Dynamic Retail Assortment Models with Demand Learning for Seasonal Consumer Goods

by

Felipe Caro

Civil Industrial Engineer, University of Chile (1999)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Author .....
Sloan School of Management
April 29, 2005

Certified by .....
Jérémie Gallien
J. Spencer Standish Career Development Professor
Thesis Supervisor

Accepted by .....
Birger Wernerfelt
Professor of Management Science
Chair, Doctoral Program

Dynamic Retail Assortment Models with Demand Learning for Seasonal Consumer Goods

by

Felipe Caro

Submitted to the Sloan School of Management
on April 29, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Management

Abstract

The main research question we explore in this dissertation is: How should a retailer modify its product assortment over time in order to maximize overall profits for a given selling season?

Historically, long development, procurement, and production lead times have 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. As a result, 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.

Recently however, a few innovative firms, such as Spain-based Zara, Mango and Japan-based World Co. (referred to as "Fast Fashion" retailers), 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.

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 crucial 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? The problem just described seems challenging, in part because it relates to the classical trade-off known as exploration versus exploitation, usually represented via the multiarmed bandit problem.

In this thesis we analyze the dynamic assortment problem under different sets of assumptions, including: (i) without lost sales; (ii) with lost sales but observable demand; (iii) with lost sales and censored information; and (iv) with time varying demand rates. In each case we formulate an appropriate model and suggest a (near-optimal) policy that can be implemented in practice, together with associated suboptimality bounds. We also study the incorporation of substitution effects and

the extension of the models to a generic family of demand distributions. The common solution approach involves the Lagrangian relaxation and the decomposition of weakly coupled dynamic programs.

The dissertation makes three contributions: (1) it is the first attempt in providing mathematical optimization models with near-optimal solutions for the dynamic assortment problem faced by a fast-fashion retailer; (2) our analysis contributes to the literature on the multiarmed bandit problem, in particular for its finite-horizon version, we derive a general closed-form dynamic index policy that performs remarkably well; and (3) the solution approach contributes to the emerging literature on duality in dynamic programming.

Thesis Supervisor: Jérémie Gallien
Title: J. Spencer Standish Career Development Professor

Acknowledgments

I have reached the end of my studies at MIT and I cannot submit this thesis without expressing my gratefulness to all the people that have been next to me during these splendid years in Cambridge.

First, I would like to thank my advisor Jérémie Gallien for getting me involved in the fascinating research project that became my thesis. This work owes a lot to his support and guidance, and for me it has been career defining. I am still amazed on how fast everything worked out. I guess he had it all clear and well planned in his mind. Lucky me. Merci beaucoup!

I also want to thank the other members of my Thesis Committee, Professors Stephen Graves and Gabriel Bitran. It was an honor to have you in my committee and it was a privilege to have my office (or closet?) right next to yours. I admire your academic work and I really appreciate all the feedback and experience you shared with me.

At MIT I had the opportunity to learn from and interact with many faculty and staff members that made these years a formidable and enriching experience. In particular, I thank Professor David Simchi-Levi for his support during my first years. I hope that we continue to collaborate in the future.

MIT is a remarkable place and I would say that its best treasure is the quality of the students it attracts. I have learned almost as much from my classmates than from my professor. But besides the research, MIT is also a great place to make friends, and I made lots of them. To begin with, I have to thank Herman Bennett for being an excellent comrade from day one. My friend Victor Martínez-de-Albéniz deserves a similar distinction. Then I want to thank my OM classmates and the smart guys from the OR Center. The list is long (should I put an online appendix?).

Beyond the MIT campus I also met many people that cheered up the evenings and weekends. In particular I thank the partygoers that attended the Pub Tours on Thursdays: Cornelius, Matteo, Manu, Eva, Jitkee, Hui, Caroline, Marta, Raf, Seb, and many others with whom I shared a beer. Yeah, despite the stress, we had a lot of fun. Outside the party scene, I thank Alejandro Conejero, Daniel Hojman and Paulina Achurra for the good company, and Joe Doherty for being like family here.

Well, I have said enough about MIT and how much I loved living in Boston, but now I go down to Earth. For any achievement in my life, I must thank my parents, Rodrigo y Ximena, my sister Loreto and my brother Javier. Your love has allowed me to get where I am, without forgetting what are the really important things in life. Los quiero mucho.

Coming from Chile, I also want to thank the Department of Industrial Engineering of the University of Chile, in particular, my colleagues and academic mentors Andrés Weintraub and Rafael Epstein.

Finally, last but not least, the most important acknowledgement is for my dearly loved Marcela. I am extremely happy that you became part of this story, and with you I want to celebrate the most. Needless to say, my Ph.D. dissertation is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 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 responsive-

ness 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 1-1 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 1-1 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.

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

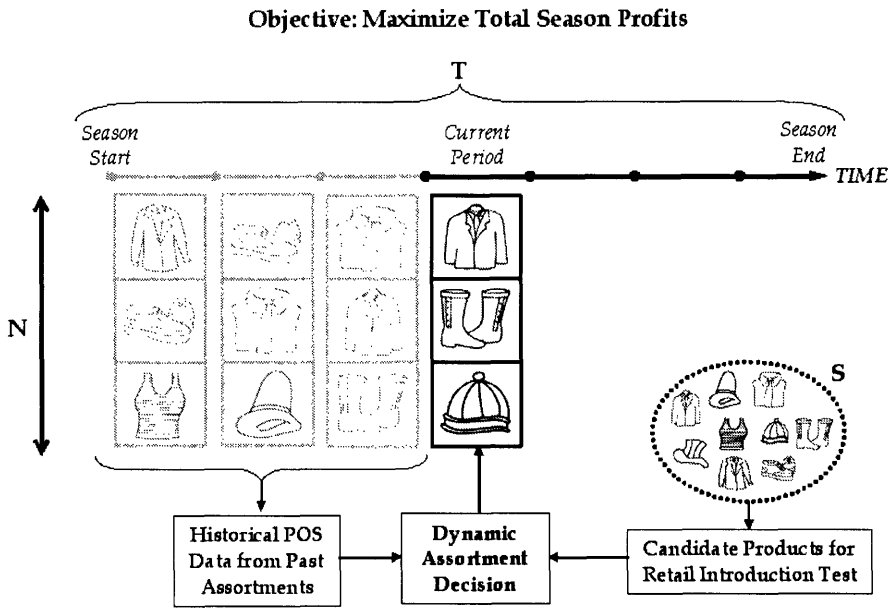

Figure 1-1: The dynamic assortment problem.

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: in Chapter 2 we provide all the duality results for dynamic programming that are used throughout the thesis. In Chapter 3 we present the basic dynamic assortment model under the following salient assumptions: (i) the demand process for each individual product is independent of the other products; (ii) there are no lost sales; and (iii) the demand rates remain constant during the selling season. Chapter 4 shows how substitution effects can be taken into account, and Chapter 5 covers the models that consider inventory decisions. Finally, in Chapter 6 we provide concluding remarks and discuss other model extensions.

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 1-1, 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).

In §3.3.5 we introduce an assortment implementation lag to our model so it becomes equivalent to a finite horizon multiarmed bandit with a response delay. The amount of literature available for this variant of the classic problem is rather limited. Most of the papers come from the statistics community that is interested in the application to clinical trials. However, the typical model involves only two arms with Bernoulli rewards. A good example is the recent work by Hardwick *et al.* (2005),

where the response delay has an exponential distribution (in our case the lag is constant and measured in number of periods).

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.

Chapter 2

Duality Results for Dynamic Programs

Dynamic programming (DP) is the natural methodology used to model and solve sequential decision problems. However, despite its versatility, in the vast majority of cases when a closed form optimal policy is not available, the numerical solution of the model involves computational requirements that quickly become overwhelming. This fact is known to be the *curse of dimensionality* of DP (see Bertsekas 2001), and as a consequence, approximate solution methods are in order.

The DP models developed in this thesis are subject to the curse of dimensionality and the approximate solution approach we follow is based on Lagrangian relaxation and the decomposition of weakly coupled dynamic programs. 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). The approach dates at least from the late 80's with the independent work done by Karmarkar (1987) on a finite-horizon multilocation inventory problem, and the seminal paper of Whittle (1988) on restless bandits. The rest of the literature reporting successful applications of this methodology is rather recent, see for instance Castañon (1997), Talluri and van Ryzin (1998), Yost and Washburn (2000), Rajaram and Karmarkar (2002), Hawkins (2003), Bertsimas and Mersereau (2004), and references therein.

The results shown in the present chapter can be seen as a generalization and extension of the individual applications found in the literature.

2.1 The Dual DP

Under the same framework as in Puterman (1990) and Bertsekas (2001), consider the following generic finite horizon Bellman equation, in which periods are counted backwards:[1]

$$J_t^*(\boldsymbol{x}) = \max_{\substack{\boldsymbol{u}\in\mathcal{U}:\\ \boldsymbol{a}'\boldsymbol{u}\leq N}} g_t(\boldsymbol{x},\boldsymbol{u}) + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[J_{t-1}^*(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \quad \boldsymbol{x}\in\Omega \tag{2.1}$$

with $J_0^*(\boldsymbol{x}) = 0 \;\forall\; \boldsymbol{x}\in\Omega$.

In terms of notation, bold symbols represent vectors, and in particular, $\boldsymbol{x}$, $\boldsymbol{u}$, and $\boldsymbol{n}$ correspond to the state, control, and random disturbance vectors respectively (note that index t for these quantities has been omitted for ease of notation). The state space is Ω, and the control space is given by the intersection of $\mathcal{U}$ (an arbitrary set) with all the control vectors that satisfy the linear constraint $\boldsymbol{a}'\boldsymbol{u} \leq N$. The transition function $\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n})$ captures the dynamics of the model from period t to period $t-1$. The proofs we provide assume that the disturbance $\boldsymbol{n}$ takes values in a countable space (in order to avoid unnecessary technical details, see the discussion in §1.5 of Bertsekas 2001).

In our case, the linear constraint $\boldsymbol{a}'\boldsymbol{u} \leq N$ corresponds to a shelf space constraint but in general it can be regarded as some "coupling constraint" that is conveniently relaxed, 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 $\lambda_t(\boldsymbol{x})$ 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 $\boldsymbol{\lambda_t} = (\lambda_t(\cdot), \lambda_{t-1}(\cdot), \ldots, \lambda_1(\cdot))$.

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

$$H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{x}) = N\lambda_t(\boldsymbol{x}) + \max_{\mathbf{u}\in\mathcal{U}} \quad g_t(\boldsymbol{x},\boldsymbol{u}) - \lambda_t(\boldsymbol{x})\boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[H_{t-1}^{\boldsymbol{\lambda_{t-1}}}(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \tag{2.2}$$

with $H_0^{\boldsymbol{\lambda_0}}(\boldsymbol{x}) = 0 \;\forall\; \boldsymbol{x}\in\Omega$. In words, a dual policy gives a price (Lagrange multiplier) for a unit of shelf space for each period and each possible state.

[1] We present our results in a finite-horizon framework that fits the DP model to be introduced in the next chapter. Analogous results can be derived for other settings, in particular, for the infinite horizon case.

A dual policy $\boldsymbol{\lambda_t}$ is *optimal* if it minimizes the right hand side of (2.2) for any initial state. In line with standard dynamic programming theory, we recursively define $\lambda_t^*(\boldsymbol{x})$ to be the smallest solution of the following dual problem:

$$H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{x}) = \min_{\lambda_t \geq 0} \quad N\lambda_t + \max_{\mathbf{u} \in \mathcal{U}} \quad g_t(\boldsymbol{x}, \boldsymbol{u}) - \lambda_t \boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[H_{t-1}^{\boldsymbol{\lambda_{t-1}}}(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n}))\right], \quad (2.3)$$

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

The following proposition is an intuitive result that relates the primal and dual DP's; 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 $\boldsymbol{x}$: $J_t^*(\boldsymbol{x}) \leq H_t^*(\boldsymbol{x}) \leq H_t^{\boldsymbol{\lambda_t}}(\boldsymbol{x})$.*

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 partially solved by the following proposition:

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

$$f_\tau(\boldsymbol{x}'; C) = \max_{\substack{\mathbf{u} \in \mathcal{U}: \\ \boldsymbol{a}'\boldsymbol{u} = C}} \; g_t(\boldsymbol{x}', \boldsymbol{u}) + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x}',\boldsymbol{u})}\left[J_{\tau-1}^*(\varphi_t(\boldsymbol{x}', \boldsymbol{u}, \boldsymbol{n}))\right] \quad (2.4)$$

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

In contrast with (2.1), the parametric function defined by (2.4) 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.

Via a counterexample, it will be shown in the next chapter that strong DP duality does not always hold. However, in the rather few cases of the dynamic assortment problem when it does not apply, the duality gap is small.

2.2 Open-Loop Dual Policies

Solving the dual DP problem given by equation (2.3) seems just as hard as solving the original primal problem (2.1), motivating further simplifications. Specifically, we now

restrict our attention to *open-loop dual policies*, 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. We use the name *open-loop* to be consistent with the usual concepts in DP theory that makes a difference between the policies that depend on the system state (closed-loop) and those that do not (see p. 4 in Bertsekas 2001). Castañon (1997) calls the closed-loop dual policies *stochastic multipliers* and the open-loop policies *deterministic multipliers*. Karmarkar (1987) refers to the latter as "restricted Lagrangian".

In the following, we will use the notation $H_t^{\boldsymbol{\lambda}}(\cdot)$, instead of $H_t^{\boldsymbol{\lambda_t}}(\cdot)$, to denote the profit-to-go corresponding to an open-loop dual policy $\boldsymbol{\lambda}$.

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

$$J_t^*(\boldsymbol{x}) \leq \min_{\boldsymbol{\lambda} \geq \mathbf{0}} H_t^{\boldsymbol{\lambda}}(\boldsymbol{x}) \tag{2.5}$$

A better bound follows from using the best open-loop dual policy to approximate (for each state) the profit-to-go $J_{t-1}^*(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n}))$ in the Bellman equation (2.1), that is:[2]

$$J_t^*(\boldsymbol{x}) \leq \max_{\substack{\boldsymbol{u} \in \mathcal{U}: \\ \boldsymbol{a}'\boldsymbol{u} \leq N}} g_t(\boldsymbol{x}, \boldsymbol{u}) + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\big[\min_{\boldsymbol{\lambda} \geq \mathbf{0}} H_{t-1}^{\boldsymbol{\lambda}}(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n}))\big] \tag{2.6}$$

However, the expectation in (2.6) is not separable and its calculation seems very computationally intensive. By interchanging the order of the minimization and maximization operators in (2.6) we still have an upper bound and the problem becomes separable but also then equivalent to solving (2.5). The minimization with respect to $\boldsymbol{\lambda}$ in (2.5) can be solved with any convex non-differentiable optimization method, and yields the upper performance bound that we use extensively in this thesis.

It is interesting to note that finding the best open-loop dual policy, i.e. solving (2.5), is equivalent to solve the original (primal) problem but requiring the coupling constraint to be satisfied *on average* in each period, instead of having it satisfied for each possible sample-path. In other words, for each period, the constraint $\boldsymbol{a}'\boldsymbol{u} \leq N$ is replaced by $\mathbb{E}[\boldsymbol{a}'\boldsymbol{u} \leq N]$, where the expectation is with respect to all possible states weighted by the probability of reaching each one of them under a given (primal)

[2]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 $H_{t-1}^{\boldsymbol{\lambda}}(\cdot)$, $\boldsymbol{\lambda}$ is a vector with $(t-1)$ components.

policy. This fact has been observed by several authors in their particular applications (see Whittle 1988, Castañon 1997, and Talluri and van Ryzin 1998). We will show the equivalence by means of an example relevant to the dynamic assortment problem, that is the finite horizon multiarmed bandit problem with several plays per stage.

Consider S independent bandit machines. Let Ω_s be the set of all possible states of bandit s, that for any practical purposes can be assumed to be finite. The player has T periods in which he can play (at most) N arm. If a given state $i \in \Omega_s$ the arm of bandit s is pulled, then the player receives a reward equal to R_i and a transition to state $j \in \Omega_s$ occurs with probability p_{ij}. For simplicity (and also in agreement with the dynamic assortment problem) we assume that $R_i > 0 \ \forall i \in \bigcup_{s=1}^{S} \Omega_s$. The objective of the player is to maximize the total reward over T.

We consider the problem description in the previous paragraph to be the "original primal problem". Suppose now that the player is allowed to play (at most) N arms on average in each period. This relaxed version of the problem can be formulated as a linear program (LP). In fact, let $\boldsymbol{\pi}$ be any admissible policy (for the relaxed problem). For a given state $j \in \Omega_s$, let $\mathbb{I}_j(t)$ be the indicator function that equals one if the player pulls the arm of bandit s in state j at time t, and let $y_j^t(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}[\mathbb{I}_j(t)]$. Then $y_j^t(\boldsymbol{\pi})$ is a frequency measure that represents the expected number of times the player pulls the arm of bandit j at time t under policy $\boldsymbol{\pi}$. Let $\boldsymbol{x}^T$ be the initial state ($x_j^T = 1$ if bandit s starts at state $j \in \Omega_s$), then the player solves the following problem:

$$J_T^r(\boldsymbol{x}^T) = \max_{\boldsymbol{\pi}} \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{j \in \Omega_s} R_j y_j^t(\boldsymbol{\pi}) \tag{2.7}$$

It can be shown that problem (2.7) is equivalent to solving the following LP (see Bertsimas and Ñiño-Mora 2000):

$$
\begin{aligned}
J_T^r(\boldsymbol{x}^T) \;=\; & \max \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{j\in\Omega_s} R_j y_j^t && (2.8)\\
& \text{subject to} \\
& y_j^{t-1} + \overline{y}_j^{t-1} = \sum_{i\in\Omega_s} p_{ij} y_i^t + \overline{y}_j^t \qquad \forall s, \forall j \in \Omega_s, t = 2,\ldots,T \\
& y_j^T + \overline{y}_j^T = x_j^T \qquad \forall s, \forall j \in \Omega_s \\
& \sum_{s=1}^{S}\sum_{j\in\Omega_s} y_j^t \le N \qquad t = 1,\ldots,T && (2.9)\\
& y_j^t, \overline{y}_j^t \ge 0 \qquad \forall s, \forall j \in \Omega_s, t = 1,\ldots,T
\end{aligned}
$$

The constraints (2.9) ensure that the on average the player does not pull more than N arms in every period. We now relax ("dualize") that constraint using a vector of multipliers $\boldsymbol{\lambda}$. Let $H_T^{\lambda}(\boldsymbol{x}^T)$ be the optimal value of the LP with the new objective function subject to all the other constraints. From standard LP duality theory we have that $J_T^r(\boldsymbol{x}^T) = \min_{\boldsymbol{\lambda}\ge\mathbf{0}} H_T^{\lambda}(\boldsymbol{x}^T)$. Following the same steps that led to the LP formulation (2.8), it is easy to verify that $H_T^{\lambda}(\boldsymbol{x}^T)$ corresponds to the profit-to-go reported by the dual DP (2.2) under the open-loop policy $\boldsymbol{\lambda}$. Hence, via an example (that can be generalized), we have shown that when finding the best open-loop policy we are actually solving a relaxed version of the original problem in which the coupling constraint only has to be satisfied on average each period.

