OPERATIONS RESEARCH
Vol. 57, No. 2, March–April 2009, pp. 327–341
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5702 | 0327

DOI 10.1287/opre.1080.0577
© 2009 INFORMS

Dynamic Pricing with Online Learning and Strategic Consumers: An Application of the Aggregating Algorithm

Tatsiana Levina, Yuri Levin, Jeff McGill, Mikhail Nediak
School of Business, Queen's University, Kingston, Ontario, Canada K7L 3N6 {tlevin@business.queensu.ca, ylevin@business.queensu.ca, jmcgill@business.queensu.ca, mnediak@business.queensu.ca}

We study the problem faced by a monopolistic company that is dynamically pricing a perishable product or service and simultaneously learning the demand characteristics of its customers. In the learning procedure, the company observes the sales history over consecutive learning stages and predicts consumer demand by applying an *aggregating algorithm* (AA) to a pool of online stochastic predictors. Numerical implementation uses finite-sample distribution approximations that are periodically updated using the most recent sales data. These are subsequently altered with a random step characterizing the stochastic predictors. The company's pricing policy is optimized with a simulation-based procedure integrated with AA. The methodology of the paper is general and independent of specific distributional assumptions. We illustrate this procedure on a demand model for a market in which customers are aware that pricing is dynamic, may time their purchases strategically, and compete for a limited product supply. We derive the form of this demand model using a game-theoretic consumer choice model and study its structural properties. Numerical experiments demonstrate that the learning procedure is robust to deviations of the actual market from the model of the market used in learning.

Subject classifications: marketing: pricing; games: stochastic; artificial intelligence: online learning.
Area of review: Revenue Management.
History: Received March 2006; revisions received September 2006, April 2007, August 2007, December 2007; accepted January 2008. Published online in *Articles in Advance* January 21, 2009.

1. Introduction

One of the fundamental challenges faced by any company is that of assessing customer response to changes in the price of the products or services it sells, and this task is particularly important for organizations that are experimenting with controlled dynamic pricing. Fortunately, frequent price changes also produce frequent opportunities for measurement of customer responses and, in principle, the possibility of obtaining near real-time estimates for customer demand models. In this paper, we develop an approach to this type of *online learning* of customer behavior; that is, learning that takes place as sales unfold. The approach works with a discrete-time approximation to the sales process and can be applied to learning the parameters of any demand model that produces estimates of consumer purchase probability at each time step of the approximation. An important aspect of the method is that learning is integrated with pricing so that pricing policy formation and consumer demand prediction proceed concurrently.

Our approach is based on the *Aggregating Algorithm* (AA)—a particularly general method for online learning developed by Vovk (1990). To illustrate its generality, we apply it to a model of consumer behavior that allows for *strategic consumers* who know that pricing is dynamic and may delay their purchases to times of anticipated lower price. The possibility of such strategic behavior has been of increasing interest recently because of the rapid growth in information available to customers through Internet sales channels and "price-shopper" websites. In many cases, customers are able to monitor both prices and availabilities of products over time and may develop accurate guesses about a company's future prices. Ironically, companies that employ carefully controlled dynamic pricing may be more vulnerable to strategic consumer behavior than companies employing ad hoc price adjustments because controlled dynamic pricing can lead to pricing policies with regular features; for example, monotone-decreasing prices in situations where "price skimming" appears to be optimal (see Besanko and Winston 1990).

Dynamic pricing is properly viewed as one approach to the general problem of revenue management, and there is now an extensive literature on revenue management and related practices. For surveys, see Bitran and Caldentey (2003), Elmaghraby and Keskinocak (2003), and McGill and van Ryzin (1999). Broad discussions of revenue management can be found in recent books by Talluri and van Ryzin (2004) and Phillips (2005). Many revenue management applications depend on forecasts of consumer behavior that are generated from stochastic models of demand, for example, demand as a function of time and price. Unfortunately, such stochastic demand-response models typically assume characteristics of demand that cannot be known precisely in practice. This uncertainty in demand characteristics has long been recognized in economics, marketing, pricing, inventory management, and revenue management, and there have been efforts to develop methods for learning of demand-response functions over time. For example, Balvers and Cosimano (1990) study a pricing problem with learning of demand that is a linear function of price. The model does not consider any limits on sales due to inventory levels. Carvalho and Puterman (2005) also study learning and pricing when capacity is unlimited. The authors consider a finite time horizon and focus on specific parametric forms of the customer arrival distribution and the probability of sale (both the number of arrivals and the actual sales are observable). The parameters are assumed to be fixed and unknown. The authors explore a trade-off between learning and pricing using a "one step look ahead" heuristic based on a two-period version of the problem. Petruzzi and Dada (2002) consider a stocking and pricing model with a fixed but unknown perturbation of some given demand function. Papers by Bertsimas and Perakis (2006), Aviv and Pazgal (2005a), and Lin (2006) study the pricing of a fixed stock of items over a finite horizon with demand learning. Bertsimas and Perakis (2006) consider learning of all demand characteristics, including the price sensitivity, but assume a linear demand model with normal perturbations. The other two papers assume a known reservation price distribution. Aviv and Pazgal (2005b) present a general framework for dynamic pricing when stochastic properties of demand are affected by the current *state of the world*. The number of possible states considered by the authors is finite. They use partially observed Markov decision processes as a modeling basis and information-structure modification heuristics to provide a tractable implementation. A recent work by Besbes and Zeevi (2007) considers a joint learning and pricing method for a network revenue management problem involving multiple products utilizing multiple resources. The approach assumes a Poisson model with demand rates determined by unknown functions of price. While the model of demand in their paper is nonparametric, the authors simplify the problem by only considering demand rates that do not depend explicitly on time (in contrast to this paper). The policies considered involve a "brief" period of learning (experimentation with prices selected from a grid of prices) followed by static pricing. The authors establish asymptotic optimality of the policy given that the resource capacities and the demand rates simultaneously tend to infinity.

All the learning-focused papers cited above consider restricted forms of demand models that do not consider potentially complex consumer behavior. The prior work on dynamic pricing, which assumes *known* demand models, has allowed for varying degrees of consumer sophistication. For example, the classical model by Gallego and van Ryzin (1994) assumes *myopic* consumers who make a purchase as soon as the price is below their valuation for the product, whereas other models allow for strategic consumers who may benefit by delaying their purchase decisions (see Besanko and Winston 1990, Elmaghraby et al. 2008, Aviv and Pazgal 2008, Liu and van Ryzin 2008, Su 2007, and Levin et al. 2005). In the case of strategic consumers, the demand model should also capture *competition* between the customers if the product supply is limited. Although the case of myopic consumers is amenable to existing learning approaches, it is difficult to extend these approaches to the instances of more complex consumer behavior, in particular, strategic behavior. Indeed, one of the typical approaches in dynamic pricing with demand uncertainty (with or without learning) is policy optimization by dynamic programming techniques, but the complexity of demand learning with strategic consumers renders an exact dynamic programming approach computationally intractable.

The main contribution of this paper is the presentation of an integrated procedure to both determine prices and estimate customer behavior under general parametric uncertainty. We accomplish this with an adaptation of the *Aggregating Algorithm* (AA) of Vovk (1999). The AA methodology belongs to the class of *online methods* and was originally developed to address the problem of combining expert advice (Vovk 1999). Similar techniques have been applied to the problem of online portfolio selection since the work of Cover (1991).

In our online approach, the company observes the sales history over consecutive learning stages and predicts future demand by applying the AA to a pool of stochastic predictors. Numerical implementation uses finite-sample approximations to the pool of predictors. These are periodically updated using the most recent sales data and are subsequently altered by a random step that maintains diversity of the predictors. This is similar to a method applied to online portfolio selection by Levina (2004). The company's pricing policy is optimized by a simulation-based method that is integrated with AA.

We illustrate the versatility of this integrated procedure on a demand model for a complex market in which customers are aware that pricing is dynamic, may time their purchases strategically, and are competing for a limited product supply. The model of consumer demand used in this illustration is adapted from a game-theoretic, strategic consumer-choice model described in Levin et al. (2005). In that model, a fully rational consumer's decision is characterized as a probability of purchase at each time and state of the sales process. Summation of these probabilities across consumers defines the demand model. A key departure from that model in this paper is that we assume limited rationality of consumers with respect to anticipated future prices.

A number of structural properties of the revised consumer-choice model are relevant to the implementation of online learning described here. In particular, we show that strategic consumer response to price is inherently dependent on time and the remaining capacity of the firm. We also show that this model supports an intuitively appealing decision rule for strategic consumers—that they will attempt to purchase when their consumer surplus from an immediate purchase is greater than the discounted *expected* surplus from all future purchasing opportunities. The expected surplus is thus identified as a key component of strategic consumer behavior. We then derive important properties for the expected surplus that can be used to construct an *empirical* consumer demand model. (Such an empirical approximation is needed because exact computation of the expected surplus, when combined with online learning, is not practical in problems of realistic size.) Numerical experiments demonstrate that the learning procedure is robust to discrepancies between the detailed strategic market response and the model of that response used in learning.

This paper is organized as follows. In §2, we discuss a general class of time, inventory, and price-dependent demand models that includes the case of strategic consumers. In §2.2, we outline a simple Bayesian approach to online learning of the parameters of the general demand model for any pricing policy. In §2.3, we discuss a specialization of the AA that implements demand learning using a general Bayesian approach with finite-sample approximations. Pricing policy optimization is addressed in subsequent sections. In §3, we identify restricted pricing policy classes that are practical to implement and facilitate dynamic pricing with demand learning. In §4, we show how learning can be integrated with optimization of the pricing policy through an online procedure that utilizes the AA for learning and simulation-based optimization for pricing. In §5, we discuss the application of this procedure to the case of strategic consumers and our numerical experience with it. We summarize the main contributions of the paper in §6. In the online appendix to this paper, we provide an outline of the game-theoretic strategic consumer-choice model and its detailed analysis. An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

2. Demand Model and Its Learning

2.1. Sales Process

Consider a product with limited availability sold by a monopolistic company over several planning horizons, each comprising T decision periods, $\{0, 1, \ldots, T-1\}$. At time 0 in each planning horizon, the initial inventory of the product is Y, with no replenishment possible during the planning horizon. All planning horizons start with the same initial inventory. At time T, the product expires and all unsold items are lost. We assume that the company wishes to recompute its pricing decisions after each sale or offer of sale, and that the sales process unfolds in an "orderly" fashion; that is, the company presents items for sale sequentially, one per time period. In each period t, a single sale may or may not occur. This assumption of at most one sale per decision period is a discrete-time approximation to the continuous-time Poisson demand model frequently assumed in revenue management literature—the probability of more than one sale in a time period becomes negligible if we consider sufficiently short time intervals. If the unit of time measurement equals the length of a single decision period, then the probability of one sale in each decision period corresponds to the demand intensity for the entire market in this decision period. The *sale probability* depends on a number of quantities, some of which may be unknown to the company, and its exact functional form is determined by a model of consumer behavior. In this paper, we consider demand models in which the sale probability is a function of time t, remaining inventory level y, and the current price p.

The consumer behavior model is specified by a constant parameter vector $\mathbf{x}$ that may include both known and unknown components. The existence of a parameter vector that specifies the model is implicit whenever a specific instance of a general demand model is described. For example, in the context of the classical dynamic pricing model of Gallego and van Ryzin (1994), $\mathbf{x}$ would include the parameters of a model of customer arrival intensity as a function of time and parameters of the reservation price distribution. In this paper, we expand this set to include other parameters describing consumer behavior. The only restriction is that the set of parameters is finite.

The resulting sales process is, then, a discrete-time counting (Bernoulli) process with the sale probability given by a known function $\Lambda^{\mathbf{x}}(t, y, p)$ of $t \in \{0, 1, \ldots, T-1\}$, $y \in \{1, \ldots, Y\}$, and $p \in \Pi$, where Π is a set of admissible prices. These probabilities define the demand model.

Example 1 (Myopic Consumers). This example outlines one particular set of assumptions that leads to a specific form of the sale probability $\Lambda^{\mathbf{x}}(t, y, p)$. We assume that: (1) the consumer population is homogeneous and finite; (2) each customer will purchase at most one item; and (3) all customers are present in the market from the beginning of the planning horizon and remain present until they make a purchase, the company runs out of inventory, or the planning horizon ends. For the purposes of this example, suppose also that customers are myopic and have uncertain valuations for the product at time t given by the random variable $B(t)$. The customers cannot control the precise timing of their purchases, but can adjust the intensity of their efforts to acquire an item (*shopping intensity* in the terminology of Levin et al. 2005). This shopping intensity is proportional to the probability that the valuation $B(t)$ exceeds the current price p. We assume that the demand intensity for the entire market is the sum of shopping intensities of individual customers. Let N be the initial market size (at the beginning of the planning horizon) and $\bar{\lambda}$ be the maximum shopping intensity of each myopic customer. Following the model developed in Levin et al.

(2005), the demand intensity for this entire market (the sale probability) will be given by

$$\Lambda^{\mathbf{x}}(t, y, p) = \bar{\lambda}(N - (Y - y))P^{\mathbf{x}}(B(t) \geqslant p) \qquad (1)$$

when there are y items left for sale (implying that there are $N - (Y - y)$ customers left in the market). The vector $\mathbf{x}$ in this case includes $\bar{\lambda}$ and parameters of the distribution of $B(t)$ and its evolution over time. In §5, we present an intuitively reasonable generalization of the above expression to the case of strategic consumers. The parameter vector is then extended to include parameters describing strategic behavior. In the next section, we show how a demand model like (1) can be used to learn the values of the elements of $\mathbf{x}$.

2.2. Learning from Sales

We now assume that the company has a sale probability model $\Lambda^{\mathbf{x}}(t, y, p)$ that describes collective consumer behavior, and focus on the problem of learning the parameters $\mathbf{x}$ under any pricing policy as a necessary first step toward optimizing the pricing policy. We consider demand learning on the basis of sales only, although exogenous information such as the results of consumer surveys can be introduced as part of a prior distribution for $\mathbf{x}$. If $\mathbf{x}$ is known exactly, the company can predict customer response to a pricing policy because the probability of sale $\Lambda^{\mathbf{x}}(t, y, p)$ then becomes a function of t, y, and p only. A key challenge in this learning process is that the demand is state and time dependent—a condition that rules out use of many conventional learning methods.

