MANAGEMENT SCIENCE
Vol. 53, No. 2, February 2007, pp. 276–292
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5302 | 0276

informs

DOI 10.1287/mnsc.1060.0613

Dynamic Assortment with Demand Learning for Seasonal Consumer Goods

Felipe Caro
Anderson School of Management, University of California, Los Angeles, California 90095,
fcaro@anderson.ucla.edu

Jérémie Gallien
Sloan School of Management, Massachusetts Institute of Technology, 30 Wadsworth Street,
Cambridge, Massachusetts 02142, jgallien@mit.edu

Companies such as Zara and World Co. have recently implemented novel product development processes and supply chain architectures enabling them to make more product design and assortment decisions during the selling season, when actual demand information becomes available. How should such retail firms modify their product assortment over time in order to maximize overall profits for a given selling season? Focusing on a stylized version of this problem, we study a finite horizon multiarmed bandit model with several plays per stage and Bayesian learning. Our analysis involves the Lagrangian relaxation of weakly coupled dynamic programs (DPs), results contributing to the emerging theory of DP duality, and various approximations. It yields a closed-form dynamic index policy capturing the key exploration versus exploitation trade-off and associated suboptimality bounds. In numerical experiments its performance proves comparable to that of other closed-form heuristics described in the literature, but this policy is particularly easy to implement and interpret. This last feature enables extensions to more realistic versions of the motivating dynamic assortment problem that include implementation delays, switching costs, and demand substitution effects.

Key words: retail assortment; dynamic programming duality; bayesian learning; multiarmed bandit
History: Accepted by Paul H. Zipkin, operations and supply chain management; received January 13, 2005.
This paper was with the authors $8\frac{1}{2}$ months for 2 revisions.

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, and having to use markdowns to sell the many unpopular products still accumulating in their stores (see Fisher et al. 2000).

Recently, a few innovative firms including Spain-based Zara, Mango, and Japan-based World Co. (sometimes referred to as "fast-fashion" companies) have implemented 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 are partly achieved through an increased reliance on more costly local production relative to the supply networks of traditional retailers. The contrast between these two supply chain design alternatives seems particularly drastic, as shown in Table 1.

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: Given the constantly evolving demand information available, which products should be included in the assortment at each point in time? Figure 1 provides a conceptual representation of this operational challenge. In each period over a finite season (T), the retailer must decide the subset (N) of products that will be offered from a larger set (S) of all candidates. As sales occur, the retailer gathers new demand information about each particular product included in the latest assortment, which may be combined with prior historical demand information to select a future assortment.

This problem seems challenging 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

Table 1 Retail Industry Benchmark

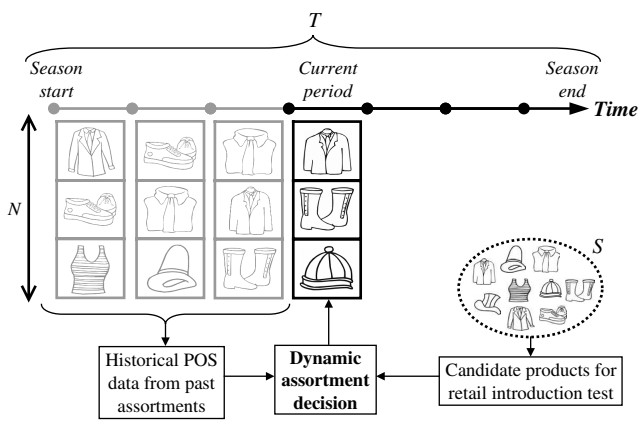
	Time to market	In-house production after season starts (%)	Different products manufactured per year	Products sold with discount (%)	Average discount (%)
Traditional retailer	6–9 months	0–20	2,000–4,000	30–40	30
Zara	2–5 weeks	85	∼11,000	15–20	15

Source. Ghemawat and Nueno (2003).

like to gather more demand information (exploration) and may be more profitable in the long run. In other words, how to balance learning with immediate profit. Our main objective in this paper is to develop and analyze a stylized optimization model capturing the main features of this dynamic assortment problem. Our results and contributions include (i) a closed-form solution for this model with learning that only requires knowledge of the two first moments of demand; (ii) an extension of that solution accounting for lead times, switching costs, and substitution effects; and (iii) a better understanding of when the ability to learn has more impact in this setting.

Our approach consists of two steps. After an overview of the relevant literature in §1.2, we consider in §2 a first version of our motivating dynamic assortment problem, which as described in §2.1, amounts to a finite-horizon multiarmed bandit model with several plays per stage and Bayesian learning. The associated analysis shown in §2.3 results in a closed-form policy. While we later report (in §4) that its performance is near-optimal and on par with that of other comparable heuristics available in the literature, our policy is particularly easy to implement and interpret. This last feature enables, as a second step presented in §3, the extension of our policy to more realistic versions of our motivating dynamic assortment problem that include implementation delays (§3.1), switching costs (§3.2), and demand substitution effects (§3.3). In §4, we describe the numerical experiments we executed assess the performance of the suggested policies and provide indications on the primary drivers of this performance. Finally, §5 contains our concluding remarks, and the proofs are available in the appendix.

Figure 1 The Dynamic Assortment Problem



1.2. Literature Review

We first discuss contextually related papers focusing on assortment problems. A first subset is found in the marketing literature, in which several studies, typically motivated by supermarkets, consider static assortment problems formulated as deterministic nonlinear optimization models. See Bultez and Naert (1988) for a classical example in which the demand of a product depends on the allocated shelf space, and the overall space available is a limited resource. In the operations management literature, van Ryzin and Mahajan (1999) and Smith and Agrawal (2000) also consider 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 at 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. For additional references on assortment problems, we refer the reader to Kök and Fisher (2004).

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. Although papers associated with the quick response initiative, such as the seminal work by Fisher and Raman (1996), do emphasize learning and exploiting early sales data, the demand information acquired over time is primarily used by the manufacturer to improve ordering and production, as opposed to product design or assortment decisions.

From a methodological standpoint, the stylized model we analyze in the first part of our paper is based on the following version of the much studied multiarmed bandit problem (Berry and Fristedt 1985): 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.

A remarkable result for the multiarmed bandit problem with infinite horizon ($T = \infty$), one arm pulled per stage ($N = 1$), and a discount factor strictly smaller than one, is due to Gittins (see Gittins 1979). The so-called Gittins index of an arm s is defined as the lump reward expected by a player indifferent between retiring or playing arm s individually. The optimal policy is then to play in each stage the arm with the highest Gittins index. In the finite horizon case ($T < \infty$), however, it is known that Gittins' index policy is generally not optimal (Berry and Fristedt 1985), and much research has focused on developing near-optimal heuristics. In particular, the policies developed by Ginebra and Clayton (1995) and Brezzi and Lai (2002) constitute natural benchmarks to our policy. In contrast, although the allocation rule proposed by Anantharam et al. (1987) for a frequentist version of that problem is asymptotically efficient, it does not seem directly applicable to our specific environment. Our analysis, based on Lagrangian decomposition, is also closely related to the work in Bertsimas and Mersereau (2004) for an adaptive sampling problem. However, they assume a Beta-Bernoulli learning model, whereas we use the Gamma-Poisson model, and they do not provide a suboptimality bound for their derived policy. Finally, although Whittle (1988) focuses on restless bandits, we owe much to his work in that he also uses a Lagrangian relaxation and considers multiple plays per stage.

The more realistic models that we consider in the second part of our paper are extensions of the basic multiarmed bandit just described. Bandit models with response delays seem to have received only moderate attention in the past (see Hardwick et al. 2006 and references therein), and we are in particular not aware of any other closed-form policy (let alone suboptimality bounds) described in the literature for this problem. There are likewise several papers considering infinite horizon bandit problems with switching costs (e.g. Agrawal et al. 1988, Brezzi and Lai 2002). Although Gittins' index policy is no longer optimal for such a model, other policies are known to be asymptotically optimal.[1] However, we have not found studies considering a multiarmed bandit problem with switching costs over a finite horizon. Finally, a limited amount of research has been done for bandit problems with dependent arms (see, for instance, Presman and Sonin 1990), but we do not know of any heuristic described elsewhere for the case in which, as in this paper, rewards follow a correlation structure given by a demand substitution model similar to that of Smith and Agrawal (2000) and Kök and Fisher (2004).

[1] The asymptotic regime considered is defined by the discount factor tending to one.

2. A Stylized Multiarmed Bandit Model

The material presented in this section constitutes an intermediary step towards the construction in §3, of more realistic policies for our motivating dynamic assortment problem. In §2.1, we first introduce a stylized multiarmed bandit model. Then in §2.2, we discuss modeling assumptions and motivate subsequent extensions. The model analysis and the derivation of our closed-form index policy is presented in §2.3.

In this paper periods are counted backwards, boldface symbols represent vectors, subscripts represent the components of a vector, superscripts represent elements in a sequence, and r.h.s. means "right-hand side."

2.1. Problem Definition

We consider a finite horizon multiarmed bandit model in which a player gets to pull N arms out of S in each one of T periods.[2] The reward per period for each pulled arm s is equal to $r_s n_s$, where r_s is a known positive constant and n_s is an independent Poisson random variable with constant but unknown mean γ_s. The objective is to maximize the total expected reward from all arms pulled over all periods.

Our basic dynamic assortment model is equivalent to the multiarmed bandit just described by means of the following analogy: Each arm corresponds to a product. Pulling an arm is the same as including that product in the assortment. The random variable n_s is the number of sales of product s (if it is included in the assortment), and r_s is the unit gross margin. The length of the selling season is given by T, and the assortment decisions are made by a retailer who wants to maximize total season profits. We continue to use this language in the remainder of the paper.

