Dynamic Assortment with Demand Learning for Seasonal Consumer Goods

Felipe Caro [*] Jérémie Gallien [†]

January 10, 2005

Abstract

Companies such as Zara and World Co. have recently implemented novel product development processes and supply chain architectures enabling them to make many more product design and assortment decisions during the selling season, when actual demand information becomes available. How should such retail firms modify their product assortment over time in order to maximize overall profits for a given selling season? We formulate this problem as a finite horizon multiarmed bandit with several plays per stage, Bayesian learning and response lag. Our analysis involves the Lagrangian relaxation of weakly coupled dynamic programs, results contributing to the emerging theory of DP duality, and various approximations; it yields a closed-form dynamic index policy capturing the key exploration vs. exploitation trade-off, and associated suboptimality bounds. Numerical experiments suggest that our index policy is near-optimal, and outperforms the greedy policy with passive learning; its relative superiority seems particularly significant in environments with little prior information and long design-to-shelf leadtimes.

[*] MIT Sloan School of Management, Cambridge, MA 02142, fcaro@mit.edu

[†] MIT Sloan School of Management, Cambridge, MA 02142, jgallien@mit.edu

1 Introduction

1.1 Motivation

Long development, procurement, and production lead times resulting in part from a widespread reliance on overseas suppliers have traditionally constrained fashion retailers to make supply and assortment decisions well in advance of the selling season, when only limited and uncertain demand information is available. With only little ability to modify product assortments and order quantities after the season starts and demand forecasts can be refined, many retailers are seemingly cursed with simultaneously missing sales for want of popular products, while having to use markdowns in order to sell the many unpopular products still accumulating in their stores (see Fisher et al. 2000).

Since the late 1980's an industry-wide initiative known as "Quick Response" (see Hammond 1990 for a more detailed description) has focused on attenuating that curse, meeting some success. Leveraging information technologies, improved product designs and manufacturing schemes as well as faster transportation modes, some of its followers have significantly improved the flexibility of their overseas supply networks, thus managing to postpone part of their production until more demand information can be gathered.

Recently however, a few innovative firms including Spain-based Zara, Mango and Japan-based World Co. (sometimes referred to as "Fast Fashion" companies) have gone substantially further, implementing product development processes and supply chain architectures allowing them to make *most* product design and assortment decisions *during* the selling season. Remarkably, their higher flexibility and responsiveness is partly achieved through an increased reliance on more costly local production relative to the supply networks of more traditional retailers. The contrast between these two supply-chain design alternatives seems particularly drastic: Zara's design-to-shelf lead time range for new or modified product is $2-5$ weeks, versus $6-9$ months for a more traditional retailer; in-house production during the season is reported to be approximately 85% for Zara, versus less than 20% for other retailers; Zara manufactures about $11,000$ different products per year (excluding variations in color, size and fabric), compared to only $2,000-4,000$ items for key competitors; only $15-20\%$ of Zara's sales are typically generated at marked-down prices, compared with $30-40\%$ for most of its European peers, furthermore the percentage discount for their marked-down items was estimated as roughly half of the 30% average for other European apparel retailers (see Ghemawat and Nueno 2003).

At the operational level, leveraging the ability to introduce and test new products once the season has started motivates a new and important decision problem, which seems key to the success of these fast-fashion companies: given the constantly evolving demand information available, which products should be included in the assortment at each point in time? Figure 1 provides a conceptual representation of this operational challenge: in each period over a finite horizon (representing the whole season T), the retailer must decide the subset (N) of products that will be offered from a larger set (S) of all retail introduction candidates. As sales occur, the retailer gathers new demand information about each particular product that was included in the latest assortment, which may be combined with prior historical demand information to select the next assortment – although not shown in Figure 1 for simplicity, it must be noted that the assortment decision can typically only be implemented after a lag (ℓ) corresponding to the design-to-shelves lead time.

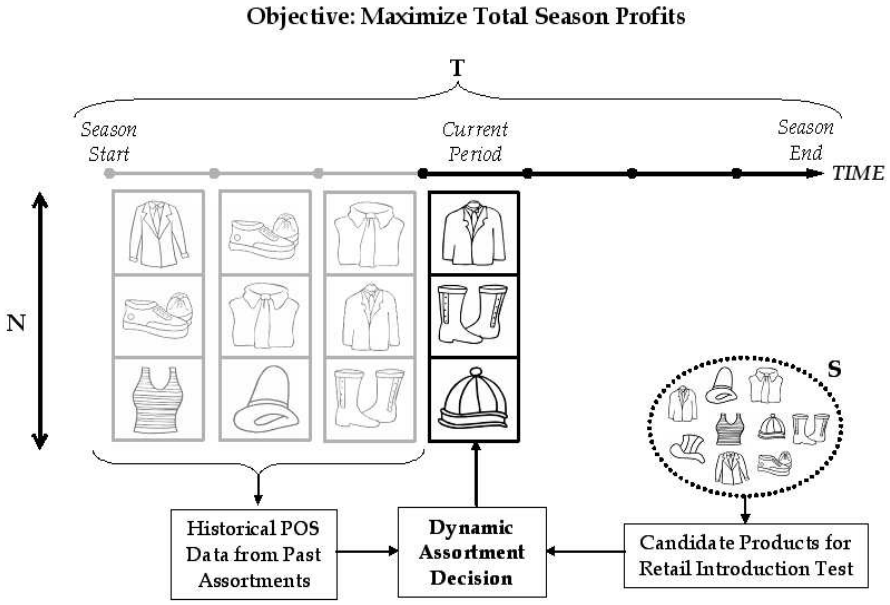
Objective: Maximize Total Season Profits
T
Season Start
Current Period
Season End
TIME
N
S
Historical POS Data from Past Assortments
Dynamic Assortment Decision
Candidate Products for Retail Introduction Test

Figure 1: The dynamic assortment problem.

The problem just described seems challenging, in part because it relates to the classical trade-off known as exploration versus exploitation: in each period the retailer must choose between including in the assortment products for which he has a “good sense” that they are profitable (exploitation), or products for which he would like to gather more demand information (exploration); that is, he must decide between being “greedy” based on his current information, or try to learn more about product demand (which might be more profitable in the future). In addition, this problem poses itself frequently, for a high number of products, and involves a large amount of data. Incidentally, we only have limited understanding at present of how these companies actually solve this dynamic assortment problem in practice, and all studies focusing on fast-fashion companies we are aware of (e.g. Fisher et al. 2000, Ghemawat and Nueno 2003, and Ferdows et al. 2003) only describe this challenge in qualitative terms. Our main objective in the present paper is thus to develop and analyze a quantitative optimization model capturing the main features of this dynamic assortment problem, with a view towards eventually creating an operational decision support system.

The remainder is organized as follows: the next subsection §1.2 is an overview of the relevant literature, and we present and discuss our mathematical model in §2. Section §3 is devoted to the analysis, and contains in particular the derivation of our proposed dynamic index policy as well as an upper performance bound. An extensive simulation study is reported in §4, Section §5 focuses on possible model extensions, and Section §6 contains our concluding remarks. All proofs are given in the Appendix.

1.2 Literature Review

We first discuss papers focusing on assortment problems. A first subset is found in the Marketing literature, where several studies, typically motivated by supermarkets, consider static assortment problems formulated as deterministic nonlinear optimization models in which the demand of a product depends on the allocated shelf space, and the overall space available is a limited resource. A classical example in this vein is Bultez and Naert (1988); for more recent work see Kök and Fisher (2004) and references therein. In the Operations Management literature, van Ryzin and Mahajan (1999) and Smith and Agrawal (2000) are two papers also considering static assortment problems, but with a stochastic demand model and static product substitution. That is, customer demand reflects aggregated substitution effects depending on the initial assortment decision, but not on the actual inventory levels observed by individual customers once arrived to the store. In contrast, Mahajan and van Ryzin (2001) describe a more detailed assortment model capturing dynamic substitutions, that is substitutions due to stockouts experienced by individual customers, and analyze it using sample path methods.

None of the papers just cited considers demand learning, and accordingly the assortment problems they investigate are static, not dynamic. Presumably because of the relative novelty of fast fashion companies, we have in fact not found in the literature any dynamic assortment model explicitly described as such. While papers underlying the quick response initiative described in the previous section do place much emphasis on learning and exploiting early sales information, the demand information acquired over time is primarily exploited by the manufacturer to make better ordering and production quantities decisions, as opposed to product design or assortment decisions; the seminal paper by Fisher and Raman (1996), motivated by skiwear manufacturer Sport Obermeyer, presents a two-stage stochastic programming model in which initial production commitments are made before any sales occur, but further production decisions are made in a second stage after receiving some customer orders and refining total sales forecasts. Note that the trade-off between exploration and exploitation is not present in the problem just described, where in fact the optimal policy consists of postponing the ordering of products for which demand is most uncertain.

As may already be clear from Figure 1, our work is closely related to the multiarmed bandit problem, which has been extensively studied in the literature (see Berry and Fristedt 1985, Kumar 1985, and Brezzi and Lai 2002). In the discrete-time version, a player chooses N arms to pull out of a total of S available in each one of T periods. Whenever pulled, each arm generates a stochastic reward following an arm-dependent distribution, which is initially unknown but can be inferred with experience as successive rewards are observed; the player's objective is to maximize total reward over the game horizon. In the present paper, pulling an arm is equivalent to including in the assortment the product to which it is associated.

A remarkable result for the multiarmed bandit problem is due to Gittins (see Gittins and Jones 1974, and Gittins 1979). It involves the definition of the so-called Gittins' index for each arm s, equal to the lump sum that would make the player indifferent between retiring or playing arm s individually, ignoring the other arms (cf. Bertsekas 2001, Vol. II, pp. 60-70). Assuming independent arms, infinite horizon ($T = \infty$), exactly one arm pulled in each stage ($N = 1$) and a discount factor strictly smaller than one, the optimal policy is to play in each stage the arm with the highest Gittins' index. Among several subsequent extensions to Gittins' result we highlight the work on restless bandits by Whittle (1988), whose analysis is related to ours in that it also involves Lagrangian multipliers.

In the finite horizon case ($T < \infty$), it is known that Gittins' index policy is in general not optimal. Relevant references include the book by Berry and Fristedt (1985), which presents analytical techniques similar to the ones we use in the next sections. Lai (1987) develops a policy (or allocation rule) based on the calculation of an upper confidence bound for each arm (which can also be seen as an index). For the case with multiple plays per stage, Anantharam et al. (1987) consider the frequentist version of the problem, where the objective is to minimizing regret. While the allocation rule they propose is asymptotically efficient, it does not seem directly applicable to our problem because it requires a setup phase of at least $S \times N$ periods in order to have N initial observations per arm, and does not allow for a response lag (stemming in our context from the design-to-shelf lead times).

Finally, the paper by Bertsimas and Mersereau (2004), which focuses on an adaptive sampling problem, is the reference that is methodologically closest to our work – their model is a finite horizon version of the multiarmed bandit problem, and their analysis also involves Lagrangian decomposition. However, they do not consider response lags and assume a Beta-Bernoulli learning model, while we use the Gamma-Poisson model. Besides, in contrast to that paper we provide a suboptimality bound for the policy we derive.

2 Model Definition and Discussion

We now formulate our dynamic assortment model in §2.1, then discuss its applicability and justify our assumptions in §2.2. Throughout the remaining of the paper all symbols in boldface represent vectors, subscripts represent the components of a vector, and superscripts represent elements in a sequence.

2.1 Model Definition

2.1.1 Supply

Consider a retailer selling products in a store during a limited selling season. The set of all products that the retailer may potentially sell is denoted by $\mathcal{S} = \{1, 2, \ldots, S\}$; this set includes both the products already available when the season starts and all the variants and new products that may be designed during the season. The net margin r_s of product $s \in \mathcal{S}$ is assumed to be exogenously given, positive, and constant. In line with the features of fast fashion companies described in §1, we assume that the selling season can be divided into T periods, and that at the beginning of each of these periods the product assortment in the store may be revised; time is counted backwards and denoted by the index t (thus representing the number of periods remaining before the end of the season). Due to design, production and distribution delays, there may be a lag ℓ between the period t when an assortment decision is made and the period $t - \ell$ at which this assortment is actually implemented in the store (this also occurs at the beginning of period $t - \ell$). However, our approach in this paper is to perform our analysis in subsections §3.1 to §3.5 under the assumption that the lag is zero ($\ell = 0$), then adapt the policy and performance upper bound we derive to the case with a positive lag $\ell > 0$ in subsection §3.6.

The store's limited shelf space (or desire to limit in-store product variety due to other considerations) is captured by the constraint that the assortment in each period may include at most N different products out of the S available; we are thus implicitly assuming that all products require the same shelf space. We also assume a perfect inventory replenishment process during each assortment period, so that there are no stockouts or lost sales. Consequently, in our model, realized sales equal total demand, and we focus for each product on assortment inclusion or exclusion as opposed to order quantity. Finally, holding costs are ignored in our formulation.

2.1.2 Demand

In our model, demand for each product in the assortment is exogenous and stationary but stochastic, and we do not capture substitution effects. Specifically, we assume that customers willing to buy one unit of each product s in the assortment arrive to the store according to a Poisson process with an unknown but constant rate γ_s. That is, the underlying arrival rate γ_s is assumed to remain constant throughout the entire season, but the resulting actual demand for product s may only be observed in the periods when that product is included in the assortment. In addition, the arrival processes corresponding to different products are assumed to be independent.

We adopt a standard Gamma-Poisson Bayesian learning mechanism (also used for instance in Aviv and Pazgal 2002): The underlying demand rate γ_s for each product s is initially unknown to the retailer, however he starts each period with a prior belief on the value of that parameter represented by a Gamma distribution with shape parameter m_s and scale parameter α_s (m_s and α_s must be positive, and m_s is assumed to be integer[1]). Redefining time units if necessary, we can assume with no loss of generality that the length of each assortment period is 1; the predictive demand distribution under that belief for selling n_s units of product s in the upcoming assortment period is then given by:

[1] The model can be extended to consider non integer values of m_s but the binomial coefficient in equation (1) must be replaced with the corresponding $\Gamma(\cdot)$ terms, and the interpretation as a negative binomial (to be given) would not be valid.

$$\Pr(n_s) = \binom{n_s + m_s - 1}{m_s - 1} \Big(\frac{1}{\alpha_s + 1}\Big)^{n_s} \Big(\frac{\alpha_s}{\alpha_s + 1}\Big)^{m_s}, \tag{1}$$

which is a negative binomial distribution with parameters m_s and $\alpha_s(\alpha_s{+}1)^{-1}$. When necessary, we will write $n_s(m_s, \alpha_s)$ to make the parameter dependence explicit. If now product s is included in the assortment and n_s actual sales are observed in that period, it follows from Bayes' rule that the posterior distribution of γ_s has a Gamma distribution with shape parameter $(m_s + n_s)$ and scale parameter $(\alpha_s + 1)$. In summary, for each product s and period t, the parameters of the prior distribution on γ_s are updated as follows:

$$(m_s, \alpha_s) \longrightarrow \left\{ \begin{array}{ll} (m_s + n_s, \alpha_s + 1) & \text{If product } s \text{ is in the assortment and } n_s \text{ sales} \\ & \text{are observed during period } t \\ (m_s, \alpha_s) & \text{If product } s \text{ is not in the assortment} \\ & \end{array} \right. . \tag{2}$$

The intuition for the update procedure (2) is straightforward: the retailer initially believes that m_s units of product s will sell in α_s periods on average, so that the expected sales rate is $\mathbb{E}[\gamma_s] = m_s/\alpha_s$; after observing then n_s sales of product s he subsequently expects $(m_s + n_s)$ units of product s to sell in $(\alpha_s + 1)$ periods. Note that the retailer's beliefs become more accurate with the number of observed sales, since the variance of the prior is $\mathbb{V}[\gamma_s] = m_s/\alpha_s^2$ so that its coefficient of variation equals $1/\sqrt{m_s}$.

