assortment capacity C (as here exploration is conducted on a product basis), and on the cardinality of the set $\overline{\mathcal{N}}$. The latter matches exactly the order of the result in Theorem 1: unlike policy π_2, the customized policy π_3 prevents optimal products from being offered in suboptimal assortments. Since estimation is conducting using information arising from both exploration and exploitation phases, one would expect a better empirical performance from the Logit customized policy. Note that the result implicitly states that strictly-suboptimal products will be tested on a finite number of customer, in expectation. The following corollary, whose proof can be found in the e-companion to this paper, is the MNL-customized version of Corollary 1.

Corollary 2. *For any strictly-suboptimal product $i \in \underline{\mathcal{N}}$ and for any selling horizon T*

$$\mathbb{E}_{\pi}[T_i(T)] \leq K_3,$$

for a positive finite constant K_3, independent of T.

Example 3: Performance of the MNL-customized policy π_3. Consider the set up of Example 1 in section 5.1. Note that $S^*(\nu) = A_2$, i.e., the optimal assortment matches one of the test assortments. As a result, strictly suboptimal detection is conducted in finite time for both policies π_2 and π_3, and hence any gain in performance for policy π_3 over π_2 is tied in to the ability of the former incorporate information gathered during both exploitation and exploration phases. Figure 3 depicts the average performance of policies π_2 and π_3 over 500 replications, using $\mathcal{K} = \mathcal{M} = 20$, and considering selling horizons ranging from $T = 1000$ to $T = 10000$. Customization to a logit nets significant, roughly 10-fold, improvement in performance of π_3 relative to π_2. Overall, the logit-customized policy π_3 only offers suboptimal assortments to less than a handful of customers, regardless of the horizon of the problem. This provides "picture proof" that the regret (=number of suboptimal sales) is finite. In particular, since $\underline{\mathcal{N}} = \emptyset$ Theorem 4 predicts a finite regret. This

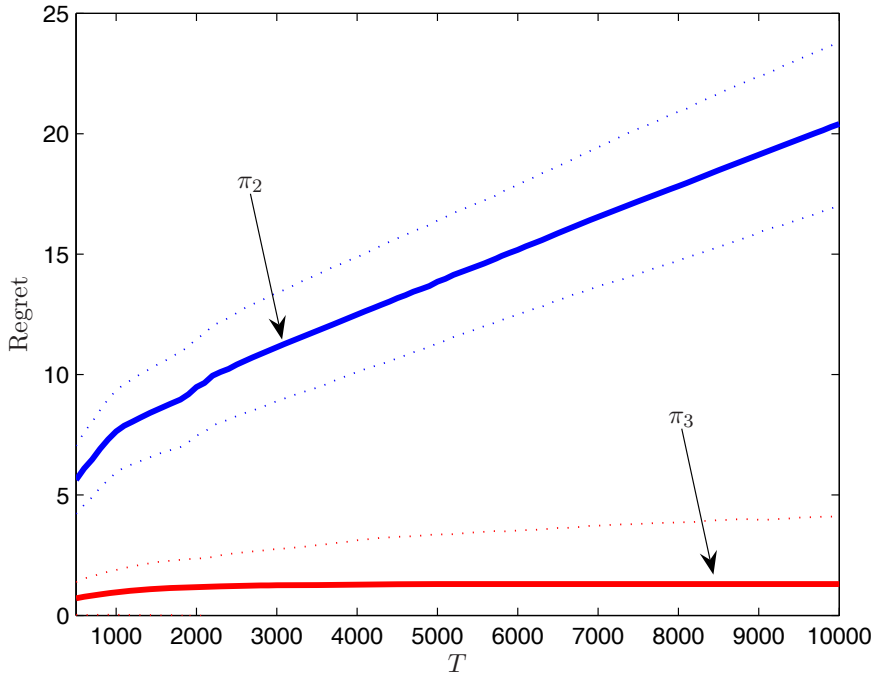

Figure 3: **Performance of the MNL-customized policy π_3.** The graph compares the more general policy π_2 to its Logit-customized version π_3, in terms of regret dependence on T. The dotted lines represent 95% confidence intervals for the simulation result.

suggests that difference in performance is mainly due to errors made in the exploitation phase. This elucidates the reason why the Logit customized policy π_3 outperforms the more general policy π_2: the probability of error decays much faster in the Logit customized version. If all previous exploitation efforts were successful, and assuming correct strictly-suboptimal product detection,

the probability of error decays exponentially for the customized policy (π_3) and polynomially for the more general policy (π_2); see proof for further details.

6 Concluding Remarks

Complexity of the dynamic assortment problem. Theorem 1 provides a lower bound for the regret of an optimal policy for the dynamic assortment problem. We have shown that this lower bound can be achieved, up to constant terms, when the noise distribution on the utility of each customer is known. In particular, we proposed an assortment-exploration-based algorithm whose regret scales optimally in the selling horizon T and, exhibits the "right" dependence on the number of possible optimal products $\left|\overline{\mathcal{N}}\right|$. (In addition our proposed policies do not require a priori knowledge of the length the selling horizon.)

Comparison of our policy with benchmark results. Our results significantly improve on and generalize the policy proposed by Rusmevichientong, Shen, and Shmoys (2008), where an order-$(N^2(\log T)^2)$ upper bound is presented for the case of an MNL choice model. Recall the regret of our policy exhibits order-$\left|\overline{\mathcal{N}}\right| \log T$ performance, and we show that this can not be improved upon. We note that the policy of Rusmevichientong, Shen, and Shmoys (2008) is a more direct adaptation of multi armed bandit ideas an hence does not detect strict-suboptimal products and does not limit exploration on them. We illustrate this with a simple numerical example

Consider again Example 1 in section 5.1. Figure 4 compares the average performance of our proposed policies with that of (Rusmevichientong, Shen, and Shmoys 2008),denoted RSS for short, over 500 replications, using $\mathcal{C} = \mathcal{K} = \mathcal{M} = 20$, and considering selling horizons ranging from $T = 1000$ to $T = 10000$. From graph (a) one can see that the performance of the benchmark behaves quadratically with $\log T$, while the performance of our proposed policies grow linearly.

Several factors explain the difference in performance. First, we consider a set of roughly N test assortments while in RSS this set contains roughly N^2 items. This explains why even the naive separation-based policy π_1 outperforms RSS. Panel (a) in Figure 4 shows that the RSS policy loses sales on about $20 - 25\%$ of the customers, while policy π_1 never loses more than 10%, the loss diminishes as the horizon increases to around 2.5%. Since policies π_2 and π_3 limit exploration on strictly-suboptimal products, a feature absent in both RSS and in the naive separation-based policy π_1, they exhibit far superior performance compared to either one of those benchmarks as illustrated in panel (b) of Figure 4. Finally, our MNL-customized policy π_3 uses all information gathered for computing parameter estimates, while the policy in RSS only uses the information collected during the exploration phase. The improvement in performance due to this feature is also

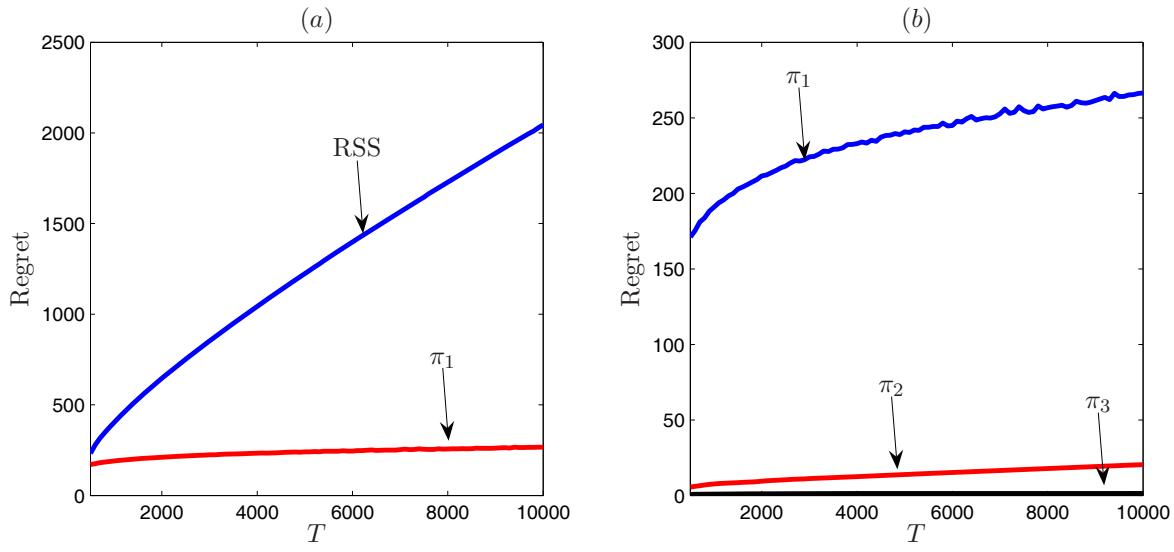

Figure 4: **Benchmark performance.** The graph in (a) compares the separation-based policy π_1 to the benchmark policy RSS, in terms of regret dependence on T. The graph in (b) compares the separation-based policy π_1, the proposed policy π_2 and its Logit-customized version π_3 in terms of regret dependence on T.

illustrated in panel (b) of Figure 4. The overall effect is that policy π_3 improves performance by a factor of 200-1000 compared to RSS, and is able to zero in on the optimal assortment almost instantaneously with a regret that is bounded independent of the horizon T.