Assume that initial knowledge about the parameter vector $\mathbf{x}$ is contained in a prior distribution, which is continuous and specified by a given density function. As additional sales information becomes available, the distribution reflecting the company's knowledge about $\mathbf{x}$ is updated to a posterior distribution, which is also continuous. At the beginning of decision period t, complete histories of the sales and price processes from previous decision periods are available. We define:

- $\mathcal{N}_t$ as the set of decision periods during which a sale was made, and
- $\mathcal{P}_t = \{p_0, \dots, p_{t-1}\}$ as the list of all prices used previously.

At time t, the list $\mathcal{P}_t$ has length t, and the set $\mathcal{N}_t$ has cardinality $Y - y$, where y is the current level of inventory. We denote the random vector distributed according to the prior density as $\mathbf{x}(\varnothing, \varnothing)$. The parameter vector distributed according to the posterior density at time t corresponding to histories $\mathcal{N}_t, \mathcal{P}_t$ is $\mathbf{x}(\mathcal{N}_t, \mathcal{P}_t)$. The posterior density at $\mathbf{x}$ is obtained (up to a normalizing constant) by multiplying the prior density at $\mathbf{x}$ and the likelihood of the observed sales history for a particular parameter value $\mathbf{x}$. In the likelihood expression, we use the auxiliary notation:

- $y_\tau = Y - |\mathcal{N}_\tau|$ is the number of units left at the beginning of decision period τ (note that for any $t \geqslant \tau$, $y_\tau = Y - |\mathcal{N}_t \cap [0, \tau - 1]|$), and
- $\bar{\mathcal{N}}_t = \{0, \dots, t-1\} \setminus \mathcal{N}_t$ is the set of decision periods during which no sale was made.

The likelihood function at the beginning of decision period t is given by

$$L(\mathcal{N}_t, \mathcal{P}_t \mid \mathbf{x}) = \prod_{\tau \in \mathcal{N}_t} \Lambda^{\mathbf{x}}(\tau, y_\tau, p_\tau) \prod_{\tau \in \bar{\mathcal{N}}_t} (1 - \Lambda^{\mathbf{x}}(\tau, y_\tau, p_\tau)). \qquad (2)$$

Based on the observed histories and using the posterior distribution, the company can estimate the probability of future sales as a function of price paths. For example, the probability of a sale occurring in decision period t is

$$d(\mathcal{N}_t, \mathcal{P}_t, p) = E_{\mathbf{x}(\mathcal{N}_t, \mathcal{P}_t)}[\Lambda^{\mathbf{x}(\mathcal{N}_t, \mathcal{P}_t)}(t, Y - |\mathcal{N}_t|, p)], \qquad (3)$$

where p is the price set by the company.

2.3. Aggregating Algorithm for Demand Learning

The previous section shows that, given an appropriate general demand model, predicting consumer response is reduced to estimating the parameter vector $\mathbf{x}$ that uniquely specifies this model. The uncertainty in prediction and the current state of knowledge of $\mathbf{x}$ is represented by a posterior distribution for $\mathbf{x}$, and learning the parameters consists in updating this posterior distribution as sales and price information becomes available. We refer to the time between such updates as a *learning stage*. In the absence of specific distributional assumptions, learning through updating the posterior distribution is handled numerically using finite-sample approximations.

To contain the growth in computational intensity of posterior distribution updates as the volume of collected data grows, we use our experience with the AA of Vovk (1999) (see also Levina 2006). Here, we give a brief overview of the AA to establish basic terminology. The methodology of AA hinges upon the notion of an *elementary predictor*—any method that produces a prediction in every learning stage. The concept of an elementary predictor is very general; for example, it can include human "expert" opinion, a conventional forecasting technique like regression, or the realizations of an arbitrary stochastic process (with the value in each learning stage being the prediction). The key is that elementary predictors produce sequences of predictions over time. A *pool* Θ of predictors is a given collection of such methods, and AA is a general approach for aggregating predictions, $\boldsymbol{\xi}_s(\theta)$, $\theta \in \Theta$ from a pool into a single prediction in a given learning stage s. Each elementary predictor in the pool has a *weight* that reflects its past performance. The weights at the beginning of stage s are represented, in general, by a measure $P_{s-1}(d\theta)$ on Θ. The weights induce a probability distribution on the pool, which is obtained by normalizing the weights to sum to one: $P_{s-1}(d\theta)/P_{s-1}(\Theta)$. After stage s predictions are made, the outcome of *reality*, γ_s, is observed, and each predictor's weight is updated by multiplying it by a nonnegative factor reflecting its current performance:

$$P_s(d\theta) = e^{-\lambda_s(\boldsymbol{\xi}_s(\theta), \gamma_s)} P_{s-1}(d\theta),$$

where $\lambda_s(\mathbf{x}, \gamma_s)$ is a *loss* suffered in stage s if prediction $\mathbf{x}$ is made and the outcome of reality is γ_s. The aggregation into a single prediction in stage s is often accomplished with a weighted average of predictions using the current weights:

$$\frac{\int_\Theta \boldsymbol{\xi}_s(\theta) P_{s-1}(d\theta)}{\int_\Theta P_{s-1}(d\theta)},$$

but other aggregation methods are possible. In a practical application of AA, the following components have to be defined: the predictor pool Θ, the loss function, a numerical representation of the current predictor weights $P_{s-1}(d\theta)$ and a procedure for their update, the initial weights $P_0(d\theta)$, and, finally, the aggregation method. In the rest of this section, we describe the first three components.

In our problem, we identify predictions with possible values of the parameter vector $\mathbf{x}$ and use elementary predictors that, in every learning stage, *alter* their predictions according to the realizations of a given discrete-time Markov process on the space of parameters. (In our implementation, we draw the next realization of the process from the mixture of two distributions: a step of Gaussian random walk away from the current parameter values, and a prior distribution. However, other Markov processes could be used.) Elementary predictors of this general type, sometimes called Markov switching, have been used in other problems to expand the pool of predictors. For example, Levina (2004) describes their application in online portfolio selection. At any point in the learning process, such a pool Θ is infinite because its elements are possible future realizations of the predictor process. The Markovian structure of predictors ensures that the current distribution of predictions completely determines any future distribution of predictions by the pool.

For each prediction $\mathbf{x}$ given by the pool, we determine the corresponding likelihood of the sales data accumulated since the last update. We use that likelihood as a factor to update the weight of any predictor that makes prediction $\mathbf{x}$. That is, our loss function is equal to the negative log-likelihood of the sales data obtained during the current learning stage. This choice of update, together with the Markovian structure, implies that the combined weight of all elementary predictors producing prediction $\mathbf{x}$ is an approximation to the posterior density at $\mathbf{x}$ corresponding to observed sales data. The accuracy of this approximation is affected by the choice of the predictor process. If alterations in predictions from one learning stage to the next are, on average, small (the magnitude of the random walk step is small), then the level of "noise" introduced into the approximation is also small. On the other hand, the magnitude of the step should be sufficiently large to ensure diversification of learning and faster exploration of the parameter space.

Because the pool of predictors is infinite, we maintain their weights in the form of a finite-sample approximation to the distribution of predictions. Specifically, the proportion of a finite sample of predictions that fall in a particular area of the parameter space approximates the combined weight of all elementary predictors that deliver predictions in that area. An accept-reject bootstrap-resampling procedure (described in the online appendix) approximates the weight update by ensuring that prediction sample points with greater likelihood are sampled with proportionally higher probability. The resulting updated sample represents new weights of predictors in the pool. Of course, the resulting sample will have many duplicate points because of the bootstrap feature. However, even if two predictors provide the same parameter vector at the end of the current learning stage, the probability that their predictions will be identical in the next stage is zero because predictors correspond to realizations of Gaussian random walk.

Our choice of initial weights $P_0(d\theta)$, which correspond to a prior distribution, is uniform on a rectangular set in the parameter space.

The final component, aggregation, does not have to provide a single prediction of the parameter vector in our case. On the contrary, the uncertainty in current predictions provides valuable information for pricing because we can make the pricing policy more robust by taking this uncertainty into account. Thus, in our application, predictions are aggregated by passing individual sample elements to the policy optimization procedure. Before providing the statement of the integrated learning-policy optimization algorithm, we discuss policy class restrictions.

3. Implementable Pricing Strategies: Policy Class Restriction

The company's objective is to maximize its expected revenues by selecting an optimal pricing policy from an appropriate class. If $\mathbf{x}$ is known, the probability of sale $\Lambda^{\mathbf{x}}(t, y, p)$ is a function of t, y, and p only. Then, the state of the system at each time t can be described by the current inventory level y, and one can solve the problem using dynamic programming with pricing policies of the form $p = p(t, y)$. If even a single component of the parameter vector $\mathbf{x}$ is unknown, a state description of the form (t, y) is inadequate. Indeed, any control problem with unknown parameters belongs to a class of problems with imperfect information, well known for their difficulty. In the present case, the difficulty arises because not only the time t and the current inventory level $Y - |\mathcal{N}_t|$, but the entire history, affect the probability of a sale given in (3). Consequently, the optimal price will also depend on the histories; that is, $p = p(\mathcal{N}_t, \mathcal{P}_t)$. Obtaining an optimal policy of this form, even for a moderately sized problem, is computationally intractable (at least in the general case) due to the size of the state space.

This motivates us to consider policies in the class $p = p(t, y)$, while assuming that the company would like to maximize total expected revenues until the end of each planning horizon. Expectations must be taken both over all possible sales paths and over the most recent parameter distribution update. Policies of the form $p = p(t, y)$ have been common in the dynamic pricing literature since Gallego and van Ryzin (1994). Note that under this policy structure, the price path $\mathscr{P}_t$ can always be computed from the sales path $\mathscr{N}_t$ because

$$p_t = p(t, Y - |\mathscr{N}_t|). \tag{4}$$

Denote the price path corresponding to the sales history $\mathscr{N}_t$ and policy $p(\cdot)$ as $\mathscr{P}_t(p(\cdot), \mathscr{N}_t)$, and the corresponding revenues as

$$R(p(\cdot), \mathscr{N}_t) = \sum_{\tau \in \mathscr{N}_t} p(\tau, Y - |\mathscr{N}_t \cap [0, \tau - 1]|).$$

Also, let $\mathfrak{N}^\Delta$ denote the set of all sales histories $\mathscr{N}_t$ terminating in either a sold-out condition $|\mathscr{N}_t| = Y$ or the end-of-horizon condition $t = T$. Then, maximization of total expected revenues can be expressed as

$$\max_{p(\cdot)} E_{\mathbf{x}(\varnothing,\varnothing)}\left[\sum_{\mathscr{N}_t \in \mathfrak{N}^\Delta} R(p(\cdot), \mathscr{N}_t) L(\mathscr{N}_t, \mathscr{P}_t(p(\cdot), \mathscr{N}_t) \,|\, \mathbf{x}(\varnothing, \varnothing))\right]. \tag{5}$$

The inner sum in (5) corresponds to the expectation over all possible sample paths conditional on the value of the parameter vector. The exact (either numerical or analytic) computation of this objective function is difficult in most situations. We remark, however, that such an objective function is amenable to an approximate computation through simulation. Moreover, for each fixed pricing policy, the simulation procedure is straightforward and can be accomplished by drawing a sufficiently large number of samples as follows:

(i) simulate a parameter vector from $\mathbf{x}(\varnothing, \varnothing)$, and

(ii) conditional on the value of the sampled parameter vector, simulate a sales path using $\Lambda^{\mathbf{x}}(t, y, p)$ as probabilities of sale.

The objective function is then computed as the sample average of revenues corresponding to all simulated sample paths. The average over repetitions of Step (i) corresponds to taking the outer expectation in (5), and the average over step (ii) evaluates the inner sum in (5).

We seek to determine the policy of the firm that maximizes (5). However, we have to restrict our search to classes of policies that can be described by a small number of continuous variables. We do so in order to use the DFO algorithm (see Conn et al. 1997), a general derivative-free optimization method that can handle "noisy" objective functions. Such a method is required because the objective function is computed by simulation, and its derivatives are not readily available. Subclasses of policies that can be described by only a few variables are, for example:

(1) an *open-loop policy*, for which price depends on time only and remains fixed on each of m prespecified partitions of $\{0, 1, \ldots, T-1\}$ (this policy is described by m variables);

(2) an *open-loop, single-threshold policy*, for which the price on each of m partitions depends on whether the inventory y exceeds a fixed threshold y^* (described by $2m$ variables);

(3) a *single-threshold linear policy*, whose variables are coefficients of two linear functions of time $p_1(t) = v_1 + w_1 t$ and $p_2(t) = v_2 + w_2 t$, such that the price is $p_1(t)$ if $y \geqslant y^*$ and $p_2(t)$, otherwise (described by four variables);

(4) a *single-ratio threshold linear policy* that is similar to the previous one except for a threshold of the ratio form $y/(T-t) \geqslant \rho$ for some fixed $\rho \geqslant 0$.

In numerical experiments provided later, we confine ourselves to policies in these classes. We note that, aside from computational tractability, policies of this type are also more easily implemented than more general policies.

4. Integrated Learning and Pricing Policy Optimization

The current state of knowledge about the demand model is represented by the posterior distribution of model parameters as described in §2. The specialization of the AA given in §2.3 keeps a finite-sample approximation of the posterior distribution up to date. On the other hand, as shown in §3, policy selection is naturally accomplished by a simulation-based optimization method that requires a sample from the posterior. Thus, we integrate learning with optimization by using the sample of predictions produced by AA inside the policy optimization. This step replaces a direct aggregation of predictions by averaging their contributions to simulation-based calculation of a new optimal policy. A schematic representation of our procedure, its main steps, and corresponding changes in the sample of predictions for a single learning stage are shown in Figure 1.