2.1.3 Dynamic Programming Formulation

Given the discrete and sequential character of our problem, the natural solution approach is dynamic programming (DP); the state at time t is given in our model by the parameter vector $\mathbf{I}^t = (\boldsymbol{m}, \boldsymbol{\alpha})$, which summarizes all relevant information including past assortments and observed sales[2] (cf. Bertsekas 2001, Vol I. Chapter 6). In each period, the decision to include product s in the assortment or not can be represented by a binary variable $u_s \in \{0, 1\}$, where $u_s = 1$ means that product s is included. The set $\mathcal{U}$ of all feasible assortments (i.e. the control space) corresponding to the shelf space constraint described above can then be defined as $\mathcal{U} = \big\{\boldsymbol{u} \in \{0,1\}^S : \sum_{s=1}^S u_s \leq N\big\}$.

The optimal profit-to-go function $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$ given state $(\boldsymbol{m}, \boldsymbol{\alpha})$ and t remaining periods must then satisfy the following Bellman equation:

$$J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = \max_{\substack{\boldsymbol{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big], \tag{3}$$

where $\boldsymbol{v} \cdot \boldsymbol{u}$ represents the componentwise product of two vectors, and the terminal condition is $J_0^*(\boldsymbol{m}, \boldsymbol{\alpha}) = 0$ for all states; the expectation $\mathbb{E}_{\boldsymbol{n}}[\cdot]$ is with respect to the product demand vector $\boldsymbol{n}$ with distribution $\prod_{s=1}^S \Pr(n_s)$, where $\Pr(n_s)$ is given by equation (1).

[2] For ease of notation, we omit the dependence of $\boldsymbol{m}$ and $\boldsymbol{\alpha}$ on t.

Note that the only link between consecutive periods in this model is the information acquired about demand, and that different products are only coupled at a given period through the shelf space constraint $\sum_{s=1}^{S} u_s \leq N$ (clearly $S > N$, otherwise the retailer would always include all available products in the assortment); this type of problem is known as a *weakly coupled* DP. Observe also that the summation on the right hand side of (3) includes the immediate expected profit associated with each product and represents the exploitation component, while the expectation term that follows captures the future benefits from exploration.

2.2 Model Discussion

This subsection begins with a discussion of the model realism grounded in a potential application to the company Zara, and ends with comments on what we believe to be our three most salient assumptions (independent products, no lost sales and stationary demand).

At Zara, assortment periods (i.e. the time between two consecutive assortment decisions) seem to correspond to one week (Ghemawat and Nueno 2003), and the length T of the whole selling season thus falls between 12 and 24 periods (Zara has only two seasons Spring/Summer and Fall/Winter); incidentally the assumption that all periods have equal length can easily be relaxed in our model. A typical Zara store is divided into three essentially independent sections (Women, Men and Children), and each section is further divided into categories. As an example, the categories for the Women section include: lower garment, upper garment, underwear, footware, accessories, and suits. Within a category, the number N of different products seems to roughly vary between 20 and 60.[3] These numbers do not take into account differences in size, color and fabric however; more generally in our model a product may represent an individual stock keeping unit (SKU) or a family of related SKUs (e.g. different sizes or colors aggregated). Our shelf space constraint may reflect the amount of space available for each section and category driven by the physical layout of actual stores, but it may also result from deliberate operational or marketing decisions. The assumption that all products require the same shelf space, which is somewhat analogous to the equal capacity requirements assumed in the Sport Obermeyer study (Fisher and Raman 1996), could be relaxed at the cost of increased model complexity. We note however that this assumption does seem realistic in the case of a separate application of our model to each individual category as suggested above, since products within the same category indeed have similar shapes.

Based on figures reported in Ghemawat and Nueno (2003), we estimate the total number S of potential products in a category for the whole season to be of the order of T times N, or 720, for Zara. While our formulation assumes that the corresponding set $\mathcal{S}$ is known at the beginning of the season and does not change further on, in any practical implementation new products may be added to $\mathcal{S}$ as they become available; at Zara, new products are indeed designed during the selling season based on customer feedback reported by store managers.

[3] These observations are based on information provided on the company's website as well as visits to various stores by the first author.

We now focus on what we think are the three most salient model assumptions:

Independent Products In contrast with most of the (static) assortment studies discussed in the literature review §1.2, our basic model ignores all product substitution and complementarity effects. In support of that assumption, the absence of dynamic substitutions due to stockouts is consistent with the perfect inventory replenishment process we assume (see below). However, this also saliently implies that the underlying customer demand for all products offered is completely independent from the other products constituting the assortment, a requirement clearly damaging realism. In practice, there may be significant substitution effects between products from the same category (e.g. two slightly different shirts may cannibalize each other when both introduced in the assortment) and/or complementarity effects between products from different categories (e.g. matching lower garments and upper garments). From that standpoint, the demand learning model we use is relatively coarse; we observe however that the current set of available tools for inferring demand dynamically in the presence of substitution effects is very limited (see discussion in §5).

No Lost Sales For the sake of model simplicity and tractability, we assume that the inventory replenishment process (which we do not describe) is perfect, in the sense that there are no lost sales under any assortment; we may thus focus on assortment decisions as opposed to other operational issues such as inventory ordering and service levels. In practice, Zara replenishes its stores twice a week and seems to indeed experience fewer lost sales than other more traditional retailers (Ghemawat and Nueno 2003). However, that assumption is clearly very strong, and in fact Zara deliberately introduces some lost sales in order to generate a feeling of "scarcity" among consumers (cf. Ferdows et al. 2003, p. 66), a phenomenon which is not captured by our model where demand is exogenous (see below). In this setting, ignoring holding costs seems consistent with the assumption that inventory levels are exogenous as described just above. More generally, we observe that holding costs are often ignored in the case of seasonal products (see, for instance, Aviv and Pazgal 2002).

Constant Demand Rates In practice, the demand rate for fashionable products usually follows some asymmetric "bell shaped" curve over time. However, our model assumes that it is constant, mostly for tractability reasons – this is key in particular to the fact that all relevant state information is captured by the pair $(\boldsymbol{m}, \boldsymbol{\alpha})$. While demand stationarity may be a particularly strong assumption in some settings, we observe that it is consistent with some of our other assumptions. Specifically, an important reason why demand nonstationarity may arise in practice is the use of dynamic pricing, but we assume that prices remain constant throughout the season (the margin r_s of every product s is fixed); note that this is partly justified by the figures reported in §1 showing that fast-fashion retailers rely less frequently on markdown policies, and that when they do so their price markdowns are also lower. Likewise, another important driver for demand nonstationarity may be stockouts, but these do not occur in our model since we assume a perfect replenishment process. Finally, our model can be easily generalized to the case where all demand rates are multiplied by the same deterministic time-varying factor, since this is equivalent to having periods of different lengths.

While we consider the above three assumptions to be quite strong, our approach is partly motivated by the belief that the closed-form policy they allow to derive (in §3) constitutes a useful starting point for designing heuristics or developing extensions in more complex environments, as discussed in Section §5. For example, we describe in §5.1 a heuristic procedure for capturing substitution effects that is based on the analysis of our basic model.

3 Analysis

3.1 Properties of the Profit-to-go Function

In this subsection we state two simple and intuitive properties of the profit-to-go function of our assortment problem. The first result confirms the intuition that the expected profit should increase if the prior beliefs are higher (i.e. the expected sales rate for a product is larger), or more accurate (i.e. the coefficient of variation is smaller); this follows mathematically from the fact that the negative binomial (1) is stochastically increasing in m_s and decreasing in α_s, so that the random vector $\boldsymbol{n}(\boldsymbol{m}, \boldsymbol{\alpha})$ inherits the same properties[4] (see Ross 1996). This is formalized by the following Lemma, which will be used later on to establish further results:

Lemma 1 *If $\boldsymbol{m}'' \geq \boldsymbol{m}'$ and $\boldsymbol{\alpha}'' \leq \boldsymbol{\alpha}'$, then $J_t^*(\boldsymbol{m}'', \boldsymbol{\alpha}'') \geq J_t^*(\boldsymbol{m}', \boldsymbol{\alpha}')$, for all t. The last inequality is strict if any of the former is strict.*

The second result shows that dynamic assortment will do no worse on average than implementing the optimal static assortment at the beginning of the season, and no better than the optimal assortment under perfect information (see Aviv and Pazgal 2002 p. 25 for a comparable result):

Lemma 2 *For every state $(\boldsymbol{m}, \boldsymbol{\alpha})$ and period t:*

$$\max_{\sum_{s=1}^{S} u_s \leq N} \sum_{s=1}^{S} r_s \mathbb{E}[\gamma_s] u_s \;\leq\; \frac{J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})}{t} \;\leq\; \mathbb{E}_{\boldsymbol{\gamma}(\boldsymbol{m},\boldsymbol{\alpha})}\Big[\max_{\sum_{s=1}^{S} u_s \leq N} \sum_{s=1}^{S} r_s \gamma_s u_s \Big], \qquad (4)$$

where the s-th component of random vector $\boldsymbol{\gamma}(\boldsymbol{m}, \boldsymbol{\alpha})$ follows a Gamma distribution with parameters (m_s, α_s).

Incidentally, the difference between $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$ and the upper bound of (4) times t is known as the Bayes risk or regret (see Lai 1987, p. 1092). It can be further shown that $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})/t$ is monotonically increasing in t, defining a bounded monotone sequence which therefore converges when

[4] For two vectors we write $\boldsymbol{v}_1 \geq \boldsymbol{v}_2$ to denote that the given inequality holds componentwise.

the planning horizon goes to infinity. Empirical evidence and intuition suggest that it converges to the right hand side of (4); we have not attempted to prove that conjecture however, since we are primarily motivated here by situations where the opportunity to learn about demand is severely limited by a finite selling horizon.

3.2 The Dual Dynamic Program

The optimal dynamic assortment policy may conceptually be derived from the dynamic programming equation (3). The associated computational requirements are overwhelming however, except for very small problem instances; even with a truncated state space, only calculating the expectation in the right hand side of equation (3) (which constitutes in fact the objective function of a discrete nonconcave optimization problem for which there is currently no standard solution method) is an intensive numerical task. Therefore, we do not aim to solve the dynamic assortment problem optimally; our motivation is rather to find a simple near-optimal policy that can be easily implemented in practice.

The approximate solution method described in this subsection is based on Lagrangian relaxation and the decomposition of weakly coupled dynamic programs. While the literature reporting successful applications of this methodology is rather recent (see Castañon 1997, Hawkins 2003, Bertsimas and Mersereau 2004, and references therein), the underlying concepts involved are similar to those of the well-established theory of duality for general nonlinear optimization problems (see for instance Bertsekas 1999).

Specifically, we relax the shelf space constraint, which leads to the definition of *dual policies* that will later prove to be useful in finding near-optimal *primal* policies and upper bounds for the optimal profit-to-go: Let $\lambda_t(\boldsymbol{m}, \boldsymbol{\alpha})$ denote any function associated with period t that maps the state space into the set of nonnegative real values; we define a *dual policy* to be any vector a functions $\boldsymbol{\lambda_t} = (\lambda_t(\cdot), \lambda_{t-1}(\cdot), \dots, \lambda_1(\cdot))$.

For any dual policy $\boldsymbol{\lambda_t}$ and any initial state $(\boldsymbol{m}, \boldsymbol{\alpha})$, the corresponding profit-to-go is obtained by solving the *dual dynamic program* given by:

$$H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{m}, \boldsymbol{\alpha}) = N\lambda_t(\boldsymbol{m}, \boldsymbol{\alpha}) + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^{S} \Big(r_s \frac{m_s}{\alpha_s} - \lambda_t(\boldsymbol{m}, \boldsymbol{\alpha}) \Big) u_s + \mathbb{E}_{\boldsymbol{n}} \big[H_{t-1}^{\boldsymbol{\lambda_{t-1}}}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u}) \big], \quad (5)$$

with $H_0^{\boldsymbol{\lambda_0}}(\boldsymbol{m}, \boldsymbol{\alpha}) = 0 \ \forall (\boldsymbol{m}, \boldsymbol{\alpha})$. In words, a dual policy gives the price of a unit of shelf space for each period and each possible state.

A dual policy $\boldsymbol{\lambda_t}$ is *optimal* if it minimizes the right hand side of (5) for any initial state. In line with standard dynamic programming theory, we recursively define $\lambda_t^*(\mathbf{m}, \boldsymbol{\alpha})$ to be the smallest solution of the following dual problem:

$$H_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = \min_{\lambda_t \geq 0} \quad N\lambda_t + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^{S} \Big(r_s \frac{m_s}{\alpha_s} - \lambda_t \Big) u_s + \mathbb{E}_{\boldsymbol{n}} \big[H_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u}) \big], \quad (6)$$

and it can be verified through straightforward induction that the policy $\boldsymbol{\lambda_t^*}$ is indeed optimal.

The following proposition is the main result in this subsection; a similar result for *open-loop dual policies* (to be defined shortly) can be found in Hawkins (2003).

Proposition 1 ***(Weak DP Duality)*** *For any period t, any dual policy $\boldsymbol{\lambda_t}$ and any given initial state $(\boldsymbol{m}, \boldsymbol{\alpha})$: $J_t^*(\mathbf{m}, \boldsymbol{\alpha}) \leq H_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) \leq H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{m}, \boldsymbol{\alpha})$.*

As in classical duality theory, an interesting theoretical question is to determine if the first inequality in Proposition 1 ever holds as an equality; this question is partly resolved by the following proposition:

Proposition 2 ***(Strong DP Duality)*** *Consider the following parametric function:*

$$f_\tau(\boldsymbol{m}', \boldsymbol{\alpha}'; C) = \max_{\substack{\mathbf{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s = C}} \sum_{s=1}^S r_s \frac{m'_s}{\alpha'_s} u_s + \mathbb{E}_{\boldsymbol{n}}\big[J^*_{\tau-1}(\boldsymbol{m}' + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha}' + \boldsymbol{u})\big] \tag{7}$$

If $f_\tau(\boldsymbol{m}', \boldsymbol{\alpha}'; C)$ is concave in C for all $\tau = t, \ldots, 1$ and states $(\boldsymbol{m}', \boldsymbol{\alpha}')$ reachable from $(\boldsymbol{m}, \boldsymbol{\alpha})$ in period τ, then $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = H_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$.

