Dynamic Pricing and Learning with Finite Inventories

Arnoud den Boer¹, Bert Zwart¹,²

¹Centrum Wiskunde & Informatica (CWI)
Science Park 123, 1098 XG Amsterdam, The Netherlands

²VU University Amsterdam, Department of Mathematics
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

November 24, 2011

Abstract

We study a dynamic pricing problem with finite inventory and parametric uncertainty on the demand distribution. Products are sold during selling seasons of finite length, and inventory that is unsold at the end of a selling season, perishes. The goal of the seller is to determine a pricing strategy that maximizes the expected revenue. Inference on the unknown parameters is made by maximum likelihood estimation. We propose a pricing strategy for this problem, and show that the Regret - which is the expected revenue loss due to not using the optimal prices - after \( T \) selling seasons is \( O(\log(T)^2) \). Apart from a small modification, our pricing strategy is a certainty equivalent pricing strategy, which means that at each moment, the price is chosen that is optimal w.r.t. the current parameter estimates. The good performance of our strategy is caused by an ‘endogenous learning’ property: using any pricing policy which is sufficiently close to the optimal one, leads to a.s. convergence of the parameter estimates to the true, unknown parameter.

1 Introduction

The emergence of Internet as a sales channel has made it very easy for companies to experiment with selling prices. Where in the past costs and effort were needed to change prices, for example by issuing a new catalogue or replacing price tags, and consequently prices where fixed for longer periods of time, nowadays a webshop can adapt their prices with a proverbial flick of the switch, without any additional costs or efforts. This flexibility in pricing is one of the main drivers for research on dynamic pricing: the study of determining optimal selling prices under changing circumstances.

A much-studied situation is a firm who sells limited amounts of products during finite selling periods, after which all unsold products perishes. Examples of products with this property are flight tickets, hotel rooms, car rental reservations, and concert tickets. Various dynamic pricing models are already applied in these branches (see Talluri and van Ryzin, 2004). Other products that fall in this framework but for which dynamic pricing is not (yet) commonplace, are newspapers, magazines, and food at a grocery store. The emergence of digital price tags however may change this in the near future, see Kalyanam et al. (2006).
An important insight from the literature on dynamic pricing, is that the optimal selling price of these type of products depends on the remaining inventory and the length of the remaining selling period, see e.g. Gallego and van Ryzin (1994). The optimal price is thus not a single value, but a collection of prices, one for each combination of remaining inventory and remaining length of the selling period. To determine these optimal prices it is essential to know the relation between the demand and the selling price. In most literature from the 90-ies on dynamic pricing, it is assumed that this relation is exactly known to the seller, but in practice exact information on consumer behavior is generally not available. It is therefore not surprising that the review on dynamic pricing by Bitran and Caldentey (2003) mentions dynamic pricing with demand learning as an important future research direction. The presence of digital sales data enables a data-driven approach of dynamic pricing, where the selling firm not only determines optimal prices, but also learns how changing prices affects the demand. Ideally, this learning will eventually lead to optimal pricing decisions.

Since then, a considerable number of studies on this subject have appeared, most of which are reviewed in Araman and Caldentey (2011). We also mention the related studies by Kleinberg and Leighton (2003), Broder and Rusmevichientong (2010), den Boer and Zwart (2010), Harrison et al. (2011a), who consider dynamic pricing in a slightly different setting, namely with infinite inventory. This significantly changes the structure of the learning behavior, as further discussed in Section 4.

A common feature of the studies on dynamic pricing with finite inventory, is the restriction to a single selling season during which learning and optimization takes place. To assess the performance of proposed pricing strategies, one often considers an asymptotic regime where the demand rate and the initial amount of inventory grow to infinity (e.g. Besbes and Zeevi, 2009, Wang et al., 2011). Such an asymptotic regime may have practical value if demand, initial inventory, and the length of the selling season are relatively large. In many situations, however, this is not the case. For example, in the hotel rooms industry (Talluri and van Ryzin, 2004, section 10.2, Weatherford and Kimes, 2003), a product may be modeled as a combination of arrival date and length-of-stay. Different products may have different, overlapping selling periods, and similar demand characteristics. It would therefore be unwise to learn the consumer behavior for each product and selling period separately. In addition, the average demand, initial capacity and length of a selling period may be quite low, which makes the asymptotic regime not a suitable setting to study the performance of pricing strategies.

Motivated by this we investigate in the present study dynamic pricing of perishable products with finite initial inventory, during multiple consecutive selling seasons of finite duration.

We propose a pricing strategy that is structurally very intuitive, and easy to understand by price managers. The demand function is assumed to belong to a parametric family, where the true parameter is unknown to the firm. At every moment where prices can be changed, the firm calculates a statistical estimate of the unknown parameter. Subsequently, the price is determined that would be optimal if this parameter estimate were correct, and this price is used until the next decision moment. In other words, at each decision moment the firm acts as if being certain about the parameter estimates. Only in the last period of a selling season that inventory is positive, a small deviation on this price may be prescribed by our pricing strategy.

This type of strategy for sequential decision problems under uncertainty, is known under different names in the literature: certainty equivalent control, myopic control, passive learning, and the principle of estimation and control. There are problems for which certainty equivalent control is not a good strategy, e.g. the multi-period control problem (Anderson and Taylor, 1976, Lai and Robbins, 1982), and dynamic pricing with infinite inventory (Broder and Rusmevichientong, 2010, Harrison et al., 2011b, den Boer and Zwart, 2010). In these two examples, passive learning is
not sufficient to learn the parameters: the decision maker should actively account for the fact that he is not only optimizing prices, but also tries to 'optimize' the learning process. This implies that sometimes decisions should be taken that seem suboptimal on a short term. In the dynamic pricing problem with infinite inventory, this can be accomplished by the controlled variance policy of den Boer and Zwart (2010) or the MLE-cycle policy of Broder and Rusmevichientong (2010).

In the situation that we study in this article, dynamic pricing with finite inventory and finite selling periods, certainty equivalent control does perform well: the parameter estimates converge with probability one to the correct values, and the prices converge to the optimal prices. The Regret($T$), which measures the expected amount of revenue loss in the first $T$ selling seasons, due to not using the optimal prices, is $O(\log(T)^2)$. This is considerably better than $\sqrt{T}$, which is the best achievable bound on the Regret for the problem with infinite inventory (in different settings, this is shown by Kleinberg and Leighton (2003), Besbes and Zeevi (2011), Broder and Rusmevichientong (2010)).

Thus, the Regret, which can be interpreted as the 'cost for learning', behaves structurally different in these two models. This difference in qualitative behavior can be explained as follows. In the infinite inventory model, prices and parameter estimates can get stuck in what Harrison et al. (2011a) call an 'indeterminate equilibrium'. This means that for some values of the parameter estimates, the expected observed demand at the certainty equivalent price is equal to what the parameter estimates would predict; in other words, the observations confirm the correctness of the (incorrect) parameter estimates. As a result, certainty equivalent control induces insufficient dispersion in the chosen selling prices to eventually learn the true value of the parameters.

Such cannot occur in the setting with finite inventories and finite selling seasons. An optimal price - optimal w.r.t. certain parameter estimates - is namely not a fixed number, but changes depending on the remaining inventory and the remaining length of the selling season. Thus, an optimal policy naturally induces endogenous price dispersion, and prices cannot get stuck in an 'indeterminate equilibrium'. On the contrary, the large amount of price dispersion implies that the unknown parameters are learned quite fast, and consequently that the Regret($T$) is only $O(\log(T)^2)$.

