Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Dynamic Assortment Planning with Demand Learning

Denis Sauré, Assaf Zeevi,

To cite this article:
Denis Sauré, Assaf Zeevi, (2013) Optimal Dynamic Assortment Planning with Demand Learning. Manufacturing & Service Operations Management 15(3):387-404. http://dx.doi.org/10.1287/msom.2013.0429

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 3, Summer 2013, pp. 387–404
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2013.0429

Optimal Dynamic Assortment Planning with Demand Learning

Denis Sauré
Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, dsaure@pitt.edu

Assaf Zeevi
Graduate School of Business, Columbia University, New York, New York 10027, assaf@gsb.columbia.edu

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 trade-off 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.

Key words: assortment planning; online algorithm; demand learning
History: Received: November 25, 2009; accepted: November 11, 2012. Published online in *Articles in Advance* April 11, 2013.

1. Introduction

1.1. 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, especially for companies that 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. That is the essence of the assortment planning problem; see Kök et al. (2008) for an overview. Our interest lies in *dynamic* instances of the problem, where assortment planning decisions can be revisited frequently, and consumer preferences for products are not known a priori and need to be learned over the course of the selling horizon. These instances will be referred to as *dynamic assortment planning* problems. 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 Company, 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 these retailer's 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: *Online advertising.* This emerging area of business is the single most important source of revenues for thousands of websites. Giants such as Yahoo! and Google depend almost completely on online advertisement to subsist. One of the most prevalent business models here builds on the cost-per-click statistic: Advertisers pay the website (a "publisher") only when a user clicks on their advertisements (henceforth, 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 revenue within the life span of the contract with the advertiser.

The above motivating applications share common features: (i) a priori information on consumer purchase/click behavior is scarce or nonexistent; (ii) products/ads can be substituted one for the other, but may differ in the profit they generate, and demand for individual product/ad is affected by the assortment decision, which is subject to display constraints; (iii) assortment decisions can be done in a dynamic fashion.

The purpose of this paper is to study a stylized version of the dynamic assortment planning problem that incorporates these salient features. Central to this study is the trade-off between information collection (exploration), which leads to a clearer picture of demand, and revenue maximization (exploitation), that strives to make optimal assortment decisions at each point in time. In this context, the longer a retailer 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 the revenue maximization objective.

To isolate the role assortment planning plays in balancing information collection and revenue maximization, our stylized model ignores a variety of operational considerations, such as pricing decisions, inventory replenishment, assortment sequencing and switching costs, availability of users' profile information, etc; further discussion of these aspects can be found in §7. The main salient feature that we build into our stylized model is limited display capacity, as such a constraint is a defining feature of assortment planning problems (see Fisher and Vaidyanathan 2009 for a discussion), and our work will elucidate the manner in which it impacts the complexity of the *dynamic* assortment problem.

Although our focus is on a revenue management objective via assortment decisions, we assume that product prices are fixed throughout the selling season. Such an assumption is common in the assortment planning literature and facilitates analysis. We note in passing that dynamic pricing has been studied as a stand-alone mechanism in the context of choice-driven demand with limited prior information (see, e.g., Rusmevichientong and Broder 2012), but incorporating a pricing dimension into our formulation would obscure insights regarding the role of assortment experimentation in demand inference.

As stated above, our main focus is on learning consumer behavior via suitable assortment experimentation, and doing this in a manner that guarantees revenue maximization over the selling horizon. For that purpose, we consider a population of utility maximizing customers: each customer assigns a (random) utility to each offered product, and purchases the product that maximizes his or her utility. The retailer needs to devise an assortment policy to maximize revenues over the relevant time horizon by properly adapting the assortment offered based on observed customer purchase decisions and subject to capacity constraints that limit the size of the assortment.

1.2. Key Insights and Qualitative Results

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 nonanticipating policies). Performance of such policies will be measured in terms of the expected revenue loss relative to an oracle that knows the product utility distributions in advance, that 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 nonanticipating assortment policy.

The main findings of this paper are summarized below:

(i) We establish fundamental bounds on the performance of any "good" policy (we formalize this in §4). Specifically, we identify the magnitude of loss relative to the oracle performance that *any* policy must incur and characterize 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 §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 in finite time (i.e., independent of the length of the selling horizon) with high probability; see Theorems 1 and 3. We show that our proposed policy successfully limits the extent at which such products are offered (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 reduce the complexity of the assortment problem.

The above results establish that an oracle with advance knowledge of customer behavior gains only a relatively modest additional revenues relative to policies that do not have such prior knowledge. To ensure this modest gap, the policies in question must adhere to a critical rate of assortment experimentation. An interesting feature of these policies is that they can limit exploration on a certain subset of products (in particular, these products need only be offered to a bounded number of customers *independent* of the time horizon). This result differs markedly from most of the literature on sequential decision-making problems under uncertainty; see further discussion in §2.

1.3. Remainder of This Paper

Section 2 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 §5.3 customizes our proposed algorithm for the most widely used customer choice model, namely, the logit. Section 6 presents a comparison with benchmark results. Finally, §7 presents our concluding remarks. Proofs are relegated to Appendix A and to the online companion (available as supplemental material at http://dx.doi.org/10.1287/msom.2013.0429). Appendix B contains further details pertaining to some estimation methods used in this paper.

2. Literature Review

2.1. Static Assortment Planning

The literature here 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 Kök et al. (2008) for a review of the state of the art in static assortment optimization. Van Ryzin and Mahajan (1999) formulate the assortment planning problem using a multinomial logit (MNL) model of consumer choice. Assuming that customers do not look for a substitute if their choice is stocked out, they prove that the optimal assortment is always in the "popular assortment set" and establish structural properties of the optimal assortment and ordering quantities. In the same setting, Gaur and Honhon (2006) use the locational choice model and characterize properties of the optimal assortment. In a recent paper, Goyal et al. (2009) prove that the static assortment problem is NP-hard when customers look for a substitute if their choice is stocked out, and propose a near-optimal heuristic for a particular choice model; see Mahajan and van Ryzin (2001), Honhon et al. (2009), and Hopp and Xu (2008) for formulations considering stockout-based substitution.

Our formulation assumes perfect inventory replenishment (thus eliminating stockout-based substitution) while considering limited display capacity. Fisher and Vaidyanathan (2009) studies assortment planning under display constraints and highlights how these arise in practice. Although the single-sale profit maximization problem remains NP-hard under the perfect replenishment assumption, Rusmevichientong et al. (2010) presents a polynomial-time algorithm for the single-sale profit maximization problem when consumer preferences are represented using particular choice models; hence, at least in certain instances the single-sale problem can be solved efficiently.

2.2. Dynamic Assortment Planning

This problem setting allows revisiting assortment decisions at each point in time as more information is collected about initially unknown demand/consumer preferences. Caro and Gallien (2007), to the best of our knowledge, 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 independent of demand and availability of other products, the rate of demand is constant throughout the selling season, and perfect inventory replenishment is assumed. Taking a Bayesian approach to demand learning, the problem is studied using a dynamic programming formulation: Caro and Gallien (2007) derive bounds on the value function and propose an index-based policy that is shown to be *near* optimal when there is some prior information on demand. Closer to our formulation is the work by Rusmevichientong et al. (2010). There, utility maximizing customers make purchase decisions according to the MNL choice model (a special case of the more general setting treated here), and an adaptive algorithm for joint parameter estimation and assortment optimization is developed; see further discussion below. A different formulation is advanced by Honhon et al. (2012), who study a dynamic assortment problem using the locational choice model.

2.3. Related Work in Dynamic Optimization with Limited Demand Information

Uncertainty at demand-model level has been considered previously in revenue management settings, in the context of dynamic pricing. Araman and Caldentey (2009) and Farias and Van Roy (2010), for example, present dynamic programming formulations with Bayesian updating of initially unknown parameters; see also Lim and Shanthikumar (2007). Closer to the current paper is the work by Besbes and Zeevi (2009) that considers the dynamic pricing formulation in Gallego and van Ryzin (1994) when

the demand function is initially unknown and no prior information is available. In a slightly simpler setting, Rusmevichientong and Broder (2012) analyze the case where demand is given by a parametric choice model. Roughly speaking, the latter two papers are instances of online stochastic convex optimization problem (either with or without pathwise constraints). As such, the methodology used to study them differs from the discrete and combinatorial nature of the assortment decision problem.

2.4. Connection to the Multiarmed Bandit Literature

The multiarmed bandit problem is one of the earliest instances of the aforementioned exploration versus exploitation trade-off. Introduced by Thompson (1933) and Robbins (1952), in its basic formulation, a decision maker seeks to maximize cumulative reward by pulling arms (of a slot machine) sequentially over time (one at each time) when limited prior information on reward distributions is available. The dynamic assortment planning setup can be viewed as a multiarmed bandit problem via the following analogy: each arm corresponds to a feasible assortment; hence, pulling an arm is the same as offering the assortment to a consumer. Reward distributions are determined by the purchase probabilities, which are initially unknown, and product profit margins. Application of standard multiarmed bandit algorithms would result in a regret (we define this concept in the next section) of order $\binom{N}{C}\log T$, where N is the total number of products available, C is the assortment capacity, and T is the length of the planning horizon. However, such an approach fails to incorporate two features that separates dynamic assortment planning from the multiarmed setting: (i) assortment rewards are not independent (a key assumption in the classical multiarmed setting), and (ii) assortments are not a priori identical since product profit margins are not necessarily equal (see the discussion below).