References

Anderson, S. P., A. De Palma, and J. F. Thisse (1992). *Discrete Choice Theory of Product Differentiation*. MIT Press.

Auer, P., N. Cesa-Bianchi, and P. Fisher (2002). Finite-time analysis of the multiarmed bandit problem. In *Machine Learning*, pp. 235–256.

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

Farias, V. and R. Madan (2009). The irrevocable multi-armed bandit problem. Working paper.

Gaur, V. and D. Honhon (2006). Assortment planning and inventory decisions under a locational choice model. *Manage. Sci. 52*(10), 1528–1543.

Goyal, V., R. Levi, and D. Segev (2009). Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. Working paper.

Honhon, D., V. Gaur, and S. Seshadri (2009). Assortment planning and inventory decisions under stoc-out based substitution. Working paper.

Hopp, W. and X. Xu (2008). A atatic approximation for dynamic substitution with applications in a competitive market. *Operations Reseach 56*(3), 630 – 645.

Kok, A., M. Fisher, and R. Vaidyanathan (2006). Assortment planning: Review of literature and industry practice. In *Retail Supply Chain Management*. Kluwers.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6*(1), 4 – 22.

Mahajan, S. and G. van Ryzin (2001). Stocking retail assortments under dynamic consumer substitution. *Oper. Res. 49*(3), 334–351.

Megiddo, N. (1979). Combinatorial optimization with rational objective functions. *Mathematics of Operations Reseach 4*(4), 414 – 424.

Rusmevichientong, P., Z.-J. M. Shen, and D. Shmoys (2008). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. Working paper.

van Ryzin, G. and S. Mahajan (1999). On the relationship between inventory costs and variety benefits in retail assortments. *Manage. Sci. 45*(11), 1496–1509.

A Proof of Main Results

Proof of Theorem 1. The lower bound is trivial when $\overline{\mathcal{N}} = \emptyset$, so assume $\left|\overline{\mathcal{N}}\right| > 0$. For $i \in \mathcal{N}$ define $T_i(t)$ as the number of customers product i has been offered to, before customer t's arrival,

$$T_i(t) := \sum_{u=1}^{t-1} \mathbf{1}\left\{i \in S_u\right\}\ ,\ t \geq 1.$$

Similarly, for $n \geq 1$ define $t_i(n)$ as the customer to whom product i is offered for the n-th time,

$$t_i(n) := \inf\left\{t \geq 1\ :\ T_i(t+1) = n\right\}\ ,\ n \geq 1.$$

For $i \in \overline{\mathcal{N}}$, define Θ_i as the set of mean utility vectors for which product i is in the optimal assortment, but that differs from μ only on its i-th coordinate. That is,

$$\Theta_i := \left\{\nu \in \mathbb{R}_+^N\ :\ \nu_i \neq \mu_i\,,\, \nu_j = \mu_j \quad \forall\, j \in \mathcal{N} \setminus \{i\}\ ,\ i \in S^*(\nu)\right\}.$$

We will use $\mathbb{E}_\pi^\nu$ and $\mathbb{P}_\pi^\nu$ to denote expectations and probabilities of random variables, when the assortment policy $\pi \in \mathcal{P}$ is used, and the mean utilities are given by the vector ν. Let $\mathcal{I}_i(\mu\|\nu)$ denote the Kullback-Leibler divergence between $F(\cdot - \mu_i)$ and $F(\cdot - \nu_i)$,

$$\mathcal{I}_i(\mu\|\nu) := \int_{-\infty}^{\infty} \left[\log\left(dF(x-\mu_i)/dF(x-\nu_i)\right)\right] dF(x-\mu_i).$$

This quantity measures the "distance" between $\mathbb{P}_\pi^\mu$ and $\mathbb{P}_\pi^\nu$. We have that $0 < \mathcal{I}_i(\mu\|\nu) < \infty$ for all $\nu \neq \mu$, $i \in \mathcal{N}$. Fix $i \in \overline{\mathcal{N}}$ and consider a configuration $\nu \in \Theta_i$. For $n \geq 1$ define the log-likelihood function

$$\mathcal{L}_i(n) := \sum_{u=1}^{n} \left[\log(dF(U_i^{t_i(u)} - \mu_i)/dF(U_i^{t_i(u)} - \nu_i))\right].$$

Note that $\mathcal{L}_i(\cdot)$ is defined in terms of utility realizations that are unobservable to the retailer. Define $\delta(\eta)$ as the minimum (relative) optimality gap when the mean utility vector is given by $\eta \in \mathbb{R}_+^N$,

$$\delta(\eta) := \inf\left\{1 - f(S,\eta)/f(S^*(\eta),\eta) > 0 \,:\, S \in \mathcal{S}\right\}. \tag{9}$$

Fix $\alpha \in (0,1)$. For any consistent policy π one has that for any $\epsilon > 0$,

$$\begin{aligned}
\mathcal{R}^\pi(T,\nu) &\geq \delta(\nu)\mathbb{E}_\pi^\nu\left\{T - T_i(T)\right\} \\
&\geq \delta(\nu)\left(T - \frac{(1-\epsilon)}{\mathcal{I}_i(\mu\|\nu)}\log T\right)\mathbb{P}_\pi^\nu\left\{T_i(T) < (1-\epsilon)\log T/\mathcal{I}_i(\mu\|\nu)\right\},
\end{aligned}$$

and by assumption on π $\mathcal{R}^\pi(T,\nu) = o(T^\alpha)$. From the above, we have that

$$\mathbb{P}_\pi^\nu\left\{T_i(T) < (1-\epsilon)\log T/\mathcal{I}_i(\mu\|\nu)\right\} = o(T^{\alpha-1}). \tag{10}$$

Define the event

$$\beta_i := \left\{T_i(T) \leq \frac{(1-\epsilon)}{\mathcal{I}_i(\mu\|\nu)}\log T\,,\, \mathcal{L}_i(T_i(T)) \leq (1-\alpha)\log T\right\}.$$

From the independence of utilities across products and the definition of β_i, we have that

$$\begin{aligned}
\mathbb{P}_\pi^\nu\left\{\beta_i\right\} &= \int_{\omega\in\beta_i} d\mathbb{P}_\pi^\nu \\
&= \int_{\omega\in\beta_i} \prod_{u=1}^{T-1}\prod_{i\in S_u} dF(U_i^u - \nu_i) \\
&= \int_{\omega\in\beta_i} \prod_{u=1}^{T-1}\prod_{i\in S_u} \frac{dF(U_i^u - \nu_i)}{dF(U_i^u - \mu_i)} d\mathbb{P}_\pi^\mu \\
&= \int_{\omega\in\beta_i} \prod_{n=1}^{T_i(T)} \frac{dF(U_i^{t_i(n)} - \nu_i)}{dF(U_i^{t_i(n)} - \mu_i)} d\mathbb{P}_\pi^\mu \\
&= \int_{\omega\in\beta_i} \exp(-\mathcal{L}_i(T_i(T)))d\mathbb{P}_\pi^\mu \\
&\geq \exp(-(1-\alpha)\log T)\mathbb{P}_\pi^\mu\left\{\beta_i\right\}.
\end{aligned}$$

From (10) one has that $\mathbb{P}_\pi^\nu\left\{\beta_i\right\} = o(T^{\alpha-1})$. It follows by (10) that as $T \to \infty$

$$\mathbb{P}_\pi^\mu\left\{\beta_i\right\} \leq \mathbb{P}_\pi^\nu\left\{\beta_i\right\}/T^{\alpha-1} \to 0. \tag{11}$$

Indexed by n, $\mathcal{L}_i(n)$ is the sum of finite mean identically distributed independent random variables, therefore, by the strong law of large numbers (SLLN).

$$\limsup_{n\to\infty} \frac{\max\{\mathcal{L}_i(l) \,:\, l \le n\}}{n} \le \frac{\mathcal{I}_i(\mu\|\nu)}{(1-\alpha)} \quad \mathbb{P}_\pi^\mu \, a.s.,$$

i.e., the log-likelihood function grows no faster than linearly with slope $\mathcal{I}_i(\mu||\nu)$. This implies that

$$\limsup_{n\to\infty} \mathbb{P}_\pi^\mu \{\exists\, l \le n\,,\, \mathcal{L}_i(l) > n\mathcal{I}_i(\mu\|\nu)/(1-\epsilon)\} = 0.$$

In particular,

$$\lim_{T\to\infty} \mathbb{P}_\pi^\mu \left\{ T_i(T) < \frac{(1-\epsilon)}{I_i(\mu\|\nu)} \log T \,,\, \mathcal{L}_i(T_i(T)) > \frac{(1-\epsilon)}{1-\alpha} \log T \right\} = 0.$$

Taking $\alpha < \epsilon$ small enough, and combining with (11) one has that

$$\lim_{T\to\infty} \mathbb{P}_\pi^\mu \left\{ T_i(T) < \frac{(1-\epsilon)}{I_i(\mu\|\nu)} \log T) \right\} = 0.$$

Finally, defining the positive finite constant $H_i^\mu := \inf\{\mathcal{I}(\mu\|\nu) \,:\, \nu \in \Theta_i\}$, it follows that

$$\lim_{T\to\infty} \mathbb{P}_\pi^\mu \{T_i(T) \ge (1-\epsilon)\log T/H_i^\mu)\} = 1.$$

By Markov's inequality, and letting ϵ shrink to zero we get

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\pi^\mu \{T_i(T)\}}{\log T} \ge \frac{1}{H_i^\mu}. \tag{12}$$