The dynamic pricing problem that we study falls in the framework of adaptive control in Markov Decision Problems (Hernández-Lerma, 1989, Kumar, 1985, chapter 12 of Kumar and Varaiya, 1986). An important feature that distinguishes our work from many previous literature in this area, is the following. Hernández-Lerma and Cavazos-Cadena (1990), Gordienko and Minjárez-Sosa (1998) assume that the "next" state $x_{t+1}$ at period $t+1$ is determined by the "current" state $x_t$, action $a_t$, and a random component $\xi_t$. These random components are assumed to be independent and identically distributed. In our setting, the randomness in state transitions is completely determined by the demand realizations. These are neither identically distributed (their distribution depends on the chosen prices), nor independent (chosen prices may depend on all previously chosen prices and observed demand realizations, and consequently, demand in different time periods is not independent). In other literature, such as Altman and Shwartz (1991), unknown transition probabilities are estimated by the empirically observed relative frequencies. In our setting, all uncertainty is captured by a single unknown parameter, and transition probabilities are estimated simultaneously. Furthermore, we consider a compact continuous action space, in contrast to e.g. Burnetas and Katehakis (1997), Chang et al. (2005) who assume a finite action space, which links the adaptive control problem to the multi-armed bandit problem.

Summarizing, the contributions of this paper are as follows:

(i) We formulate the problem of dynamic pricing with finite inventories during multiple, consecutive selling seasons of finite duration, with parametric uncertainty in the demand func-
(ii) We propose a simple and intuitive pricing strategy, based on the idea of subsequently estimating the unknown parameters and choosing the selling price that would be optimal if this parameter estimate were correct.

(iii) We show that the problem satisfies an 'endogenous learning' property, which means that the use of policies that are optimal w.r.t. parameter estimates, leads to online learning of the true, unknown parameter.

(iv) We prove that this pricing strategy implies convergence of the parameter estimates to the true value, and we show \( \text{Regret}(T) = O(\log(T)^2) \).

(v) We provide a numerical example to illustrate our results.

The rest of this paper is organized as follows: Section 2 introduces the mathematical model, and discusses the optimal pricing policy, estimation methods, and the definition of regret. We introduce our pricing strategy in Section 3, and discuss the endogenous-learning property in Section 4. In Section 5 we prove that our pricing strategy has \( \text{Regret}(T) = O(\log(T)^2) \). Section 6 contains a numerical illustration. A discussion of the results and possible extensions of this paper is provided in Section 7. The Appendix contains proofs of two auxiliary lemmas.

Notation The interior of a set \( U \subset \mathbb{R}^n \) is denoted by \( \text{int}(U) \). If \( v \) is a vector, \( ||v|| \) denotes the Euclidean norm. If \( A \) is an \( m \times n \) matrix, \( ||A|| = \max_{x \in \mathbb{R}^n, ||x||=1} ||Ax|| \) denotes the induced matrix norm of \( A \), and \( \lambda_{\min}(A) \) denotes the smallest eigenvalue of \( A \). For \( x \in \mathbb{R}_+ \), \( \lfloor x \rfloor \) denotes the largest integer which is smaller than or equal to \( x \).

2 Model, full-information solution, and parameter estimation

Model formulation. We consider a monopolist seller of perishable products which are sold during consecutive selling seasons. Each selling season consists of \( S \in \mathbb{N} \) discrete time periods. The \( i \)-th selling season starts at period \( 1 + (i-1)S \), and last until period \( iS \), for \( i \in \mathbb{N} \). For \( t \in \mathbb{N} \), let \( SS_t = 1 + \lfloor (t-1)/S \rfloor \) denote the selling season corresponding to period \( t \). The inventory or capacity level of the firm at the beginning of period \( t \in \mathbb{N} \) is denoted by \( c_t \). At the start of each selling season, the seller has \( C \in \mathbb{N} \) discrete units of inventory at his disposal; thus \( c_{1+(i-1)S} = C \) for all \( i \in \mathbb{N} \). In each time period \( t \) the seller determines a selling price \( p_t \in \mathcal{P} = [p_l, p_h] \), where \( \mathcal{P} \) is the set of admissible prices, and \( 0 < p_l < p_h \) are the lowest and highest price admissible to the firm. Subsequently the seller observes a realization \( d_t \) of the random variable \( D(p_t) \), which models the demand, and collects revenue \( p_t \cdot d_t \). The inventory level changes from \( c_t \) to \( \max\{c_t - d_t, 0\} \). All inventory which is unsold at the end of the selling season, perishes. The purpose of the seller is to maximize the average expected revenue over an infinite time horizon.

Demand distribution. The demand in a single time period with selling price \( p \), is a realization of the random variable \( D(p) \). We assume that \( D(p) \) is Bernoulli distributed with expectation \( h(\beta_0 + \beta_1 p) \). Here \( z \mapsto h(z) \) is a thrice continuously differentiable function, and \( \beta = (\beta_0, \beta_1) \in \mathbb{R}^2 \) are unknown parameters. The true value of \( \beta \) is denoted by \( \beta^{(0)} \), and is unknown to the seller. It is assumed to lie in the interior of a compact known set \( B \subset \mathbb{R}^2 \), which is called the parameter set. If we wish to emphasize the dependence of the demand on the unknown parameters \( \beta \), we write
To assure existence and uniqueness of revenue-maximizing selling prices, we make a number of additional assumptions on $h$. In particular, we assume that $h(z)$ is log-concave, $h(\beta_0 + \beta_1 p) \in (0, 1)$ for all $\beta \in B$, $p \in P$, and the derivative $h'(z)$ of $h(z)$ is strictly positive. This latter assumption implies that $E[D(p, \beta)]$ is decreasing in $p$, for all $\beta \in B$.

Write $r_{1,S}^{*}[\beta^{(0)}] = \max_{p \in P} r(p, \beta^{(0)})$, and

$$g_{a,\beta}(p) = -(p - a)\beta_1 \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)}, \quad ((a, p) \in \mathbb{R} \times B \times P).$$

We assume that $g_{a,\beta^{(0)}}(p_k) < 1$, $g_{a,\beta^{(0)}}(p_k) > 1$, and $g_{a,\beta^{(0)}}(p)$ is strictly increasing in $p$, for all $0 \leq a \leq r_{1,S}^{*}[\beta^{(0)}]$. These conditions, which for $a = 0$ coincide with the assumptions in Lariviére (2006, page 602), are to ensure existence and uniqueness of optimal prices in the interior of $P$.

Examples of functions $h$ that satisfy the assumptions (with appropriate conditions on $B$ and $P$), are $h(z) = \exp(z)$, $h(z) = z$, and $h(z) = \logit(z) = \frac{\exp(z)}{1 + \exp(z)}$.

Full-information optimal solution. If the value of $\beta$ is known, the optimal pricing policy can be determined by solving a Markov Decision Problem (MDP). Since each selling season corresponds to the same MDP, the optimal policy for the infinite-horizon average reward criterion is to repeatedly use the optimal policy for a single selling season. The state space of this MDP is $X = \{(c, s) \mid c = 0, \ldots, C, s = 1, \ldots, S\}$, where $(c, s)$ means that there are $c$ units of remaining inventory at the beginning of the $s$-th period of the selling season. The action space is $P$, and the expected immediate reward function $r((c, s), p) = E[p \cdot \min\{D(p), c\}] = ph(\beta_0 + \beta_1 p)$ if $c \geq 1$, and equals zero if $c = 0$. A (stationary deterministic) policy $\pi \in \Pi := \mathcal{P}^{(C+1) \times S}$.

Given a policy $\pi \in \Pi$, let $V[\pi, \beta](c, s)$ be the expected revenue-to-go function starting in state $(c, s)$ under a policy $\pi \in \Pi$, for $(c, s) \in X$. Then $V[\pi, \beta](c, s)$ satisfies the following recursion:

\begin{align*}
V[\pi, \beta](c, s) &= P(D(\pi(c, s), \beta) = 0) \cdot (V[\pi, \beta](c, s + 1)) \\
&\quad + P(D(\pi(c, s), \beta) = 1) \cdot (V[c, \pi + \beta](c - 1, s + 1)), \quad (c \geq 1), \\
V[\pi, \beta](0, s) &= 0,
\end{align*}

where we write $V[\pi, \beta](c, S + 1) = 0$.