To address (i), it is possible to take advantage of the underlying reward structure. This is essentially the approach in Rusmevichientong et al. (2010), where the authors exploit the connection between the solution to the single-sale profit maximization problem and the underlying model parameters to limit the number of arms (assortments) worthy of consideration. In particular, they identify order-N arms among which the optimal one is found with high probability, and these arms are fed to a standard multiarmed bandit algorithm. The proposed algorithm works in cycles, and explores order-N^2 assortments on each of them. As a consequence, the overall procedure results in a regret of order $(N\log T)^2$. Alternatively, one can envision the dynamic assortment planning problem as a multiarmed bandit problem with multiple simultaneous plays; each product constitutes an arm by itself, and the decision maker can select multiple arms at each time. Indeed, this is the approach of Caro and Gallien (2007), who use a dynamic programming formulation and Bayesian learning approach to solve the exploration versus exploitation trade-off optimally. (See also Farias and Madan 2011 for a similar bandit formulation with multiple simultaneous plays under a more restricted class of policies.) In this paper, we show how one can restrict exploration to *at most* order-N assortments, hence significantly reducing the combinatorial complexity $\binom{N}{C}$, which would characterize the problem if standard bandit approaches were used.

Regarding (ii), note that in the bandit setting, arms are ex ante identical; hence, there is always the possibility that a poorly explored arm is in fact optimal. (In their seminal work, Lai and Robbins 1985 showed that any "good" policy should explore each arm at least order-$\log T$ times.) In the assortment planning setting, arms (either assortment or products, depending on the arm analogy being used) are *not* ex ante identical, and revenue is capped by the products' profit margins. In §4, we show how this observation can be exploited to limit exploration on certain *strictly* 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 multiarmed bandit literature, most notably the lower bound technique developed by Lai and Robbins (1985), but it also exploits salient features of the assortment problem in constructing optimal algorithms and highlighting key differences from traditional bandit results; this will become evident as we flesh out our main results and return to discuss these connections in §7.

3. Problem Formulation

3.1. 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 r_i and c_i denote the price and the marginal cost of product i, respectively. As mentioned in §1, we assume both prices and marginal costs are fixed and constant throughout the selling horizon. For $i \in \mathcal{N}$, let $w_i := r_i - c_i > 0$ denote the marginal profit resulting from selling one unit of the product, and let $w := (w_1, \ldots, w_N)$ denote the vector of profit margins. Due to display space constraints, the retailer can offer at most C products simultaneously. We assume, without loss of generality, that $C \le N$.

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 $t=T$ to the last. We assume a perfect inventory replenishment policy, and that the retailer has 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 tractability and allow us to extract structural insights.)

We 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

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

where $\mu_i \in \mathbb{R}$ denotes the mean utility assigned to product i, $\zeta_i^1, \ldots, \zeta_i^T$ are independent and identically distributed random variables drawn from a distribution F, and product 0 represents a no-purchase alternative. (See §7 for a discussion of an alternative utility specification.)

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 $\mu_0 := 0$.

3.2. The Single-Sale Profit Maximization Problem

Let $\mathcal{S}$ denote the set of possible assortments, that is, $\mathcal{S} := \{S \subseteq \mathcal{N} : |S| \le C\}$, where $|S|$ denotes the cardinality of the set $S \subseteq \mathcal{N}$. 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 given by

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

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

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

We let $S^*(\mu)$ denote the assortment that maximizes the single-sale profit; that is,

$$S^*(\mu) \in \arg\max_{S \in \mathcal{S}} r(S,\mu). \tag{1}$$

In what follows we will assume that the solution to the single-sale problem is unique. (This assumption greatly simplifies our exposition, in particular our performance bounds. Such bounds can be generalized to the case of multiple solutions, and we briefly indicate how one might do so in the proof of Theorem 1 in Appendix A.) We assume that the retailer can compute $S^*(\mu)$ for any vector μ; solving problem (1) efficiently is beyond the scope this paper.

Remark 1 (On Solving a Special Case). The MNL is by far the most commonly used choice model in the literature. Rusmevichientong et al. (2010) present an order-N^2 algorithm to solve the single-sale problem when such a choice model is assumed, that is, when F is assumed to be a standard Gumbel distribution (with location parameter 0 and scale parameter 1) for all $i \in \mathcal{N}$. The algorithm, based on a more general solution concept developed by Megiddo (1979), can in fact be used to solve the single-sale problem efficiently for any attraction-based choice model (these are choice models for which $p_i(S) = \nu_i/(\sum_{j \in S} \nu_j)$ for a vector $\nu \in \mathbb{R}_+^N$, and any $S \subseteq \mathcal{N}$; see, for example, Anderson et al. (1992).

3.3. The Dynamic Optimization Problem

We assume that the retailer knows F, the distribution that generates the idiosyncrasies of customer utilities, but *does not know* the mean vector μ. The retailer is able to observe purchase/no-purchase decisions made by each customer. He or she needs to decide what assortment to offer to each customer, taking into account all information gathered up to that point in time, to maximize expected cumulative profits. More formally, let $(S_t \in \mathcal{S} : 1 \le t \le T)$ denote an *assortment process*, with $S_t \in \mathcal{S}$ for all $t \le T$. Let

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

denote the purchase decision of customer t regarding product $i \in S_t$, where here, and in what follows, $\mathbf{1}\{A\}$ denotes the indicator function of a set A, that is, $Z_i^t = 1$ indicates that customer t decided to purchase product i, and $Z_i^t = 0$ otherwise. Also, let $Z_0^t := \mathbf{1}\{U_0 > U_j,\, j \in S_t\}$ denote the overall purchase decision of customer t, where $Z_0^t = 1$ if customer t opted not to purchase any product, and $Z_0^t = 0$ otherwise. We denote by $Z^t := (Z_0^t, Z_1^t, \ldots, Z_N^t)$ the vector of purchase decisions of customer t. Let $\mathcal{F}_t = \sigma((S_u, Z^u), 1 \le u \le t)$ $t = 1, \ldots, T$, denote the filtration (history) associated with the assortment process and purchase decisions up to (including) time t, with $\mathcal{F}_0 = \varnothing$. An admissible assortment policy π is a mapping from past history to assortment decisions such that the associated assortment process $(S_t \in \mathcal{S} : 1 \le t \le T)$ is nonanticipating (i.e., S_t is $\mathcal{F}_{t-1}$-measurable, for all t). 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

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

If the mean utility vector μ is known at the start of the selling season, the retailer would offer the assortment that maximizes the single-sale profit, $S^*(\mu)$, to every customer. The corresponding expected cumulative revenues, denoted by $J^*(T,\mu)$, would be

$$J^*(T,\mu) := Tr(S^*(\mu),\mu).$$

This quantity provides an upper bound on expected revenues generated by *any* admissible policy, that is, $J^*(T,\mu) \geq J^\pi(T,\mu)$ for all $\pi \in \mathcal{P}$. Define the *regret* associated with a policy π to be

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

The regret of a policy π is a normalized measure of revenue loss due to the lack of a priori knowledge of consumer behavior, and it can be roughly thought of as the number of customers to whom nonoptimal assortments are offered over $\{1,\dots,T\}$.

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 to *guess* the optimal assortment, $S^*(\mu)$, with high probability. On the other hand, the longer the retailer explores, the less consumers will be offered a supposedly optimal assortment; therefore, the retailer has incentives to reduce the exploration efforts in favor of exploitation.

4. Fundamental Limits on Achievable Performance

4.1. A Lower Bound on the Performance of Any Admissible Policy

We begin this section narrowing down the set of policies worthy of consideration. We say that an admissible policy is *consistent* if for all $\mu \in \mathbb{R}^N$

$$\frac{\mathcal{R}^\pi(T,\mu)}{T^a} \to 0, \tag{2}$$

as $T \to \infty$, for every $a > 0$. In other words, the long-run single-sale profit of consistent policies converges to the profit generated by offering the optimal assortment, for all possible mean utility vectors. (The condition in (2) restricts the rate of such convergence in T.) Let $\mathcal{P}' \subseteq \mathcal{P}$ denote the set of nonanticipating consistent assortment policies.

Suppose the retailer knows up front the value of the components of μ associated with products in $S^*(\mu)$, whereas the other components remain unknown: We say a product is *potentially optimal* if it cannot be discarded solely on the basis of such prior information; this means that a product i is potentially optimal if there exists an alternative mean utility vector $\gamma \in \mathbb{R}^N$ for which product i is optimal (i.e., $i \in S^*(\gamma)$), and that coincides with the original one, μ, on the components of products in $S^*(\mu)$. (Note that this definition does not consider changes in w, the vector of profit margins.) Define $\bar{\mathcal{N}}(\mu)$ as the set of potentially optimal products; that is,

$$\bar{\mathcal{N}}(\mu) := \{j \in \mathcal{N}\colon j \in S^*(\gamma) \text{ for some } \gamma \in \mathbb{R}^N \text{ such that } \mu_i = \gamma_i \ \forall\, i \in S^*(\mu)\}.$$

Similarly, we say a product is *strictly suboptimal* if it is not potentially optimal, that is, if it can be discarded as suboptimal based on partial knowledge of the mean utility vector; in other words, these products would not be included in the optimal assortment under any alternative mean utility vector among those that do not change mean utilities of products in $S^*(\mu)$. We define $\underline{\mathcal{N}} := \mathcal{N}\backslash\bar{\mathcal{N}}$ as the set of strictly suboptimal products. (In a slight abuse of notation, we drop dependencies on μ when possible.)

It is worth noting that this classification (potential optimality versus strict suboptimality) depends on (i) the vector of profit margins, which is observable at all times; and (ii) mean utilities of optimal products, which are initially unknown. Hence, the retailer cannot separate these two classes upfront with certainty.