By the definition of the regret, we have that for any consistent policy $\pi \in \mathcal{P}'$,

$$\begin{aligned} \mathcal{R}^\pi(T,\mu) &\overset{(a)}{\ge} \delta(\mu)\, \mathbb{E}_\pi^\mu \left[\sum_{t=1}^T \mathbb{P}_\pi^\mu \mathbf{1}\{S_t \ne S^*(\mu)\} \right] \\ &\overset{(b)}{\ge} \delta(\mu) \frac{1}{C} \sum_{i\in\overline{\mathcal{N}}} \mathbb{E}_\pi^\mu \left[T_i(T)\right]. \end{aligned}$$

where (a) follows from the non-optimal assortments contributing at least $\delta(\mu)$ to the regret, and (b) follows by assuming non-optimal products are always tested in batches of size C, discarding non-optimal products in $\underline{\mathcal{N}}$. Thus

$$\sum_{u=1}^T \mathbf{1}\{S_u \ne S^*(\mu)\} \ge \sum_{u=1}^T \mathbf{1}\left\{S_u \cap \overline{\mathcal{N}} \ne \emptyset\right\} \ge \frac{1}{C} \sum_{i\in\overline{\mathcal{N}}} \sum_{u=1}^T \mathbf{1}\{i \in S_u\} = \frac{1}{C} \sum_{i\in\overline{\mathcal{N}}} T_i(T).$$

Combining the above with (12) we have that, asymptotically,

$$\mathcal{R}^\pi(T,\mu) \ge \delta(\mu) \frac{1}{C} \left(\sum_{i\in\overline{\mathcal{N}}} \frac{1}{H_i^\mu} \right) \log T + K_2(\mu),$$

for a finite positive constant K_2. Taking $K_1(\mu) := \delta(\mu) \min_{i \in \overline{\mathcal{N}}} \left\{ (H_i^\mu)^{-1} \right\}$ gives the desired result. ■

Proof of Theorem 2. We prove the result in 3 steps. First, we compute an upper bound on the probability of the estimates deviating from the true mean utilities. Second, we address the quality of the solution to the static problem, when using estimated mean utilities. Finally, we combine the above and analyze the regret. For purposes of this proof, let $\mathbb{P}$ denote probability of random variables when the assortment policy π_1 is used, and the mean utilities are given by the vector μ. With a slight abuse of notation define $p_i := \{p_i(A_j, \mu) : A_j \in \mathcal{A} \text{ s.t. } i \in A_j\}$, for $i \in \mathcal{N}$, and $p := (p_1, \ldots, p_{\mathcal{N}})$.

Step1. Define $T^j(t)$ to be the number of customers A_j has been offered to, up to customer $t-1$, for $A_j \in \mathcal{A}$. That is,

$$T^j(t) = \sum_{u=1}^{t-1} \mathbf{1}\left\{S_u = A_j\right\} \ , \ j = 1, \ldots, |\mathcal{A}| \, .$$

We will need the following side lemma, whose proof is deferred to Appendix B.

Lemma 1. *Fix $j \leq |\mathcal{A}|$ and $i \in A_j$. Then, for any $n \geq 1$ and $\epsilon > 0$*

$$\mathbb{P}\left\{ \left| \sum_{u=1}^{t-1} \left(X_i^u - p_i(A_j, \mu)\right) \mathbf{1}\left\{S_u = A_j\right\} \right| \geq \epsilon T^j(t) \, , \, T^j(t) \geq n \right\} \leq 2 \exp(-c(\epsilon) n),$$

for a positive constant $c(\epsilon) < \infty$.

For any vector $\nu \in \mathbb{R}_+^N$ and set $A \subseteq \mathcal{N}$ define $\|\nu\|_A = \max\{\nu_i : i \in A\}$. Consider $\epsilon > 0$ and fix $t \geq 1$. By Assumption 1 we have that for any assortment $A_j \subseteq \mathcal{A}$

$$\|\mu - \hat{\mu}_t\|_{A_j} \leq \kappa(\epsilon) \|p - \hat{p}_t\|_{A_j}, \tag{13}$$

for some constant $1 < \kappa(\epsilon) < \infty$, whenever $\|p - \hat{p}_t\|_{A_j} < \epsilon$. We have that, for $n \geq 1$,

$$\begin{aligned}
\mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon\,,\, T^j(t) \geq n\right\} &= \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon\,,\, \|p - \hat{p}_t\|_{A_j} \geq \epsilon\,,\, T^j(t) \geq n\right\} + \\
&\quad \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon\,,\, \|p - \hat{p}_t\|_{A_j} < \epsilon\,,\, T^j(t) \geq n\right\} \\
&\leq \mathbb{P}\left\{\|p - \hat{p}_t\|_{A_j} \geq \epsilon\,,\, T^j(t) \geq n\right\} + \\
&\quad \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon\,,\, \|p - \hat{p}_t\|_{A_j} < \epsilon\,,\, T^j(t) \geq n\right\} \\
&\overset{(a)}{\leq} \mathbb{P}\left\{\|p - \hat{p}_t\|_{A_j} \geq \epsilon\,,\, T^j(t) \geq n\right\} + \mathbb{P}\left\{\|p - \hat{p}_t\|_{A_j} > \epsilon/\kappa(\epsilon)\,,\, T^j(t) \geq n\right\} \\
&\leq 2\mathbb{P}\left\{\|p - \hat{p}_t\|_{A_j} \geq \epsilon/\kappa(\epsilon)\,,\, T^j(t) \geq n\right\} \\
&\leq 2\sum_{i \in A_j} \mathbb{P}\left\{|p_i(A_j, \mu) - \hat{p}_{i,t}| \geq \epsilon/\kappa(\epsilon)\,,\, T^j(t) \geq n\right\} \\
&\overset{(b)}{=} 2\sum_{i \in A_j} \mathbb{P}\left\{\left|\sum_{s=1}^{t} (X_i^s - p_i(A_j, \mu))\,\mathbf{1}\{S_t = A_j\}\right| \geq T^j(t)\epsilon/\kappa(\epsilon)\,,\, T^j(t) \geq n\right\} \\
&\overset{(c)}{\leq} 2\,|A_j| \exp(-c(\epsilon/\kappa(\epsilon))n), \qquad (14)
\end{aligned}$$

where (a) follows from (13), (b) follows from the definition of $\hat{p}_{i,t}$, and (c) follows from Lemma 1.

Step 2. Fix an assortment $S \in \mathcal{S}$. By the Lipschitz-continuity of $p(S, \cdot)$ we have that, for $t \geq 1$,

$$\max\{|p_i(S, \mu) - p_i(S, \hat{\mu}_t)| \,:\, i \in S\} \leq K\|\mu - \hat{\mu}_t\|_S,$$

for a positive constant $K < \infty$, and therefore

$$|f(S, \mu) - f(S, \hat{\mu}_t)| \leq \|w\|_\infty\, K\, C\|\mu - \hat{\mu}_t\|_S. \qquad (15)$$

From here, we conclude that

$$\begin{aligned}
f(S^*(\hat{\mu}_t), \mu) &\geq f(S^*(\hat{\mu}_t), \hat{\mu}_t) - \|w\|_\infty\, K\, C\|\mu - \hat{\mu}_t\|_{S^*(\hat{\mu}_t)} \\
&\geq f(S^*(\mu), \hat{\mu}_t) - \|w\|_\infty\, K\, C\|\mu - \hat{\mu}_t\|_{S^*(\hat{\mu}_t)} \\
&\geq f(S^*(\mu), \mu) - 2\|w\|_\infty\, K\, C\|\mu - \hat{\mu}_t\|_{(S^*(\mu) \cup S^*(\hat{\mu}_t))}.
\end{aligned}$$

As a consequence, if

$$\|\mu - \hat{\mu}_t\|_{(S^*(\mu) \cup S^*(\hat{\mu}_t))} < (2\|w\|_\infty\, K\, C)^{-1}\delta(\mu) f(S^*(\mu), \mu)$$

then $S^*(\mu) = S^*(\hat{\mu}_t)$, where $\delta(\mu)$ is the minimum (relative) optimality gap (see (9) in proof of Theorem 1). This means that if the mean utility estimates are uniformly close to the underlying mean utility values, then solving the static problem using estimates returns the same optimal assortment as when solving the static problem with the true parameters. In particular we will use the following relation:

$$\{S^*(\mu) \neq S^*(\hat{\mu}_t)\} \subseteq \left\{\|\mu - \hat{\mu}_t\|_{(S^*(\mu) \cup S^*(\hat{\mu}_t))} \geq (2\|w\|_\infty\, K\, C)^{-1}\delta(\mu) f(S^*(\mu), \mu)\right\}. \qquad (16)$$