We adopt a standard Gamma-Poisson Bayesian learning mechanism (also used, for instance, in Aviv and Pazgal 2002). The retailer starts each period with a prior belief of the value of the unknown demand rate γ_s represented by a Gamma distribution with (positive) shape parameter m_s and (positive) scale parameter α_s, so that $\mathbb{E}[\gamma_s] = m_s/\alpha_s$ and $\mathbb{V}[\gamma_s] = m_s/\alpha_s^2$. The predictive distribution for n_s is then a negative binomial with parameters m_s and $\alpha_s(\alpha_s+1)^{-1}$given by

$$\Pr(n_s) = \binom{n_s + m_s - 1}{m_s - 1}\left(\frac{1}{\alpha_s+1}\right)^{n_s}\left(\frac{\alpha_s}{\alpha_s+1}\right)^{m_s}. \quad (1)$$

When necessary, we write $n_s(m_s, \alpha_s)$ to make the parameter dependence explicit. According to Bayes' rule, the posterior distribution on γ_s after observing n_s sales still has a Gamma distribution. For each arm s,

[2] With no loss of generality, we assume the lengths of these periods to be identical.

the parameters (m_s, α_s) of the belief distribution on γ_s are updated between consecutive periods as

$$(m_s, \alpha_s) \longrightarrow \begin{cases} (m_s + n_s, \alpha_s + 1) \\ \quad \text{if product } s \text{ is in the assortment,} \\ \quad\quad \text{and } n_s \text{ sales are observed} \\ (m_s, \alpha_s) \\ \quad \text{if product } s \text{ is not in} \\ \quad\quad \text{the assortment.} \end{cases} \tag{2}$$

The parameter vector $\mathbf{I}^t = (\mathbf{m}, \boldsymbol{\alpha})$ provides a natural dynamic programming state representation for each decision period t, following the dynamics described by (2).[3] The decision to include product s in the assortment can be represented by a binary variable $u_s \in \{0, 1\}$, where $u_s = 1$ means that product s is included. The set $\mathcal{U}$ representing all feasible actions (i.e., the control space) can then be defined as $\mathcal{U} = \{\mathbf{u} \in \{0, 1\}^S : \sum_{s=1}^S u_s \le N\}$, and the optimal profit-to-go function $J_t^*(\mathbf{m}, \boldsymbol{\alpha})$ given state $(\mathbf{m}, \boldsymbol{\alpha})$ and t remaining periods satisfies the following Bellman equation:

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

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

Note that the only link between consecutive periods in this model is the information acquired about the observed sales n_s, and that different products are coupled only through the constraint $\sum_{s=1}^S u_s \le N$. This type of problem is known as a weakly coupled dynamic program (DP). Clearly we must have $S > N$, otherwise the retailer would always include all available products. Observe also that the summation on the r.h.s. of (3) includes the immediate expected reward associated with each product and represents the exploitation component, while the expectation term that follows captures the future benefits from exploration.

2.2. Model Discussion

The analogy between the stylized multiarmed bandit model and the dynamic assortment problem introduced at the beginning of §2.1 implies a number of assumptions about the retailer's environment. We now comment on the most salient ones.

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

Firstly, we note that in the basic multiarmed bandit model, model decisions become effective immediately. This assumption seems particularly strong since design, production, and transportation delays may, in reality, induce an implementation lag of several weeks. For this reason, in §3.1 we present an extension of our proposed policy accounting for positive lead times.

Another strong underlying assumption is that assortments may be changed at no cost. In reality, introducing new products may entail some additional design, production, and store setup costs, while destocking existing products may entail additional transportation and inventory salvage costs. We thus consider in §3.2 an extension of our basic model capturing such switching costs.

Third, customer demands for similar but distinct products in the same period are typically not independent as assumed earlier. For example, most customers set on a specific style would choose only one color among several available, introducing negative demand correlations across products. In §3.3, we show how our assortment policy may be modified to account for such substitution effects.

In addition, our model assumes that the underlying demand rate of each product is constant and exogenous throughout the season, mostly for tractability reasons. While demand stationarity may be a strong assumption in some settings, we observe that an important reason why demand nonstationarity may arise in practice is the use of dynamic pricing. However, we assume that prices remain constant throughout the season (the margin r_s of every product s is fixed), which seems partly justified by the figures reported in §1 showing that fast-fashion retailers rely less frequently on markdown policies, and that when they do so their price markdowns are also lower.

The store's limited shelf space (or desire to limit in-store product variety as a result of deliberate operational or marketing decisions) is captured by the constraint that the assortment in each period may include at most a fixed number (N) of different products. We are thus implicitly assuming that all products require the same shelf space. Also, although the set of all (S) candidate products would include in practice both the products already available when the season starts and all the variants and new products that may be designed during the season, our analysis does not recognize that this set may change over time. The policy we develop may, however, still be implemented by ignoring the impact on present decisions of future changes in the set of candidate products.

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, we focus

for each product on assortment inclusion or exclusion as opposed to order quantity, and inventory holding costs are ignored. We observe that assortment design seems a higher level consideration than inventory management, partly justifying this modeling choice. It may still be a strong assumption in many settings, however, and we refer the reader to Caro (2005) for a modified problem formulation including order quantity decisions and censorship of demand information from stockouts.

Finally, our model considers only the assortment problem faced by a single store. Considering several stores would require a richer demand structure, specifying, in particular, how sales observed in one store should impact the demand forecasts of other stores. From a mathematical perspective, the resulting DP would no longer be weakly coupled unless demands for different stores are assumed independent, so the analysis would likely follow a very different path. Our policy might still serve as a good starting point for that case, but this development is not addressed here.

We turn now to the analysis of the stylized bandit model described in §2.1.

2.3. Analysis

2.3.1. DP Duality. In light of the computational complexity associated with solving Equation (3) exactly, our objective is not to characterize the optimal solution but rather to find one that is near-optimal, simple, and easily interpretable, so it can be extended later to capture the additional complexities of the dynamic assortment problem. We thus use an approximate solution method, 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 Bertsekas 1999). The approach dates back to at least the late 1980s 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. For more accounts of successful applications of this methodology, see Castañon (1997), Bertsimas and Mersereau (2004), and the references therein.

Specifically, let $\lambda_t(\mathbf{m}, \boldsymbol{\alpha})$ denote any function associated with period t that maps the state space into the set of nonnegative real values. We define a dual policy to be any vector of 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 $(\mathbf{m}, \boldsymbol{\alpha})$, the corresponding profit-to-go is obtained by solving the dual dynamic program given by

$$H_t^{\boldsymbol{\lambda}_t}(\mathbf{m}, \boldsymbol{\alpha}) = N\lambda_t(\mathbf{m}, \boldsymbol{\alpha}) + \max_{\mathbf{u}\in\{0,1\}^S} \sum_{s=1}^{S} \left(r_s \frac{m_s}{\alpha_s} - \lambda_t(\mathbf{m}, \boldsymbol{\alpha}) \right) u_s + \mathbb{E}_{\mathbf{n}}[H_{t-1}^{\boldsymbol{\lambda}_{t-1}}(\mathbf{m} + \mathbf{n}\cdot\mathbf{u}, \boldsymbol{\alpha} + \mathbf{u})], \quad (4)$$

with $H_0^{\boldsymbol{\lambda}_0}(\mathbf{m}, \boldsymbol{\alpha}) = 0 \ \forall (\mathbf{m}, \boldsymbol{\alpha})$. In words, a dual policy gives the price of pulling one arm for each period and each possible state. A dual policy $\boldsymbol{\lambda}_t$ is optimal if it minimizes the r.h.s. of (4) for any initial state. Let $H_t^*(\mathbf{m}, \boldsymbol{\alpha})$ be the profit-to-go of the optimal dual policy for a given state $(\mathbf{m}, \boldsymbol{\alpha})$, which can be obtained recursively using standard dynamic programming theory. Our first result is the following proposition, which we use later, in particular, when establishing an upper bound on the optimal profit-to-go:[4]

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

As in classical duality theory, an interesting theoretical question is to determine whether the first inequality in Proposition 1 ever holds as an equality. This question is studied in Caro (2005), where conditions are given for strong DP duality to hold.

2.3.2. Problem Decomposition and Upper Bound. Solving the dual DP problem (4) seems about as hard as solving the original primal problem (3), motivating further simplifications. Specifically, we now restrict our attention to *open-loop dual policies*, in which the shadow price of the coupling constraint 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.[5] In the following, we will refer to the profit-to-go corresponding to an open-loop dual policy $\boldsymbol{\lambda}$ as $H_t^{\boldsymbol{\lambda}}(\cdot)$ instead of the previous notation $H_t^{\boldsymbol{\lambda}_t}(\cdot)$. The next lemma shows that with open-loop policies the dual DP decomposes into S single-product subproblems.

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

$$H_t^{\boldsymbol{\lambda}}(\mathbf{m}, \boldsymbol{\alpha}) = N \sum_{\tau=1}^{t} \lambda_\tau + \sum_{s=1}^{S} H_{t,s}^{\boldsymbol{\lambda}}(m_s, \alpha_s), \quad (5)$$

[4] Results similar to Proposition 1 and Lemma 1 (to be introduced shortly) can be found in Hawkins (2003).

[5] Open-loop policies are called *deterministic multipliers* and *restricted Lagrangian* by Castañon (1997) and Karmarkar (1987), respectively.

where

$$H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s) = \max\left\{ r_s\frac{m_s}{\alpha_s} - \lambda_t + \mathbb{E}_{n_s}[H^{\boldsymbol{\lambda}}_{t-1,s}(m_s+n_s,\alpha_s+1)], \; H^{\boldsymbol{\lambda}}_{t-1,s}(m_s,\alpha_s)\right\}, \quad (6)$$

and the first term in the r.h.s. corresponds to $u_s = 1$, *the second term to* $u_s = 0$.