In constructing bounds on achievable performance, we will consider a subclass of potentially optimal products, namely, those that become optimal under some *unilateral* change on their mean utilities. Define

$$\tilde{\mathcal{N}} := \{i \in \mathcal{N}\colon i \in S^*(\gamma), \gamma := (\mu_1,\dots,\mu_{i-1},v,\mu_{i+1},\dots,\mu_N) \text{ for some } v \in \mathbb{R}\}.$$

Potentially optimal products are, by definition, those that become optimal when alternative mean utility vectors, differing possibly on several coordinates, are considered: For a product in $\tilde{\mathcal{N}}$ such an alternative mean utility configuration differs from μ only on its jth component. (It follows that $\tilde{\mathcal{N}} \subseteq \bar{\mathcal{N}}$.)

We assume 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. The result below establishes a fundamental limit on what can be achieved by any consistent assortment policy. Recall that $|S|$ denotes the cardinality of a set $S \subseteq \mathcal{N}$.

THEOREM 1. *For any $\pi \in \mathcal{P}'$, and any $\mu \in \mathbb{R}^N$, there exist finite constants $\underline{K}$ and $\underline{K}'$, such that*

$$\mathcal{R}^\pi(T,\mu) \geq \underline{K}(|\tilde{\mathcal{N}}\backslash S^*(\mu)|/C)\log T + \underline{K}',$$

for all T.

Recall that $\tilde{\mathcal{N}}$ is a subset of potentially optimal products, and $\tilde{\mathcal{N}}\backslash S^*(\mu)$ is the result of removing $S^*(\mu)$ from that set. Expressions for the constants $\underline{K}$ and $\underline{K}'$ are given in Appendix A. Note that if one were to treat each possible assortment as a different arm and appeal to standard bandit-type algorithms, the regret would scale linearly with a combinatorial term of order $\binom{N}{C}$, instead of the much smaller constant $(|\tilde{\mathcal{N}}\backslash S^*(\mu)|/C)$ appearing above. Theorem 1 also suggests that when all nonoptimal products are strictly suboptimal (and hence $\tilde{\mathcal{N}} = S^*(\mu)$), a finite regret may be attainable. It is worth noting that $\bar{\mathcal{N}} = \tilde{\mathcal{N}}$ for Luce-type choice models, the MNL being a special case (this also holds for other choice models under certain conditions). When this is not the case, one can adapt our results to provide a tighter bound where the regret scales linearly with $(|\mathcal{N}'\backslash S^*(\mu)|/C)$ for a set $\mathcal{N}' \subseteq \mathcal{N}$ such that $\tilde{\mathcal{N}} \subseteq \mathcal{N}' \subseteq \bar{\mathcal{N}}$.

Remark 2 (Implications for Design of "Good" Policies). The proof of Theorem 1, which is outlined below, suggests certain desirable properties for assortment policies: (i) potentially optimal products are to be tested on order-$\log T$ customers, and (ii) product experimentation should be conducted in batches of size C and only on potentially optimal products. In addition, Theorem 1 does not impose a priori constraints on the number of customers to whom strictly suboptimal products are offered to. This suggests that strictly- suboptimal products may only be tested on a finite number of customers (in expectation), *independent* of T. This will be proved in what follows (see Corollary 1).

4.2. Proof Outline and Intuition Behind Theorem 1

The proof of Theorem 1 exploits the connection between the regret and testing of suboptimal assortments. In particular, it bounds the regret by computing lower bounds on the expected number of tests involving some potentially optimal products that are not optimal (those in $\tilde{\mathcal{N}}\backslash S^*(\mu)$): each time such a product is offered, the corresponding assortment must be suboptimal, contributing directly to the policy's regret.

To bound the number of tests involving nonoptimal products, we use a change-of-measure argument introduced by Lai and Robbins (1985) for proving an analogous result for a multiarmed bandit problem. To adapt this idea, we need to address the fact that realizations of the underlying random variables (i.e., product utilities) are nonobservable in the assortment setting, which differs from the bandit setting where reward realizations are observed directly. Our argument can be roughly described as follows. By construction, any nonoptimal product $i \in \tilde{\mathcal{N}}$ is in the optimal assortment for at least one alternative (suitable chosen) mean utility vector. When such an alternative vector is considered, any consistent policy π must offer product i to all but a subpolynomial (in T) number of customers. If this alternative vector does not differ in a "significant manner" from the original, a notion that is made precise in Appendix A, then one would expect this product to be offered to a large number of customers under the original mean utility vector μ. In particular, for a product i in $\tilde{\mathcal{N}}$, the alternative vector differs from μ only on the parameter associated with product i: One can use this observation to show that for any policy π,

$$\mathbb{P}_\pi\{T_i(T) \le \log T/K_i\} \to 0 \tag{3}$$

as $T \to \infty$, $i \in \tilde{\mathcal{N}}$, 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. Note that this asymptotic minimum-testing requirement is inversely proportional to K_i, which turns out to be a measure of how close the vector μ is to a configuration that makes product i be part of the optimal assortment. One can use the above to bound the expected number of times nonoptimal products in such a class are tested: Using Markov's inequality, we have that, for any $i \in \tilde{\mathcal{N}}\backslash S^*(\mu)$,

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

The result in Theorem 1 follows directly from the 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 insights gleaned from Theorem 1. Our policy is based on the idea that performance of a product in a given assortment, measured in terms of frequency of purchase, should provide information on the performance of the same product in other assortments. More formally, one might recover mean utilities of products on a given assortment by observing the frequency at which products are purchased when such an assortment is offered. With this in mind, we introduce the following assumption.

Assumption 1 (Identifiability). *For any vector $\rho \in \mathbb{R}_+^N$ such that $\sum_{i\in\mathcal{N}} \rho_i < 1$, there exists a unique vector $\eta(\rho)$ such that $p_i(\mathcal{N}, \eta(\rho)) = \rho_i$, for all $i \in \mathcal{N}$. In addition, $p(\mathcal{N}, \cdot)$ is Lipschitz continuous, and $\eta(\cdot)_i$ is locally Lipschitz continuous in the neighborhood of ρ, when $\rho_i > 0$.*

Note that because F is absolutely continuous, any $i \in \mathcal{N}$ with $\rho_i = 0$ might be regarded as *infinitely*

unattractive to consumers (i.e., $\eta(\rho)_i = -\infty$), and thus can be ignored. Under this assumption, one can recover mean utilities for products in a given assortment from the associated purchase probability vector. We exploit this when estimating the mean utility vector μ: we first estimate purchase probabilities by observing consumer purchase decisions; then, we use those probabilities to reconstruct a mean utility vector that is consistent with 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, we first consider a policy that separates exploration from exploitation. Assuming prior knowledge of T, such a policy first engages in an exploration phase, where $\lceil N/C \rceil$ assortments, encompassing all products, are offered sequentially to order-log T customers ($\lceil n \rceil$ denotes the smallest integer larger than a real number n); the intuition for this scale comes from Theorem 1. Then, an estimator for μ is computed based on observed purchase decisions. Later, in the exploitation phase, this estimator is used to compute a proxy for the optimal assortment, which is then offered to the remaining customers. Define the set of test assortments $\mathcal{A} := \{A_1, \ldots, A_{\lceil N/C \rceil}\}$ used in the exploration phase, where

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

Suppose $t-1$ customers have arrived: For each $A_j \in \mathcal{A}$, we use $\hat{p}_{i,t}(A_j)$ to estimate $p_i(A_j, \mu)$, where

$$\hat{p}_{i,t}(A_j) := \frac{\sum_{u=1}^{t-1} Z_i^u \mathbf{1}\{S_u = A_j\}}{\sum_{u=1}^{t-1} \mathbf{1}\{S_u = A_j\}} \tag{4}$$

for $i \in A_j \cup \{0\}$, and $\hat{p}_{i,t}(A_j) = 0$ otherwise. Let $\hat{p}_t(A_j) := (\hat{p}_{1,t}(A_j), \ldots, \hat{p}_{N,t}(A_j))$ denote the vector of estimated purchase probabilities associated with test assortment A_j.

For $i \in \mathcal{N}$, we use $\hat{\mu}_{t,i}$ to estimate μ_i, with

$$\hat{\mu}_{t,i} := (\eta(\hat{p}_t(A_j)))_i,$$

where A_j corresponds to the unique test assortment including product i, and $(a)_i$ denotes the ith component of the vector a. The procedure above allows one to separate estimation across subsets of products, as opposed to estimating *all* parameters simultaneously (which is computationally expensive for large problem instances). However, the procedure does not allow refining parameter estimates using information collected from offerings beyond the exploration phase. Let $\hat{\mu}_t := (\hat{\mu}_{t,1}, \ldots, \hat{\mu}_{t,N})$ denote the vector of mean utility estimates. One can show that when Assumption 1 holds, our method is an instance of maximum-likelihood estimation (MLE); see Daganzo (1979, p. 118). See further discussion of key features and possible limitations of our approach in §7.

The idea behind the separation-based policy is the following: When an assortment $A_j \in \mathcal{A}$ has been offered to a large number of customers, one would expect $\hat{p}_t(A_j)$ to be close to $p(A_j, \mu)$. If this is the case for all assortments in $\mathcal{A}$, by Assumption 1, one would also expect $\hat{\mu}_t$ to be close to μ. The separation-based policy, summarized for convenience in Algorithm 1, is defined through a positive constant κ_1 that regulates the length of the exploration phase.

Algorithm 1 ($\pi_1 = \pi(\kappa_1, T, w)$)

Step 1. Exploration:
Offer each test assortment in $\mathcal{A}$ to $\lceil \kappa_1 \log T \rceil$ customers. [Exploration]

Step 2. Exploitation:
Compute estimate $\hat{\mu}_t := \{\hat{\mu}_{t,1}, \ldots, \hat{\mu}_{t,N}\}$.
Offer $S^*(\hat{\mu}_t)$ to all remaining customers. [Exploitation]

This policy is constructed to guarantee that the expected revenue loss during the exploitation phase balances that stemming from exploration efforts, which is of order-log T. This, in turn, translates into an order-$(\lceil N/C \rceil \log T)$ regret. The next result formalizes this.