In the next chapters we will also be interested in finding the best *stationary* open-loop policy (i.e. $\boldsymbol{\lambda} = (\lambda, \lambda, \ldots, \lambda)$, for some scalar $\lambda \ge 0$). It is now easy to see that this is equivalent to solving the original problem but requiring the coupling constraint to be satisfied on average over the whole horizon T. The reasoning is the same as above but replacing (2.9) with the constraint:

$$\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{j\in\Omega_s} y_j^t \le N \cdot T. \tag{2.10}$$

As mentioned before, the upper bound obtained from considering open-loop dual policies will be used later to asses the suboptimality of some heuristic policies. Then knowing the quality of the bound would be relevant. In that respect, Adelman and Mersereau (2004) provide an alternative LP-based bound that is shown to be tighter

(or no worse) than the bound given by equation (2.5). However, the computation of their bound is more demanding. Finally, there is some evidence showing that the open-loop dual policy bound can be "asymptotically" tight. In fact, Weber and Weiss (1990) prove this result (under certain regularity conditions, not easily verified) for the average reward, infinite horizon, restless bandit. In their case, the asymptotic regime corresponds to N and S tending to infinity while the ratio N/S remains fixed. The parameters N and S are the same as in the example given above (which can also be seen as a restless bandit but with finite horizon). When the regularity conditions are not met, they claim that the "size of the suboptimality which one might expect is minuscule". In a finite horizon network revenue management setting, Talluri and van Ryzin (1998) also show that the bound (2.5) is asymptotically tight when the initial leg capacities and sales volumes are scaled to infinity. An interesting open research topic would be to find general conditions for this asymptotic result to hold.

Chapter 3

Model without Lost Sales

In this chapter, we formulate the basic dynamic assortment model in §3.1, then discuss its applicability and justify our assumptions in §3.2. Throughout the remaining of the thesis all symbols in boldface represent vectors, subscripts represent the components of a vector, and superscripts represent elements in a sequence.

3.1 Model Definition

3.1.1 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 the introduction, 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 chapter is to perform our analysis in subsections §3.3.1 to §3.3.4 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 §3.3.5.

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.

3.1.2 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. As a consequence, the learning process for a given product is not affected by the other products that might be included in the assortment.

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 (3.1) must be replaced with the corresponding $\Gamma(\cdot)$ terms, and the interpretation as a negative binomial (to be given) would not be valid.

$$\Pr(n_s) = \binom{n_s + m_s - 1}{m_s - 1} \left(\frac{1}{\alpha_s + 1}\right)^{n_s} \left(\frac{\alpha_s}{\alpha_s + 1}\right)^{m_s}, \tag{3.1}$$

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:

$$(m_s, \alpha_s) \longrightarrow \begin{cases} (m_s + n_s, \alpha_s + 1) & \text{If product } s \text{ is in the assortment and } n_s \text{ sales} \\ & \text{are observed during period } t \\ (m_s, \alpha_s) & \text{If product } s \text{ is not in the assortment} \end{cases}. \tag{3.2}$$

The intuition for the update procedure (3.2) 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 $1/\sqrt{m_s}$.

3.1.3 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 $\mathcal{U} = \{\boldsymbol{u} \in$

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

$\{0,1\}^S : \sum_{s=1}^S u_s \leq N\}$.

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:

$$J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) = \max_{\substack{\boldsymbol{u}\in\{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\right], \tag{3.3}$$

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 (3.1).

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 (3.3) 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.

3.2 Model Discussion

This section 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.

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

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

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. We refer the reader to the discussion in Chapter 4, where we also show how to use the index policy (to be derived) in a setting with substitution effects.

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). In Chapter 5 we introduce models that do consider lost sales and there we resume the current discussion.

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 Chapter **??** 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. In Chapter 6 a model with variable demand rates is presented and briefly discussed.

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 §3.3) constitutes a useful starting point for designing heuristics or developing extensions in more complex environments, as discussed in the next chapters. For example, we describe in Chapter 4 a heuristic procedure for capturing substitution effects that is based on the analysis of our basic model.

3.3 Analysis

3.3.1 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 (3.1) 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):

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

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

$$\max_{\sum_{s=1}^{S} u_s \leq N} \sum_{s=1}^{S} r_s \mathbb{E}[\gamma_s] u_s \;\leq\; \frac{J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})}{t} \;\leq\; \mathbb{E}_{\boldsymbol{\gamma}(\boldsymbol{m}, \boldsymbol{\alpha})}\Big[\max_{\sum_{s=1}^{S} u_s \leq N} \sum_{s=1}^{S} r_s \gamma_s u_s \Big], \qquad (3.4)$$

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 (3.4) 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 the planning horizon goes to infinity. Empirical evidence and intuition suggest that it converges to the right hand side of (3.4); 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.

3.3.2 Remarks on the Dual Dynamic Program

The optimal dynamic assortment policy may conceptually be derived from the dynamic programming equation (3.3). 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 (3.3) (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.

We follow the solution approach outlined in Chapter 2. In the current subsection we give some further remarks related to the particular dual DP obtained in the dynamic assortment problem, and in the next subsection we show how the problem decomposes when open-loop dual policies are considered.

Consider the parametric function defined in Proposition 2 (strong DP duality). The following lemma shows that at least the increasing monotonicity is guaranteed.

Lemma 3 *If $r_s > 0\ \forall s$, then $f_t(\boldsymbol{m}, \boldsymbol{\alpha}; C)$ is a strictly increasing function of C, with $C \leq S$, for any state $(\boldsymbol{m}, \boldsymbol{\alpha})$.*

The lemma reflects 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 pending concavity 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 counterexample showing that strong duality does not hold in general absent such a condition: For $t = 2$, $S = 2$, $N = 1$, $r_1 = r_2 = 1$, $m_1 = 44$, $m_2 = 4$, $\alpha_1 = 10$, 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 (2.4) 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 (2.4) 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 (2.4) 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 (2.3), 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.

3.3.3 Single Product Subproblem

The next Lemma shows that with open-loop policies the dual DP decomposes into S single-product subproblems:

Lemma 4 *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:*

$$H_t^{\boldsymbol{\lambda}}(\boldsymbol{m}, \boldsymbol{\alpha}) = N \sum_{\tau=1}^{t} \lambda_\tau + \sum_{s=1}^{S} H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) \tag{3.5}$$

where:

$$H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) = \max \left\{ \underbrace{r_s \frac{m_s}{\alpha_s} - \lambda_t + \mathbb{E}_{n_s}\left[H_{t-1,s}^{\boldsymbol{\lambda}}(m_s + n_s, \alpha_s + 1)\right]}_{u_s=1}, \underbrace{H_{t-1,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)}_{u_s=0} \right\} \tag{3.6}$$

The single-product subproblem defined by (3.6) 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. For a given open-loop dual policy $\boldsymbol{\lambda}$, the values $H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s)$ can be calculated efficiently in a standard recursive fashion. That is how we proceeded in our numerical experiments, but an alternative would be to adapt the (polynomially solvable) LP formulation obtained by Bertsimas and Ñiño-Mora (1996) for the infinite horizon case.

It is clear from (3.6) 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 (3.6)), 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 $\beta^{\boldsymbol{\lambda}}_{t,s}(\cdot)$ such that at state (m_s, α_s) the optimal policy for the single-product subproblem (3.6) is: $u_s = 1 \Longleftrightarrow \alpha_s \leq \beta^{\boldsymbol{\lambda}}_{t,s}(m_s)$*

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 $B^{\boldsymbol{\lambda}}_{t,s} \subseteq B^{\boldsymbol{\lambda}}_{t-1,s}$.*

When $\lambda_t \leq \lambda_{t-1}$ however, Propositions 3 and 4 imply that the optimal policy for (3.6) 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 (3.6) 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).

The inclusions of the stopping sets $B^{\boldsymbol{\lambda}}_{t,s}$ are not reverted when $\lambda_t > \lambda_{t-1}$ since the threshold functions $\beta^{\boldsymbol{\lambda}}_{\cdot,s}(m_s)$ might cross then. However, this can only happen for low values of m_s. In fact, the following Corollary (stated without proof) shows that the threshold functions are linear for large values of m_s:

Corollary 1 *If $\lambda_t > \lambda_q\ \forall q \leq t-1$, and $m_s \geq \frac{\lambda_q(t-q)}{r_s\left(1-\frac{\lambda_q}{\lambda_t}\right)}\ \forall q \leq t-1$, then $\beta^{\boldsymbol{\lambda}}_{t,s}(m_s) = r_s m_s/\lambda_t$.*

In Figure 3-1 we plot the threshold functions for a 3-period problem with $r_s = 1$. The top graph has $(\lambda_1, \lambda_2, \lambda_3) = (4, 5, 6)$ and the bottom one has $(\lambda_1, \lambda_2, \lambda_3) = (4.6, 4.8, 5.0)$. The continuous lines are obtained by interpolating the function values for the integer points. In the bottom graph the threshold functions intersect around $m_s \approx 13$.

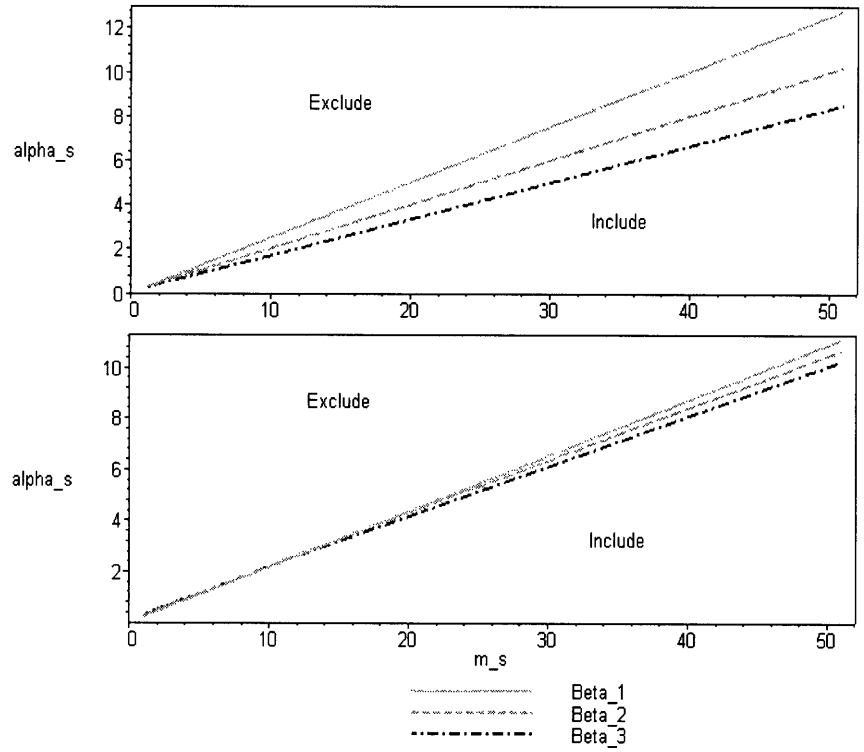

Figure 3-1: The threshold functions when $\lambda_3 > \lambda_2 > \lambda_1$.

3.3.4 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 (3.6) is called in that case by Gittins a standard arm, and it follows from Proposition 4 that:

$$H^{\lambda}_{t,s}(m_s, \alpha_s) = \max\left\{r_s \frac{m_s}{\alpha_s} - \lambda + \mathbb{E}_{n_s}\left[H^{\lambda}_{t-1,s}(m_s + n_s, \alpha_s + 1)\right], 0\right\}. \quad (3.7)$$

- 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:

$$\tilde{H}^{\lambda}_{t-1,s}(m_s, \alpha_s) = (t-1) \cdot \max\left\{ r_s \frac{m_s}{\alpha_s} - \lambda, 0 \right\}. \tag{3.8}$$

Substituting (3.8) in (3.7) 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:

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

$$\text{where } b^{\lambda}_s = \left(\frac{\lambda}{r_s} - \frac{m_s}{\alpha_s} \right) \frac{\alpha_s \sqrt{\alpha_s + 1}}{\sqrt{m_s}}, \ \mathbb{E}[n_s] = \frac{m_s}{\alpha_s}, \text{ and } \mathbb{V}[n_s] = \mathbb{E}[n_s]\left(\frac{\alpha_s + 1}{\alpha_s} \right). \tag{3.9}$$

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:

$$\tilde{d}^{\lambda}_{t,s}(m_s, \alpha_s) \approx \frac{r_s \sqrt{m_s}}{\alpha_s \sqrt{\alpha_s + 1}} \left((t-1) \cdot \Psi(b^{\lambda}_s) - b^{\lambda}_s \right), \tag{3.10}$$

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

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

247), the equation

$$(t-1)\cdot\Psi(z_t) = z_t \tag{3.11}$$

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 3.1 and Figure 3-2 for the first few numerical values of z_t with four digits accuracy.

The policy for problem (3.6) 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:

$$H_{t,s}^\lambda(m_s,\alpha_s) \approx \frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}}\left[(t-1)\cdot\Psi(b_s^\lambda) - b_s^\lambda\right]^+. \tag{3.12}$$

t	z_t	t	z_t
2	0.2760	8	0.8168
3	0.4363	9	0.8616
4	0.5492	10	0.9014
5	0.6360	11	0.9373
6	0.7065	12	0.9700
7	0.7658	13	0.9999

Table 3.1: First values of z_t.

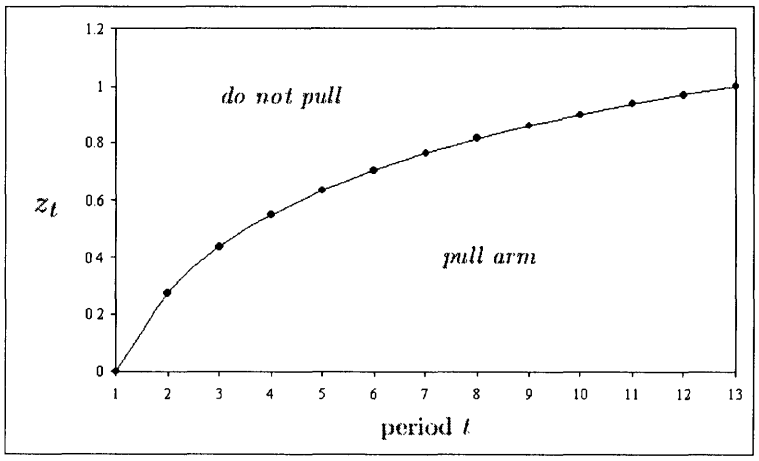

Figure 3-2: Plot of z_t as a function of t.

Second Step: Linear Search in λ

We now adapt to our problem a heuristic solution method initially developed by Castañon (1997) that is essentially analogous to the heuristic method proposed by Whittle (1998) for the infinite horizon restless bandit. 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 (3.7). 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 (3.7) 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:

$$
\begin{array}{lll}
& u_{t,s}^{\lambda} = 1 & \\
\Leftrightarrow & \frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}}\left[(t-1)\cdot\Psi(b_s^{\lambda}) - b_s^{\lambda}\right] \geq 0 & \text{by definition of } \tilde{d}_{t,s}^{\lambda}(m_s,\alpha_s) \text{ in (3.10);} \\
\Leftrightarrow & b_s^{\lambda} \leq z_t & \text{by definition of } z_t \text{ in (3.11);} \\
\Leftrightarrow & \lambda \leq r_s\frac{m_s}{\alpha_s} + z_t\frac{r_s\sqrt{m_s}}{\alpha_s\sqrt{\alpha_s+1}} & \text{by definition of } b_s^{\lambda} \text{ in (3.9).}
\end{array}
\tag{3.13}
$$

Substituting in the last expression of (3.13) the moments of γ_s (given at the end of §3.1.2) and introducing the length of each period δ (assumed with no loss of generality to equal one everywhere else) in order to avoid the appearance of a unit inconsistency, we finally obtain the following approximation for index $\eta_{t,s}$:

$$
\eta_{t,s} \approx r_s\mathbb{E}[\gamma_s]\delta + z_t\frac{r_s\mathbb{V}[\gamma_s]\delta}{\sqrt{\mathbb{V}[\gamma_s] + \frac{\mathbb{E}[\gamma_s]}{\delta}}} \tag{3.14}
$$

$$
= r_s\left(\mathbb{E}[\gamma_s]\delta + z_t\frac{\mathbb{V}[\gamma_s]\delta^2}{\sqrt{\mathbb{V}[n_s]}}\right) \tag{3.15}
$$

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 (3.14). 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 (3.14) 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 (3.14).

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.

When the index formula is expressed as in equation (3.15) a salient feature is made evident: learning term plays a role only if the uncertainty regarding the arrival rate ($\mathbb{V}[\gamma_s]$) is relevant with respect to the total uncertainty on demand ($\mathbb{V}[n_s]$). The latter is affected by the fact that arrival rate γ_s is unknown, and also by the inherent randomness of having stochastic demand, but only the first type of uncertainty can be resolved by means of an dynamic assortment policy.

As an example, Table 3.2 describes some initial state data for which Figure 3-3 illustrates the behavior of our proposed index policy. In particular, Figure 3-3 shows that the index policy generates different assortments depending on the length of the planning horizon, as opposed to the greedy policy that always select products as if there were only one period left to go. With $N = 2$ for instance, the index policy

s	r_s	m_s	α_s	$\mathbb{E}[\gamma_s]$	$\mathbb{V}[\gamma_s]$
1	4.0	4	0.98	4.1	4.2
2	4.0	6	1.44	4.2	2.9
3	4.0	9	2.10	4.3	2.0
4	3.9	3	0.74	4.1	5.5

Table 3.2: Data of the example in Figure 3-3.

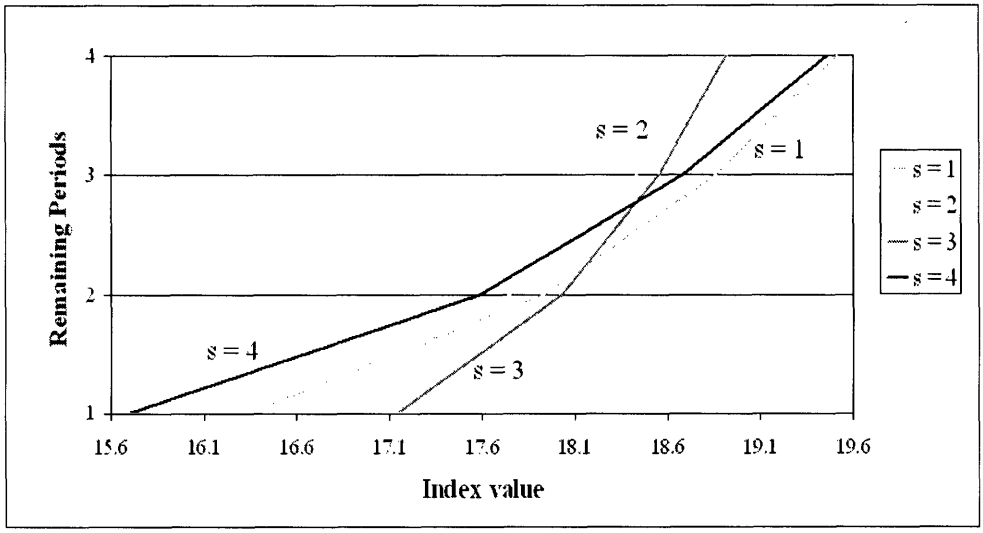

Figure 3-3: Graphical representation of the proposed index policy.

would select products 2 and 3 if $t = 1$, products 1 and 3 if $t = 2$, and products 1 and 4 if $t = 3$. In contrast, the greedy policy would always select products 2 and 3, regardless of the number of periods remaining and the variance of the priors.