In contrast with (3), the parametric function defined by (7) requires the shelf space constraint of the current period to be satisfied as an equality; the shelf space constraints for the subsequent periods remain unaltered however. It can be shown (see the Appendix) that $f_t(\mathbf{m}, \boldsymbol{\alpha}; C)$ is strictly increasing in C reflecting the fact that the retailer can only do better given additional shelf space, and $f_t(\boldsymbol{m}, \boldsymbol{\alpha}; N) = J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$.

Except when $t = 1$, the condition required by Proposition 2 may seem restrictive and difficult to verify. While finding weaker or simpler conditions is the matter of future research, we have still found instances that provably satisfy the one stated in Proposition 2, and we have also found a counter-example showing that strong duality does not hold in general absent such a condition: For $t = 2$, $S = 2$, $N = 1$, $m_1 = 44$, $m_2 = 4$, $\alpha_1 = 10$, and $\alpha_2 = 1$, it is easy to verify that $f_t(\boldsymbol{m}, \boldsymbol{\alpha}; C)$ is not concave in C and $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) < H_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$. As an interesting observation, Proposition 2 does apply for any other value of m_1, keeping the other parameters constant. Moreover for $C = 1$ and $m_1 \leq 44$ the optimal action in the right hand side of equation (7) is to include product 2, but the optimal choice switches to product 1 when $m_1 > 44$. We have observed that the non-concavity of (7) always comes in hand with a similar discrete change in the optimal action of the corresponding parametric optimization problem. However, the reverse is not true: parameter values at which the optimal action changes do not imply $f_t(\boldsymbol{m}, \boldsymbol{\alpha}; C)$ being non-concave.

More generally, both our intuition and (limited) empirical observations suggest that the cases where the parametric profit-to-go defined by (7) is non-concave are somewhat pathological, and correspond to situations when both S and NT are small and some of the initial beliefs have a high variance. In those cases, marginally increasing the value of the shelf-space parameter C from a certain level may suddenly allow to access both exploration and exploitation modes and result in a higher marginal gain than the same increase from a smaller value of C, when only exploitation makes sense. Because our subsequent analysis relies on an approximate solution to the dual DP (6), our overall error will be the sum of the duality gap and an approximation error. Proposition 2 and this discussion thus suggest that the latter error term will dominate in most cases of practical interest.

3.3 Open-loop Dual Policies

Solving the dual DP problem given by equation (6) seems just as hard as solving the original primal problem (3), motivating further simplifications. Specifically, we now restrict our attention to *open-loop dual policies*[5], in which the shadow price on shelf space is constant across all states for each period; formally an open-loop dual policy $\boldsymbol{\lambda}$ is a constant vector $(\lambda_t, \lambda_{t-1}, \ldots, \lambda_1)$, rather than a vector of functions. In the following, we will refer to the profit-to-go corresponding to an open-loop dual policy $\boldsymbol{\lambda}$ as $H_t^{\boldsymbol{\lambda}}(\cdot)$ instead of the previous notation $H_t^{\boldsymbol{\lambda_t}}(\cdot)$. The next Lemma shows that with open-loop policies the dual DP decomposes into S single-product subproblems:

Lemma 3 *Consider an open-loop dual policy $\boldsymbol{\lambda} = (\lambda_t, \lambda_{t-1}, \ldots, \lambda_1)$, then the profit-to-go can be written as:*

$$H_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\alpha}) = N \sum_{\tau=1}^{t} \lambda_\tau + \sum_{s=1}^{S} H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) \tag{8}$$

where:

$$H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) = \max \left\{ \underbrace{r_s \frac{m_s}{\alpha_s} - \lambda_t + \mathbb{E}_{n_s}\Big[H_{t-1,s}^{\boldsymbol{\lambda}}(m_s + n_s, \alpha_s + 1)\Big]}_{u_s=1}, \underbrace{H_{t-1,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)}_{u_s=0} \right\} \tag{9}$$

The single-product subproblem defined by (9) is equivalent to a two-armed bandit in which one arm provides a stochastic (unknown) reward, while the other is deterministic and provides in each period t a reward equal to λ_t. It is clear from (9) that for any fixed state (m_s, α_s), $H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ in nondecreasing with t . Also, it can be shown that $H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ is a convex and piecewise linear function of $(\lambda_t, \ldots, \lambda_1)$, and the proof of Lemma 1 can be repeated replacing $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$ with $H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$, establishing the same monotonicity property with respect to m_s and α_s.

We now focus on the single-product subproblem and characterize its solution; the following properties are insightful and can be used to reduce numerical computations. For any open-loop dual policy $\boldsymbol{\lambda}$, let $A_{t,s}^{\boldsymbol{\lambda}}$ be the set of all states (m_s, α_s) such that it is optimal to include product s in the assortment in period t (i.e. $u_s = 1$ is optimal in (9)), and define $B_{t,s}^{\boldsymbol{\lambda}}$ as its complement (e.g., the stopping set in period t). The next Proposition shows that $A_{t,s}^{\boldsymbol{\lambda}}$ is a connected set which is separated from $B_{t,s}^{\boldsymbol{\lambda}}$ by a strictly increasing threshold function of m_s.

Proposition 3 *Let $\lambda_t > 0$ $\forall t$. For each period t there exists a strictly increasing function $\beta_{t,s}^{\boldsymbol{\lambda}}(\cdot)$ such that at state (m_s, α_s) the optimal policy for the single-product subproblem (9) is: $u_s = 1 \iff \alpha_s \leq \beta_{t,s}^{\boldsymbol{\lambda}}(m_s)$*

[5] Open-loop and close-loop are standard concepts in DP theory (see Bertsekas 2001). Castañon (1997) calls the closed-loop policies *stochastic multipliers* and the open-loop policies *deterministic multipliers.*

The next Proposition shows that the stopping sets decrease when the corresponding shadow prices on shelf space increase:

Proposition 4 *If $\lambda_t \leq \lambda_{t-1}$, then $B^{\boldsymbol{\lambda}}_{t,s} \subseteq B^{\boldsymbol{\lambda}}_{t-1,s}$.*

Note that the inclusion of the stopping sets are not reverted when $\lambda_t > \lambda_{t-1}$, since the threshold functions $\beta^{\boldsymbol{\lambda}}_{t,s}(m_s)$ and $\beta^{\boldsymbol{\lambda}}_{t-1,s}(m_s)$ might cross then. When $\lambda_t \leq \lambda_{t-1}$ however, Propositions 3 and 4 imply that the optimal policy for (9) is characterized by thresholds satisfying $\beta^{\boldsymbol{\lambda}}_{t,s}(m_s) \geq \beta^{\boldsymbol{\lambda}}_{t-1,s}(m_s)$ for all m_s. As a result, when $\lambda_t \leq \lambda_{t-1}$ for all t subproblem (9) then becomes an *optimal stopping problem* (cf. Bertsekas 2001, Vol. I p. 168). That is, for every initial state there is a stochastic time $0 \leq t^*_s \leq t$ at which it is optimal to forever remove product s from the shelf. If we further assume $\lambda_t = \lambda$ for all t, this becomes equivalent to the two-armed bandit problem with one known arm (cf. Berry and Fristedt 1985, p. 92).

3.4 A Suboptimality Bound

Since the exact calculation of $J^*_t(\boldsymbol{m}, \boldsymbol{\alpha})$ seems challenging except in trivial instances, we now develop an upper bound using the duality results of subsections §3.2 and §3.3; we then use that bound in subsequent sections to quantify the suboptimality of some heuristic policies.

Proposition 1 implies that an upper bound for the optimal expected profit is obtained by considering the best open-loop dual policy:

$$J^*_t(\boldsymbol{m}, \boldsymbol{\alpha}) \;\leq\; \min_{\boldsymbol{\lambda} \geq \boldsymbol{0}} \; H^{\boldsymbol{\lambda}}_t(\boldsymbol{m}, \boldsymbol{\alpha}), \tag{10}$$

where an explicit expression for the right-hand side is provided by (8) and (9). A better bound follows from using the best open-loop dual policy to approximate (for each state) the profit-to-go $J^*_{t-1}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})$ in the Bellman equation (3), that is:[6]

$$J^*_t(\boldsymbol{m}, \boldsymbol{\alpha}) \leq \max_{\substack{\mathbf{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\Big[\min_{\boldsymbol{\lambda} \geq \boldsymbol{0}} H^{\boldsymbol{\lambda}}_{t-1}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u}) \Big]. \tag{11}$$

However, the expectation in (11) is not separable and its calculation seems very computationally intensive. By interchanging the order of the minimization and maximization operators in (11) we still have an upper bound and the problem becomes separable, but becomes then equivalent to solving (10). The minimization with respect to $\boldsymbol{\lambda}$ in (10) can be solved with any convex non-differentiable optimization method, and yields the upper performance bound we will use in the remainder of this paper.

[6] In what follows there is a slight abuse of notation: we write $\boldsymbol{\lambda}$ to denote a vector but the number of components depends on the context, for example when writing $H^{\boldsymbol{\lambda}}_{t-1}(\cdot)$, $\boldsymbol{\lambda}$ is a vector with $(t-1)$ components.

3.5 The Index Policy

In this subsection we derive a heuristic index policy for the dynamic assortment problem. This is done in two steps:

First Step: a Closed Form Approximation for the Single-Product Profit-to-go

- First, we impose $\lambda_t = \lambda$ for all t, i.e. the shelf space opportunity cost is assumed to be the same in all periods. The known arm in (9) is called in that case by Gittins a standard arm, and it follows from Proposition 4 that:

$$H^{\lambda}_{t,s}(m_s, \alpha_s) = \max\left\{ r_s \frac{m_s}{\alpha_s} - \lambda + \mathbb{E}_{n_s}\Big[H^{\lambda}_{t-1,s}(m_s + n_s, \alpha_s + 1) \Big], 0 \right\}. \qquad (12)$$

- Second, we implement a lookahead horizon of length one (see Bertsekas 2001). That is, in the recursive calculation of the expected profit at period t the profit-to-go of period $t-1$ is approximated by the profit-to-go of stage 1. Formally, the profit-to-go $H^{\lambda}_{t-1,s}(m_s, \alpha_s)$ is thus approximated by:

$$\tilde{H}^{\lambda}_{t-1,s}(m_s, \alpha_s) = (t-1) \cdot \max\left\{ r_s \frac{m_s}{\alpha_s} - \lambda, 0 \right\}. \qquad (13)$$

Substituting (13) in (12) and using $[x]^+$ to denote the positive side of x, we see that the optimal strategy at period t in the approximate problem depends on the sign of:

$$\begin{aligned} \tilde{d}^{\lambda}_{t,s}(m_s, \alpha_s) &= r_s \frac{m_s}{\alpha_s} - \lambda + (t-1) \cdot \mathbb{E}_{n_s}\left[\left[r_s \frac{m_s + n_s}{\alpha_s + 1} - \lambda \right]^+ \right] \\ &= \frac{r_s \sqrt{m_s}}{\alpha_s \sqrt{\alpha_s + 1}} \left((t-1) \cdot \mathbb{E}_{n_s}\left[\left[\frac{n_s - \mathbb{E}[n_s]}{\sqrt{\mathbb{V}[n_s]}} - b^{\lambda}_s \right]^+ \right] - b^{\lambda}_s \right), \end{aligned}$$

$$\text{where } b^{\lambda}_s = \left(\frac{\lambda}{r_s} - \frac{m_s}{\alpha_s} \right) \frac{\alpha_s \sqrt{\alpha_s + 1}}{\sqrt{m_s}}, \ \mathbb{E}[n_s] = \frac{m_s}{\alpha_s}, \text{ and } \mathbb{V}[n_s] = \mathbb{E}[n_s]\left(\frac{\alpha_s + 1}{\alpha_s} \right). \qquad (14)$$

The second equality above is obtained through direct algebraic manipulation (similar to the example on p.12 in Berry and Fristedt 1985).

- Third, as a negative binomial with parameters m_s and $\alpha_s(\alpha_s + 1)^{-1}$, n_s is the sum of m_s independent geometric random variables; we thus approximate n_s by a normal distribution with the same mean and variance, which is asymptotically exact as m_s increases by the Central Limit Theorem. This yields:

$$\tilde{d}^{\lambda}_{t,s}(m_s, \alpha_s) \approx \frac{r_s \sqrt{m_s}}{\alpha_s \sqrt{\alpha_s + 1}} \Big((t-1) \cdot \Psi(b^{\lambda}_s) - b^{\lambda}_s \Big), \qquad (15)$$

where $\Psi(z) = \int_z^{\infty} (x - z)\phi(x) dx$ is the loss function of a standard normal.

Since $\Psi(z)$ is continuous, positive and strictly decreasing (cf. DeGroot 1970, p. 247), the equation

$$(t-1)\cdot\Psi(z_t) = z_t \tag{16}$$

has a unique solution for all $t \geq 2$ (in the following, we let $z_1 \equiv 0$ for completeness). Moreover, the values z_t, which are independent of the problem data, are increasing and concave in t – see Table 1 for the first few numerical values of z_t with four digits accuracy.

t	1	2	3	4	5	6	7	8
z_t	0.0000	0.2760	0.4363	0.5492	0.6360	0.7065	0.7658	0.8168

Table 1: First values of z_t.