Theorem 2. *Let $\pi_1 := \pi(\kappa_1, T, w)$ be defined by Algorithm* 1*, and let Assumption* 1 *hold. There exist finite constants $\bar{K}_1$ and $\bar{\kappa}_1$ such that the regret associated with π_1 is bounded as follows*:

$$\mathcal{R}^{\pi_1}(T, \mu) \le \kappa_1(\lceil N/C \rceil) \log T + \bar{K}_1,$$

for all T, provided that $\kappa_1 > \bar{\kappa}_1$.

Constants $\bar{K}_1$ and $\bar{\kappa}_1$ are instance specific, but do not depend on the length of the selling horizon. Proof of Theorem 2 elucidates that $\bar{K}_1$ bounds the expected cumulative revenue loss incurred during the exploitation phase, whereas $\bar{\kappa}_1$ represents the minimum value of κ_1 that makes such a bound finite and independent of T. The bound presented in Theorem 2 is essentially the one in Theorem 1, with N replacing $|\tilde{\mathcal{N}} \setminus S^*(\mu)|$. This indicates that (i) imposing the right order (in T) of exploration is enough to obtain the right dependence (in T) of the regret, and (ii) achieving the lower bound requires limiting the exploration on strictly suboptimal products.

Remark 3 (Selection of the Tuning Parameter κ_1). We have established that the lower bound in Theorem 1 can be achieved, in terms of its dependence on T, for proper choice of κ_1. However, Theorem 2 requires κ_1 to be greater than $\bar{\kappa}_1$, whose value is not known a priori. In particular, setting κ_1 below

the specified threshold might compromise the performance guarantee in Theorem 2. To avoid the risk of misspecifying κ_1, one can increase the length of the exploration phase to, say, $|\mathcal{A}|\kappa_1(\log t)^{1+\alpha}$, for any $\alpha > 0$. With this, the upper bound above would read

$$\mathcal{R}^{\pi}(T, \mu) \leq \kappa_1 \lceil N/C \rceil (\log T)^{1+\alpha} + \bar{K}_1,$$

for any κ_1, and the policy becomes optimal up to a $(\log T)^{\alpha}$-term.

Next, we illustrate the performance of Algorithm 1 in two examples that consider the most prevalent choice models in the literature, that is, the logit and probit models.

Example 1 (Performance of the Separation-Based Policy π_1 for an MNL Choice Model). Consider $N = 10$ and $C = 4$, with

$$w = (0.98, 0.88, 0.82, 0.77, 0.71, 0.60, 0.57, 0.16, 0.04, 0.02),$$

$$\mu = (0.36, 0.84, 0.62, 0.64, 0.80, 0.31, 0.84, 0.78, 0.38, 0.34),$$

and assume $\{\zeta_i^t\}$ have a standard Gumbel distribution, for all $i \in \mathcal{N}$ and all $t \geq 1$, that is, we consider the MNL choice model. One can verify that $S^*(\mu) = \{1, 2, 3, 4\}$ and $r(S^*(\mu), \mu) = 0.76$. Test assortments are given by $A_1 = \{1, 2, 3, 4\}$, $A_2 = \{5, 6, 7, 8\}$, and $A_3 = \{9, 10\}$.

Example 2 (Performance of the Separation-Based Policy π_1 for a Multinomial Probit Choice Model). Consider $N = 6$ and $C = 2$, with

$$w = (2.00, 1.80, 1.50, 1.40, 1.20, 1.00),$$

$$\mu = (0.20, 0.30, 0.35, 0.45, 0.50, 0.55),$$

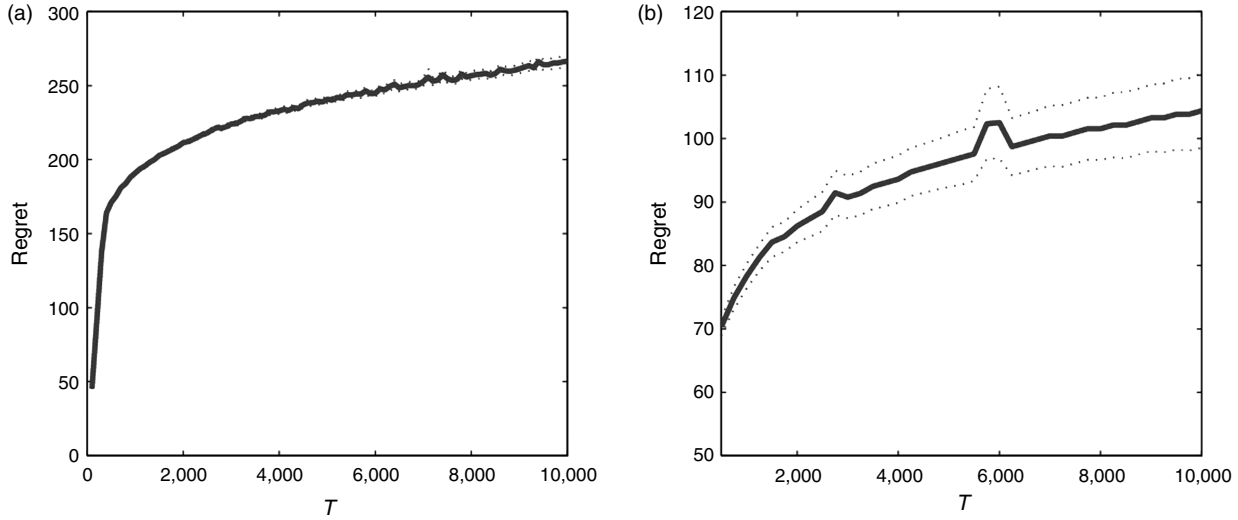
and assume $\{\zeta_i^t\}$ have a standard normal distribution, for all $i \in \mathcal{N}$ and all $t \geq 1$, that is, we consider the multinomial probit choice model. One can verify that $S^*(\mu) = \{1, 2\}$ and $r(S^*(\mu), \mu) = 1.39$. Test assortments are given by $A_1 = \{1, 2\}$, $A_2 = \{3, 4\}$, and $A_3 = \{5, 6\}$.

Panels (a) and (b) in Figure 1 depict the average performance of policy π_1 for instances in Examples 1 and 2, respectively. In Example 1, parameter estimates are computed via $\eta(\hat{p}_t(A_j))$ using the closed-form expression for $\eta(\cdot)$ given in (6). In Example 2, $\eta(\cdot)$ cannot be expressed in closed form, and we use maximum simulated likelihood estimation. (Recall that when Assumption 1 holds, our method is an instance of MLE.) Note that existence of an inverse mapping for the trinomial probit model is well known; see Daganzo (1979). Additional details on parameter estimation can be found in Appendix B.

In Examples 1 and 2, we solve the single-sale profit maximization problem via enumeration. Simulation results were conducted over 500 replications, using $\kappa_1 = 20$ and considering selling horizons ranging from $T = 500$ to $T = 10{,}000$. Dotted lines represent 95% confidence intervals for the simulation results. Note that the regret in both panels seems to be of order $\log T$, as predicted by Theorem 2. Also, note that policy π_1 makes suboptimal decisions on a diminishing fraction of customers; for example, in panel (a) it ranges from around 10% when the horizon is 2,000 sales attempts, and diminishes to around 2.5% for a horizon of 10,000. (Recall that the regret is directly linked to the number of suboptimal sales.)

In the case of Example 1, one can show that $\bar{\mathcal{N}} = \{1, 2, 3, 4\}$ (see §5.3). We observe that, by construction, in this setting assortments A_2 and A_3 are offered to order-$\log T$ customers, despite being composed

Figure 1 Performance of the Separation-Based Policy π_1



Notes. The graphs (a) and (b) illustrate the dependence of the regret on T for instances in Examples 1 and 2, respectively. Dotted lines represent 95% confidence intervals for the simulation results.

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

Ideally, a policy should offer suboptimal assortments to at most order-$\log T$ consumers, and those assortments should not include strictly suboptimal products. Thus, such a policy should be able to "identify" strictly suboptimal products when there is no information about the mean utility provided by *any* of these products. We observe that, in general, there exists a threshold value, $\omega(\mu) < r(S^*(\mu), \mu)$, such that

$$\underline{\mathcal{N}} = \{i \in \mathcal{N}\colon w_i < \omega(\mu)\};$$

that is, any product with margin less than this threshold value is strictly suboptimal and vice versa. This observation follows from noting that products are ex ante differentiated only through their profit margins; hence, it is not possible for a potentially optimal product to have a lower profit margin than a strictly suboptimal one. One can use this observation in the design of test assortments: Consider the set of test assortments $\mathcal{A} := \{A_1, \ldots, A_{\lceil N/C \rceil}\}$, where

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

and $i_{(k)}$ corresponds to the product with the kth highest profit margin. Suppose one has a proxy for $\omega(\mu)$. 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 assortments containing only strictly suboptimal products. Note that in practice, $\omega(\mu)$ must be computed numerically for most choice models. This procedure is greatly simplified when $\bar{\mathcal{N}} = \tilde{\mathcal{N}}$, so that assessing potential optimality is equivalent to solving a one-dimensional, single-sale profit maximization problem.

Next, we propose a policy that limits exploration on strictly suboptimal products and show that it performs well for any value of T. The policy executes the following logic upon arrival of customer t: Using $\hat{\mu}_t$, the current estimate of μ, it solves for $S_t = S^*(\hat{\mu}_t)$ and computes the threshold value $\omega_t = \omega(\hat{\mu}_t)$. If all assortments in $\mathcal{A}$ containing products with margins greater than or equal to ω_t have been tested on a minimum number of customers, then assortment S_t is offered to customer t. Otherwise, we select, arbitrarily, an *undertested* assortment in $\mathcal{A}$ containing at least one product with margin greater than or equal to ω_t and offer it to the current customer. (The term undertested means tested on less than order-$\log t$ customers prior to the arrival of customer t.) Note that this logic will enforce the correct order of exploration for any value of T.