It might seem counterintuitive that the second term in the index formula (3.14) is decreasing with respect to the expected demand rate $\mathbb{E}[\gamma_s]$. However, a similar situation is observed in other approximations of the Gittins index (for instance, see equation (16) in Brezzi and Lai 2002), and it is due to the particular relation between the expectation and variance for a Poisson distribution.[5] Moreover, it is easy to prove that, for $\mathbb{E}[\gamma_s] \geq \frac{z_t^2}{27\delta_t}$, the index formula is increasing in the expectation of the demand rate (regardless of the variance). Note that this covers all reasonable values of the

[5]The approximation to the Gittins index given by equation (16) in Brezzi and Lai 2002 is derived for the general infinite horizon case with independent arms. If the rewards follow a Poisson distribution, the learning term in the approximate index formula is increasing in the variance and decreasing in the expectation of the unknown parameter.

expectation. In fact, for $t = 24$ (and $\delta_t = 1$), the index formula is strictly increasing in the range $\mathbb{E}[\gamma_s] \geq 0.06$. Then, given two products with the same net margin and variance, the retailer would prefer the one with a higher expected demand.

In the case when periods are not equal, for example when all demand rates are affected by the same inflation/deflation factors during the season, the derivation of the index formula is the same as above but using the following limited lookahead horizon approximation:

$$\bar{H}^{\lambda}_{t-1,s}(m_s, \alpha_s) = \frac{\sum_{\tau=1}^{t-1} \delta_\tau}{\delta_t} \max\left\{ r_s \frac{m_s}{\alpha_s} - \frac{\lambda}{\delta_t}, 0 \right\}$$

where δ_t is the length of period t. We obtain the same index formula as in equation (3.14) but with δ_t replacing δ, and z_t comes from solving the equation $z_t = \frac{\sum_{\tau=1}^{t-1} \delta_\tau}{\delta_t} \cdot \Psi(z_t)$.

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.*

3.3.5 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".

The existence of a positive lead time makes the learning process slower since anything that has been learned about demand can only be implemented ℓ periods later. As a consequence, the number of remaining effective "learning periods" is reduced to $t - \ell - 1$. This has a direct impact on expected profits as it can be seen in the following Lemma (the proof is by induction):

Lemma 5 *Let $J_t^\ell(\boldsymbol{m}, \boldsymbol{\alpha})$ denote the optimal expected profit-to-go when the lead time is equal to ℓ periods. Let $\ell' = \left(t - \lfloor \frac{t}{\ell+1} \rfloor (\ell+1)\right)$. Then for any state $(\boldsymbol{m}, \boldsymbol{\alpha})$ and horizon length t:*

$$J_t^0(\boldsymbol{m}, \boldsymbol{\alpha}) \geq J_t^\ell(\boldsymbol{m}, \boldsymbol{\alpha}) \geq \ell' J^0_{\lceil \frac{t}{\ell+1} \rceil}(\boldsymbol{m}, \boldsymbol{\alpha}) + (\ell - \ell' + 1) J^0_{\lfloor \frac{t}{\ell+1} \rfloor}(\boldsymbol{m}, \boldsymbol{\alpha}) \tag{3.16}$$

The left hand side of (3.16) confirms the fact that with a positive lead time the expected profit can only deteriorate, and the right hand side shows that the store can expect to do better (but never worse) than solving ℓ' and $(\ell - \ell' + 1)$ independent subproblems with zero lead time and planning horizons equal to $\lceil \frac{t}{\ell+1} \rceil$ and $\lfloor \frac{t}{\ell+1} \rfloor$ respectively.[6]

The incorporation of lead times is a common practice in the Operations Management literature. The standard procedure is to expand the state definition. In some few cases, a transformation is possible so that the problem with lead times can be reduced to one with essentially no lead times. The most well known example is the single installation inventory control problem with i.i.d. demands and backlogs (see Clark and Scarf 1960). However, in that case the transformation relies heavily on the backlog assumption and the fact that demand distributions are known in advance. In the dynamic assortment problem, such a transformation is not available and we must extend the state definition to keep track of past decisions.

We now reformulate the dynamic program. With a positive lead time ℓ the state is 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

[6] In terms of notation, $\lfloor x \rfloor$ is equal to the largest integer lower than or equal to x, and $\lceil x \rceil$ is equal to x if x is integer, otherwise $\lceil x \rceil = \lfloor x \rfloor + 1$.

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:

$$J_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) = \sum_{s=1}^{S} \sum_{\tau=t-\ell+1}^{t} r_s \frac{m_s}{\alpha_s} v_s^\tau + W_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) \quad (3.17)$$

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

$$W_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) = \max_{\sum_{s=1}^{S} u_s \leq N} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\Big[W_{t-1}^*(\boldsymbol{v}^{t-1}, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{u}, \boldsymbol{m}+\boldsymbol{n}\cdot\boldsymbol{v}^t, \boldsymbol{\alpha}+\boldsymbol{v}^t)\Big] \quad (3.18)$$

The summation in the right hand side of (3.17) 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 (3.18) reduces then to the recursion (3.3) 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 (3.18):

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

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

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

If the components of the open-loop dual policy are nondecreasing (i.e. $\lambda_t \leq \lambda_{t-1}\ \forall t$), then a *stopping time* property similar to Proposition 4 can be formulated:

Proposition 5 *Consider a nondecreasing open-loop dual policy. At period t, if product s will not be part of the assortment in the next ℓ periods (i.e. $v_s^{t-\ell+1} = \ldots = v_s^t = 0$), and it is not optimal to include it in the assortment of period $t-\ell$ (i.e. the second term in the right hand side of the equation above is optimal), then it is also optimal not to include product s in the assortment of any period beyond $t-\ell$.*

In other words, Proposition 5 says that, if product s will not be part of the assortment in the next ℓ periods, and it is optimal not to include product s in period $t-\ell$, then it will be also optimal not to include product s in the assortment of periods beyond $t-\ell$. If product s will be included at least once in the next ℓ periods (equivalently, if the condition $v_s^{t-\ell+1} = \ldots = v_s^t = 0$ does not hold), then Proposition 5 is no longer valid. If fact, consider for example an instance with three periods ($t = 3$), one period of design-to-shelf lead time ($\ell = 1$), an open-loop dual policy with $\lambda_2 \leq \lambda_1$, and parameters for a product s such that $r_s \frac{m_s}{\alpha_s} < \lambda_2$. Assume that product s will be part of the assortment in period 3 ($v_s^3 = 1$). Since the expected immediate profit $r_s \frac{m_s}{\alpha_s}$ is less than the opportunity cost λ_2, and since after period 2 there are no more decisions to be made, it can be seen (from solving single-product maximization above for $t = 3$) that product s should not be included in the assortment of period 2 (i.e. $u_s = 0$ is optimal). However, the optimal decision in period 2 (for period 1) will depend on n_s, the sales of product s observed during period 3. Clearly, if n_s is large enough so that $r_s \frac{m_s+n_s}{\alpha_s+1} > \lambda_1$, then it will be optimal to include product s in period 1. By the contrary, if $v_s^3 = 0$, meaning that product s will not be part of the assortment in period 3, then $u_s = 0$ is optimal in periods 3 and 2 (for periods 2 and 1 respectively) as stated by Proposition 5.

We can invoke arguments similar to the ones used in section §2.2 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:

$$\min_{\boldsymbol{\lambda}} N \sum_{\tau=1}^{T} \lambda_\tau + \sum_{s=1}^{S} \max_{\substack{v_s^T, \ldots, v_s^{T-\ell+1} \\ \in \{0,1\}}} \left(\sum_{\tau=T-\ell+1}^{T} \left(r_s \frac{m_s}{\alpha_s} - \lambda_\tau \right) v_s^\tau + H_{t,s}^{\boldsymbol{\lambda}}(v_s^T, \ldots, v_s^{T-\ell+1}, m_s, \alpha_s) \right), \tag{3.19}$$

which provides the upper bound that we will report for the performance of various policies simulated in Section §3.5 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 (3.14):

1. First, we substitute the term z_t in (3.14) with

$$z_t \quad \longrightarrow \quad z_{L(t)}, \tag{3.20}$$

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 (3.14) concerns the variance $\mathbb{V}[\gamma_s]$. Recall from section §3.1.2 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:

$$\mathbb{V}[\gamma_s] = \frac{m_s}{\alpha_s^2} \quad \longrightarrow \quad \mathbb{V}[\gamma_s] = \frac{m_s + \frac{m_s}{\alpha_s} \sum_{\tau=t-\ell+1}^{t} v_s^\tau}{(\alpha_s + \sum_{\tau=t-\ell+1}^{t} v_s^\tau)^2}, \tag{3.21}$$

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 (3.21) 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 (3.21), the second term in the index formula (3.14) 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 section §3.5 we report the performance achieved by the heuristic policy just described in various numerical experiments.

3.4 Demand Distribution from the Exponential Family

The DP model introduced in section §3.1.3 was derived under the assumption that the demand learning process has a Poisson-Gamma structure. We will show that the model is essentially the same for a much broader class of demand distributions and initial priors.

Consider the exponential family of distributions with one unknown parameter. Suppose that the demand distribution of each product s belongs to that family, and let γ_s be the unknown parameter. In other words, the probability (density) function of product s can be written as:

$$dF(n_s|\gamma_s) = \exp\big(a_s(n_s) + b_s(\gamma_s) + c_s(n_s)d_s(\gamma_s)\big) \tag{3.22}$$

where $a_s(\cdot), b_s(\cdot), c_s(\cdot)$, and $d_s(\cdot)$ are known functions. If $c_s(n_s) = n_s$, then the distribution is said to be in *canonical form.*

The following distributions belong to this family and are in canonical form: (1) the normal distribution with known variance (γ_s is the mean); (2) the binomial and negative binomial distributions, in this case γ_s corresponds to the success probability of the Bernoulli trials; (3) the Poisson distribution, with γ_s being the rate (or mean); (4) the Gamma distribution with an unknown scale parameter γ_s. The lognormal and Weibull distributions are also part of the exponential family but they are not in canonical form.

Demand distributions from the exponential family have been widely used in inventory/retailing problems that involve learning (see for instance Azoury 1985 and Eppen and Iyer 1997), the main reason being the following property that can be easily

verified: Let $n^1, n^2, \ldots, n^\alpha, n^{\alpha+1}$ be a sample of i.i.d. random variables from a distribution that belongs to the exponential family. Let γ be the corresponding unknown parameter, and let $W(\gamma)$ be the (initial) prior distribution of γ with support Θ. Then we have that:

$$\begin{aligned}\Pr\left(n^{\alpha+1} \middle|\, n^1, \ldots, n^\alpha\right) &= \frac{\int_\Theta \exp\left(a(n^{\alpha+1}) + (\alpha+1)\cdot b(\gamma) + d(\gamma)\sum_{\tau=1}^{\alpha+1} c(n^\tau)\right) W(\gamma)d\gamma}{\int_\Theta \exp\left(\alpha\cdot b(\gamma) + d(\gamma)\sum_{\tau=1}^{\alpha} c(n^\tau)\right) W(\gamma)d\gamma} \\ &= \Pr\left(n^{\alpha+1} \middle|\, \sum_{\tau=1}^{\alpha} c(n^\tau), \alpha\right)\end{aligned}$$

From the equation above it is clear that, for any initial prior $W(\gamma)$, the pair $\left(\sum_{\tau=1}^{\alpha} c(n^\tau), \alpha\right)$ is a sufficient statistic for the posterior distribution of $n^{\alpha+1}$ given a sample of past observations of size α. This result allows us to extend our basic dynamic assortment model to any demand distribution from the exponential family with arbitrary priors. In the Bellman equation (3.3) the fraction m_s/α_s must be replaced by $\mathbb{E}[\nu(\gamma_s)]$, where $\nu(\gamma_s) = \mathbb{E}[n_s|\gamma_s]$, and the update rule is now:

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

The duality results developed in Chapter 2 still apply and the Lagrangian relaxation can be used to calculate the corresponding Gittins indices and the upper bound, thought the solving the single-product subproblems can be more demanding. We might not be able to follow the same step used in the derivation of the closed-form index formula (3.14). However, we can easily adapt it to this more general case. In fact, let $\sigma^2(\gamma_s) = \mathbb{E}[n_s^2|\gamma_s] - \nu^2(\gamma_s)$ be the (inherent) variance of the stochastic demand. Note that if the demand distribution is in canonical form, then we have that $\sigma^2(\gamma_s) = \frac{d_s''(\gamma_s)b_s'(\gamma_s) - d_s'(\gamma_s)b_s''(\gamma_s)}{d_s'(\gamma_s)^3}$. The equivalent index formula can be written as follows:

$$\eta_{t,s} \approx r_s\mathbb{E}[\nu(\gamma_s)] + z_t\frac{r_s\mathbb{V}[\nu(\gamma_s)]}{\sqrt{\mathbb{E}[\sigma^2(\gamma_s)] + \mathbb{V}[\nu(\gamma_s)]}} \tag{3.23}$$

$$= r_s\left(\mathbb{E}[\nu(\gamma_s)] + z_t\sqrt{\mathbb{V}[\nu(\gamma_s)]}\cdot\left(\sqrt{\frac{\mathbb{V}[\nu(\gamma_s)]}{\mathbb{V}[n_s]}}\right)\right). \tag{3.24}$$

As an example, suppose that demand for product s is normal with unknown mean γ_s and known variance ρ^2. In this case, the index formula would be:

$$\eta_{t,s} \approx r_s\mathbb{E}[\gamma_s] + z_t\frac{r_s\mathbb{V}[\gamma_s]}{\sqrt{\rho^2 + \mathbb{V}[\gamma_s]}}.$$

3.5 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 §2.2 and §3.3.5. Throughout this section we assume the Poisson demand model. We describe our methodology in §3.5.1, then discuss our experimental results in §3.5.2 to §3.5.6.

3.5.1 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 §3.5.2 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 (3.1)) 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 (3.3), independently of the Bayesian framework under which it has been derived.

Subsection §3.5.3 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).

For completeness, we define the frequentist approach in more formal terms. Let $\boldsymbol{\pi} = (\boldsymbol{\mu}^t, \boldsymbol{\mu}^{t-1}, \ldots, \boldsymbol{\mu}^1)$ be any feasible policy for the dynamic assortment problem.[7] For a given vector of demand rates $\boldsymbol{\gamma}$, a given policy $\boldsymbol{\pi}$, and an initial information vector $\mathbf{I}^t$, we define the (frequentist) regret as the difference between the expected profit of the optimal assortment with full information (i.e. knowing $\boldsymbol{\gamma}$) and the performance of policy $\boldsymbol{\pi}$. Formally, we define:

$$R_t(\boldsymbol{\gamma}) = t\mu^*(\boldsymbol{\gamma}) - S_t^{\boldsymbol{\pi}}(\boldsymbol{\gamma}) \tag{3.25}$$

$$\text{where } \mu^*(\boldsymbol{\gamma}) = \max_{\boldsymbol{u}\in\mathcal{U}} \sum_{s=1}^{S} r_s\gamma_s u_s, \text{ and } S_t^{\boldsymbol{\pi}}(\boldsymbol{\gamma}) = \mathbb{E}\Big[\sum_{\tau=1}^{t}\sum_{s=1}^{S} r_s n_s^\tau \mu_s^\tau(\mathbf{I}^\tau)\Big|\boldsymbol{\gamma}\Big].$$

For each product s, the random variables n_s^τ are i.i.d. following a Poisson distribution with mean γ_s. The information vector is updated according to the equation $\mathbf{I}^{\tau-1} = \mathbf{I}^\tau \cup \{\boldsymbol{n}^\tau, \boldsymbol{\mu}^\tau(\mathbf{I}^\tau)\}$. The Bayesian formulation, analyzed in the previous sections, is obtained by integrating $\int S_t^{\boldsymbol{\pi}}(\boldsymbol{\gamma})dF(\boldsymbol{\gamma})$, where $F(\boldsymbol{\gamma})$ is a prior distribution for the unknown demand rates.

The goal of the frequentist version of the multiarmed bandit problem is minimize regret, which is the same as maximizing the expectation on the right of equation (3.25). Note that it is impossible to achieve this objective uniformly over all parameter configurations (cf. Anantharam *et al.* 1987, p. 969), therefore usually the search is narrowed to policies that are *uniformly good.* A policy is uniformly good if $R_t(\boldsymbol{\gamma}) = \mathcal{O}(\log t)$. Many simple policies are not uniformly good. For example, the policy that always plays the first N arms is $\mathcal{O}(t)$. The greedy policy is also $\mathcal{O}(t)$ (see Kumar 1985). For the case of the multiarmed bandit with multiple plays per state, Anatharam *et al.* (1987) derive a policy (or allocation rule, in their terms) that is

[7] To be precise, $\boldsymbol{\pi}$ is a sequence of random vectors with values in $\mathcal{U}$ such that the event $\{\boldsymbol{\mu}^t = \boldsymbol{u}\}$ with $\boldsymbol{u} \in \mathcal{U}$ is measurable with respect to the σ-field generated by all past observations and assortments.

asymptotically uniformly good (as t goes to infinity), and which is also asymptotically efficient, meaning that it achieves a lower bound for all uniformly good policies.

Even though our models are formulated in a Bayesian framework, and developing a theoretical extension to the frequentist viewpoint goes beyond the scope of this thesis, we are still interested in using this approach as another way of testing the goodness of our index policy versus the greedy rule. We did not implement the policy suggested by Anatharam *et al.* (1987) since it requires at least $S \cdot N$ periods in order to have N initial observations per arm (cf. p. 972). In the assortment problem we have described, typically the number of periods T is in the order of S/N.

We used similar data sets for the experiments reported in §3.5.2 and §3.5.3. 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 §3.2). 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 3.3.

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

Table 3.3: 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 open-loop dual policy upper bounds 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. (2.5)) 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 (2.5) correspond to open-loop dual policies). In some instances we tried different starting points for this algorithm, and report then the best bounds we have found.

3.5.2 Bayesian Experiments

Table 3.4 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.

Over the range of scenarios considered in Table 3.4, 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 3.4 appears more clearly in Figure 3-4, 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

$\mathbb{V}[\gamma_s]$	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 3.4: Index policy vs. greedy rule (Bayesian approach).

implementation lead time ($\ell = 0$).