It is clear from (6) that for any fixed state (m_s, α_s), $H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)$ is nondecreasing with t. Also, it can be shown that $H^{\boldsymbol{\lambda}}_{t,s}(m_s,\alpha_s)$ is a convex and piecewise linear function of $(\lambda_t,\ldots,\lambda_1)$ that is strictly increasing in m_s and strictly decreasing in α_s. The optimal policy for this single-product subproblem can be characterized by a collection of T threshold functions (one per period). Besides, when the shadow prices are nondecreasing (i.e., $\lambda_t \le \lambda_{t-1}$), then the stopping set at period t is a subset of the stopping set at period $t-1$. Formal statements and proofs of these properties can be found in Caro (2005).

The weak duality result (Proposition 1) implies that an upper bound for the optimal expected profit obtained by considering the best open-loop dual policy:

$$J^*_t(\mathbf{m},\boldsymbol{\alpha}) \le \min_{\boldsymbol{\lambda}\ge\mathbf{0}} H^{\boldsymbol{\lambda}}_t(\mathbf{m},\boldsymbol{\alpha}), \quad (7)$$

where (5) and (6) can be substituted in the r.h.s., and the associated minimization problem can be solved with a convex nondifferentiable optimization algorithm. This method yields the upper performance bound we use later to assess the suboptimality of our index policy.

Note that finding the best open-loop dual policy, i.e., solving (7), is equivalent to solving the original (primal) problem when the coupling constraint is no longer required to be satisfied for each possible sample path, but only on average in each period. That is, in each period the constraint $\sum_{s=1}^{S} u_s \le N$ is replaced by $\mathbb{E}[\sum_{s=1}^{S} u_s \le N]$, where the expectation is with respect to all possible states weighted by the probability of reaching each one of them under a given (primal) policy. This fact has been observed by several authors in various other settings (e.g., Whittle 1988, Castañon 1997), and a proof of this equivalence for the finite horizon multiarmed bandit is available in Caro (2005). We also point out that Adelman and Mersereau (2004) provide an alternative linear programming-based bound that is shown to be tighter (or not worse) than (7), but requires more extensive computations. Finally, bounds similarly based on optimal open-loop dual policies have been proven to be asymptotically tight in other settings (Weber and Weiss 1990).

2.3.3. A Closed-Form Dynamic Index Policy. The decomposition into single-product subproblems defined by (6) enables us to derive a closed-from index policy for our multiarmed bandit problem. We proceed in two steps.

Step 1. *A General Framework for Index Policies* (*Whittle's Heuristic*). We impose $\lambda_t = \lambda$ for all t, i.e., the opportunity cost of pulling an arm is assumed to be the same in all periods (and all states). In that case, it is easy to show that:

$$H^{\lambda}_{t,s}(m_s,\alpha_s) = \max\{d^{\lambda}_{t,s}(m_s,\alpha_s),0\} \quad \text{with } d^{\lambda}_{t,s}(m_s,\alpha_s) = r_s\frac{m_s}{\alpha_s} - \lambda + \mathbb{E}_{n_s}[H^{\lambda}_{t-1,s}(m_s+n_s,\alpha_s+1)]. \quad (8)$$

Let $u^{\lambda}_{t,s}$ be the optimal decision in the single-product subproblem defined by (8). For any product s, we have that $\lim_{\lambda\to 0} u^{\lambda}_{t,s} = 1$ and $\lim_{\lambda\to\infty} u^{\lambda}_{t,s} = 0$. Moreover, it follows from (8) that $H^{\lambda}_{t,s}(m_s,\alpha_s)$ is nonnegative and nonincreasing in λ. Consequently, there must exist $\eta_{t,s} \ge 0$ such that $u^{\lambda}_{t,s} = 1$ if and only if $\lambda \le \eta_{t,s}$. The shelf space opportunity cost threshold $\eta_{t,s}$ is thus well-defined and its value can be obtained by solving the nonlinear equation $d^{\lambda}_{t,s}(m_s,\alpha_s) = 0$ for λ. Moreover, following the definition given in §1.2, the threshold $\eta_{t,s}$ multiplied by t is in fact the "Gittins index" for our version of the bandit problem.

We interpret the threshold $\eta_{t,s}$ as the degree to which it is desirable to include product s in the assortment when the information state is $(\mathbf{m},\boldsymbol{\alpha})$ at time t. Hence, from now on we refer to $\eta_{t,s}$ as the *exact* desirability index, where the adjective "exact" distinguishes it from the approximation to be described shortly.

Given the previous interpretation, a natural policy for the dynamic assortment problem is to put in the store the N most desirable products. That is, at any time t, include in the assortment the N products with the largest desirability indices $\eta_{t,s}$. Such type of heuristic policy was first introduced by Whittle (1988) for an infinite horizon model with restless bandits. The specific index he proposes is also derived as a break-even Lagrange multiplier, and its exact computation is a complicated task that can only be done numerically, as is the case in our model. The policy we suggest also consists of selecting the N most desirable products, but based instead on an approximation for $\eta_{t,s}$ that we derive next.

Step 2. *Derivation of a Closed Form Index Formula*. The derivation of our closed-form approximation for $\eta_{t,s}$ is based on two simple ideas:

- First, we implement a 1-step lookahead (1-sla) approximation. That is, the profit-to-go $H^{\lambda}_{t-1,s}(m_s,\alpha_s)$ is approximated using the single-period profit by

$$\tilde{H}^{\lambda}_{t-1,s}(m_s,\alpha_s) \triangleq (t-1)\cdot\max\left\{ r_s\frac{m_s}{\alpha_s} - \lambda, 0\right\}. \quad (9)$$

Substituting $H^{\lambda}_{t-1,s}(m_s, \alpha_s)$ with (9) in (8), we can then approximate $d^{\lambda}_{t,s}(m_s, \alpha_s)$ by

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

where

$$\begin{aligned} b^{\lambda}_s &= \left(\frac{\lambda}{r_s} - \mathbb{E}[\gamma_s]\right)\frac{\sqrt{\mathbb{V}[n_s]}}{\mathbb{V}[\gamma_s]}, \\ \mathbb{E}[n_s] &= \frac{m_s}{\alpha_s}, \quad \text{and} \quad \mathbb{V}[n_s] = \mathbb{E}[n_s]\left(\frac{\alpha_s + 1}{\alpha_s}\right). \end{aligned} \tag{10}$$

The second equality is obtained through direct algebraic manipulation (similar to the example in Berry and Fristedt 1985, p. 12). The moments of γ_s are given in §2.1.

• Second, 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, so we approximate it by a normal distribution with the same mean and variance. By the Central Limit Theorem, the approximation is asymptotically exact as m_s increases. This yields

$$\tilde{d}^{\lambda}_{t,s}(m_s, \alpha_s) \approx r_s \frac{\mathbb{V}[\gamma_s]}{\sqrt{\mathbb{V}[n_s]}}((t-1)\cdot \Psi(b^{\lambda}_s) - b^{\lambda}_s), \tag{11}$$

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

Because $\Psi(z)$ is continuous, positive and strictly decreasing (cf. DeGroot 1970, p. 247), the equation $(t-1)\cdot\Psi(z_t) = z_t$ has a unique solution for all $t \geq 1$. Moreover, the values z_t, which are independent of the problem data, are increasing and concave in t (see Table 2 for the first few numerical values of z_t with four-digit accuracy). It is easy to verify that the previous 1-sla approximation (9) underestimates the profit-to-go. Therefore, the values of z_t are rather conservative when t is large. In the numerical §4.2 we show a possible amendment.

Recall that the exact desirability index $\eta_{t,s}$ comes from solving the equation $d^{\lambda}_{t,s}(m_s, \alpha_s) = 0$ for λ. If instead we use the approximation $\tilde{d}^{\lambda}_{t,s}(m_s, \alpha_s)$ given by (11), then we obtain the following approximate expression for $\eta_{t,s}$:

$$\eta^{CG}_{t,s} \triangleq r_s\left(\mathbb{E}[\gamma_s] + z_t \frac{\mathbb{V}[\gamma_s]}{\sqrt{\mathbb{V}[n_s]}}\right). \tag{12}$$

Table 2 **First Values of z_t**

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

Even though it is an approximation, we still refer to Equation (12) as the desirability index. To avoid any confusion with the true value $\eta_{t,s}$, we will refer to the latter in the following as the *exact* desirability index.

Note that the first term in the r.h.s. of (12) favors exploitation of the current demand forecast $\mathbb{E}[\gamma_s]$, whereas the second term favors exploration, because it is increasing in both the variance of γ_s (i.e., the uncertainty of that forecast) and the number of remaining periods (through z_t). Intuitively, when uncertainty about demand for product s (captured by $\mathbb{V}[\gamma_s]$) is high, there is more benefit to learn from including s because of the upside potential from future gains. Furthermore, the last term in Equation (12) shows that our policy favors such learning strategy only to the extent that the uncertainty associated with the demand rate estimate ($\mathbb{V}[\gamma_s]$) is high relative to the total uncertainty on demand ($\mathbb{V}[n_s]$), which also includes the structural uncertainty associated with the underlying stochastic (Poisson) demand process. Because only the first type of uncertainty can be resolved, this feature effectively amounts to separating out the stochastic noise introduced by the intrinsic demand randomness when assessing the desirability of learning. Because resolving the estimation uncertainty does take some time, however, one may not be able to benefit from this learning with only a few periods left before the end of the season. 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 (12). Finally, the fact that our policy depends on only the first two moments of the demand rates γ_s is a desirable feature from an implementation standpoint. In particular, the estimation procedure based on experts opinions described by Fisher and Raman (1996) could be used to estimate the initial priors.

Alternative approaches that also provide closed-form approximations similar to Equation (12) have been described in the literature. The development is typically for infinite horizon bandits but can be adapted to our case. In particular, Ginebra and Clayton (1995) develop a formula assuming a priori that the exact desirability index $\eta_{t,s}$ is normally distributed, and Brezzi and Lai (2002) obtain another closed-form expression using a diffusion approximation. The policies resulting from the last two papers are compared with ours in the numerical experiments (see §4.2).