By Proposition 4.4.3 of Puterman (1994), for each $\beta \in B$ there is a $\beta$-optimal policy $\pi[\beta] \in \Pi$. This policy can be calculated using backward induction. We shorthand write $V[\beta](c, s) = V[\pi[\beta], \beta](c, s)$ for the optimal revenue-to-go function. $V[\beta](c, s)$ and $\pi[\beta](c, s)$, for $(c, s) \in X$, can be calculated with the following recursion.

\begin{align*}
V[\beta](c, s) &= \max_{p \in P} (r(p, \beta) - \Delta V[\beta](c, s + 1)h(\beta_0 + \beta_1 p) + V[\beta](c, s + 1)), \\
\pi[\beta](c, s) &\in \arg \max_{p \in P} (r(p, \beta) - \Delta V[\beta](c, s + 1)h(\beta_0 + \beta_1 p)),\quad (3)
\end{align*}

where we define $\Delta V[\beta](c, s) = V[\beta](c, s) - V[\beta](c - 1, s)$ and $\Delta V[\beta](0, s) = 0$ for all $1 \leq s \leq S$.

The optimal average reward of the MDP is equal to $V[\beta](C, 1)$, and the ‘true’ optimal average reward is equal to $V[\beta^{(0)}](C, 1)$.
The score-function, which is the derivative of reward in a certain state, is given by

\[ L_t(\beta) = \sum_{i=1}^{t} \log \left[ \frac{h(\beta_0 + \beta_1 p_i)}{\beta_0 + \beta_1 p_i} \left( 1 - h(\beta_0 + \beta_1 p_i) \right)^{1-d_i} \right]. \]

The score-function, which is the derivative of \( L_t(\beta) \) w.r.t. \( \beta \), equals

\[ l_t(\beta) = \sum_{i=1}^{t} \frac{h(\beta_0 + \beta_1 p_i)}{h(\beta_0 + \beta_1 p_i)(1 - h(\beta_0 + \beta_1 p_i))} \left( \frac{1}{p_i} \right) (d_i - h(\beta_0 + \beta_1 p_i)). \quad (4) \]

We let \( \hat{\beta}_t \) be a solution to \( l_t(\beta) = 0 \). If no solution exists, we define \( \hat{\beta}_t = \beta^{(1)} \), for some predefined \( \beta^{(1)} \in B \). If a solution to \( l_t(\beta) = 0 \) exists but lies outside \( B \), we define \( \hat{\beta}_t \) as the projection of this solution on \( B \). In Proposition 1 we show that eventually, for \( t \) sufficiently large, \( \hat{\beta}_t \) exists in a neighborhood of the true value \( \beta^{(0)} \).

**Pricing strategy.** At each time period \( t \in \mathbb{N} \), the seller needs to determine a policy \( \pi_t \in \Pi \). This policy may depend on all past demand realizations \( d_1, \ldots, d_{t-1} \), chosen policies \( \pi_1, \ldots, \pi_{t-1} \), and visited states \( (c_1, s_1), \ldots, (c_{t-1}, s_{t-1}) \). More formally, if \( \mathcal{F}_t = \sigma(\pi_1, \ldots, \pi_t, d_1, \ldots, d_t, (c_1, s_1), \ldots, (c_t, s_t)) \) denotes the \( \sigma \)-algebra generated by the policies \( \pi_1, \ldots, \pi_t \), demand realizations \( d_1, \ldots, d_t \), and states \( (c_1, s_1), \ldots, (c_t, s_t) \) up to and including time \( t \), then we require \( \pi_t \in \mathcal{F}_{t-1} \). The sequence \( \{\pi_t\}_{t \in \mathbb{N}} \) is called a pricing strategy. Using this strategy means that in each period \( t \) the price \( p_t = \pi_t(c_t, s_t) \) is used. The purpose of the seller is to find a pricing strategy that maximizes the expected revenue over an infinite time horizon.

To assess the quality of a pricing strategy \( \psi \), we define the Regret. This quantity measures the expected amount of money which is lost due to not using the optimal prices. The Regret of \( \psi \) after the first \( T \) selling seasons is defined as

\[ \text{Regret}(\psi, T) = E \left[ \sum_{t=1}^{T} V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_{1+(i-1)S}, \ldots, \pi_t S, \beta^{(0)}](C, 1) \right]. \quad (5) \]

Here \( V[\pi_{k+S}, \beta^{(0)}](C, 1) \) denotes the expected reward of the policy that consecutively uses the actions from \( \pi_{k+1}, \ldots, \pi_{k+S} \in \Pi \), for \( k \in \mathbb{N} \); i.e. in state \((c, s)\), action \( \pi_{k+S}(c, s) \) is used. The expectation operator in (5) is because \( \pi_{1+(i-1)S}, \ldots, \pi_t S \) may be random variables. Observe that the regret cannot directly be used by the seller to find the optimal strategy, since this depends on \( \beta^{(0)} \), which is unknown to the seller. Also note that we calculate the regret per selling period, and not per time period \( t \in \mathbb{N} \). The reason is that the policy \( \pi[\beta^{(0)}] \) is optimized over a whole selling season, and not for each individual state; a chosen price \( p_t \) may induce a higher instant reward in a certain state \((c_t, s_t)\) than the optimal price \( \pi[\beta^{(0)}](c_t, s_t) \). This effect is averaged out by looking at the optimal expected reward in a complete selling season, i.e. at \( V[\pi, \beta^{(0)}](C, 1) \).

## 3 Pricing strategy

We propose a pricing strategy based on the following principle: in each period, estimate the unknown parameters with maximum likelihood estimation, and subsequently use the action from
the policy that is optimal w.r.t. this estimate. Let \((e_t, s_t)_{t \in \mathbb{N}}\) denote the state process and \((p_t)_{t \in \mathbb{N}}\) the price sequence. Our proposed pricing strategy is as follows:

\[
\text{Pricing strategy } \Phi(\epsilon).
\]

Choose \(0 < \epsilon < \frac{1}{2}(p_h - p_l)\), and initial prices \(p_1, p_2 \in \mathcal{P}\), with \(p_1 \neq p_2\).

For all \(t \geq 2:\)

1) Determine \(\hat{\beta}_t\), and let \(p_{\text{ceqp}} = \pi[\hat{\beta}_t]\).

IIa) If \(\max\{|p_t - p_{\text{ceqp}}| : i : SS_i = SS_t+1\} < \epsilon\), and \(c_{t+1} = 1\) or \(s_{t+1} = S\), choose \(p_{t+1} \in \{p_{\text{ceqp}} + 2\epsilon, p_{\text{ceqp}} - 2\epsilon\} \cap \mathcal{P}\), and \(\pi_{t+1}(c, s) = \pi_{\text{ceqp}} 1(c, s) \neq (c_{t+1}, s_{t+1}) + p_{t+1} 1(c, s) = (c_{t+1}, s_{t+1})\).

IIb) Else, set \(p_{t+1} = p_{\text{ceqp}}, \pi_{t+1} = \pi_{\text{ceqp}}\).

The pricing strategy \(\Phi(\epsilon)\) always chooses the price \(p_{\text{ceqp}}\), except possibly when \((e_{t+1}, s_{t+1})\) is in \(\{(c, s) : c = 1\text{ or } s = S\}\). This set contains all states that, with positive probability, are the last states in the selling season in which products are sold (either because the selling season has finished, or because the inventory is depleted). In these states, the price \(p_{t+1}\) deviates from \(p_{\text{ceqp}}\) if the variation in the other selling prices from the same selling season, measured by \(\max\{|p_t - p_{\text{ceqp}}| : i : SS_i = SS_t+1\}\), is smaller than \(\epsilon\). This deviation guarantees a minimum amount of price dispersion, during a selling season.

In the proof of Theorem 1, we show that for sufficiently small \(\epsilon\), IIa occurs only finitely many times. Then for sufficiently large \(t\), the pricing strategy \(\Phi(\epsilon)\) acts as a ‘passive learning’ or ‘certainty equivalent’ pricing strategy. The pricing decisions in IIb are driven by optimizing instant revenue, and do not reckon with the objective of optimizing the quality of the parameter estimates \(\hat{\beta}_t\). It turns out that learning the parameter values happens on the fly, without active effort.