Step 3. Let $NO(t)$ denote the event that a non-optimal assortment is offered to customer t. That is

$$NO(t) := \{S_t \neq S^*(\mu)\},$$

Define $\xi := (2\|w\|_\infty K\, C)^{-1}\delta(\mu) f(S^*(\mu), \mu)$. For $t \geq |\mathcal{A}|\, \lceil \mathcal{C} \log T \rceil$ one has that

$$\begin{aligned}
\mathbb{P}\{NO(t)\} &\overset{(a)}{\leq} \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{(S^*(\mu)\cup S^*(\hat{\mu}_t))} \geq \xi\right\} \\
&\leq \sum_{A_j \in \mathcal{A}} \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} \geq \xi\right\} \\
&= \sum_{A_j \in \mathcal{A}} \mathbb{P}\left\{\|\mu - \hat{\mu}_t\|_{A_j} \geq \xi\,,\, T^j(t) \geq \mathcal{C} \log T\right\} \\
&\overset{(b)}{\leq} \sum_{A_j \in \mathcal{A}} 2\,|A_j|\, T^{-\mathcal{C} c(\xi/\kappa(\xi))}, \qquad (17)
\end{aligned}$$

were (a) follows from (16) and (b) follows from (14). Considering $\mathcal{C} > c(\xi/\kappa(\xi))^{-1}$ results in the following bound for the regret:

$$\begin{aligned}
\mathcal{R}^\pi(T, \mu) &\leq \sum_{t=1}^{T} \mathbb{P}\{NO(t)\} \\
&\leq |\mathcal{A}|\, \lceil \mathcal{C} \log T \rceil + \sum_{t > |\mathcal{A}| \lceil \mathcal{C} \log T \rceil}^{\infty} \sum_{A_j \in \mathcal{A}} 2\,|A_j|\, T^{-\mathcal{C} c(\xi/\kappa(\xi))} \\
&\leq |\mathcal{A}|\, \mathcal{C} \log T + C_1 \\
&= \lceil N/C \rceil \mathcal{C} \log T + C_1,
\end{aligned}$$

for a finite constant C_1. Setting $C_2 = c(\xi/\kappa(\xi))^{-1}$ gives the desired result. ■

Proof of Theorem 3. The proof follows the arguments of the proof of Theorem 2. Steps 1 and 2 are identical.

Step 3. Let $NO(t)$ denote the event that a non-optimal assortment is offered to customer t, and $G(t)$ the event that there is no forced testing for customer t. That is,

$$\begin{aligned}
NO(t) &:= \{S_t \neq S^*(\mu)\}, \\
G(t) &:= \left\{T^j(t) \geq \mathcal{K} \log t\,,\, j \leq |\mathcal{A}| \text{ such that } \|w\|_{A_j} \geq f(S^*(\hat{\mu}_t), \hat{\mu}_t)\right\}. \qquad (18)
\end{aligned}$$

Define $\xi := (2\|w\|_\infty K\, C)^{-1}\delta(\mu) f(S^*(\mu),\mu)$. We have that

$$\begin{aligned}
\mathbb{P}\left\{NO(t)\,,\, G(t)\right\} &\overset{(a)}{\leq} \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{(S^*(\mu)\cup S^*(\hat{\mu}_t))} > \xi\,,\, G(t)\right\} \\
&\leq \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{S^*(\mu)} > \xi\,,\, G(t)\right\} + \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{S^*(\hat{\mu}_t)} > \xi\,,\, G(t)\right\} \\
&\overset{(b)}{\leq} \sum_{j:A_j\cap S^*(\hat{\mu}_t)\neq\emptyset} \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_j} > \xi\,,\, T^j(t) > \mathcal{K}\log t\right\} + \\
&\qquad \sum_{j\,:\,A_j\cap\in S^*(\mu)\neq\emptyset} \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_j} > \xi\,,\, G(t)\right\} \\
&\overset{(c)}{\leq} \sum_{j:A_j\cap S^*(\hat{\mu}_t)\neq\emptyset} 2\,|A_j|\, t^{-c(\xi/\kappa(\xi))\mathcal{K}} + \sum_{j\,:\,A_j\cap\in S^*(\mu)\neq\emptyset} \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_j} > \xi\,,\, G(t)\right\},
\end{aligned}$$

where: (a) follows from (16); (b) follows from the fact that Assumption 2 guarantees $w_i \geq f(S^*(\nu),\nu)$ for all $i \in S^*(\nu)$ for any vector $\nu \in \mathbb{R}^N$; and (c) follows from (14).

Fix j such that $A_j \cap S^*(\mu) \neq \emptyset$. For such an assortment we have that

$$\mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_j} > \xi\,,\, G(t)\right\} \;\leq\; \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_j} > \xi\,,\, T^j(t) \geq \mathcal{K}\log t\,,\, G(t)\right\} + \mathbb{P}\left\{T^j(t) < \mathcal{K}\log t\,,\, G(t)\right\}.$$

The first term on the right-hand-side above can be bounded using (14). For the second one, note that $\left\{T^j(t) < \mathcal{K}\log t\,,\, G(t)\right\} \subseteq \left\{\|w\|_{A_j} < f(S^*(\hat{\mu}_t),\hat{\mu}_t)\right\}$, and that

$$\begin{aligned}
\|w\|_{A_j} - f(S^*(\mu),\mu)\delta(\mu)/2 &\overset{(a)}{\geq} f(S^*(\mu),\mu)(1-\delta(\mu)/2) \\
&\overset{(b)}{\geq} f(S^*(\hat{\mu}_t),\mu) \\
&\overset{(c)}{\geq} f(S^*(\hat{\mu}_t),\hat{\mu}_t) - \|w\|_\infty KC\|\mu-\hat{\mu}_t\|_{S^*(\hat{\mu}_t)},
\end{aligned}$$

where: (a) follows from Assumption 2; (b) follows from the definition of $\delta(\mu)$; and (c) follows from (15). The above implies that $\left\{\|w\|_{A_j} < f(S^*(\hat{\mu}_t),\hat{\mu}_t)\right\} \subseteq \left\{\|\mu-\hat{\mu}_t\|_{S^*(\hat{\mu}_t)} > \xi\right\}$, i.e.,

$$\begin{aligned}
\mathbb{P}\left\{T^j(t) < \mathcal{K}\log t\,,\, G(t)\right\} &\leq \mathbb{P}\left\{\|w\|_{A_j} < f(S^*(\hat{\mu}_t),\hat{\mu}_t)\right\} \\
&\leq \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{S^*(\hat{\mu}_t)} > \xi\,,\, G(t)\right\} \\
&\leq \sum_{k\,:\,A_k\cap S^*(\hat{\mu}_t)\neq\emptyset} \mathbb{P}\left\{\|\mu-\hat{\mu}_t\|_{A_k} > \xi\,,\, G(t)\right\} \\
&\leq \sum_{k\,:\,A_k\cap S^*(\hat{\mu}_t)\neq\emptyset} 2\,|A_k|\, t^{-c(\xi/\kappa(\xi))\mathcal{K}},
\end{aligned}$$

where the last step follows from (14). Using the above we have that

$$\begin{aligned}
\mathbb{P}\left\{NO(t)\,,\, G(t)\right\} &\leq \sum_{j:A_j\cap S^*(\hat{\mu}_t)\neq\emptyset} 2\,|A_j|\, t^{-c(\xi/\kappa(\xi))\mathcal{K}} + \\
&\qquad \sum_{j\,:\,A_j\cap\in S^*(\mu)\neq\emptyset} \left(2\,|A_j|\, t^{-c(\xi/\kappa(\xi))\mathcal{K}} + \sum_{k\,:\,A_k\cap S^*(\hat{\mu}_t)\neq\emptyset} 2\,|A_k|\, t^{-c(\xi/\kappa(\xi))\mathcal{K}}\right) \\
&\leq 2C^2(2+C)t^{-c(\xi/\kappa(\xi))\mathcal{K}}. \qquad (19)
\end{aligned}$$

Consequently, we have that

$$\begin{aligned}\mathbb{P}\left\{NO(t)\,,\,G(t)^c\right\} &\leq \sum_{A_j\in\mathcal{A}} \mathbb{P}\left\{S_t = A_j\,,\,G(t)^c\right\} \\ &= \sum_{j\,:\,\|w\|_{A_j}\geq f(S^*(\mu),\mu)} \mathbb{P}\left\{S_t = A_j\,,\,G(t)^c\right\} + \sum_{j\,:\,\|w\|_{A_j}< f(S^*(\mu),\mu)} \mathbb{P}\left\{S_t = A_j\,,\,G(t)^c\right\}.\end{aligned}$$

For the first term above, we have from the policy specification that

$$\sum_{u=1}^{T} \sum_{j\,:\,\|w\|_{A_j}\geq f(S^*(\mu),\mu)} \mathbb{P}\left\{S_u = A_j\,,\,G(u)^c\right\} \leq \lceil \overline{\mathcal{N}}/C\rceil\,(\mathcal{K}\log T + 1)\,. \tag{20}$$

To analyze the second term, fix j such that $\|w\|_{A_j} < f(S^*(\mu),\mu)$, and define $L(t)$ as the last customer (previous to customer t) to whom the empirical optimal assortment (according to estimated mean utilities) was offered. That is

$$L(t) := \sup\left\{u \leq t-1\,:\,G(u)\right\},$$

with $G(u)$ given in (18). Note that $L(t) \in \{t - \lfloor |\mathcal{A}|\,\mathcal{K}\log t\rfloor, \ldots, t-1\}$ for $t \geq \tau$, where τ is given by

$$\tau := \inf\left\{u \geq 1\,:\,\log(u - \lfloor |A|\,\mathcal{K}\log u\rfloor) + \mathcal{K}^{-1} > \log u\right\}.$$