3. Extensions for a Dynamic Assortment Problem

We now present in §§3.1, 3.2, and 3.3 some extensions of our proposed policy for the stylized multiarmed

bandit model to more realistic environments. As will soon be clear, all three extensions may easily be performed simultaneously, although for clarity of exposition we describe only marginal modifications relative to our proposed policy defined in (12).

3.1. Assortment Implementation Lead Time

We now assume a lag l between the period t when an assortment decision is made and the period $t-l$ at which this assortment is actually implemented in the store. Although this implementation lag l arises in practice from delays associated with all process steps between design and storage on the shelf, in the following we will refer to l as the "lead time."

The state space in the DP model must now be extended to keep track of past decisions yet to be implemented. Specifically, the state is now given by the vector $(\mathbf{v}^t, \ldots, \mathbf{v}^{t-l+1}, \mathbf{m}, \boldsymbol{\alpha})$, where $\mathbf{v}^t, \ldots, \mathbf{v}^{t-l+1}$ are the assortments that will be offered from the current period t down to period $t-l+1$, and $(\mathbf{m}, \boldsymbol{\alpha})$ are the distribution parameters of the beliefs about demand at time t. The decision made at time $t \in \{T+l, \ldots, l+1\}$ is the assortment that will be implemented at time $t-l$, and the first l assortments $\mathbf{v}^T, \ldots, \mathbf{v}^{T-l+1}$ 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:

$$J_t^*(\mathbf{v}^t, \ldots, \mathbf{v}^{t-l+1}, \mathbf{m}, \boldsymbol{\alpha}) = \sum_{s=1}^{S} \sum_{\tau=t-l+1}^{t} r_s \frac{m_s}{\alpha_s} v_s^\tau + W_t^*(\mathbf{v}^t, \ldots, \mathbf{v}^{t-l+1}, \mathbf{m}, \boldsymbol{\alpha}) \quad (13)$$

where $W_0^* = \cdots = W_l^* = 0$ for any state, and $W_t^*(\cdot)$ satisfies for $t > l$:

$$W_t^*(\mathbf{v}^t, \ldots, \mathbf{v}^{t-l+1}, \mathbf{m}, \boldsymbol{\alpha}) = \max_{\sum_{s=1}^{S} u_s \le N} \sum_{s=1}^{S} r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\mathbf{n}}\left[W_{t-1}^*(\mathbf{v}^{t-1}, \ldots, \mathbf{v}^{t-l+1}, \mathbf{u}, \mathbf{m} + \mathbf{n} \cdot \mathbf{v}^t, \boldsymbol{\alpha} + \mathbf{v}^t)\right]. \quad (14)$$

The summation in the r.h.s. of (13) 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 $t-l-1$. Note that if $l = 0$, then $J_t^*(\mathbf{m}, \boldsymbol{\alpha}) = W_t^*(\mathbf{m}, \boldsymbol{\alpha})$ and (14) reduces then to the recursion (3) studied in the previous subsections. As is clear from the expansion of the state space by a factor of $2^{S \times l}$, 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 (14):

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

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

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

Moreover, we can invoke arguments similar to the ones used in §2.3.2 to obtain the following upper bound for the maximization of $J_T^*(\mathbf{v}^T, \ldots, \mathbf{v}^{T-l+1}, \mathbf{m}, \boldsymbol{\alpha})$ with respect to $(\mathbf{v}^T, \ldots, \mathbf{v}^{T-l+1})$, subject to the corresponding binary and shelf space constraints:

$$\min_{\boldsymbol{\lambda}} N \sum_{\tau=1}^{T} \lambda_\tau + \sum_{s=1}^{S} \max_{\substack{v_s^T, \ldots, v_s^{T-l+1} \\ \in \{0,1\}}} \left(\sum_{\tau=T-l+1}^{T} \left(r_s \frac{m_s}{\alpha_s} - \lambda_\tau \right) v_s^\tau + H_{t,s}^{\boldsymbol{\lambda}}(v_s^T, \ldots, v_s^{T-l+1}, m_s, \alpha_s) \right), \quad (15)$$

which provides the upper bound that we will report in our numerical experiments for the performance of various policies simulated in environments with a positive lead time.

Finally, our proposed policy may be heuristically adapted by introducing the two following modifications to the desirability index $\eta_{t,s}^{CG}$ given by Equation (12):

1. First, we substitute the term z_t in (12) with $z_{L(t)}$, where $L(t) = \max\{t - 2l, 1\}$. The rationale is that in period t the retailer must decide the assortment of period $(t-l)$, and from then on he has l fewer periods to learn about demand. In particular, if $l \ge (t-1)/2$, then $z_{L(t)} = 0$ so that the adapted index policy coincides with the greedy policy (see §4.2.1), which can be shown to generate optimal actions in that case. Note that if $l \ge T - 1$, 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 case would correspond to the "traditional retailer" described in §1.1.

2. The second modification in (12) concerns the variance $\mathbb{V}[\gamma_s]$. Recall from §2.1 that the priors become more accurate as more sales are observed. Hence, the prediction made at time t for the variance of γ_s at time $t-l$ must take into account whether product s is committed as part of the assortment in

any of the l periods in between. Specifically, we use the following expression for the variance of γ_s:

$$\mathbb{V}[\gamma_s] = \frac{m_s + (m_s/\alpha_s)\sum_{\tau=t-l+1}^{t} v_s^\tau}{(\alpha_s + \sum_{\tau=t-l+1}^{t} v_s^\tau)^2}, \tag{16}$$

where, as before, $\sum_{\tau=t-l+1}^{t} v_s^\tau$ 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 $t-l$, considering how many times product s will have been part of the assortment by then. Intuitively, substitution (16) 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 (16), the second term in the desirability index formula (12) now decreases with the sum $\sum_\tau v_s^\tau$, reflecting that when designing the assortment for period $t-l$, the incentive to explore the demand for product s reduces when it already has a large presence in the next l assortments.

3.2. Switching Costs

We now consider the case in which an additional cost $\omega_{t,s}$ is incurred whenever a product s not included in the assortment in period $t+1$ becomes part of it in period t (i.e., the following period). Assuming with seemingly little loss of generality that the profit and cost parameters are such that the retailer will always use all available shelf space, the parameter $\omega_{t,s}$ may also include the cost associated with removing the product replaced by product s in the assortment (the portion of $\omega_{t,s}$ accounting for such removal cost would be product-independent however).

In the presence of such switching costs, the state space of the DP must be extended so as to keep track of the assortment $\mathbf{v}^{t+1}$ implemented in the previous period (with $\mathbf{v}^{T+1} = \mathbf{0}$). The Bellman equation thus becomes

$$\begin{aligned} J_t^*(\mathbf{v}^{t+1}, \mathbf{m}, \boldsymbol{\alpha}) &= \max_{\mathbf{u}\in\{0,1\}^S:\sum_{s=1}^S u_s \le N} \sum_{s=1}^{S} \left(r_s \frac{m_s}{\alpha_s} - \omega_{t,s}(1 - v_s^{t+1}) \right) u_s \\ &\quad + \mathbb{E}_{\mathbf{n}}[J_{t-1}^*(\mathbf{u}, \mathbf{m} + \mathbf{n}\cdot\mathbf{u}, \boldsymbol{\alpha} + \mathbf{u})]. \end{aligned} \tag{17}$$

As in the previous subsection, the duality concepts developed in §§2.3.1 and 2.3.2 still apply, and in particular, an upper bound on the optimal profit-to-go can be computed from the following single-product subproblem:

$$\begin{aligned} H_{t,s}^{\boldsymbol{\lambda}}(v_s^{t+1}, m_s, \alpha_s) = \max\Big\{ & r_s \frac{m_s}{\alpha_s} - \omega_{t,s}(1 - v_s^{t+1}) - \lambda_t \\ & + \mathbb{E}_{n_s}\big[H_{t-1,s}^{\boldsymbol{\lambda}}(1, m_s + n_s, \alpha_s + 1)\big], H_{t-1,s}^{\boldsymbol{\lambda}}(0, m_s, \alpha_s) \Big\}. \end{aligned} \tag{18}$$

Finally, we can heuristically adapt our proposed policy so that it accounts for switching costs through the following modification of the desirability index formula (12):

$$\eta_{t,s}^{CG} = r_s \left(\mathbb{E}[\gamma_s] + z_t \frac{\mathbb{V}[\gamma_s]}{\sqrt{\mathbb{V}[n_s]}} \right) - \frac{\omega_{t,s}(1 - v_s^{t+1})}{t}, \tag{19}$$

where the switching cost $\omega_{t,s}$, amortized over t periods, is subtracted if product s was not part of the previous assortment ($v_s^{t+1} = 0$). In that situation, it is intuitively clear that the presence of a switching cost makes it less attractive to include product s, and that the desirability index $\eta_{t,s}^{CG}$ should then be reduced somehow. However, attributing the switching cost in its entirety to the period in which it is incurred would ignore the fact that the product might stay in the assortment for more than one period and so the cost should be shared. Equation (19) specifically amortizes the additional cost over t periods, an accounting rule that would be correct if the product remained in the assortment until the end of the season. Since this is not always the case, that rule underestimates the cost of introducing products in the assortment. However, we have found that such cost underestimation is balanced by the underestimation of the future benefits from learning due to the limited lookahead approximation (see §2.3.3), so that in the end policy (19) performs relatively well (see §4.4).

3.3. Substitution Effects

Designing and analyzing a dynamic assortment model in which learning concerns not only the demand rates of individual products but also their correlation structure is a challenging task. Even if a Bellman equation similar to (3) could be written for such a model, the corresponding DP would no longer be weakly coupled because of the many relationships between different products introduced by the correlation structure, so that our duality-based decomposition approach would likely break down.