In line with previous results, the expected profit per period shown in Figure 3-4 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 similar findings. In addition, while the performance of both policies appearing in Figure 3-4 for a single decision period ($T = 1$) is by definition exactly identical to that of the static assortment reported in Table 3.3, 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

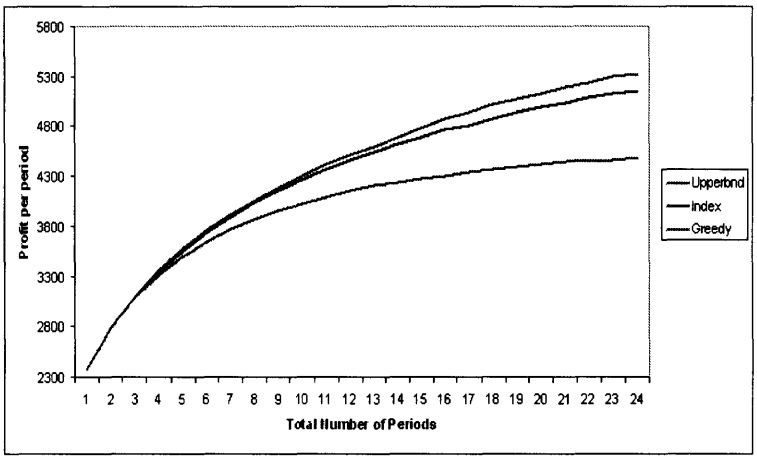

Figure 3-4: Relative policy performance for various horizon lengths.

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 3.4 and Figure 3-4 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 3.3 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 3.4 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 3-5 the performance of the index and greedy policies as well as the corresponding upper bound 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 §3.2).

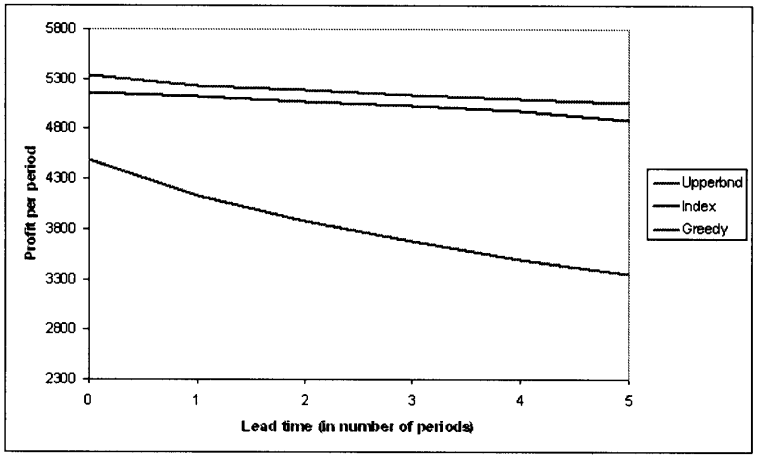

Figure 3-5: Relative policy performance for various lead times.

The performance of both policies as well as the upper bound values shown in Figure 3-5 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 3-4 thus indirectly follow from the same phenomenon. Also, the results shown in Figure 3-5 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 (3.21)), thus yielding a larger relative advantage to active learning over passive learning.

3.5.3 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 3.5 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.

The results shown in the seventh column of Table 3.5 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 3.5 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 3.5 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

$\mathbb{V}[\gamma_s]$	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 3.5: Index policy vs. greedy rule (frequentist approach).

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 3.6, which has the same structure as Table 3.5, shows our results for $M = 3$.

As shown in Table 3.6, 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 3.5. The main insight we thus draw from Table 3.6 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

$\mathbb{V}[\gamma_s]$	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 3.6: Relative policy performance with improved accuracy of initial information.

confronted with similar demand patterns may be very large in practice (see Chapter 4 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 3.7.

$\mathbb{V}[\gamma_s]$	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 3.7: Relative policy performance with biased initial information.

The performance of the greedy and index policies reported in Table 3.7 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 3.7 are substantially worse than their corresponding values in Table 3.5 (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 3.6, 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.

3.5.4 Assortment Rotation

From a qualitative perspective it is interesting to measure how much "assortment rotation" is induced by the suggested index policy. We define the assortment rotation as the expected percentage of the assortment that changes in a given period with respect to the previous one. For example, if in period t, the assortment has products A, B, and C, and in period $t+1$ it was A, D, and E, then the assortment rotation with t periods to go is 2/3 (66%). For a given policy, the expectation is with respect to the probability of reaching each possible state under that policy, so we can calculate the assortment rotation as the arithmetic average of the assortment change in each period during our simulated experiments.

Table 3.8 shows the assortment rotation for different values of the initial variance (assuming all products with the same initial prior), and several season lengths T. For instance, for a planning horizon of 20 periods and starting with an initial variance equal to 5, under the index policy, half way through the season only 13% of the assortment is changed on average. Figure 3-6 plots the assortment rotation of the index and greedy policies as a function of the periods to go for the case $\mathbb{V}[\gamma_s] = 100$ and $T = 40$.

$\mathbb{V}[\gamma_s]$	Policy	T	Periods to go			
			T-1	2T/3	T/2	1
5	Grdy	10	45%	19%	14%	8%
		20	45%	10%	7%	3%
		40	45%	4%	3%	1%
	Indx	10	52%	25%	21%	8%
		20	54%	18%	12%	3%
		40	55%	12%	8%	1%
50	Grdy	10	53%	17%	12%	4%
		20	53%	7%	3%	1%
		40	53%	2%	1%	0%
	Indx	10	68%	34%	28%	5%
		20	73%	27%	15%	1%
		40	75%	14%	6%	0%
100	Grdy	10	57%	19%	14%	4%
		20	57%	8%	3%	1%
		40	57%	1%	1%	0%
	Indx	10	73%	38%	31%	4%
		20	75%	31%	16%	1%
		40	78%	15%	6%	0%

Table 3.8: Assortment rotation.

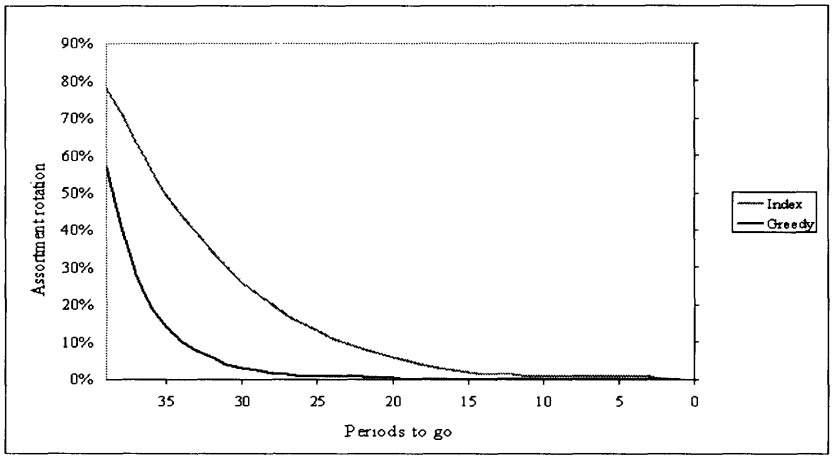

Figure 3-6: Assortment rotation with active and passive learning ($N = 30$).

As expected, the index policy induces a higher assortment rotation than the greedy policy, which is consistent with the fact that the former has active learning. The assortment rotation is also higher when the initial prior variance is higher, since there is more uncertainty to resolve via exploration. As a rule of thumb, the assortment rotation should be high at the beginning, drop to the half of the initial value after the first third of the season, and then quickly converge to a value close to zero. In general, 5 to 10 periods are enough to learn about demand. Note that this rule was derived under the assumptions of our model. In practice, a fast-fashion retailers usually maintains a non-zero assortment rotation since that contributes to attract customers to the store. This phenomenon cannot be captured by our model since we assume exogenous demand. However, our model and the index policy can be easily modified so that the assortment rotation never goes below a certain threshold.

3.5.5 Sensibility Analysis with Respect to S and N

It is intuitively clear that the improvement of the index rule upon the greedy policy should be more relevant when the set of potential products is larger with respect to the amount of shelf space.[8] In Figure 3-7 we show a set of numerical experiments that capture this effect. The simulations where done under the Bayesian approach ignoring lead times and using the same data set as in the previous subsection but with the horizon length T equal to 24. The number of potential products S remained fixed at 720 and the amount of shelf space N took the following values: $720, 120, 60, 40, 30, 20, 10$.

The top curve in Figure 3-7 represents the relative improvement of the index policy compared to the greedy rule as a function of the S/N ratio. The monotonicity and concavity of the curve implies that the marginal improvement is positive and decreasing with respect to S/N (or equivalently, is negative and increasing with respect to N). We obtained a similar curve by setting the shelf space N equal to 10 and then letting S take the values 10, 60, 120, 180, 240, 360, and 720. This suggests that the performance of the index policy depends on the ratio S/N rather than specific values of S and N.

The lower curve in Figure 3-7 corresponds to the suboptimality gap as a function of S/N. We observe that by using the index policy the improvement upon the greedy rule is much more relevant than the possible regret from not implementing the actual

[8] In the remaining part of this chapter we will refer to the index policy as the one that uses the original index formula (3.14 with $\delta_t = 1$).

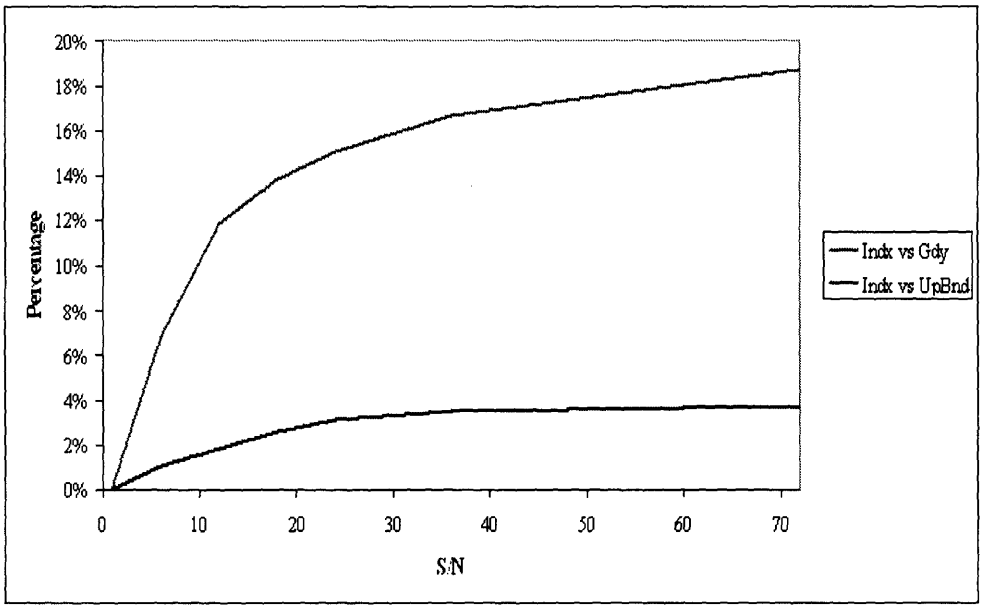

Figure 3-7: Sensibility analysis with respect to S and N.

optimal policy. Recall that the descent method we use does not guarantee finding the optimal open-loop dual policy. Hence, the true regret might be even lower.

3.5.6 Response Surface Bandits

In this subsection we compare our index policy with other heuristic index policies that try to explicitly capture the exploitation vs. exploration trade-off. In particular, we would like to asses the performance of the index policy that uses the following index formula:

$$\eta_{t,s}^{RS} = r_s\big(\mathbb{E}[\gamma_s] + k\sqrt{\mathbb{V}[\gamma_s]}\big) \tag{3.26}$$

where k is a parameter that must be defined by the decision maker. Equation (3.26) is a heuristic index formula widely used in the literature. In particular, it has the same structure as the index formula that is studied in Ginebra and Clayton 1995 for the response surface bandit (that is why we denote the index by 'RS'). In general, the index given by equation (3.26) seems to perform well. However, it has the complication that the parameter k must be calibrated somehow (see Cope 2004). Note as well that index formula (3.26) can be obtained as an approximation from our

index formula (3.14) by ignoring the expectation in denominator of the learning term, and by replacing the finite-horizon factor z_t with the parameter k (we again assume $\delta_t = 1$).

T	k	Rule	Grdy	Indx	$\frac{Indx-Grdy}{Grdy} \cdot 100$
10	z_t	AG	4031.50	4273.81	6.01%
		RS	4031.50	4259.88	5.66%
	$\overline{z}$	AG	4031.50	4245.55	5.31%
		RS	4031.50	4234.67	5.04%
	1	AG	4031.50	4219.47	4.66%
		RS	4031.50	4223.16	4.75%
	2	AG	4031.50	3914.63	-2.90%
		RS	4031.50	4024.40	-0.18%
20	z_t	AG	4420.36	4985.29	12.78%
		RS	4420.36	4969.87	12.43%
	$\overline{z}$	AG	4420.36	4952.63	12.04%
		RS	4420.36	4938.90	11.73%
	1	AG	4420.36	4983.32	12.74%
		RS	4420.36	4966.80	12.36%
	2	AG	4420.36	4819.90	9.04%
		RS	4420.36	4894.15	10.72%
40	z_t	AG	4646.64	5632.36	21.21%
		RS	4646.64	5577.03	20.02%
	$\overline{z}$	AG	4646.64	5578.40	20.05%
		RS	4646.64	5537.60	19.17%
	1	AG	4646.64	5552.47	19.49%
		RS	4646.64	5505.96	18.49%
	2	AG	4646.64	5686.93	22.39%
		RS	4646.64	5686.31	22.37%
2000	z_t	AG	4898.52	7014.20	43.19%
		RS	4898.52	7012.12	43.15%
	$\overline{z}$	AG	4898.52	6977.56	42.44%
		RS	4898.52	6974.42	42.38%
	1	AG	4898.52	6262.23	27.84%
		RS	4898.52	6219.92	26.98%
	2	AG	4898.52	6882.27	40.50%
		RS	4898.52	6873.71	40.32%

Table 3.9: Approximate Gittins index vs response surface index.

In Table 3.9 we compare the performance of the index policy using two different formulas for the indices. The first one is based on the approximation to the Gittins index given by equation (3.15) and has the following structure:

$$\eta_{t,s}^{AG} = r_s\left(\mathbb{E}[\gamma_s] + k\frac{\mathbb{V}[\gamma_s]}{\sqrt{\mathbb{V}[n_s]}}\right) \tag{3.27}$$

The second index formula is the RS rule given by equation (3.26). The simulations were done under the Bayesian approach assuming all initial priors equal to $(1, 1/10)$ (i.e. $\mathbb{V}[\gamma_s] = 100$), $N = 30$, and using the same net margins r_s as before. The table compares both index rules considering different values for the k parameter and three horizon lengths: 10, 20 and 40. Under the column labelled k, z_t means that k was replaced with the time-dependent factors that are the solutions to the equation $z_t = (t-1)\cdot\Psi(z_t)$, and $\overline{z}$ means that k was replaced with the average value $(\sum_{\tau=1}^{T} z_t)/T$.

From Table 3.9 it can be seen that the for the horizon length T equal 10 and 20 the AG rule with the time-dependent z_t factor outperforms all the other policies. For $T = 40$ the best performance is achieved with the AG rule and $k = 2$, but when $T = 2000$ the AG rule with the z_t factor is again the best. This indicates that $k = 2$ might be a good factor (or is well calibrated) only for the case $T = 40$. Hence, the data-independent z_t factors are definitely more convenient. Note that the full information upper bound (cf. Lemma 2) for this instance is equal to 7180.96, meaning that when $T = 2000$ the suboptimality gap of our heuristic policy is less than 3% showing also a very good "asymptotic" performance.

As a final comment on the z_t factors, it is easy to show by induction that the limited lookahead horizon approximation used when deriving the index formula (cf. equation 3.8) actually underestimates the profit-to-go function. Therefore, the z_t factors are rather conservative and might not be the most appropriate weight for the learning term in (3.27) when T is large. A possible is to consider the following profit-to-go approximation instead of equation (3.8):

$$\bar{H}_{t-1,s}^{\lambda}(m_s, \alpha_s) = \frac{\nu(t)}{\delta_t}\max\left\{r_s\frac{m_s}{\alpha_s} - \frac{\lambda}{\delta_t}, 0\right\} \tag{3.28}$$

where $\nu(t)$ is a function that increases faster than linearly in t. Then the z_t factors would be obtained as the unique solution of the equation $z_t = \frac{\nu(t)}{\delta_t}\cdot\Psi(z_t)$ for each t. Note that originally we had $\nu(t) = t - 1$ (for $\delta_t = 1$). If $\nu(t)$ increases faster than linearly, then the corresponding z_t factors will give more weight to the learning term when t is larger.

Chapter 4

Incorporating Substitution Effects

4.1 Heuristic Procedure

As argued in §3.2, 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 (3.3) 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 §3.2) 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 §3.3.4 and §3.3.5, 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

$$\widetilde{n}_s = n_s \cdot \left(\frac{\frac{m_s}{\alpha_s}}{\frac{m_s}{\alpha_s} + \sum_{i \neq s} q_{is} \frac{m_i}{\alpha_i}(1 - u_i)} \right). \tag{4.1}$$

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 (4.1).

2. The index $\eta_{t,s}$ derived in §3.3.4 (and extended to the case of positive lead time in §3.3.5) 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:

$$\max_{\substack{\boldsymbol{u}\in\{0,1\}^S:\\ \sum_{s=1}^{S} u_s \leq N}} \sum_{s=1}^{S} \Big(\eta_{t,s} + r_s \sum_{i\neq s} q_{is} \frac{m_i}{\alpha_i}(1-u_i)\Big) u_s. \tag{4.2}$$

In words, the objective in (4.2) 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 (4.2) results in our original index policy.

4.2 Numerical Experiments

In this section we test the performance of the heuristic procedure suggested above.

Let $L_i = \sum_{s\neq i} q_{is} < 1$ be the probability that a customer wanting product i does not substitute at all when that product is not included in the assortment. Following

Smith and Agrawal (2000), we assume that $L_i = L$ and consider three particular substitution structures:

One-Item Substitution: this would be the case when there is a particular item that serves as the substitute by default for any other item (some sort of "vanilla flavor" that any one might take when they do not find there first choice).

Adjacent Substitution: here the assumption is that the products can be sorted according to some attribute and every customer would consider as a substitute the product that is slightly better and the one that is slightly worse than their first choice.

Random Substitution: in the case the retailer assumes that when a customer does not find her first choice she will randomly substitute among those that are available at the store.

The corresponding substitution matrices for these three cases are shown in Figure 4-1. Note that the only parameter that must be estimated is L, the probability that no substitution occurs.

One-Item Substitution Matrix

$$\begin{pmatrix} 0 & 0 & 1-L & 0 & 0 \\ 0 & 0 & 1-L & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-L & 0 & 0 \\ 0 & 0 & 1-L & 0 & 0 \end{pmatrix}$$

Adjacent Substitution Matrix

$$\begin{pmatrix} 0 & 1-L & 0 & 0 & 0 \\ \frac{1-L}{2} & 0 & \frac{1-L}{2} & 0 & 0 \\ 0 & \frac{1-L}{2} & 0 & \frac{1-L}{2} & 0 \\ 0 & 0 & \frac{1-L}{2} & 0 & \frac{1-L}{2} \\ 0 & 0 & 0 & 1-L & 0 \end{pmatrix}$$

Random Substitution Matrix

$$\begin{pmatrix} 0 & \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} \\ \frac{1-L}{n-1} & 0 & \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} \\ \frac{1-L}{n-1} & \frac{1-L}{n-1} & 0 & \frac{1-L}{n-1} & \frac{1-L}{n-1} \\ \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} & 0 & \frac{1-L}{n-1} \\ \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} & \frac{1-L}{n-1} & 0 \end{pmatrix}$$

Figure 4-1: Simple substitution structures.

