158

DISCUSSION PAPER

Originally published by Ipea in September 2005 as number 1117 of the series Texto para Discussão.

ipea

DYNAMIC OPTIMIZATION AND LEARNING: HOW SHOULD A MANAGER SET PRICES WHEN THE DEMAND FUNCTION IS UNKNOWN?

Alexandre X. Carvalho
Martin L. Puterman

ipea

158
DISCUSSION PAPER
Originally published by Ipea in September 2005 as number 1117 of the series Texto para Discussão.
Brasília, January 2015

DYNAMIC OPTIMIZATION AND LEARNING: HOW SHOULD A MANAGER SET PRICES WHEN THE DEMAND FUNCTION IS UNKNOWN?[1]

Alexandre X. Carvalho[2]
Martin L. Puterman[3]

1. Agradecimentos/Acknowledgments
This research was partially supported by NSERC and the MITACS NCE (Canada). We wish to thank Dan Adelman for his helpful comments on an earlier draft of this manuscript. We are also thankful for the suggestions provided by participants at the Canadian Operations Research Society Annual Meeting 2003, Vancouver, and at the INFORMS Annual Meeting 2003, Atlanta. All errors are the authors'. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors.
2. Directorate of Regional and Urban Studies, Institute of Applied Economic Research (Ipea)/Diretoria de Estudos Regionais e Urbanos (Dirur), Instituto de Pesquisa Econômica Aplicada.
3. Director of Centre for Health Care Management, Sauder School of Business, University of British Columbia.

DISCUSSION PAPER

A publication to disseminate the findings of research directly or indirectly conducted by the Institute for Applied Economic Research (Ipea). Due to their relevance, they provide information to specialists and encourage contributions.

Discussion paper / Institute for Applied Economic

Research.- Brasília : Rio de Janeiro : Ipea, 1990-

ISSN 1415-4765

1. Brazil. 2. Economic Aspects. 3. Social Aspects. I. Institute for Applied Economic Research.

CDD 330.908

JEL: C44; C61; C63; D81

SUMÁRIO

SINOPSE

Neste artigo, nós estudamos o problema de escolher preços sequencialmente, de forma a maximizar a receita esperada, em um ambiente onde os parâmetros da função de demanda são desconhecidos, e o horizonte de vendas é finito. Vários métodos de otimização seqüencial são discutidos, onde os preços e as vendas resultantes anteriores são utilizados para determinar o preço no período atual. Expansões de Taylor são empregadas para construir aproximações da função valor, explicitando a relação de compromisso entre maximização de receita no curto-prazo e maior ganho de informação para obter maior receita agregada no longo-prazo. A partir dessas expansões, nós derivamos estratégias promissoras, denominadas políticas de *one-step look-ahead* que combinam otimização e aquisição de informação dinamicamente. Simulações de Monte Carlo são apresentadas, onde constatamos a superioridade das políticas de *one-step look-ahead*, quando comparadas a diversas outras regras de otimização seqüencial. Finalmente, nós discutimos problemas de endogeneidade, onde fazemos um paralelo com a teoria de controle adaptativo, e discutimos a validade das regras de *one-step look-ahead*, mesmo quando endogeneidade é observada.

ABSTRACT

This paper considers the problem of changing prices over time to maximize expected revenues in the presence of unknown demand distribution parameters. It provides and compares several methods that use the sequence of past prices and observed demands to set price in the current period. A Taylor series expansion of the future reward function explicitly illustrates the tradeoff between short term revenue maximization and future information gain and suggests a promising pricing policy referred to as a one-step look-ahead rule. An in-depth Monte Carlo study compares several different pricing strategies and shows that the one-step look-ahead rules dominate other heuristic policies and produce good short term performance. The reasons for the observed bias of parameter estimates are also investigated.

1 Introduction

A commonly encountered problem in the optimization literature is how to sequentially set prices so as to maximize the cumulative expected revenues. Martinez (2003), for example, presents an application where analysts working for Intrawest Corporation, in Vancouver, had to develop approaches to set prices for ski lift tickets to increase the company's revenue. In that project, the analysis team quickly found that historical data was not sufficient to determine the effect of price changes on demand. At this point, the project focus shifted to determining data requirements and developing tools and methods to capture data to investigate the effect of price on demand. But even if relevant data had been available, the question of how management should vary its prices to maximize revenue remained. This paper provides insights into how to do this by developing and evaluating several implementable price setting mechanisms. Further it measures the benefits of using these methods and provides implementable recommendations for how to use these approaches.

In a nutshell, the main messages of this paper are:

- Managers can increase revenue by changing prices over time. This benefit comes through using variable prices to learn about the relationship between price and demand.
- Managers should collect and save pricing and sales data and use it to guide pricing decisions.
- Managers can increase total revenue by intermittently choosing prices in a random manner.
- Managers can increase total revenue by implementing a systematic approach to choosing prices and using it in real time.
- The Internet and its host of e-commerce tools are ideal for using the approaches proposed in this paper.

In the latter sections of this paper, we will expand on and quantify the benefits from following these recommendations. The results and techniques presented in

this paper can be extended to a great variety of situations where one needs to set control variables in order to maximize a partially known objective function.

This paper considers the following pricing problem. Each period a manager (who we refer to as "he") sets the price of a good (for example, a ski ticket of a particular type) and observes the total demand (skier visits) for that good during the period. He seeks to choose prices so as to maximize total expected revenues over a fixed *finite* horizon of length T which might represent the length of a season or the lifetime of a good.

We assume an *unlimited* inventory of goods to simplify analysis and develop generalizable principals that may apply in wider contexts. Clearly this assumption is appropriate in the ski industry as well as in software and other information good industries. Also it applies to settings in which pricing and inventory decisions are made in separate units within the organization. From a technical perspective, by focussing on pricing only we avoid joint optimization of inventory levels and prices and can provide clearer recommendations. We have chosen not to take a revenue management approach in which the inventory is finite and prices are set to maximize the total expected revenue from selling this inventory of goods. Conceptually either of these extensions would not change our approach, they would only alter the dynamic programming equations on which our analysis is based.

We assume demand is stochastic and its mean is described by a demand function which relates demand to price and possibly other factors such as season, time on market and competitive factors. We further assume that the demand function form is known, for example, its logarithm may be linear in price and quadratic in the day of the season. When the parameters of the demand function are also known, the choice of an optimal price reduces to a simple stochastic optimization problem. However, when the parameters are not known, which is the setting we will focus on, the manager may benefit from using variable pricing to learn them. Initially he may rely on prior information or intuition to guide pricing decisions. After several periods, he can use statistical methods to estimate the demand function and use this information to guide his price setting for the rest of the planning horizon.

The simplest and most widely used approach to pricing is to set a single price at the beginning of the planning horizon and use that price throughout the lifetime of the product. To do this well requires complete knowledge of the demand curve for the product. Alternatively the decision maker may specify an "open loop" price

schedule which determines how to vary prices over the planning horizon independent of any information that may be acquired throughout the planning period. We will show that these approaches are not attractive when there is uncertainty about the demand model.

We focus on adaptive or "closed-loop" price setting, that is when the decision maker varies prices on the basis of his record of historical demand and prices chosen. He can do this "myopically" by setting the price p_t at each period t, in order to maximize the immediate expected revenue R_t. Obviously, at time t, he can use all past price and demand data to choose p_t. As will be illustrated in this paper, this strategy will *not* yield the maximum total expected revenues by the end of the planning horizon. In fact, this myopic strategy may turn out to be far from optimal. This is because the price set today will impact not only the immediate expected return R_t, but will also affect the amount of information about the demand function that the retailer gains. In fact, we show that prices that do not maximize immediate revenue will lead to better estimation of the demand curve and better future price decisions.