An entire learning process unfolds as follows. We fix the maximum number of decision periods in a single learning stage: a number S between 1 and T (*periodicity*). With periodicity of one, distribution updates occur after every time step (online), whereas with periodicity of T, they occur after the completion of a sales season (offline). Moreover, because a fully online mode may suffer from excessive computational overhead, we consider intermediate situations ($1 < S < T$), in which learning occurs multiple times per planning horizon, but not as frequently as every decision period. A distribution update also occurs at the end of the planning horizon or if all items are sold out. The process begins with an initial sample from the prior distribution. Then, for each time horizon, the process:

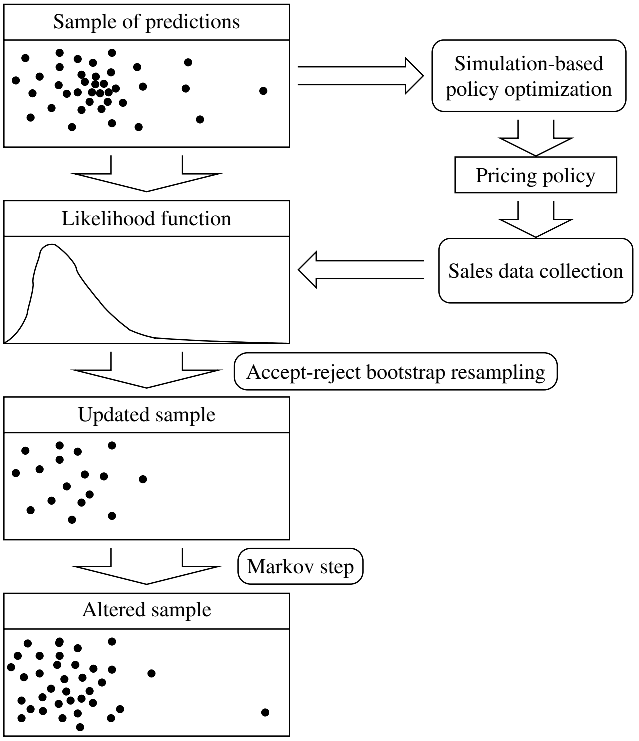
(1) recomputes the pricing policy until the end of the current horizon using a simulation-based optimization of the objective in (6);

(2) observes the sales for up to S decision periods, or until the end of horizon is reached, or all items are sold;

(3) updates the finite-sample approximation to the posterior parameter distribution using bootstrap resampling;

(4) alters the parameter sample using a random Markov step;

Figure 1. Main steps of the integrated learning-policy optimization procedure.



(5) if the end of horizon is reached, or all items are sold, moves to the next time horizon; and

(6) returns to Step 1.

We next provide some details of the resulting numerical procedure by concentrating on its three components: simulation-based pricing policy optimization, finite-sample posterior distribution update, and sample alteration by a random Markov step. Additional implementation information is provided in the online appendix.

In this procedure, the pricing policy is recomputed after each update of the posterior distribution. When the policy is recomputed at an intermediate point t' of the planning horizon, the objective function (5) has to be modified by taking into account the expected revenues over $[t', T]$ only, and by replacing the prior distribution $\mathbf{x}(\varnothing, \varnothing)$ with the posterior $\tilde{\mathbf{x}} = \mathbf{x}(\mathcal{N}_{t'}, \mathcal{P}_{t'}(p'(\cdot), \mathcal{N}_{t'}))$ corresponding to the observed histories $\mathcal{N}_{t'}, \mathcal{P}_{t'}(p'(\cdot), \mathcal{N}_{t'})$ obtained under the previous policy. The modified objective is to maximize future expected revenues:

$$\max_{p(\cdot)} E_{\tilde{\mathbf{x}}}\Bigg[\sum_{\mathcal{N}_t \in \Re^{\Delta}:\, \mathcal{N}_t \cap [0,\, t'-1] = \mathcal{N}_{t'}} (R(p(\cdot), \mathcal{N}_t) - R(p'(\cdot), \mathcal{N}_{t'}))\, \frac{L(\mathcal{N}_t, \mathcal{P}_t(p(\cdot), \mathcal{N}_t) \mid \tilde{\mathbf{x}})}{L(\mathcal{N}_{t'}, \mathcal{P}_{t'}(p'(\cdot), \mathcal{N}_{t'}) \mid \tilde{\mathbf{x}})}\Bigg], \qquad (6)$$

where $R(p'(\cdot), \mathcal{N}_{t'})$ represents past revenues (a constant), $L(\mathcal{N}_t, \mathcal{P}_t(p(\cdot), \mathcal{N}_t) \mid \tilde{\mathbf{x}})/L(\mathcal{N}_{t'}, \mathcal{P}_{t'}(p'(\cdot), \mathcal{N}_{t'}) \mid \tilde{\mathbf{x}})$ represents the portion of the likelihood corresponding to the future segment of the sales and price process histories, and $p(\cdot)$ is a future pricing policy (coinciding with $p'(\cdot)$ up to time t').

Because an analytic calculation of the objective (6) is impossible in general, we use numerical approximation via Monte Carlo integration over the most recent posterior density and the future sales process paths. The two-step simulation approach to objective evaluation has been outlined in §3. In that procedure, the first step involves sampling parameter vectors. The key to computational efficiency of the integrated learning and policy evaluation procedure is that this sample is already available from the AA. It is only necessary to simulate the sample paths using a given pricing policy and the probabilities of sale corresponding to sample elements.

The second component of the integrated procedure, the finite-sample posterior distribution update, is handled by bootstrap resampling, which may result in many duplicates in the updated sample, particularly for highly weighted predictors. However, the third component, sample alteration, breaks up ties between different sample points. A reasonable choice for the elementary predictors of the Markov switching type are realizations of Gaussian random walks with a step of mean zero. The standard deviation parameter of the step can be used to adjust the average magnitude of the step. To make the method more robust to changes in the environment, we also allow occasional "restarts"—a small (random) number of vectors in the new sample are sampled from a prior distribution rather than from a random walk around the value from the old sample.

The resulting learning procedure can be viewed from a genetic algorithm perspective. The sample represents a population of parameter vectors. The fitness of each parameter vector is the likelihood of the observed sales and price history given this parameter vector. The number of offspring of the vector in the new sample is proportional to its fitness. Each offspring also undergoes a mutation. This mutation is usually small (Gaussian random walk), but sometimes a strong deviation from the parent point appears (reset to a prior distribution). This achieves diversity in the sample and an appropriate trade-off between exploitation of previous good values and exploration of new ones.

In many settings, it is possible to construct a learning system using the general approach of *reinforcement learning* (see, for example, Sutton and Barto 1998)—a methodology closely related to Markov decision processes and focused on estimation of a value function. However, reinforcement learning methods typically depend on the assumption that the "environment" is time stationary. In the case of dynamic pricing, the key component of the environment is consumer demand and because the demand models considered by us are inherently nonstationary within each planning horizon, an application of value-function-based reinforcement learning techniques becomes particularly difficult. Generally, any method based on directly learning the optimal pricing policy is likely to be inappropriate for a dynamic pricing problem with an unknown demand model if consumer response to a policy is time and state dependent. Such methods require the whole planning horizon to collect the sales data needed to evaluate the performance of a single pricing policy. Moreover, after observing the effects of a given policy, it is impossible to know the performance of other policies until new sales data are collected. On the other hand, a prediction approach does not pose the same difficulty because it is possible to form alternative predictions of the customer response to a given pricing policy and to simultaneously evaluate how they match the observed data. Therefore, we have chosen to formulate learning in terms of improving predictions of customer response to selected pricing policies.

Next, we report our numerical experience with this integrated learning-optimization procedure. We specify the demand models used in §§5.1 and 5.2 and report the results of computer simulations of the learning process in §5.3.

5. Application: Markets with Strategic Consumers

Myopic customers compare the current price with their valuation and make a purchase if their valuation exceeds the price; that is, as soon as their current consumer surplus is positive. They disregard future prices or product availability. In contrast, a strategic customer will compare the current purchasing opportunity to potential future opportunities and decide whether to purchase now or to wait. A model that captures *strategicity* of customers should therefore specify customer beliefs about future product prices and availability.

In this section, we apply the general learning methodology to a demand model that captures such strategic behavior.

5.1. Demand Model That Captures Consumer Strategicity

The demand model for strategic consumers is based on a game-theoretic consumer-choice model of Levin et al. (2005) that leads to specific forms for the sale probability $\Lambda^{\mathbf{x}}(t, y, p)$. In the online appendix, we present a set of assumptions ensuring that all of the information necessary for a customer to form his beliefs at time t is contained in the current inventory level y, price p, and, perhaps, some additional constant parameters. The assumption that customers can observe the remaining inventory is reasonable in many settings. For example, users of online travel-booking websites can view aircraft layouts showing available seats, and many online booksellers and other retailers show the number of items remaining in stock.

We extend the basic setup of Example 1 to the case of strategic customers as follows. A customer with valuation b who makes a purchase at price p, evaluates it in terms of the surplus $b - p$. The value of the surplus for an item purchased in the future is discounted by a factor $\beta \in [0, 1]$ per time period, which can be interpreted as the degree of strategicity of the customer. This is a natural interpretation because customers with $\beta = 0$ will place no value on future surpluses (behave myopically), whereas customers with $\beta > 0$ will consider the possibility of future surpluses (behave strategically). Because the population is homogeneous, market participants (company and customers) need to track the number of remaining customers $n = N - (Y - y)$ to make optimal strategy decisions. This is relatively easy for companies that are monitoring "hits" to their website, but consumers may have to rely on the speed at which the inventory is dropping to estimate the total number of customers.

In this paper, we also make a key simplifying assumption that *customers' rationality is limited*, and they treat the future price realizations as random values (moves of nature) rather than as values strategically selected by a rational player (the company). Specifically, we assume that they use a common *anticipated price process* $\tilde{p}(t)$, which is a Markov process on Π with finite expectation for all t. The initial value of this Markov process at time t is the last price seen by the customers. That is, if the company uses price p at time t, the distribution of prices anticipated by the customers at time $t + 1$ is that of $[\tilde{p}(t+1) \mid \tilde{p}(t) = p]$. This model of consumer anticipation with respect to future prices has been used, for example, by Assuncao and Meyer (1993). The customers still treat each other as rational participants in the resulting stochastic dynamic game. There are several simple models that satisfy the Markovian assumption. For example, the $\tilde{p}(t)$s can be independent for different t. Alternatively, when the set of feasible prices Π is discrete, we can assume (as we do in our benchmark numerical examples) that $\tilde{p}(t)$ is a random walk with probability q of moving to the next-higher value in Π, probability r of moving to the next-lower value in Π, and probability $1 - q - r$ of staying at the same value. Once the walk reaches an extreme value in Π, it can only stay constant or move to interior values; that is, it will stay constant with probability $1 - r$ if it is at a maximum, or probability $1 - q$ at a minimum. This model of beliefs about prices assumes customer knowledge of the set Π. If Π is a continuous set, the assumed knowledge of possible prices is simply the minimum and maximum of an interval. In the case of discrete prices, customers can often guess likely prices. For example, it is quite common in retail sales to use prices of the form \$99, \$109, \$119, \$129, ... or discounts of the form 5%, 10%, 15%, ..., and consumers are well aware of this.

Let $S^{\mathbf{x}}(t, y, n, p)$ denote the expected present value of the customer surplus *at time t* given the knowledge of y, n, and the price p used by the company *in the previous* decision period and *before* the current period's price is observed. In the online appendix, we derive the following generalization of (1):

$$\Lambda^{\mathbf{x}}(t, y, p) = \bar{\lambda}(N - (Y - y))P^{\mathbf{x}}(B(t) \geqslant p + \beta S^{\mathbf{x}}(t+1, y, N - (Y - y), p)). \tag{7}$$

The quantity $\beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$, interpreted as the present value of the expected future surplus at time t, captures the effects of consumer strategicity on demand: it can be viewed as adjusting consumer reaction to price relative to the myopic case. Note that, aside from the valuation distribution, the sale probability given by (7) only requires knowledge of the *adjustment term* $\beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$. For learning purposes, the value of the adjustment term can be modelled with a detailed consumer-choice model such as the one presented in the online appendix, or an empirical approximation (to be discussed below). In either case, the parameter vector $\mathbf{x}$ of Example 1 must be extended with the parameters of the corresponding model.

Although we assume that customers' behavior is consistent with knowledge of the vector $\mathbf{x}$, the company will typically not have information about at least some of its components. However, the company may possess prior estimates for some of these parameters through, for example, consumer surveys. Such estimates can be introduced naturally to the learning process as part of a prior distribution for $\mathbf{x}$.

5.2. Empirical Model for the Discounted Expected Surplus

When the assumptions underlying the consumer-choice model hold, we can view a company's learning process as a gradual improvement of knowledge of all unknown model parameters in $\mathbf{x}$. This corresponds to learning $\bar{\lambda}$, β, the distribution $F_t(\cdot)$ of $B(t)$ for each time t, and the set of parameters that determine the behavior of the anticipated price process $\tilde{p}(t)$. Although this may be possible in principle, it is unlikely to be computationally feasible in practice with problems of realistic size; thus, approximations are required.

One obvious approximation is to replace the freely time-varying valuation distributions $F_t(\cdot)$ with a constant distribution $F(\cdot)$, or introduce a model for the time variation. In our numerical experiments, we use the constant distribution approximation. Now, because the sale probability given by (7) only requires knowledge of the adjustment term $\beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$, we can replace learning of the remaining parameters in $\mathbf{x}$ with simply learning the parameters of an approximation to this adjustment term. We will call such an approximation an *empirical model* for strategic consumers. In selecting the form of an empirical model, it is important to understand typical characteristics of $\beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$ as a function of time t, remaining inventory y, and price p. The structural results derived in the online appendix provide some guidance in selecting the empirical model. Numerical examination of typical behavior of $\beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$ resulting from the consumer-choice model can also assist in selection.

Based on the structural results, we may expect the expected surplus $S^{\mathbf{x}}(t, y, N-(Y-y), p)$ to be a decreasing function of t and an increasing function of y, and the expression $p + \beta S^{\mathbf{x}}(t+1, y, N-(Y-y), p)$ to be an increasing function of p. Two additional properties are difficult to prove in general, but are both intuitively reasonable and supported by numerical experiments. Although this has not been proved as a structural result, we may expect from standard "decreasing marginal return" considerations that the expected surplus will typically be concave in t and y. In addition, if the future prices anticipated by the customers increase monotonically in the last price observed by them, then it is reasonable to expect that the expected surplus is decreasing in prices, and drops to nearly zero when the price is high enough.

All of these dependencies are indeed present in the following numerical illustration of the expected surplus determined by the detailed consumer choice model. The parameters used in the choice model are:

- initial inventory $Y = 20$,
- initial number of customers $N = 30$,
- planning horizon $T = 200$,
- $\bar{\lambda}T = 4$, $\beta = 1$,
- valuation distribution $\mathtt{Normal}(4, 2)$,
- a discrete set of 50 prices $\{0.2, 0.4, \ldots, 10\}$, and
- a random walk with probability 0.05 of moving higher or lower than current price on this price set as an anticipated price process model.