Algorithm 2 ($\pi_2 = \pi(\kappa_2, w)$)

Step 1. Initialization:
Offer each test assortment in $\mathcal{A}$ to a customer. [Initial test]

Step 2. Joint exploration and assortment optimization:

for customer t **do**
Compute estimate $\hat{\mu}_t := \{\hat{\mu}_{t,1}, \ldots, \hat{\mu}_{t,N}\}$, and $\omega_t = \omega(\hat{\mu}_t)$.
Set $\mathcal{A}_t = \{A_j \in \mathcal{A}\colon \max\{w_i\colon i \in A_j\} \geq \omega_t\}$. [Test assortments]
if some $A_j \in \mathcal{A}_t$ has been offered to less than $\kappa_2 \log t$ customers **then**
Offer such A_j to customer t. [Exploration]
else
Offer $S^*(\hat{\mu}_t)$ to customer t. [Exploitation]
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 a minimum exploration frequency 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 κ_2 to balance exploration (which contributes directly to the regret), and the expected revenue loss in the exploitation phase.

The next result characterizes the performance of the proposed assortment policy. Recall that $\lceil n \rceil$ denotes the smallest integer larger than a real number n.

Theorem 3. *Let $\pi_2 = \pi(\kappa_2, w)$ be defined by Algorithm 2, and let Assumption 1 hold. There exist finite constants $\bar{K}_2$ and $\bar{\kappa}_2$, such that the regret associated with π_2 is bounded as follows*:

$$\mathcal{R}^{\pi}(T, \mu) \leq \kappa_2(\lceil |\bar{\mathcal{N}}|/C \rceil) \log T + \bar{K}_2,$$

for all T, provided that $\kappa_2 > \bar{\kappa}_2$.

The performance guarantee in Theorem 3 manifests the correct dependence on both T and $\bar{\mathcal{N}}$, as per Theorem 1 (up to the size of the optimal assortment, and the difference between $\bar{\mathcal{N}}$ and $\tilde{\mathcal{N}}$). The result essentially shows that focusing exploration efforts on a set of products "rich enough" to provide an optimality guarantee for the incumbent optimal solution suffices for identifying the optimal assortment with high probability. Note that the argument in Remark 3 remains valid in regard to the selection of κ_2. Theorem 3 also states

(implicitly) that assortments containing only strictly suboptimal products will be tested on a finite number of customers (in expectation). The following corollary formalizes this statement. Recall that $T_i(t)$ denotes the number of customers product i has been offered to, up to the arrival of customer t.

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

$$\mathbb{E}_\pi[T_i(T)] \leq K_2,$$

for all $i \in A_j$, *where* K_2 *is a finite positive constant independent of* T.

Remark 4 (Relationship to Bandit Problems). The result in Corollary 1 stands in contrast to typical multiarmed 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, the contribution of a product to the overall profit is bounded, independent of its mean utility. More importantly, this feature makes some products a priori *better* than others. Such characteristic is not present in the typical bandit problem, and the above result illustrates some of its implications.

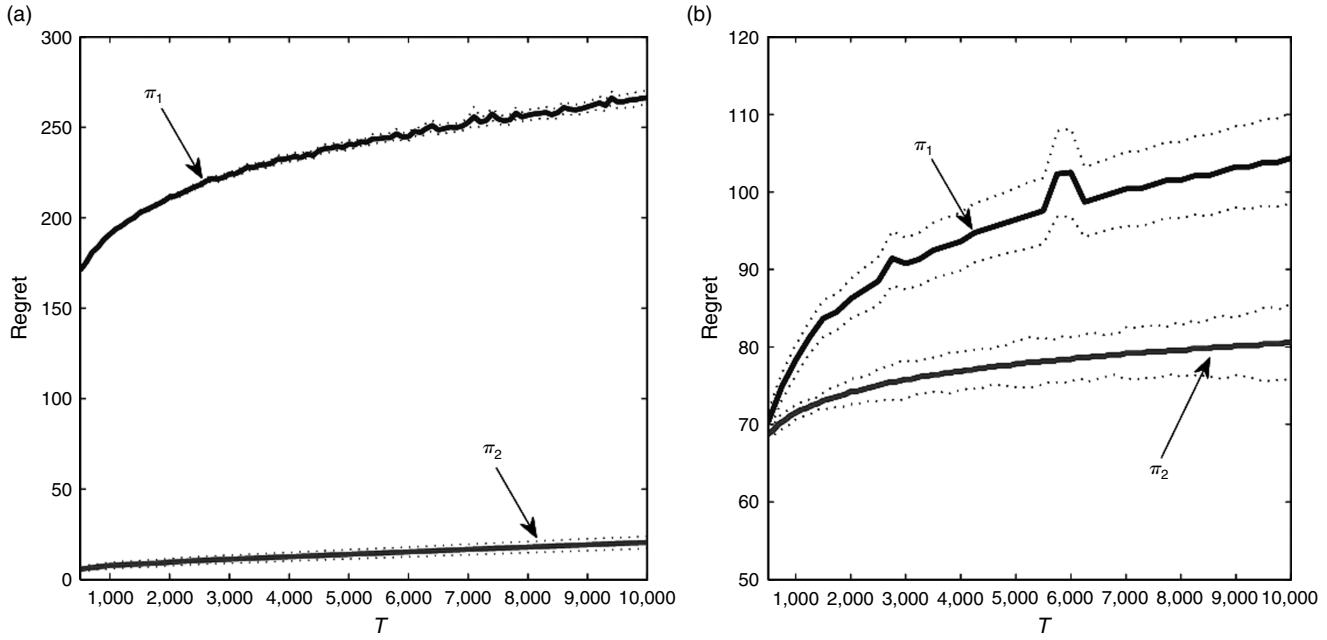
Next, we illustrate the performance of the proposed algorithm for the examples of §5.1.

Example 1 (Continued) (Performance of the Policy π_2 for the MNL Choice Model). Consider the setting of Example 1 in §5.1. In §5.3, we show that $\omega(\mu) = r(S^*(\mu), \mu)$; hence, one has that $\underline{\mathcal{N}} = A_2 \cup A_3$. Note that $\bar{\mathcal{N}} = A_1$, thus one would expect Algorithm 2 to offer suboptimal assortments to a finite number of consumers, independent of T.

Example 2 (Continued) (Performance of the Policy π_2 for the Probit Choice Model). Consider the setting of Example 2 in §5.1. One can check (numerically) that $\underline{\mathcal{N}} = A_3$. Because test assortment A_2 is suboptimal, but contains potentially optimal products, Algorithm 2 should offer it to order-log T consumers.

Panels (a) and (b) in Figure 2 depict the average performance of policies π_1 and π_2 for instances in Examples 1 and 2, respectively. Parameter estimation and single-sale profit maximization are conducted as in Examples 1 and 2. The threshold value $\omega(\mu)$ is computed through its closed-form expression for the case of Example 1 above. For the case of Example 2, one can show that $\bar{\mathcal{N}} = \tilde{\mathcal{N}}$; thus, computing $\mathcal{A}_t$ requires verifying potential optimality for only one product per test assortment. Such a step is conducted through numerical maximization and simulation. (Simulation is used to approximate purchase probabilities, and numerical maximization is used to find an alternative mean utility configuration that improves the optimal single-sale profit.)

Figure 2 Performance of the Refined Policy π_2



Notes. The graphs (a) and (b) compare the separation-based policy π_1, given by Algorithm1, and the proposed policy π_2, in terms of regret-dependence on T, for instances in Examples 1 and 2, respectively. Dotted lines represent 95% confidence intervals for the simulation results.

Simulation results were conducted over 500 replications, using $\kappa_1 = \kappa_2 = 20$. Graphs (a) and (b) compare the regret produced by the separation-based policy π_1 and the proposed policy π_2 for selling horizons ranging from $T = 500$ to $T = 10{,}000$. Dotted lines represent 95% confidence intervals for the simulation results. We observe that policy π_2 outperforms substantially the separation-based policy π_1. In particular, for the instance in Example 1, π_1 results in lost sales in the range of 2.5%–10% (200–260 customers are offered nonoptimal choices) depending on the length of selling horizon, whereas for π_2 we observe suboptimal 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. Such an improvement in performance can be explained as follows: π_2 identifies that both A_2 and A_3 contain only strictly suboptimal products, with increasing probability as t grows large. As a result, exploration efforts are eventually directed exclusively to the optimal assortment. Because incorrect choices in the exploitation phase are also controlled by π_2, we expect the regret to be finite. This is supported by the numerical results displayed in Figure 2.

5.3. A Policy Customized to the Multinomial Logit Choice Model

In general, purchase probabilities depend on the offered assortment in a nontrivial way. With no trivial way to combine information collected from offering different assortments, it is not clear how to use data gathered in the exploitation phase efficiently. Next, we illustrate how to modify parameter estimation to include exploitation-based information in the case of an MNL choice model. Note that Rusmevichientong et al. (2010) present an efficient algorithm for solving the single-sale optimization problem in this setting. (As indicated previously, the results in this section extend directly to Luce-type choice models.)