On the other hand, there is a vast statistical body of literature (see Draper and Smith, 1998) that addresses how to vary experimental conditions, which in the setting of this paper are the prices, to best learn a parametric function that relates outcomes to experimental conditions. Such optimal design approaches do not usually focus on the impact of outcomes, herein the revenue gained, during the experimental process and use optimality criteria based on covariance matrices of parameters.

The trade-off between immediate reward maximization and learning has been extensively studied in many areas. We draw inspiration from the reinforcement learning (RL) literature where most of the techniques are suited for problems where the horizon is infinite and the state space is of high dimension. In these cases, there is a great interest in finding the optimal strategy, and much attention is devoted to construct methods to uncover the optimal policy in the limit. RL methods seek to approximate the value function with high accuracy and many algorithms have been proposed to do this (see Anderson and Hong, 1994, Dietterich and Wang, 2003, Forbes and Andre, 2000, Sutton and Barto, 1998, Tsitsiklis, 1997).

The literature on learning and pricing appears to date back to Rothschild (1974). He represents the problem of maximizing the firm's revenue by choosing

between two prices as a two armed bandit model and shows among other results that the manager can choose the wrong price infinitely often. Kiefer and Nyarko (1989) study the general problem of learning and control for a linear system with a general utility function. They formulate this as an infinite horizon discounted Bayesian dynamic program; show that a stationary optimal policy exists and that the posterior process converges with probability one but that the limit need not be concentrated on the true parameter. Related work appears in Easley and Kiefer (1988,1989).

Balvers and Cosimano (1990) apply the Kiefer and Nyarko framework to the specific problem of a manager who sets prices dynamically to maximize expected total revenue. They use a dynamic programming approach to "gain some insight into why it is important for the firm to learn". They derive a specific expression for the optimal price and then explore its implications. They conclude among other results, that when varying price, an anticipated change in demand leads to a small price change while an unanticipated shift in demand leads to greater changes in price and that the effect of learning persists into the future. Their focus is qualitative and does not quantify the potential benefits that can be gained from using learning nor show to do it in practice.

Aviv and Pazgal (2002a and 2002b) introduce learning into a pricing model formulated by Gallego and van Ryzin (1994) in which customers arrive singly according to a Poisson process depending with rate depending on the price. Aviv and Pazgal (2002a) are concerned with deriving a closed form optimal control policy while Aviv and Pazgal (2002b) use a partially observed MDP framework to consider a model with a finite number of scenarios. Petruzzi and Dada (2002) also study learning, pricing and inventory control. In their model, the inventory level censors demand and once the inventory level is sufficiently high or demand is low so that it is not censored, the demand function parameters can be determined with certainty and revenue can be maximized.

Lobo and Boyd (2003) consider a similar model to that in this paper. They assume demand is linear in price and derive a Taylor series approximation for the one-period difference between the expected reward under the policy that uses the true optimal price and that based on the myopic policy. Although their Taylor series expansion is similar to ours, they use it for a different purpose; that is, to formulate a semi-definite convex program in which the objective function is

the sum of discounted one-period Taylor series approximations. They then solve the problem over short planning horizons (10 and 20 periods) and compare the average revenue from several approaches. In our view, the contribution of this paper is primarily methodological; it does not focus on the managerial implications of learning.

Cope (2004) and Carvalho and Puterman (2005) study a related dynamic pricing problem. Their setting is as follows. Each period, the manager sets a price and observes the *number* of people who arrive and the number who purchase the product at that price. In these papers the focus is on estimating the relationship between the *probability* of purchase and the price set. Cope uses a non-parametric non-increasing model to relate price to the probability of purchase while Carvalho and Puterman use logistic regression. Further, Cope uses Bayesian methods while Carvalho and Puterman use maximum likelihood estimation. From a managerial perspective this work of Cope and Carvalho and Puterman focus on low demand items or a setting in which the prices can be changed frequently, while the focus of this paper is on settings in which demand is high or prices can be altered less frequently.

The work in Carvalho and Puterman (2005) differs from the work in this paper in the following ways: (1) while Carvalho and Puterman assumes the retailer observes both the number of arriving customers and their decision to buy or not, in this paper the retailer observes only the total number of items sold in each period; (2) the parametric model in Carvalho and Puterman is based on a logistic regression for the purchase probability and a Poisson model for the arrival process, while in this paper, we consider a log-linear regression model for the demand function; (3) in this paper, we use Kalman filter closed form equations to update the parameter estimates, whereas Carvalho and Puterman uses plain MLE because of the impossibility of obtaining nice closed form expressions for sequential estimation. (4) this paper presents the general theoretical foundations for the problem of pricing and learning; (5) we provide a thorough discussion about endogeneity issues, where we show the relation between the dynamic pricing problem and adaptive control; (6) the Monte Carlo experiment presented here is more extensive, and we compare our suggested one-step-ahead strategy to a more complete set of alternative policies; (7) finally, even in the presence of parameter estimate bias, we provide a discussion about why the one-step-ahead policy is still valid.

Raju, Narahari and Kumar (2004) study the use of reinforcement learning to set prices in a retail market consisting of multiple competitive sellers and heterogeneous buyers who have different preferences to price and lead time. They use a continuous time queuing model for their system and explore the effect of using several different algorithms for computing optimal policies.

The effect of learning in other contexts have been considered by Scarf (1960), Azoury (1985), Lovejoy (1990), and Treharne and Sox (2002) in the inventory context and Hu, Lovejoy and Shafer (1996) in the area of drug therapy. In the papers of Scarf, Azoury, Lovejoy and Treharne and Sox, learning is *passive* in the sense that the policy selected by the decision maker does not impact the information received. In particular, in these papers the decision maker observes the demand each period regardless of the inventory level set by the decision maker. In contrast in this paper and the remaining papers above, learning is *active*, that is, the policy set by the decision maker impacts the information that is received. In the newsvendor models studied by Lariviere and Porteus (1999), Ding, Puterman and Bisi (2002) the inventory levels censors demand so when setting inventory levels the decision maker is faced with the tradeoff of myopic optimality and learning. In Hu, Lovejoy and Shafer (1996) the dosage level impacts both the health of the patient and subsequent parameter estimates so the decision maker must again tradeoff short term optimality with learning.

The remainder of the paper is organized as follows. In Section 2 we formulate our model, illustrate the trade-off between immediate revenue maximization and learning and discuss demand distribution parameter updating. In Sections 3 and 4, we provide a formal treatment for the problem of dynamic pricing and learning for arbitrary demand parametric models. We present Monte Carlo simulation results to evaluate the performance of several heuristic methods on Section 5. In Section 6 we describe why model parameter estimates may be biased. Conclusions and recommendations appear in the final section. Proofs are presented in the Appendix.

2 Model Formulation

Consider a manager who has an unlimited quantity of a product to sell at each time period t, $t = 1, \ldots, T$, where T is finite. The demand q_t for the product is

represented by a continuous random variable and is assumed to follow a log-linear model. Such a model was recommended by Kalyanam (1996) on the basis of an analysis of marketing data. Consequently, we assume that the demand in period t, is related to price set in that period, p_t through the equation

$$q_t = \exp\big[\alpha + \beta p_t + \epsilon_t\big], \tag{1}$$

where α and β are unknown regression parameters, and ϵ_t is a random disturbance term that is normally distributed with mean 0 and variance σ^2. We will start by focusing on this model but many generalizations are possible including letting the regression parameters vary over time with both random and systematic components, or we can add other factors to the model; for example if we assume a quadratic time trend in the model, then (1) becomes

$$q_t = \exp\big[\alpha + \beta p_t + \gamma t + \eta t^2 + \epsilon_t\big], \tag{2}$$

A further enhancement of this model would be to include an interaction between trend and price.