Consider $t \geq \tau$ and $u \in \{t - \lfloor |\mathcal{A}|\,\mathcal{K}\log t\rfloor, \ldots, t-1\}$. Then

$$\begin{aligned}\mathbb{P}\left\{S_t = A_j\,,\,G(t)^c\,,\,L(t)=u\right\} &\leq \mathbb{P}\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\,,\,G(t)^c\,,\,L(t)=u\right\} \\ &\leq \mathbb{P}\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\,,\,G(t)^c\,,\,G(u)\right\} \\ &= \mathbb{P}\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\,,\,G(t)^c\,,\,G(u),NO(u)\right\} + \\ &\quad \mathbb{P}\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\,,\,G(t)^c\,,\,G(u),NO(u)^c\right\} \\ &\leq \mathbb{P}\left\{NO(u)\,,\,G(u)\right\} + \\ &\quad \mathbb{P}\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\,,\,T^k(t) \geq \mathcal{K}\log t\,,\,\forall\,k \text{ s.t. } A_k\cap S^*(\mu)\neq\emptyset\right\},\end{aligned}$$

where the last step follows from the fact that offering $S^*(\mu)$ to customer u implies (from $G(u)$) that $T^j(u) \geq \mathcal{K}\log u$, and therefore (for $t \geq \tau$) that $T^j(t) \geq \mathcal{K}\log t$, for all j such that $A_j \cap S^*(\mu) \neq \emptyset$. The first term in the last inequality can be bounded using (19). For the second, observe that

$$f(S^*(\mu),\hat{\mu}_t) - \|w\|_{A_j} \geq f(S^*(\mu),\mu) - \|w\|_\infty KC\|\mu - \hat{\mu}_t\|_{S^*(\mu)} - \|w\|_{A_j},$$

which follows from (15). Define $\delta := \inf\left\{(\|w\|_\infty KC)^{-1}\left(1 - \|w\|_{A_j}/f(S^*(\mu),\mu)\right) > 0\,:\,A_j \in \mathcal{A}\right\}$. From the above,

$$\left\{\|w\|_{A_j} \geq f(S^*(\hat{\mu}_t),\hat{\mu}_t)\right\} \subseteq \left\{\|w\|_{A_j} \geq f(S^*(\mu),\hat{\mu}_t)\right\} \subseteq \left\{\|\mu - \hat{\mu}_t\|_{S^*(\mu)} > \delta f(S^*(\mu),\mu)\right\}.$$

Define the event $\Xi = \left\{ \|w\|_{A_j} \geq f(S^*(\hat{\mu}_t), \hat{\mu}_t)\,,\, T^k(t) \geq \mathcal{K} \log t\,,\, \forall\, k \text{ s.t. } A_k \cap S^*(\mu) \neq \emptyset \right\}$ and $\bar{\delta} := \delta f(S^*(\mu), \mu)$. One has that

$$\begin{aligned}
\mathbb{P}\{\Xi\} &\leq \mathbb{P}\left\{ \|\mu - \hat{\mu}_t\|_{S^*(\mu)} > \bar{\delta}\,,\, T^k(t) \geq \mathcal{K}\log t\,,\, \forall\, k \text{ s.t. } A_k \cap S^*(\mu) \neq \emptyset \right\} \\
&\leq \sum_{k\,:\,A_k \cap S^*(\mu) \neq \emptyset} \mathbb{P}\left\{ \|\mu - \hat{\mu}_t\|_{A_k} > \bar{\delta}\,,\, T^k(t) \geq \mathcal{K}\log t \right\} \\
&\leq \sum_{k\,:\,A_k \cap S^*(\mu) \neq \emptyset} 2\,|A_k|\, t^{-c(\bar{\delta}/\kappa(\bar{\delta}))\mathcal{K}}.
\end{aligned}$$

Using Lemma 2 , we have that, when $\mathcal{K} > c(\bar{\delta}/\kappa(\bar{\delta}))$,

$$\begin{aligned}
\mathbb{P}\{S_t = A_j\,,\, G(t)^c\,,\, L(t) = u\} &\leq C^2(2+C)u^{-c(\xi/\kappa(\xi))\mathcal{K}} + \sum_{k\,:\,A_k \cap S^*(\mu) \neq \emptyset} 2\,|A_k|\, t^{-c(\bar{\delta}/\kappa(\bar{\delta}))\mathcal{K}} \\
&\leq C^2(2+C)\left(t - \lfloor |\mathcal{A}|\,\mathcal{K}\log t \rfloor\right)^{-c(\xi/\kappa(\xi))\mathcal{K}} + C^2 t^{-c(\bar{\delta}/\kappa(\bar{\delta}))\mathcal{K}}.
\end{aligned}$$

Since the right hand side above is independent of u, one has that

$$\mathbb{P}\{S_t = A_j\,,\, G(t)^c\} \leq C^2(2+C)\left(t - \lfloor |\mathcal{A}|\,\mathcal{K}\log t \rfloor\right)^{-c(\xi/\kappa(\xi))\mathcal{K}} + C^2 t^{-c(\bar{\delta}/\kappa(\bar{\delta}))\mathcal{K}}, \tag{21}$$

for j such that $\|w\|_{A_j} < f(S^*(\mu), \mu)$, and $t \geq \tau$. Considering $\mathcal{K} > \max\left\{ c(\xi/\kappa(\xi))^{-1}, c(\bar{\delta}/\kappa(\bar{\delta}))^{-1} \right\}$ results in the following bound for the regret:

$$\begin{aligned}
\mathcal{R}^\pi(T,\mu) &\leq \sum_{t=1}^{T} \mathbb{P}\{NO(t), G(t)\} + \sum_{t=1}^{T} \mathbb{P}\{NO(t)\,,\, G(t)^c\} \\
&\leq \sum_{t=1}^{T} \mathbb{P}\{NO(t), G(t)\} + \\
&\quad \sum_{t=1}^{T} \sum_{j\,:\,\|w\|_{A_j} \geq f(S^*(\mu),\mu)} \mathbb{P}\{S_t = A_j\,,\, G(t)^c\} + \sum_{t=1}^{T} \sum_{j\,:\,\|w\|_{A_j} < f(S^*(\mu),\mu)} \mathbb{P}\{S_t = A_j\,,\, G(t)^c\} \\
&\overset{(a)}{\leq} \sum_{t=1}^{\infty} C^2(2+C)u^{-c(\xi/\kappa(\xi))\mathcal{K}} + \lceil \overline{\mathcal{N}}/C \rceil (\mathcal{K}\log(T) + 1) + \\
&\quad \sum_{t=1}^{\infty} \sum_{j\,:\,\|w\|_{A_j} < f(S^*(\mu),\mu)} C^2(2+C)\left(t - \lfloor |\mathcal{A}|\,\mathcal{K}\log t \rfloor\right)^{-c(\xi/\kappa(\xi))\mathcal{K}} + C^2 t^{-c(\bar{\delta}/\kappa(\bar{\delta}))\mathcal{K}} \\
&\overset{(b)}{\leq} \lceil \left| \overline{\mathcal{N}} \cup S^*(\mu) \right| / C \rceil \mathcal{K}\log T + K_1,
\end{aligned}$$

for a finite constant $K_1 < \infty$, where: (a) follows from (19), (20) and (21); and (b) uses the summability of the series, implied by the terms in (19) and (21). Taking $K_2 > (\max\left\{ c(\xi/\kappa(\xi))^{-1}, c(\bar{\delta}/\kappa(\bar{\delta}))^{-1} \right\}$ provides the desired result. ■

Proof of Corollary 1. Fix $i \in \underline{\mathcal{N}}$, and fix $j = \{k \leq |\mathcal{A}| \,:\, i \in A_k\}$. We have that

$$\begin{aligned}
\mathbb{E}_\pi[T_i(T)] &\leq \tau + \sum_{t=\tau+1}^{T} \mathbb{P}[NO(t)\,,\, G(t)] + \mathbb{P}[S_t = A_j\,,\, G(t)^c] \\
&\leq K_3,
\end{aligned}$$

for a finite constant K_3, where we have used the summability of the terms in (19) and (21). This complete the proof. ■

Proof of Theorem 4. The proof is an adaptation of the one for Theorem 3, customized for the MNL choice model. However, we provide a explanation version of each step with the objective of highlighting how the structure of the MNL model is exploited.

Step 1. We will need the following side lemma, whose proof is deferred to Appendix B.

Lemma 2. *Fix $i \in \mathcal{N}$. For any $n \geq 1$ and $\epsilon > 0$ one has*

$$\mathbb{P}\left\{\left|\sum_{u=1}^{t-1}\left(X_j^u - \mathbb{E}\left\{X_j^u\right\}\right)\mathbf{1}\left\{i \in S_u\right\}\right| \geq \epsilon\, T_i(t)\,,\, T_i(t) \geq n\right\} \leq 2\exp(-c(\epsilon)n),$$

for $j \in \{i, 0\}$ and a positive constant $c(\epsilon) < \infty$.