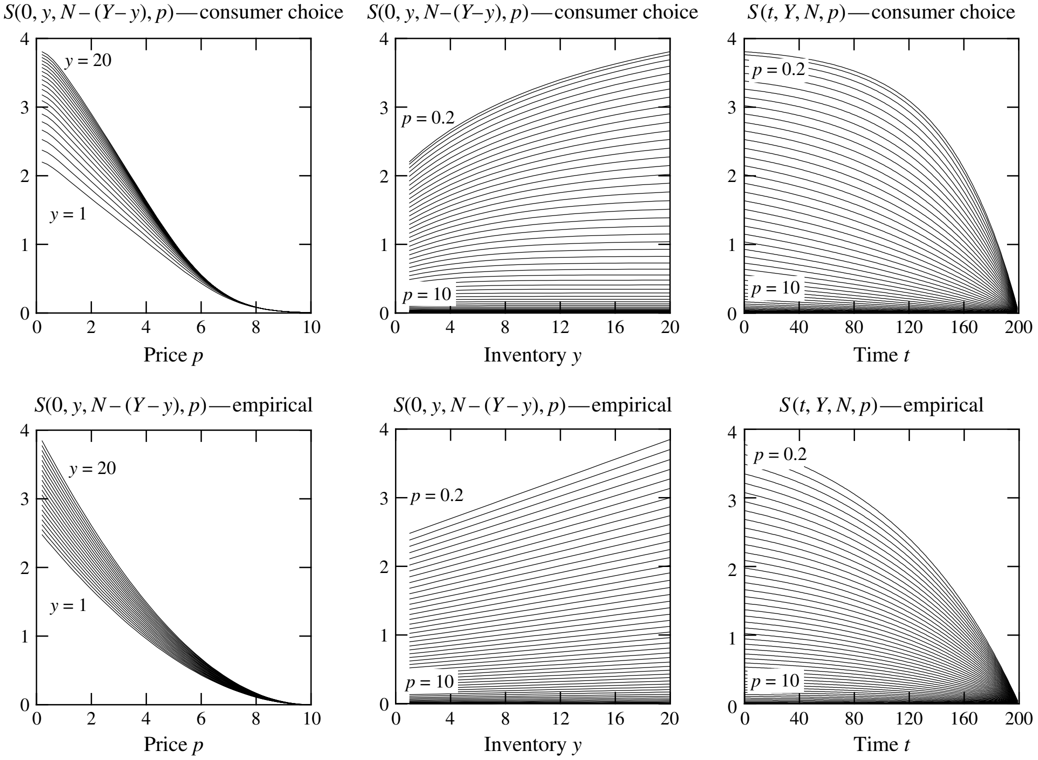
The top three plots in Figure 2 show the graphs of expected surplus $S^{\mathbf{x}}(0, y, N-(Y-y), p)$ at time 0 for different levels of inventory y as functions of price p and for different prices p as functions of inventory y, as well as of $S^{\mathbf{x}}(t, Y, N, p)$ as functions of time t for different prices p. The expected surplus is, approximately, hyperbolic as a function of price; linear increasing as a function of inventory, and tail negative exponential as a function of time. This, together with properties outlined above, motivates the following empirical model to approximate the discounted expected future surplus term in the sale probability model (7):

$$\begin{aligned}&\tilde{S}^{a,b,c,d}(t, y, N-(Y-y), p)\\&\quad = c\Big(1 + d\frac{y}{Y}\Big)\frac{\sqrt{(1-p/\max\Pi)^2 + a^2} - a}{\sqrt{1+a^2} - a}\\&\qquad \times \frac{1 - e^{-b(1-t/T)}}{1 - e^{-b}}.\end{aligned} \tag{8}$$

This model approximates the effects of many of the parameters in the detailed consumer-choice model; namely, β, and the set of parameters for the anticipated price process $\tilde{p}(t)$. Estimation of the parameter vector $\mathbf{x}$ is thereby simplified to estimation of the four parameters a, b, c, d in the empirical model plus the parameters of the valuation distributions $B(t)$.

Examination of graphs of $\tilde{S}^{a,b,c,d}(0, y, N-(Y-y), p)$ and of $\tilde{S}^{a,b,c,d}(t, Y, N, p)$ for parameter values $a = b = 2$,

Figure 2. Expected surplus and its empirical model: Graphs of $S(0, y, N-(Y-y), p)$ as functions of price p for different inventory levels y, and as functions of inventory y for different prices p, as well as the graphs of $S(t, Y, N, p)$ as functions of time t for different prices p.



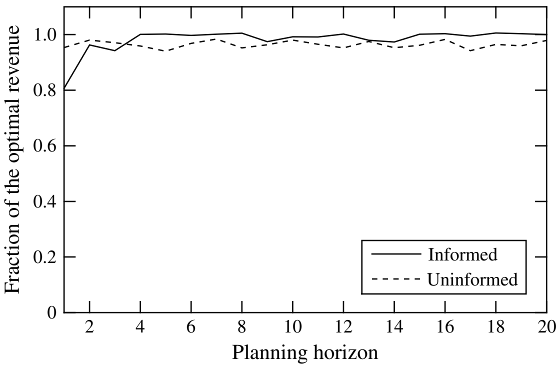
$c = 2$, $d = 0.6$ (three bottom plots in Figure 2) confirms that the empirical model exhibits behavior sufficiently similar to that of the consumer-choice model to justify its use as an approximation in our numerical experiments. Clearly, in any specific application, more elaborate models could be devised to accomplish this approximation.

5.3. Results of Computer Simulations

In this section, we numerically examine the effects of strategic consumer behavior in a learning environment. We assume that the actual customer population behaves according to the consumer-choice model discussed in the online appendix, which leads directly to the probability of sale of the form (7). The numerical parameter values are the same as in §5.2, with the exception of the valuation distribution mean, the population size N, the initial inventory Y, and the time horizon T. The company, however, can only use an empirical model of the form (8) to adjust for consumer behavior. The pricing policy of the company in these experiments belongs to one of three classes: open-loop, single-threshold linear, and single-ratio threshold linear. The policies are otherwise unrestricted and can give rise to arbitrary nonnegative prices in a given time interval (not only in the set $\{0.2, 0.4, \ldots, 9.8, 10.0\}$ used for the random walk model of the anticipated price $\tilde{p}(t)$). Therefore, in simulation of customer behavior, we approximate the expected surplus for general price p by the expected surplus for the highest price from $\{0.2, 0.4, \ldots, 9.8, 10.0\}$ less than or equal to p. The size of the sample of parameter vectors is $K = 10{,}000$. In each evaluation of the function (6), we only use 10% of the sample of parameter vectors (selected randomly, with replacement) and, for each vector sampled, we generate a single sales realization. This decreases the accuracy of approximation in each particular function evaluation, but may be unimportant overall, because the optimization algorithm used in our experiments (DFO) creates an interpolation model of the simulated function.

In the first experiment, $Y = 100$, $N = 150$, and $T = 1{,}000$ over 20 planning horizons. Updates are done "offline" (one distribution update per planning horizon). In this experiment we compare the performance of the learning algorithm under two scenarios. In an "uninformed" scenario, the company only tries to learn the parameters of the valuation (mean and standard deviation for the normal). In an "informed" scenario, the company is aware that customers may behave strategically and also tries to learn the four parameters of the empirical model. The a, b, c, d prior distribution under this scenario is uniform over the four-dimensional hyperrectangle $[0.01, 10]^2 \times [0, 10]^2$. In both scenarios, the prior for the valuation mean is uniform on $[2, 8]$ whereas the prior for the valuation standard deviation is uniform on $[0.5, 3]$. All other parameters are assumed to be known. The distribution of the switching step in the parameter sample update is Gaussian (truncated so that the

Figure 3. Average observed revenues as a fraction of the optimal revenues for the true model over 10 replications of 20 consecutive planning horizons with offline learning and $Y = 100$, $N = 150$, and $T = 1{,}000$.



resulting parameter vector remains within the support set of the prior distribution). The steps for different parameters are independent with zero means and standard deviations of 0.05. The probability that a vector is reset back to the prior distribution is 0.001. The true valuation distribution is $\texttt{Normal}(4, 2)$. In this experiment, the pricing policy of the company is in the single-threshold linear class with a threshold at 50. Figure 3 shows each respective planning horizon's total revenue as a fraction of the optimal expected revenues for the detailed customer behavior model (computed by the standard dynamic programming approach). The results are averaged over 10 replications. We see that the performance of an informed company quickly dominates the performance of an uninformed company. Overall, the average performance for the informed scenario is 98.2% of the optimal expected revenue, and the performance for the uninformed scenario is 96.4%.

In the second experiment, we examine the effect of learning periodicity on the algorithm's performance. We increase the number of replications to 40, but decrease the scale of the example to $Y = 20$, $N = 30$, and $T = 200$. With this many replications, differences in average performance greater than 1.7% are significant at the 5% level (based on a one-sided t-test, and given the observed standard deviations). This also applies to subsequent experiments. The true valuation distribution remains $\texttt{Normal}(4, 2)$. We compare the same two learning scenarios, but with different values of periodicity: every 20, 40, 100, and 200 decision periods (10, 5, 2, and 1 distribution updates and policy calculations per planning horizon, respectively). The switching distribution's standard deviation is scaled by the square root of (T/periodicity), and the reset probability is divided by (T/periodicity). Two pricing policies are tested: single-threshold linear with a threshold at 10, and an open-loop policy in which the entire time horizon is partitioned into five equal subintervals and the price remains constant in every subinterval. The two policies are labelled TL10 and OL5, respectively. Each policy is recalculated after every distribution update. Although an open-loop policy does not depend on the inventory level explicitly, its recalculation after a distribution update allows the company to incorporate a limited form of state feedback. Table 1 presents the average of the observed revenues in all planning horizons and all replications as a fraction of the optimal expected revenues for the true model, together with the standard deviation of the average of this fraction per replication. We see that decreasing periodicity (that is, making learning more frequent) significantly increases the average performance for the informed scenario to the point that it approaches 99%. On the other hand, under the uninformed scenario, the gap in performance (of about 3.5% and 2% of the optimum for TL10 and OL5 policies, respectively) persists and does not close with increasing periodicity of learning. This illustrates the importance of having an appropriate (even if not absolutely accurate) model of consumer demand in the online learning framework. A slight advantage of the uninformed company when periodicity is 200 (offline mode) can be explained by the gain in learning efficiency in the first few time horizons when the number of parameters is smaller.

Table 1. Average observed revenues for TL10 and OL5 policies as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, and varied periodicity.

Policy	Periodicity	Informed	Uninformed
TL10	20	0.988 ± 0.028	0.950 ± 0.025
TL10	40	0.981 ± 0.030	0.946 ± 0.026
TL10	100	0.981 ± 0.024	0.962 ± 0.023
TL10	200	0.948 ± 0.027	0.960 ± 0.034
OL5	20	0.991 ± 0.028	0.973 ± 0.032
OL5	40	0.994 ± 0.029	0.971 ± 0.031
OL5	100	0.976 ± 0.021	0.973 ± 0.029
OL5	200	0.973 ± 0.035	0.976 ± 0.033

Although the gap of 2%–4% between the informed and uninformed scenarios is significant, it is interesting to know what factors may affect it, and whether it can be larger. As we have already pointed out, a marketplace with strategic consumers results in a nonstationary and state-dependent demand intensity. Our model reflects this in the time- and state-dependent form of the expected surplus, which is completely ignored by an uninformed company. The expected consumer surplus starts at some large value in the beginning of the planning horizon and subsequently drops to zero at the end. Equation (7) reveals that any particular price will result in larger demand at the end of the horizon than at the beginning. This is because, at the end of the horizon, there is a higher chance that the valuation-price difference will exceed the (smaller) present value of the surplus. Thus, one may expect the performance gap to be larger when the maximum value of the surplus at typical price levels is larger. The two factors that may significantly affect the maximum surplus are the uncertainty of the valuation relative to its constant part (represented by the coefficient of variation $CV = \sigma/\mu$), and the availability of the product. Note that the constant part of the valuation distribution is typically incorporated in the overall price and is not likely to affect the maximum surplus observed during the planning horizon. We next compare the performance of the open-loop policy with five intervals (OL5) under the informed and uninformed scenarios, two periodicity values (40 and 200), two values of the coefficient of variation (0.5 and 1, corresponding to different valuation means 4 and 2 but the same standard deviation of 2), and two levels of the initial inventory (20 and 30). All other settings are the same as in the previous experiment. The averages and the standard deviations of the revenues per replication as a fraction of the optimum are presented in Table 2. In this experiment, we focus on the long-term performance and exclude the first time horizon in every replication. The results confirm our intuition. The performance gap between the two scenarios increases with an increase in the relative uncertainty in the valuation and with the initial inventory. We also note that the standard deviation of the performance ratio increases with the relative uncertainty in the valuation. This is natural because demand becomes less predictable when the uncertainty in the valuation is higher. The most difficult situation is $Y = 30$ and coefficient of variation $CV = 1$ for either scenario. However, the uncertainty of results and performance deterioration is particularly high for an uninformed company. It is also interesting to note that more frequent learning leads to deterioration of performance in the uninformed case. This is most likely due to the phenomenon of "overfitting" of the myopic model to the sales data during a particular segment of the planning horizon. The myopic model fitted, for example, in the beginning of the planning horizon will be inadequate for prediction at the end of the planning horizon.

Table 2. Average observed revenues (excluding the first horizon) for OL5 policy as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $N = 30$, $T = 200$, and varied periodicity, Y and the valuation's coefficient of variation CV.