We argue here that from a modeling perspective parametric models are preferable to non-parametric models in this finite horizon setting. Since our primary focus is optimize revenues during a limited number of periods, say $T = 100$ or $T = 200$, it would be difficult to obtain a reasonable approximation to a non-parametric demand function with such little data without strong prior assumptions. Further, this parametric formulation enables the user to easily include additional information in the demand equation, such as in (2) above, which would be extremely difficult in a non-parametric setting.

The manager's objective is to adaptively choose a sequence of prices $\{p_t : t = 1, \ldots, T\}$ to maximize the total expected revenue $\sum_{t=1}^{T} p_t E[q_t]$, over T periods. We now focus on the model in (1). If the retailer knows the model parameters α and β, he can set the price p_t to the value that maximizes the revenue $R_t(p_t) = p_t E[q_t]$ in each period t. The optimum price in this case is $p_t^* = -1/\beta$, which does not depend on α, and the maximum expected revenues are $\sum_{t=1}^{T} R_t^* = \sum_{t=1}^{T} R_t(p_t^*) = -(T/\beta)e^{\alpha-1}M$, where $M = E[e^{\epsilon_t}] = \exp[\sigma^2/2]$.

Of course, in any real setting, but especially for new products, the decision

maker does not know the true parameter values α and β so they must be estimated from the data that is acquired during the planning horizon. The key concept on which this paper is based is that the prices the manager sets influence *both* the data received and the ongoing revenues. Thus the manager wants to use prices both strategically, that is to learn about the demand function and optimally to maximize immediate revenue. This paper is about this tradeoff.

The information flow in this system is linked through a feedback process. Each period, the manager uses his current estimate of the demand function parameters to set the price, then he observes demand in the period and finally he updates his estimates of the demand function parameters. Then he repeats this process. The paper now proceeds along two paths. First we discuss how to update the parameter estimates given an additional demand observation and second we discuss how to choose a price given estimates of the demand distribution parameters.

2.1 Updating Demand Distribution Parameter Estimates

This section describes how to update the demand function parameter estimates once the demand in a particular period has been observed. We assume at time $t = 0$, which means before we set the first price p_1 and observe the demand q_1, that the manager has specified a prior distribution for the regression parameters. The initial prior may either be subjective, derived from product information and past history or based on some preliminary sales data for the current product. We assume that the vector of regression coefficients $\theta = [\alpha\ \beta]' \sim N(\theta_0, \sigma^2 P_0)$, where θ_0 is the 2×1 prior mean vector and $\sigma^2 P_0$ is a 2×2 is the prior covariance matrix. To simplify our initial analysis, we assume known σ^2.

At period t, based on the available estimates α_{t-1} and β_{t-1}, the manager sets p_t (hopefully by methods suggested below) and observes the demand q_t. From (1), we can rewrite the model as $y_t = \log(q_t) = \alpha + \beta p_t + \epsilon_t$. To simplify notation and allow easy generalization of results we write $y_t = \theta' z_t + \epsilon_t$, where $z_t = [1\ p_t]'$ and the covariance of the estimates of regression parameters as $\sigma^2 P_t$ where P_t is an explicit function of the prices up to and including decision epoch t.

Using properties of the normal distribution and the fact that the normal prior is conjugate to the normal distribution, we can easily derive the posterior distribution of the parameters α and β using standard methods. It has been widely

established (see, for example, Harvey, 1994 or Fahrmeir and Tutz, 1994) that the parameters of the posterior distribution are related to the parameters of the prior distribution through the recursive equations which can be expressed in matrix and vector notation as

$$\theta_t = \theta_{t-1} + P_{t-1} z_t' F_t^{-1} [y_t - \delta_{t-1}' z_t] \tag{3}$$

$$P_t = P_{t-1} - P_{t-1} z_t' F_t^{-1} z_t P_{t-1}, \tag{4}$$

$$F_t = z_t P_{t-1} z_t' + H_t, \tag{5}$$

where, in our model $H_t = 1$ and all other quantities are defined above. At time t, the variance of the estimate β_t is given by $\sigma^2 P_{t,2,2}$, where $P_{t,2,2}$ is the second element in the diagonal of P_t. Therefore, at the end of period t, after observing q_t, we have updated estimates α_t and β_t for the parameters. These equations are referred to as the *Kalman Filter*. They were originally developed in an engineering context for a different purpose and are widely used in time series analysis and control theory.

2.2 Choosing a Price

We investigate the trade-off between optimization and learning by focussing on the case where we have observed the demands $\{q_t : t = 1, \ldots, T-2\}$, corresponding to the sequence of prices $\{p_t : t = 1, \ldots, T-2\}$, and now we seek to maximize the expected revenues in the last two periods $E[R_{T-1}(p_{T-1})] + E[R_T(p_T)]$. This can be formulated as a two-period dynamic program; our analysis is motivated by that observation. At the beginning of period $t = T-1$, the information for θ is provided by the estimate θ_{T-2} which is normally distributed with mean θ and covariance matrix $\sigma^2 P_{T-2}$. After fixing the price p_{T-1} and observing q_{T-1}, we update θ_{T-2} and P_{T-2} using the Kalman filter recursions (3) - (5). At the last period $t = T$, learning has no subsequent benefit so the manager should choose a price in that period to maximize immediate expected revenue in period T. Using simple calculus, the optimum price, considering the estimate β_{T-1}, will be $p_T = -1/\beta_{T-1}$.

The estimate β_{T-1} will be normally distributed with variance $\sigma^2_{\beta_{T-1}} = \sigma^2 P_{T-1,2,2}$. Here $P_{T-1,2,2}$ denotes the coefficient of the matrix P_{T-1} in the 2nd row and 2nd column. Because β_{T-1} is subject to estimation error, we expect the optimum price

$p_T = -1/\beta_{T-1}$ to deviate from its true optimum $-1/\beta$. Therefore, we may anticipate some loss in the revenue at time T given the uncertainty about β. For an estimate β_{T-1}, if we use the rule $p_T(\beta_{T-1}) = -1/\beta_{T-1}$, the expected revenue in the last period can be written as

$$R_T^*(p_T(\beta_{T-1})) = E\Big[-\frac{1}{\beta_{T-1}} e^{(\alpha-\beta/\beta_{T-1}+\epsilon_T)}\Big] = -M\frac{1}{\beta_{T-1}} e^{(\alpha-\beta/\beta_{T-1})}, \qquad (6)$$

where the expectation above is taken with respect to the noise ϵ_T. Obviously $R_T^*(p_T(\beta_{T-1}))$ is maximum when β_{T-1} is the true value β. Theorem 1 below provides an approximation for the expectation $E[R_T^*(p_T(\beta_{T-1}))]$ based on a Taylor expansion. The proof is presented in the Appendix and generalized considerably in Section 3.

Theorem 1 *Suppose that the price in period T is set equal to $p_T = -1/\beta_{T-1}$. Then the expected revenue in period T can be approximated by*

$$E[R_T^*(p_T(\beta_{T-1}))] = R_T^*(p_T(\beta)) + \frac{1}{2}\frac{Me^{(\alpha-1)}}{\beta^3}\sigma^2_{\beta_{T-1}} + O((T-2)^{-2}), \qquad (7)$$

where the expectation above is calculated with respect to the distribution of the random variable β_{T-1}.