An important reason why this is the case, is the presence of endogenous learning in \(\beta\)-optimal policies: if actions from \(\pi[\beta]\) are used, for \(\beta\) sufficiently close to \(\beta(0)\), the estimate \(\hat{\beta}_t\) will a.s. converge to \(\beta(0)\). This is elaborated in Section 4.

Our main result about the pricing policy \(\Phi(\epsilon)\) is the following upper bound on the Regret.

**Theorem 1.** There is an \(\epsilon^* > 0\), such that \(\text{Regret}(\Phi(\epsilon), T) = O((\log(T))^2)\), provided \(\epsilon < \epsilon^*\).

The proof is contained in Section 5.

### 4 Endogenous learning of \(\beta\)-optimal policies

The amount of learning after \(t \in \mathbb{N}\) periods can be quantified by the smallest eigenvalue of the design matrix \(P_t\), which is defined as

\[
P_t = \sum_{i=1}^{t} \left( \begin{array}{c} 1 \\ p_i \end{array} \right) (1, p_i).
\]

The justification for this measure comes from den Boer and Zwart (2011), where bounds on \(E\left[\|\hat{\beta}_t - \beta(0)\|^2\right]\) are obtained in terms of (a non-random lower bound on) \(\lambda_{\min}(P_t)\).
It is not difficult to derive that $P_t$ is positive semi-definite, and positive definite if $p_1, \ldots, p_t$ are not all equal. Moreover, $\lambda_{\min}(P_t) \geq \lambda_{\min}(P_r) + \lambda_{\min}(P_s)$ for all integers $r$ and $s$ such that $r + s = t$ (Bhatia, 1997, Corollary III.2.2, page 63).

The main result of this section is that if the prices from a $\beta$-optimal policy are used during a selling season $k \in \mathbb{N}$, $(k > 1)$, and $\beta$ is sufficiently close to $\beta^{(0)}$, then $\lambda_{\min}(P_{kS}) - \lambda_{\min}(P_{(k-1)S}) \geq w_0 \text{ a.s.}$ for some non-random $w_0 > 0$ independent of $k$ and $\beta$. Using such a $\beta$-optimal policy thus strictly increases the amount of learning.

First, notice the following. The continuity of $g$ and $\dot{g}$ implies that the assumptions $g_{a,\beta}(p_{\beta}) < 1$, $g_{a,\beta}(h_{\beta}) > 1$ and $g_{a,\beta}(p)$ strictly increasing in $p$, do not only hold for $(a, \beta) \in [0, r_{1,S}^{(0)}] \times \{\beta^{(0)}\}$, but for $(a, \beta) \in U_a \times U_\beta$, where $U_a \subset \mathbb{R}$ is an open neighborhood containing the interval $[0, r_{1,S}^{(0)}]$, and $U_\beta \subset B$ is an open neighborhood containing $\beta^{(0)}$. We choose $U_a, U_\beta$ such that

$$\sup_{\beta \in U_\beta} \max_{p \in P} r(p, \beta) \in U_a. \quad (7)$$

**Theorem 2.** Let $1 < C < S$, $\beta \in U_\beta$, and suppose that during a selling season $k \in \mathbb{N}$, the prices from $\pi[\beta]$ are used. Then there exists a non-random $w_0 > 0$, independent of $\beta$ and $k$, and an open neighborhood $U_\beta' \subset U_\beta$ of $\beta^{(0)}$, such that $\lambda_{\min}(P_{kS}) - \lambda_{\min}(P_{(k-1)S}) \geq w_0 \text{ a.s.}$ provided $\beta \in U_\beta'$.

In Theorem 2, the requirement $C < S$ is crucial. If $C > S$, then clearly $C - S$ items cannot be sold during the selling season. The selling of the remaining $S$ items can be interpreted as that each item can be sold only in a single period. There is then no interaction between individual items, and the pricing problem is equivalent to the setting with $C = S = 1$. In this case, there is no endogenous learning. On the contrary, the seller needs to actively experiment with selling prices, in order to learn the unknown parameters. A pricing strategy for this setting is elaborated by den Boer and Zwart (2010).

The proof of Theorem 2 makes use of the following two lemmas, which establish structural properties of the optimization problem (3) and its solution $\pi[\beta]$. The proofs of these two lemmas are contained in the Appendix.

**Lemma 1.** For $(a, \beta) \in U_a \times U_\beta$, define the function $f : P \to \mathbb{R}$, $f_{a,\beta}(p) = (p - a)h(\beta_0 + \beta_1p)$. Write $\dot{f}_{a,\beta}(p)$ and $\ddot{f}_{a,\beta}(p)$ for the first and second derivative of $f_{a,\beta}(p)$ w.r.t. $p$, and let $p_{a,\beta}^* = \arg \max_{p \in P} f_{a,\beta}(p)$.

(i) $p_{a,\beta}^*$ is the unique solution to $\dot{f}_{a,\beta}(p) = 0$, lies in the interior of $P$, and in addition satisfies $\ddot{f}_{a,\beta}(p_{a,\beta}^*) < 0$.

(ii) $p_{a,\beta}^*$ is continuously differentiable in $a$ and $\beta$, and strictly increasing in $a$. Furthermore, for all $\beta \in U_\beta$, $f_{a,\beta}(p_{a,\beta}^*)$ is strictly decreasing in $a$.

(iii) There is a $C > 0$ such that for all $(a, \beta) \in U_a \times U_\beta$ and $p \in P$,

$$f_{a,\beta}(p) \leq f_{a,\beta}(p_{a,\beta}^*) - C(p - p_{a,\beta}^*)^2.$$

**Lemma 2.** For each $\beta \in U_\beta$, $\pi[\beta]$ is uniquely defined and continuous in $\beta$.

**Proof of Theorem 2.** Fix $\beta \in U_\beta$. Let $(c_1, 1), (c_2, 2), \ldots, (c_S, S)$ be a path in the state-space, with $c_1 = C$. We show that there are $1 \leq s, s' \leq S$ and constant $v_0 > 0$ independent of $\beta$, such that
\[ |\pi[\beta](c_s, s) - \pi[\beta](c_s', s')| \geq \nu_0. \] Then, writing \( q' = \pi[\beta](c_s, s), q' = \pi[\beta](c_s', s') \),

\[
\lambda_{\min}(P_{kS}) - \lambda_{\min}(P_{(k-1)S}) \\
\geq \lambda_{\min} \left( \left( \begin{array}{c} 1 \\ q \end{array} \right) (1, q) + \left( \begin{array}{c} 1 \\ q' \end{array} \right) (1, q') \right) \\
\geq (1 + p_k^2)^{-1} \cdot 2 \cdot \frac{(q - q')^2}{4} \\
\geq (1 + p_k^2)^{-1} \frac{\nu_0}{2} > 0.
\] (8)

Here the second inequality follows from den Boer and Zwart (2010, Lemma 1) together with the fact that the sample variance of \( \{q, q'\} \) equals \( \frac{(q-q')^2}{4} \).

Define

\[
\Delta = \{(c, s) \mid S + 1 - C \leq s \leq S, S + 1 - s \leq c \leq C \}. \tag{9}
\]

See Figure 1 for an illustration of \( \Delta \) in the state space \( X \). Notice that since \((C, 1) \notin \Delta \) (by the assumption \( C < S \)), the path \((c_s, s)_{1 \leq s \leq S} \) may or may not hit \( \Delta \). We show that in both cases, at least two different selling prices occur on the path \((c_s, s)_{1 \leq s \leq S} \).