Y	CV	Periodicity	Informed	Uninformed
20	0.5	40	1.000 ± 0.030	0.971 ± 0.031
20	0.5	200	0.989 ± 0.035	0.979 ± 0.031
20	1.0	40	0.986 ± 0.034	0.935 ± 0.052
20	1.0	200	0.969 ± 0.037	0.952 ± 0.069
30	0.5	40	0.989 ± 0.041	0.907 ± 0.049
30	0.5	200	0.993 ± 0.030	0.945 ± 0.028
30	1.0	40	0.981 ± 0.045	0.865 ± 0.064
30	1.0	200	0.964 ± 0.045	0.883 ± 0.062

The next experiment examines the sensitivity of performance of the learning procedure to the parameters of the predictors. Specifically, we examine whether the restart of the predictor to a vector drawn from the prior distribution adds any value, and what the effects of the magnitude of the switching step are. The overall setup is similar to the previous experiment, but we only examine a valuation mean of 2 ($CV = 1$) and an inventory level $Y = 20$. In addition to the informed and uninformed scenarios and periodicities of 40 and 200, we examine: the setup with no restarts, and two setups with restarts where the magnitude of the switching step was on the average five times higher and five times lower than in the previous experiment. The results are presented in Table 3 where, for convenience, we also include lines 3 and 4 of Table 2. We see that the absence of restarts does not significantly affect the performance in the informed scenario (the difference in averages is less than 1%), but makes the method less robust in the uninformed scenario where the company's model of demand is less adequate. On the other hand, under the uninformed scenario, a smaller magnitude of the switching step reduced the problem of overfitting (and the larger magnitude increased it). The performance in the informed scenario is more robust to the predictor selections due to a more adequate model.

It is also interesting to see how *policy selection* affects performance. Using the consumer-choice model parameters as in the previous experiment and the default predictors, we examine the following policies: the open-loop policy with two prices (OL2), the single-threshold linear policy with thresholds 5, 10, and 15 (TL5, TL10, and TL15), and the single-ratio threshold policy with thresholds $0.75Y/T$, Y/T, and $1.25Y/T$ (RTL0.75, RTL1.0, RTL1.25). The results of these runs are summarized in Table 4. Again, for comparison purposes, lines 3 and 4 of Table 2 are repeated in this table. First, we see that the OL2 policy results in performance deterioration in the uninformed case compared to the OL5 policy showing that two prices are insufficient.

Table 3. Average observed revenues (excluding the first horizon) for OL5 policy as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, $CV = 1$, varied periodicity, and various predictor selections.

Predictors	Periodicity	Informed	Uninformed
Default	40	0.986 ± 0.034	0.935 ± 0.052
Default	200	0.969 ± 0.037	0.952 ± 0.069
No resets	40	0.995 ± 0.039	0.925 ± 0.066
No resets	200	0.968 ± 0.038	0.937 ± 0.061
Smaller switches	40	0.979 ± 0.040	0.955 ± 0.052
Smaller switches	200	0.973 ± 0.038	0.956 ± 0.052
Larger switches	40	0.978 ± 0.038	0.894 ± 0.045
Larger switches	200	0.955 ± 0.035	0.922 ± 0.051

Table 4. Average observed revenues (excluding the first horizon) for different policies as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, $CV = 1$, and varied periodicity.

Policy	Periodicity	Informed	Uninformed
OL2	40	0.975 ± 0.051	0.853 ± 0.054
OL2	200	0.981 ± 0.041	0.904 ± 0.064
OL5	40	0.986 ± 0.034	0.935 ± 0.052
OL5	200	0.969 ± 0.037	0.952 ± 0.069
TL5	40	0.997 ± 0.033	0.980 ± 0.032
TL5	200	0.976 ± 0.040	0.964 ± 0.038
TL10	40	0.990 ± 0.045	0.960 ± 0.034
TL10	200	0.968 ± 0.047	0.958 ± 0.041
TL15	40	0.981 ± 0.038	0.930 ± 0.041
TL15	200	0.973 ± 0.042	0.959 ± 0.035
RTL0.75	40	0.984 ± 0.037	0.979 ± 0.036
RTL0.75	200	0.989 ± 0.034	0.985 ± 0.035
RTL1.0	40	0.974 ± 0.033	0.947 ± 0.035
RTL1.0	200	0.970 ± 0.029	0.945 ± 0.038
RTL1.25	40	0.969 ± 0.037	0.929 ± 0.038
RTL1.25	200	0.956 ± 0.034	0.905 ± 0.055

The policies of the RTL type show a better performance for a smaller threshold value 0.75, whereas the policies of the TL type show a better performance for a threshold of 5. The overall "winner" appears to be TL5. For both periodicities, a smaller threshold (resulting in a dramatic price change for lower values of inventory) leads to better performance. This can be explained by the need for greater control over prices near the end of the planning horizon when there are usually only a few items left in the inventory. The end of the horizon is especially important because the expected customer surplus is quite low and the remaining few items can collect much higher prices than in the beginning of the season. On the other hand, if sales were not very active, and there are many items left in the inventory, the company will prefer to keep prices low. If the threshold is high, it simply pushes prices higher too early. We also see that the performance gap between the informed and uninformed scenarios persists across all policy types.

In all of the above experiments, we see that an informed company, even one using an approximate empirical model, has a significant advantage over an uninformed one. As an additional robustness test, we examine what happens if customer perceptions of the pricing policy change from one planning horizon to the next. This partially relaxes the assumption of the constant parameter vector **x**. In the first robustness experiment, we assume that the parameters q, r of the anticipated price process (initially equal to 0.05) are perturbed from one planning horizon to the next using a $\texttt{Normal}(0, 0.02)$ random step truncated so that the resulting values satisfy $q, r \geqslant 0$ and $q + r \leqslant 1$. This random perturbation in parameters results in approximately 4% variability in the optimal expected revenues between consecutive planning horizons. The results of the experiment using the OL5 policy, $Y = 20$, $N = 30$, $T = 200$, $CV = 1$ and default predictor settings are presented in Table 5. We do not see deterioration in performance in the informed case while the gap between the informed and uninformed scenarios increases. This shows that the learning procedure is sufficiently flexible to handle moderate random changes occurring in the marketplace.

Table 5. Drift in the random walk parameters: average observed revenues (excluding the first horizon) for OL5 policy as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, $CV = 1$, and varied periodicity.

Periodicity	Informed	Uninformed
40	0.993 ± 0.032	0.929 ± 0.060
200	0.977 ± 0.038	0.912 ± 0.064

In the second robustness experiment, similar to the previous one, we examine what happens if customer perceptions of the pricing policy change in a systematic fashion. The customers adjust their values of q and r from one time horizon to the next using the exponential smoothing formulas

$$q := \alpha \hat{q} + (1 - \alpha) q,$$
$$r := \alpha \hat{r} + (1 - \alpha) r,$$

where $\alpha = 0.1$ is a smoothing coefficient, and $\hat{q}$, $\hat{r}$ are the estimates based on the complete price history $\mathcal{P}_{t^*} = \{p_1, p_2, \ldots, p_{t^*}\}$ in the current time horizon. Specifically,

$$\hat{q} = \frac{1}{t^* \Delta p} \sum_{t=1}^{t^*} (p_t - p_{t-1})_+,$$
$$\hat{r} = \frac{1}{t^* \Delta p} \sum_{t=1}^{t^*} (p_{t-1} - p_t)_+,$$

where $\Delta p = 0.2$ is the step size of the price grid. These estimates satisfy the relation that the total up and down changes in the price process are equal to the expected up and down changes $t^* \Delta p q$ and $t^* \Delta p r$, respectively, in the value of the random walk with step Δp. Such systematic changes in the random walk parameters constitute a limited case of customer learning. The results of the experiment are presented in Table 6. Again, we do not see a deterioration in performance in the informed case while there is a larger gap between informed and uninformed cases. Thus, the demand learning procedure can also handle some degree of customer learning.

As a final robustness test of the learning procedure, we examine its performance for the case of unknown $\bar{\lambda}$. Specifically, we assume that the company has a prior on $\bar{\lambda}$ that

Table 6. Systematic change in the random walk parameters: average observed revenues (excluding the first horizon) for OL5 policy as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, $CV = 1$, and varied periodicity.

Periodicity	Informed	Uninformed
40	0.984 ± 0.039	0.905 ± 0.066
200	0.984 ± 0.038	0.922 ± 0.051

is uniform on the interval from one-half to double of its true value. The number of decision periods in this experiment is increased to $T = 400$ to ensure that the maximum possible sale probability remains sufficiently small even for the maximum $\bar{\lambda}$ in the range of the prior distribution. The results for the TL5 policy (other settings are the same as in the previous experiments) are given in Table 7. We exclude the first two time horizons in averages to account for slower learning resulting from a larger parameter space. On the remaining horizons, the procedure shows similar behavior in terms of the highest observed percentage and the gap between the two scenarios (99% performance for the informed case and 1.7% gap for higher learning frequency).

6. Conclusions

This paper presents a new and robust procedure for the learning of customer demand characteristics that is integrated with dynamic pricing. The demand-learning and pricing optimization methodology is general and independent of specific distributional assumptions. This adaptive procedure permits learning of consumer response through observation of sales over successive learning stages. The learning component draws on the ideas of the general aggregating algorithm and can learn all characteristics of demand simultaneously. The pricing policy is optimized by a simulation-based procedure. We provide an efficient implementation of the method and illustrate its use with a particularly complex dynamic pricing problem faced by a monopolist whose customers are strategic.

Table 7. Unknown $\bar{\lambda}$: average observed revenues (excluding the first two horizons) for TL5 policy as a fraction of the optimal expected revenues for the true model ± standard deviation of the average of this fraction per replication for two scenarios in the case of $Y = 20$, $N = 30$, $T = 200$, $CV = 1$, and varied periodicity.

Periodicity	Informed	Uninformed
40	0.990 ± 0.037	0.983 ± 0.044
200	0.988 ± 0.035	0.973 ± 0.045

We also demonstrate that the proposed learning approach is robust in a statistical sense. The observed performance of the method is very close to that of the optimal dynamic pricing policy for the case when the demand model is known exactly despite the use of an approximate model of consumer behavior in learning.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgments

The authors thank the associate editor and two anonymous referees for their many constructive suggestions that greatly helped to improve the exposition of this paper. This research was supported by the Natural Sciences and Engineering Research Council of Canada (grant numbers 261512-04 and 341412-07).

References

Assuncao, J., R. Meyer. 1993. The rational effect of price promotions on sales and consumption. *Management Sci.* **39**(5) 517–535.

Aviv, Y., A. Pazgal. 2005a. Pricing of short-cycle products through active learning. Working paper, Washington University, St. Louis.

Aviv, Y., A. Pazgal. 2005b. A partially observed Markov decision process for dynamic pricing. *Management Sci.* **51**(9) 1400–1416.

Aviv, Y., A. Pazgal. 2008. Optimal pricing of seasonal products in the presence of forward-looking consumers. *Manufacturing Service Oper. Management* **10**(3) 339–359.

Balvers, R., T. Cosimano. 1990. Actively learning about demand and the dynamics of price adjustment. *Econom. J.* **100** 882–898.

Bertsimas, D., G. Perakis. 2006. Dynamic pricing: A learning approach. *Mathematical and Computational Models for Congestion Charging*, Vol. 101. *Applied Optimization*. Springer, New York, 45–79.

Besanko, D., W. Winston. 1990. Optimal price skimming by a monopolist facing rational consumers. *Management Sci.* **36**(5) 555–567.

Besbes, O., A. Zeevi. 2007. Blind network revenue management. Working paper, Columbia University, New York.

Bitran, G., R. Caldentey. 2003. An overview of pricing models for revenue management. *Manufacturing Service Oper. Management* **5**(3) 203–229.

Carvalho, A., M. Puterman. 2005. Learning and pricing in an Internet environment with binomial demand. *J. Revenue Pricing Management* **3**(4) 320–336.

Conn, A., K. Scheinberg, P. Toint. 1997. Recent progress in unconstrained nonlinear optimization without derivatives. *Math. Programming* **79**(1–3, Ser. B) 397–414.

Cover, T. 1991. Universal portfolios. *Math. Finance* **1** 1–29.

Elmaghraby, W., P. Keskinocak. 2003. Dynamic pricing in the presence of inventory considerations: Research overview, current practices and future directions. *Management Sci.* **49**(10) 1287–1309.

Elmaghraby, W., A. Gülcü, P. Keskinocak. 2008. Designing optimal preannounced markdowns in the presence of rational customers with multi-unit demands. *Manufacturing Service Oper. Management* **10**(1) 126–148.

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* **40**(8) 999–1020.

Levin, Y., J. McGill, M. Nediak. 2005. Optimal dynamic pricing of perishable items by a monopolist facing strategic consumers. Working paper, Queen's University, Kingston, Ontario, Canada.

Levina, T. 2004. Using the aggregating algorithm for portfolio selection. Ph.D. thesis, Rutgers University, Newark, NJ.

Levina, T. 2006. Online methods for portfolio selection. K. Voges, N. Pope, eds. *Business Applications and Computational Intelligence*. Idea Group, Inc., Hershey, PA, 431–459.

Lin, K. 2006. Dynamic pricing with real-time demand learning. *Eur. J. Oper. Res.* **174**(1) 522–538.

Liu, Q., G. van Ryzin. 2008. Strategic capacity rationing to induce early purchases. *Management Sci.* **54**(6) 1115–1131.

McGill, J., G. van Ryzin. 1999. Revenue management: Research overview and prospects. *Transportation Sci.* **33**(2) 233–256.

Petruzzi, N., M. Dada. 2002. Dynamic pricing and inventory control with learning. *Naval Res. Logist.* **49** 303–325.

Phillips, R. 2005. *Pricing and Revenue Optimization*. Stanford University Press, Stanford, CA.

Shaked, M., J. Shanthikumar. 1994. *Stochastic Orders and Their Applications. Probability and Mathematical Statistics*. Academic Press, Boston.

Su, X. 2007. Inter-temporal pricing with strategic customer behavior. *Management Sci.* **53**(5) 726–741.

Sutton, R., A. Barto. 1998. *Reinforcement Learning*: *An Introduction*. MIT Press, Boston.

Talluri, K., G. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Norwell, MA.

Vovk, V. 1990. Aggregating strategies. *Proc. Third Annual Workshop on Computational Learning Theory*. Morgan Kaufmann, San Mateo, CA, 371–383.

Vovk, V. 1999. Derandomizing stochastic prediction strategies. *Machine Learning* **35** 247–282.

OPERATIONS RESEARCH
DOI 10.1287/opre.1080.0577ec
pp. ec1–ec14

INFORMS
© 2009 INFORMS

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Dynamic Pricing with Online Learning and Strategic Consumers: An Application of the Aggregating Algorithm" by Tatsiana Levina, Yuri Levin, Jeff McGill, and Mikhail Nediak, *Operations Research*, DOI 10.1287/opre.1080.0577.