One of the crucial assumptions in Theorem 1 is that the sequence of prices $\{p_1, p_2, \ldots, p_{T-2}\}$ is non-random, or, if it is random, the dependence between p_t and p_{t+k} vanishes as $k \to \infty$. In practice, when the prices are updated recursively, this hypothesis does not necessarily hold. The implications of violating this assumption will be discussed in Subsection 4.2 and in Section 6.

Based on the result above, the expected loss at period T associated with the uncertainty in β_{T-1} can be approximated by

$$-\frac{1}{2}\frac{Me^{(\alpha-1)}}{\beta^3}\sigma^2_{\beta_{T-1}} > 0. \qquad (8)$$

At the time $t = T-1$, the fixed price p_{T-1} will affect not only the immediate return $R(p_{T-1})$ but will also affect the variance of β_{T-1}, $\sigma^2_{\beta_{T-1}} = \sigma^2_{\beta_{T-1}}(p_{T-1})$. Therefore, we can choose the price p_{T-1} that maximizes the expression

$$F_{T-1}(p_{T-1}) = p_{T-1}e^{\alpha+\beta p_{T-1}}M + \frac{1}{2}\frac{Me^{(\alpha-1)}}{\beta^3}\sigma^2_{\beta_{T-1}}(p_{T-1}), \qquad (9)$$

which explicitly shows the trade-off between learning and maximization at time $t = T - 1$. Note that (9) involves the parameters α and β, which are exactly what we want to estimate, but in computation we replace them by their current estimates.

When the managers seeks to maximize revenue in an horizon exceeding 2 periods, writing down expressions equivalent to (9) is not an easy task and the dynamic programming formulations become quickly intractable. However, as a result of (9) and general observations about the growth of value functions over time, we propose a simple heuristic pricing policy, hereafter referred to as one-step-ahead strategy, to optimize the the total expected revenue over longer horizons. In this case, the one-step-ahead policy consists of choosing the price p_t, at each time period t, that maximizes the objective function

$$\hat{F}_t(p_t) = p_t e^{\alpha_{t-1} + p_t \beta_{t-1}} M_{t-1} + \frac{G(t)}{2} \frac{M_{t-1} e^{-1} e^{(\alpha_{t-1})}}{\beta_{t-1}^3} \sigma_{\beta_t}^2(p_t). \tag{10}$$

Note that we replaced the unknown regression coefficients α and β by their available estimates α_{t-1} and β_{t-1} at the beginning of period t (before observing the demand q_t). Analogously, $M_{t-1} = e^{\sigma_{t-1}^2/2}$, where σ_{t-1}^2 is the estimate of σ^2 at the beginning of the period t. The first term in the objective function above corresponds to the immediate revenue maximization, while the second term corresponds to maximizing the information (minimizing the variance) about the model parameters. We include a multiplicative term $G(t)$ in the second expression to reflect the time remaining in the planning horizon. For a finite horizon T, we can make $G(t)$ a decreasing function in t, with $G(T) = 0$, since we do not have to learn anymore at the last stage. We can assume, for example, a piecewise linear form $G(t) = (T_c - t)$ for $t < T_c$, and $G(t) = 0$ when $t \geq T_c$, where T_c does not necessarily equal the horizon T. Alternatively, we can use an exponential decaying form $G(t) = Ke^{-t\rho} - Ke^{-K\rho}$ for $t < K$, and $G(t) = 0$ when $t \geq K$.

3 A Formal Analysis of the Trade-Off between Optimization and Learning

We now provide a formal treatment for the trade-off between optimization and learning for an arbitrary parametric model, generalizing the results in the previous

section. The treatment here covers for example the binomial demand model in Carvalho and Puterman (2005). This Section focusses on the case where we have only two periods to go before the end of the sales season. Section 4 considers the general N period model.

Similarly to Subsection 2.2, assume we have observed the responses $\{q_t : t = 1, \ldots, T-2\}$, for the sequence of prices $\{p_t : t = 1, \ldots, T-2\}$, and now we seek to maximize the expected revenues in the last two periods $E[R_{T-1}(p_{T-1})]+E[R_T(p_T)]$. At the last period $t = T$, we are not interested in further learning, because there is no future to consider. Therefore, the price is set to optimize the revenue at period T. Recall that for the log-linear model in this paper, the optimum price, considering the estimate β_{T-1}, will be $p_T = -1/\beta_{T-1}$. The optimum price at period T does not necessarily depend on the entire unknown parameter vector ϑ, but only on a sub vector θ. We will write the vector ϑ as $\vartheta = (\theta', \varphi')' \in \Re^K \times \Re^J$, $K, J \geq 1$, where the optimum price at period T can be written $p_T = p^*(\theta)$. In the log-linear model, $\theta = \beta$ and $\vartheta = (\alpha, \beta, \sigma^2)'$. For the binomial demand model in Carvalho and Puterman (2005), $\theta = (\alpha, \beta)'$ and $\vartheta = (\alpha, \beta, \lambda)'$. The optimum price at epoch decision T is given by $p_T = p^*(\alpha_{T-1}, \beta_{T-1})$, with $p^*(\alpha_{T-1}, \beta_{T-1})$ satisfying $p^*(\alpha_{T-1}, \beta_{T-1}) = \arg\max_{p>0}[p\, e^{\alpha_{T-1}+\beta_{T-1}p}]/[1 + e^{\alpha_{T-1}+\beta_{T-1}p}]$.

At the beginning of period $t = T-1$, the information for θ is provided by the estimate θ_{T-2}, whose distribution, under certain regularity conditions (see Section 4), is approximately normal with mean θ and covariance matrix Σ_{T-2}. After fixing the price p_{T-1} and observing q_{T-1}, we obtain the updated estimate θ_{T-1}, with covariance matrix Σ_{T-1}. Because the θ_{T-1} is subject to estimation error, we expect the optimum price $p_T = p^*(\theta_{T-1})$ to deviate from the true optimum $p^*(\theta)$. Therefore, we may anticipate some loss in the revenue at time T given the uncertainty about θ. For an estimate θ_{T-1}, if we use the rule $p_T(\theta_{T-1}) = p^*(\theta_{T-1})$, the expected revenue in the last period can be written as

$$R_T^*(p^*(\theta_{T-1})) = E\Big[p^*(\theta_{T-1}) \times q(p^*(\theta_{T-1}), \epsilon_T; \vartheta)\Big], \tag{11}$$

where the expectation above is taken with respect to the noise ϵ_T. Obviously $R_T^*(p^*(\theta_{T-1}))$ is maximum when $\theta_{T-1} = \theta$. Theorem 2 below provides an approximation for the expectation $E[R_T^*(p^*(\theta_{T-1}))]$ based on a Taylor expansion. The proof follows the same steps in the proof of Theorem 3, and will be omitted.

Theorem 2 *Given the uncertainty in the parameter estimate θ_{T-1}, the expected revenue at period T, for the price $p_T = p^*(\theta_{T-1})$, can be approximated as*

$$E[R_T^*(p^*(\theta_{T-1}))] = R_T^*(p^*(\theta)) + \frac{1}{2}\text{trace}\Big[\partial_{\theta_{T-1}}\partial_{\theta'_{T-1}} R_T^*(p^*(\theta_{T-1}))\Big|_{\theta_{T-1}=\theta} \times \Sigma_{T-1}\Big] + O((T-2)^{-2}),$$

where the expectation above is calculated with respect to the random variable θ_{T-1}.