The policy for problem (9) resulting from these approximations is simple: if $b_s^\lambda \leq z_t$ at period t, then include product s in the assortment (i.e. "pull arm s"), otherwise do not include it. The corresponding profit-to-go is given by:

$$H_{t,s}^\lambda(m_s, \alpha_s) \approx \frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}}\Big[(t-1)\cdot\Psi(b_s^\lambda) - b_s^\lambda\Big]^+. \tag{17}$$

Second Step: Linear Search in λ

We now adapt to our problem a heuristic solution method initially developed by Castañon (1997). Assume $\lambda_t = \lambda$ for all t as before, and let $u_{t,s}^\lambda$ be the optimal decision in the single-product subproblem $H_{t,s}^\lambda(m_s, \alpha_s)$ defined by (12). For any product s, we have that $\lim_{\lambda\to 0} u_{t,s}^\lambda = 1$ and $\lim_{\lambda\to\infty} u_{t,s}^\lambda = 0$; moreover, it follows from (12) that $H_{t,s}^\lambda(m_s, \alpha_s)$ is nonnegative and nonincreasing in λ. Consequently, there must exist $\eta_{t,s} \geq 0$ such that $u_{t,s}^\lambda = 1$ if and only if $\lambda \leq \eta_{t,s}$. The threshold $\eta_{t,s}$ (multiplied by t) is exactly the equivalent of Gittins' index for our version of the multiarmed bandit problem (where Gittins' index is defined as the lump sum described in §1.2). Using the approximation derived in the first step above, we obtain:

$$\begin{array}{lll} & u_{t,s}^\lambda = 1 & \\ \Leftrightarrow & \frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}}\Big[(t-1)\cdot\Psi(b_s^\lambda) - b_s^\lambda\Big] \geq 0 & \text{by definition of } \tilde{d}_{t,s}^\lambda(m_s,\alpha_s) \text{ in (15);} \\ \Leftrightarrow & b_s^\lambda \leq z_t & \text{by definition of } z_t \text{ in (16);} \\ \Leftrightarrow & \lambda \leq r_s\frac{m_s}{\alpha_s} + z_t\frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}} & \text{by definition of } b_s^\lambda \text{ in (14).} \end{array} \tag{18}$$

Substituting the moments of γ_s given at the end of §2.1.2 in the last expression of (18), we finally obtain the following approximation for index $\eta_{t,s}$:

$$\eta_{t,s} \approx r_s\mathbb{E}[\gamma_s] + z_t\frac{r_s\mathbb{V}[\gamma_s]}{\sqrt{\mathbb{V}[\gamma_s]+\mathbb{E}[\gamma_s]}} \tag{19}$$

In order to find a feasible policy for the original problem, Castañon suggests a linear search on the value of λ so that the coupling constraint (in our case, $\sum_{s=1}^S u_{t,s}^\lambda \leq N$) is satisfied as an equality (ties can be solved with a lexicographic rule); in our problem the resulting approximate policy therefore consists of selecting the N products with the largest indices $\eta_{t,s}$, calculated according to equation (19). In words, the index $\eta_{t,s}$ represents the highest price at which one should be willing to rent some shelf space in order to display (and sell) product s there; it is thus a measure of the desirability of including each individual product in the assortment, and from that standpoint the rationale behind Castañon's heuristic is to fill all shelf space with the most desirable products. Note that the first term in the index expression (19) favors exploitation, and the second term favors exploration, since it is increasing in both the variance of γ_s and the number of remaining periods (through z_t). Intuitively, when uncertainty about demand for a product s (captured by $\mathbb{V}[\gamma_s]$) is high, there is more benefit to learn from including s in the assortment because of the upside potential from future sales. Because resolving this uncertainty does take some time however, one may not be able to benefit from this learning with only few periods left before the end of the season, since the associated upside potential then remains limited. That is, one should increasingly favor exploitation over exploration as the remaining planning horizon (and opportunity for leveraging exploration) shortens, which is captured by the decrease with t of the multiplicative factor z_t in (19).

Note that our index $\eta_{t,s}$ takes the form of immediate expected profit plus some function of the variance, and resembles in that way other indices defining policies suggested for different versions of the multiarmed bandit problem by Ginebra and Clayton (1995) and Brezzi and Lai (2002) for example. The fact that our policy thus depends on only the first two moments of expected demand may be a desirable feature from an implementation standpoint; in particular, the estimation procedure based on experts opinions developed by Fisher and Raman (1996) for Sport Obermeyer could be used to estimate the initial priors. Our reader may however notice the apparent unit inconsistency that in the exploration term of (19) an expectation is added to a variance (as opposed to say, a standard deviation). This results in fact from our rescaling time units in order to work with a period length equal to one, effectively hidding appropriate unit conversion factors; a more detailed explanation can be found in the Appendix.

Finally, when assessing the performance of the index policy defined above, our primary benchmark will be the *greedy policy*, which consists of selecting in each period the N products with the highest immediate expected profit $r_s\mathbb{E}[\gamma_s]$ (thus greedily favoring exploitation over exploration). The greedy policy is also known in the multiarmed bandit literature as *play-the-leader* rule; note that it still involves learning despite its myopic nature, since priors are still updated in each period with observed demand with that policy, only the impact of assortment decisions on future learning is ignored. As a result, several authors (e.g. Aviv and Pazgal 2002) also refer to it as *passive learning.*

3.6 Assortment Implementation Lead Time

In this subsection we remove the assumption that the assortment decisions can be implemented in the same period when they are made. Instead, we assume that there is more generally a constant lag of ℓ periods between the time when the assortment decision is made and the time when it becomes effective in the store. That is, an assortment decision made in period t will impact the store in period $t-\ell$. In the case of Zara, the implementation lag ℓ would likely be an integer value between 2 and 5, representing the same number of weeks since assortment decisions seem to be made on a weekly basis. Although this implementation lag ℓ arises in practice from delays associated with all process steps between design and storage on the shelf (e.g., drawing, procurement, sewing, distribution, etc.), in the following we will only refer to ℓ as the "lead time".

With a positive lead time, the state space in the DP model must be extended in order to keep track of past decisions yet to be implemented. Specifically, the state is now given by the vector $(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha})$, where $\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}$ are the assortments that will be offered from the current period t down to period $t-\ell+1$, and $(\boldsymbol{m}, \boldsymbol{\alpha})$ are the distribution parameters of the beliefs about demand at time t. The decision made at time $t \in \{T+\ell, \ldots, \ell+1\}$ is the assortment that will be implemented at time $t-\ell$, and the first ℓ assortments $\boldsymbol{v}^T, \ldots, \boldsymbol{v}^{T-\ell+1}$ must all be determined upfront (i.e. before the season starts at time T) with the only knowledge of the initial prior on demand. The optimal profit-to-go for a given initial state can be then obtained through the following recursion:

$$J_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) = \sum_{s=1}^{S} \sum_{\tau=t-\ell+1}^{t} r_s \frac{m_s}{\alpha_s} v_s^\tau + W_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) \tag{20}$$

where $W_0^* = \ldots = W_\ell^* = 0$ for any state, and $W_t^*(.)$ satisfies for $t > \ell$:

$$W_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) = \max_{\sum_{s=1}^{S} u_s \le N} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\Big[W_{t-1}^*(\boldsymbol{v}^{t-1}, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{u}, \boldsymbol{m}+\boldsymbol{n}\cdot\boldsymbol{v}^t, \boldsymbol{\alpha}+\boldsymbol{v}^t)\Big] \tag{21}$$

The summation in the right hand side of (20) shows explicitly that the expected profit of the next ℓ periods cannot be affected. Intuitively, the existence of a positive lead time slows the learning process down (since any learning about demand may only have an impact ℓ periods later), and the number of remaining learning periods at t effectively reduces to $t-\ell-1$. Note that if $\ell = 0$ then $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = W_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$ and (21) reduces then to the recursion (3) studied in the previous subsections.

As is clear from the expansion of the state space by a factor of $2^{S\times\ell}$, the existence of a positive lead time increases the complexity of our dynamic program. However, the duality concepts introduced earlier still apply and may be used to generate the following upper bound for equation (21):

$$W_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) \;\le\; \min_{\boldsymbol{\lambda}} \; N \sum_{\tau=1}^{t-\ell} \lambda_\tau + \sum_{s=1}^{S} \; H_{t,s}^{\boldsymbol{\lambda}}(v_s^t, \ldots, v_s^{t-\ell+1}, m_s, \alpha_s),$$

where $H_{0,s}^{\boldsymbol{\lambda}} = \ldots = H_{\ell,s}^{\boldsymbol{\lambda}} = 0$ and for $t > \ell$:

$$H_{t,s}^{\boldsymbol{\lambda}}(v_s^t, \ldots, v_s^{t-\ell+1}, m_s, \alpha_s) = \max\left\{ r_s \tfrac{m_s}{\alpha_s} - \lambda_{t-\ell} + \mathbb{E}_{n_s}\left[H_{t-1,s}^{\boldsymbol{\lambda}}(v_s^{t-1}\ldots, v_s^{t-\ell+1}, 1, m_s + n_s \cdot v_s^t, \alpha_s + v_s^t)\right], \right.$$
$$\left. \mathbb{E}_{n_s}\left[H_{t-1,s}^{\boldsymbol{\lambda}}(v_s^{t-1}\ldots, v_s^{t-\ell+1}, 0, m_s + n_s \cdot v_s^t, \alpha_s + v_s^t)\right] \right\}.$$

Moreover, we can invoke arguments similar to the ones used in §3.4 to obtain the following upper bound for the maximization of $J_T^*(\boldsymbol{v}^T, \ldots, \boldsymbol{v}^{T-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha})$ with respect to $(\boldsymbol{v}^T, \ldots, \boldsymbol{v}^{T-\ell+1})$ subject to the corresponding binary and shelf space constraints:

$$\min_{\boldsymbol{\lambda}} \; N \sum_{\tau=1}^{T} \lambda_\tau + \sum_{s=1}^{S} \max_{\substack{v_s^T, \ldots, v_s^{T-\ell+1} \\ \in \{0,1\}}} \left(\sum_{\tau=T-\ell+1}^{T} \left(r_s \frac{m_s}{\alpha_s} - \lambda_\tau \right) v_s^\tau + H_{t,s}^{\boldsymbol{\lambda}}(v_s^T, \ldots, v_s^{T-\ell+1}, m_s, \alpha_s) \right), \quad (22)$$

which provides the upper bound that we will report for the performance of various policies simulated in Section §4 in environments with a positive lead time.

Finally, our proposed policy may be heuristically adapted by introducing the two following modifications to the index definition given by equation (19):

1. First, we substitute the term z_t in (19) with

$$z_t \quad \longrightarrow \quad z_{L(t)}, \quad (23)$$

   where $L(t) = \max\{t - 2\ell, 1\}$. The rationale is that in period t the retailer must decide the assortment of period $(t - \ell)$, and from then on he has ℓ fewer periods to learn about demand. In particular, if $\ell \geq \frac{t-1}{2}$ then $z_{L(t)} = 0$ so that the adapted index policy coincides then with the greedy policy, which can be shown to generate optimal actions in that case. Note that if $\ell \geq T - 1$ then no learning is possible and the best the retailer can do is to implement the optimal static assortment for the next T periods; this would exactly corresponds to the "traditional retailer" described earlier in §1.1.

2. The second modification in (19) concerns the variance $\mathbb{V}[\gamma_s]$. Recall from section §2.1.2 that the prior becomes more accurate as more sales are observed. Hence, the prediction made at time t for the variance of γ_s at time $t - \ell$ must take into account whether product s is committed as part of the assortment in any of the ℓ periods in between. Specifically, we substitute the variance term in the index formula with:

$$\mathbb{V}[\gamma_s] = \frac{m_s}{\alpha_s^2} \quad \longrightarrow \quad \mathbb{V}[\gamma_s] = \frac{m_s + \frac{m_s}{\alpha_s} \sum\limits_{\tau=t-\ell+1}^{t} v_s^\tau}{(\alpha_s + \sum\limits_{\tau=t-\ell+1}^{t} v_s^\tau)^2}, \quad (24)$$

   where as before $\sum_{\tau=t-\ell+1}^{t} v_s^\tau$ is the number of times that product s is included in the assortment during the interval of ℓ periods starting with period t. Note that m_s and α_s are thus replaced by a prediction of what their values will be at time $t-\ell$, considering how many times product s will have been part of the assortment by then. Intuitively, substitution (24) captures the predicted gain in information quality (or equivalently reduction in prior variance) resulting from the assortments already decided but not yet implemented. As a consequence of (24), the second term in the index formula (19) now decreases with the sum $\sum_{\tau} v_s^{\tau}$, expressing that when designing the assortment for period $t-\ell$ the incentive to explore the demand for product s reduces when it already has a large presence in the next ℓ assortments.

In the next section, we report the performance achieved by the heuristic policy just described in various numerical experiments.

4 Numerical Experiments

The objective of the simulation study we report in this section is to assess the relative performance in various environments of our proposed index policy against the greedy policy and the dual upper bounds derived in §3.4 and §3.6. We describe our methodology in §4.1, then discuss our experimental results in §4.2 and §4.3.

4.1 Methodology

There seems to be two accepted methodologies for evaluating policy performance in environments involving learning, and in the two next subsections we adopt each one in turn. Subsection §4.2 follows what is known in the multiarmed bandit literature as the *Bayesian* approach, also adopted for example in Aviv and Pazgal (2002). It relies on the assumption that the predictive Bayesian distribution updated in each period (in our case, the negative binomial distribution characterized by equation (1)) is essentially correct. In simulations, actual demand in each period is generated from that negative binomial distribution (as opposed to a Poisson distribution), and those experiments do not require the specification of any underlying demand rates. These experiments thus allow to focus on the quality of the index policy as a solution to the self-contained dynamic programming formulation (3), independently of the Bayesian framework under which it has been derived.

Subsection §4.3 follows the *frequentist* approach (see Lai 1987 and Brezzi and Lai 2002), also adopted for example in Bertsimas and Mersereau (2004). In contrast, this method relies on the specification of the real underlying distribution parameters (in our case, the demand rates γ_s), and actual demand for each product in each period is generated in simulations from the corresponding Poisson distribution. This approach therefore allows to characterize how the relative performance of different policies may be affected by the quality of the information initially available (e.g. accuracy and bias).

We used similar data sets for the experiments reported in §4.2 and §4.3. Specifically, we assumed that the available shelf space N is equal to 30 and that the number of potential products S is equal to 720, roughly matching our estimates of these quantities for one category of products (e.g. Women's upper garments) in a Zara store (see our discussion in §2.2). We ran most experiments for values of the season length T equal to 10, 20 and 40, and values of the assortment implementation lead time ℓ equal to 0 and 5. We generated upfront the net margin r_s for each product $s \in S$ through independent draws from a Uniform distribution $U[2, 8]$, and used these numbers throughout. We also assumed that the retailer had the same initial prior for all products. In particular, we fixed the initial expected demand rate $\mathbb{E}[\gamma_s]$ at 10 products per period, but we tested three different values for the initial variance $\mathbb{V}[\gamma_s]$: 5, 50, and 100, corresponding to values for the distribution parameters (m_s, α_s) equal to $(1, 1/10)$, $(2, 1/5)$, and $(20, 2)$ respectively. The lower and upper bounds given by Lemma 2 for the expected total profits generated by the optimal policy for these data sets are provided in Table 2.

$\mathbb{V}[\gamma_s]$	Static Assortment	Bayesian Full Info.
5	2376.10	3042.16
50	2376.10	5424.06
100	2376.10	7176.11

Table 2: Bounds of Lemma 2.

Finally, all numerical experiments were performed on a personal computer with a 1.6 GHz Pentium processor with 768 MB of RAM. The simulations and the upper bound optimization problem were coded in the C programming language. We ran $11,000$ replications for each simulation data point, which was sufficient to ensure that all reported results have an absolute relative error smaller than 0.5% for a confidence level of 95%. The running time of one simulation point (i.e. $11,000$ replications) increased with the horizon length T, reaching about 5 minutes for $T = 40$. When computing the upper bounds derived in §3.4 and §3.6, the support of the negative binomial distribution was truncated at values with probability less than 10^{-6}. Solutions to the corresponding non differentiable optimization problem (cf. (10)) were computed using the Nelder-Mead simplex method. While this algorithm is not generally guaranteed to converge to the minimum (see Lagarias et al. 1998), it does maintain a best solution found to date, which in our case still yields a valid bound (this follows from weak duality since solutions to (10) correspond to open-loop dual policies, see §3.2 and §3.4). In some instances we tried different starting points for this algorithm, and report then the best bounds we have found.