**Case 1.** The path \((c_s, s)_{1 \leq s \leq S} \) hits \( \Delta \). Then there is an \( s \) such that \((c_s, s) \in \Delta \) and \((c_s, s-1) \notin \Delta \). In particular, \((c_s, s-1) \in (\mathcal{L}\Delta) = \{(1, S-1), (2, S-2), \ldots, (C-1, S-C+1), (C, S-C)\} \), where \((\mathcal{L}\Delta) \) denotes the points \((c, s)\) immediately left to \( \Delta \) in Figure 1. The sets \( \Delta \) and \((\mathcal{L}\Delta) \) satisfy the following properties:

**P1.** If \((c, s) \in \Delta \) then \( \Delta V[\beta](c, s+1) = 0, \pi[\beta](c, s) = \arg \max_{p \in P} r(p, \beta), \) and

\[ V[\beta](c, s) = (S - s + 1) \cdot V[\beta](1, S). \]

**P2.** If \((c, s) \in (\mathcal{L}\Delta) \), then \( \pi[\beta](c, s) \neq \pi[\beta](c, s+1) \) and \( \Delta V[\beta](c+1, S-c) \neq 0 \) (provided \( c < C \)).

Proof of (P1): Backward induction on \( s \). If \( s = S \) and \((c, s) \notin \Delta \), then the assertions follow immediately. Let \( s < S \). Then \( \Delta V[\beta](c, s+1) = V[\beta](c, s+1) - V[\beta](c-1, s+1) = 0, \pi[\beta](c, s) = \arg \max_{p \in P} r(p, \beta) \) and \( V[\beta](c, s) = \max_{p \in P} r(p, \beta) + V[\beta](c, s+1) = (S - s + 1) \cdot V[\beta](1, S) \), by (3) and the induction hypothesis. This proves (P1).

Proof of (P2). Induction on \( c \). If \( c = 1 \) and \((c, s) \in (\mathcal{L}\Delta) \), then \((c, s) = (1, S-1) \). Since \( \Delta V[\beta](1, S) = V(1, S) > 0 \), it follows from Lemma 1 and (3) that \( \pi[\beta](1, S-1) \neq \pi[\beta](1, S). \) In addition,

\[
V[\beta](2, S-1) = \max_{p \in P} (r(p, \beta) - \Delta V[\beta](2, S) h(\beta_0 + \beta_1 p) + V[\beta](2, S)), \tag{10}
\]

\[
V[\beta](1, S-1) = \max_{p \in P} (r(p, \beta) - \Delta V[\beta](1, S) h(\beta_0 + \beta_1 p) + V[\beta](1, S)). \tag{11}
\]

Property (P1) implies \( V[\beta](2, S) = V[\beta](1, S) \) and \( \Delta V[\beta](2, S) = 0, \Delta V[\beta](1, S) = V[\beta](1, S) > 0 \), and thus by Lemma 1, \( \Delta V[\beta](2, S-1) = V[\beta](2, S-1) - V[\beta](1, S-1) \neq 0. \)

Let \( c > 1 \) and \((c, s) \in (\mathcal{L}\Delta) \). Then \((c, s) = (c, S-c) \). By the induction hypothesis we have
Figure 1: Schematic picture of $\triangle$

$\Delta V[\beta](c, S - c + 1) \neq 0$, and thus

$$\pi[\beta](c, S - c) = \arg \max_{p \in P} r(p, \beta) - \Delta V[\beta](c, S - c + 1) \cdot h(\beta_0 + \beta_1 p)$$

(12)

$$\neq \arg \max_{p \in P} r(p, \beta) = \pi[\beta](c, S - c + 1),$$

(13)

where we used Lemma 1 for the first inequality, and (P.1) for the second equality. It remains to show $\Delta V[\beta](c + 1, S - c) \neq 0$, when $c < C$. Note that

$$V[\beta](c + 1, S - c) = \max_{p \in P} r(p, \beta) - \Delta V[\beta](c + 1, S - c + 1) \cdot h(\beta_0 + \beta_1 p) + V[\beta](c + 1, S - c + 1),$$

$$V[\beta](c, S - c) = \max_{p \in P} r(p, \beta) - \Delta V[\beta](c, S - c + 1) \cdot h(\beta_0 + \beta_1 p) + V[\beta](c, S - c + 1).$$

Since $(c + 1, S - c + 1) \in \triangle$ and $(c, S - c + 1) \in \triangle$, (P.1) implies $V[\beta](c + 1, S - c + 1) = V[\beta](c, S - c + 1).$ In addition, $c < C$ implies $(c + 1, S - c) \in \triangle$, and thus $\Delta V[\beta](c + 1, S - c + 1) = 0$ by (P.1). The induction hypothesis implies $\Delta V[\beta](c, S - c + 1) \neq 0.$ Then Lemma 1 implies $V[\beta](c + 1, S - c) \neq V[\beta](c, S - c).$ This proves (P.2), and shows that a price-change occurs when $\triangle$ is entered.

**Case 2.** The path $(c_s, s)_{1 \leq s \leq S}$ does not hit $\triangle$. Then there is an $s$ such that $c_s = 2$ and $c_{s+1} = 1$. We show $\pi[\beta](2, s) \neq \pi[\beta](1, s + 1)$, for all $1 \leq s \leq S - 2$.

$$\pi[\beta](2, s) = \arg \max_{p \in P} r(p, \beta) - \Delta V[\beta](2, s + 1) \cdot h(\beta_0 + \beta_1 p),$$

(14)

$$\pi[\beta](1, s + 1) = \arg \max_{p \in P} r(p, \beta) - \Delta V[\beta](1, s + 2) \cdot h(\beta_0 + \beta_1 p),$$

(15)

By Lemma 1, and the fact that $\pi[\beta](2, s)$ and $\pi[\beta](1, s + 1)$ are both in $\text{int}(P)$, it suffices to show $\Delta V[\beta](2, s + 1) \neq \Delta V[\beta](1, s + 2).$ We show by backwards induction that $V[\beta](2, s) - V[\beta](1, s) \neq \Delta V[\beta](2, s + 1) - \Delta V[\beta](1, s + 1).$
\( V[\beta](1, s + 1) \) for all \( 2 \leq s \leq S - 1 \). Let \( 2 \leq s \leq S - 1 \).

\[
V[\beta](2, s) = \max_{p \in P} (r(p, \beta) - \Delta V[\beta](2, s + 1) \cdot h(\beta_0 + \beta_1 p) + V[\beta](2, s + 1)) ,
\]

\[
V[\beta](1, s) = \max_{p \in P} (r(p, \beta) - \Delta V[\beta](1, s + 1) \cdot h(\beta_0 + \beta_1 p) + V[\beta](1, s + 1)) ,
\]

\[
V[\beta](1, s + 1) = \max_{p \in P} (r(p, \beta) - \Delta V[\beta](1, s + 2) \cdot h(\beta_0 + \beta_1 p) + V[\beta](1, s + 2)) .
\]

Using \( V[\beta](1, s + 1) \geq \left[ r^*_1, s - \Delta V[\beta](1, s + 2) d^*_2, s + V[\beta](1, s + 2) \right] \), we have

\[
V[\beta](2, s) - V[\beta](1, s) - V[\beta](1, s + 1) \\
\leq \left[ r^*_1, s - \Delta V[\beta](2, s + 1) d^*_2, s + V[\beta](2, s + 1) \\
- \left[ r^*_1, s - \Delta V[\beta](1, s + 1) d^*_1, s + V[\beta](1, s + 1) \\
- \left[ r^*_2, s - \Delta V[\beta](1, s + 2) d^*_2, s + V[\beta](1, s + 2) \\
= \left[ V[\beta](2, s + 1) - V[\beta](1, s + 1) \right] d^*_2, s + V[\beta](2, s + 1) \\
+ V[\beta](1, s + 1) d^*_1, s - V[\beta](1, s + 1) + V[\beta](1, s + 2) d^*_2, s - V[\beta](1, s + 2) \\
= \left[ V[\beta](2, s + 1) - V[\beta](1, s + 1) \right] - \left[ V[\beta](1, s + 1) d^*_1, s \\
\leq - r^*_1, s + V[\beta](1, s + 1) d^*_1, s \\
= V[\beta](1, s + 1) - V[\beta](1, s).
\]

The last inequality is implied by \( \left[ V[\beta](2, s + 1) - V[\beta](1, s + 1) - V[\beta](1, s + 2) \right] \leq 0 \), which for \( s = S - 1 \) follows from \( (P.1) \), and for \( s < S - 1 \) follows from the induction hypothesis. By \( (29) \), \( p h - V[\beta](1, s + 1) > 0 \), and thus \( V[\beta](1, s) \geq (p h - V[\beta](1, s + 1)) \cdot h(\beta_0 + \beta_1 p h) + V[\beta](1, s + 1) > V[\beta](1, s + 1) \), which proves \( V[\beta](2, s) - V[\beta](1, s) - V[\beta](1, s + 1) < 0 \).