Based on the result above, the expected loss at period T associated with the uncertainty in θ_{T-1} can be approximated by

$$R_T^*(p^*(\theta)) - E[R_T^*(p^*(\theta_{T-1}))] \doteq -\frac{1}{2}\text{trace}\Big[\partial_{\theta_{T-1}}\partial_{\theta'_{T-1}} R_T^*(p^*(\theta_{T-1}))\Big|_{\theta_{T-1}=\theta} \times \Sigma_{T-1}\Big]. \tag{12}$$

At $t = T-1$, the fixed price p_{T-1} will affect not only the immediate return $R(p_{T-1})$ but will also affect the covariance matrix $\Sigma_{T-1} = \Sigma_{T-1}(p_{T-1})$ of parameter estimate θ_{T-1}. Therefore, we can choose the price p_{T-1} to maximize

$$F_{T-1}(p_{T-1}) = R_{T-1}^*(p_{T-1}) + \frac{1}{2}\text{trace}\Big[\partial_{\theta_{T-1}}\partial_{\theta'_{T-1}} R_T^*(p^*(\theta_{T-1}))\Big|_{\theta_{T-1}=\theta} \times \Sigma_{T-1}(p_{T-1})\Big]. \tag{13}$$

Note that the first term in the right-hand-side of (13) represents the expected revenue in period $T-1$, while the second term is related to the potential future revenues due to more precise parameters estimates. We may choose a price p_{T-1} which not necessarily provides the maximum revenues at the current period $T-1$, but implies a reduction in the covariance matrix Σ_{T-1}, reducing the average revenue loss in the last period $t = T$.

Equation (13) extends the representation in (9) to a general framework. In fact, for the log-linear demand function, the expected revenue $R_{T-1}^*(p_{T-1})$ at period $T-1$, for price p_{T-1}, is given by $R_{T-1}^*(p_{T-1}) = p_{T-1}e^{\alpha+\beta p_{T-1}}M$. Besides, $p^*(\theta_{T-1}) = p^*(\beta_{T-1}) = -1/\beta_{T-1}$, $R_T^*(p^*(\theta_{T-1})) = -\frac{1}{\beta_{T-1}}Me^{\alpha-\beta/\beta_{T-1}}$ (recall that $\theta_{T-1} = \beta_{T-1}$), and

$$\partial_{\theta_{T-1}}\partial_{\theta'_{T-1}} R_T^*(p^*(\theta_{T-1}))\Big|_{\theta_{T-1}=\theta} = \partial^2_{\beta_{T-1}} R_T^*(p^*(\beta_{T-1}))\Big|_{\beta_{T-1}=\beta} = \frac{Me^{\alpha-1}}{\beta^3}. \tag{14}$$

For the derivation of (14), see the proof of Theorem 1 in the Appendix. Finally, $\Sigma_{T-1}(p_{T-1}) = \sigma^2_{\beta_{T-1}}$ and we obtain (9) from (13).

The objective function in (13) may be used not only to explain the trade-off between revenue optimization and demand learning, but it also can be employed in practice to provide algorithms for dynamic pricing setting when the demand is unknown. Similarly to equation (9), equation (13) involves the parameter vector θ, which is exactly what we want to estimate. When implementing the algorithms in practice, we can replace it by its current estimate, as suggested in Section 2.2. The consequences of replacing θ by its estimate will be addressed in Subsection 4.2 and Section 6.

4 The general multiperiod problem

In this section, we give a formal treatment for the general problem of learning and optimization when we have more than two periods to go. Remember that the demand q_t is governed by a parametric probabilistic model, indexed by the true parameter vector $\vartheta = (\theta', \varphi')' \in \Re^K \times \Re^J$, $K, J \geq 1$, with $(\theta', \varphi')'$ unknown. We can write $q_t = q(p_t, \epsilon_t)$, where ϵ_t, $t = 1, 2, \ldots, T$, are independent random innovations. The decision maker has to fix the price p_t at period t in order to maximize the immediate revenues and gather more information about $(\theta', \varphi')'$ at the same time. Assume that his decision depends on the estimate θ_{t-1} of θ and on the covariance matrix Σ_{t-1} of θ_{t-1}. In the log-linear model above, $\theta = \beta$ and $\varphi = (\alpha \ \sigma^2)'$, and in the binomial demand model $\theta = (\alpha \ \beta)'$ and $\varphi = \lambda$. The policy the decision maker will use is based on the specific rule $p_t = h_t(\theta_{t-1}, \Sigma_{t-1})$, so that the sum of expected revenues from period t to the final period T is given by

$$V_t(h_t(\theta_{t-1}, \Sigma_{t-1})) = R_t(h_t(\theta_{t-1}, \Sigma_{t-1})) + \sum_{k=t+1}^{T} R_k^*(p_k | h_t(\theta_{t-1}, \Sigma_{t-1})), \qquad (15)$$

where $R_t(h_t(\theta_{t-1}, \Sigma_{t-1})) = E_{\epsilon_t}[q(h_t(\theta_{t-1}, \Sigma_{t-1}), \epsilon_t)]$ is the expected revenue at period t, with expectation corresponding to the noise ϵ_t. Similarly, $R_k^*(p_k | h_t(\theta_{t-1}, \Sigma_{t-1}))$ is the maximum expected revenue at time k, $k = t+1, \ldots, T$, given that the price $p_t = h_t(\theta_{t-1}, \Sigma_{t-1})$ is chosen at epoch t. We assume that rule $p_t = h_t(\theta_{t-1}, \Sigma_{t-1})$ is *unbiased*, according to the definition below.

Definition 1 *An optimization policy* $p_t = h_t(\theta_{t-1}, \Sigma_{t-1})$ *is called unbiased if the value function at period* t, $V_t(h_t(\theta_{t-1}, \Sigma_{t-1}))$, *is maximum when* $\theta_{t-1} = \theta$ *(true parameter) and* $\Sigma_{t-1} = \mathbf{0}$.

In the next two subsections, we provide a discussion about the optimization/learning solutions to the dynamic problem in (15). Initially, we treat the case where the retailer wants to set the price at time t, in order to optimize the present and future revenues, provided that the information available at the beginning of t is based on a sequence of fixed prices $\{p_k : k = 1, \ldots, t-1\}$. We will refer to this analysis as static solution, because it only applies to the optimization at a specific period t, given that the previous prices are fixed. If we repeat the same solution at period $t+1$ for example, the analysis in Subsection 4.1 will not be valid anymore, because p_t, in the history $\{p_k : k = 1, \ldots, t\}$ will be a random variable, as will be discussed in Subsection 4.2. Nonetheless, Subsection 4.1 provides the basis for the more realistic analysis in Subsection 4.2, where the sequence of prices $\{p_k : k = 1, \ldots, t\}$ is random.

4.1 Static solution based on fixed prices

Initially, we will consider the optimization/learning problem at period t, where the retailer has to fix the price p_t in order to maximize the present and future revenues, based on a sequence of fixed prices $\{p_k : k = 1, \ldots, t-1\}$ and on a sequence of observed demands $\{q_k : k = 1, \ldots, t-1\}$. A discussion about the implication of relaxing the hypothesis of fixed previous prices is discussed in Subsection 4.2. After setting the price p_t and observing the demand q_t, the retailer obtains the estimate θ_t with covariance matrix Σ_t. We will assume that both the bias and the covariance matrix of the estimator θ_t are of order $O(t^{-1})$, so they go to zero at rate t^{-1}. This is normally the case for usual estimators in parametric models (the reader can refer to Lehmann, 1999, pp. 233, for more details). After obtaining θ_t, the retailer fix the price p_{t+1} according to the policy $p_{t+1} = h_{t+1}(\theta_t, \Sigma_t)$. Because θ_t is a random variable, so is $V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))$, and from the distribution of $h_{t+1}(\theta_t, \Sigma_t)$ it is possible to find the distribution of $V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))$. To derive the price decision at epoch t, consider the following theorem, which provides an approximation for the expectation $E_{\theta_t}\{V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\}$ of the value function,

based on a second order Taylor expansion. The proof is provided in the Appendix.