4.2 Bayesian Experiments

Table 3 summarizes our numerical results for this first set of experiments. The total expected profit divided by the number of periods (hereafter referred to as ”expected profit per period”) is shown for the greedy rule and our index policy in its fourth and fifth columns respectively. The sixth column provides the upper bound for these quantities derived using DP duality. The seventh column reports the relative improvement achieved by the index policy over the greedy policy, and the eight column provides the associated suboptimality gap for the index policy.

$\mathbb{V}[\gamma_s]$	T	ℓ	Grdy	Indx	UpBnd	$\frac{Indx-Grdy}{Grdy}\cdot 100$	$\frac{UpBnd-Indx}{Indx}\cdot 100$
5	10	0	2598.35	2604.19	2608.05	0.22%	0.15%
	20	0	2670.37	2686.78	2693.97	0.61%	0.27%
	40	0	2726.53	2766.50	2819.91	1.47%	1.93%
5	10	5	2429.44	2441.42	2456.12	0.49%	0.60%
	20	5	2522.01	2588.84	2608.58	2.65%	0.76%
	40	5	2617.38	2709.41	2753.84	3.52%	1.64%
50	10	0	3498.76	3635.11	3656.37	3.90%	0.58%
	20	0	3753.40	4082.60	4133.26	8.77%	1.24%
	40	0	3910.34	4479.50	4714.70	14.56%	5.20%
50	10	5	2609.78	2861.14	2864.40	9.63%	0.11%
	20	5	2961.80	3791.60	3945.55	28.02%	4.06%
	40	5	3334.55	4396.98	4625.55	31.86%	5.20%
100	10	0	4031.50	4273.81	4311.70	6.01%	0.89%
	20	0	4420.36	4985.29	5130.00	12.78%	2.90%
	40	0	4646.64	5632.36	5883.58	21.21%	4.46%
100	10	5	2706.58	3095.76	3206.80	14.38%	3.59%
	20	5	3198.91	4580.76	4787.70	43.20%	4.52%
	40	5	3757.42	5530.75	5754.43	47.20%	4.04%

Table 3: Index policy vs. greedy rule (Bayesian approach).

Over the range of scenarios considered in Table 3, the relative gap between the performance of the index policy and the dual upper bound is typically small, reaching a maximum value of 5.2%. This not only suggests that the index policy is in fact near optimal, but also that the upper bound is quite tight.

We also observe that the proposed index policy always outperforms the greedy policy, and that its relative advantage increases with the number of periods and prior variances. Our interpretation is that increases in the season length and initial prior variances respectively increase the opportunity to learn about demand and the payoff from doing so, both favoring the index policy which implements a more elaborate (active) learning strategy than the (passive) learning used by the greedy policy. The impact of the season length shown in Table 3 appears more clearly in Figure 2, which specifically plots the expected profit per period of the index and greedy policies as well as the corresponding upper bound against the total number of periods T for an initial state equal to $(1, 1/10)$ (i.e. $\mathbb{E}[\gamma_s] = 10$ and $\mathbb{V}[\gamma_s] = 100$) and no implementation lead time ($\ell = 0$).

In line with previous results, the expected profit per period shown in Figure 2 increases with the total number of periods faster overall for the index policy than it does for the greedy policy. An important observation however is that the performance advantage of the index policy relative to the greedy policy only becomes significant when the number of periods is large enough (in this case $T > 6$): ripping the benefits of active learning seems to require a minimum number of decision and observation periods, below which the greedy policy does just as well – other studies involving Bayesian learning models (e.g. Aviv and Pazgal 2002, or Brezzi and Lai 2002) report

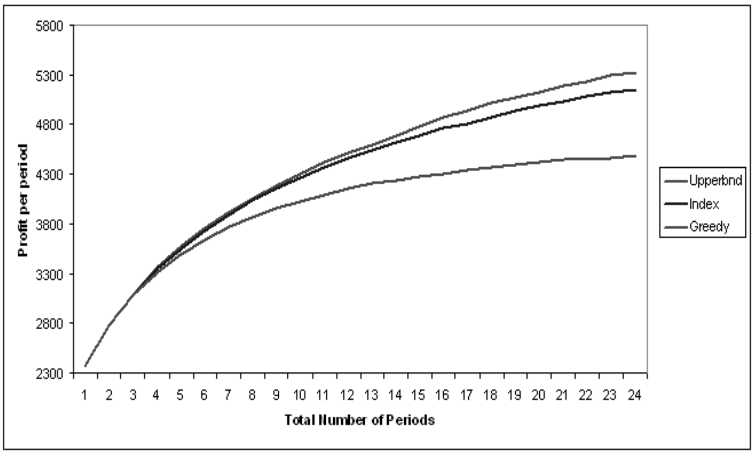

Figure 2: Relative policy performance for various horizon lengths.

similar findings. In addition, while the performance of both policies appearing in Figure 2 for a single decision period ($T = 1$) is by definition exactly identical to that of the static assortment reported in Table 2, the greedy policy (and a fortiori the index policy) significantly outperforms the static assortment with two or more periods to go. Specifically, the performance gain over the static assortment from implementing passive learning with a single additional period of observation (i.e. $T = 2$) is about 21%: passive learning is considerably better for this data set than no learning at all. However, while that finding may apply to many situations of practical interest, it does not have any obvious theoretical grounding: consider an environment with a first group of more than N products having known average profit rates, and a second group with uncertain demand and lower predicted profit rates but high prior variances, reflecting that some of the products in this second group may in fact have higher underlying profit rates; the greedy policy would then never include any of the products from the second group in the assortment, thus never learning anything about their demand, and its performance would then remain identical to that of the static assortment regardless of the season length.

Although very long season lengths appear unlikely in the retail setting that initially motivated this study, one may legitimately wonder how the results of Table 3 and Figure 2 would change in the limit where the number of periods T is very large, which is also the object of the brief discussion after Lemma 2. Other experiments conducted for $T = 500$ (not reported here) support the conjecture that the expected profit per period of the index policy converges to the full information upper bound appearing in Table 2 as T goes to infinity. Note that the greedy policy does not have this property in general, as illustrated by the environment described in the previous paragraph.

Table 3 also suggests that the relative advantage of the index policy over the greedy policy becomes even more significant with an assortment implementation lead time ($\ell > 0$). To focus on this issue we plot in Figure 3 the performance of the index and greedy policies as well as the corresponding upper bound (derived in §3.4 and §3.6) against the lead time ℓ for an initial state equal to $(\mathbb{E}[\gamma_s], \mathbb{V}[\gamma_s]) = (10, 100)$ as before, and a season length T equal to 24 periods. Note that the range of lead times considered ($\{0, ..., 5\}$) as well as the season length assumed (about six months) roughly correspond to our estimates for the corresponding quantities at Zara (see §2.2).

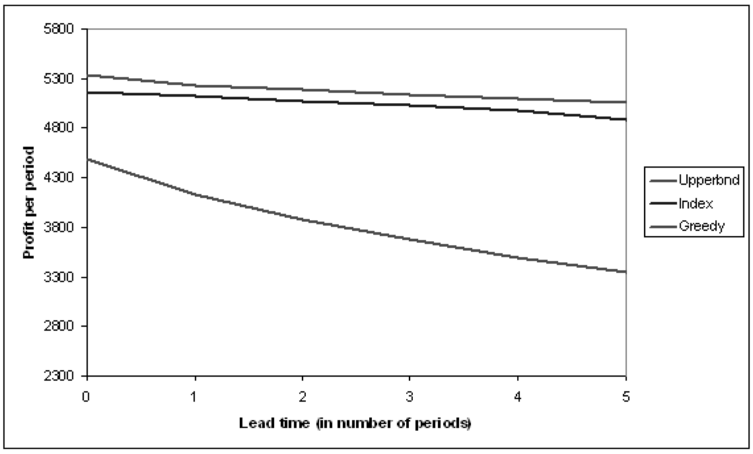

Figure 3: Relative policy performance for various lead times.

The performance of both policies as well as the upper bound values shown in Figure 3 all exhibit a general decreasing trend. Increasing the lead time while holding the season length constant effectively reduces the number of periods where demand can be observed and acted upon, and therefore the potential to learn throughout the season; this decreasing trend and the overall increase of performance with T appearing in Figure 2 thus indirectly follow from the same phenomenon. Also, the results shown in Figure 3 confirm that the performance of the greedy policy relative to both the index policy and the upper bound quickly deteriorates when the lead time increases. We believe that the distinction between active and passive learning is key to this phenomenon. Specifically, increasing the lead time augments the magnitude of future changes in information quality (i.e. expected reduction of prior variances resulting from the next ℓ assortments) that the greedy policy ignores but the index policy captures (through (24)), thus yielding a larger relative advantage to active learning over passive learning.

4.3 Frequentist Experiments

The goal of our frequentist experiments was to assess how the relative performance of the index and greedy policies is affected by the quality of the demand information initially available, both in terms of accuracy and bias. We used the same data sets as in the Bayesian experiments, but used instead real underlying demand rates and associated Poisson distributions when generating actual demand in simulations.

The objective of our first set of experiments was to examine policy performance in environments where the initial priors were unbiased and had various degree of accuracy, in the following sense: we generated upfront three sets of underlying demand rates through independent draws from a

Gamma distribution with the same parameters (m_s, α_s) as the three different initial Gamma priors characterizing the retailer's initial beliefs we assumed; furthermore, when performing a simulation run with given initial priors we used the corresponding set of underlying demand rates.

Table 4 shows in its fourth and fifth columns the expected profit per period of the greedy and index policies obtained in those experiments. The sixth column gives the full information upper bound, i.e. the expected profit achievable by a decision-maker with knowledge of the underlying demand rates that were generated as described above. The seventh column reports the improvement of the index policy upon the greedy rule, and finally the eight column shows the performance gap of the index policy relative to the full information upper bound, or relative regret.

$\mathbb{V}[\gamma_s]$	T	ℓ	Grdy	Indx	Full	$\frac{Indx-Grdy}{Grdy} \cdot 100$	$\frac{Full-Indx}{Indx} \cdot 100$
5	10	0	2722.38	2732.87	3166.81	0.39%	15.88%
	20	0	2802.85	2819.59	3166.81	0.60%	12.31%
	40	0	2864.43	2892.12	3166.81	0.97%	9.50%
5	10	5	2533.15	2544.88	3166.81	0.46%	24.44%
	20	5	2635.37	2716.01	3166.81	3.06%	16.60%
	40	5	2731.94	2840.27	3166.81	3.97%	11.50%
50	10	0	3330.73	3577.49	5366.44	7.41%	50.01%
	20	0	3602.94	4048.13	5366.44	12.36%	32.57%
	40	0	3763.54	4450.54	5366.44	18.25%	20.58%
50	10	5	2414.05	2779.42	5366.44	15.14%	93.08%
	20	5	2754.51	3755.01	5366.44	36.32%	42.91%
	40	5	3142.28	4382.27	5366.44	39.46%	22.46%
100	10	0	3872.19	4112.01	7102.50	6.19%	72.73%
	20	0	4121.11	4822.96	7102.50	17.03%	47.26%
	40	0	4276.16	5422.32	7102.50	26.80%	30.99%
100	10	5	2825.25	3078.95	7102.50	8.98%	130.68%
	20	5	3274.56	4450.35	7102.50	35.91%	59.59%
	40	5	3694.57	5351.85	7102.50	44.86%	32.71%

Table 4: Index policy vs. greedy rule (frequentist approach).

The results shown in the seventh column of Table 4 confirm the earlier finding that the index policy performs better than the greedy policy over a range of environments and that this superiority is particularly significant for large initial prior variance, large number of periods and long lead times, indicating that this finding is quite robust. This relative advantage seems to always increases with the leadtime ℓ as before, and the results in Table 4 suggest that the same holds for the total number of periods T. We interpret the relative regret of the index policy reported in the last column of Table 4 as follows: the benefit of having full information relative to using the index policy increases with the initial prior variance (which measure the quality of the partial information initially available), decreases with the number of periods (because longer horizons provide for more opportunity to learn), and increases with the lead time (which effectively reduces the number of periods when demand observations can be acted upon).

The goal of our second set of frequentist experiments was to estimate the impact of improved prior accuracy on policy performance. As in Bertsimas and Mersereau (2004), we assumed that the retailer could perform some preliminary off-line experiments before the beginning of the season in order to strengthen his initial priors. That is, we generated for each product M random observations from a Poisson distribution with a mean equal to the real underlying demand rate, and performed the corresponding Bayesian updates to obtain the priors from which we started our simulations. Table 5, which has the same structure as Table 4, shows our results for $M = 3$.

$\mathbb{V}[\gamma_s]$	T	ℓ	Grdy	Indx	Full	$\frac{Indx-Grdy}{Grdy} \cdot 100$	$\frac{Full-Indx}{Indx} \cdot 100$
5	10	0	3039.11	3041.21	3166.81	0.07%	4.13%
	20	0	3060.50	3062.36	3166.81	0.06%	3.41%
	40	0	3079.01	3080.09	3166.81	0.04%	2.82%
5	10	5	3007.48	3006.30	3166.81	-0.04%	5.34%
	20	5	3023.18	3040.09	3166.81	0.56%	4.17%
	40	5	3056.82	3069.62	3166.81	0.42%	3.17%
50	10	0	5278.74	5278.74	5366.44	0.00%	1.66%
	20	0	5288.60	5289.18	5366.44	0.01%	1.46%
	40	0	5299.09	5299.08	5366.44	0.00%	1.27%
50	10	5	5263.65	5267.45	5366.44	0.07%	1.88%
	20	5	5272.17	5272.97	5366.44	0.02%	1.77%
	40	5	5285.92	5291.31	5366.44	0.10%	1.42%
100	10	0	7019.44	7022.49	7102.50	0.04%	1.14%
	20	0	7028.40	7035.03	7102.50	0.09%	0.96%
	40	0	7035.50	7045.87	7102.50	0.15%	0.80%
100	10	5	6995.09	6995.83	7102.50	0.01%	1.52%
	20	5	7012.79	7017.94	7102.50	0.07%	1.20%
	40	5	7026.63	7038.89	7102.50	0.17%	0.90%

Table 5: Relative policy performance with improved accuracy of initial information.

As shown in Table 5, the performance of the greedy and index policies become statistically indistinguishable when the quality of the information initially available is improved as described above – in this environment where the payoff from learning is significantly reduced, sophisticated learning strategies do not yield any advantage over simpler ones. In addition, the regret associated with both policies (i.e. the performance gap relative to the full information upper bound) is drastically reduced compared to the values in Table 4. The main insight we thus draw from Table 5 is the speed at which estimation accuracy and policy performance improve with the number of preliminary offline observations. This experimental finding suggests that the potential benefits associated with leveraging sales data across multiple stores confronted with similar demand patterns may be very large in practice (see §5.1 for a related discussion).

Finally, in our third set of experiments we explored the impact of introducing some bias in the initial demand information on policy performance. Specifically, we first generated another three sets of *biased* demand rate estimations γ_s' (one set for each possible type of initial prior information) using the exact same procedure followed to generate the real demand rates γ_s as described above. Secondly, we assumed now that the $M = 3$ preliminary demand observations were generated from Poisson distributions with mean equal to the biased demand estimates γ'_s, instead of the true demand rates γ_s used at this stage in the previous set of experiments, performed the corresponding Bayesian updates, and started each simulation with the resulting priors. The results for $T = 40$ and $\ell = 0$ are shown in Table 6.