Online Appendix A: additional implementation details

The first component of the integrated policy optimization-learning approach is a simulation-based policy optimization step. While the optimization algorithm is a widely available derivative-free method, we need to implement a special subroutine that evaluates the objective function by simulation. Let the most recent distribution of the parameter vector $\tilde{\mathbf{x}}$ be approximated by a discrete sample $\mathcal{W} = \{\mathbf{x}_i\}_{i=1}^K$. In the terminology of AA, each element of $\mathcal{W}$ corresponds to a prediction of consumer demand by a particular predictor. The second step is to simulate M_i realizations of the future sales process sample path for each $\mathbf{x}_i$ using the pricing policy and the corresponding $\Lambda^{\mathbf{x}_i}(t, y, p)$ as the probability of sale for each t, y, p. If we use the entire sample $\mathcal{W}$ then M_i should be the same fixed value for all sample points. However, using the entire sample will usually entail excessive computation. In this case, we can simulate sample paths for a randomly selected subset of $\mathcal{W}$ (with or without replacement) resulting in random but identically distributed M_i's. Let the corresponding sales process histories be labelled as $\mathcal{N}_{t_{ij}}^{ij}$, $j = 1, \ldots, M_i$ (where t_{ij} represents T or the time after the sale of the Y'th item in this sales process path). Then the expectation in (6) can be approximated by the average over $\mathcal{W}$ and the simulated sample paths:

$$\frac{1}{M} \sum_{i=1}^{K} \sum_{j=1}^{M_i} (R(\mathcal{N}_{t_{ij}}^{ij}) - R(\mathcal{N}_{t'})),$$

where $M = \sum_{i=1}^{K} M_i$ is the total number of sample paths.

The second component is an update procedure for the finite-sample approximation $\mathcal{W}$ to the distribution $\mathbf{x}(\mathcal{N}_{t'}, \mathcal{P}_{t'})$. This update produces a finite-sample approximation $\mathcal{W}'$ to the posterior distribution $\mathbf{x}(\mathcal{N}_t, \mathcal{P}_t)$, where $\mathcal{N}_t \cap [0, t'-1] = \mathcal{N}_{t'}$ and $\mathcal{P}_{t'}$ is a sublist of prices in $\mathcal{P}_t$ up to time $t'-1$, inclusive. This is done by a Monte-Carlo accept-reject algorithm with bootstrap-like resampling of the elements of $\mathcal{W}$ (see, for example, Levina (2006)). The elements of the new sample are labelled $\mathbf{x}'_k$, where k is the counter of points in the new sample $\mathcal{W}'$.

Algorithm: accept-reject with bootstrap resampling

```
Inputs: finite-sample (size $K$) approximation $\mathcal{W}$ of $\mathbf{x}(\mathcal{N}_{t'}, \mathcal{P}_{t'})$,
    sales and price process history $(\mathcal{N}_t, \mathcal{P}_t)$
Output: finite-sample (size $K$) approximation $\mathcal{W}'$ of $\mathbf{x}(\mathcal{N}_t, \mathcal{P}_t)$
    Set $k := 0$
    Set $L_{\max} := \max_{i=1,\ldots,K} L(\mathcal{N}_t, \mathcal{P}_t \mid \mathbf{x}_i)$
    while $k < K$ do
        Choose $u$ from $U[0,1]$
        Choose $j$ randomly from $\{1,\dots,K\}$
        if $u \leq \frac{L(\mathcal{N}_t,\mathcal{P}_t\,|\,\mathbf{x}_j)}{L_{\max}}$ then
            Set $k := k+1$
            Set $\mathbf{x}'_k := \mathbf{x}_j$
        endif
    enddo
```

In selecting the new sample $\mathcal{W}'$, the algorithm favors those elements of $\mathcal{W}$ with a high likelihood ratio.

Online Appendix B: game-theoretic consumer choice model

The consumer behavior model used in this paper shares a number of assumptions with the model described in Levin et al. (2005). However, that model assumed full consumer rationality in the 'game' against the company — an assumption that is not reasonable when the information available to the company is imperfect, and customers cannot know the learning mechanism used by the company. Thus, a fundamental departure in the present paper is the assumption of limited rationality of customers. This assumption is practical and allows us to reduce the complexity of consumer behavior being modeled and to derive a number of intuitive structural results. Such results cannot be established in general in the full rationality case.

We emphasize that, like any model, the present one contains assumptions that may not hold in many real markets. However, the assumptions may hold approximately, and our main goal in presenting the detailed model is to establish a theoretical justification for the demand model (7), that establishes the importance of the expected consumer surplus.

As in Example 1, we let N be the initial market size. The assumption of a finite population is appropriate for modeling strategic consumers since, in this setting, it is necessary to describe individual consumer behavior. Consumer presence from the beginning of the planning horizon represents a marketplace in which each consumer engages in strategic planning and can make a purchase at any time during the selling season. This assumption can be relaxed by incorporating random consumer departures and arrivals (up to some maximum population size). However, such a generalization complicates the analysis and is not crucial for an initial presentation of the methodology.

We also assume that the consumer population is homogeneous (as in Example 1), and this assumption has some important consequences. First, it reduces information requirements for computing the optimal strategy. If the population is nonhomogeneous, then consumers and the company would need to track the dynamics of the population *distribution* over nonhomogeneous characteristics. We show that, with a homogeneous population, it is sufficient to track only the current number of customers, n, who have not yet acquired an item. Second, homogeneity justifies a customer belief that other customers behave in the same way. The knowledge of n may be assumed in electronic markets requiring registration, and when web-sites track and report the number of unique 'hits' to their site. A company has an incentive to report the total market size to its customers to reinforce the perception of competition. Settings where both y and n may be public information include, for example, cruise lines, which sell their own tickets through the internet, as well as charter flights, stadiums and concert halls. Both of these quantities will be known as long as the available seats are shown and the market size is reported on the website.

Stochastic aspects of the model aim at capturing consumer uncertainty: about future prices, about future willingness to pay, and about timing of purchases. Timing uncertainty also encompasses acquisition uncertainty since customers cannot be sure that they will acquire the product if supply is limited — a phenomenon frequently occurring in practice.

Next, we summarize elements of the consumer choice model that are similar to those in Levin et al. (2005).

1. Customer decisions are derived in terms of *eagerness to purchase*, which is a value between 0 and 1 with 0 signifying the absence of desire, and 1 signifying the desire to purchase the product as quickly as possible. The eagerness controls the intensity of purchase opportunities for the customer. The expected number of purchase opportunities for an eager customer during the planning horizon of length T is given by $\bar{\lambda}T$. Because of the choice of the time units of the discrete-time model, the upper bound on the intensity of such events $\bar{\lambda}$ is equal to the probability that an eager customer will actually be able to acquire an item in a given decision period.
2. The probability that a sale to *some* customer occurs is the sum of the probabilities of purchase by individual customers. For example, in the case of k customers eager to purchase an item, the probability of a sale occurring is $k\bar{\lambda}$. Summing purchase probabilities in the discrete-time model corresponds to summing the shopping intensities of individual customers to obtain the total demand intensity for the entire market in a continuous-time model.
3. Each customer has a *valuation* for the product: the maximum amount he/she is willing to spend in the current decision period. A customer with valuation b who makes a purchase at price p, evaluates it in terms of the surplus $b-p$. The value of the surplus for an item purchased in the future is discounted by a factor $\beta \in [0,1]$ per time period, which can be interpreted as the degree

of *strategicity* of the customer. There is no penalty, other than opportunity loss, for failure to acquire an item.

4. Consumers do not know their exact valuations for the product at the time of purchase. The uncertainty in valuation of any customer at time $t \in \{0,1,\dots,T-1\}$ is represented by the same random variable $B(t)$ with a distribution function $F_t(b)$, reflecting homogeneity of the consumer population. We assume that the actual customer eagerness to purchase is the average of the eagerness derived for each possible value b of $B(t)$ over $F_t(b)$.
5. At the time of their decisions, all customers know the current number of customers n, the remaining inventory y, current price p, and their valuation distribution function $F_t(b)$. Customer eagerness at time t for a given inventory level y, number of remaining customers n, price p and valuation b, is denoted as $e^{\mathbf{x}}(t,n,y,p,b) \in [0,1]$. (Note that, since customers know the size of the customer population, the number of remaining customers can be computed from the current inventory level as $n = N - (Y - y)$). The decisions are identical for all customers because of population homogeneity. The superscript $\mathbf{x}$ emphasizes the obvious dependence of customer decisions on the model parameters in $\mathbf{x}$.
6. Customer response to a pricing policy is determined by a stochastic dynamic game among the customers. A round of the game at time t, given y and n, proceeds as follows: customers observe the price p and simultaneously respond by their eagerness to purchase averaged over the valuation distribution. The probability of purchase $\lambda^{\mathbf{x}}(t,n,y,p)$ by the customer in the current time interval is *proportional* to the customer's average eagerness $E_{B(t)|\mathbf{x}}[e^{\mathbf{x}}(t,n,y,p,B(t))]$, and the coefficient of proportionality is $\bar{\lambda}$. The customers make their decisions as if their payoff is a fraction of the market's payoff which is also computed as the average over $B(t)$.

The potential unknowns to the company that determine customer response to prices are: the customer's maximum demand intensity $\bar{\lambda}$, the customer discount factor β, the distribution of $B(t)$ at each time t, $F_t(\cdot)$, and the set of parameters that determine the behavior of the anticipated price process $\tilde{p}(t)$ (for example, the random walk transition probabilities q, r, or distribution parameters for the independent $\tilde{p}(t)$'s case). If the distribution of $B(t)$ is parametric (determined by a finite number of parameters), we can include all of this information in the parameter vector $\mathbf{x}$.

In the online appendix C, we describe how this model, for a known $\mathbf{x}$, predicts customer response (purchase probability) for given y, n, p at time t. A key insight from this analysis is the importance of the consumer's *expected surplus* corresponding to a given state of the process.

Online Appendix C: properties of solutions determined by the consumer choice model

Recall that $S^{\mathbf{x}}(t,y,n,p)$ denotes the equilibrium expected present value of the customer surplus at time t given the knowledge of y,n and the price p used by the company *in the previous* decision period and *before* the current period's price is observed. The following proposition describes a recursion for calculating the equilibrium customer surplus and the resulting purchase probability. We provide proofs of this and the other propositions in this section in online appendix D.

Proposition 1. *Suppose that $E[|B(t)|]$ and $E[|\tilde{p}(t)|]$ are finite for given parameter values $\mathbf{x}$. Then a subgame-perfect equilibrium in the game between the customers exists. The expected payoffs of all customers in information states at time t before experiencing price values at time t and given the observed price at time $t-1$ are identical and uniquely determined by recursion*

$$\begin{aligned} S^{\mathbf{x}}(t,y,n,p) = E_{\tilde{p}(t)|\tilde{p}(t-1)=p,\mathbf{x}}\Big\{ & E_{B(t)|\mathbf{x}}\big[\bar{\lambda}\big(B(t)-\tilde{p}(t)-\beta S^{\mathbf{x}}(t+1,y,n,\tilde{p}(t))\big)^{+}\big] \\ & +\beta\Big((n-1)\lambda^{\mathbf{x}}(t,n,y,\tilde{p}(t))\big(S^{\mathbf{x}}(t+1,y-1,n-1,\tilde{p}(t))-S^{\mathbf{x}}(t+1,y,n,\tilde{p}(t))\big) \\ & +S^{\mathbf{x}}(t+1,y,n,\tilde{p}(t))\Big)\Big\},\ n-y=N-Y,\ y\in\{1,\dots,Y\},\ t\in\{1,\dots,T-1\},\ p\in\Pi \end{aligned} \quad (9)$$

with the terminal conditions

$$S^{\mathbf{x}}(T,y,n,p) = 0,\ n-y=N-Y,\ y\in\{1,\dots,Y\},\ p\in\Pi, \quad (10)$$

$$S^{\mathbf{x}}(t,0,N-Y,p) = 0,\ t\in\{1,\dots,T-1\},\ p\in\Pi. \quad (11)$$

The corresponding equilibrium strategies (identical for all customers) are given by

$$\lambda^{\mathbf{x}}(t,y,n,p)=\bar{\lambda}P^{\mathbf{x}}(B(t)-p\geq\beta S^{\mathbf{x}}(t+1,y,n,p)). \quad (12)$$

The interpretation of this result is clear: a customer will be eager to purchase whenever the valuation/price difference is greater than or equal to the discounted expected surplus for purchasing an item in the future, given that the inventory and the market size do not change. Since the purchase probability by the customer is proportional to his/her average eagerness, we get a relation of the form (12). Note that (12) generalizes the model with myopic consumers ($\beta=0$), who attempt a purchase whenever $b-p\geq 0$, to the case of general β by means of the adjustment term $\beta S^{\mathbf{x}}(t+1,y,n,p)$.

The corollary stated next is important for application of the consumer choice model to the pricing problem of the company. The corollary follows directly from (12) by summing purchase probabilities of all remaining customers:

Corollary 1. *Assuming that the customers behave according to their equilibrium strategies, the sale probability $\Lambda^{\mathbf{x}}(t,y,p)$ can be computed as (7).*

The above corollary shows that, under certain simplifying assumptions, the probability of sale in a given decision period is a function of time, inventory level and the announced price. Moreover, the corollary describes the structure of this functional dependence. While Corollary 1 offers a useful insight, this structure needs further study. One of the issues is dependence of $\Lambda^{\mathbf{x}}(t,y,p)$ on price p. In the model with myopic customers, which is obtained by taking $\beta = 0$, $\Lambda^{\mathbf{x}}(t,y,p) = \bar{\lambda}(N-(Y-y))P^{\mathbf{x}}(B(t) \geq p)$ is a non-increasing function of p. While this is expected to be true in demand models for most (non-Giffin) goods, the non-increasing property is not immediately obvious in our model for general β. Thus, we need to analyze the expected surplus $S^{\mathbf{x}}(t+1,y,n,p)$ resulting from a subgame-perfect equilibrium in the stochastic dynamic game between customers who assume that the future price follows an exogenous Markov process $\tilde{p}(t)$. In the discussion to follow, we consider general values of y and n, not necessarily those appearing in a realization of a sales process starting with fixed Y and N.