Theorem 3 *Consider the optimization/learning decision at time $t+1$ based on the decision rule $p_{t+1} = h_{t+1}(\theta_t, \Sigma_t)$, where θ_t is the estimate for θ and Σ_t is the corresponding covariance matrix, based on the prices $p_1, p_2, \ldots, p_t$. Assume that:*
(A) the prices $p_1, p_2, \ldots, p_t$ are nonstochastic;
(B) the policy $p_{t+1} = h_{t+1}(\theta_t, \Sigma_t)$ is unbiased;
(C) the function $V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))$ is twice continuously differentiable at $\theta_t = \theta$ and $\Sigma_t = \mathbf{0}$;
(D) the estimator θ_t has both bias $E\{\theta_t - \theta\}$ and covariance matrix Σ_t of order $O(t^{-1})$.
Then, we can write the approximation

$$E_{\theta_t}\{V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\} = V_{t+1}(h_{t+1}(\theta, \mathbf{0})) + \frac{1}{2}\text{trace}[\Sigma_t \mathbf{A}_{t+1}] + O((t-1)^{-2}), \quad (16)$$

where

$$\mathbf{A}_{t+1} = \left[\partial_{\theta_t}\partial_{\theta_t'} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta}\right]. \quad (17)$$

If θ is real, the second term in the right-hand side of (16) becomes

$$\frac{1}{2}\text{trace}[\Sigma_t \mathbf{A}_{t+1}] = \frac{1}{2}\left[\partial^2_{\theta_t} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta}\right]\sigma^2_{\theta_t}, \quad (18)$$

where $\sigma^2_{\theta_t}$ is the variance of θ_t. This is the case for the simple log-linear model in Section 2. Because of the unbiasedness assumption for the rule $h_{t+1}(\theta_t, \Sigma_t)$, the value function $V_{t+1}(h_{t+1}(\theta_t, \mathbf{0}))$ is concave at the true θ, and the expected loss in terms of total expected revenues, given the uncertainty (variance) in the estimate θ_t, can be approximated by

$$L_{t+1} = V_{t+1}(h_{t+1}(\theta, \mathbf{0})) - E_{\theta_t}\{V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\} \doteq -0.5\text{trace}[\Sigma_t \mathbf{A}_{t+1}]. \quad (19)$$

The loss L_{t+1} given in (19) depends on the covariance matrix of θ_t, which depends on the prices from period $k = 1$ up to $k = t$. Therefore, the pricing problem at epoch t can be solved by choosing the price p_t that maximizes the objective function

$$G_t = R_t(p_t) + 0.5\text{trace}[\Sigma_t(p_t)\mathbf{A}_{t+1}]. \quad (20)$$

In equation (20), we wrote $\Sigma_t = \Sigma_t(p_t)$ to make explicit the dependence of Σ_t on the price set at epoch t. If $\theta \in \Re$, we can rewrite (20) as

$$G_t = R_t(p_t) + 0.5\Big[\partial^2_{\theta_t} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta}\Big]\sigma^2_{\theta_t}(p_t). \tag{21}$$

The inclusion of a discount factor $\gamma \in (0,1)$ in the analysis can be accommodated by replacing (15) and (20) by

$$V_{t+1}(h_{t+1}(\theta_t, \Sigma_t)) = R_{t+1}(h_{t+1}(\theta_t, \Sigma_t)) + \sum_{k=t+2}^{T} \gamma^{k-t+3} R^*_k(p_k|h_{t+1}(\theta_t, \Sigma_t)) \tag{22}$$

and

$$G_t = R_t(p_t) + 0.5\gamma \text{trace}[\Sigma_t(p_t)\mathbf{A}_{t+1}]. \tag{23}$$

Equations (20), (21) and (23) provide the solution for the trade-off learning versus optimization at time t, assuming the prices $p_1, \ldots, p_{t-1}$ are fixed. Even when p_t is random (it may depend on the previous estimate $(\theta'_{t-1}, \hat{\varphi}'_{t-1})'$), the approximation (19), for the expected revenue loss, is still valid, since the importance of p_t in the sequence $\{p_1, \ldots, p_{t-1}, p_t\}$ vanishes as $t \to \infty$. Subsection 4.2 presents a further discussion for the issue of nonstochastic versus random prices. Section 6 discusses the reasons why the one-step-ahead pricing policy is still valid even when the prices are random.

For general parametric forms for the demand function q_t, it is possible to write down the variance of θ_t as a function of p_t, given the history $p_1, \ldots, p_{t-1}$. Therefore, the only term missing in the implementation of the static solution at time t, based on (20) or (23), is the matrix $\mathbf{A}_{t+1}$ of second partial derivatives of the value function $V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))$ at the true θ. In Carvalho and Puterman (2005) and in Section 5, heuristic approximations based on piecewise linear and exponential decaying forms for $G(t)$ are employed. Similarly to Carvalho and Puterman (2005), the simulations in Section 5 suggest the validity of the one-step-ahead pricing strategy.

4.2 Dynamic solution based on random prices

The optimization/learning solution at time t discussed in the previous subsection assumes that sequence of prices $\{p_k : k = 1, \ldots, t-1\}$ is fixed. However, the objective functions in (20) and (23) are also functions of the unknown parameter

vector $(\theta', \varphi')'$, and to find the optimum price p_t, we have to replace $(\theta', \varphi')'$ by its estimate $(\theta'_{t-1}, \hat{\varphi}'_{t-1})'$. Therefore, the price p_t chosen by maximizing (20) or (23) will also be a function of the estimates $(\theta'_{t-1}, \hat{\varphi}'_{t-1})'$, which are random variables depending on the history of the random sequence $\{q_k : k = 1, \ldots, t-1\}$. It implies that the price p_t is also a random variable, and the sequence $\{p_1, \ldots, p_{t-1}, p_t\}$ now contains $t-1$ fixed elements and 1 random component. This fact does not invalidate the approximations based on Theorem 3, since the importance of the pair (p_t, q_t) in the estimate θ_t, and its covariance matrix Σ_t, vanishes as t goes to infinity.

If we employ the optimization rule recursively, by maximizing the objective functions (20) or (23), at periods $k = t$ to $k = T$, with $(\theta', \varphi')'$ replaced by $(\theta'_{k-1}, \hat{\varphi}_{k-1})'$, the whole sequence of prices $p_t, \ldots, p_T$ is random. Therefore, the importance of the random prices on the sequence $\{p_k : k = 1, \ldots, t-1, t, \ldots, s\}$ increases as s becomes large relatively to $t-1$. In reality, because we want to find a solution to the optimization/learning problem from period $t = 1$ to period $t = T$, the whole sequence of prices $\{p_t : t = 1, \ldots, T\}$ will be random, assuming the initial estimate $(\theta'_0, \hat{\varphi}'_0)'$ is random[1]. The randomness of the price sequence compromises some of the hypothesis in Theorem 3. In fact, as will be discussed in Section 6, the estimator θ_t presents a bias $E\{\theta_t - \theta\}$, which does not go to zero as t increases. Fortunately, although the assumptions in Theorem 3 do hold anymore, Subsection 6.2 presents arguments to show that a recursive policy based on the sequential optimization of the objective functions (20) or (23) may still be valid.