$\mathbb{V}[\gamma_s]$	T	ℓ	Grdy	Indx	Full	$\frac{Indx-Grdy}{Grdy} \cdot 100$	$\frac{Full-Indx}{Indx} \cdot 100$
5	40	0	2649.41	2672.12	3166.81	0.86%	18.51%
50	40	0	3414.14	3457.27	5366.44	1.26%	55.22%
100	40	0	3626.21	3666.61	7102.50	1.11%	93.71%

Table 6: Relative policy performance with biased initial information.

The performance of the greedy and index policies reported in Table 6 are almost identical. This suggests that in the presence of bias, there is no advantage from performing active learning over passive learning – these two strategies distinguish themselves from the relevance of what information is acquired over time, not from their ability to detect erroneous prior information. This observation may motivate the development of more robust learning models including the ability to challenge existing priors, for example through dynamic goodness-of-fit tests.

Remarkably, for both policies the performance results in terms of regret shown in Table 6 are substantially worse than their corresponding values in Table 4 (where the gaps of the index policy relative to the full information bound are only 9.50%, 20.58% and 30.99% in the three corresponding scenarios). That is, the retailer would have been better off without doing any experiments at all, regardless of which policy is followed – while preliminary demand observations can be extremely valuable as shown in Table 5, it is particularly important to ensure that they are not biased. If such additional sales data is obtained by observing demand in another store for example, it is paramount to establish that these stores indeed face similar customer populations, or at least that any systematic bias is corrected.

5 Model Extensions

Completing the discussion initiated in §2.2, we now comment on possible ways to relax the three assumptions made in our analysis that we consider to be most restrictive: independent product demands (§5.1); no lost sales (§5.2); and constant demand rates (§5.3).

5.1 Substitution and Complementarity Effects

As argued in §2.2, our model would gain realism if demands for different products were no longer assumed independent, capturing instead substitution effects between products from the same category, and possibly complementarity effects between products from different categories. However, designing and analyzing a dynamic assortment model where learning concerns not only the demand rates of individual products but also their correlation structure seems very challenging for at least two reasons. First, even if a Bellman equation similar to (3) could be written for such a model, the corresponding DP would predictably no longer be weakly coupled because of the many relationships between different products introduced by the correlation structure, so that our decomposition approach would likely break down. Second and perhaps more fundamentally, the number of parameters required to characterize such a correlation structure would be a priori in the order of S^2; a high value of S relative to $N \times T$ (the total number of demand observations available) may thus create a discrepancy between the amount of data required for estimation and the speed at which it can be acquired – this is related to the problem known as "overfitting" in the Machine Learning literature (i.e. the model is too complex with respect to the available data). Indeed, our rough estimates of these parameters in the case of Zara (see §2.2) indicate that this problem could be an important one in practice. It is also revealing that (static) assortment studies proposing practical methods for estimating demand correlation structures (e.g. Kök and Fisher 2004 , Anunpindi et al. 1998) typically rely on sales history from multiple stores with different assortments assumed to face the same demand characteristics, that is substantially more learning data than the single store observations we consider. While coordinating dynamic assortment decisions across multiple stores and leveraging the resulting data constitutes an important avenue for future research in our view, we caution that studies such as Fisher and Rajaram (2000) have established that demand characteristics faced by different stores of the same firm may in practice be quite different.

But we believe that the dynamic assortment policy presented in §3.5 and §3.6, even though its derivation required the assumption of independence, may still provide a useful starting point when designing heuristics capturing substitution effects. One such possible design path, which we now develop, is to assume that the correlation structure across products is known (or can at least be estimated upfront), while the individual demand rates of individual products must be estimated dynamically as before. As in the substitution models of Smith and Agrawal (2000) and Kök and Fisher (2004) we can use the concept of the *original* demand for each product, defined as the demand that would be observed for that product if all the other products were also included in the assortment. In addition, we also assume that the retailer knows the probability q_{is} that a customer switches to product s given that he originally wanted product i but it was not available in the assortment – as in the last two papers cited, this model assumes that each customer only makes one such substitution attempt, and $\sum_{s \neq i} q_{is} < 1$ capturing the fact that customer might leave without buying. Our dynamic index policy can then be adapted heuristically by performing the following two modifications:

1. The retailer now maintains Gamma Bayesian priors with parameters $(\mathbf{m}, \boldsymbol{\alpha})$ on the *original* demand rates for each product, so the information updating rule must be modified to reflect that observed sales for a given product may include some to customers who only bought it because their favorite choice was not part of the assortment. Let $\boldsymbol{u} \in \mathcal{U}$ represent the assortment that was available in the store at period t, s be a product that was part of the

assortment (i.e. $u_s = 1$), and n_s be the sales observed for s. An estimate of the original sales $\widetilde{n}_s$ of product s is then given by

$$\widetilde{n}_s = n_s \cdot \left(\frac{\frac{m_s}{\alpha_s}}{\frac{m_s}{\alpha_s} + \sum_{i \neq s} q_{is} \frac{m_i}{\alpha_i}(1 - u_i)} \right). \tag{25}$$

In words, the fraction of original observed sales is estimated as the ratio between the expected contribution of the original demand for product s and the total expected demand considering substitution. The information state for each included product s is then updated from m_s to $m_s + \widetilde{n}_s$, and α_s is updated to $\alpha_s + 1$ as before. The demand estimates for products not included in the assortment remain unchanged in this proposal, although an alternative approach could consist of also updating priors based on the fraction of sales that is discarded through equation (25).

2. The index $\eta_{t,s}$ derived in §3.5 (and extended to the case of positive lead time in §3.6) is a measure of the desirability of independently including each product in the assortment, defined as the opportunity cost of the corresponding shelf space. In the presence of substitutions, the desirability of including a product must also take into account whether it is a good substitute for other products not included in the assortment. The selection of the N most desirable products becomes then a combinatorial problem, which we propose to address through the following quadratic integer program:

$$\max_{\substack{\boldsymbol{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^{S} \Big(\eta_{t,s} + r_s \sum_{i \neq s} q_{is} \frac{m_i}{\alpha_i}(1 - u_i) \Big) u_s. \tag{26}$$

In words, the objective in (26) evaluates the profitability of including each product s in the assortment at t by adding to the initial desirability index $\eta_{t,s}$ the expected profits following from substitutions to product s from all products i not included in the assortment (represented by the inner summation term). This formulation thus still captures the essential trade-off between exploration and exploitation, but corrects the exploitation term for the expected sales resulting from substitutions. Note that when substitution effects are ignored (i.e. $q_{is} = 0 \ \forall i, s$), solving (26) results in our original index policy.

5.2 Models with Lost Sales

In a model with lost sales, the product inventory levels become important to capture, and in addition to assortment inclusion or exclusion decisions one should seemingly also consider order quantity decisions. Furthermore, different assumptions about the type of demand information available to the retailer can be made, and we have formulated accordingly the following models and associated Bellman equations: (i) lost sales are observable for products included in the assortment; (ii) lost sales are not observable, but the point in time when a stockout occurs (when applicable) is known for every product in the assortment; and (iii) the only information available about lost sales is whether or not some of them did occur. In formulating problem (iii) with censored information defined above, we used a significantly different demand model than the one assumed in the present paper. Specifically, we have adapted to our problem the Bayesian learning model with censored observations initially developed by Lariviere and Porteus (1999), where the existence of unobserved lost sales is explicitly taken into account when updating information. Because the underlying demand in that model is restricted to a rather narrow family of distributions however, we fear that the resulting assortment model may only be useful to obtain insights rather than for a practical implementation. More generally, we are hoping to report analytical results for all three aforementioned models in the future.

5.3 Variable Demand Rates

In our model, the unknown demand rates γ_s remain constant during the season, which results in a partially observed Markov decision process (POMDP) in which the underlying state is fixed. Situations where product life-cycles are really short compared to the season length (e.g. a couple of weeks versus six months) may however be more faithfully described by time-varying demand rates. This feature could be captured by a POMDP where the real underlying state would change over time with some given transition probabilities; this would basically amount to extending our model in the same way that Aviv and Pazgal (2004) extend their initial dynamic pricing problem (Aviv and Pazgal 2002). While the theory of POMDPs allows for a transformation of the partially observed state problem into one with perfect state information, this comes at the expense of increase state space dimension, so that further approximations would likely have to be made.

6 Conclusions

We have developed in this paper a discrete-time DP model for the dynamic assortment problem faced by a fast-fashion retailer refining his estimate of consumer demand for his products over time. The main assumptions made were: (i) independent products; (ii) no lost sales; and (iii) constant demand rates. Under these assumptions we have formulated this dynamic assortment problem as a multiarmed bandit with finite horizon and multiple plays per stage. Using the Lagrangian decomposition of weakly coupled DPs, we have derived a closed form index policy characterized by equation (19) that depends on only the first two moments of the priors on demand rates. Despite its simple form, our proposed index policy captures two key features of the dynamic assortment problem, namely the trade-off between exploration and exploitation and the finite horizon effect, and is amenable to an extension for the case with positive design-to-shelf lead times. Also based on DP duality, we have derived an upper bound for the optimal profit-to-go, which allows to assess the suboptimality gap of the suggested index policy.

Our simulation study indicates that the index policy always performs at least as well as the greedy policy (or passive learning), and significantly outperforms it in scenarios with diffuse or biased prior demand information. Also, numerical computations of the bound mentioned above suggest that the index policy is close to optimal. In general, the improvement of the suggested index policy upon the greedy rule increases with the planning horizon length, the variance of the initial priors, and the lead time.

Although the three major assumptions listed above may be particularly strong in some environments, our approach was partly motivated by the belief that the closed-form policy they allow to derive constitutes a useful starting point for designing heuristics or developing extensions in more complex environments. In the present paper we have thus proposed a heuristic for capturing substitution effects between products, and discussed possible ways for relaxing the last two major assumptions as part of future work. Another interesting extension, motivated in part by our experimental findings, would consider the coordination of dynamic assortment decisions across multiple stores.

Finally, although the model presented here focuses essentially on operational issues, we point out that it may also have some design implications. Specifically, the current financial success of fast-fashion firms like Zara suggests that the relative benefits of increased supply flexibility, while considerably harder to quantify at the design stage than the relative costs of local and overseas production, may still be very large. Could it be that many traditional fashion retail firms have been mistaken for years when assessing the trade-off between costs of production and benefits of flexibility? A legitimate hypothesis is that the heavy historical reliance of the fashion industry on overseas suppliers may have resulted in part from a lack of appropriate quantitative models enabling to correctly predict the potential gains associated with local production and a responsive supply network. In our model, the design-to-shelf leadtime ℓ may precisely reflect the procurement delays resulting from a given supply-chain configuration, and studying the variation of retailer's profits with that parameter (as shown in Figure 3) may thus inform the assessment of such trade-off. We thus hope that our model may also be useful to some practitioners when designing supply-chains.

Acknowledgments

We are most grateful to Pr. Yossi Sheffi, Larry Lapide and the MIT-Zaragoza Logistics Program for funding this project and providing useful feedback. We are also grateful to Pr. Yves Dallery and the Laboratoire de Génie Industriel of the École Centrale Paris for hosting the first author during the Summer 2004, and providing many helpful comments. We also thank Gabriel Bitran, Steve Graves and the participants of the OM seminar at the Sloan School of Management for offering insightful remarks about this work. Finally, the second author is indebted to Dr. Martha Nieto for a conversation about Zara that sparked his interest in fast fashion companies and was key to the genesis of this project.

A Appendix

A.1 A Short Comment on the Index Formula

If δ denote the length of a time period, the expression for the index of product s in period t is then more generally:

$$\eta_{t,s} = r_s \mathbb{E}[\gamma_s]\delta + z_t \frac{r_s \mathbb{V}[\gamma_s]\delta}{\sqrt{\mathbb{V}[\gamma_s] + \frac{\mathbb{E}[\gamma_s]}{\delta}}} \tag{27}$$

Since γ_s represents an arrival rate it must have the same units as $1/\delta$; the right hand side of (27) is therefore consistent (in terms of units) and does not depend on the choice of the period length. Rescaling the time units so that $\delta = 1$ yields (19). Alternatively, one may redefine γ_s as $\gamma_s\delta$, making this quantity a scalar with no physical dimension or units. Likewise, direct substitution in (27) gives equation (19).

A.2 Proof of Lemma 1

We proceed by induction on t. The property is trivial for $t = 0$ so we assume it holds for $t - 1$, with $t \geq 1$. Consider any vector $\mathbf{u} \in \{0,1\}^S$ such that $\sum_{s=1}^{S} u_s \leq N$. Let $\boldsymbol{n}'' = \boldsymbol{n}(\boldsymbol{m}'', \boldsymbol{\alpha}'')$ and $\boldsymbol{n}' = \boldsymbol{n}(\boldsymbol{m}', \boldsymbol{\alpha}')$. From the induction hypothesis $J^*_{t-1}(\boldsymbol{m}'' + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha}'' + \boldsymbol{u}) \geq J^*_{t-1}(\boldsymbol{m}' + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha}' + \boldsymbol{u})$ for any $\boldsymbol{n} \in \mathbb{N}^S$, which in turn implies that:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{n}''}\Big[J^*_{t-1}(\boldsymbol{m}'' + \boldsymbol{n}'' \cdot \boldsymbol{u}, \boldsymbol{\alpha}'' + \boldsymbol{u})\Big] &\geq \mathbb{E}_{\boldsymbol{n}''}\Big[J^*_{t-1}(\boldsymbol{m}' + \boldsymbol{n}'' \cdot \boldsymbol{u}, \boldsymbol{\alpha}' + \boldsymbol{u})\Big] \\ &\geq \mathbb{E}_{\boldsymbol{n}'}\Big[J^*_{t-1}(\boldsymbol{m}' + \boldsymbol{n}' \cdot \boldsymbol{u}, \boldsymbol{\alpha}'\boldsymbol{u})\Big]. \end{aligned}$$

The first inequality is strict if for any product s, $m''_s > m'_s$ or $\alpha''_s < \alpha'_s$. The last inequality follows from the fact that $J^*_{t-1}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})$ is a (componentwise) increasing function of $\boldsymbol{n}$ (by the induction hypothesis), and from the relative stochastic ordering of $\boldsymbol{n}(\boldsymbol{m}, \boldsymbol{\alpha})$. It follows that:

$$\sum_{s=1}^{S} r_s \frac{m''_s}{\alpha''_s} u_s + \mathbb{E}_{\boldsymbol{n}''}\left[J^*_{t-1}(\boldsymbol{m}'' + \boldsymbol{n}'' \cdot \boldsymbol{u}, \boldsymbol{\alpha}''\boldsymbol{u})\right] \geq \sum_{s=1}^{S} r_s \frac{m'_s}{\alpha'_s} u_s + \mathbb{E}_{\boldsymbol{n}'}\left[J^*_{t-1}(\boldsymbol{m}' + \boldsymbol{n}' \cdot \boldsymbol{u}, \boldsymbol{\alpha}' + \boldsymbol{u})\right]$$

Since the above inequality is valid for any feasible action $\boldsymbol{u}$, invoking the definition of the profit-to-go function (3) completes the proof. □

A.3 Proof of Lemma 2

The lower bound follows from the fact that $J^*_t(\boldsymbol{m}, \boldsymbol{\alpha})$ is the expected profit-to-go of the optimal dynamic assortment policy. In particular, the optimal policy performs at least as well as a static policy implementing in each period the assortment given by $\operatorname{argmax}_{\boldsymbol{u} \in \mathcal{U}} \sum_{s=1}^{S} r_s \mathbb{E}[\gamma_s] u_s$.