An alternative (admittedly simpler) design path, which we now develop, is to assume that the correlation structure across products is known (or can at least be estimated upfront and then passively updated), and the individual demand rates of individual products must be estimated dynamically as before. In the marketing literature, substitution effects are often captured through parametric models like the multinomial logit (see Ben-Akiva and Lerman 1985). However, we adopt a substitution model similar to Smith and Agrawal (2000) and Kök and Fisher (2004). That is, we 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, we assume that each customer makes only one such substitution attempt. Defining the (total) substitution probability as $L_i \triangleq \sum_{s \neq i} q_{is}$, note that the case $L_i < 1$ would capture customers leaving without buying any substitute to product i. Our dynamic index policy defined by (12) can then be adapted heuristically by performing the two following modifications:

1. 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 $\mathbf{u} \in \mathcal{U}$ 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 $\tilde{n}_s$ of product s is then given by:

$$\tilde{n}_s = n_s \cdot \left(\frac{\mathbb{E}[\gamma_s]}{\mathbb{E}[\gamma_s] + \sum_{i \neq s} q_{is} \mathbb{E}[\gamma_i](1 - u_i)} \right). \quad (20)$$

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 $m_s + \tilde{n}_s$, 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 (20).

2. The index $\eta^{CG}_{t,s}$ derived in §2.3.3 is a measure of the desirability of 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 then becomes a combinatorial problem, which we propose to address through the following quadratic integer program (QIP):

$$\max_{\mathbf{u} \in \{0,1\}^S : \sum_{s=1}^S u_s \leq N} \sum_{s=1}^{S} \left(\eta^{CG}_{t,s} + r_s \sum_{i \neq s} q_{is} \mathbb{E}[\gamma_i](1 - u_i) \right) u_s. \quad (21)$$

In words, the objective in (21) evaluates the profitability of including each product s in the assortment at t by adding to the initial desirability index $\eta^{CG}_{t,s}$ 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 (21) results in our original index policy.

4. Numerical Experiments

We pursue two goals with our numerical experiments: (i) evaluate the policies proposed in this paper for the basic multiarmed bandit model considered (cf. §4.2) and its extensions to environments with an assortment implementation lead time (cf. §4.3), switching costs (cf. §4.4), and substitution effects (cf. §4.5); and (ii) understand the factors that most sensitively affect performance.

4.1. Methodology

There seems to be two accepted methodologies for evaluating policy performance in environments involving learning. The first one, known as the *Bayesian approach* and adopted for example in Aviv and Pazgal (2002), relies on the assumption that the predictive Bayesian distribution updated in each period (in our case, the negative binomial distribution (1)) is 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 us to specifically focus on the quality of the index policy as a solution to the self-contained dynamic program given by the Bellman Equation (3). This is the approach we follow in §§4.2, 4.3, and 4.4.

The second methodology, known as the *frequentist approach*, and adopted for example in Bertsimas and Mersereau (2004), relies on the specification of real underlying distribution parameters (in our case, the demand rates γ_s). In simulations, actual demand for each product in each period is generated from the corresponding (Poisson) distributions. These experiments thus allow to specifically characterize how the relative performance of different policies may be affected by the quality of the information initially available (e.g., accuracy and bias). This is the method we follow in §4.5, because we have not rigorously formulated a Bayesian learning model or DP for the environment with substitution effects. See Caro (2005) for the frequentist version of §§4.2 and 4.3.

The simulation and upper bound optimization code we used was written in C, and is available from the first author upon request. In all simulation experiments, we ran a number of replications sufficient to obtain a relative estimation error smaller than 0.5% for a confidence level of 95%. When computing the duality-based upper bounds mentioned in §§2 and 3, the support of the negative binomial distribution

was truncated to exclude values with probability lower than 10^{-6}. Solutions to the corresponding nondifferentiable minimization problem were computed using the Nelder-Mead simplex method. While this algorithm is not generally guaranteed to converge to an optimal solution (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 DP duality). One iteration of this method could take up to several minutes for the most complicated instance described below ($T = 40$ and $\mathbb{V}[\gamma_s] = 100$). In some cases we tried different starting points, and report then the best bound we have found.

4.2. Basic Multiarmed Bandit Model

4.2.1. Experiments Description. For this first set of experiments, we used a data set with $N = 30$, $S = 720$, and $T \in \{10, 20, 40\}$. The unit profits r_s used throughout were obtained through one set of S independent draws from a uniform distribution $U[2, 8]$. We considered identical initial priors across all products to represent the case when the retailer has poor initial information. The results with different priors are qualitatively the same but more difficult to interpret. We set the initial expected demand rate $\mathbb{E}[\gamma_s]$ equal to 10 units per period, but tested three scenarios corresponding to an initial prior variance $\mathbb{V}[\gamma_s] \in \{5, 50, 100\}$.[6]

We were interested in comparing the performance of our proposed policy with that of others described in the literature for this problem and also in assessing the impact of the various approximations made when deriving (12). We thus considered the following policies:

$CG^{\text{1-sla}}_{\text{norm}}$ (1-sla and normal approximations). This is the basic policy derived in §2.3.3 that relies on the desirability index formula (12).

$CG^{\text{regr}}_{\text{norm}}$ (normal approximation and linear regression). This modified version of the previous policy was designed to assess the impact of the underestimation of future profits associated with the 1-sla approximation. The amendment we have explored consists of replacing the factor $(t-1)$ appearing in both Equation (9) and the equation $z_t = (t-1)\cdot\Psi(z_t)$ defining z_t by a function of $t-1$ increasing faster than linearly, specifically $\nu(t) \triangleq a(t-1)^b$. To find good values for parameters a and b, we have first computed numerically the exact desirability index $\eta_{s,t}$ for a small sample of instances that we can then extrapolate. In particular, we chose $t = 3, \ldots, 40$ and $(m_s, \alpha_s) = (1, 1/10)$, $(2, 1/5)$, $(10, 1)$, and $(20, 2)$. Second, we have fitted these exact values to the regression model

$$r_s\frac{m_s}{\alpha_s} - \eta_{s,t}(m_s, \alpha_s) + a(t-1)^b\left(\frac{m_s}{\alpha_s^2}\right)^c \cdot \mathbb{E}_{n_s}\left[\left[r_s\frac{m_s+n_s}{\alpha_s+1} - \eta_{s,t}(m_s, \alpha_s)\right]^+\right] = 0, \quad (22)$$

where $\eta_{s,t}(m_s, \alpha_s)$ are the independent variables and (a, b, c) are the parameters to be determined. Note that (22) can be converted to a linear regression model through a simple logarithmic transformation. The fit was very good ($R^2 = 0.976$), and all parameters were significant, with $a = 1.868$, $b = 1.212$, and $c = -0.262$. The factors z_t used in the desirability index formula (12) were finally replaced with the (larger) solutions $\tilde{z}_t$ of the modified equation $\tilde{z}_t = \nu(t)\cdot\Psi(\tilde{z}_t)$.

$CG^{\text{1-sla}}_{\text{neg. bin}}$ (1-sla approximation only). This modified version of our earlier policy $CG^{\text{1-sla}}_{\text{norm}}$ consists of computing numerically the indices by solving in each period and for each product s the relevant nonlinear equation obtained from (10) when n_s is no longer assumed to be normally distributed, but rather follows the original negative binomial distribution defined in (1).

GC (Ginebra-Clayton policy). This policy selects in each period the N products with the highest values of the indices $\eta^{GC}_{t,s}$ calculated dynamically with the formula

$$\eta^{GC}_{t,s} \triangleq r_s\left(\mathbb{E}[\gamma_s] + k\sqrt{\mathbb{V}[\gamma_s]}\right), \quad (23)$$

where k is a constant parameter. Equation (23) is the natural adaptation to our setting of the heuristic index formula developed by Ginebra and Clayton (1995) for the response surface bandit. Note that its performance is sensitive to the value chosen for parameter k, which must be calibrated somehow. To enable a meaningful comparison, we have performed a simulation-based linear search in k for each data set considered and report the results corresponding to the optimal choice of k thus identified.

BL (Brezzi-Lai policy). This dynamic index policy relies on the closed-form approximation of Gittins' index described in Brezzi and Lai (2002). Because it is derived for a discounted infinite horizon model, we adapt it to our finite horizon setting by using the natural correspondence $\beta = 1 - 1/t$, where β is the discount factor appearing in their original formula. The resulting index formula is

$$\eta^{BL}_{t,s} \triangleq r_s\left(\mathbb{E}[\gamma_s] + \psi\left(\frac{\mathbb{V}[\gamma_s]}{\log\left(\frac{t}{t-1}\right)\mathbb{E}[\gamma_s]}\right)\right), \quad (24)$$

where $\psi(\cdot)$ is the increasing function with $\psi(0) = 0$ defined in Brezzi and Lai (2002, p. 93).

GDY (Greedy policy). This policy consists of selecting in each period the N products with the highest immediate expected profit $r_s\mathbb{E}[\gamma_s]$ (thus greedily

[6] These scenarios are equivalent to (m_s, α_s) equal to $(20, 2)$, $(2, 1/5)$, and $(1, 1/10)$, respectively.

favoring exploitation over exploration). Note that it does involve learning despite its myopic nature, since priors are still updated in each period with observed demand, only the impact of assortment decisions on learning is ignored. As a result, several authors (e.g., Aviv and Pazgal 2002) also refer to this policy as *passive learning*.

Note that we are only including in this benchmark study policies from the literature that can be implemented using closed-form expressions. The main reason for this is that it is not clear at all how other policies can be adapted to our specific environment and furthermore extended to include lead times, switching costs, or substitution effects. These considerations led us to exclude in particular the policies described in Anantharam et al. (1987) and Whittle (1988). One could also think of policies based on the dual upper bound (7). However, the long running time currently required to calculate this upper bound would jeopardize any practical implementation of such policy, and assessing its performance through simulation would be even more prohibitive. Despite our focus on the narrow set of policies just defined, we observe that the small suboptimality bounds reported in the next section suggest that the performance superiority of any excluded policy should be small in most instances and in practice well within the range of data estimation errors.