The dynamic pricing algorithms based on the maximization of the objective function in (20) and (23) can be viewed as an adaptive control problem. In a general adaptive control model, the control variable x_t (p_t in our case) is derived from a control law $x_t = f(\theta_{t-1})$ which assumes that θ_{t-1} is an estimate for the unknown parameter θ based on a sequence $x_1, \ldots, x_{t-1}$ assumed fixed. However, because $x_1, \ldots, x_{t-1}$ are not random and depend on the previous estimates $\theta_0, \theta_1, \ldots, \theta_{t-2}$, the estimator θ_t does not converge in probability or almost surely to the true parameter vector θ, as discussed in Campi and Kumar (1998), Chen and Guo (1988), Kumar (1990) and Sternby (1977). Subsection 6.1 extends this discussion to the log-linear demand model discussed in Sections 2 and 5.

[1]Even if the initial estimates are not random, they will eventually be random for some small t, so that the general conclusions remain the same.

5 Monte Carlo Simulation

In this section, we present and discuss results of an extensive Monte Carlo simulation that investigates the performance of the one-step ahead policy and other heuristic pricing strategies. We begin with a discussion of the simulation setup.

5.1 Simulation Design

We describe the classes of price selection rules that will be compared in the simulation study.

1. *Myopic rule.* The simplest strategy is the myopic rule, which at each period t sets the price $p_t = -1/\beta_{t-1}$, where β_{t-1} is the most recent estimate of the regression slope. We will see that this strategy produces prices which get "stuck" at a level far away from the optimum $p^* = -1/\beta$ and do not benefit from learning.

2. *Myopic rules with random exploration.* An alternative to the myopic rule is to choose the optimum price $p_t = -1/\beta_{t-1}$ with probability $1 - \eta_t$ and choose a random price with probability η_t (Sutton and Barto, 1998). Because learning is more important at initial periods, we let $\eta_t \to K_0$ when $t \to \infty$, where K_0 equals zero or a very small value, in the case we wish to continue experimenting indefinitely. We used an exponential decay function, $\eta_t = K_0 + K_1 e^{-K_2 t}$, with different values for the parameters K_0, K_1 and K_2. Some care must be taken here, because when implementing the proposed methodology in practice, prices must be chosen in an economically viable range. Further, since the proposed parametric model is only an approximation for the real data generating process, the approximation may be reasonable only for a limited range of prices. In light of this when learning, we choose random prices periods from a uniform distribution on a pre-specified interval $[p_l, p_u]$.

3. *Softmax exploration rules.* An alternative to the myopic rule with random exploration is to use the softmax exploration rule described in Sutton and Barto (1998). The basic idea is to draw, at each time period t, the price p_t from the distribution with density

$$f(p_t) \propto \exp\big\{[p_t e^{\alpha_{t-1}+\beta_{t-1}p_t} e^{\sigma^2_{t-1}/2}]/\tau_t\big\}, \tag{24}$$

with $\tau_t \to 0$ as $t \to \infty$. The density in (24) has a single mode at $-1/\beta_{t-1}$ and,

as $\tau_t \to 0$, it becomes more concentrated around the mode, so that, in the limit, we only select the price $p_t = -1/\beta_{t-1}$, and the softmax rule becomes equivalent to the myopic policy. The same way as before, we used $\tau_t = K_0 + K_1 e^{-K_2 t}$.

4. *Optimal design rules.* Another approach to price selection is to choose a "statistically" optimal design in terms of model estimation during the first C periods, and then apply the myopic rule for the rest of the process. We may think of this as acting as a "statistician" from $t = 1$ to $t = C$, and as an "optimizer" from $t = C + 1$ to $t = T$. Therefore, for $t = 1, \ldots, C$, we select $p_t = p_u$ if t is odd and $p_t = p_l$ if t is even. From $t = C + 1$ to $t = T$, we use $p_t = -1/\beta_{t-1}$.

5. *One-step look ahead rules.* Less arbitrary strategies are based on the one-step ahead rule, which explicitly account for the trade-off between learning and revenue maximization through (10). As noted in Section 2, we use functions $G(t)$ with piecewise linear, $G(t) = \max\{T_c - t, 0\}$, and exponential decaying, $G(t) = \max\{Ke^{-t\rho} - Ke^{-K\rho}, 0\}$, functional forms. Alternatively, to overcome the lack of consistency of the ordinary least squares estimator, discussed in Subsection 6.1, we also simulated modified versions of the one-step ahead rules, by performing random exploration with constant probability $\eta_t = 0.01$. Similarly to the myopic policy with random exploration, at the exploration stage, the prices p_t were drawn from a uniform distribution on $[p_l, p_u]$. We refer to these strategies as unconstrained one-step ahead rules (to differentiate from the policies described below) or simply one-step ahead rules.

6. *Price constrained one-step ahead rules.* Figure 4 below shows that the prices vary considerably at the beginning of the process, in order to allow for faster learning. In practice, a manager may wish to avoid such abrupt price changes. We explored this possibility by imposing a limit on period to period price changes. Given the price p_t, the price p_{t+1} at decision epoch $t + 1$ is restricted to be in the interval $[p_t - 0.25p_t, p_t + 0.25p_t]$.

In all the above policies, we restricted the prices to be within the range $[p_l, p_u]$. Therefore, if, at a certain period t, the calculated optimum price is $p_t > p_u$, we used $p_t = p_u$. Analogously, if the calculated optimum price is $p_t < p_l$, we used $p_t = p_l$. In the different policies described above, we tried different values for K_0, K_1, K_2, C, T_c, K and ρ, and the results reported here correspond to the configurations providing the best performances.

For all the strategies described above, the Kalman filter updating equations (3) - (5) require initial values $\delta_0 = [\alpha_0 \ \beta_0]'$ for the regression parameters $\delta = [\alpha \ \beta]'$ and for the matrix P_0. Besides, all strategies except the optimum design rules require an initial value β_0 to set the price p_1. To avoid any bias related to wrong prior information, we assumed that we had information from two previous data points $[\log q_{-2} \ p_{-2}]'$ and $[\log q_{-1} \ p_{-1}]'$, with $\log q_{-i} = \alpha + \beta p_{-i} + \epsilon_{-i}$, $\epsilon_{-i} \sim \mathrm{N}(0, \sigma^2)$, $i = 1, 2$ [2]. Avoiding bias in the initial values is particularly important when studying the bias in the ordinary least squares estimator. If incorrect prior information were used, one may argue that the bias observed in the estimates $\hat{\beta}_t$ is due to this initial misleading set up.

Based on the discussion above, we set the initial matrix $P_0 = (Z'_{-1}Z_{-1})^{-1}$, for Z_{-1} a two by two matrix, $Z_{-1} = \left[[1 \ p_{-2}]', [1 \ p_{-1}]'\right]'$, with p_{-2} and p_{-1} the same for all simulation replicates. The vector δ_0 is equal to $(Z'_{-1}Z_{-1})^{-1}Z'_{-1}Y_{-1}$, with $Y_{-1} = [\log q_{-2} \ \log q_{-1}]'$. Note that, although δ_0 is a random vector, it has expectation equal to the true vector δ and covariance matrix equal to $\sigma^2 P_0$, so that we are not biasing the conclusions due to wrong priors. In the simulation results presented in this paper, we fixed $p_{-2} = p_u$ and $p_{-1} = p_l$. We also tried other values for p_{-2} and p_{-1}, but the conclusions remained the same. It is important to mention that all the rules considered in the simulation, including the "optimal design rule", benefited from the fact that we used correct information about δ_0 and P_0.