We have shown that in Case 1, \( \pi[\beta](c, S - c) \neq \pi[\beta](c, S - c + 1) \), and in Case 2, \( \pi[\beta](2, s) \neq \pi[\beta](1, s + 1) \). In the notation of Lemma 1, \( \pi[\beta](c, S - c) = \Delta V[\beta](c, S - c + 1) \neq \pi[\beta](c, S - c + 1) \) and \( \pi[\beta](2, s) = \Delta V[\beta](1, s + 1) \neq \pi[\beta](1, s + 1) \).

This implies that on any path \( (c, s), 1 \leq s \leq S \) in \( X \), starting at \( (C, 1) \), there are \( s, s' \) such that

\[
|\pi[\beta](c, s) - \pi[\beta](c, s')| \geq v_0[\beta],
\]

where we define

\[
v_0[\beta] = \min \left\{ \min_{1 \leq s' \leq C} |p^*_\Delta V(c, S - c + 1), \beta - p^*_0, \beta|, \min_{1 \leq s' \leq S - 2} |p^*_\Delta V(2, s + 1), \beta - p^*_\Delta V(1, s + 2), \beta| \right\}.
\]

Since \( (a, \beta) \mapsto p^*_\beta \) is continuous on \( U_a \times U_\beta \), and \( \Delta V[\beta](c, s) \) is continuous in \( \beta \) for all \( (c, s) \in X \), it follows that \( v_0[\beta] \) is continuous in \( \beta \). For each \( \beta \in U_\beta, v_0[\beta] > 0 \), and thus there exists an open neighborhood \( U'_\beta \subset U_\beta \) of \( \beta^{(0)} \), such that \( v_0 = \inf_{\beta \in U_1} v_0[\beta] > 0 \).
5 Performance of pricing strategy

In this section we prove Theorem 1, which says that for some $\epsilon^*$ and any $0 < \epsilon < \epsilon^*$,

$$\text{Regret}(\Phi(\epsilon), T) = O(\log(T)^2).$$

The proof of Theorem 1 is divided in several steps. First we consider convergence of the parameter estimates $\hat{\beta}_t$ to $\beta^{(0)}$. For $\rho > 0$, define

$$T_\rho = \sup \left\{ t \in \mathbb{N} \mid \text{there is no } \beta \in B \text{ with } \left\| \beta - \beta^{(0)} \right\| \leq \rho \text{ and } u_t(\beta) = 0 \right\}. \quad (19)$$

**Proposition 1** (Convergence of parameter estimates). There is a $\rho_1 > 0$ such that for all $0 < \rho \leq \rho_1$, $T_\rho < \infty$ a.s., and $E[T_\rho] < \infty$. Furthermore, $E \left[ \left\| \hat{\beta}_t - \beta^{(0)} \right\|^2 1_{t>T_\rho} \right] = O \left( \frac{\log(t)}{t} \right)$. 

**Proof.** Let $U'_\beta$ and $w_0$ be as in Theorem 2. Choose $\rho_1 \in (0, \rho_0)$, where $\rho_0$ is as in den Boer and Zwart (2011, Theorem 1), and $\{ \beta \in B \mid \left\| \beta - \beta^{(0)} \right\| \leq \rho_1 \} \subset U'_\beta$. Let $k \in \mathbb{N}$. If $t \leq T_{\rho_1}$, then by IIa and an argument similar to (8),

$$\lambda_{\min}(P_{kS}) - \lambda_{\min}(P_{(k-1)S}) \geq (1 + p_h^2)^{-1} \frac{c^2}{2}. \quad (20)$$

If $t > T_{\rho_1}$, then by Theorem 2,

$$\lambda_{\min}(P_{kS}) - \lambda_{\min}(P_{(k-1)S}) \geq w_0. \quad (21)$$

Then for all $k \in \mathbb{N}$, $\lambda_{\min}(P_{kS}) \geq c \cdot k$, where $c = \min \{ w_0, (1 + p_h^2)^{-1} \frac{c^2}{2} \}$, and thus for all $t > 2S$,

$$\lambda_{\min}(P_t) \geq \lambda_{\min}(P_{(SS_t-1)S}) \geq c(SS_t - 1) \geq c \frac{t-S}{S} \geq \frac{c}{2} t, \text{ since } t - S \geq t/2 \text{ for } t > 2S.$$

The Proposition now follows immediately from Theorem 1, Theorem 2 and Remark 2 of den Boer and Zwart (2011). \qed

The following proposition considers the sensitivity of $\pi[\beta](c, s)$ w.r.t. $\beta$, and of $V[\pi, \beta^{(0)}]$ w.r.t. $\pi$.

**Proposition 2.**

(i) For all $(c, s) \in X$ there exists a $K_{c,s} > 0$ such that for all $\beta \in U_\beta$, $|\pi[\beta](c, s) - \pi[\beta^{(0)}](c, s)| \leq K_{c,s} \left\| \beta - \beta^{(0)} \right\|$. 

(ii) There exists a $C > 0$ such that for all $(c, s) \in X$ and $\pi \in \Pi$,

$$\left| V[\pi[\beta^{(0)}], \beta^{(0)}](c, s) - V[\pi, \beta^{(0)}](c, s) \right| \leq C \left\| \pi[\beta^{(0)}] - \pi \right\|^2.$$ 

**Proof.** Fix $(c, s) \in X$, and let $\beta \in U_\beta$.

$$\pi[\beta](c, s) - \pi[\beta^{(0)}](c, s) = p_\Delta^0 V[\beta](c, s+1), \beta - p_\Delta^* V[\beta^{(0)}](c, s+1).$$

By Lemma 1, the mapping $(a, \beta) \mapsto p_{a,\beta}^*$ is continuously differentiable on $U_a \times U_\beta$. One can easily show with backwards induction on $s$, using (3), that $\beta \mapsto \Delta V[\beta](c, s+1)$ is also continuously differentiable on $U_\beta$. The sets $U_a, U_\beta$ can be replaced by open sets $V_a, V_\beta$, such that continuous differentiability still hold on $V_a, V_\beta$, and $U_a \subset V_a, U_\beta \subset V_\beta$. As a consequence, we can conclude that the derivatives of $p_{a,\beta}^*$ and $\Delta V[\beta](c, s+1)$ are bounded on $U_a \times U_\beta$. A first-order Taylor
expansion implies the existence of a constant $K_{c,s} > 0$ such that for all \( \beta \in U_\beta \),
\[
|\pi[\beta](c, s) - \pi[\beta(0)](c, s)| \leq K_{c,s} \left| \beta - \beta(0) \right|.
\]