Taking F to have a standard Gumbel distribution (see, e.g., Anderson et al. 1992),

$$p_i(S,\mu) = \frac{\nu_i}{1+\sum_{j\in S}\nu_j}, \quad i \in S, \text{ for any } S \in \mathcal{S}, \tag{5}$$

where $\nu_i := \exp(\mu_i)$, $i \in \mathcal{N}$, and $\nu := (\nu_1, \ldots, \nu_N)$. For a vector $\rho \in \mathbb{R}_+^N$ such that $\sum_{i\in\mathcal{N}}\rho_i < 1$, we have that $\eta(\rho)$, the unique solution to $\{\rho_i = p_i(\mathcal{N},\mu),\ i \in \mathcal{N}\}$, is given by

$$\eta_i(\rho) = \begin{cases} \ln\left(\rho_i\left(1-\sum_{j\in\mathcal{N}}\rho_j\right)^{-1}\right) & \rho_i > 0, \\ -\infty & \rho_i = 0, \end{cases} \tag{6}$$

$i \in \mathcal{N}$. One can check that (5) and (6) imply that $\bar{\mathcal{N}} = \tilde{\mathcal{N}}$. Indeed, solving the single-sale optimization problem in this setting is equivalent to finding the *largest* value of λ such that

$$\sum_{i\in S}\nu_i(w_i - \lambda) \geq \lambda, \tag{7}$$

for some $S \in \mathcal{S}$; thus, one can characterize the set of strictly suboptimal products as

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

This implies that $\omega(\mu) = r(S^*(\mu),\mu)$ for the MNL model.

We propose a policy, denoted π_3, customized for the MNL choice model. The policy, summarized for convenience in Algorithm 3, maintains the general structure of Algorithm 2; however, parameter estimation is conducted at the product level. As in the previous sections, the policy is defined through a positive constant κ_3 that regulates the length of the exploration phase.

Suppose $t-1$ customers have shown up so far. We use $\hat{\nu}_{i,t}$ to estimate ν_i, where

$$\hat{\nu}_{i,t} := \frac{\sum_{u=1}^{t-1} Z_i^u \mathbf{1}\{i \in S_u\}}{\sum_{u=1}^{t-1} Z_0^u \mathbf{1}\{i \in S_u\}}, \quad i \in \mathcal{N}, \tag{8}$$

and define $\hat{\mu}_{i,t} := \ln(\hat{\nu}_{i,t})$. The estimate above exploits the independence of irrelevant alternatives (IIA) property of the logit model, which states that the ratio between purchase probabilities of any two products is independent of the assortment in which they are offered; that is,

$$\frac{p_i(S,\mu)}{p_j(S,\mu)} = \frac{\nu_i}{\nu_j} \quad \text{for all products } i, j \in \mathcal{N}\cup\{0\}, \text{ for all } S \in \mathcal{S}.$$

Indeed, the IIA property allows us to perform parameter estimation on subsets of products without using predetermined assortments (see Chapter 3 in Train 2009 for further details). In our case, we perform separate estimation on each pair product/no-purchase alternative. As a result, all information collected (both from exploration and exploitation phases) is used to construct the parameter estimates. It is worth noting that a policy exploiting this feature might help correct errors made in the exploitation phase faster than the previous type of policy. In particular, estimates of expected single-sale profits for suboptimal assortments offered during exploitation are anticipated to converge faster to their actual values; thus, optimal products are likely to be identified as such at earlier stages (see the discussion following Example 3).

The next result characterizes the performance of the proposed assortment policy.

Algorithm 3 ($\pi_3 = \pi(\kappa_3, w)$)
Step 1. Initialization:
Offer each product $i \in \mathcal{N}$ by itself until a no-purchase occurs. [Initial test]
Step 2. Joint exploration and assortment optimization:
for customer t **do**
Compute estimates $\hat{\mu}_t := \{\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{N,t}\}$ and set $\omega_t = r(S^*(\hat{\mu}_t), \hat{\mu}_t)$.
Set $\bar{\mathcal{N}}_t = \{i \in \mathcal{N}: w_i \geq \omega_t\}$. [Potentially optimal products]
if some $i \in \bar{\mathcal{N}}_t$ has been offered to less than $\kappa_3 \log t$ customers **then**
Offer $S \in \{i \in \bar{\mathcal{N}}_t: T_i(t) \leq \kappa_3 \log t\} \cap \mathcal{S}$ to customer t. [Exploration]
else
Offer $S^*(\hat{\mu}_t)$ to customer t. [Exploitation]
end if
end for

Theorem 4. *Let $\pi_3 = \pi(\kappa_3, w)$ be defined by Algorithm* 3. *There exist finite constants $\bar{K}_3$ and $\bar{\kappa}_3$, such that the regret associated with π_3 is bounded as follows*:

$$\mathcal{R}^{\pi}(T, \mu) \leq \kappa_3(|\bar{\mathcal{N}} \backslash S^*(\mu)|) \log T + \bar{K}_3,$$

for all T, provided that $\kappa_3 > \bar{\kappa}_3$.

Theorem 4 is essentially the equivalent of Theorem 3, customized to the logit case, with the exception of the dependence on the assortment capacity C (as here exploration is conducted on a product basis) and the dependence on the set $\bar{\mathcal{N}}$. 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 conducted using information arising from both the exploration and exploitation phases, one would expect a better empirical performance from the logit-customized policy. In particular, strictly suboptimal products will be tested on a finite number of customers, in expectation, as shown in the following corollary.

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.

Regarding selection of the parameter κ_3, note that the argument in Remark 3 remains valid.

Example 1 (Continued) (Performance of the MNL-Customized Policy π_3). Consider the setup of Example 1 in §5.1. Note that $S^*(\mu) = A_2$; that is, the optimal assortment matches one of the test assortments. Moreover, one has that $\bar{\mathcal{N}} = S^*(\mu)$. As a result, strictly suboptimal detection is conducted in finite time for both policies π_2 and π_3; hence, any gain in performance for policy π_3 over π_2 is tied in to the ability of the former to incorporate information gathered during both exploration and exploitation phases.

Example 3 (Performance of the MNL-Customized Policy Revisited). Consider the setup of Example 1 in §5.1, but when

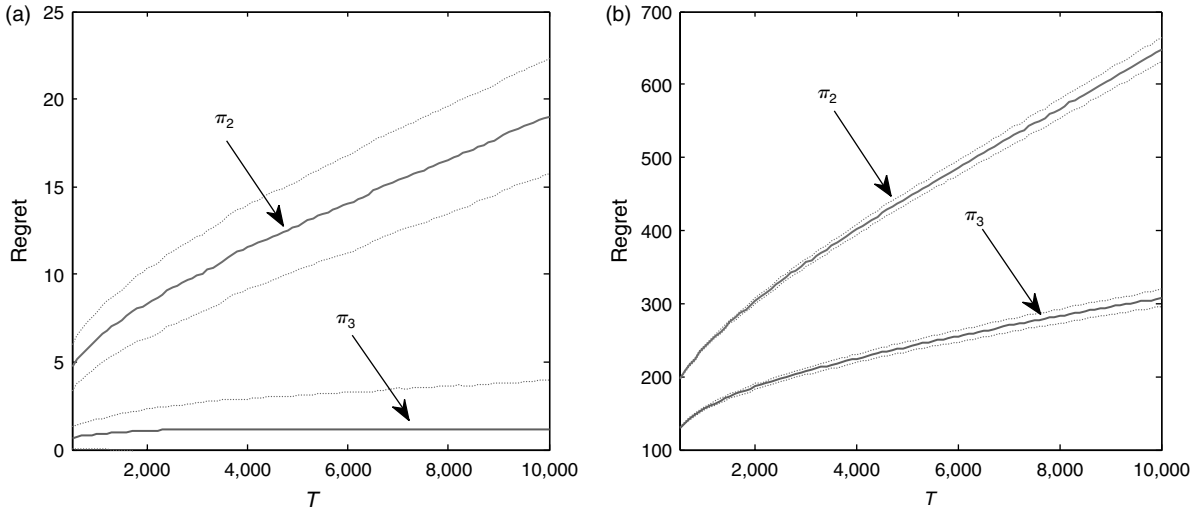
$$w = (0.95, 0.81, 0.75, 0.72, 0.68, 0.60, 0.58, 0.41, 0.35, 0.21),$$

$$\mu = -(2.83, 3.96, 5.50, 2.90, 2.60, 2.80, 3.20, 4.27, 4.60, 2.78).$$

This corresponds to a setting in which all products are less attractive than the no-purchase alternative. One can verify that $S^*(\mu) = \{1, 4, 5, 6\}$ and $r(S^*(\mu), \mu) = 0.147$. Note that $\underline{\mathcal{N}} = \varnothing$; thus, the difference in performance between π_2 and π_3 emanates mainly from the manner in which information collected during the exploitation and exploration phases is used.

Panels (a) and (b) in Figure 3 depict the average performance of policies π_2 and π_3 for instances in Examples 1 and 3, respectively. Simulation results were conducted over 500 replications, using $\kappa_2 = \kappa_3 = 20$ and considering selling horizons ranging from $T = 1{,}000$ to $T = 10{,}000$. Parameter estimation is conducted according to (8), and single-sale profit maximization is carried out by enumeration. The graphs compare the more general policy π_2 with its logit-customized version π_3 in terms of regret dependence on T. Dotted lines represent 95% confidence intervals for the simulation results. In graph (a), one can see that 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 for any T in the case of Example 1, as predicted by Theorem 4. This also suggests that differences in performance are mainly due to errors in the exploitation phase. This is reinforced by the results in graph (b), where we see that the logit-customized policy π_3 outperforms the more general policy π_2, confirming that the probability of error decays faster in the logit-customized version. Note that when all exploitation efforts are successful, and

Figure 3 Performance of the MNL-Customized Policy π_3



Notes. Graphs (a) and (b) compare the more general policy π_2 with its logit-customized version π_3, in terms of regret-dependence on T, for instances in Examples 1 and 3, respectively. Dotted lines represent 95% confidence intervals for the simulation results.

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 further details in the proof of Theorem 4 in the online companion.