Recall that the heuristic procedure involves two modifications: (i) a new state update rule (we refer to it as 'NU') that deflates the observed sales taking into account possible substitutions (cf. equation (4.1)), and (ii) a quadratic integer program (we call it 'QP') that provides the assortment selection in each period (cf. equation (4.2)). In order to capture the individual impact of each modification, for each substitution structure we simulated the following policies for different values of L:

- IndexNUQP: this is exactly the modified index policy described in §4.1.

- IndexQP: this policy decides the assortment by solving QP but then updates the state information based on the total observed sales instead of using NU.
- IndexNU: opposed to the previous one, this index policy uses the new update rule NU but the assortment decision ignores substitution, i.e. the N products with the highest indices are included in assortment instead of solving QP.
- Index: this corresponds to the original index policy that ignores substitution and does not deflate observed sales.
- GreedyNUQP: same as IndexNUQP but the linear term in the objective function of the quadratic integer program is given by the immediate expected return $r_s \frac{m_s}{\alpha_s}$ instead of the active learning index $\eta_{t,s}$.
- Greedy: the original greedy policy that ignores substitution and does not deflate observed sales.

We also provide the corresponding full information upper bound, which also involves solving a quadratic program. As a minor observation, note that when the state information is updated using the deflated observed sales, the m_s component is no longer an integer value.

We tested the performance of the policies for each one of the substitution structure described above. We followed the frequentist approach since we do not have a DP formulation for the problem with substitution (therefore, the Bayesian approach would not be appropriate). As before, we used the data set with all initial priors equal to $(1, 1/10)$ (but now the "real" demand rates γ_s come into play). For the one-item and adjacent substitution the products were sorted by the net margins r_s (in descending order). This would be equivalent to the case when the production costs are the same for all products (in one category) and customers substitute according to price. The horizon length was set equal to 24 and we assumed a zero lead time. The quadratic integer program QP was solved using the callable library of CPLEX 9.3. We noticed that the running times increased dramatically in the number of products (i.e. the size of the quadratic problem). As a consequence, we restricted our study to instances with $S = 144$ and $N = 6$ (note that $S/N = T$, which would be a reasonable ratio for the case of Zara). In each simulation we performed 100 replications, 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 next three figures show the numerical results. In each figure, the order of the policies in the legend corresponds to the order of the curves in the graph.

For the one-item substitution case product number 80 was assumed to be the substitute by default, representing the situation when customers might choose a rather inexpensive alternative when their first choice is not available. In this case the performance of the IndexNUQP and IndexQP policies is the same and is represented by the second curve with a slope (cf. Figure 4-2). The IndexNU and Index policies also have an identical performance given by the first dark horizonal line. This shows that, when there is a substitute by default, the key step of the heuristic procedure is to solve QP since then the "vanilla flavor" is most likely to be included in the assortment. It is intuitively clear that this action should be optimal, and the proximity of the QP policies to the full information upper bound confirms this observation. Note as well that the relative performance of the policies is invariant with respect to the no substitution probability L.

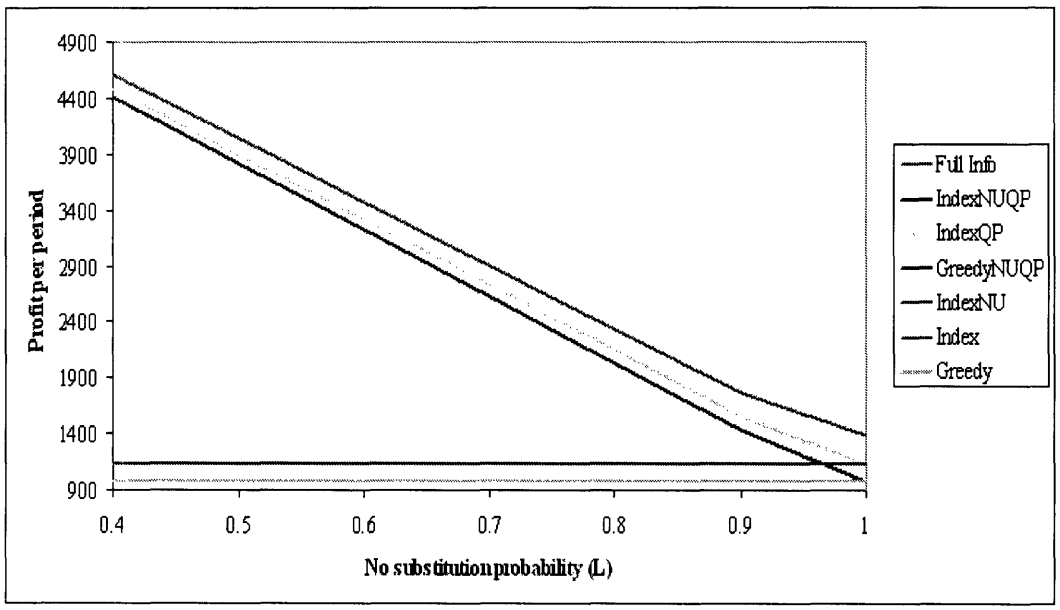

Figure 4-2: One-item substitution.

Figure 4-3 shows the results under adjacent substitution. In this case both modifications to the original index policy are relevant. The most important observation is that IndexNUQP remarkably outperforms GreedyNUQP for all values of L, meaning that active learning can provide a substantial improvement upon passive learning.

However, the large gap with respect to the full information curve shows that there would still be plenty of room for more learning if T were larger.

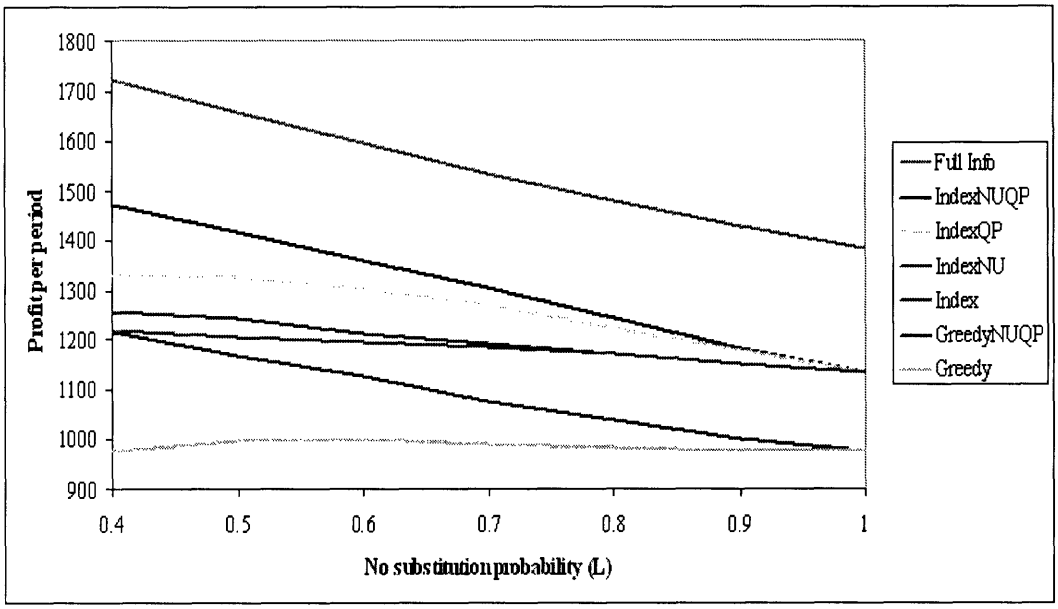

Figure 4-3: Adjacent substitution.

The case with random substitution is shown in Figure 4-4. Opposed to the one-item substitution case, the IndexNUQP-IndexNU and IndexQP-Index policy pairs perform almost identically. This means that now the key step of the heuristic procedure is to update the state using the deflated observed sales. Note that the gap with the full information upper bound is large (as in the adjacent case), but the improvement of IndexNUQP upon GreedyNUQP is still quite significant (favoring active learning).

The previous numerical results confirm what intuition would suggest in sense that the one-item and random substitution structures are two extremes of all feasible configurations. In the first case the second modification of the heuristic procedure is crucial, while in the second case the first modification is the main driver. The adjacent substitution fits right in between the other two and is finally the most interesting and difficult case. In fact, Table 4.1 shows the corresponding running times (the computational requirements for the other two substitution structures is at least an order of magnitude lower). Note that the GreedyNUQP policy is harder to compute possibly because the quadratic term in QP has a larger relative weight (compared to

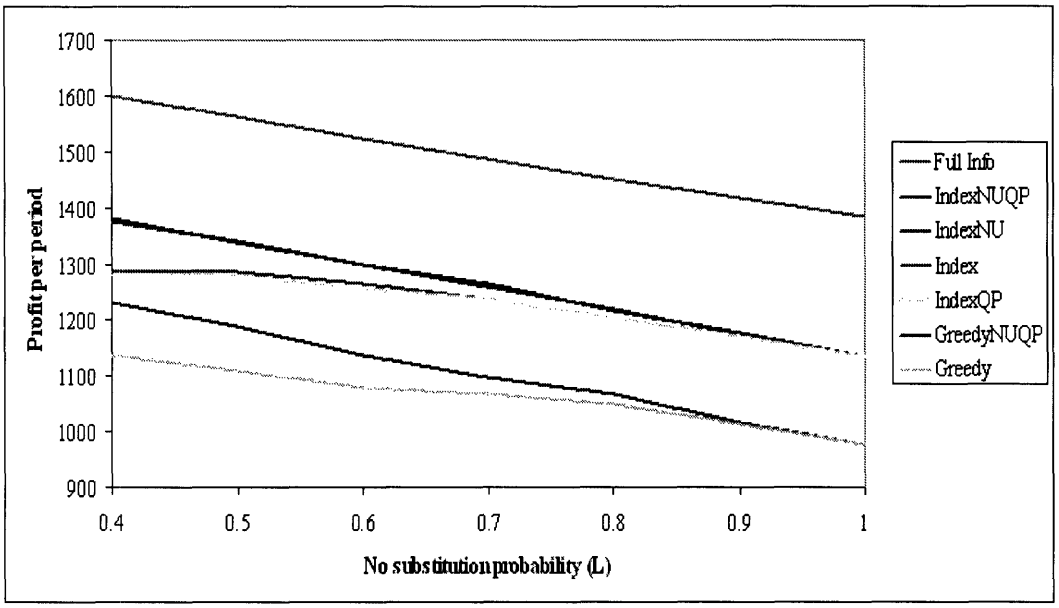

Figure 4-4: Random substitution.

IndexNUQP).

L	IndexNUNLP	IndexNLP	IndexNU	GreedyNUNLP
0.4	1050	428	12	2806
0.5	654	293	12	1545
0.6	350	216	12	852
0.7	199	162	12	395
0.8	116	104	12	157
0.9	74	70	12	60
1.0	0	0	0	0

Table 4.1: Simulation running times for adjacent substitution (rounded to seconds).

Chapter 5

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; and (ii) the only information available about lost sales is whether or not they occur. A third model that assumes that the retailer can register the epochs when stock occur is shown in Section §6.2 as an extension.

In the following models the assortment is controlled through the stock level of each product. A particular product is excluded from the assortment by setting its stock level equal to zero. Inventory decision are now endogenous and a if a customer arrives wanting to buy product s and there are none available, then she leaves and the sale is lost. We assume that, besides the loss in revenue, lost sales are not penalized since it is not clear what would be an appropriate value for the penalty parameter, and moreover, as mentioned in the model discussion of chapter 3, penalizing lost sales is not a usual practice among fast-fashion retailers.

In terms of cost, the procurement expenditure is only taken into account implicitly via the net margins r_s. As before, we are ignoring any holding cost since our goal is to focus on the assortment decisions. Similarly, we assume that the retailer can dispose (at no extra cost) any amount of product that was not sold at the end of a given period. The justification is that a retailer will never waste shelf space with products that are not selling. If the assortment selected in the previous period was

inappropriate, then the most likely is that the retailer will remove those products that are overstocked. In other words, we can assume that in each period the stock-level decisions are made regardless of product leftovers. We are aware that this assumption would be valid in the case of Zara. In fact, Zara stores do not have a "back room" where excess inventory can be stored, and products that sell slowly are "ruthlessly weeded out by store managers with incentives to do so" (see Ghemawat, P. and Nueno 2003, and McAfee *et al.* 2004).

Under the previous assumption, our setting is closely related to inventory models for unstorable/perishable products or with a *full returns policy* (i.e. when the retailer can return any amount of unsold items at a cost equal to the procurement cost).

Throughout this chapter the decision variable u_s represents the inventory on hand after ordering. We assume a zero lead time so that u_s units of product s are available at the beginning of the period. We also keep the independence assumption from Chapter 3 for the demand of different products. For simplicity we assume that at the store level all products have the same space requirement. The total amount of inventory that the store can handle is restricted to N. Obviously, if demand is discrete, then the control u_s should also be restricted to discrete values. Finally, the models can be easily extended to the case in which there are upper limits on the amount of shelf space that can be assigned to a particular product.

5.1 Total Demand is Observable

The first model assumes that, for those products included in the assortment, lost sales are observable. That is to say, any customer that leaves due to a stockout is registered at the point of sale. This is the case, for example, when sales are made online, via catalogs, or whenever the seller can record customer requests for missing items. As a consequence, demand is not censored and that information can be used in the learning process. This assumption is a usual one in Bayesian learning models for tractability reasons (see for instance Chen and Plambeck 2004).

5.1.1 Model Definition

As before, we assume period of equal length, and the demand for product s is Poisson with unknown mean γ_s, though the extension to the exponential family described in the previous chapter also applies. We have that in this case, the Bellman equation takes the following form:

$$\begin{aligned} J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}) &= \max_{\substack{\boldsymbol{u} \in \mathcal{U}: \\ \sum_{s=1}^S u_s \leq N}} \mathbb{E}_{\boldsymbol{n}}\left[\sum_{s=1}^S (n_s \wedge u_s) r_s + J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \mathbb{I}(\boldsymbol{u}), \boldsymbol{\alpha} + \mathbb{I}(\boldsymbol{u})) \right] \\ &= \max_{\substack{\boldsymbol{u} \in \mathcal{U}: \\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s \mathbb{E}_{n_s}\left[n_s \wedge u_s\right] + \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \mathbb{I}(\boldsymbol{u}), \boldsymbol{\alpha} + \mathbb{I}(\boldsymbol{u})\right] \end{aligned} \tag{5.1}$$

where $\wedge$ represents the (componentwise) minimum operator, and $\mathbb{I}(\boldsymbol{u})$ is a vector such that the i-th component is equal to 1 when $u_s > 0$, meaning that the demand information update is done only for those products that were offered in that period (i.e. were included in the assortment). Then learning is only affected by the binary decision of including or not the product in the assortment regardless of the actual amount of product that is stocked.

In the model (5.1) we have not specified the control set $\mathcal{U}$. It could seem natural to have $\mathcal{U}$ equal to the positive octant. However, in that case, even though the feasible set in (5.1) is compact, the maximization might no be well defined since the objective function is discontinuous. To avoid this technical complication, we introduce the constraints $u_s \geq \mathbb{I}(u_s)\varepsilon \ \forall s$, with $\varepsilon > 0$. Note that these constraints have a direct interpretation: if the retailer decides to stock product s in the current period, he must stock at least ε units, this is the minimum amount that would allow the retailer to learn about the respective product demand.

We can now relax the coupling constraint $\sum_{s=1}^S u_s \leq N$ in order to decompose the problem by product. The analogue to equation (3.6) is given by:

$$H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) = \max \begin{cases} r_s \mathbb{E}_{n_s}\left[n_s \wedge u_s^o\right] - \lambda_t u_s^o + \mathbb{E}_{n_s}\left[H_{t-1,s}^{\boldsymbol{\lambda}}(m_s + n_s, \alpha_s + 1)\right] \\ H_{t-1,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s) \end{cases} \tag{5.2}$$

where $u_s^o(\lambda_t) = \max\{\varepsilon, u_s^v(\lambda_t)\}$, and $u_s^v(\lambda_t)$ is the largest solution to the newsvendor

inequality $\Pr(n_s \geq u_s^v) \geq \lambda_t / r_s$. Clearly only $\lambda_t \leq r_s$ makes sense. Note that once the coupling constraint is relaxed, it is reasonable to add the (originally redundant) constraints $u_s \leq N \; \forall s$, otherwise there is no limit to the amount ordered in the single product subproblem and that affects the tightness of the dual upper bound. With the additional constraint we have that $u_s^o(\lambda_t) = \min\big\{ \max\{\varepsilon, u_s^v(\lambda_t)\}, N \big\}$.

It can be verified that when stationary open-loop dual policies are considered (i.e. $\lambda_t = \lambda \; \forall t$), then a *stopping time policy* as in Proposition 4 is optimal. Moreover, as discussed in §3.3.4, there is a (unique) value of λ at which the retailer is indifferent between the 'include' and 'not include' actions in the single-product subproblem (5.2). Let $\eta_{t,s}$ correspond to that breakpoint. It is easy to verify that $\eta_{t,s} < r_s \; \forall s$, and that $\lim_{\varepsilon \to 0} \eta_{t,s} = r_s$.

It is important to point out that even though we have followed the same steps as in Chapter 3 in order to decompose the problem by product, the open-loop dual policies have different interpretations. In Chapter 3 the Lagrange multipliers represent the opportunity cost of carrying a certain product in the store with no restriction on the amount that can be sold. Hence, the decision is what product lines to consider and the shelf-space constraint says that at most N can be handled. The indices $\eta_{t,s}$ in that case can be seen as a *proxy* for the profitability (per period) of the product line as a whole. As such, there is no limit on the value the indices can take. The situation in the present chapter is completely different. The Lagrange multipliers represent the opportunity cost for one unit of stock in the store. Each unit has to pay that cost (instead of the product line), so exploring becomes more expensive. The fact that the shelf-space constraint restricts total sales to no more than N units limits the opportunities to learn but also undermines its benefits and makes the "explore now to exploit later" rule less effective. In a figurative way, if the retailer explores and finds a "gold mine", he cannot take full advantage of it since he can sell (exploit) at most N units in each period.

In order to find a feasible action for the current period in the original DP model (5.1), we suggest solving a knapsack problem similar in spirit to the one in Chapter 4, which in turn was a generalization of the index policy of Chapter 3. We assume that demand is sold in discrete amounts. If demand is originally continuous, then it must be discretized (possibly at a finer scale than the natural numbers). However, we will describe the procedure having an integer demand process in mind (for example, the Poisson model of Chapter 3).