The upper bound follows from the fact that the frequentist regret is nonnegative for any nonnegative parameter vector $\boldsymbol{\gamma}$ (cf. Lai 1987, p.1092). The proof is complete. □

A.4 Proof of Proposition 1

From the definition, it is clear that $H_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) \leq H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{m}, \boldsymbol{\alpha})$ for any dual policy $\boldsymbol{\lambda_t}$, therefore we only need to prove the first inequality. We proceed by induction on t. Assume that $J_{t-1}^*(\mathbf{m}, \boldsymbol{\alpha}) \leq H_{t-1}^*(\boldsymbol{m}, \boldsymbol{\alpha})$ for all states $(\mathbf{m}, \boldsymbol{\alpha})$, then for any $\lambda_t \geq 0$:

$$\begin{aligned}
J_t^*(\mathbf{m}, \boldsymbol{\alpha}) &= \max_{\substack{\mathbf{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big] \\
&\leq N\lambda_t + \max_{\substack{\mathbf{u} \in \{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S \Big(r_s \frac{m_s}{\alpha_s} - \lambda_t\Big) u_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big] \\
&\leq N\lambda_t + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^S \Big(r_s \frac{m_s}{\alpha_s} - \lambda_t\Big) u_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big] \\
&\leq N\lambda_t + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^S \Big(r_s \frac{m_s}{\alpha_s} - \lambda_t\Big) u_s + \mathbb{E}_{\boldsymbol{n}}\big[H_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big] \qquad (28)
\end{aligned}$$

The first inequality follows from the fact that $\lambda_t \geq 0$, and the second holds because the feasible set is larger. The third inequality relies on the induction hypothesis. Considering now the minimum of the right hand side of (28) yields the desired result. □

A.5 Proof of Proposition 2

We will need the following lemmas that are interesting per se:

Lemma 4 *Let $(\boldsymbol{m}, \boldsymbol{\alpha})$ be the system state at period t. For any $i \in \mathcal{S}$ the following holds:*

$$\mathbb{E}_{n_i}\big[J_t^*(\boldsymbol{m} + n_i \boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})\big] \geq J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}), \qquad (29)$$

where n_i is a negative binomial with parameters (m_i, α_i).

Proof: We proceed by induction on t. Assume that (29) is true for some $t-1 \geq 0$. For any (random) vector $\boldsymbol{v}$, let $\boldsymbol{v_{-i}} = \boldsymbol{v} - v_i \boldsymbol{e_i}$. For any given decision vector $\boldsymbol{u} \in \{0,1\}^S$ we denote the respective profit by:

$$g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha}) = \sum_{s=1}^S r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n'}}\big[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big]. \qquad (30)$$

We will show that $\mathbb{E}_{n_i}[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i \boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})] \geq g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha})$ by considering two cases. First, assume that $u_i = 0$, then we have that:

$$\begin{aligned}
\mathbb{E}_{n_i}[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})] &= \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{n_i}\Big[\mathbb{E}_{\boldsymbol{n'_{-i}}}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\big]\Big] \\
&= \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n'_{-i}}}\Big[\mathbb{E}_{n_i}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\big]\Big] \\
&\geq \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n'_{-i}}}\big[J^*_{t-1}(\boldsymbol{m} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\big] \\
&= g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha})
\end{aligned}$$

The first equality follows from (30) and the fact that we are assuming $u_i = 0$. The expectation interchange in the second equality is a consequence of demands among products being independent and Fubini's Theorem (all terms are nonnegative). In the third step we used the induction hypothesis, and then in the last step we used again (30) and $u_i = 0$.

For the second case assume that $u_i = 1$ and fix n_i at a given (nonnegative) integer value. Then we have the following inequality:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{n'}}\Big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\Big] &= \mathbb{E}_{\boldsymbol{n'_{-i}}}\Big[\mathbb{E}_{n'_i}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\big]\Big] \\
&\geq \mathbb{E}_{\boldsymbol{n'_{-i}}}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u_{-i}}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u_{-i}})\big]
\end{aligned} \tag{31}$$

where n'_i is a negative binomial random variable with parameters $(m_i + n_i, \alpha_i + 1)$, and in the second inequality we use the induction hypothesis. We now have that:

$$\begin{aligned}
\mathbb{E}_{n_i}[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})] &= \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{n_i}\Big[\mathbb{E}_{\boldsymbol{n'}}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\big]\Big] \\
&\geq \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{n_i}\Big[\mathbb{E}_{\boldsymbol{n'_{-i}}}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u_{-i}}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u_{-i}})\big]\Big] \\
&= g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha})
\end{aligned}$$

The first equality follows from (30) and the fact that $\mathbb{E}_{n_i}\left[\frac{m_i+n_i}{\alpha_i+1}\right] = \frac{m_i}{\alpha_i}$. In the second inequality we used (31), and the last step is also given by (30) and the independence among product demands.

So we can conclude that:

$$\mathbb{E}_{n_i}[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})] \geq g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha}) \quad \forall \boldsymbol{u} \in \{0, 1\}^S \tag{32}$$

We can now prove the inequality of the lemma. In fact, we have the following:

$$\begin{aligned}
\mathbb{E}_{n_i}\big[J_t^*(\boldsymbol{m}+n_i\boldsymbol{e_i},\boldsymbol{\alpha}+\boldsymbol{e_i})\big] &= \mathbb{E}_{n_i}\Bigg[\max_{\substack{\mathbf{u}\in\{0,1\}^S:\\ \sum_{s=1}^S u_s\le C}} g_t(\boldsymbol{u},\boldsymbol{m}+n_i\boldsymbol{e_i},\boldsymbol{\alpha}+\boldsymbol{e_i})\Bigg]\\
&\ge \max_{\substack{\mathbf{u}\in\{0,1\}^S:\\ \sum_{s=1}^S u_s\le C}} \mathbb{E}_{n_i}\big[g_t(\boldsymbol{u},\boldsymbol{m}+n_i\boldsymbol{e_i},\boldsymbol{\alpha}+\boldsymbol{e_i})\big]\\
&\ge \max_{\substack{\mathbf{u}\in\{0,1\}^S:\\ \sum_{s=1}^S u_s\le C}} g_t(\boldsymbol{u},\boldsymbol{m},\boldsymbol{\alpha})\\
&= J_t^*(\boldsymbol{m},\boldsymbol{\alpha})
\end{aligned}$$

The first and last equality are given by the definition of $J_t^*(\cdot)$ and (30). The second inequality can be seen as a consequence of Jensen's inequality and the fact that the maximum norm is convex, and the third inequality follows from (32). Then the proof is complete. □

Lemma 5 *If $r_s > 0\ \forall s$, then $f_t(\boldsymbol{m},\boldsymbol{\alpha};C)$ is a strictly increasing function of C, with $C \le S$, for any state $(\boldsymbol{m},\boldsymbol{\alpha})$.*

Proof: Consider $C < S$. Let $\boldsymbol{u}^*$ be an optimal solution of the maximization problem in the definition of $f_t(\boldsymbol{m},\boldsymbol{\alpha};C)$ (cf. (7)), and let i be such that $u_i^* = 0$. Then we have that:

$$f_t(\boldsymbol{m},\boldsymbol{\alpha};C) = \sum_{s=1}^S r_s\frac{m_s}{\alpha_s}u_s^* + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{n}\cdot\boldsymbol{u}^*,\boldsymbol{\alpha}+\boldsymbol{u}^*)\big] \tag{33}$$

Let $\overline{\boldsymbol{u}} = \boldsymbol{u}^* + \boldsymbol{e_i}$, where $\boldsymbol{e_i}$ is the i-th unit vector. By conditioning on all n_s with $s \neq i$ and using Lemma 4 we have that:

$$\mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{n}\cdot\overline{\boldsymbol{u}},\boldsymbol{\alpha}+\overline{\boldsymbol{u}})\big] \ge \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{n}\cdot\boldsymbol{u}^*,\boldsymbol{\alpha}+\boldsymbol{u}^*)\big]. \tag{34}$$

Since $r_i > 0$, from (34) we get a strict inequality relating the objective values of $\overline{\boldsymbol{u}}$ and $\boldsymbol{u}^*$:

$$\sum_{s=1}^S r_s\frac{m_s}{\alpha_s}\overline{u}_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{n}\cdot\overline{\boldsymbol{u}},\boldsymbol{\alpha}+\overline{\boldsymbol{u}})\big] > f_t(\boldsymbol{m},\boldsymbol{\alpha};C). \tag{35}$$

Since $\sum_{s=1}^S \overline{u}_s = C+1$, from (35) we have that $f_t(\boldsymbol{m},\boldsymbol{\alpha};C+1) > f_t(\boldsymbol{m},\boldsymbol{\alpha};C)$, i.e. $f_t(\boldsymbol{m},\boldsymbol{\alpha};C)$ is a strictly increasing function of C. □

Lemma 6 *For a given state $(\boldsymbol{m},\boldsymbol{\alpha})$ at period t, consider the following dual function:*

$$h_t(\lambda_t,\mathbf{m},\boldsymbol{\alpha}) = N\cdot\lambda_t + \max_{\mathbf{u}\in\{0,1\}^S}\ \big(\sum_{s=1}^S r_s\frac{m_s}{\alpha_s} - \lambda_t\big)u_s + \mathbb{E}_{\boldsymbol{n}}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{n}\cdot\boldsymbol{u},\boldsymbol{\alpha}+\boldsymbol{u})\big]$$

Let $h_t^*(\mathbf{m},\boldsymbol{\alpha}) = \min_{\lambda_t\ge 0} h_t(\lambda_t,\mathbf{m},\boldsymbol{\alpha})$. If $f_t(\mathbf{m},\boldsymbol{\alpha};C)$ is concave in C, then $J_t^*(\mathbf{m},\boldsymbol{\alpha}) = h_t^*(\mathbf{m},\boldsymbol{\alpha})$.

Proof: Since the state $(\mathbf{m}, \boldsymbol{\alpha})$ is fixed throughout the proof it will be omitted in the notation.

Instead of following a standard duality proof (for example using a hyperplane separation theorem, see Bertsekas (1999), we provide a short direct corroboration.

Let λ_t^* be such that $f_t(N+1) - f_t(N) \leq \lambda_t^* \leq f_t(N) - f_t(N-1)$. The existence of λ_t^* is guaranteed from the concavity of $f_t(C)$ with respect to C, and also λ_t^* is nonnegative from Lemma 5. We will show that λ_t^* is a Lagrangian multiplier in the sense that $J_t^* = h_t(\lambda_t^*) = h_t^*$.

First, note that the dual function can be written as:

$$h_t(\lambda_t) = N \cdot \lambda_t + \max_{C \in \ \mathbb{N}} f_t(C) - C \cdot \lambda_t. \tag{36}$$

Suppose that for λ_t^* the maximum on the right hand side of (36) is attained strictly at some $C > N$. This means that $f_t(C) - C \cdot \lambda_t^* > f_t(N) - N \cdot \lambda_t^*$, or equivalently, $f_t(C) - f_t(N) > (C-N) \cdot \lambda_t^*$.

On the other hand, from the concavity of $f_t(C)$ we have that:

$$f_t(C) - f_t(N) = f_t(C) - f_t(C-1) + f_t(C-1) - f_t(C-2) + \ldots + f_t(N+1) - f_t(N) \leq (C-N) \cdot \lambda_t^*,$$

which is contradiction. If we now suppose that the maximum on the right hand side of (36) is attained strictly at some $C < N$, then a similar contradiction is obtained, and therefore we must have that $h_t(\lambda_t^*) = f_t(N)$.

To conclude, we know that $J_t^* = f_t(N)$ because $f_t(C)$ is nondecreasing (cf. Lemma 5), and also $J_t^* \leq h_t(\lambda_t)$. Then $J_t^* = h_t(\lambda_t^*) = \min_{\lambda_t \geq 0} h_t(\lambda_t) = h_t^*$, and the proof is complete. □

Finally, to prove Proposition 2 we proceed by induction on t. The case $t = 1$ is trivial so we assume that the property holds for $t - 1 > 0$ and that $f_\tau(\boldsymbol{m}', \boldsymbol{\alpha}'; C)$ is concave in C for all $\tau = t, \ldots, 1$ and states $(\boldsymbol{m}', \boldsymbol{\alpha}')$ reachable from $(\boldsymbol{m}, \boldsymbol{\alpha})$ in period τ. For any $\boldsymbol{u} \in \mathcal{U}$ and any vector $\boldsymbol{n} \in \mathbb{N}^S$ we have that $(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})$ is reachable from $(\boldsymbol{m}, \boldsymbol{\alpha})$ in period $t-1$. Then, by the induction hypothesis we have that $J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u}) = H_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})$. Using the latter we see that the last inequality in the proof of Proposition 1 (cf. 28) is actually an equality. If we now minimize with respect to λ_t, from Lemma 6 and the definition of the optimal dual policy (cf. (6)) we have that $J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = H_t^*(\boldsymbol{m}, \boldsymbol{\alpha})$, and the proof is complete. □

A.6 Proof of Lemma 3

We proceed by induction. Consider $t \geq 1$ and assume that (8) holds for $t-1$. Then, from equation (5):

$$\begin{aligned} H_t^{\boldsymbol{\lambda}}(\mathbf{m}, \boldsymbol{\alpha}) &= N\lambda_t + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^{S} (r_s \frac{m_s}{\alpha_s} - \lambda_t) u_s + \mathbb{E}_{\boldsymbol{n}} \big[H_{t-1}^{\boldsymbol{\lambda}}(\boldsymbol{m} + \boldsymbol{n} * \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u}) \big] \\ &= N\lambda_t + \max_{\mathbf{u} \in \{0,1\}^S} \sum_{s=1}^{S} (r_s \frac{m_s}{\alpha_s} - \lambda_t) u_s + \mathbb{E}_{\boldsymbol{n}} \big[N \sum_{\tau=1}^{t-1} \lambda_\tau + \sum_{s=1}^{S} H_{t-1,s}^{\boldsymbol{\lambda}}(m_s + n_s u_s, \alpha_s + u_s) \big] \end{aligned}$$

$$
\begin{aligned}
&= N\sum_{\tau=1}^{t}\lambda_\tau + \max_{\mathbf{u}\in\{0,1\}^S}\sum_{s=1}^{S}(r_s\frac{m_s}{\alpha_s}-\lambda_t)u_s + \sum_{s=1}^{S}\mathbb{E}_{n_s}\left[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_su_s,\alpha_s+u_s)\right]\\
&= N\sum_{\tau=1}^{t}\lambda_\tau + \sum_{s=1}^{S}\left(\max_{u_s\in\{0,1\}}(r_s\frac{m_s}{\alpha_s}-\lambda_t)u_s + \mathbb{E}_{n_s}\left[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_su_s,\alpha_s+u_s)\right]\right)\\
&= N\sum_{\tau=1}^{t}\lambda_\tau + \sum_{s=1}^{S}H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)
\end{aligned}
$$

The second equation uses the induction hypothesis. The third equation comes from the fact that all products are independent so the expectation is simplified, and the final two equations rearrange terms in order to obtain the desired result. □

A.7 Proof of Proposition 3

We first need the following two additional lemmas:

Lemma 7 $H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s) \le (\frac{r_s m_s}{\alpha_s})t \quad \forall(m_s,\alpha_s)$.