The following proposition can be interpreted as follows: at any time t, the expected customer surplus is smaller when a sale has just occurred than when it has not, since customer competition for remaining items increases after a sale occurs.

Proposition 2. *For all t,y,n,p, we have $S^{\mathbf{x}}(t,y-1,n-1,p) \leq S^{\mathbf{x}}(t,y,n,p)$.*

The next proposition shows that the surplus is a non-increasing function in the number of remaining customers. Again, this result is natural since competition increases for the same inventory when the number of customers is larger:

Proposition 3. *For all t,y,n,p, we have $S^{\mathbf{x}}(t,y,n,p) \leq S^{\mathbf{x}}(t,y,n-1,p)$.*

Propositions 2 and 3 immediately imply that, for a fixed population of customers, the surplus is a non-decreasing function of the remaining inventory y (also natural since smaller inventory for the same number of customers creates higher competition):

Corollary 2. *For all t,y,n,p, we have $S^{\mathbf{x}}(t,y-1,n,p) \leq S^{\mathbf{x}}(t,y,n,p)$.*

The next proposition provides a useful bound on the difference in expected customer surpluses corresponding to different observed prices and leads directly to a proof of monotonicity of purchase probability with respect to price. The formulation uses the following notion of stochastic order (see, for example Shaked and Shanthikumar (1994)): a random variable X is said to be stochastically smaller than a random variable Y (denoted $X \leq_{\mathrm{st}} Y$) if $P(X > u) \leq P(Y > u)$ for all $u \in \mathbb{R}$. A characterization of the stochastic order relation $X \leq_{\text{st}} Y$ is that for all increasing functions $\phi(\cdot)$, $E[\phi(X)] \leq E[\phi(Y)]$, provided that expectations exist. The bounding relation is defined in terms of a sequence of constants κ_t, $t = 0, \dots, T$ such that

$$\kappa_t = \beta(\bar{\lambda} + (1 - \bar{\lambda})\kappa_{t+1}), \tag{13}$$

$$\kappa_T = 0. \tag{14}$$

The sequence is decreasing in t and has the property that $0 \leq \kappa_t < \bar{\kappa}$, where

$$\bar{\kappa} = \frac{\beta\bar{\lambda}}{1 - \beta(1 - \bar{\lambda})} \leq 1.$$

Proposition 4. *Suppose that for all $t > 0$ and $p, p' \in \Pi$ such that $p < p'$, the anticipated future price distribution is such that $[\tilde{p}(t) \,|\, (\tilde{p}(t-1) = p)]$ is stochastically smaller than $[\tilde{p}(t) \,|\, (\tilde{p}(t-1) = p')]$ and that $E[\tilde{p}(t) - p' \,|\, \tilde{p}(t-1) = p', \mathbf{x}] \leq E[\tilde{p}(t) - p \,|\, \tilde{p}(t-1) = p, \mathbf{x}]$ with probability 1. Then, for all t, y, n and $p, p' \in \Pi$ such that $p < p'$, we have*

$$\beta S^{\mathbf{x}}(t, y, n, p) - \beta S^{\mathbf{x}}(t, y, n, p') \leq \kappa_t(p' - p). \tag{15}$$

Since $\kappa_t < 1$, the absolute difference in discounted surpluses corresponding to two different prices is smaller than the difference in these prices. Then, as long as the stochastic order assumption is satisfied, the following important corollary establishing monotonicity of $\Lambda^{\mathbf{x}}(t, y, p)$ in p holds:

Corollary 3. *Under the assumptions of the above proposition, for all t, y, n and $p, p' \in \Pi$ such that $p < p'$, we have*

$$p + \beta S^{\mathbf{x}}(t, y, n, p) < p' + \beta S^{\mathbf{x}}(t, y, n, p'), \text{ and} \tag{16}$$

$$\Lambda^{\mathbf{x}}(t, y, p) \geq \Lambda^{\mathbf{x}}(t, y, p'). \tag{17}$$

The latter inequality is strict if $P^{\mathbf{x}}(p + \beta S^{\mathbf{x}}(t, y, n, p) \leq B(t) < p' + \beta S^{\mathbf{x}}(t, y, n, p')) > 0$.

The assumptions of Proposition 4 concerning stochastic properties of $\tilde{p}(t)$ are reasonable restrictions on the customer perception of the policy and can be interpreted as follows: when $p < p'$, $[\tilde{p}(t) \,|\, (\tilde{p}(t-1) = p)]$ is stochastically smaller than $[\tilde{p}(t) \,|\, (\tilde{p}(t-1) = p')]$, since a customer assumes that the future price tends to be higher if the past price is higher. Also, when $p < p'$, $E[\tilde{p}(t) - p' \,|\, \tilde{p}(t-1) = p', \mathbf{x}] \leq E[\tilde{p}(t) - p \,|\, \tilde{p}(t-1) = p, \mathbf{x}]$ with probability 1 since a customer expects the difference in the future and past prices to be lower when the past price is higher. These assumptions are satisfied both when $\tilde{p}(t)$ is a random walk on equally spaced prices (see the online appendix B for details), and when $\tilde{p}(t)$ is independent of $\tilde{p}(t-1)$.

The next result shows that the surplus is a non-increasing function of time when the valuation distribution is stationary over time. This result is natural since shorter remaining time implies fewer purchase opportunities for the customers.

Proposition 5. *Suppose that for all t and $p \in \Pi$, the distributions of the valuation $B(t)$ and the anticipated future price $[\tilde{p}(t+1) \mid \tilde{p}(t) = p]$ are independent of t (stationary). Then, for all t, y, n and $p \in \Pi$, we have $S^{\mathbf{x}}(t,y,n,p) \geq S^{\mathbf{x}}(t+1,y,n,p)$.*

We point out that it is not necessary to accept all of the assumptions of the consumer choice model described above. Instead, one can assume that the simple decision rule given by (12) is a plausible approximation to the behavior of risk-neutral consumers and that the structural results given above are reasonable assumptions about the properties of the expected surplus.

Online Appendix D: proofs of structural results for the consumer choice model

Proof of Proposition 1

The proof is by reverse induction on t. Suppose that the statement holds for $t+1$ and subsequent decision periods. Given the observed price p, suppose the valuation is equal to b. Let the expected present value of surplus for a customer be $V^{\mathbf{x}}(t,y,n,p,b)$. The expected present value of the customer's surplus at time $t+1$ for given y, n *before* experiencing the price used by the company in step $t+1$ (which a consumer anticipates to be $\tilde{p}(t+1)$) averaged over the valuation $B(t+1)$ is

$$S^{\mathbf{x}}(t+1,y,n,p) = E_{\tilde{p}(t+1),B(t+1)|\tilde{p}(t)=p,\mathbf{x}}[V^{\mathbf{x}}(t+1,y,n,\tilde{p}(t+1),B(t+1))].$$

We next describe the customer decision at time t. The quantity $V^{\mathbf{x}}(t,y,n,p,b)$ is computed as the expected present value over all possible transitions (sales to one of n remaining customers) in the sales process:

$$\begin{aligned} V^{\mathbf{x}}(t,y,n,p,b) = \max_{0 \leq e \leq 1} \Big\{ & \bar{\lambda} e(b-p) + \beta(n-1)\lambda^{\mathbf{x}}(t,y,n,p)S^{\mathbf{x}}(t+1,y-1,n-1,p) \\ & + \beta\Big(1 - \bar{\lambda}e - (n-1)\lambda^{\mathbf{x}}(t,y,n,p)\Big)S^{\mathbf{x}}(t+1,y,n,p)\Big\}. \end{aligned}$$

After collecting terms, this expression can be rewritten as

$$\begin{aligned} V^{\mathbf{x}}(t,y,n,p,b) = & \max_{0 \leq e \leq 1} \{\bar{\lambda} e(b - p - \beta S^{\mathbf{x}}(t+1,y,n,p))\} \\ & + \beta\Big((n-1)\lambda^{\mathbf{x}}(t,y,n,p)\big(S^{\mathbf{x}}(t+1,y-1,n-1,p) - S^{\mathbf{x}}(t+1,y,n,p)\big) + S^{\mathbf{x}}(t+1,y,n,p)\Big). \end{aligned}$$

Recall that the customer eagerness to purchase is denoted as $e^{\mathbf{x}}(t,y,n,p,b)$. Noting the linearity of the objective in e, we obtain the optimal eagerness

$$e^{\mathbf{x}}(t,y,n,p,b) = I[b - p - \beta S^{\mathbf{x}}(t+1,y,n,p)].$$

After averaging over $B(t)$ we get relation (12) for the customer purchase probability, which is valid for each of the n remaining customers. A recursive relation (9) for $S^{\mathbf{x}}(t,y,n,p)$ is obtained by averaging $V^{\mathbf{x}}(t,y,n,\tilde{p}(t),B(t))$ over $\tilde{p}(t)|\tilde{p}(t-1) = p$ and $B(t)$.

Reformulation of (9)

In the subsequent analysis, we use the following reformulation of (9):

$$S^{\mathbf{x}}(t,y,n,p)=E_{\tilde{p}(t),B(t)|\tilde{p}(t-1)=p,\mathbf{x}}[s^{\mathbf{x}}(t,y,n,\tilde{p}(t),B(t))], \tag{18}$$

where

$$\begin{aligned}&s^{\mathbf{x}}(t,y,n,p,b)=\bar{\lambda}\big(b-p-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)^{+}\\ &+\bar{\lambda}(n-1)I[b\geq p+\beta S^{\mathbf{x}}(t+1,y,n,p)]\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)+\beta S^{\mathbf{x}}(t+1,y,n,p).\end{aligned} \tag{19}$$

Technical lemma used in the proof of Proposition 2

LEMMA 1. *For any t and p, if the inequality $S^{\mathbf{x}}(t+1,y-1,n-1,p)\leq S^{\mathbf{x}}(t+1,y,n,p)$ holds for all y,n then $s^{\mathbf{x}}(t,y-1,n-1,p,b)\leq s^{\mathbf{x}}(t,y,n,p,b)$ holds for all y,n,b.*

Proof. Let y,n be arbitrary and consider the following three cases split according to the possible ranges of b.

Case 1: $b<p+\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)$. Then

$$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t,y-1,n-1,p,b)=\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\geq 0.$$

Case 2: $p+\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\leq b<p+\beta S^{\mathbf{x}}(t+1,y,n,p)$. Then

$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t,y-1,n-1,p,b)$

$$\begin{aligned}&=\beta S^{\mathbf{x}}(t+1,y,n,p)-\bar{\lambda}\big(b-p-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\big)\\ &\quad-\bar{\lambda}(n-2)\big(\beta S^{\mathbf{x}}(t+1,y-2,n-2,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\big)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\end{aligned}$$

where we use $b<p+\beta S^{\mathbf{x}}(t+1,y,n,p)$ to obtain

$$\begin{aligned}&>\beta S^{\mathbf{x}}(t+1,y,n,p)-\bar{\lambda}\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\big)\\ &\quad-\bar{\lambda}(n-2)\big(\beta S^{\mathbf{x}}(t+1,y-2,n-2,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\big)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\end{aligned}$$

and, rearranging the terms, finally get

$$\begin{aligned}&=(1-\bar{\lambda})\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)\big)\\ &\quad+\bar{\lambda}(n-2)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y-2,n-2,p)\big)\geq 0.\end{aligned}$$

Case 3: $b\geq p+\beta S^{\mathbf{x}}(t+1,y,n,p)$. Then

$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t,y-1,n-1,p,b)$

$$=\bar{\lambda}\big(b-p-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)$$

$$\begin{aligned}
&\quad + \bar{\lambda}(n-1)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big) + \beta S^{\mathsf{x}}(t+1,y,n,p)\\
&\quad - \bar{\lambda}\big(b-p-\beta S^{\mathsf{x}}(t+1,y-1,n-1,p)\big)\\
&\quad - \bar{\lambda}(n-2)\big(\beta S^{\mathsf{x}}(t+1,y-2,n-2,p) - \beta S^{\mathsf{x}}(t+1,y-1,n-1,p)\big) - \beta S^{\mathsf{x}}(t+1,y-1,n-1,p)\\
&= (1-\bar{\lambda}n)\big(\beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y-1,n-1,p)\big)\\
&\quad + \bar{\lambda}(n-2)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y-2,n-2,p)\big) \geq 0.
\end{aligned}$$

Proof of Proposition 2

The statement is obtained by inverse induction on t. The basis of induction (at $t = T$) is the boundary conditions. Each induction step is an immediate application of Lemma 1 and equation (18) for all y,n,p.

Technical lemma used in the proof of Proposition 3

LEMMA 2. *For any t and p, if the inequalities $S^{\mathsf{x}}(t+1,y-1,n-1,p) \leq S^{\mathsf{x}}(t+1,y,n,p)$ and $S^{\mathsf{x}}(t+1,y,n,p) \leq S^{\mathsf{x}}(t+1,y,n-1,p)$ hold for all y,n then $s^{\mathsf{x}}(t,y,n,p,b) \leq s^{\mathsf{x}}(t,y,n-1,p,b)$ holds for all y,n,b.*

Proof. Let y,n be arbitrary and consider the following three cases.

Case 1: $b < p + \beta S^{\mathsf{x}}(t+1,y,n,p)$. Then,

$$s^{\mathsf{x}}(t,y,n,p,b) - s^{\mathsf{x}}(t,y,n-1,p,b) = \beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p) \leq 0.$$

Case 2: $p + \beta S^{\mathsf{x}}(t+1,y,n,p) \leq b < p + \beta S^{\mathsf{x}}(t+1,y,n-1,p)$. Then
$s^{\mathsf{x}}(t,y,n,p,b) - s^{\mathsf{x}}(t,y,n-1,p,b)$

$$\begin{aligned}
&= \bar{\lambda}\big(b-p-\beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\quad + \bar{\lambda}(n-1)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\quad + \beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)
\end{aligned}$$

where we use $b < p + \beta S^{\mathsf{x}}(t+1,y,n-1,p)$ to obtain

$$\begin{aligned}
&< \bar{\lambda}\big(\beta S^{\mathsf{x}}(t+1,y,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\quad + \bar{\lambda}(n-1)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\quad + \beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\\
&= (1-\bar{\lambda})\big(\beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\big)\\
&\quad + \bar{\lambda}(n-1)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big) \leq 0,
\end{aligned}$$

since both terms in the last expression are nonpositive.