Consider the following (generic) knapsack problem: Let $a_{s,\varepsilon}$ be the benefit from including ε units of product s in the assortment. We assume that ε is a positive integer value. Let $a_{i,\varepsilon+i}$ with $i = 1, \ldots, N-\varepsilon$ be the marginal benefit of adding the $(\varepsilon+i)$-th unit of product s to the assortment. Let $x_{s,\varepsilon}$ and $x_{s,\varepsilon+i}$ be binary variables that are equal to one if the corresponding units of product s are included in the assortment. With this data structure, the optimal assortment can be found by solving a knapsack problem with precedence constraints:

$$\max \sum_{s=1}^{S} \sum_{i=0}^{N-\varepsilon} a_{s,\varepsilon+i} x_{s,\varepsilon+i} \tag{5.3}$$

subject to

$$\sum_{s=1}^{S} \varepsilon x_{s,\varepsilon} + \sum_{s=1}^{S} \sum_{i=1}^{N-\varepsilon} x_{s,\varepsilon+i} \leq N \tag{5.4}$$

$$x_{s,\varepsilon+i} \geq x_{s,\varepsilon+i+1} \quad \forall s = 1, \ldots, S,\ i = 0, \ldots, N-\varepsilon-1 \tag{5.5}$$

$$x_{s,\varepsilon+i} \in \{0,1\} \quad \forall s = 1, \ldots, S,\ i = 0, \ldots, N-\varepsilon \tag{5.6}$$

The objective is to maximize profits subject to the available shelf-space N. Constraint (5.5) ensures that the $(\varepsilon+i)$-th unit of product s is included in the assortment before the $(\varepsilon+i+1)$-th. The combinatorial problem (5.3)-(5.6) is NP-Hard but several good approximation algorithms are available (see for instance Samphaiboon and Yamada 2000).

We now relate the knapsack problem described above to our dynamic assortment problem. At time t, it is clear that the marginal benefit $a_{s,\varepsilon+i}$ with $i = 1, \ldots, N-\varepsilon$ obtained from including the $(\varepsilon+i)$-th unit of product s in the assortment is equal to $r_s \Pr(n_s \geq \varepsilon+i)$, where the probability is calculated according to the distribution dictated by the current state (m_s, α_s). What is not obvious is the appropriate value for $a_{s,\varepsilon}$, since for the first ε units of product s the retailer not only receives the direct benefit from selling those units but also the benefits from being able to observe total demand. As in previous chapters, we distinguish two possibilities that lead to two different policies. In the first heuristic, $a_{s,\varepsilon}$ is equal to $r_s \sum_{i=1}^{\varepsilon} \Pr(n_s \geq i)$, and we call it the greedy or passive learning policy in this context since the assortment decision is made ignoring future benefits from learning. The second heuristic, which we call the active learning policy, is to set $a_{s,\varepsilon}$ equal to the breakpoint $\eta_{t,s}$. It is easy to show

that $\eta_{t,s} \geq r_s \sum_{i=1}^{\varepsilon} \Pr(n_s \geq i)$, and the difference comes from the fact that $\eta_{t,s}$ is calculated taking into account the profit-to-go in the single-product subproblem.

Once the parameter values are set, the knapsack problem (5.3)-(5.6) can be (approximately) solved, and the solution determines the assortment to be implemented in the current period. Note that if $t = 1$, then the assortment problem (5.1) is nothing but a multiproduct newsvendor problem with a shared resource (see §6-4 in Hadley and Whitin 1963). In that case, the active and passive policies provide the same solution, which is optimal.

5.1.2 Numerical Experiments

We performed some numerical experiments in order to test the active and passive learning policies. We used the Poisson demand model as in §3.5, and we assumed that $\varepsilon = 1$ since in that case the precedence constraints (5.5) are redundant and we obtain an easy knapsack problem that can be solved with a simple greedy algorithm (see Martello and Toth 1990). We considered the data set given in the Sport Obermeyer case (see Hammond and Raman 1994), so the number of products is ten ($N = 10$) and the net margin r_s is equal to the product price times 0.24. The number of replications in each simulation was equal to 2,500 in order to guarantee an absolute error less than $\pm 0.5\%$ with a 95% confidence.

$\mathbb{V}[\gamma_s]$	N	Passive	Active	UpBnd	$\frac{Active-Passive}{Passive} \cdot 100$	$\frac{UpBnd-Active}{Active} \cdot 100$
10	10	383.43	383.44	384.75	0.00%	0.34%
	30	986.96	987.84	988.23	0.09%	0.04%
	60	1674.73	1677.11	1678.43	0.14%	0.08%
	120	2525.99	2525.99	2527.76	0.00%	0.07%
50	10	369.44	369.52	376.00	0.02%	1.76%
	30	955.02	959.74	976.42	0.49%	1.74%
	60	1630.50	1629.87	1656.46	-0.04%	1.63%
	120	2429.54	2439.65	2491.08	0.42%	2.11%
100	10	358.94	359.02	367.89	0.02%	2.47%
	30	940.53	935.54	971.37	-0.53%	3.83%
	60	1600.81	1601.98	1656.10	0.07%	3.38%
	120	2348.84	2394.92	2487.69	1.96%	3.87%

Table 5.1: Active vs. passive learning with lost sales.

Table 5.1 shows the simulated profit per period for both policies and the value of the dual upper bound for a planning horizon of ten periods ($T = 10$). As in

§3.5 we assume the same initial priors for all the products. Several observations can be made. First, note that the performance of both policies and the upper bound decreases with a larger variance. This means that the retailer is better off with less uncertainty about the demand rates. This is a usual fact in newsvendor problems. However, it notoriously contrasts with the result of the model with no lost sales (see Table 3.4), in which a larger variance means more possible benefits from exploration and therefore the expected profit larger. Second, the active learning policy performs slightly better than the policy with passive learning, though the difference is far from being substantial, and actually in many cases it can be regarded as being equal to zero.[1] And third, the small improvement of active upon passive learning might suggest that the former is not a good policy. Nevertheless, the last column shows that there is actually very little room for a better rule, and the active and passive learning policies are both near optimal. This confirms the previous discussion in the sense that exploring is not a major component of the optimal policy in the lost sales model (5.1) with observed demand.

5.2 Censored Information

We now address the case when the retailer can only observe total sales (instead of demand) at the end of a period.

In formulating the problem with censored information, we use a significantly different demand model. 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.

In what follows, we assume that demand per period n_s for product s has a *newsvendor distribution* with an homogenous unknown rate (an implicit consequence is that all periods have the same length). Note that demand is now continuous so the procurement decisions are also continuous (but still constrained by N). The family of

[1] A similar result between active and passive learning is obtained in Bertsimas and Mersereau 2004 under the frequentist approach for a direct marketing model.

newsvendor distributions is a subclass of the exponential family and was introduced by Braden and Freimer (1991). The functional form of the density is the following:

$$dF(n_s|\gamma_s) = \gamma_s c'(n_s) \exp\big(-c_s(n_s)\gamma_s\big) \tag{5.7}$$

where $c_s(n_s)$ is positive, differentiable, and increasing. For simplicity, we assume that $c_s(0) = 0$.

In their paper, Braden and Freimer show that the newsvendor family can be characterized as the class of distributions that admit a conjugate prior under censored observations. The conjugate prior turns out to be a Gamma distribution. This means that, in the case of censored information, if we want to take advantage of state space reduction and sufficient statistics using a conjugate pair, we must restrict ourselves to the class of newsvendor distributions. The downside is that this precludes us from using the normal, Poisson and negative binomial distributions which seem to fit well the sales pattern of a fashion retailer (see Agrawal and Smith 1996, and Eppen and Iyer 1997).

An alternative approach would be to forgo the desire of having a sufficient statistic and work directly with the posterior distributions as the state variable. We could formulate a model similar in nature to the censored newsvendor model in Ding *et al.* (2002), but then we would have a major loss in tractability and it seems unlikely that the results would be easy to implement in practice.

Given a Gamma prior with shape parameter α_s and scale parameter m_s, we have that the predictive (unconditional) demand density of product s is given by:

$$\Pr(n_s) = \frac{\alpha_s m_s^{\alpha_s} d'(n_s)}{[m_s + d_s(n_s)]^{\alpha_s+1}}$$

The information updating rule in this case works as follows:

$$(m_s, \alpha_s) \longrightarrow \begin{cases} (m_s + c_s(n_s), \alpha_s + 1) & \text{If product } s \text{ is in the assortment, } n_s \text{ sales are} \\ & \text{observed in period } t, \text{ and no stockout occurs} \\ (m_s + c_s(n_s), \alpha_s) & \text{Idem but with stockout} \\ (m_s, \alpha_s) & \text{If product } s \text{ is not in the assortment} \end{cases}.$$

The state components are the scale and shape parameters respectively of the Gamma posterior belief for γ_s. Note that the roles of the scale and shape parameters

are switched with respect to the previous models, i.e. now the scale parameter is updated with the sales information and the shape parameter captures the notion of time. This means that the coefficient of variation of the Gamma prior (or posterior) is equal to $1/\sqrt{\alpha_s}$. Then the retailer gains precision in his estimation of the unknown parameter γ_s only in the first case of the updating rule, i.e. when the product is included in the assortment and it does not stock out. This clearly contrasts with the learning model of Chapter 3 in which precision was gained with a larger amount of observed sales.

We now write the corresponding Bellman equation for this case:

$$J_t^*(\boldsymbol{m},\boldsymbol{\alpha}) = \max_{\substack{\boldsymbol{u}\geq\boldsymbol{0}:\\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s \mathbb{E}_{n_s}\left[n_s \wedge u_s\right] + \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m}+c(\boldsymbol{n}\wedge\boldsymbol{u}),\boldsymbol{\alpha}+\mathbb{I}(\boldsymbol{n},\boldsymbol{u}))\right] \quad (5.8)$$

where $c(\boldsymbol{n}) = (c_1(n_1),\ldots,c_S(n_S))$ and the s-th component of $\mathbb{I}(\boldsymbol{u},\boldsymbol{n})$ is equal to one unless $u_s = 0$ or $u_s < n_s$. In this case the minimum order quantity ε is not strictly required.

Depending on the explicit functional form of $c_s(n_s)$, the expectations $\mathbb{E}_{n_s}\left[n_s \wedge u_s\right]$ might have a closed form formula. For example, in the exponential case $c_s(n_s) = n_s$, we get $\mathbb{E}_{n_s}\left[n_s \wedge u_s\right] = \frac{m_s}{\alpha_s - 1}\left(1 - \left(\frac{m_s}{m_s+u_s}\right)^{\alpha_s - 1}\right)$. Note that $\alpha_s > 1$ is required for the expectation to be finite.

Let $G_t(\boldsymbol{u},\boldsymbol{n}|\boldsymbol{m},\boldsymbol{\alpha}) = \sum_{s=1}^S (u_s \wedge n_s) r_s + J_{t+1}^*(\boldsymbol{m} + c(\boldsymbol{u}\wedge\boldsymbol{n}),\boldsymbol{\alpha}+\mathbb{I}(\boldsymbol{u},\boldsymbol{n}))$, and as in the previous cases, consider the single-product subproblem and an open-loop dual policy $\boldsymbol{\lambda}$. The next proposition adapts the results of Lariviere and Porteus (1999) to the dynamic assortment problem considered in this section:

Proposition 6 *(a) $G_t(\boldsymbol{u},\boldsymbol{n}|\boldsymbol{m},\boldsymbol{\alpha})$ is (componentwise) increasing in n_s and m_s, and decreasing in α_s.*

(b) $J_t^*(\boldsymbol{m},\boldsymbol{\alpha})$ is increasing in m_s and decreasing in α_s.

(c) If the underlying demand distributions is Weibull, that is $c_s(n_s) = n_s^k$ for some known k, then the state reduction of Azoury (1985) applies to the single-product subproblem.

(d) For the exponential case ($c_s(n_s) = n_s$) with $\lambda_t \leq r_s\ \forall t$, $J_{t,s}^{\boldsymbol{\lambda}}(m_s,\alpha_s) = \frac{m_s}{\alpha_s - 1}\rho_{t,s}^{\boldsymbol{\lambda}}(\alpha_s)$, where $\rho_{t,s}(\alpha_s)$ is obtained recursively from the equation:

$$\rho^{\boldsymbol{\lambda}}_{t,s}(\alpha_s) = r_s - \lambda_t - \alpha_s \lambda_t y^{\boldsymbol{\lambda}}_{t,s}(\alpha_s) + \rho^{\boldsymbol{\lambda}}_{t-1,s}(\alpha_s + 1)$$

$$y^{\boldsymbol{\lambda}}_{t,s}(\alpha_s) = \left(\frac{r_s + \rho^{\boldsymbol{\lambda}}_{t-1,s}(\alpha_s + 1) - \rho^{\boldsymbol{\lambda}}_{t-1,s}(\alpha_s)}{\lambda_t} \right)^{1/\alpha_s} - 1$$

Part (a) of Proposition 6 is used to prove part (b), which in turn is equivalent to Lemma 1 in the no lost sales model. Parts (c) and (d) can be used to facilitate the calculation of the performance upper bound and the parameters of the knapsack problem (to be described) for particular cases of the function $c_s(n_s)$.

To solve the lost sales model with censored information (5.8) we suggest a procedure almost identical to the one in the previous section. First, demand must be discretized. For simplicity we consider the integer numbers, but a finer partition would work as well. At a given period, in a given state, the assortment decision is made by solving the knapsack problem (5.3)-(5.6) with $a_{s,i+1} = b_{s,i+1} - b_{s,i}$, and $b_{s,i}$ is equal to the root (breakpoint) of the following equation expressed in terms of the unknown λ:

$$r_s \mathbb{E}_{n_s}\left[n_s \wedge i\right] - \lambda i + \mathbb{E}_{n_s}\left[H^{\lambda}_{t-1,s}(m_s + c_s(n_s \wedge i), \alpha_s + \mathbb{I}(n_s, i))\right] = 0$$

where $H^{\lambda}_{t-1,s}(\cdot)$ is the single product subproblem obtained from (5.8) when the shelf-space constraint is relaxed using a stationary open-loop dual policy ($\lambda_t = \lambda \; \forall t$).

Chapter 6

Conclusions and Extensions

In this final chapter we first provide concluding remarks on the models introduced and analyzed in the thesis, and then we discuss some possible extensions and future research.

6.1 Concluding Remarks

We have developed in this thesis several discrete-time DP models for the dynamic assortment problem faced by a fast-fashion retailer refining his estimate of consumer demand for his products over time. In Chapter 3 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 multi-armed bandit with finite horizon and multiple plays per stage. Using the Lagrangian decomposition of weakly coupled DPs that was described in Chapter 2, we have derived a closed form index policy characterized by equation (3.14) 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. The index formula (3.14) and the (numerical) performance guarantee are applicable in general to any finite-horizon multiarmed bandit with Bayesian learning.

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. As a rule of thumb, the assortment rotation should be high at the beginning of the planning horizon, drop to the half of the initial value after the first third of the season, and then quickly converge to a value close to zero. In general, 5 to 10 periods are enough to learn about demand.

In Chapter 3 we have also considered the case when there is a lag in the implementation of the assortment decisions. This case can be seen as a multiarmed bandit problem with a delayed response, which has not yet received much attention in the literature. We adapt our index policy to fit the new setting via two modifications that provide insights about the relevant factors that must be taken into account under the presence of a positive lead time: (1) there is less time to learn and, (2) the prediction of the future variability in demand must consider those assortment that are "on their way to the store" since for those products exploration is already committed and some learning will necessarily occur.

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 Chapter 4 we have thus proposed a heuristic for capturing substitution effects between products, and we show how it performs under different substitution structures. In particular, if customers substitute randomly or there is a clear substitute that any one would take, then the dynamic assortment problem is relatively easy to solve. However, if product substitution follows a more complex pattern, for instance when customers substitute to products that are "adjacent" (in terms of the attributes) to their initial choice, then selecting product assortment becomes a combinatorial problem that does not have a simple solution.

In Chapter 5 we present models that consider the stocking decision and deal with lost sales. The first model assumes that, despite lost sales, total demand is observable, and the second model has censored information, that is, the retailer can only observe sales but is aware when a stock out occurs. The inventory decision makes

the models more complex, especially under censored information, but furthermore, since the shelf-space constraint is expressed in terms of stock units (instead of product lines), learning becomes more expensive and is also limited by the fact that the total sales are bounded. Therefore, the characteristics of these models are quite different from the multiarmed bandit ones in Chapter 3, and the improvement of active upon passive learning is less substantial.

Finally, although the models presented here focuses essentially on operational issues, we point out that they 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 models, 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 3-5) may thus inform the assessment of such trade-off. We thus conclude that our models may also be useful to some practitioners when designing supply-chains.

6.2 Model Extensions and Future Work

We conclude the thesis by briefly commenting on other potential applications of the models and possible extensions.

6.2.1 The Multiarmed Bandit Beyond Retailing

In Chapter 3 we showed that the dynamic assortment problem is basically a version of the finite horizon multiarmed bandit problem with several plays per stage. A similar analogy is valid in other setting beyond the retailing industry, and the same analysis would go through. The following example of "fashion" bedding products was

provided by Professor Harvey Wagner from UNC Chapel Hill:

The example of "fashion" bedding (sheets, pillowcases, draperies, etc.) is different in several respects from fashion retailing. The span of time for sales is nearly two years. The decision is not whether to stock a SKU in a specific location (store), but whether or not to "support" a family of items by replenishing (that is, manufacturing) warehouse stock. This decision involves allocating (scarce) production capacity to replenishing inventories. In the manufacturing setting, the slow moving inventory sits on the shelf until it is eventually remaindered; there is no necessity to dispose of it right away. However, the manufacturer keeps focused on N families of products, and as more sales information is obtained, may switch the components of N.

6.2.2 Lost Sales Model when Stock-out Epochs are Observable

Here we describe a dynamic assortment model with the assumption of Poisson demand and that the retailer cannot observe lost sales but can register the point in time when a stock-out occurs. The demand information must be updated accordingly. In fact, suppose that product s runs out of stock in period t, and it happens δ units of time after the beginning of the period, with δ less than δ_t, the duration of period t. If u_s is the amount of inventory that was available after ordering, the posterior distribution of γ_s is obtained by letting δ play the role of δ_t, and the expectations must be taken not only over demand but also with respect to the points in time in which the retailer might run out of stock. The corresponding Bellman equation can be expressed as follows:

$$J_t^*(\boldsymbol{m},\boldsymbol{\alpha}) = \max_{\substack{\boldsymbol{u}\geq\boldsymbol{0}:\\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^S r_s E_{n_s}\left[n_s \wedge u_s\right] + \mathbb{E}_{\boldsymbol{n}}\Big[\mathbb{E}_{\boldsymbol{\xi}(\boldsymbol{n},\boldsymbol{u})}\big[J_{t-1}^*(\boldsymbol{m}+\boldsymbol{u}\wedge\boldsymbol{n}, \boldsymbol{\alpha}+\boldsymbol{\xi}\cdot\mathbb{I}(\boldsymbol{u}))\big]\Big] \tag{6.1}$$

where $\boldsymbol{\xi}(\boldsymbol{n},\boldsymbol{u})$ is a random vector with density equal to $\prod_{s=1}^S y_t(n_s,u_s,\xi_s)$, and

$$y_t(n,u,\xi) = \begin{cases} \Delta(\xi-\delta_t) & \text{if } u \geq n \\ \left(\frac{u}{\delta_t^n}\binom{n}{u}\xi^{u-1}(\delta_t-\xi)^{n-u}\right)\mathbb{I}_{[0\leqslant\xi\leqslant\delta_t]} & \text{if } u < n. \end{cases} \tag{6.2}$$

The upper term in (6.2) is the usual Dirac delta function and the lower term is the marginal density of the u-th arrival conditional on $n > u$ arrivals happening in

an interval of length δ_t (obviously for a Poisson process).