4.2.2. Results. Table 3 reports under the heading *DualBnd* the upper bound for total expected profit derived using DP duality (see §2.3.1) divided by the number of periods, and the simulated performance gap of each policy relative to that bound. For each data set, we highlight in bold the suboptimality bound of the best policy.

Our main observations about these results are the following:

1. A first legitimate comparison involves policies *GDY*, *BL*, and $CG^{\text{1-sla}}_{\text{norm}}$, the only ones that are readily implementable across all data sets using only closed-form expressions. *GDY* performed worst, and $CG^{\text{1-sla}}_{\text{norm}}$ slightly outperformed *BL*, although this was not statistically significant.

2. A second natural comparison involves policies *GC* and $CG^{\text{regr}}_{\text{norm}}$, which are still closed form but both require some computation at the outset before implementation. Calibrating factor k of *GC* required a simulation-based search for each data set. On the contrary, computing the finite time horizon factors $\tilde{z}_t$ of $CG^{\text{regr}}_{\text{norm}}$ involved calculating a sample of 152 exact desirability indices and then performing a linear regression. Note, however, that this last procedure was done only once for all data sets shown in Table 3. Moreover, policy $CG^{\text{regr}}_{\text{norm}}$ outperformed *GC* in all nine data sets tested. This seems to indicate that policy $CG^{\text{regr}}_{\text{norm}}$ dominates policy *GC* not only in ease of computation, but also in terms of performance.

3. Policy $CG^{\text{1-sla}}_{\text{neg. bin}}$ also performed consistently better than $CG^{\text{1-sla}}_{\text{norm}}$. These results suggest that both the normal and the 1-sla approximations mentioned in §2.3.3 result in a small but clearly measurable loss of performance. In addition, the performance loss associated with the 1-sla approximation seems slightly larger than that resulting from the normal approximation.

4. The fact that the average relative suboptimality bound of the best policy is 1% (with a maximum of 2.87%) over all scenarios considered not only shows the good performance of that policy, but it also suggests that the duality-based upper bound (7) used to compute it is quite tight.

5. Finally, we also observe that all active learning policies outperform the greedy policy *GDY*, and that their relative superiorities increase with the number of periods and initial prior variances. Our interpretation is that larger initial prior variances and larger numbers of periods increase the opportunities to leverage demand learning. On the contrary, for data sets with particularly short horizons ($T < 6$), we have observed that the greedy policy performs nearly identically to all other policies. Other studies involving Bayesian learning models (e.g., Aviv and Pazgal 2002, Brezzi and Lai 2002) report similar findings.

Table 3 Performance Benchmark for the Stylized Multiarmed Bandit Problem

$\mathbb{V}[\gamma_s]$	T	*GDY* (%)	*BL* (%)	$CG^{\text{1-sla}}_{\text{norm}}$ (%)	*GC* (%)	$CG^{\text{1-sla}}_{\text{neg. bin}}$ (%)	$CG^{\text{regr}}_{\text{norm}}$ (%)	*DualBnd*
5	10	0.40	0.07	0.15	0.13	0.08	**0.06**	2,608.05
	20	0.86	0.31	0.25	0.23	0.29	**0.14**	2,693.97
	40	3.28	2.06	1.90	1.78	1.83	**1.68**	2,819.91
50	10	4.27	0.37	0.59	1.20	0.42	**0.06**	3,656.37
	20	9.20	1.35	1.33	1.13	0.64	**0.46**	4,133.26
	40	16.15	3.75	4.00	3.61	3.14	**2.87**	4,664.85
100	10	6.58	2.79	0.91	1.71	0.42	**0.24**	4,311.70
	20	13.83	4.57	2.75	2.82	1.80	**1.46**	5,130.00
	40	21.14	3.91	4.33	3.20	2.52	**2.41**	5,883.58
Average		8.41	2.13	1.80	1.76	1.24	**1.04**	
Max.		21.14	4.57	4.33	3.61	3.14	**2.87**	

4.3. Assortment Implementation Lead Time

4.3.1. Experiments Description. For this second series of experiments, we used the same data sets as described in §4.2.1, but we considered an assortment implementation lead time l equal to five periods. In addition to $CG^{\text{1-sla}}_{\text{norm}}$ and *GDY* (see §4.2.1), we also simulated the following policies:

CG-LT. Our proposed policy in environments with an assortment implementation lead time, obtained by

Table 4 Performance Benchmark in Environments with a Positive Lead Time ($l = 5$)

$\mathbb{V}[\gamma_s]$	T	*GDY* (%)	$CG^{\text{1-sla}}_{\text{norm}}$ (%)	*CG-LT*$_{z_t}$ (%)	*CG-LT*$_{\text{Var}}$ (%)	*CG-LT* (%)	*DualBnd*
5	10	1.09	1.15	1.09	1.47	**0.60**	2,456.12
	20	3.32	3.46	3.25	1.88	**0.76**	2,608.58
	40	4.96	4.36	4.25	2.08	**1.61**	2,753.84
50	10	8.89	9.69	8.89	11.54	**0.11**	2,864.40
	20	24.93	26.31	24.37	14.82	**3.90**	3,945.55
	40	27.35	26.37	24.88	9.27	**4.20**	4,589.80
100	10	15.60	16.90	15.60	18.57	**3.46**	3,206.80
	20	33.18	34.76	32.40	19.10	**4.32**	4,787.70
	40	34.70	33.76	31.97	10.52	**3.89**	5,754.43
Average		17.11	17.42	16.30	9.92	**2.54**	
Max.		34.70	34.76	32.40	19.10	**4.32**	

applying to $\eta^{CG}_{s,t}$ the two modifications described at the end of §3.1.

CG-LT$_{z_t}$ (respectively, *CG-LT*$_{\text{Var}}$). The policy obtained by only applying to our basic dynamic index policy $CG^{\text{1-sla}}_{\text{norm}}$ the first (respectively, second) modification described in §3.1, that is a correction of the weighting factor z_t for the remaining horizon (respectively, a correction of the predicted prior variance after l periods).

4.3.2. Results. Table 4 reports the duality-based upper bound (still noted *DualBnd*) on the expected profit per period provided by (15) and the simulated performance gap of all policies just mentioned relative to that bound.

The observations we draw from Table 4 are the following:

1. Our proposed policy *CG-LT* still performs close to optimal in these new (admittedly more complex) environments. Its average suboptimality bound averages 2.54% with a maximum of 4.32% over the range of scenarios considered. As before, these numbers also indicate that the duality-based bound (15) is relatively tight.

2. Both modifications made to policy $CG^{\text{1-sla}}_{\text{norm}}$ when constructing policy *CG-LT* result in clearly measurable and, particularly when combined, very significant performance improvements, as evidenced by the relative standings of the policies $CG^{\text{1-sla}}_{\text{norm}}$, *CG-LT*$_{z_t}$, *CG-LT*$_{\text{Var}}$, and *CG-LT*. With positive implementation lead times, it thus seems important to take into account the resulting reduced potential for leveraging information (modification of z_t), and even more important to appropriately change predictions of future information quality (modification of $\mathbb{V}[\gamma_s]$).

3. The relative performance of the greedy policy deteriorates with the lead time, as seen by comparing the suboptimality of *GDY* in Table 4 (where $l = 5$) with that shown in Table 3 (where $l = 0$). This is because the greedy policy in this setting not only ignores the future benefits from learning, but also disregards the l assortments that are on their way to the store.

4.4. Switching Costs

4.4.1. Experiments Description. For these experiments, we considered a high switching cost scenario with $\omega_{t,s} = \omega = 50$, a low switching cost scenario with $\omega_{t,s} = \omega = 20$, and data sets otherwise identical to those described in §4.2.1. In addition to $CG^{\text{1-sla}}_{\text{norm}}$ and *GDY* (see §4.2.1), we simulated the following policies:

CG-SC. Our proposed dynamic index policy for this environment, characterized by Equation (19).

CG-SC$_\omega$. Same as the previous policy, except that the total switching cost $\omega(1 - v^{t+1})$ is subtracted in the desirability index formula (19) instead of the amortized one.

GDY-SC. The adaptation of the greedy policy to this environment, including in the assortment the N products with the highest value of the greedy index $r_s\mathbb{E}[\gamma_s] - \omega(1 - v_s^{t+1})/t$, which predicts immediate profits corrected for amortized switching costs as in Equation (19), but omits the learning term.

4.4.2. Results. Table 5 reports the duality-based upper bound on expected profit per period, now computed from (18) but still noted *DualBnd*, and the simulated performance gap of the policies just described relative to that bound.

We make the following observations about the results shown in Table 5:

1. Policy *CG-SC* is near-optimal, and both its average and worst-case performance are superior to that of all other policies considered.

2. The fact that the policy *CG-SC*$_\omega$ is dominated by our proposed policy *CG-SC* confirms that switching costs should not be attributed exclusively to the

Table 5 Performance Benchmark in Environments with a Switching Cost

$\mathbb{V}[\gamma_s]$	T	ω	*GDY* (%)	*GDY-SC* (%)	*CG-SC*$_\omega$ (%)	$CG^{\text{1-sla}}_{\text{norm}}$ (%)	*CG-SC* (%)	*DualBnd*
5	20	20	0.60	**0.31**	4.00	1.24	0.60	2,590.55
	40	20	2.81	2.83	6.37	2.66	**2.33**	2,752.98
5	20	50	1.79	**0.10**	7.22	4.57	1.32	2,478.31
	40	50	2.84	**2.43**	12.47	4.66	3.18	2,674.06
50	20	20	7.11	7.24	1.56	0.74	**0.40**	3,946.96
	40	20	14.44	14.49	4.71	**3.30**	3.35	4,519.96
50	20	50	4.85	5.43	6.58	1.06	**0.17**	3,716.61
	40	50	11.99	12.19	9.96	2.53	**2.47**	4,319.20
100	20	20	12.10	12.15	2.75	2.18	**2.05**	4,926.07
	40	20	19.41	19.27	4.06	**3.23**	3.27	5,699.98
100	20	50	11.04	11.64	7.05	3.19	**2.81**	4,714.05
	40	50	18.49	18.66	8.48	3.69	**3.44**	5,550.45
Average			8.96	8.89	6.27	2.75	**2.12**	
Max.			19.41	19.27	12.47	4.66	**3.44**	

Figure 2 Substitution Structures Considered

One-item substitution matrix

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

Random substitution matrix

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

Adjacent substitution matrix

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

period in which the product is introduced, as discussed in §3.2.