To prove (ii), it suffices to show that for all \((c, s) \in X\) there is a $C_{c,s} > 0$ such that
\[
\left| V[\pi[\beta(0)], \beta(0)](c, s) - V[\pi, \beta(0)](c, s) \right| \leq C_{c,s} \left| \pi - \pi[\beta(0)] \right|^2,
\]
for all $\pi \in \Pi$. We show this by backwards induction on $s$. For $s = S + 1$ the assertion is trivial; assume $s \leq S$. Then for all $1 \leq c \leq C$,
\[
\begin{align*}
V[\pi[\beta(0)], \beta(0)](c, s) - V[\pi, \beta(0)](c, s) &= f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s) \\
&= f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) + f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s) \\
&+ f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s) - f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) \\
&+ V[\pi[\beta(0)], \beta(0)](c, s, 1) - V[\pi, \beta(0)](c, s, 1)
\end{align*}
\]
Since $\Delta V[\pi[\beta(0)], \beta(0)](c, s + 1) \in U_\alpha$, we can apply Lemma 1(iii) to obtain
\[
\left| f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) + f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s) - f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) \right|
\leq C \left( p_{\Delta V}[\pi[\beta(0)], \beta(0)] - \pi(c, s) \right)^2.
\]
Furthermore, for fixed $p \in P$, $f_{a, \beta(0)}(p)$ is Lipschitz continuous in $a$ on $U_\alpha$, and thus for some $K_{c,s+1}^\prime > 0$,
\[
\begin{align*}
&\left| f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) + f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s) - f_{\Delta V}[\pi[\beta(0)], \beta(0)](c, s, 1) \right| \\
&\leq K_{c,s+1}^\prime \left| \Delta V[\pi[\beta(0)], \beta(0)](c, s + 1) - \Delta V[\pi, \beta(0)](c, s + 1) \right| \\
&\leq K_{c,s+1}^\prime \left| V[\pi[\beta(0)], \beta(0)](c, s + 1) - V[\pi, \beta(0)](c, s + 1) \right| \\
&+ K_{c,s+1}^\prime \left| V[\pi[\beta(0)], \beta(0)](c - 1, s + 1) - V[\pi, \beta(0)](c - 1, s + 1) \right| \\
&\leq (K_{c,s+1}^\prime C_{c,s+1} + K_{c,s+1}^\prime C_{c-1,s+1}) \left| \pi - \pi[\beta(0)] \right|^2,
\end{align*}
\]
by the induction hypothesis. It follows that
\[
\left| V[\pi[\beta(0)], \beta(0)](c, s) - V[\pi, \beta(0)](c, s) \right| \leq C_{c,s} \left| \pi - \pi[\beta(0)] \right|^2,
\]
for some $C_{c,s} > 0$ independent of $\pi$. \(\square\)

**Proof of Theorem 1.** Let $v_0$ be as in the proof of Theorem 2, $\rho_1$ be as in Proposition 1 and $U_\beta^\prime$ as in Theorem 2. Let $\rho \in (0, \rho_1)$. Then $\beta_t \in U_\beta$ for all $t > T_\rho$, and the proof of Theorem 2 make clear that $\max\{|p_i - p_{\text{coop}}| \mid i : SS = SS_t + 1 \} \geq v_0$ for all $t > T_\rho$. If we define $\epsilon = v_0$ and $\epsilon < \epsilon^*$, then clearly IIA does not occur for all $t > T_\rho$. By Propositions 1 and 2, there exist $K > 0$ and $K^\prime > 0$
such that
\[
E \left[ V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_t, \beta^{(0)}](C, 1) \right] \\
= E \left[ \left( V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_t, \beta^{(0)}](C, 1) \right) 1_{t \leq T_\rho} \right] \\
+ E \left[ \left( V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_t, \beta^{(0)}](C, 1) \right) 1_{t > T_\rho} \right] \\
\leq V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) \cdot P(t \leq T_\rho) + C \cdot E \left[ \left| \|\beta^{(0)}\| - \pi[\hat{\theta}] \right|^2 1_{t > T_\rho} \right] \\
\leq V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) \cdot \frac{E[T_\rho]}{t} + K \cdot E \left[ \left| \|\beta^{(0)}\| - \hat{\theta}_t \right|^2 1_{t > T_\rho} \right] \\
\leq K' \frac{\log(t)}{t}.
\]
Consequently, for each selling season \(i > 1\) there is a \(K > 0\) such that
\[
E \left[ V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_{1+i-1S}, \ldots, \pi_iS, \beta^{(0)}](C, 1) \right] \\
\leq \sum_{j=1+i-1S}^{iS} E \left[ V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_j, \beta^{(0)}](C, 1) \right] \\
\leq K \frac{\log(iS)}{iS},
\]
and thus
\[
\text{Regret}(T) = \sum_{i=1}^{T} E \left[ V[\pi[\beta^{(0)}], \beta^{(0)}](C, 1) - V[\pi_{1+i-1S}, \ldots, \pi_iS, \beta^{(0)}](C, 1) \right] \\
\leq \sum_{i=1}^{T} K \frac{\log(iS)}{iS} \\
= O(\log(T)^2),
\]
since \(\frac{1}{2} \int_{x=1}^{t} x^{-1} \log(x) dx = \log(t)^2\).
6  Numerical illustration

To illustrate the analytical results that we have derived, we provide a numerical illustration. We let $C = 10$, $S = 20$, $p_l = 1$, $p_h = 20$, $\beta_0^{(0)} = 2$, $\beta_1^{(0)} = -0.4$, and $h(z) = \text{logit}(z)$. The optimal expected revenue per selling season, $V[\pi[\beta^{(0)}], \beta_1^{(0)}](C, 1)$, is equal to 47.79. We consider a time span of 100 selling periods. The following figures show a sample path of $||\hat{\beta}_t - \beta^{(0)}||$ for $t = 1, \ldots, 100S$, Regret($T$), and the relative regret $\frac{\text{Regret}(T)}{V[\pi[\beta^{(0)}], \beta_1^{(0)}](C, 1)} \times 100\%$, for $T = 1, \ldots, 100$.

7  Conclusions and future research

In this paper we consider a dynamic pricing problem with uncertain parametric demand, with finite inventories and multiple consecutive selling seasons. The purpose of the seller is to determine selling prices that optimize the expected revenue. We propose a pricing strategy, based on the intuitive idea to estimate at each decision moment the unknown parameters using Maximum Likelihood Estimation, and subsequently use the price that would be optimal if this estimate were correct. We show that the parameter estimates converge to the correct value, and show that the Regret after $T$ selling periods satisfies $\text{Regret}(T) = O(\log(T)^2)$.

To avoid cumbersome notation, we perform the analysis in a setting of non-overlapping selling seasons. The analysis can however easily be extended to the case of multiple overlapping selling seasons; these may even be of different length or have different initial inventory levels. The same pricing strategy $\Phi(\epsilon)$ can be used, leading to $\text{Regret}(T) = O(\log(T)^2)$.

The regret bound of Theorem 1 is only valid if $\epsilon$ does not exceed $\epsilon^*$. The value of $\epsilon^*$ depends
on the value of \( v_0 \) in the proof of Theorem 2, which in principle can be calculated explicitly. Its value does however depend on the unknown \( \beta^{(0)} \). On the other hand, in practical situations the assumptions on \( g \) have to be made for all \( \beta \in B \), not only for \( \beta^{(0)} \), since \( \beta^{(0)} \) is unknown. Consequently one can calculate \( \inf_{\beta \in B} v_0[\beta] \), where \( v_0[\beta] \) is as in the proof of Theorem 2, and choose \( \epsilon^* < \inf_{\beta \in B} v_0[\beta] \). Another solution is to try to alter \( \Phi(\epsilon) \) by decreasing the value of \( \epsilon \) each time that IIa is invoked, until \( \epsilon < v_0 \), and thus construct a pricing strategy which is independent of the parameter \( \epsilon \). This is an interesting question for future research.

The dynamic pricing problem that we study is an example of a sequential decision problem under parametric uncertainty, where optimal decisions are the solution of a Markov Decision Problem (with compact action space and finite state space). Our solution approach may be applicable to many more such problems, provided some self-learning result analogous to Theorem 2 holds. An interesting direction for future research is to design a framework for these type of sequential decision problems, and apply our solution approach.
Appendix

Proof of Lemma 1. Since
\[ f_{a,\beta}(p) = h(\beta_0 + \beta_1 p) \left[ 1 + (p - a)\beta_1 \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)} \right] = h(\beta_0 + \beta_1 p) \left[ 1 - g_{a,\beta}(p) \right] \]
and \( h(\beta_0 + \beta_1 p) > 0 \) for all \( \beta \in U_\beta, p \in \mathcal{P} \), we have \( f_{a,\beta}(p) = 0 \) if and only if \( g_{a,\beta}(p) = 1 \). The assumptions on \( g_{a,\beta} \) imply that for all \((a, \beta) \in U_a \times U_\beta\), there is a unique \( p^*_{a,\beta} \in \text{int}(\mathcal{P}) \) such that \( g_{a,\beta}(p^*_{a,\beta}) = 1 \). From
\[ f_{a,\beta}(p) = \frac{\partial}{\partial p} \left[ h(\beta_0 + \beta_1 p)(1 - g_{a,\beta}(p)) \right] \]
follows
\[ f_{a,\beta}(p^*_{a,\beta}) = -h(\beta_0 + \beta_1 p^*_{a,\beta}) \frac{\partial}{\partial p} g_{a,\beta}(p^*_{a,\beta}) < 0, \]
since \( g_{a,\beta} \) is strictly increasing in \( p \). This proves (i).