When the shelf constraint is relaxed we obtain the following single-product subproblems:

$$H^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) = \max_{0 \leqslant u_s \leqslant N} r_s \mathbb{E}_{n_s}\left[n_s \wedge u_s\right] - \lambda_t u_s + \mathbb{E}_{n_s}\left[\mathbb{E}_{\xi(n_s,u_s)}\left[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s + n_s \wedge u_s, \alpha_s + \xi \cdot \mathbb{I}_s(u_s))\right]\right] \tag{6.3}$$

and $\xi(n_s, u_s)$ has density $y_t(n_s, u_s, \xi)$.

Note that the policy described in Chapter 5 based on solving the knapsack problem (5.3)-(5.6) is also applicable to the this case. However, solving the single-product subproblem (6.3) via backwards induction is a non-trivial computational task because now the second component of the state space takes values in a continuous range. Then calculating the breakpoints might not be easy in practice.

6.2.3 Model with Variable Demand Rates

In our models, 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 basically amounts to extending our model in the same way that Aviv and Pazgal (2004) extend their initial dynamic pricing problem (Aviv and Pazgal 2002).

We assume that for each product s, the demand environment follows an independent Markov chain $M_s(t)$, called the core process, on a state space Ω_s. The transition probabilities $A^{u_s}_{s,j,k} \quad j, k \in \Omega_s$ can depend on whether the product is included in the assortment or not. In the latter case we obtain a "restless" bandit and it should be verified that the *indexability* property holds (see Whittle 1988).

The system state is given by all the available information. In particular, for product s, the retailer knows the probability distribution $p_{t,s,k} = \Pr(M_s(t) = k) \quad k \in \Omega_s$, and each in period the information is updated according to Bayes rule:

$$p_{t-1,s,k} = \varphi_{t,s}(\boldsymbol{p}_{t,s}, u_s, n_s) = \frac{\sum_{j\in\Omega_s} A^{u_s}_{s,j,k} \cdot \Pr(n_s|M_s(t)=j,u_s)\cdot p_{t,s,j}}{\sum_{j\in\Omega_s} \Pr(n_s|M_s(t)=j,u_s)\cdot p_{t,s,j}} \quad k\in\Omega_s.$$

The retailer can learn about the core process of product s only if the product is included in the assortment. Hence, if $u_s = 0$, then $\Pr(n_s|M_s(t)=j,u_s)$ is independent of $M_s(t)$.

We can now state the Bellman equation for the dynamic assortment problem with variable demand rates:

$$J^*_t(\boldsymbol{p}_t) = \max_{\substack{\boldsymbol{u}\geq\boldsymbol{0}:\\ \sum_{s=1}^S u_s \leq N}} \sum_{s=1}^{S} r_s\mathbb{E}[n_s] + \mathbb{E}_{\boldsymbol{n}}\left[J^*_{t-1}(\varphi_t(\boldsymbol{p}_t,\boldsymbol{u},\boldsymbol{n}))\right] \tag{6.4}$$

where $\varphi_t(\boldsymbol{p}_t,\boldsymbol{u},\boldsymbol{n}))$ is a vector function such that the s-th component is equal to $\varphi_{t,s}(\boldsymbol{p}_{t,s},u_s,n_s)$, and the expectations are calculated by conditioning on the states of the core processes. Note that $\boldsymbol{p}_t$ is a matrix (or a vector of vectors) that can have infinite dimension, and we are assuming that the retailer knows the set of possible states Ω_s for each core process.

As before, we relax the shelf-space constraint by introducing open-loop dual policies so the problem decouples into single product POMDPs, and the usual dual upper bound is available (at least in theory).

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 even the single product subproblems might be hard to solve and further approximations would be necessary (as in Aviv and Pazgal 2004). However, the dynamic assortment policy based on our index (cf. equation (3.14)) is still applicable and easy to implement.

6.2.4 Multiple Stores, Endogenous Demand, and Other Extensions

The dynamic assortment models in the thesis are based on the operations of fast-fashion retailers (like Zara), and we claim that this is the first attempt in following a quantitative/optimization approach. We introduced the basic model in Chapter 3 and then devoted the rest of the thesis to discuss, analyze, and provide guidance on

how to remove what we considered to be the three major assumptions we made (see §3.2). We now finally comment on model extensions based on the removal of other assumptions that could also be important in the case of a fast-fashion retailer.

Multiple Stores It is a fact that fast-fashion retailers manage simultaneously several store, some of them located within the same city or region, and therefore what is learned about demand in one store might be relevant information for another store with similar characteristics. In §4.1 we also discussed that considering multiple stores could be a path to follow when trying to overcome the lack of data required in the estimation of demand rates and correlation parameters. Then it seems natural to extend our model to multiple stores. The difficulties we see are related to the fact that we would no longer have a weakly coupled DP since there would be more than one coupling constraint, but most importantly, the independence assumption would not hold. As we mentioned, the interesting case is when the demands for the same product in different store are correlated, otherwise we go back to the framework developed in this thesis. Another potential issue that could arise is how to make local decision based on aggregated data. We visualize this topic as rich source of unanswered research questions.

Endogenous Demand In §3.5.4 we showed that in our model the assortment rotation tends to zero. In other words, at a certain point in time the assortment stabilizes and remains almost fix until de end of season. This result is reasonable given our assumption that demand rates are constant and exogenous. The previous section discussed the extension to variable demand rates, but still assuming that the actions of the retailer cannot stimulate demand. The Zara managers claim that having a permanent assortment rotation attracts customers to the store since they know that every four weeks they will find a new selection of products. Our model could be modify to accommodate a constraint that requires a minimum fraction of new products to be introduced in each period. However, it would be interesting to study the case when the frequency of changes in the assortment affects the arrival rate to the store. Then demand would be endogenous and in some sense it would be a model for repeated interactions between the retailer and his customers. Such type of models have started to emerge and receive attention in the literature, see for instance the pricing paper of Popescu and Wu 2005.

Creation of New Products During the Selling Season In our model we use the set $\mathcal{S}$ to represent all the potential products that can be included in the store during the selling season. This is rather a conceptual construct that we need to formulate the DP, but it implicity assumes that the retailer knows beforehand all the available products, which is not exactly the case of a fast fashion retailer that creates new products based on experts opinions and the feedback from customers. Then it might seem more appropriate to allow the set S to evolve over time. Such situation clearly resembles the "arm-acquiring bandit" studied in Whittle 1981. However, the author states on the first page that *the new projects shall be regarded as being very much variants as the old ones, occurring in a statistically homogeneous stream... by its nature, creative research (meaning innovative products in our context) cannot be formalized.* In other words, the set $\mathcal{S}$ cannot change unpredictably, which certainly supports our assumption. Moreover, our index policy does not depend on the set $\mathcal{S}$, but it would be interesting to see how it relates to the framework developed by Whittle.

Assortment Switching Costs We have assumed throughout the thesis that the assortment can be changed at no cost. This seems reasonable for a fast fashion retailer like Zara. If the switching cost were high, then it would be hard to understand how they manage to introduce 11,000 different product per year (compared to only 4,000 for a more traditional retailer). However, in other contexts the switching cost might be relevant, for instance in the bedding example described in §6.2.1. In that case we would have a multiarmed bandit with switching cost, a variant that has been studied in the literature, mostly under infinite horizon (see Agrawal *et al.* 1988). It has been shown that a bandit with switching cost achieves the same asymptotic performance as the original bandit without any costs. This is made possible by grouping together the samples in a certain fashion. We presume that such result does not translate to our setting because of the finite horizon, and then it would be important to study how much the switching cost affects the exploring capability.

Appendix A

On the concavity of $f_t(C)$

We want to study the (discrete) concavity of the following parametric function with respect to C:

$$f_t(\mathbf{m}, \boldsymbol{\alpha}; C) = \max_{\substack{\mathbf{u}\in\{0,1\}^S: \\ \sum_{s=1}^S u_s = C}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n}\cdot\boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\right] \quad \text{(A.1)}$$

For simplicity we will write $f_t(C)$. Provided that $r_s > 0\ \forall s$, from Lemma 3 we know that $f_t(C)$ is strictly increasing in C, with $C \leq S$. We want to verify the validity of the following inequality:

$$f_t(C+2) - f_t(C+1) \leq f_t(C+1) - f_t(C) \quad \forall C \in \{0, 1, \ldots, (S-2)\}. \quad \text{(A.2)}$$

Recall that N is the shelf space available in all the next periods (after t), and S is the total number of products.

For $t = 1$, we have that (A.2) holds since for that case the greedy solution is optimal, and therefore the marginal profit given by one additional unit of shelf space must be nonincreasing.

The simplest (non-trivial) case is then with two periods, two products (that provide the same net revenue) and enough shelf space for only one product. In other words, consider $t = 2$, $S = 2$, $r_1 = r_2 = 1$, and $N = 1$. Then we have:

$$
\begin{aligned}
f_2(0) &= \max\left\{\frac{m_1}{\alpha_1}, \frac{m_2}{\alpha_2}\right\} \\
f_2(1) &= \max\left\{\frac{m_1}{\alpha_1} + \mathbb{E}_{n_1}\left[\max\left\{\frac{m_1+n_1}{\alpha_1+1}, \frac{m_2}{\alpha_2}\right\}\right], \frac{m_2}{\alpha_2} + \mathbb{E}_{n_2}\left[\max\left\{\frac{m_1}{\alpha_1}, \frac{m_2+n_2}{\alpha_2+1}\right\}\right]\right\} \\
f_3(2) &= \frac{m_1}{\alpha_1} + \frac{m_2}{\alpha_2} + \mathbb{E}_{\boldsymbol{n}}\left[\max\left\{\frac{m_1+n_1}{\alpha_1+1}, \frac{m_2+n_2}{\alpha_2+1}\right\}\right]
\end{aligned}
$$

Using the fact that the products have independent demands and all random variables are nonnegative, we can rewrite the expectations above as:

$$
\begin{aligned}
\mathbb{E}_{n_1}\left[\max\left\{\frac{m_1+n_1}{\alpha_1+1}, \frac{m_2}{\alpha_2}\right\}\right] &= \frac{m_2}{\alpha_2} + \int_{\frac{m_2}{\alpha_2}}^{\infty} \Pr\left(\frac{m_1+n_1}{\alpha_1+1} > x\right) dx \\
\mathbb{E}_{n_2}\left[\max\left\{\frac{m_1}{\alpha_1}, \frac{m_2+n_2}{\alpha_2+1}\right\}\right] &= \frac{m_1}{\alpha_1} + \int_{\frac{m_1}{\alpha_1}}^{\infty} \Pr\left(\frac{m_2+n_2}{\alpha_2+1} > x\right) dx \\
\mathbb{E}_{\boldsymbol{n}}\left[\max\left\{\frac{m_1+n_1}{\alpha_1+1}, \frac{m_2+n_2}{\alpha_2+1}\right\}\right] &= \int_0^{\infty} \left(1 - \Pr\left(\frac{m_1+n_1}{\alpha_1+1} \leq x\right) \Pr\left(\frac{m_2+n_2}{\alpha_2+1} \leq x\right)\right) dx
\end{aligned}
$$

With no loss of generality, assume that $m_1 \geq m_2$. Recall that if $\alpha_1 \leq \alpha_2$, then n_1 is stochastically larger than n_2. Moreover, for that case we can show that:

- $\frac{m_1+n_1}{\alpha_1+1}$ is stochastically larger than $\frac{m_2+n_2}{\alpha_2+1}$.
- It is optimal in period $t = 2$ to select product 1. This means that the maximum in $f_2(1)$ is achieved by the first term.
- Inequality (A.2) is satisfied. This follows from using the two previous observations and the explicit formulas for the expectations given above.

Then a sufficient condition for (A.2) to hold is that n_1 and n_2 can be stochastically ordered. By the contrary, when $\alpha_1 > \alpha_2$, the function $f_t(C)$ is in general not concave in C. In fact, we can provide a counterexample. Consider $m_1 = 44$, $m_2 = 4$, $\alpha_1 = 10$, and $\alpha_2 = 1$. For this case it can be verified that $f_2(2) - 2f_2(1) + f_2(0) > 0$. The corresponding dual function is given by $q_2(\lambda) = \lambda + \max\{f_2(2) - 2\lambda, f_2(1) - \lambda, f_2(0)\}$. The next figure is a closeup of the graph of $q_2(\lambda)$ around its minimum.

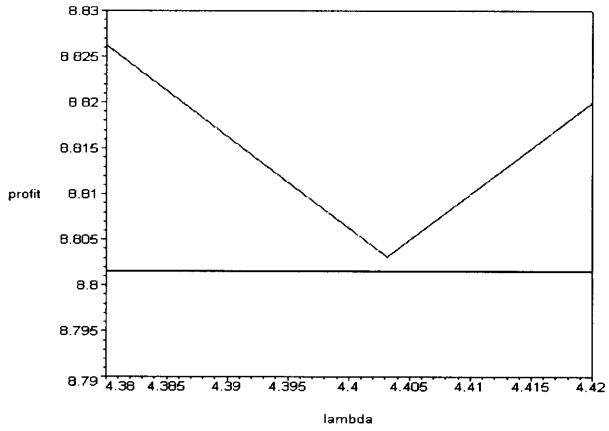

Figure A-1: The dual function $q_2(\lambda)$.

The horizontal line in Figure A-1 is the optimal profit of the two-period problem, which is equal to $f_2(1)$. From the figure is can be seen that there is a (small) gap, and then in this case Proposition 1 holds as a strict inequality.

As a final note, this small example also serves as a counterexample to show that the Gittins index is not optimal in the finite horizon case. In fact, in period $t = 2$, the optimal action is to include product 2 in the assortment. However, the Gittins indices are $\eta_{2,1} = 4.4556$ and $\eta_{2,2} = 4.4011$, so if we choose the highest one we would include product 1. Note that index values obtained using the closed-form approximation (3.14) are 4.4552 and 4.3904 for products 1 and 2 respectively.

Appendix B

Proofs

B.1 Proof of Proposition 1

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

$$\begin{aligned}
J_t^*(\boldsymbol{x}) &= \max_{\substack{\boldsymbol{u}\in\mathcal{U}:\\ \boldsymbol{a}'\boldsymbol{u}\leq N}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[J_{t-1}^*(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \\
&\leq N\lambda_t + \max_{\substack{\boldsymbol{u}\in\mathcal{U}:\\ \boldsymbol{a}'\boldsymbol{u}\leq N}} g_t(\boldsymbol{x},\boldsymbol{u}) - \lambda_t \boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[J_{t-1}^*(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \\
&\leq N\lambda_t + \max_{\boldsymbol{u}\in\mathcal{U}} g_t(\boldsymbol{x},\boldsymbol{u}) - \lambda_t \boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[J_{t-1}^*(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \\
&\leq N\lambda_t + \max_{\boldsymbol{u}\in\mathcal{U}} g_t(\boldsymbol{x},\boldsymbol{u}) - \lambda_t \boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[H_{t-1}^*(\varphi_t(\boldsymbol{x},\boldsymbol{u},\boldsymbol{n}))\right] \qquad \text{(B.1)}
\end{aligned}$$

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 (B.1) yields the desired result. □

B.2 Proof of Proposition 2

We start with an intuitive duality lemma:

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

$$h_t(\lambda_t, \boldsymbol{x}) = N \cdot \lambda_t + \max_{\boldsymbol{u}\in\mathcal{U}} \; g_t(\boldsymbol{x}, \boldsymbol{u}) - \lambda_t \boldsymbol{a}'\boldsymbol{u} + \mathbb{E}_{\boldsymbol{n}(\boldsymbol{x},\boldsymbol{u})}\left[J^*_{t-1}(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n}))\right]$$

Let $h^*_t(\boldsymbol{x}) = \min_{\lambda_t \geq 0} h_t(\lambda_t, \boldsymbol{x})$. If $f_t(\boldsymbol{x}; C)$ is concave and increasing in C, then $J^*_t(\boldsymbol{x}) = h^*_t(\boldsymbol{x})$.

Proof: Since the state $\boldsymbol{x}$ 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 $f_t(C)$ is increasing. 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:

$$h_t(\lambda_t) = N \cdot \lambda_t + \max_{C\in\,\mathbb{N}} f_t(C) - C \cdot \lambda_t. \tag{B.2}$$

Suppose that for λ^*_t the maximum on the right hand side of (B.2) 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 (B.2) 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 3), 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 $f_\tau(\boldsymbol{x}'; C)$ is concave in C for all $\tau = t-1, \ldots, 1$ and states $\boldsymbol{x}'$ reachable from $\boldsymbol{x}$ in period τ. For any $\boldsymbol{u} \in \mathcal{U}$ and any vector $\boldsymbol{n}$ we have that $\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n})$ is reachable from $\boldsymbol{x}$ in period $t-1$. Then, by the induction hypothesis we have that $J^*_{t-1}(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n})) = H^*_{t-1}(\varphi_t(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{n}))$. Using the latter we see that the last inequality in the proof of Proposition 1 (cf. B.1) 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. (2.3)) we have that $J^*_t(\boldsymbol{x}) = H^*_t(\boldsymbol{x})$, and the proof is complete. □

B.3 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 (3.3) completes the proof. □

B.4 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 $\text{argmax}_{\boldsymbol{u}\in\mathcal{U}} \sum_{s=1}^{S} r_s \mathbb{E}[\gamma_s] u_s$.

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. □

B.5 Proof of Lemma 3

We need the following lemma:

Lemma 7 *Let $(\boldsymbol{m}, \boldsymbol{\alpha})$ be the system state at period t. For any $i \in \mathcal{S}$ the following holds:*

$$\mathbb{E}_{n_i}\left[J_t^*(\boldsymbol{m} + n_i \boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})\right] \geq J_t^*(\boldsymbol{m}, \boldsymbol{\alpha}), \tag{B.3}$$

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

Proof: We proceed by induction on t. Assume that (B.3) 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:

$$g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha}) = \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\boldsymbol{n'}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{u})\right]. \tag{B.4}$$

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 (B.4) 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 (B.4) 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:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{n'}}\Big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\Big] &= \mathbb{E}_{\boldsymbol{n'_{-i}}}\Big[\mathbb{E}_{n'_i}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'} \cdot \boldsymbol{u}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u})\big]\Big] \\
&\geq \mathbb{E}_{\boldsymbol{n'_{-i}}}\big[J^*_{t-1}(\boldsymbol{m} + n_i\boldsymbol{e_i} + \boldsymbol{n'_{-i}} \cdot \boldsymbol{u_{-i}}, \boldsymbol{\alpha} + \boldsymbol{e_i} + \boldsymbol{u_{-i}})\big]
\end{aligned} \tag{B.5}$$

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:

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

The first equality follows from (B.4) 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 (B.5), and the last step is also given by (B.4) and the independence among product demands.