3. When the initial variance is small ($\mathbb{V}[\gamma_s] = 5$), the greedy policy with amortized switching costs (*GDY-SC*) performs slightly better than *CG-SC*, particularly when switching costs are high ($\omega = 50$). Our interpretation is that, in these scenarios, the underestimation of the switching costs associated with the amortized term $\omega(1 - v^{t+1})/t$ in (19) comes into play (see the discussion at the end of §3.2). In this situation, ignoring the (relatively small) future benefits from learning (second term of (19)), as in the greedy index formula characterizing *GDY-SC*, provides an appropriate compensation for that bias.

4. When the initial variance is high ($\mathbb{V}[\gamma_s] \in \{50, 100\}$), the horizon length is large ($T = 40$) and the switching costs are low ($\omega = 20$), the original index policy $CG^{\text{1-sla}}_{\text{norm}}$, which ignores switching costs, performs slightly better than the modified one *CG-SC*. The significant relative superiority of $CG^{\text{regr}}_{\text{norm}}$ over $CG^{\text{1-sla}}_{\text{norm}}$ seen in Table 3 for the corresponding environments ($\mathbb{V}[\gamma_s] \in \{50, 100\}$, $T = 40$) shows that the underestimation of the learning benefits associated with the term z_t used by both $CG^{\text{1-sla}}_{\text{norm}}$ and *CG-SC* is particularly significant then. An even more drastic compensation is then needed to correct that bias, which in the case of $CG^{\text{1-sla}}_{\text{norm}}$ is obtained by ignoring the switching costs altogether. Note, however, that even in those cases the incremental improvement of the original index policy $CG^{\text{1-sla}}_{\text{norm}}$ on the modified *CG-SC* is rather small.

4.5. Substitution Effects

4.5.1. Experiments Description. For these experiments, we considered a data set with $T = 24$, $S = 144$, $N = 6$, unit profits r_s were generated upfront through S independent draws from distribution $U[2, 6]$ as before, and products were indexed from 1 to S by decreasing unit profit values. We used the frequentist approach as described in §4.1, and used a single set of real underlying rates γ_s for the original demand that were independently drawn upfront from a Gamma distribution with parameters $(1, 1/10)$. The actual sales resulting from this original demand and the assortments implemented were simulated according to the substitution structure defined in §3.3. More specifically, following Smith and Agrawal (2000) we assumed that $L_i = L$ for all i and considered the following three scenarios:

One-item substitution. A particular product ($s = 80$ in our experiments) is the substitute by default for all other products, assuming substitution does occur.

Random substitution. When a first original product choice is not available, all other products may serve as substitutes with equal probability.

Adjacent substitution. Customers consider as possible substitutes, with equal probability, the two products that have the closest unit prices relative to their original choice. The two products with the lowest and highest unit prices have only one possible substitute.[7]

Figure 2 shows the corresponding substitution matrices $(q_{is})_{(i,s)\in\mathbb{S}^2}$, which are completely characterized by the value of the total substitution probability L. Finally, we considered identical initial priors characterized by $(\mathbb{E}[\gamma_s], \mathbb{V}[\gamma_s]) = (10, 100)$, or equivalently, $(m_s, \alpha_s) = (1, 1/10)$.

In addition to a basic performance evaluation, we were interested in assessing the relative impact of the new prior update rule (20) and the QIP (21) described in §3.3. In addition to policy $CG^{\text{1-sla}}_{\text{norm}}$ (which essentially ignores substitution effects), we thus simulated the following policies:

CG-SUB. Our proposed policy here, obtained by modifying $CG^{\text{1-sla}}_{\text{norm}}$ as described in §3.3.

CG-SUB$_{\text{NU}}$ (respectively, *CG-SUB*$_{\text{QIP}}$). The policy obtained by applying only to $CG^{\text{1-sla}}_{\text{norm}}$ the first (respectively, second) modification described in §3.3. That is, *CG-SUB*$_{\text{NU}}$ uses the new update rule (20) but does not otherwise account for substitutions when comparing products, whereas *CG-SUB*$_{\text{QIP}}$ uses the original

[7] We assume here for ease of interpretation that all products have the same cost, so that unit profits correspond to unit prices. More generally, this substitution structure may be considered for any other attribute defining some "proximity" among products.

update rule (2), but accounts for substitutions by solving the QIP (21).

GDY-SUB. Same policy as *CG-SUB*, except that the linear term in the objective function of the QIP (21) is given by the immediate expected return $r_s\mathbb{E}[\gamma_s]$ instead of the desirability index $\eta^{CG}_{t,s}$.

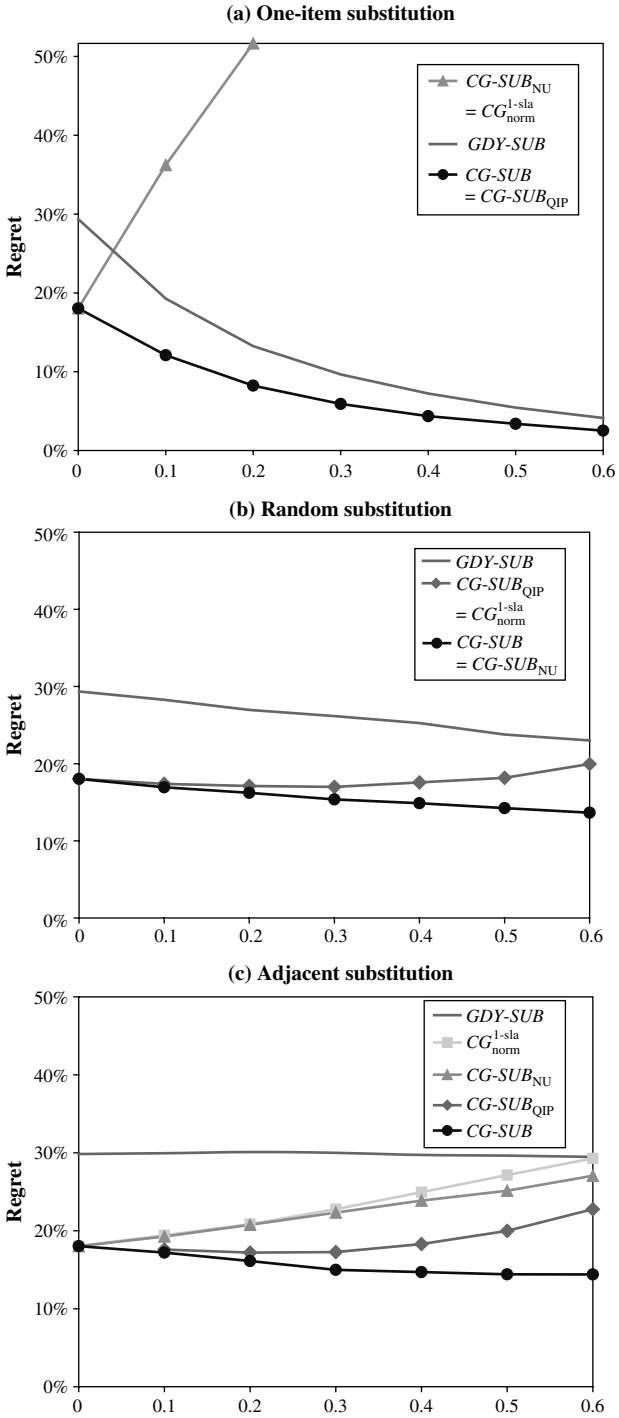
4.5.2. Results. Figures 3(a)–(c) show the profit per period of the policies just mentioned relative to the full information upper bound obtained by substituting $\eta^{CG}_{t,s}$ with $r_s\gamma_s$ and m_i/α_i with γ_i in (21).

1. In the one-item substitution case (Figure 3(a)), the performances of *CG-SUB* and *CG-SUB*$_{\text{QIP}}$ are identical and very close to optimal (within 5.2% of the full information bound on average for $L > 0.1$). Policy *GDY-SUB* also performs well when the substitution probability is not too low. In contrast, *CG-SUB*$_{\text{NU}}$ and $CG^{\text{1-sla}}_{\text{norm}}$ perform identically and very poorly. These results show that the key performance driver of policies in the one-item substitution case is whether they include in all assortments the product serving as a universal substitute (the QIP supporting *CG-SUB*, *CG-SUB*$_{\text{QIP}}$, and *GDY-SUB* ensures this). Policies *CG-SUB*$_{\text{NU}}$ and $CG^{\text{1-sla}}_{\text{norm}}$ are unlikely to include the universal product because it has the 80th unit price and the assortment size is $N = 6$ by design. The low benefit associated with learning, either active (*CG-SUB*) or passive (*GDY-SUB*), is also evidenced here by the proximity of both policies to the full information upper bound.

2. In the random substitution case (Figure 3(b)), policy *CG-SUB* now performs identically to *CG-SUB*$_{\text{NU}}$ and policy *CG-SUB*$_{\text{QIP}}$ performs identically to $CG^{\text{1-sla}}_{\text{norm}}$, with the former pair of policies clearly outperforming the latter for $L > 0.2$. Contrary to the previous case, the most significant performance driver in this environment is the ability to accurately estimate the original demand rates. This is done effectively by polices *CG-SUB* and *CG-SUB*$_{\text{NU}}$ since they update the state information using the sales estimate $\tilde{n}_s$ given by Equation (20). Note also that in this case active learning is relevant since policy *CG-SUB* outperforms its greedy counterpart *GDY-SUB*.