Consider $\epsilon > 0$ and fix $t \geq 1$ and $i \in \mathcal{N}$. Define $\varrho = 1/2\,(1+C\|w\|_\infty)^{-1}$. From Assumption 2 we have that $p_0(S,\mu) \geq 2\varrho$, for all $S \in \mathcal{S}$. For $n \geq 1$ define the event $\Xi := \{|\nu_i - \hat{\nu}_{i,t}| > \epsilon\,,\, T_i(t) \geq n\}$. We have that

$$\begin{aligned}
\mathbb{P}\{\Xi\} &= \mathbb{P}\left\{\left|\frac{\sum_{u=1}^{t-1} X_i^u \mathbf{1}\{i \in S_u\}}{\sum_{u=1}^{t-1} X_0^u \mathbf{1}\{i \in S_u\}} - \nu_i\right| > \epsilon\,,\, T_i(t) \geq n\right\} \\
&\leq \mathbb{P}\left\{\left|\frac{\sum_{u=1}^{t-1} X_i^u \mathbf{1}\{i \in S_u\}}{\sum_{u=1}^{t-1} X_0^u \mathbf{1}\{i \in S_u\}} - \nu_i\right| > \epsilon\,,\, \left|\sum_{u=1}^{t-1}\left(X_0^u - \mathbb{E}\{X_0^u\}\right)\mathbf{1}\{i \in S_u\}\right| < \varrho\, T_i(t)\,,\, T_i(t) \geq n\right\} + \\
&\quad \mathbb{P}\left\{\left|\sum_{u=1}^{t-1}\left(X_0^u - \mathbb{E}\{X_0^u\}\right)\mathbf{1}\{i \in S_u\}\right| \geq \varrho\, T_i(t)\,,\, T_i(t) \geq n\right\} \\
&\overset{(a)}{\leq} \mathbb{P}\left\{\left|\sum_{u=1}^{t-1}(X_i^u - X_0^u\nu_i)\mathbf{1}\{i \in S_u\}\right| > \epsilon\varrho\, T_i(t)\,,\, T_i(t) \geq n\right\} + 2\exp(-c(\varrho)n) \\
&\overset{(b)}{\leq} \mathbb{P}\left\{\left|\sum_{u=1}^{t-1}(X_i^u - E[X_i^u])\mathbf{1}\{i \in S_u\}\right| > \epsilon\varrho/2\, T_i(t)\,,\, T_i(t) \geq n\right\} + \\
&\quad \mathbb{P}\left\{\left|\sum_{u=1}^{t-1}(X_0^u - E[X_0^u])\mathbf{1}\{i \in S_u\}\right| > \epsilon\varrho/(2\nu_i)\, T_i(t)\,,\, T_i(t) \geq n\right\} + 2\exp(-c(\varrho)n) \\
&\leq 2\exp(-c(\epsilon\varrho/2)n) + 2\exp(-c(\epsilon\varrho/(2\nu_i))n) + 2\exp(-c(\varrho)n).
\end{aligned}$$

where: (a) follows from Lemma 2 and from the fact that

$$\left|\sum_{u=1}^{t-1} X_0^u \mathbf{1}\{i \in S_u\}\right| \geq \left|\sum_{u=1}^{t-1} E[X_0^u]\mathbf{1}\{i \in S_u\}\right| - \left|\sum_{u=1}^{t-1}(X_0^u - E[X_0^u])\mathbf{1}\{i \in S_u\}\right| \geq \varrho T_i(t),$$

when $\left|\sum_{u=1}^{t-1}\left(X_0^u - \mathbb{E}\{X_0^u\}\right)\mathbf{1}\{i \in S_u\}\right| < \varrho\, T_i(t)$; and (b) follows from the fact that $\mathbb{E}X_i^u = \nu_i \mathbb{E}X_0^u$, for all $u \geq 1$ such that $i \in S_u$, $i \in \mathcal{N}$. For $\epsilon > 0$ define $\tilde{c}(\epsilon) := \min\{c(\epsilon\varrho/2)\,,\, c(\epsilon\varrho/(2\|\nu\|_{\mathcal{N}}))\,,\, c(\varrho)\}$. From above we have that for $\epsilon > 0$

$$\mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \epsilon\,,\, T_i(t) \geq n\} \leq 6\exp(\tilde{c}(\epsilon)n), \tag{22}$$

for all $i \in \mathcal{N}$.

Step 2. Consider two vectors $\upsilon, \eta \in \mathbb{R}_+^N$. From (8), for any $S \in \mathcal{S}$ one has

$$\begin{aligned} \sum_{i \in S} \upsilon_i (w_i - f(S, \upsilon)) &= f(S, \upsilon) \\ \sum_{i \in S} \eta_i (w_i - f(S, \upsilon)) &\geq f(S, \upsilon) - C\|w\|_\infty \|\upsilon - \eta\|_S \\ \sum_{i \in S} \eta_i (w_i - (f(S, \upsilon) - C\|w\|_\infty \|\upsilon - \eta\|_S) &\geq f(S, \upsilon) - C\|w\|_\infty \|\upsilon - \eta\|_S \end{aligned}$$

This implies that

$$f(S, \eta) \geq f(S, \upsilon) - C\|w\|_\infty \|\eta - \upsilon\|_S. \tag{23}$$

From the above we conclude that

$$\{S^*(\hat{\nu}_t) \neq S^*(\hat{\nu}_t)\} \subseteq \left\{ \|\nu - \hat{\nu}_t\|_{S^*(\nu) \cup S^*(\hat{\nu}_t)} \geq (2\|w\|_\infty C)^{-1} \delta(\nu) f(S^*(\mu), \mu) \right\}, \tag{24}$$

were $\delta(\nu)$ refers to the minimum optimality gap, in terms of the adjusted terms $\exp(\mu)$.

Step 3. Let $NO(t)$ denote the event that a non-optimal assortment is offered to customer t, and $G(t)$ the event that there is no "forced testing" on customer t. That is

$$\begin{aligned} NO(t) &:= \{S_t \neq S^*(\nu)\}, \\ G(t) &:= \{T_i(t) \geq \mathcal{M} \log t \,, \, \forall \, i \in \mathcal{N} \text{ such that } w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\}. \end{aligned}$$

Define $\xi := (2\|w\|_\infty C)^{-1} \delta(\nu) f(S^*(\mu), \mu)$. We have that

$$\begin{aligned} \mathbb{P}\{NO(t), G(t)\} &\overset{(a)}{\leq} \mathbb{P}\left\{ \|\nu - \hat{\nu}_t\|_{(S^*(\nu) \cup S^*(\hat{\nu}_t))} > \xi, G(t) \right\} \\ &\leq \mathbb{P}\left\{ \|\nu - \hat{\nu}_t\|_{S^*(\hat{\nu}_t)} > \xi, G(t) \right\} + \mathbb{P}\left\{ \|\nu - \hat{\nu}_t\|_{S^*(\nu)} > \xi, G(t) \right\} \\ &\overset{(b)}{\leq} \sum_{i \in S^*(\hat{\nu}_t)} \mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \xi, T_i(t) \geq \mathcal{M} \log t\} + \sum_{i \in S^*(\nu)} \mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \xi, G(t)\} \\ &\overset{(c)}{\leq} 6Ct^{-\mathcal{M}\tilde{c}(\xi)} + \sum_{i \in S^*(\nu)} \mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \xi, G(t)\} \end{aligned}$$

where: (a) follows from (24); (b) follows from the fact that Assumption 2 guarantees $w_i \geq f(S^*(\eta), \eta)$ for all $i \in S^*(\eta)$ and for any vector $\eta \in \mathbb{R}^N$; and (c) follows from (22). Fix $i \in S^*(\nu)$. We have that

$$\mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \xi, G(t)\} \leq \mathbb{P}\{|\nu_i - \hat{\nu}_{i,t}| > \xi, T_i(t) \geq \mathcal{M} \log t\} + \mathbb{P}\{G(t), T_i(t) < \mathcal{M} \log t\}.$$

The first term above can be bounded using (22). For the second one, note that $\{G(t)\,,\, T_i(t) < \mathcal{M}\log t\} \subseteq \{w_i < f(S^*(\hat{\nu}_t), \hat{\nu}_t)\}$, and that

$$\begin{aligned} w_i - f(S^*(\nu),\nu)\delta(\nu)/2 &\overset{(a)}{\geq} f(S^*(\nu),\nu)(1-\delta(\nu)/2) \\ &\overset{(b)}{\geq} f(S^*(\hat{\nu}_t),\nu) \\ &\overset{(c)}{\geq} f(S^*(\hat{\nu}_t),\hat{\nu}_t) - \|w\|_\infty C\|\nu - \hat{\nu}_t\|_{S^*(\hat{\nu}_t)}, \end{aligned}$$

where (a) follows from Assumption 2, (b) follows from the definition of $\delta(\nu)$, and (c) follows from (23). The above implies that $\{w_i < f(S^*(\hat{\nu}_t), \hat{\nu}_t)\} \subseteq \left\{\|\nu - \hat{\nu}_t\|_{S^*(\hat{\nu}_t)} > \xi\right\}$, i.e.,

$$\begin{aligned} \mathbb{P}\left\{T_i(t) < \mathcal{M}\log t\,,\, G(t)\right\} &\leq \mathbb{P}\left\{w_i < f(S^*(\hat{\nu}_t),\hat{\nu}_t), G(t)\right\} \\ &\leq \mathbb{P}\left\{\|\nu - \hat{\nu}_t\|_{S^*(\hat{\nu}_t)} > \xi\,,\, G(t)\right\} \\ &\leq \sum_{j\in S^*(\hat{\nu}_t)} \mathbb{P}\left\{|\nu_j - \hat{\nu}_{j,t}| > \xi\,,\, G(t)\right\} \\ &\leq 6Ct^{-\mathcal{M}\tilde{c}(\xi)}, \end{aligned}$$

where the last step follows from (22). Using the above we have that

$$\mathbb{P}\left\{NO(t)\,,\, G(t)\right\} \leq 6C(1+C)t^{-\mathcal{M}\tilde{c}(\xi)}. \tag{25}$$