So we can conclude that:

$$\mathbb{E}_{n_i}[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})] \geq g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha}) \quad \forall \boldsymbol{u} \in \{0,1\}^S \tag{B.6}$$

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

$$\begin{aligned}
\mathbb{E}_{n_i}\left[J_t^*(\boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})\right] &= \mathbb{E}_{n_i}\left[\max_{\substack{\mathbf{u}\in\{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})\right] \\
&\geq \max_{\substack{\mathbf{u}\in\{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} \mathbb{E}_{n_i}\left[g_t(\boldsymbol{u}, \boldsymbol{m} + n_i\boldsymbol{e_i}, \boldsymbol{\alpha} + \boldsymbol{e_i})\right] \\
&\geq \max_{\substack{\mathbf{u}\in\{0,1\}^S: \\ \sum_{s=1}^S u_s \leq N}} g_t(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\alpha}) \\
&= J_t^*(\boldsymbol{m}, \boldsymbol{\alpha})
\end{aligned}$$

The first and last equality are given by the definition of $J_t^*(\cdot)$ and (B.4). 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 (B.6). Then the proof is complete. □

We can now prove Lemma 3.

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. (2.4)), and let i be such that $u_i^* = 0$. Then we have that:

$$f_t(\boldsymbol{m}, \boldsymbol{\alpha}; C) = \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s^* + \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}^*, \boldsymbol{\alpha} + \boldsymbol{u}^*)\right] \tag{B.7}$$

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 7 we have that:

$$\mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \overline{\boldsymbol{u}}, \boldsymbol{\alpha} + \overline{\boldsymbol{u}})\right] \geq \mathbb{E}_{\boldsymbol{n}}\left[J_{t-1}^*(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{u}^*, \boldsymbol{\alpha} + \boldsymbol{u}^*)\right]. \tag{B.8}$$

Since $r_i > 0$, from (B.8) we get a strict inequality relating the objective values of $\overline{\boldsymbol{u}}$ and $\boldsymbol{u}^*$:

$$\sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} \overline{u}_s + \mathbb{E}_{\boldsymbol{n}}\big[J^*_{t-1}(\boldsymbol{m} + \boldsymbol{n} \cdot \overline{\boldsymbol{u}}, \boldsymbol{\alpha} + \overline{\boldsymbol{u}})\big] > f_t(\boldsymbol{m}, \boldsymbol{\alpha}; C). \tag{B.9}$$

Since $\sum_{s=1}^{S} \overline{u}_s = C + 1$, from (B.9) 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. □

B.6 Proof of Lemma 4

We proceed by induction. Consider $t \geq 1$ and assume that (3.5) holds for $t-1$. Then, from equation (2.2):

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

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. □

B.7 Proof of Proposition 3

We first need the following two additional lemmas:

Lemma 8 $H^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) \leq (\frac{r_s m_s}{\alpha_s})t \quad \forall (m_s, \alpha_s)$.

Proof: Direct by induction since assuming that it holds for $t-1$ we can bound both terms in the right hand side of (3.6). In fact, we have that $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s, \alpha_s) \leq (t-1)r_s m_s/\alpha_s$ and

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

□

Lemma 9 $H^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) = 0 \quad \forall (m_s, \alpha_s)$ *such that* $\big(\frac{r_s m_s}{\alpha_s}\big)\tau < \lambda_\tau \ \forall \tau = t, \ldots, 1$.

Proof: Consider $t \geq 1$ 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 8 we have that:

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

Then, from equation (3.6) 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 7 it can be shown that $\mathbb{E}_{n_s}\big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s, \alpha_s+1)\big] \geq H^{\boldsymbol{\lambda}}_{t-1,s}(m_s, \alpha_s)$. Then, for α_s sufficiently small $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) \geq r_s \frac{m_s}{\alpha_s} - \lambda_t > 0$. On the other hand, when $\alpha_s \to \infty$, from Lemma 9 we have that $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s, \alpha_s) \to 0$. From Lemmas 9 and 8 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 $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) \to -\lambda_t < 0$. If the function $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s)$ were strictly decreasing in α_s, then $\beta^{\boldsymbol{\lambda}}_{t,s}(m_s)$ could be defined as the unique solution of $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s) = 0$, and if $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s)$ were strictly increasing in m_s, then $\beta^{\boldsymbol{\lambda}}_{t,s}(m_s)$ would inherit the same monotonicity property.

We now prove by induction on t that $d^{\boldsymbol{\lambda}}_{t,s}(m_s, \alpha_s)$ is indeed strictly decreasing in α_s and strictly increasing in m_s. The claim is trivial for $t = 1$ when clearly

$\beta^{\boldsymbol{\lambda}}_{1,s}(m_s) = r_s m_s/\lambda_1$. 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 $d^{\boldsymbol{\lambda}}_t(m, \alpha)$ is continuous in α, we only need to consider three cases:

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

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

$$\begin{aligned} d^{\boldsymbol{\lambda}}_t(m,\alpha) &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[H^{\boldsymbol{\lambda}}_{t-1}(m+n, \alpha+1) - H^{\boldsymbol{\lambda}}_{t-2}(m+n, \alpha+1)\Big] \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[\max\left\{d^{\boldsymbol{\lambda}}_{t-1}(m+n, \alpha+1), 0\right\}\Big] \end{aligned} \tag{B.10}$$

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

$$\begin{aligned} \mathbb{E}_{n''}\Big[\max\left\{d^{\boldsymbol{\lambda}}_{t-1}(m''+n'', \alpha''+1), 0\right\} &\geq \mathbb{E}_{n''}\Big[\max\left\{d^{\boldsymbol{\lambda}}_{t-1}(m'+n'', \alpha'+1), 0\right\}\Big] \\ &\geq \mathbb{E}_{n'}\Big[\max\left\{d^{\boldsymbol{\lambda}}_{t-1}(m'+n', \alpha'+1), 0\right\}\Big] \end{aligned} \tag{B.11}$$

  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 (B.10) that $d^{\boldsymbol{\lambda}}_t(m', \alpha') \leq d^{\boldsymbol{\lambda}}_t(m'', \alpha'')$.

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

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

$$\begin{aligned} d^{\boldsymbol{\lambda}}_t(m,\alpha) &= r\frac{m}{\alpha} - \lambda_t + \mathbb{E}_n\Big[H^{\boldsymbol{\lambda}}_{t-1}(m+n, \alpha+1)\Big] - H^{\boldsymbol{\lambda}}_{t-2}(m, \alpha) \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[H^{\boldsymbol{\lambda}}_{t-1}(m+n, \alpha+1)\Big] - \mathbb{E}_n\Big[H^{\boldsymbol{\lambda}}_{t-2}(m+n, \alpha+1)\Big] + d^{\boldsymbol{\lambda}}_{t-1}(m, \alpha) \\ &= \lambda_{t-1} - \lambda_t + \mathbb{E}_n\Big[\max\left\{d^{\boldsymbol{\lambda}}_{t-1}(m+n, \alpha+1), 0\right\}\Big] + d^{\boldsymbol{\lambda}}_{t-1}(m, \alpha) \end{aligned}$$

  Then $d^{\boldsymbol{\lambda}}_t(m', \alpha') \leq d^{\boldsymbol{\lambda}}_t(m'', \alpha'')$ follows from (B.11) and the induction hypothesis. Again, the inequality is strict if either $\alpha' < \alpha''$ or $m' < m''$.

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

  In this case we have:

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

  The first inequality holds because $\beta^{\boldsymbol{\lambda}}_{t-1}(m') \leq \alpha' \Rightarrow d^{\boldsymbol{\lambda}}_{t-1}(m',\alpha') \leq 0$. The second inequality follows from (B.11) and is strict if either $\alpha' < \alpha''$ or $m' < m''$. The proof is now complete. □

B.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 $(m_s,\alpha_s) \in B^{\boldsymbol{\lambda}}_{t,s}$, necessarily:

$$r_s\frac{m_s}{\alpha_s} - \lambda_t + \mathbb{E}_{n_s}\Big[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)\Big] < H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s). \tag{B.12}$$

Suppose that in period $t-1$ it is optimal to have $u_s = 1$, i.e. $(m_s,\alpha_s) \notin B^{\boldsymbol{\lambda}}_{t-1,s}$. Substituting the appropriate expression for $H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s)$ in (B.12) and rearranging terms yields:

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

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

B.9 Proof of Lemma 5

Both inequalities are proved by induction, and as before, the proof for $t = 1$ is a particular case of the induction step.

For the upper bound inequality we will show that for any state $(\boldsymbol{m}, \boldsymbol{\alpha})$ and any set of committed assortments $\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}$ the following holds:

$$J_t^0(\boldsymbol{m}, \boldsymbol{\alpha}) \geq J_t^*(\boldsymbol{v}^t, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{m}, \boldsymbol{\alpha}) \tag{B.13}$$

In fact, since $\boldsymbol{v}^t$ is a feasible assortment in period t, we have that:

$$\begin{aligned} J_t^0(\boldsymbol{m}, \boldsymbol{\alpha}) &\geq \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} v_s^t + E_{\boldsymbol{n}}\Big[J_{t-1}^0(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big] \\ &\geq \sum_{s=1}^{S} \sum_{\tau=t-\ell}^{t} r_s \frac{m_s}{\alpha_s} v_s^\tau + \mathbb{E}_{\boldsymbol{n}}\Big[W_{t-1}^*(\boldsymbol{v}^{t-1}, \ldots, \boldsymbol{v}^{t-\ell+1}, \boldsymbol{v}^{t-\ell}, \boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big] \end{aligned}$$

where we have used the inductions hypothesis and the definition of $J_{t-1}^*(\boldsymbol{v}^{t-1}, \ldots, \boldsymbol{v}^{t-\ell}, \boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)$. We then maximize the second inequality with respect to $\boldsymbol{v}^{t-\ell}$ to obtain (B.13). The desired results follows from maximizing the right hand side of $(B.13)$ with respect to the committed assortments.

Now we prove the lower bound inequality in Lemma 5. In order to avoid excessive notation, we will show the result for the case when the lead time is equal to one period ($\ell = 1$). The extension to the general case is analogous.

Again, via induction, we prove that the following inequality holds:

$$W_t^*(\boldsymbol{v}^t, \boldsymbol{m}, \boldsymbol{\alpha}) \geq J_{\lceil \frac{t-1}{2} \rceil}^0(\boldsymbol{m}, \boldsymbol{\alpha}) + \mathbb{E}_{\boldsymbol{n}}\Big[J_{\lfloor \frac{t-1}{2} \rfloor}^0(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big] \tag{B.14}$$

In fact, assume that inequality (B.14) is valid for t. Then for $t + 1$ we have:

$$\begin{aligned} W^*_{t+1}(\boldsymbol{v}^{t+1}, \boldsymbol{m}, \boldsymbol{\alpha}) &= \max_{\boldsymbol{v}^t \in \mathcal{U}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} v_s^t + \mathbb{E}_{\boldsymbol{n}}\Big[W^*_t(\boldsymbol{v}^t, \boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^{t+1}, \boldsymbol{\alpha} + \boldsymbol{v}^{t+1})\Big] \\ &\geq \mathbb{E}_{\boldsymbol{n}}\Big[J^0_{\lceil \frac{t-1}{2} \rceil}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^{t+1}, \boldsymbol{\alpha} + \boldsymbol{v}^{t+1})\Big] + \\ &\quad \max_{\boldsymbol{v}^t \in \mathcal{U}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} v_s^t + \mathbb{E}_{\boldsymbol{n}}\Big[J^0_{\lfloor \frac{t-1}{2} \rfloor}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big] \\ &= \mathbb{E}_{\boldsymbol{n}}\Big[J^0_{\lfloor \frac{t}{2} \rfloor}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^{t+1}, \boldsymbol{\alpha} + \boldsymbol{v}^{t+1})\Big] + J^0_{\lceil \frac{t}{2} \rceil}(\boldsymbol{m}, \boldsymbol{\alpha}) \end{aligned}$$

In the second inequality we have used the following property:

$$\mathbb{E}_{\boldsymbol{n}'}\Big[\mathbb{E}_{\boldsymbol{n}}\big[J^0_{\lfloor \frac{t-1}{2} \rfloor}(\boldsymbol{m} + \boldsymbol{n}' \cdot \boldsymbol{v}^{t+1} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^{t+1} + \boldsymbol{v}^t)\big]\Big] \geq \mathbb{E}_{\boldsymbol{n}}\Big[J^0_{\lfloor \frac{t-1}{2} \rfloor}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big]$$

which can proved in a similar way as Lemma 7, and in the last equality we have used the identities $\lfloor \frac{t-1}{2} \rfloor + 1 = \lceil \frac{t}{2} \rceil$ and $\lceil \frac{t-1}{2} \rceil = \lfloor \frac{t}{2} \rfloor$.

Using the previous inequality we can finally prove the lower bound in Lemma 5:

$$\begin{aligned} J^1_t(\boldsymbol{m}, \boldsymbol{\alpha}) &= \max_{\boldsymbol{v}^t \in \mathcal{U}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} v_s^t + W^*_t(\boldsymbol{v}^t, \boldsymbol{m}, \boldsymbol{\alpha}) \\ &\geq J^0_{\lceil \frac{t-1}{2} \rceil}(\boldsymbol{m}, \boldsymbol{\alpha}) + \max_{\boldsymbol{v}^t \in \mathcal{U}} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} v_s^t + \mathbb{E}_{\boldsymbol{n}}\Big[J^0_{\lfloor \frac{t-1}{2} \rfloor}(\boldsymbol{m} + \boldsymbol{n} \cdot \boldsymbol{v}^t, \boldsymbol{\alpha} + \boldsymbol{v}^t)\Big] \\ &= J^0_{\lfloor \frac{t}{2} \rfloor}(\boldsymbol{m}, \boldsymbol{\alpha}) + J^0_{\lceil \frac{t}{2} \rceil}(\boldsymbol{m}, \boldsymbol{\alpha}) \end{aligned}$$

And the proof is complete. □

Bibliography

[1] Adelman, D. and A.J. Mersereau. 2004. Relaxations of Weakly Coupled Stochastic Dynamic Programs. Working Paper. Graduate School of Business. University of Chicago.

[2] Agrawal, R. M.V. Hedge, and D. Teneketzis. 1988. Asymptotically Efficient Adaptive Allocation Rules for the Multiarmed Bandit Problem with Switching Cost. *IEEE Transactions on Automatic Control.* **33**(10) pp. 899-906.

[3] Agrawal, N. and S.A. Smith. 1996. Estimating Negative Binomial Demand for Retail Inventory Management with Unobservable Lost Sales. *Navel Research Logistics.* **43** 839-861.

[4] 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.

[5] 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.

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

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

[8] Azoury, K.S. 1985. Bayes Solution to Dynamic Inventory Models Under Unknown Demand Distribution. *Management Science.* **31** 1150-1160.

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

[10] Bertsekas, D. 1999. *Nonlinear Programming.* Athena Scientific, Cambridge.

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

[12] Bertsimas, D. and A.J. Mersereau. 2004. A Learning Approach to Customized Marketing. Working Paper. Graduate School of Business. University of Chicago.

[13] Bertsimas, D. and J. Ñiño-Mora. 1996. Conservation Laws, Extended Polymatroids and Multiarmed Bandit Problems: a Unified Approach to Indexable Systems. *Mathematics of Operations Research.* **21** 257-306.

[14] Bertsimas, D. and J. Ñiño-Mora. 2000. Restless Bandits, Linear Programming Relaxations and Primal-Dual Index Heuristics. *Operations Research.* **48**(1) 80-90.

[15] Braden, D.J. and M. Freimer. 1991. Information Dynamics of Censored Observations. *Management Science.* **37** 1390-1404.

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

[17] Bultez, A. and P. Naert. 1988. SHARP: Shelf Allocation for Retailers Profit. *Marketing Science.* **7** 211-231.

[18] Cope, E. 2004. Nonparametric Strategies for Dynamic Pricing in E-Commerce. Working paper. The Sauder School of Business, University of British Columbia.

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

[20] Chen, L. and E. Plambeck. 2004. Dyanamic Bayesian Quantity Control When Some Customers Will Accept a Substitute. *Manufacturing and Service Operations Management.* **6**(1) 103-106.

[21] Clark, A.J. and H. Scarf. 1960. Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science.* **6** 475-490.

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

[23] Ding, X., M.L. Puterman, and A. Bisi. 2002. The Censored Newsvendor and the Optimal Aquisition of Information. *Operations Research.* **50**(3) 517-527.

[24] Eppen, G.D. and A.V. Iyer. 1997. Improved Fashion Buying with Bayesian Updates. *Operations Research.* **45**(6) 805-819.

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

[26] 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.

[27] 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.

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

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

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

[31] 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.

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

[33] Hadley, G. and T.M. Whitin. 1963. *Analysis of Inventory Systems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

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

[35] Hammond, J.H. and A. Raman. 1994. Sport Obermeyer Ltd. Harvard Business School Case 9-695-022.

[36] Hardwick, J., R. Oehmke, and Q.F. Stout. 2005. New Adaptive Designs for Delayed Response Models. To appear in *Journal of Sequential Planning and Inference.*

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

[38] Karmarkar U.S. 1987. The Multilocation Multiperiod Inventory Problem: Bounds and Approximations. *Management Science.* **33**(1) 86-94.

[39] Kök, A. G. and M. L. Fisher. 2004. Demand Estimation and Assortment Optimization Under Substitution: Methodology and Application. Working paper. Duke University.

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

[41] 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.

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

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

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

[45] Martello, S. and P. Toth. 1990. *Knapsack Problems: Algorithms and Computer Implementations*, John Wiley and Sons, New York, NY.

[46] McAfee, A., V. Dessain, and A. Sjöman. 2004. ZARA: IT for Fast Fashion. Harvard Business School Case 9-604-081.

[47] Puterman, M. 1990. Markov Decision Processes, in *Handbooks in Operations Research and Management Science.* Eds. D.P. Heyman and M.J. Sobel. Vol. 2. North-Holland. Amsterdam.

[48] Popescu, I. and Y. Wu. 2005. Dynamic Pricing Strategies under Repeated Interactions. Working Paper. INSEAD, Fontainebleau, France.

[49] Rajaram K. and U.S. Karmarkar. 2002. Product Cycling with Uncertain Yields: Analysis and Application to the Process Industry. *Operations Research.* **50**(4) 680-691.

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

[51] Samphaiboon, N. and T. Yamada. 2000. Heuristic and Exact Algorithms for the Precedence-Constrained Knapsack Problem. *Journal of Optimization Theory and Applications.* **105**(3) pp. 659676.

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

[53] Talluri, K. and G. van Ryzin. 1998. An Analysis of Bid-Price Controls for Network Revenue Management. *Management Science.* **44**(11) 1577-1593.

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

[55] Weber, R.R. and G. Weiss. 1990. On an Index Policy for Restless Bandits. *Journal of Applied Probability.* **27** 637-648.

[56] Whittle, P. 1981. Arm-Acquiring Bandits. *The Annals of Probability.* **9**(2) pp. 284-292.

[57] 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.

[58] Yost, K.A. and A.R. Washburn. 2000. The LP/POMDP Marriage: Optimization with Imperfect Information. *Naval Research Logistics.* **47** 607-619.