Case 3: $b \geq p + \beta S^{\mathsf{x}}(t+1, y, n-1, p)$. Then

$$
\begin{aligned}
&s^{\mathsf{x}}(t,y,n,p,b) - s^{\mathsf{x}}(t,y,n-1,p,b)\\
&\quad = \bar{\lambda}\big(b - p - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\qquad + \bar{\lambda}(n-1)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big) + \beta S^{\mathsf{x}}(t+1,y,n,p)\\
&\qquad - \bar{\lambda}\big(b - p - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\big)\\
&\qquad - \bar{\lambda}(n-2)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-2,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\big) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\\
&\quad = \bar{\lambda}(n-2)\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y-1,n-2,p)\big)\\
&\qquad + \bar{\lambda}\big(\beta S^{\mathsf{x}}(t+1,y-1,n-1,p) - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)\\
&\qquad + (1 - \bar{\lambda}(n-1))\big(\beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n-1,p)\big) \leq 0.
\end{aligned}
$$

Proof of Proposition 3

The statement is obtained by inverse induction on t. The basis of induction (at $t = T$) is the boundary conditions. Each induction step is an immediate application of Proposition 2, Lemma 2 and equation (18) for all y, n, p.

Technical lemma used in the proof of Proposition 4

Lemma 3. *For any t and any $p, p' \in \Pi$ such that $p < p'$, if the inequalities $S^{\mathsf{x}}(t+1, y-1, n-1, p) \leq S^{\mathsf{x}}(t+1, y, n, p)$ and*

$$\beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n,p') \leq \kappa_{t+1}(p'-p)$$

hold for all y, n then

$$\beta s^{\mathsf{x}}(t,y,n,p,b) - \beta s^{\mathsf{x}}(t,y,n,p',b) \leq \kappa_t(p'-p) \tag{20}$$

holds for all y, n, b.

Proof. Let y, n be arbitrary and consider the following four cases.

Case 1: $b < p + \beta S^{\mathsf{x}}(t+1, y, n, p)$ and $b < p' + \beta S^{\mathsf{x}}(t+1, y, n, p')$. Then,

$$s^{\mathsf{x}}(t,y,n,p,b) - s^{\mathsf{x}}(t,y,n,p',b) = \beta S^{\mathsf{x}}(t+1,y,n,p) - \beta S^{\mathsf{x}}(t+1,y,n,p') \leq \kappa_{t+1}(p'-p).$$

Since $p < p'$ and $\kappa_t \geq \beta\kappa_{t+1}$, it follows that $\beta s^{\mathsf{x}}(t,y,n,p,b) - \beta s^{\mathsf{x}}(t,y,n,p',b) \leq \kappa_t(p'-p)$.

Case 2: $b < p + \beta S^{\mathsf{x}}(t+1, y, n, p)$ and $b \geq p' + \beta S^{\mathsf{x}}(t+1, y, n, p')$. Then we have $p + \beta S^{\mathsf{x}}(t+1, y, n, p) > p' + \beta S^{\mathsf{x}}(t+1, y, n, p')$, a contradiction. Therefore, this case is impossible.

Case 3: $b \geq p + \beta S^{\mathsf{x}}(t+1, y, n, p)$ and $b < p' + \beta S^{\mathsf{x}}(t+1, y, n, p')$. Then,

$$
\begin{aligned}
&s^{\mathsf{x}}(t,y,n,p,b) - s^{\mathsf{x}}(t,y,n,p',b)\\
&\quad = \bar{\lambda}\big(b - p - \beta S^{\mathsf{x}}(t+1,y,n,p)\big)
\end{aligned}
$$

$$
\begin{aligned}
&+\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\
&+\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y,n,p')
\end{aligned}
$$

where we use $b<p'+\beta S^{\mathbf{x}}(t+1,y,n,p')$ to obtain

$$
\begin{aligned}
&<\bar{\lambda}\big(p'+\beta S^{\mathbf{x}}(t+1,y,n,p')-p-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\
&\quad+\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\
&\quad+\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y,n,p')\\
&=\bar{\lambda}(p'-p)+(1-\bar{\lambda})\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y,n,p')\big)\\
&\quad+\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\
&\le(\bar{\lambda}+(1-\bar{\lambda})\kappa_{t+1})(p'-p),
\end{aligned}
$$

since $\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y,n,p')\le\kappa_{t+1}(p'-p)$. Inequality (20) follows.

Case 4: $b\ge p+\beta S^{\mathbf{x}}(t+1,y,n,p)$ and $b\ge p'+\beta S^{\mathbf{x}}(t+1,y,n,p')$. Then, $s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t,y,n,p',b)$

$$
\begin{aligned}
&=\bar{\lambda}\big(b-p-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\
&\quad+\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)+\beta S^{\mathbf{x}}(t+1,y,n,p)\\
&\quad-\bar{\lambda}\big(b-p'-\beta S^{\mathbf{x}}(t+1,y,n,p')\big)\\
&\quad-\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p')-\beta S^{\mathbf{x}}(t+1,y,n,p')\big)-\beta S^{\mathbf{x}}(t+1,y,n,p')\\
&=\bar{\lambda}(p'-p)+\bar{\lambda}(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y-1,n-1,p')\big)\\
&\quad+(1-\bar{\lambda}n)\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+1,y,n,p')\big)\\
&\le\bar{\lambda}(p'-p)+\bar{\lambda}(n-1)\kappa_{t+1}(p'-p)+(1-\bar{\lambda}n)\kappa_{t+1}(p'-p)\\
&=(\bar{\lambda}+(1-\bar{\lambda})\kappa_{t+1})(p'-p).
\end{aligned}
$$

and (20) follows similarly to Case 3.

Proof of Proposition 4

The statement is obtained by inverse induction on t. The basis for induction (at $t=T$) is obtained immediately since $S^{\mathbf{x}}(T,y,n,p)=S^{\mathbf{x}}(T,y,n,p')=0$ from the boundary conditions. The induction step is based on Lemma 3. Suppose that (15) holds for all y,n,p,p' and time instances $t+1,t+2,\ldots,T$. From Proposition 2 and the induction hypothesis, it follows that conditions of Lemma 3 are satisfied and the expression $\kappa_t p+\beta s^{\mathbf{x}}(t,y,n,p,b)$ is monotone in p for each y,n. It follows that the expression

$$E_{B(t)\,|\,\mathbf{x}}[\kappa_t p+\beta s^{\mathbf{x}}(t,y,n,p,B(t))]$$

is monotone in p for each y,n. Since $\tilde{p}(t)\,|\,\tilde{p}(t-1)=p$ is stochastically smaller than $\tilde{p}(t)\,|\,\tilde{p}(t-1)=p'$, we conclude that

$$E_{\tilde{p}(t),B(t)|\tilde{p}(t-1)=p,\mathbf{x}}[\kappa_t\tilde{p}(t)+\beta s^{\mathbf{x}}(t,y,n,\tilde{p}(t),B(t))]\le E_{\tilde{p}(t),B(t)|\tilde{p}(t-1)=p',\mathbf{x}}[\kappa_t\tilde{p}(t)+\beta s^{\mathbf{x}}(t,y,n,\tilde{p}(t),B(t))]$$

for all y,n,p,p' where $p'>p$. Using the equation (18), we can rewrite this relation as

$$\begin{aligned}\beta S^{\mathbf{x}}(t,y,n,p)-\beta S^{\mathbf{x}}(t,y,n,p')&\le\kappa_t(E_{\tilde{p}(t)|\tilde{p}(t-1)=p',\mathbf{x}}[\tilde{p}(t)]-E_{\tilde{p}(t)|\tilde{p}(t-1)=p,\mathbf{x}}[\tilde{p}(t)])\\&=\kappa_t(p'-p+E_{\tilde{p}(t)|\tilde{p}(t-1)=p',\mathbf{x}}[\tilde{p}(t)-p']-E_{\tilde{p}(t)|\tilde{p}(t-1)=p',\mathbf{x}}[\tilde{p}(t)-p])\\&\le\kappa_t(p'-p).\end{aligned}$$

Test of assumptions of Proposition 4 when $\tilde{p}(t)$ is a random walk

Without loss of generality, let the separation between prices be 1 unit. We first observe that $[\tilde{p}(t)\,|\,(\tilde{p}(t-1)=p)]$ is stochastically smaller than $[\tilde{p}(t)\,|\,(\tilde{p}(t-1)=p')]$ for all $p'>p$, and $E[\tilde{p}(t)-p\,|\,\tilde{p}(t-1)=p,\mathbf{x}]$ has the same value for all internal $p\in\Pi$. It only remains to check the second assumption when $p'=\max\Pi$ or $p=\min\Pi$. Indeed, if $p'=\max\Pi$ then

$$E[\tilde{p}(t)-p'\,|\,\tilde{p}(t-1)=p',\mathbf{x}]=(1-q)\cdot 0+q\cdot(-1)=-q.$$

For $\min\Pi<p<\max\Pi$ we have

$$E[\tilde{p}(t)-p\,|\,\tilde{p}(t-1)=p,\mathbf{x}]=r\cdot 1+(1-q-r)\cdot 0+q\cdot(-1)=r-q$$

while for $p=\min\Pi$ we have

$$E[\tilde{p}(t)-p\,|\,\tilde{p}(t-1)=p,\mathbf{x}]=r\cdot 1+(1-r)\cdot 0=r.$$

We see that $E[\tilde{p}(t)-p'\,|\,\tilde{p}(t-1)=p',\mathbf{x}]\le E[\tilde{p}(t)-p\,|\,\tilde{p}(t-1)=p,\mathbf{x}]$ holds with probability 1.

Technical lemma used in the proof of Proposition 5

Lemma 4. *For any $0<t<T-1$ and $p\in\Pi$, if the inequalities $S^{\mathbf{x}}(t+1,y,n,p)\ge S^{\mathbf{x}}(t+2,y,n,p)$ and $S^{\mathbf{x}}(t+2,y-1,n-1,p)\le\beta S^{\mathbf{x}}(t+2,y,n,p)$ hold for all y,n then $s^{\mathbf{x}}(t,y,n,p,b)\ge s^{\mathbf{x}}(t+1,y,n,p,b)$ holds for all y,n,b.*

Proof. Consider arbitrary y,n and observe that $S^{\mathbf{x}}(t+1,y,n,p)\ge S^{\mathbf{x}}(t+2,y,n,p)$ implies that there are three possible cases in terms of possible ranges of b.

Case 1: $b<p+\beta S^{\mathbf{x}}(t+2,y,n,p)$. In this case,

$$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t+1,y,n,p,b)=\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\ge 0.$$

Case 2: $p+\beta S^{\mathbf{x}}(t+2,y,n,p)\le b<p+\beta S^{\mathbf{x}}(t+1,y,n,p)$. Then,
$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t+1,y,n,p,b)=$

$$\begin{aligned}&=\beta S^{\mathbf{x}}(t+1,y,n,p)-\bar\lambda\big(b-p-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\\&\quad-\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+2,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)-\beta S^{\mathbf{x}}(t+2,y,n,p)\end{aligned}$$

where we use $b<p+\beta S^{\mathbf{x}}(t+1,y,n,p)$ to obtain

$$\begin{aligned}&>\beta S^{\mathbf{x}}(t+1,y,n,p)-\bar\lambda\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\\&\quad-\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+2,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)-\beta S^{\mathbf{x}}(t+2,y,n,p)\\&=(1-\bar\lambda)\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\\&\quad-\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+2,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\ge 0.\end{aligned}$$

Case 3: $b\ge p+\beta S^{\mathbf{x}}(t+1,y,n,p)$. Then,
$s^{\mathbf{x}}(t,y,n,p,b)-s^{\mathbf{x}}(t+1,y,n,p,b)$

$$\begin{aligned}&=\bar\lambda\big(b-p-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)\\&\quad+\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+1,y,n,p)\big)+\beta S^{\mathbf{x}}(t+1,y,n,p)\\&\quad-\bar\lambda\big(b-p-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\\&\quad-\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+2,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)-\beta S^{\mathbf{x}}(t+2,y,n,p)\\&=(1-n\bar\lambda)\big(\beta S^{\mathbf{x}}(t+1,y,n,p)-\beta S^{\mathbf{x}}(t+2,y,n,p)\big)\\&\quad+\bar\lambda(n-1)\big(\beta S^{\mathbf{x}}(t+1,y-1,n-1,p)-\beta S^{\mathbf{x}}(t+2,y-1,n-1,p)\big)\ge 0.\end{aligned}$$

Proof of Proposition 5

The proof is by inverse induction on t. The base case ($t=T-1$) is immediate since $S^{\mathbf{x}}(T,y,n,p)=0$ for all y,n,p due to boundary conditions. By induction, we suppose that the statement holds for $t+1,\dots,T-1$ and prove it for t. Consider arbitrary y,n,p and examine the difference

$$\begin{aligned}&S^{\mathbf{x}}(t,y,n,p)-S^{\mathbf{x}}(t+1,y,n,p)\\&=E_{\tilde p(t),B(t)|\tilde p(t-1)=p,\mathbf{x}}[s^{\mathbf{x}}(t,y,n,\tilde p(t),B(t))]-E_{\tilde p(t+1),B(t+1)|\tilde p(t)=p,\mathbf{x}}[s^{\mathbf{x}}(t+1,y,n,\tilde p(t+1),B(t+1))].\end{aligned}$$

Since the distributions of $B(t)$ and $B(t+1)$ as well as of $\tilde p(t+1)\,|\,\tilde p(t)=p$ and $\tilde p(t)\,|\,\tilde p(t-1)=p$ are identical, we can rewrite this difference as

$$\begin{aligned}&S^{\mathbf{x}}(t,y,n,p)-S^{\mathbf{x}}(t+1,y,n,p)\\&\qquad=E_{\tilde p(t),B(t)|\tilde p(t-1)=p,\mathbf{x}}[s^{\mathbf{x}}(t,y,n,\tilde p(t),B(t))-s^{\mathbf{x}}(t+1,y,n,\tilde p(t),B(t))].\end{aligned}$$

The right-hand-side of this equation is nonnegative since, for all realizations of $B(t)$ and $\tilde p(t)$, $s^{\mathbf{x}}(t,y,n,\tilde p(t),B(t))\ge s^{\mathbf{x}}(t+1,y,n,\tilde p(t),B(t))$ (the latter follows from the induction hypothesis, Proposition 2, and Lemma 4).