For all \((a, \beta) \in U_a \times U_\beta\),
\[ \frac{\partial g_{a,\beta}(p)}{\partial p} \bigg|_{p=p^*_{a,\beta}} = -\beta_1 \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)} \bigg|_{p=p^*_{a,\beta}} - (p^*_{a,\beta} - a)\beta_1 \frac{\partial}{\partial p} \left[ \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)} \right] \bigg|_{p=p^*_{a,\beta}}. \]
The log-concavity of \( z \mapsto h(z) \) implies \( \frac{\partial}{\partial z} \log h(z) = \frac{\dot{h}(z)}{h(z)} \leq 0 \) for all \( z \). With \( \beta_1 < 0 \), this implies \( \frac{\partial}{\partial p} \left[ \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)} \right] \bigg|_{p=p^*_{a,\beta}} \geq 0 \). Furthermore, \( g(p^*_{a,\beta}) = 1 \) implies
\[ (p^*_{a,\beta} - a) = -\beta_1^{-1} \frac{\dot{h}(\beta_0 + \beta_1 p^*_{a,\beta})}{h(\beta_0 + \beta_1 p^*_{a,\beta})} > 0, \]
and thus
\[ \frac{\partial g_{a,\beta}(p)}{\partial p} \bigg|_{p=p^*_{a,\beta}} > 0 \]
for all \((a, \beta) \in U_a \times U_\beta\). The implicit function theorem (Duistermaat and Kolk, 2004) entails that the mapping \((a, \beta) \mapsto p^*_{a,\beta}\) is continuously differentiable at every point \((a, \beta) \in U_a \times U_\beta\).

For all \((a, \beta) \in U_a \times U_\beta\) and \( p \in \mathcal{P} \),
\[ \frac{\partial g_{a,\beta}(p)}{\partial a} = \beta_1 \frac{\dot{h}(\beta_0 + \beta_1 p)}{h(\beta_0 + \beta_1 p)} < 0. \]
This implies that for all \( a \in U_a, a' \in U_a \) with \( a < a' \), and all \( p \in \mathcal{P}, p \leq p^*_{a,\beta} \): \( g_{a',\beta}(p) \leq g_{a,\beta}(p) \leq 1 \). Therefore \( p^*_{a',\beta} > p^*_{a,\beta} \) for all \( a < a' \), and thus \( p^*_{a,\beta} \) is strictly monotone increasing in \( a \).

Using \((p^*_{a,\beta} - a) = (-\beta_1^{-1}) \frac{\dot{h}(\beta_0 + \beta_1 p^*_{a,\beta})}{h(\beta_0 + \beta_1 p^*_{a,\beta})}\), we have
\[ f_{a,\beta}(p^*_{a,\beta}) = (p^*_{a,\beta} - a)h(\beta_0 + \beta_1 p^*_{a,\beta}) = (-\beta_1^{-1}) \frac{\dot{h}(\beta_0 + \beta_1 p^*_{a,\beta})^2}{h(\beta_0 + \beta_1 p^*_{a,\beta})}. \]
and thus
\[
\frac{\partial}{\partial a} f_{a, \beta}(p_{a, \beta}^*) = (-\beta_1^{-1}) \left( \frac{\partial}{\partial z} \left( \frac{h(z)^2}{h(z)} \right) \right) _{z=\beta_0 + \beta_1 p_{a, \beta}^*} \beta_1 \frac{\partial}{\partial a} p_{a, \beta}^*. \tag{23}
\]

Log-concavity of \( h \) implies \( \frac{\partial^2 \log h(z)}{\partial z^2} = \frac{h(z)h'(z) - h'(z)^2}{h(z)^2} \leq 0 \), and thus
\[
\frac{\partial}{\partial z} \frac{h(z)^2}{h(z)} = \frac{2h(z)h'(z) - h'(z)^2}{h(z)^2} = h(z) \left[ 2 - \frac{h(z)h'(z)}{h(z)^2} \right] \geq h(z).
\]

Since \( \frac{\partial}{\partial a} p_{a, \beta}^* > 0 \), it follows that \( f_{a, \beta}(p_{a, \beta}^*) \) is strictly decreasing in \( a \). This completes the proof of (ii).

Let \( C = -2 \sup_{(a, \beta, p) \in U_a \times U_\beta \times \mathcal{P}} f_{a, \beta}(p) \). Since \( (a, \beta, p) \mapsto f_{a, \beta}(p) \) is twice continuously differentiable on \( \mathbb{R} \times B \times \mathcal{P} \) and \( f_{a, \beta}(p_{a, \beta}^*) < 0 \), it follows that \( 0 < C < \infty \). By a Taylor expansion, there is a \( \bar{p}_{a, \beta} \) on the line segment between \( p \) and \( p_{a, \beta}^* \), such that
\[
f_{a, \beta}(p) = f_{a, \beta}(p_{a, \beta}^*) + \bar{f}_{a, \beta}(p_{a, \beta}^*)(p - p_{a, \beta}^*) + \frac{1}{2} \bar{f}_{a, \beta}(p_{a, \beta}^*)(p - p_{a, \beta}^*)^2
\]
\[
\leq f_{a, \beta}(p_{a, \beta}^*) - C(p - p_{a, \beta}^*)^2,
\]
using \( \bar{f}_{a, \beta}(p_{a, \beta}^*) = 0 \). This proves (iii).

**Proof of Lemma 2.** Let \( \beta \in U_\beta \) be arbitrary. For \( (c, s) \in \mathcal{X} \), write \( r_{c, s}^*[\beta] = r(\pi[\beta](c, s), \beta) \) and \( d_{c, s}^*[\beta] = h(\beta_0 + \beta_1 \pi[\beta](c, s)) \). We show \( 0 \leq \Delta V[\beta](c, s) \leq r_{1, S}^*[\beta] \) for all \( (c, s) \in \mathcal{X} \). By (7), this implies \( \Delta V[\beta](c, s) \in U_\nu \). In view of (3), uniqueness and continuity of \( \pi[\beta] \) then follows from repeated application of Lemma 1(i, ii), for each \( (c, s) \in \mathcal{X} \).

If \( s = S \), then \( \Delta V[\beta](c, S) = 0 \) for \( c > 1 \) or \( c = 0 \), and \( V[\beta](1, S) = r_{1, S}^*[\beta] \). If \( s < S \), then by backwards induction,
\[
\Delta V[\beta](c, s) \geq \left[ r_{c-1, s}^* - \Delta V[\beta](c, s + 1) + d_{c-1, s}^* \right] \tag{24}
\]
\[
= \left[ r_{c-1, s}^* - \Delta V[\beta](c, s + 1) + d_{c-1, s}^* + V[\beta](c, s + 1) \right] \tag{25}
\]
\[
\Delta V[\beta](c, s) \geq 0, \tag{27}
\]
and
\[
\Delta V[\beta](c, s) \leq \left[ r_{c, s}^* - \Delta V[\beta](c, s + 1) + d_{c, s}^* \right] \tag{28}
\]
\[
= \left[ r_{c, s}^* - \Delta V[\beta](c, s + 1) + d_{c, s}^* + V[\beta](c, s + 1) \right] \tag{29}
\]
Acknowledgment

We thank Sandjai Bhulai for useful discussions and providing literature references.

References