From here, we have that

$$\begin{aligned} \mathbb{P}\left\{NO(t)\,,\, G(t)^c\right\} &\leq \sum_{i\,:\,w_i < f^*(S^*(\nu),\nu)} \mathbb{P}\left\{i\in S_t\,,\, G(t)^c\right\} + \sum_{i\,:\,w_i \geq f^*(S^*(\nu),\nu)} \mathbb{P}\left\{i\in S_t\,,\, G(t)^c\right\} \\ &\overset{(a)}{\leq} \sum_{i\,:\,w_i < f^*(S^*(\nu),\nu)} \mathbb{P}\left\{i\in S_t\,,\, G(t)^c\right\} + \left|\overline{\mathcal{N}}\right|(\mathcal{M}\log T + 1)\,. + \\ &\quad \sum_{i\in S^*(\mu)} \mathbb{P}\left\{i\in S_t\,,\, G(t)^c\right\} \end{aligned}$$

where (a) follows from the specification of the policy. Fix i such that $w_i < f(S^*(\nu),\nu)$, and define $L(t)$ as the last customer (previous to customer t) to whom the empirical optimal assortment, according to estimated mean utilities, was offered. That is

$$L(t) := \sup\left\{u \leq t-1\,:\, G(u)\right\}.$$

Note that $L(t) \in \{t - \lfloor N\mathcal{M}\log t\rfloor, \ldots, t-1\}$ for $t \geq \tau$, where τ is given by

$$\tau := \inf\left\{u \geq 1\,:\, \log(u - \lfloor N\mathcal{M}\log u\rfloor) + \mathcal{M}^{-1} > \log u\right\}.$$

Consider $t \geq \tau$ and $u \in \{t - \lfloor N\mathcal{M}\log t\rfloor, \ldots, t-1\}$. Then

$$\begin{aligned}
\mathbb{P}\left\{i \in S_t\,,\, G(t)^c\,,\, L(t) = u\right\} &\leq \mathbb{P}\left\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\,,\, G(t)^c\,,\, L(t) = u\right\} \\
&\overset{(a)}{\leq} \mathbb{P}\left\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\,,\, G(t)^c\,,\, G(u)\right\} \\
&\leq \mathbb{P}\left\{G(u), NO(u)\right\} + \mathbb{P}\left\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\,,\, G(t)^c\,,\, G(u), NO(u)^c\right\} \\
&\overset{(b)}{\leq} 6C(1+C)u^{-\mathcal{M}\tilde{c}(\xi)} + \\
&\quad \mathbb{P}\left\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\,,\, T_j(t) \geq \mathcal{M}\log t\ \forall\, j \in S^*(\nu)\right\},
\end{aligned}$$

where (a) follows from $\{L(t) = u\} \subseteq \{G(u)\}$, and (b) from (25) and the fact that offering $S^*(\nu)$ to customer u implies (from $G(u)$) that $T_j(u) \geq \mathcal{M}\log u$ and therefore (from $t \geq \tau$) that $T_j(t) \geq \mathcal{M}\log t$, for all $j \in S^*(\nu)$. From (23) we have that

$$f(S^*(\nu), \hat{\nu}_t) - w_i \geq f(S^*(\nu), \nu) - \|w\|_\infty C\|\nu - \hat{\nu}_t\|_{S^*(\nu)} - w_i.$$

Define $\delta := \inf\left\{(\|w\|_\infty C)^{-1}\left(1 - w_i/f(S^*(\nu), \nu)\right) > 0 \,:\, i \in \mathcal{N}\right\}$. From the above, we have that

$$\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\} \subseteq \{w_i \geq f(S^*(\nu), \hat{\nu}_t)\} \subseteq \left\{\|\nu - \hat{\nu}_t\|_{S^*(\nu)} > \delta f(S^*(\nu), \nu)\right\}.$$

Define $\bar{\delta} := \delta f(S^*(\nu), \nu)$. It follows that

$$\begin{aligned}
\mathbb{P}\left\{w_i \geq f(S^*(\hat{\nu}_t), \hat{\nu}_t)\,,\, T_j(t) \geq \mathcal{M}\log t\ \forall\, j \in S^*(\nu)\right\} &\leq \mathbb{P}\left\{\|\nu - \hat{\nu}_t\|_{S^*(\nu)} > \bar{\delta}\,,\, T_j(t) \geq \mathcal{M}\log t\ \forall\, j \in S^*(\nu)\right\} \\
&\leq \sum_{i \in S^*(\nu)} \mathbb{P}\left\{|\nu_i - \hat{\nu}_{t,i}| > \bar{\delta}\,,\, T_i(t) \geq \mathcal{M}\log t\right\} \\
&\leq 6Ct^{-\mathcal{M}\tilde{c}(\bar{\delta})}.
\end{aligned}$$

Using the above one gets that, when $\mathcal{M} > \tilde{c}(\xi)^{-1}$

$$\begin{aligned}
\mathbb{P}\left\{i \in S_t\,,\, G(t)^c\,,\, L(t) = u\right\} &\leq 6C(1+C)u^{-\mathcal{M}\tilde{c}(\xi)} + 6Ct^{-\mathcal{M}\tilde{c}(\bar{\delta})} \\
&\leq 6C(1+C)(t - \lfloor N\mathcal{M}\log t\rfloor)^{-\mathcal{M}\tilde{c}(\xi)} + 6Ct^{-\mathcal{M}\tilde{c}(\bar{\delta})}.
\end{aligned}$$

Since the right hand side above is independent of u, one has that

$$\mathbb{P}\left\{i \in S_t\,,\, G(t)^c\right\} \leq 6C(1+C)(t - \lfloor N\mathcal{M}\log t\rfloor)^{-\mathcal{M}\tilde{c}(\xi)} + 6Ct^{-\mathcal{M}\tilde{c}(\bar{\delta})}, \tag{26}$$

for all $i \in \mathcal{N}$ such that $w_i < f(S^*(\nu), \nu)$, and $t \geq \tau$.

Now fix $i \in S^*(\mu)$, and consider $t \geq \tau$, $u \in \{t - \lfloor N\mathcal{M}\log t\rfloor, \ldots, t-1\}$ and $\mathcal{M} > \tilde{c}(\xi)^{-1}$. Then

$$\begin{aligned}
\mathbb{P}\left\{i \in S_t\,,\, G(t)^c\,,\, L(t) = u\right\} &\leq \mathbb{P}\left\{T_i(t) < \mathcal{M}\log t\,,\, G(t)^c\,,\, L(t) = u\right\} \\
&\overset{(a)}{\leq} \mathbb{P}\left\{T_i(t) < \mathcal{M}\log t\,,\, G(u)\right\} \\
&\leq \mathbb{P}\left\{G(u), NO(u)\right\} + \mathbb{P}\left\{T_i(t) < \mathcal{M}\log t\,,\, G(u)\,,\, NO^c(u)\right\} \\
&\overset{(b)}{\leq} 6C(1+C)u^{-\mathcal{M}\tilde{c}(\xi)} \\
&\leq 6C(1+C)(t - \lfloor N\mathcal{M}\log t\rfloor)^{-\mathcal{M}\tilde{c}(\xi)},
\end{aligned}$$

where (a) follows from $\{L(t) = u\} \subseteq \{G(u)\}$, and (b) from (25) and the fact that offering $S^*(\nu)$ to customer u implies (from $G(u)$) that $T_i(u) \geq \mathcal{M} \log u$ and therefore (from $t \geq \tau$) that $T_i(t) \geq \mathcal{M} \log t$. Since the right hand side above is independent of u, one has that

$$\mathbb{P}\left\{i \in S_t\,,\, G(t)^c\right\} \leq 6C(1+C)(t - \lfloor N\mathcal{M} \log t \rfloor)^{-\mathcal{M}\tilde{c}(\xi)}, \tag{27}$$

for all $i \in S^*(\mu)$ and $t \geq \tau$.

Considering $\mathcal{M} > \max\left\{\tilde{c}(\xi)^{-1}\,,\, \tilde{c}(\bar{\delta})^{-1}\right\}$ results in the following bound for the regret

$$\begin{aligned}
\mathcal{R}^{\pi}(T,\nu) &\leq \sum_{t=1}^{T} \mathbb{P}\left\{NO(t), G(t)\right\} + \sum_{t=1}^{T} \mathbb{P}\left\{NO(t)\,,\, G(t)^c\right\} \\
&\overset{(a)}{\leq} 6C(1+C) \sum_{t=1}^{\infty} t^{-\mathcal{M}\tilde{c}(\xi)} + \left|\overline{\mathcal{N}}\right| \mathcal{M}(\log T + 1) + \tau + \\
&\quad 6C\left|\underline{\mathcal{N}} \cup S^*(\mu)\right| \sum_{t=\tau}^{\infty} (1+C)(t^{-\mathcal{M}\tilde{c}(\xi)} + (t - \lfloor N\mathcal{M}\log t \rfloor)^{-\mathcal{M}\tilde{c}(\xi)}) + t^{-\mathcal{M}\tilde{c}(\bar{\delta})} \\
&\overset{(b)}{\leq} \left|\overline{\mathcal{N}}\right| \mathcal{M} \log T + K_1,
\end{aligned}$$

for a finite constant $K_1 < \infty$, where (a) follows from (25), (26) and (27), and (b) uses the summability of the series, implied by the terms in (25), (26) and (27). Taking $K_2 > \max\left\{\tilde{c}(\xi)^{-1}\,,\, \tilde{c}(\bar{\delta})^{-1}\right\}$ provides the desired result. ■

Proof of Corollary 2. Fix $i \in \underline{\mathcal{N}}$. We have that

$$\begin{aligned}
\mathbb{E}_{\pi}[T_i(T)] &\leq \tau + \sum_{t=\tau+1}^{T} \mathbb{P}[NO(t)\,,\, G(t)] + \mathbb{P}[i \in S_t\,,\, G(t)^c] \\
&\leq K_3 < \infty,
\end{aligned}$$

for a finite constant K_3, where we have used the summability of the terms in (25) and (26). This concludes the proof. ■

B Proof of Auxiliary Results

Proof of Lemma 1.