6. Comparison with Benchmark Results

Our results significantly improve on and generalize the policy proposed by Rusmevichientong et al. (2010), where an order-$(N \log T)^2$ performance upper bound is presented for the case of an MNL choice model. Their algorithm for solving the single-sale optimization problem identifies a small set of assortments that contains the optimal one. In its dynamic formulation, the algorithm requires to test order-N^2 assortments to estimate the parameters allowing to identify such a set of candidate assortments with high probability. Note that such a dynamic policy, which operates in phases, is a more direct adaptation of multiarmed bandit ideas. Hence, it does not detect strictly suboptimal products, nor does it limit exploration on them. In addition, their policy conducts exploration efforts on order-N^2 test assortments, and periodically increases the magnitude the exploration effort while neglecting information collected in previous exploration phases. The regret of our logit-customized policy is at most of order-$|\mathcal{N} \setminus S^*(\mu)| \log T$, and we show that this cannot be improved upon.

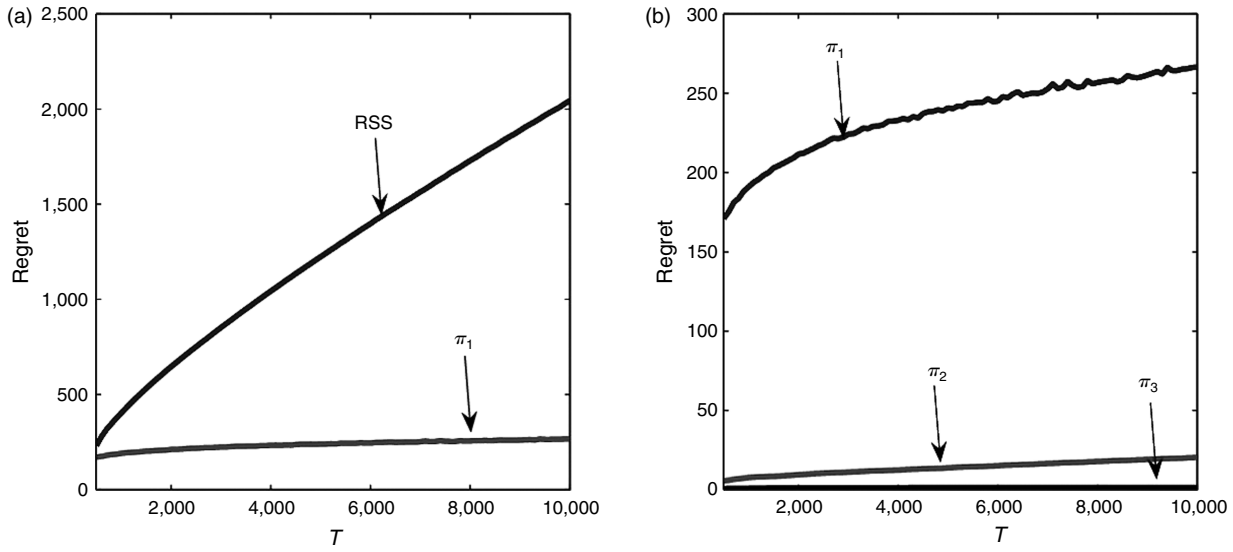
Consider again Example 1 in §5.1. Figure 4 compares the average performance of our proposed policies with that of Rusmevichientong et al. (2010), denoted RSS, over 500 replications, using $\kappa_1 = \kappa_2 = \kappa_3 = 20$, and considering selling horizons ranging from $T = 1{,}000$ to $T = 10{,}000$. The graph in (a) compares the separation-based policy π_1 with 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. A further detailed analysis of the results depicted in Figure 4 reveals that the regret of the benchmark behaves quadratically with $\log T$, as predicted. Panel (a) in Figure 4 shows that the RSS policy offers suboptimal assortments to about 20%–25% of the customers, whereas policy π_1 never exceeds 10%, and that loss diminishes as the horizon increases to around 2.5%. Because policies π_2 and π_3 limit exploration on strictly suboptimal products, a feature absent in both the RSS policy and the naive separation-based policy π_1, they exhibit far superior performance compared with either one of those benchmarks, as illustrated in panel (b) of Figure 4. Note that, unlike our logit-tailored policy, the policy in Rusmevichientong et al. (2010) uses only the information collected during the exploration phase for parameter estimation. The improvement in performance due to this feature is illustrated in panel (b) of Figure 4. The overall effect is that policy π_3 improves performance by a factor of 200–1,000 compared with RSS and is able to zero in on the optimal assortment much faster than the benchmark.

7. Discussion and Concluding Remarks

7.1. Summary and Main Insights

In this paper, we have studied the role of assortment experimentation in learning consumer preferences by

Figure 4 Comparison with a Benchmark Performance



Notes. The graph in (a) compares the separation-based policy π_1 with the benchmark policy RSS in terms of regret-dependence on T. The graph in (b) compares the separation-based policy π_1, policy π_2, and its logit-customized version π_3 in terms of regret dependence on T. Results in both panels are for Example 1.

introducing a stylized model of dynamic assortment planning. On the theoretical side, we have provided a lower bound on the performance of any consistent policy and showed that this lower bound can be achieved, up to constant terms, when the noise distribution in the utility specification is known, and a product identification condition holds. In particular, we proposed an assortment-based exploration algorithm whose regret scales optimally in the selling horizon T and exhibits the "right" dependence on the number of possible optimal products when said optimality is reached via unilateral deviations in the mean utility vector (e.g., under Luce-type models).

The problem studied in this paper, and outlined in §3, can be viewed as a multiarmed bandit problem by means of the following analogies. First, each *assortment* might constitute an arm; hence, one faces a variant of a multiarmed bandit problem with a combinatorial number of arms. Note though that arm distributions are not independent and not a priori indistinguishable, as is the case with traditional bandit formulations. Second, each *product* might be viewed as an arm; hence, one faces a variant of a multiarmed problem with multiple simultaneous plays where arms are not a priori indistinguishable (due to differences in profit margins). Our main results clearly demonstrate the inefficiency of using standard bandit methods within the former view, while elucidating ways to overcome the obstacles present in the latter view.

On the more practical side, our results suggest how to quantify the "right" amount of information one should collect on consumer preferences, so that revenue loss due to exploration balances with that stemming from errors during exploitation. Our results highlight the importance of limiting information collection by "quickly" identifying and ceasing exploration on products that are unlikely to be members of the optimal assortment.

7.2. Limitations and Future Research

As indicated in §4.1, the lower bound in Theorem 1 can be tightened when $\bar{\mathcal{N}} \neq \tilde{\mathcal{N}}$. The resulting bound, however, might not be proportional to $|\bar{\mathcal{N}}|$, which suggests even tighter bounds might be developed.

Our proposed parameter estimation method is based on MLE; thus, it inherits the associated advantages and shortcomings (e.g., asymptotic efficiency yet potential for small sample bias). One can extend the results in this paper to different estimation procedures as long as consistency is preserved; see Lemma 1 in the proof of Theorem 2. In particular, provided that a guarantee similar to that in Lemma 1 holds, such a result provides finite-sample confidence intervals for the estimation error. It is worth noting that the result does not rely on properties of MLE.

An important extension to our model is considering alternative mean-utility specifications. Studies in fields such as marketing and economics usually postulate that mean utilities are driven by product specific features. In such a setup, the retailer would strive to recover a partworth vector.

Another important area for future research is relaxing assumptions pertaining to the operational environment, especially that of perfect inventory

replenishment. Practical inventory considerations play an important role in settings such as fast fashion and display-based online advertisement, the motivating applications considered in §1. An additional important extension is to consider settings where product prices must be selected as well (for simplicity, assume prices take values in a finite set). In this regard, the work of Rusmevichientong and Broder (2012) provides an initial exploration of this possibility, though absent inventory or assortment considerations.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0429.

Appendix A. Proof of Theorem 2

We prove the result in three 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 single-sale 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 $p_i := \{p_i(A_j, \mu)\colon A_j \in \mathcal{A} \text{ s.t. } i \in A_j\}$, for $i \in \mathcal{N}$, and $p := (p_1, \ldots, p_N)$.

Step 1. 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,

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

We will need the following side lemma, whose proof is deferred to Appendix D in the online companion.

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

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

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

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

$$\|\mu - \hat{\mu}_t\|_{A_j} \le \kappa(\epsilon)\|p - \hat{p}_t\|_{A_j}, \tag{9}$$

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

$$\begin{aligned}
&\mathbb{P}\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon,\, T^j(t) \ge n\}\\
&\quad = \mathbb{P}\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon,\, \|p - \hat{p}_t\|_{A_j} \ge \epsilon,\, T^j(t) \ge n\}\\
&\qquad + \mathbb{P}\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon,\, \|p - \hat{p}_t\|_{A_j} < \epsilon,\, T^j(t) \ge n\}\\
&\quad \le \mathbb{P}\{\|p - \hat{p}_t\|_{A_j} \ge \epsilon,\, T^j(t) \ge n\}\\
&\qquad + \mathbb{P}\{\|\mu - \hat{\mu}_t\|_{A_j} > \epsilon,\, \|p - \hat{p}_t\|_{A_j} < \epsilon,\, T^j(t) \ge n\}\\
&\quad \overset{(a)}{\le} \mathbb{P}\{\|p - \hat{p}_t\|_{A_j} \ge \epsilon,\, T^j(t) \ge n\}\\
&\qquad + \mathbb{P}\{\|p - \hat{p}_t\|_{A_j} > \epsilon/\kappa(\epsilon),\, T^j(t) \ge n\}\\
&\quad \le 2\mathbb{P}\{\|p - \hat{p}_t\|_{A_j} \ge \epsilon/\kappa(\epsilon),\, T^j(t) \ge n\}\\
&\quad \le 2\sum_{i \in A_j} \mathbb{P}\{|p_i(A_j,\mu) - \hat{p}_{i,t}| \ge \epsilon/\kappa(\epsilon),\, T^j(t) \ge n\}\\
&\quad \overset{(b)}{=} 2\sum_{i \in A_j} \mathbb{P}\left\{\left|\sum_{s=1}^{t}(Z_i^s - p_i(A_j,\mu))\mathbf{1}\{S_t = A_j\}\right| \ge T^j(t)\epsilon/\kappa(\epsilon),\, T^j(t) \ge n\right\}\\
&\quad \overset{(c)}{\le} 2|A_j|\exp(-c(\epsilon/\kappa(\epsilon))n),
\end{aligned} \tag{10}$$

where (a) follows from (9), (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 $t \ge 1$,