Proof: Direct by induction since assuming that it holds for $t-1$ we can bound both terms in the right hand side of (9). In fact, we have that $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s) \le (t-1)r_s m_s/\alpha_s$ and

$$
r_s\frac{m_s}{\alpha_s}-\lambda_t+\mathbb{E}_{n_s}\Big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\Big] \le r_s\frac{m_s}{\alpha_s}+\mathbb{E}_{n_s}\Big[r_s\frac{m_s+n_s}{\alpha_s+1}(t-1)\Big] = (\frac{r_s m_s}{\alpha_s})t-\lambda_t.
$$

□

Lemma 8 $H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)=0 \quad \forall(m_s,\alpha_s)$ *such that* $\big(\frac{r_s m_s}{\alpha_s}\big)\tau<\lambda_\tau \ \forall\tau=t,\ldots,1$.

Proof: Consider $t\ge 1$ and assume that the claim holds for $t-1$. Let (m_s,α_s) be a pair that satisfies $\big(\frac{r_s m_s}{\alpha_s}\big)\tau<\lambda_q \ \forall\tau=t,\ldots,1$. Then, from the induction hypothesis, $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s)=0$, and from Lemma 7 we have that:

$$
r_s\frac{m_s}{\alpha_s}-\lambda_t+\mathbb{E}_{n_s}\Big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\Big] \le t\big(r_s\frac{m_s}{\alpha_s}\big)-\lambda_t<0.
$$

Then, from equation (9) we have that $u_s=0$ is optimal at time t and $H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)=0$, which completes the induction step. □

Now for the proof of Proposition 3, consider the following function:

$$
d^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)=r_s\frac{m_s}{\alpha_s}-\lambda_t+\mathbb{E}_{n_s}\Big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\Big]-H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s).
$$

In a similar way than in Lemma 4 it can be shown that $\mathbb{E}_{n_s}\big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\big] \ge H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s)$. Then, for α_s sufficiently small $d^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s) \ge r_s\frac{m_s}{\alpha_s}-\lambda_t>0$. On the other hand, when $\alpha_s\to\infty$, from Lemma 8 we have that $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s)\to 0$. From Lemmas 8 and 7 and the Dominated Convergence Theorem it can be seen that $\mathbb{E}_{n_s}\big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\big]\to 0$, so we have that $d_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) \to -\lambda_t < 0$. If the function $d_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ were strictly decreasing in α_s, then $\beta_{t,s}^{\boldsymbol{\lambda}}(m_s)$ could be defined as the unique solution of $d_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) = 0$, and if $d_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ were strictly increasing in m_s, then $\beta_{t,s}^{\boldsymbol{\lambda}}(m_s)$ would inherit the same monotonicity property.

We now prove by induction on t that $d_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ is indeed strictly decreasing in α_s and strictly increasing in m_s. The claim is trivial for $t = 1$ when clearly $\beta_{1,s}^{\boldsymbol{\lambda}}(m_s) = r_s m_s / \lambda_1$. Assume now that the claim is valid for $t - 1$ with $t > 1$; because no ambiguity arises here in the following we omit the subscript s for simplicity. Let $\alpha' \leq \alpha''$, $m' \leq m''$, $n' = n(m', \alpha')$, and $n'' = n(m'', \alpha'')$. Since $d_t^{\boldsymbol{\lambda}}(m, \alpha)$ is continuous in α, we only need to consider three cases:

- $\alpha' \leq \alpha'' \leq \beta_{t-1}^{\boldsymbol{\lambda}}(m') \leq \beta_{t-1}^{\boldsymbol{\lambda}}(m')$

  In general, for any $\alpha \leq \beta_{t-1}^{\boldsymbol{\lambda}}(m)$:

$$\begin{aligned} d_t^{\boldsymbol{\lambda}}(m, \alpha) &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[H_{t-1}^{\boldsymbol{\lambda}}(m+n, \alpha+1) - H_{t-2}^{\boldsymbol{\lambda}}(m+n, \alpha+1)\Big] \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m+n, \alpha+1), 0\big\}\Big] \end{aligned} \tag{37}$$

  From the induction hypothesis $\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m''+n, \alpha''+1), 0\big\} \geq \max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m'+n, \alpha'+1), 0\big\}$ for any integer n. Following now the same steps as in Lemma 1:

$$\begin{aligned} \mathbb{E}_{n''}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m''+n'', \alpha''+1), 0\big\} &\geq \mathbb{E}_{n''}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m'+n'', \alpha'+1), 0\big\}\Big] \\ &\geq \mathbb{E}_{n'}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m'+n', \alpha'+1), 0\big\}\Big] \end{aligned} \tag{38}$$

  Note that the first inequality is strict if either $\alpha' < \alpha''$ or $m' < m''$. The second inequality follows from the larger stochastic ordering of $n(m, \alpha)$. It follows then from (37) that $d_t^{\boldsymbol{\lambda}}(m', \alpha') \leq d_t^{\boldsymbol{\lambda}}(m'', \alpha'')$.

- $\beta_{t-1}^{\boldsymbol{\lambda}}(m') \leq \beta_{t-1}^{\boldsymbol{\lambda}}(m') \leq \alpha' \leq \alpha''$

  In general, for any $\alpha \geq \beta_{t-1}^{\boldsymbol{\lambda}}(m)$:

$$\begin{aligned} d_t^{\boldsymbol{\lambda}}(m, \alpha) &= r\frac{m}{\alpha} - \lambda_t + \mathbb{E}_n\Big[H_{t-1}^{\boldsymbol{\lambda}}(m+n, \alpha+1)\Big] - H_{t-2}^{\boldsymbol{\lambda}}(m, \alpha) \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[H_{t-1}^{\boldsymbol{\lambda}}(m+n, \alpha+1)\Big] - \mathbb{E}_n\Big[H_{t-2}^{\boldsymbol{\lambda}}(m+n, \alpha+1)\Big] + d_{t-1}^{\boldsymbol{\lambda}}(m, \alpha) \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m+n, \alpha+1), 0\big\}\Big] + d_{t-1}^{\boldsymbol{\lambda}}(m, \alpha) \end{aligned}$$

  Then $d_t^{\boldsymbol{\lambda}}(m', \alpha') \leq d_t^{\boldsymbol{\lambda}}(m'', \alpha'')$ follows from (38) and the induction hypothesis. Again, the inequality is strict if either $\alpha' < \alpha''$ or $m' < m''$.

- $\beta_{t-1}^{\boldsymbol{\lambda}}(m') \le \alpha' \le \alpha'' \le \beta_{t-1}^{\boldsymbol{\lambda}}(m')$

  In this case we have:

$$\begin{aligned}
d_t^{\boldsymbol{\lambda}}(m',\alpha') &= \lambda_{t-1} - \lambda_t + \mathbb{E}_{n'}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m'+n',\alpha'+1),0\big\}\Big] + d_{t-1}^{\boldsymbol{\lambda}}(m',\alpha') \\
&\le \lambda_{t-1} - \lambda_t + \mathbb{E}_{n'}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m'+n',\alpha'+1),0\big\}\Big] \\
&\le \lambda_{t-1} - \lambda_t + \mathbb{E}_{n''}\Big[\max\big\{d_{t-1}^{\boldsymbol{\lambda}}(m''+n'',\alpha''+1),0\big\}\Big] \\
&= d_t^{\boldsymbol{\lambda}}(m'',\alpha'')
\end{aligned}$$

  The first inequality holds because $\beta_{t-1}^{\boldsymbol{\lambda}}(m') \le \alpha' \Rightarrow d_{t-1}^{\boldsymbol{\lambda}}(m',\alpha') \le 0$. The second inequality follows from (38) and is strict if either $\alpha' < \alpha''$ or $m' < m''$. The proof is now complete. □

A.8 Proof of Proposition 4

In order to solve ties, we assume with no loss of generality that when the retailer is indifferent he will include the product in the assortment.

Consider a state $(m_s, \alpha_s) \in B_{t,s}^{\boldsymbol{\lambda}}$, necessarily:

$$r_s \frac{m_s}{\alpha_s} - \lambda_t + \mathbb{E}_{n_s}\Big[H_{t-1,s}^{\boldsymbol{\lambda}}(m_s+n_s,\alpha_s+1)\Big] < H_{t-1,s}^{\boldsymbol{\lambda}}(m_s,\alpha_s). \tag{39}$$

Suppose that in period $t-1$ it is optimal to have $u_s = 1$, i.e. $(m_s,\alpha_s) \notin B_{t-1,s}^{\boldsymbol{\lambda}}$. Substituting the appropriate expression for $H_{t-1,s}^{\boldsymbol{\lambda}}(m_s,\alpha_s)$ in (39) and rearranging terms yields:

$$\mathbb{E}_{n_s}\Big[H_{t-1,s}^{\boldsymbol{\lambda}}(m_s+n_s,\alpha_s+1)\Big] - \mathbb{E}_{n_s}\Big[H_{t-2,s}^{\boldsymbol{\lambda}}(m_s+n_s,\alpha_s+1)\Big] < (\lambda_t - \lambda_{t+1}) \le 0,$$

contradicting the fact that $H_{t,s}^{\boldsymbol{\lambda}}(m_s,\alpha_s)$ is nondecreasing with the horizon length. Therefore $u_s = 0$ must be optimal in period $t-1$, which completes the proof. □

References

Anantharam, V., P. Varaiya, and J. Walrand. 1987. Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays - Part I: I.I.D. Rewards. *IEEE T. Automat. Contr.* **32**(11) 968-976.

Anupindi, R., M. Dada, and S. Gupta. 1998. Estimation of consumer demand with stockout based substitution: An application to vending machine products. *Marketing Science.* **17** 406-423.

Aviv, Y. and A. Pazgal. 2002. Pricing of Short Life-Cycle Products through Active Learning. Working Paper. Washington University, St. Louis.

Aviv, Y. and A. Pazgal. 2004. A Partially Observed Markov Decision Process for Dynamic Pricing. Working Paper. Washington University, St. Louis.

Berry, D. A. and B. Fristedt. 1985. *Bandit Problems, Sequential Allocation of Experiments*, Chapman and Hall, New York.

Bertsekas, D. 1999. *Nonlinear Programming.* Athena Scientific, Cambridge.

Bertsekas, D. 2001. *Dynamic Programming and Optimal Control, Vols. I and II.* Athena Scientific. Cambridge.

Bertsimas, D. and A. Mersereau. 2004. A Learning Approach to Customized Marketing. Working Paper. The University of Chicago.

Brezzi, M. and T. L. Lai. 2002. Optimal Learning and Experimentation in Bandit Problems. *Journal of Economic Dynamics and Control.* **27** 87-108.

Bultez, A. and P. Naert. 1988. SHARP: Shelf Allocation for Retailers Profit. *Mktg Sci.* **7** 211-231.

Castañon, D.A. 1997. Approximate Dynamic Programming For Sensor Management. *Proceedings of the 36th IEEE Conference on Decision and Control*, 1202-1207.

DeGroot, M. H. 1970. *Optimal Statistical Decisions.* McGraw-Hill. New York.

Ferdows, K., M. Lewis and J. A.D. Machuca. 2003. Zara. *Supply Chain Forum, An Internatinal Journal* **4**(2) 62-67.

Fisher, M. L. and A. Raman. 1996. Reducing the Cost of Demand Uncertainty Through Accurate Response to Early Sales. *Operations Research.* **44**(1) 87-99.

Fisher, M. L., A. Raman, and A. S. McClelland. 2000. Rocket Science Retailing Is Almost Here - Are You Ready. *Harvard Business Review.* July-August 2000, 115-124.

Fisher, M. L. and K. Rajaram. 2000. Accurate Retail Testing of Fashion Merchandise: Methodology and Application. *Marketing Science.* **19**(3) 266-278.

Ghemawat, P. and Nueno J.L. 2003. ZARA: Fast Fashion. Harvard Business School Multimedia Case 9-703-416.

Ginebra, J. and M. K. Clayton. 1995. Response Surface Bandits. *J. Roy. Statist. Soc. Series B.* **57** 771-784.

Gittins, J. C. and D. M. Jones. 1974. A Dynamic Allocation Index for the Sequential Design of Experiments. *Progress in Statistics.* J. Gani, ed. North-Holland, Amsterdam, 241-266.

Gittins, J. C. 1979. Bandit Processes and Dynamic Allocation Indices. *J. Roy. Statist. Soc. Series B.* **14** 148-167.

Hammond, J. H. 1990. Quick Reponse in the Apparel Industry. Harvard Business School Note N9-690-038, Cambridge, Mass.

Hawkins, J.T. 2003. A Lagrangian Decomposition Approach to Weakly Coupled Dynamic Optimization Problems and its Applications. Ph.D. Thesis. Operations Research Center, MIT.

Kök, A. G. and M. L. Fisher. 2004. Demand Estimation and Assortment Optimization Under

Substitution: Methodology and Application. Working paper. Duke University.

Kumar, P. R. 1985. A Survey of Some Results in Stochastic Adaptive Control. *SIAM J. on Control and Optimization.* **23** 329-380.

Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright. 1998. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM J. Optim.* **9**(1) 112-147.

Lai, T. L. 1987. Adaptive Treatment Allocation and the Multiarmed Bandit Problem. *Annals of Statistics.* **15** 1091-1114.

Lariviere, M. A. and E. L. Porteus. 1999. Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales. *Management Science.* **45** 346-363.

Mahajan, S. and G. van Ryzin. 2001. Stocking Retail Assortments Under Dynamic Consumer Substitution. *Operations Research.* **49** 334-351.

Ross, S. 1996. *Stochastic Processes.* Wiley & Sons, New York.

Smith, S. A. and N. Agrawal. 2000. Management of Multi-item Retail Inventory Systems with Demand Substitution. *Operations Research.* **48** 50-64.

van Ryzin, G. and S. Mahajan. 1999. On the Relationship Between Inventory Costs and Variety Benefits in Retail Assortments. *Management Science.* **45** 1496-1509.

Whittle, P. 1988. Restless Bandits: Activity Allocation in a Changing World. J. Gani, ed. *A Celebration of Applied Probability, Journal of Applied Probability.* **25A** 287-298.