Fix $i \in \mathcal{N}$. For $\theta > 0$ consider the process $\{M_t(\theta) : t \geq 1\}$, defined as

$$M_t(\theta) := \exp\left(\sum_{u=1}^{t} \mathbf{1}\left\{S_u = A_j\right\}\left[\theta(X_i^u - p_i(A_j,\mu)) - \phi(\theta)\right]\right),$$

where

$$\phi(\theta) := \log \mathbb{E}\left\{\exp(\theta\,(X_i^u - p_i(A_j,\mu)))\right\} = -\theta p_i(A_j,\mu) + \log(p_i(A_j,\mu)\exp(\theta) + 1 - p_i(A_j,\mu)),$$

and $A_j \in \mathcal{A}$ such that $i \in A_j$. One can check that $M_t(\theta)$ is an $\mathcal{F}_t$-martingale, for any $\theta > 0$ (see §3 for the definition of $\mathcal{F}_t$). Note that

$$\exp\left(\theta \sum_{u=1}^{t} \mathbf{1}\left\{S_u = A_j\right\}\left((X_i^u - p_i(A_j,\mu)) - \epsilon\right)\right) = \sqrt{M_t(2\theta)}\exp\left(\sum_{u=1}^{t} \mathbf{1}\left\{S_u = A_j\right\}\left(\phi(2\theta)/2 - \theta\epsilon\right)\right). \tag{28}$$

Let χ_i denote the event we are interested in. That is

$$\chi_i := \left\{\sum_{u=1}^{t-1}\left(X_i^u - p_i(A_j,\mu)\right)\mathbf{1}\left\{S_u = A_j\right\} \geq T^j(t)\epsilon\,,\, T^j(t) \geq n\right\}.$$

Let $\psi(t)$ denote the choice made by the t-th user. Using the above one has that

$$\begin{aligned}
\mathbb{P}\left\{\chi_i\right\} &\overset{(a)}{\leq} \mathbb{E}\left\{\exp\left(\theta\sum_{u=1}^{t-1}\mathbf{1}\left\{S_u = A_j\right\}\left(X_i^u - p_i(A_j,\mu) - \epsilon\right)\right)\,;\, T_i(t) \geq n\right\} \\
&\overset{(b)}{\leq} \left(\mathbb{E}\left\{M_{t-1}(2\theta)\right\}\mathbb{E}\left\{\exp\left(\sum_{u=1}^{t-1}\mathbf{1}\left\{\psi(u) = i\right\}\left(\phi(2\theta) - 2\theta\epsilon\right)\right)\,;\, T_i(t) \geq n\right\}\right)^{1/2} \\
&\overset{(c)}{\leq} \left(\mathbb{E}\left\{\exp\left(\sum_{u=1}^{t-1}\mathbf{1}\left\{\psi(u) = i\right\}\left(\phi(2\theta) - 2\theta\epsilon\right)\right)\,;\, T_i(t) \geq n\right\}\right)^{1/2},
\end{aligned}$$

where: (a) follows from Chernoff's inequality; (b) follows from the Cauchy-Schwartz inequality and (28); and (c) follows from the properties of $M_t(\theta)$. Note that when $\epsilon < (1 - p_i(A_j,\mu))$ minimizing $\phi(\theta) - \theta\epsilon$ over $\theta > 0$ results on

$$\theta^* := \log\left(1 + \frac{\epsilon}{p_i(A_j,\mu)(1 - p_i(A_j,\mu) - \epsilon)}\right) > 0,$$

with

$$c(\epsilon) := \phi(2\theta^*)/2 - \theta^*\epsilon < 0.$$

Using this we have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{u=1}^{t-1}\left(X_i^u - p_i\right)\mathbf{1}\left\{S_u = A_j\right\} \geq T^j(t)\epsilon\,,\, T^j(t) \geq n\right\} &\leq \sqrt{\mathbb{E}\left\{\exp(-2c(\epsilon)T_i(t));\, T_i(t) \geq n\right\}} \\
&\leq \exp(-c(\epsilon)n).
\end{aligned}$$

Using the same arguments one has that

$$\mathbb{P}\left\{\sum_{u=1}^{t-1}\left(X_i^u - p_i\right)\mathbf{1}\left\{S_u = A_j\right\} \leq -T^j(t)\epsilon\,,\, T^j(t) \geq n\right\} \leq \exp(-c(\epsilon)n).$$

The result follows from the union bound. ■

Proof of Lemma 2

The proof follows almost verbatim the steps in the proof of Lemma 1. Fix $i \in \mathcal{N}$. For $\theta > 0$ consider the process $\left\{ M_t^j(\theta) \,:\, t \geq 1 \right\}$, defined as

$$M_t^j(\theta) := \exp\left(\sum_{u=1}^{t} \mathbf{1}\left\{ i \in S_u \right\} \left[\theta(X_j^u - p_j(S_u, \mu)) - \phi_u^j(\theta) \right] \right) \qquad j \in \{i, 0\}\,,$$

where

$$\phi_u^j(\theta) := \log \mathbb{E}\left\{ \exp(\theta(X_j^u - p_j(S_u, \mu))) \right\} = \log \mathbb{E}\left\{ \exp(-\theta p_j(S_u, \mu)) \left(\exp(\theta) p_j(S_u, \mu) + 1 - p_j(S_u, \mu) \right) \right\}.$$

One can verify that $M_t^j(\theta)$ is an $\mathcal{F}_t$-martingale, for any $\theta > 0$ and $j \in \{i, 0\}$ (see §3 for the definition of $\mathcal{F}_t$). Fix $j \in \{i, 0\}$ and note that

$$\exp\left(\theta \sum_{u=1}^{t} \mathbf{1}\left\{ i \in S_u \right\} \left((X_j^u - p_j(S_u, \mu)) - \epsilon \right) \right) = \sqrt{M_t^j(2\theta)} \exp\left(\sum_{u=1}^{t} \mathbf{1}\left\{ i \in S_u \right\} \left(\phi_u^j(2\theta)/2 - \theta\epsilon \right) \right). \tag{29}$$

Put

$$\chi_j := \left\{ \sum_{u=1}^{t-1} \left(X_j^u - p_j(S_u, \mu) \right) \mathbf{1}\left\{ i \in S_u \right\} \geq T_i(t)\epsilon\,,\, T_i(t) \geq n \right\}.$$

Let $\psi(t)$ denote the choice made by the t-th customer. Using the above one has that

$$\begin{aligned}
\mathbb{P}\left\{ \chi_j \right\} &\overset{(a)}{\leq} \mathbb{E}\left\{ \exp\left(\theta \sum_{u=1}^{t-1} \mathbf{1}\left\{ i \in S_u \right\} \left(X_j^u - p_j(S_u, \mu) - \epsilon \right) \right) ;\, T_i(t) \geq n \right\} \\
&\overset{(b)}{\leq} \left(\mathbb{E}\left\{ M_{t-1}^j(2\theta) \right\} \mathbb{E}\left\{ \exp\left(\sum_{u=1}^{t-1} \mathbf{1}\left\{ \psi(u) = j\,,\, i \in S_u \right\} \left(\phi_u^j(2\theta) - 2\theta\epsilon \right) \right) ;\, T_i(t) \geq n \right\} \right)^{1/2} \\
&\overset{(c)}{\leq} \left(\mathbb{E}\left\{ \exp\left(\sum_{u=1}^{t-1} \mathbf{1}\left\{ \psi(u) = j\,,\, i \in S_u \right\} \left(\phi_u^j(2\theta) - 2\theta\epsilon \right) \right) ;\, T_i(t) \geq n \right\} \right)^{1/2},
\end{aligned}$$

where; (a) follows from Chernoff's inequality; (b) follows from the Cauchy-Schwartz inequality and (28); and (c) follows from the properties of $M_t^j(\theta)$. Note that $\phi_s^j(\cdot)$ is continuous, $\phi_s^j(0) = 0$, $(\phi_s^j)'(0) = 0$, and $\phi_s^j(\theta) \to \infty$ when $\theta \to \infty$, for all $s \geq 1$. This implies that there exists a positive constant $c(\epsilon) < \infty$ (independent of n), and a $\theta^* > 0$, such that $\phi_s^j(2\theta^*) - 2\theta^*\epsilon < -2c(\epsilon)$ for all $s \geq 1$. Using this we have that

$$\begin{aligned}
\mathbb{P}\left\{ \sum_{u=1}^{t-1} \left(X_j^u - p_j(S_u, \mu) \right) \mathbf{1}\left\{ i \in S_u \right\} \geq T_i(t)\epsilon\,,\, T_i(t) \geq n \right\} &\leq \sqrt{\mathbb{E}\left\{ \exp(-2c(\epsilon)T_i(t));\, T_i(t) \geq n \right\}} \\
&\leq \exp(-c(\epsilon)n).
\end{aligned}$$

Using the same arguments one has that

$$\mathbb{P}\left\{ \sum_{u=1}^{t-1} \left(X_j^u - p_j(S_u, \mu) \right) \mathbf{1}\left\{ i \in S_u \right\} \leq -T_i(t)\epsilon\,,\, T_i(t) \geq n \right\} \leq \exp(-c(\epsilon)n).$$

The result follows from the union bound. ■