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

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

$$|r(S,\mu) - r(S,\hat{\mu}_t)| \le \|w\|_\infty KC\|\mu - \hat{\mu}_t\|_S. \tag{11}$$

From here, we conclude that

$$\begin{aligned}
r(S^*(\hat{\mu}_t),\mu) &\ge r(S^*(\hat{\mu}_t),\hat{\mu}_t) - \|w\|_\infty KC\|\mu - \hat{\mu}_t\|_{S^*(\hat{\mu}_t)}\\
&\ge r(S^*(\mu),\hat{\mu}_t) - \|w\|_\infty KC\|\mu - \hat{\mu}_t\|_{S^*(\hat{\mu}_t)}\\
&\ge r(S^*(\mu),\mu) - 2\|w\|_\infty KC\|\mu - \hat{\mu}_t\|_{(S^*(\mu)\cup S^*(\hat{\mu}_t))}.
\end{aligned}$$

As a consequence, if

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

then $S^*(\mu) = S^*(\hat{\mu}_t)$, where $\delta(\mu)$ is the minimum (relative) optimality gap (see (13) in the proof of Theorem 1 in the online companion). This means that if the mean utility estimates are uniformly close to the underlying mean utility values, then solving the single-sale problem using estimates returns the same optimal assortment as when solving the single-sale problem with the true parameters. In particular, we will use the following relation:

$$\begin{aligned}
\{S^*(\mu) \ne S^*(\hat{\mu}_t)\} \subseteq \big\{&\|\mu - \hat{\mu}_t\|_{(S^*(\mu)\cup S^*(\hat{\mu}_t))}\\
&\ge (2\|w\|_\infty KC)^{-1}\delta(\mu)r(S^*(\mu),\mu)\big\}.
\end{aligned} \tag{12}$$

Step 3. Let $NO(t)$ denote the event that a nonoptimal assortment is offered to customer t; that is, $NO(t) := \{S_t \ne S^*(\mu)\}$. Define $\xi := (2\|w\|_\infty KC)^{-1}\delta(\mu)r(S^*(\mu),\mu)$. For $t \ge |\mathcal{A}|\lceil \kappa_1 \log T\rceil$ one has that

$$\begin{aligned}
\mathbb{P}\{NO(t)\} &\overset{(a)}{\le} \mathbb{P}\{\|\mu - \hat{\mu}_t\|_{(S^*(\mu)\cup S^*(\hat{\mu}_t))} \ge \xi\}\\
&\le \sum_{A_j \in \mathcal{A}} \mathbb{P}\{\|\mu - \hat{\mu}_t\|_{A_j} \ge \xi,\, T^j(t) \ge \kappa_1 \log T\}\\
&\overset{(b)}{\le} \sum_{A_j \in \mathcal{A}} 2|A_j| T^{-\kappa_1 c(\xi/\kappa(\xi))},
\end{aligned}$$

where (*a*) follows from (12) and (*b*) follows from (10). Considering $\kappa_1 > c(\xi/\kappa(\xi))^{-1}$ results in the following bound for the regret:

$$\begin{aligned}\mathcal{R}^{\pi}(T,\mu) &\le \sum_{t=1}^{T}\mathbb{P}\{NO(t)\}\\ &\le |\mathcal{A}|\lceil\kappa_1\log T\rceil + \sum_{t>|\mathcal{A}|\lceil\kappa_1\log T\rceil}^{T}\ \sum_{A_j\in\mathcal{A}} 2|A_j|T^{-\kappa_1 c(\xi/\kappa(\xi))}\\ &\le |\mathcal{A}|\kappa_1\log T + 2NT^{1-\kappa_1 c(\xi/\kappa(\xi))}\\ &= \lceil N/C\rceil\kappa_1\log T + \bar{K}_1,\end{aligned}$$

where $\bar{K}_1 = 2N$. Setting $\bar{\kappa}_1 = c(\xi/\kappa(\xi))^{-1}$ gives the desired result.

Appendix B. Parameter Estimation for the Probit Model

When idiosyncratic shocks to consumer utility are normally distributed purchase probabilities are given by

$$p_i(S,\mu) = \int_{-\infty}^{\infty}\prod_{j\in S\cup\{0\}\setminus\{i\}}\Phi(x-\mu_j)\,\phi(x-\mu_i)\,dx,$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ correspond to the distribution and density of a standard normal random variable, respectively. Unfortunately, the integral above does not have a closed form and must be approximately numerically. In our numerical experiments, we approximate such an integral through simulation.

Given the empirical probabilities $\hat{p}_i(S)$, $S\in\mathcal{A}$, the average log-likelihood (LL) associated with μ is

$$\mathrm{LL}(\mu) = \sum_{i\in S\cup\{0\}}\hat{p}_i(S)\log p_i(S,\mu).$$

Because purchase probabilities cannot be computed exactly, we replace $p_i(S,\mu)$ with its simulated counterpart. In MLE, we look for the value of μ that maximizes LL. For that, we check the first-order conditions

$$\frac{\partial \mathrm{LL}(\mu)}{\partial\mu_i} = \sum_{j\in S\cup\{0\}}\hat{p}_j(S)\frac{1}{p_j(S,\mu)}\frac{\partial p_j(S,\mu)}{\partial\mu_i} = 0,\quad i\in S.$$

One can solve the system above using the Newton–Raphson method. However, such a method requires access to the Jacobian and Hessian of LL, which are not available in closed form. In our numerical experiments, we approximate these quantities numerically. The Jacobian of LL requires approximating

$$\frac{\partial p_i(S,\mu)}{\partial\mu_i} = \int_{-\infty}^{\infty}x\prod_{j\in S\cup\{0\}\setminus\{i\}}\Phi(x-\mu_j)\phi(x-\mu_i)dx - \mu_i p_i(S,\mu),\quad i\in S,$$

and

$$\frac{\partial p_j(S,\mu)}{\partial\mu_i} = -\int_{-\infty}^{\infty}\prod_{h\in S\cup\{0\}\setminus\{i,j\}}\Phi(x-\mu_h)\phi(x-\mu_i)\phi(x-\mu_j)dx,\quad j,i\in S,\ i\neq j.$$

Derivatives for $p_0(S,\mu)$ follow from the fact that purchase probabilities sum up to one. The effort required to approximate the integrals above is essentially that of approximating the purchase probabilities. (In our numerical experiments, we compute both of them simultaneously.) One can show that the same holds true for computing the Hessian of LL. For example, one has that

$$\frac{\partial^2 p_i(S,\mu)}{\partial^2\mu_i} = \int_{-\infty}^{\infty}x^2\prod_{j\in S\cup\{0\}\setminus\{i\}}\Phi(x-\mu_j)\phi(x-\mu_i)\,dx - 2\mu_i\frac{\partial p_i(S,\mu)}{\partial\mu_i} - (\mu_i^2+1)p_i(S,\mu),\quad i\in S;$$

thus, one can approximate the Hessian, Jacobian, and log-likelihood functions efficiently using Monte Carlo simulation.

In our experiments, we used a sample of size 50,000 to approximate each integral and used importance sampling to enhance the precision of our approximation; in particular, we use approximations of the normal CDF to approximate the integrals above as a sum of properly weighted components. We used incumbent parameter estimates (those computed in the previous estimation cycle) as a starting point for the Newton–Raphson method and used convergence tolerance parameter of 10^{-6}.

References

Anderson S, de Palma A, Thisse J-F (1992) *Discrete Choice Theory of Product Differentiation* (MIT Press, Cambridge, MA).

Araman V, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5): 1169–1188.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.

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

Daganzo C (1979) *Multinomial Probit: The Theory and Its Applications to Demand Forecasting* (Academic Press, New York).

Farias V, Madan R (2011) The irrevocable multi-armed bandit problem. *Oper. Res.* 59(2):383–399.

Farias V, Van Roy B (2010) Dynamic pricing with a prior on market response. *Oper. Res.* 58(1):16–29.

Fisher M, Vaidyanathan R (2009) An algorithm and demand estimation procedure for retail assortment optimization. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Gallego G, van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* 50(8):999–1020.

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

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

Honhon D, Gaur V, Seshadri S (2009) Assortment planning and inventory decisions under stock-out based substitution. *Oper. Res.* 58(5):1364–1379.

Honhon D, Ulu C, Alptekinoglu A (2012) Learning consumer tastes through dynamic assortments. *Oper. Res.* 60(4):833–849.

Hopp W, Xu X (2008) A static approximation for dynamic demand substitution with applications in a competitive market. *Oper. Res.* 56(3):630–645.

Kök AG, Fisher ML, Vaidyanathan R (2008) Assortment planning: Review of literature and industry practice. Agrawal N, Smith SA, eds. *Retail Supply Chain Management: Quantitative Models and Empirical Studies* (Springer, New York), 99–154.

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

Lim A, Shanthikumar J (2007) Relative entropy, exponential utility, and robust dynamic pricing. *Oper. Res.* 55(2):198–214.

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

Megiddo N (1979) Combinatorial optimization with rational objective functions. *Math. Oper. Res.* 4(4):414–424.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58(5):527–535.

Rusmevichientong P, Broder J (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.

Rusmevichientong P, Shen Z-JM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

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

Train K (2009) *Discrete Choice Methods with Simulation* (Cambridge University Press, New York).

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