3. The one-item and random substitution cases represent two extremes among all substitution models. They each necessitate only one of the two policy modifications described in §3.3. In contrast, the adjacent substitution case (Figure 3(c)) seems a more complex and realistic structure. Specifically, both the combinatorial complexity (addressed by solving the QIP (21)) and the learning implications of the substitution model (captured by the new updating rule (20)) are significant, as evidenced by the superiority of *CG-SUB*$_{\text{QIP}}$ and *CG-SUB*$_{\text{NU}}$ over $CG^{\text{1-sla}}_{\text{norm}}$, and that of *CG-SUB* over all other policies for $L \geq 0.3$. The poor performance of *GDY-SUB* indicates that it is also important to consider the implications of assortment decisions on the learning process.

Figure 3 Performance Benchmark with Substitution Probability



5. Conclusion

Our main goal in this paper is to shed some light on the dynamic assortment problem faced by fast fashion retailers. As a first step, we have considered a very stylized version of this problem, amounting to a finite horizon multiarmed bandit model with Bayesian learning. Using DP duality, we were able to

develop both the profit upper bound (7) and the approximate policy (12) for this first version of the problem. Although the basic form $CG^{\text{1-sla}}_{\text{norm}}$ of that policy seems to perform close to other comparable heuristics described in the literature and within 4% of optimality for all the data instances we considered, its improved form $CG^{\text{regr}}_{\text{norm}}$ does exhibit superior performance. Perhaps more important, both versions are closed form and particularly easy to implement and interpret. This last feature enabled us, as a second step, to extend this policy to more realistic versions of our motivating dynamic assortment problem accounting for implementation delays, switching costs, and demand substitution effects. Our numerical experiments showed that the performance of these policy extensions was particularly encouraging and helped uncover some important drivers of that performance. An overarching theme was the high benefit of accounting for the impact of assortment decisions on future learning, i.e., the advantage of active over passive learning. In environments with long assortment implementation lead times, we observed the importance of recognizing the resulting reduced potential for leveraging information, and appropriately predicting the quality of future information. We also observed the importance of appropriately recognizing the benefits in future periods of incurring switching costs in the current one. Finally, we found that the probabilistic structure of demand substitutions impacted the relative importance of addressing the resulting combinatorial complexity and the implications of substitutions on learning.

We believe that while this work clearly demonstrates the potential of using bandit models to address dynamic assortment problems, it only constitutes a building block towards truly operational systems. In particular, the possibility of coordinating assortment decisions across multiple stores facing similar demand patterns appears to be an exciting opportunity, both in practice and as a topic for future research.

Acknowledgments

The authors are most grateful to Yossi Sheffi, Larry Lapide, and the MIT-Zaragoza Logistics Program for funding this project and providing useful feedback. They are also grateful to Yves Dallery and the Laboratoire de Génie Industriel of the École Centrale Paris for hosting the first author during the Summer of 2004, and for providing many helpful comments. The authors thank Paul H. Zipkin (the department editor), the associate editor, and two anonymous referees for many useful suggestions and feedback. They are also grateful to Gabriel Bitran, Steve Graves, and the participants of the Operations Management Seminar at the Sloan School of Management for offering insightful remarks about this work. Finally, the second author is indebted to Martha Nieto for a conversation about Zara that sparked his interest in fast-fashion companies and was key to the genesis of this project.

Appendix

Proof of Proposition 1. From the definition, it is clear that $H^*_t(\mathbf{m},\boldsymbol{\alpha}) \le H^{\boldsymbol{\lambda}_t}_t(\mathbf{m},\boldsymbol{\alpha})$ for any dual policy $\boldsymbol{\lambda}_t$; therefore, we only need to prove the first inequality. We proceed by induction on t. Assume that $J^*_{t-1}(\mathbf{m},\boldsymbol{\alpha}) \le H^*_{t-1}(\mathbf{m},\boldsymbol{\alpha})$ for all states $(\mathbf{m},\boldsymbol{\alpha})$, then for any $\lambda_t \ge 0$:

$$\begin{aligned}
J^*_t(\mathbf{m},\boldsymbol{\alpha}) &= \max_{\substack{\mathbf{u}\in\{0,1\}^S:\\ \sum_{s=1}^S u_s \le N}} \sum_{s=1}^S r_s \frac{m_s}{\alpha_s} u_s + \mathbb{E}_{\mathbf{n}}[J^*_{t-1}(\mathbf{m}+\mathbf{n}\cdot\mathbf{u},\boldsymbol{\alpha}+\mathbf{u})] \\
&\le N\lambda_t + \max_{\substack{\mathbf{u}\in\{0,1\}^S:\\ \sum_{s=1}^S u_s \le N}} \sum_{s=1}^S \left(r_s \frac{m_s}{\alpha_s} - \lambda_t\right) u_s \\
&\quad + \mathbb{E}_{\mathbf{n}}[J^*_{t-1}(\mathbf{m}+\mathbf{n}\cdot\mathbf{u},\boldsymbol{\alpha}+\mathbf{u})] \\
&\le N\lambda_t + \max_{\mathbf{u}\in\{0,1\}^S} \sum_{s=1}^S \left(r_s \frac{m_s}{\alpha_s} - \lambda_t\right) u_s \\
&\quad + \mathbb{E}_{\mathbf{n}}[J^*_{t-1}(\mathbf{m}+\mathbf{n}\cdot\mathbf{u},\alpha+\mathbf{u})] \\
&\le N\lambda_t + \max_{\mathbf{u}\in\{0,1\}^S} \sum_{s=1}^S \left(r_s \frac{m_s}{\alpha_s} - \lambda_t\right) u_s \\
&\quad + \mathbb{E}_{\mathbf{n}}[H^*_{t-1}(\mathbf{m}+\mathbf{n}\cdot\mathbf{u},\boldsymbol{\alpha}+\mathbf{u})]. \qquad (25)
\end{aligned}$$

The first inequality follows from the fact that $\lambda_t \ge 0$, and the second holds because the feasible set is larger. The third inequality relies on the induction hypothesis. Considering now the minimum of the r.h.s. of (25) yields the desired result. □

Proof of Lemma 1. We proceed by induction. Consider $t \ge 1$ and assume that the property holds for $t-1$. Then we have that:

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

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 equation rearranges terms in order to obtain the desired result. □

References

Adelman, D., A. J. Mersereau. 2004. Relaxations of weakly coupled stochastic dynamic programs. Working paper, Graduate School of Business, The University of Chicago, Chicago, IL.

Agrawal, R., M. V. Hedge, D. Teneketzis. 1988. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Trans. Automatic Control* **33**(10) 899–906.

Anantharam, V., P. Varaiya, J. Walrand. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.i.d. rewards. *IEEE Trans. Automatic Control* **32**(11) 968–976.

Aviv, Y., A. Pazgal. 2002. Pricing of short life-cycle products through active learning. Working paper, Washington University, St. Louis, MO.

Ben-Akiva, M., S. R. Lerman. 1985. *Discrete Choice Analysis*. MIT Press, Cambridge, MA.

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

Bertsekas, D. 1999. *Nonlinear Programming*. Athena Scientific, Belmont, MA.

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

Bertsimas, D., A. Mersereau. 2004. A learning approach to customized marketing. Working paper, The University of Chicago, Chicago, IL.

Brezzi, M., T. L. Lai. 2002. Optimal learning and experimentation in bandit problems. *J. Econom. Dynam. Control* **27** 87–108.

Bultez, A., P. Naert. 1988. SHARP: Shelf allocation for retailers profit. *Marketing Sci.* **7** 211–231.

Caro, F. 2005. Dynamic retail assortment models with demand learning for seasonal consumer goods. Ph.D. thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Castañon, D. A. 1997. Approximate dynamic programming for sensor management. *Proc. 36th IEEE Conf. Decision and Control*, San Diego, CA, 1202–1207.

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

Fisher, M. L., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(1) 87–99.

Fisher, M. L., A. Raman, A. S. McClelland. 2000. Rocket science retailing is almost here—Are you ready. *Harvard Bus. Rev.* (July–August) 115–124.

Ghemawat, P., J. L. Nueno. 2003. ZARA: Fast fashion. Harvard Business School Multimedia Case 9-703-416, Harvard University, Boston, MA.

Ginebra, J., M. K. Clayton. 1995. Response surface bandits. *J. Roy. Statist. Soc. Series B* **57** 771–784.

Gittins, J. C. 1979. Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Series B* **14** 148–167.

Hardwick, J., R. Oehmke, Q. F. Stout. 2006. New adaptive designs for delayed response models. *J. Sequential Planning Inference* **136** 1940–1955.

Hawkins, J. T. 2003. A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications. Ph.D. thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Karmarkar, U. S. 1987. The multilocation multiperiod inventory problem: Bounds and approximations. *Management Sci.* **33**(1) 86–94.

Kök, A. G., M. L. Fisher. 2004. Demand estimation and assortment optimization under substitution: Methodology and application. Working paper, Duke University, Durham, NC.

Lagarias, J. C., J. A. Reeds, M. H. Wright, P. E. Wright. 1998. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optim.* **9**(1) 112–147.

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

Presman, E. L., I. N. Sonin. 1990. *Sequential Control with Incomplete Information: The Bayesian Approach to Multi-Armed Bandit Problems*. Academic Press, San Diego, CA.

Smith, S. A., N. Agrawal. 2000. Management of multi-item retail inventory systems with demand substitution. *Oper. Res.* **48** 50–64.

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

Weber, R. R., G. Weiss. 1990. On an index policy for restless bandits. *J. Appl. Probab.* **27** 637–648.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. J. Gani, ed. *A Celebration of Applied Probability. J. Appl. Probab.* **25A** 287–298.