Specifically for the one-step ahead strategies, we need, at each decision period t, $t = 1, \ldots, T$, an estimate for the variance σ^2. Because the prior information for σ^2 will affect only the one-step ahead strategies, we decided not to worry about wrong initial values for σ_0^2. The idea of having two extra data points $[\log q_{-2} \ p_{-2}]'$ and $[\log q_{-1} \ p_{-1}]'$ does not provide enough degrees of freedom to estimate σ^2. Therefore, at time $t = 1$, the one-step ahead rule was based on $\sigma_0^2 = \sigma^2/2$ (wrong prior). After fixing the price p_1 and observing $\log q_1$, we have three data points in total, what makes it possible to obtain the first estimate σ_1^2, used at the decision epoch $t = 2$. In fact, for $t = 1, \ldots, T$, we can use the ordinary least squares estimator $\sigma_t^2 = \frac{1}{t}[\hat{\epsilon}_{-2} + \hat{\epsilon}_{-1} + \sum_{k=1}^{t} \hat{\epsilon}_k^2]$, where $\hat{\epsilon}_k = y_k - \alpha_t - \beta_t p_k$, $k = -2, -1, 1, \ldots, t$. To evaluate the effect of the choice or prior σ_0^2, we also performed simulations with $\sigma_0^2 = 2\sigma^2$, but the general conclusions did not change.

[2] We used the indices p_{-2} and p_{-1}, instead of p_{-1} and p_0, to make explicit that the information is available before the first decision period $t = 1$.

To compare these different strategies, we performed $L = 10,000$ simulations for each policy and computed the cumulative revenues in each simulation

$$\mathrm{CR}(t) = \sum_{k=1}^{t} R_k(p_k), \quad t = 1, \ldots, T. \tag{25}$$

The expected cumulative revenues can then be estimated by the sample means

$$\hat{E}[\mathrm{CR}(t)] = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{t} R_k(p_k). \tag{26}$$

By plotting the path of $\hat{E}[\mathrm{CR}(t)]$ against t, $t = 1, \ldots, T$, we gain insight into how these different computational strategies perform. In general, we focus on maximizing revenues in short planning horizon ($T = 100$ or $T = 200$). On the other hand, it is interesting to look at the path of other measures as t tends to infinity. The sample paths for the estimate β_t, for example, provide insight into the long run convergence of the model parameters under each of these computational methods.

5.2 Simulation Results

The simulations show that the unconstrained one-step ahead rules provide greater mean cumulative revenues $\hat{E}[\mathrm{CR}(t)]$ than the other strategies. Figure 1 provides a comparison of a selected one-step ahead rule in which $G(t)$ is piecewise linear, and the other rules. A comparison between several one-step pricing rules is shown in Figure 2.

For these simulations the parameter values are set to $\alpha = 8.0$, $\beta = -1.5$ and $\sigma^2 = 5.0$. The optimum price in this case is $p^* = 0.667$, which implies that the maximum mean cumulative revenues equal to $8,906.5$, when $p_t = p^*$ for all $t = 1, \ldots, T$. The minimum allowed price was $p_l = 0.167$ and the maximum allowed price was $p_u = 3.00$.

After 100 periods, by using the one-step ahead rule we obtain a relative gain of at least 3.7% over all the none one-step ahead rules. This relative gain is equal to 3.0% after 200 periods and equal to 2.4% after 400 periods. Note that the myopic rules performed poorly by getting "stuck" at a price level away from the optimum. The policies with optimal statistical design at the beginning of the pricing process perform better than the policies with random exploration (myopic rule with

random exploration and the softmax rule) during the initial periods. However, as the random exploration policies keep learning about the model parameters, they eventually outperform the statistical design rules.

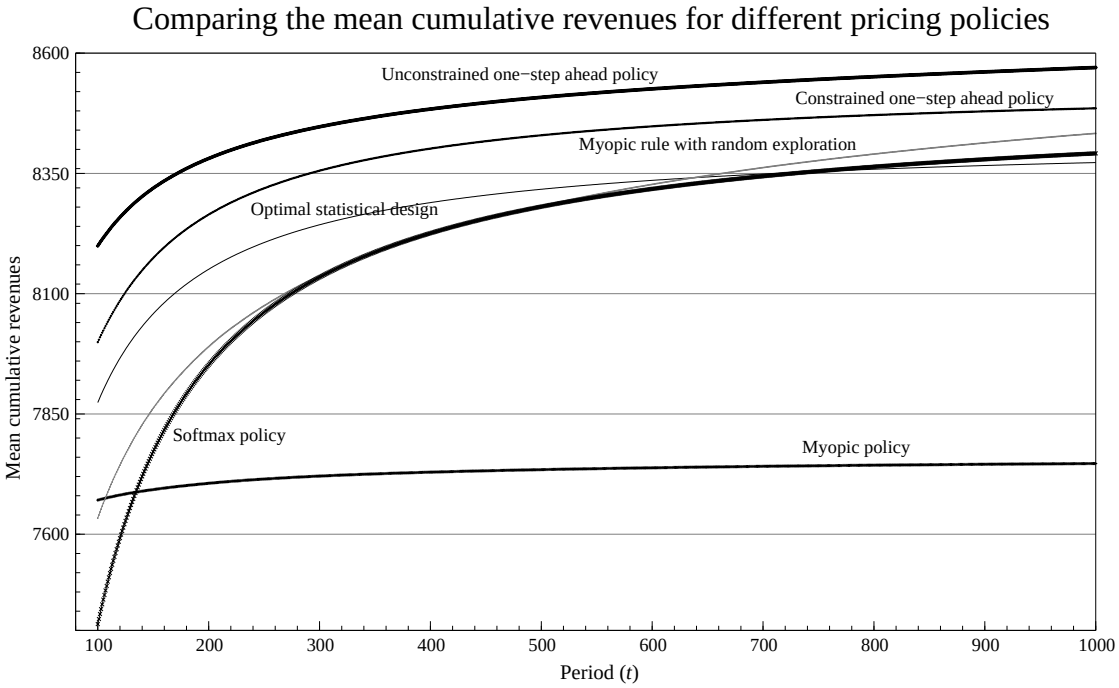

Figure 1: Comparison of expected mean cumulative revenues for several pricing policies (optimal expected revenue per period under known parameters values is $8,906$).

Figures 1 also displays the mean cumulative revenues for a one-step ahead policy with price change constraints. At each time period $t \in \{2, \ldots, T\}$, the prices were chosen by maximizing the objective function in (10) with the restriction $p_{t+1} \in [p_t - 0.25p_t, p_t + 0.25p_t]$. At the initial period $t = 1$, the price was only restricted to be within the interval $[p_l, p_u]$. We considered piecewise linear $G(t)$ with $T_c = 170$ (fast learning) and $T_c = 70$ (slow learning). To simplify the presentation, the results for the slow-learning case are shown in Figure 2. Although none of the other rules had price change restrictions, the constrained one-step ahead policies still presented a superior performance when compared to the rules other than the one-step ahead ones. As we already expected, the one-step ahead policy with fast learning performs better than the one of slow learning methods after some initial periods. To validate the analysis, we performed other simulations with different choices of model parameters, and minimum and maximum allowed prices, and the conclusions remained the same.

Figure 2 presents the mean cumulative revenues for the six one-step ahead strategies studied here. Note that there does not seem to be any significant difference between the four unconstrained policies. For $t < 400$, the rule with exponential decaying $G(t)$, without random exploration seems to slightly outperform the other ones. For $t > 400$, the policy with piecewise linear $G(t)$ and random exploration with $\eta_t = 0.01$ presents a somewhat better performance than the others. By imposing the price change constraint, the relative loss in the one-step-ahead policies is not higher than 1.2% after 100 periods, 1.4% after 200 periods, and 1.0% after 400 periods.

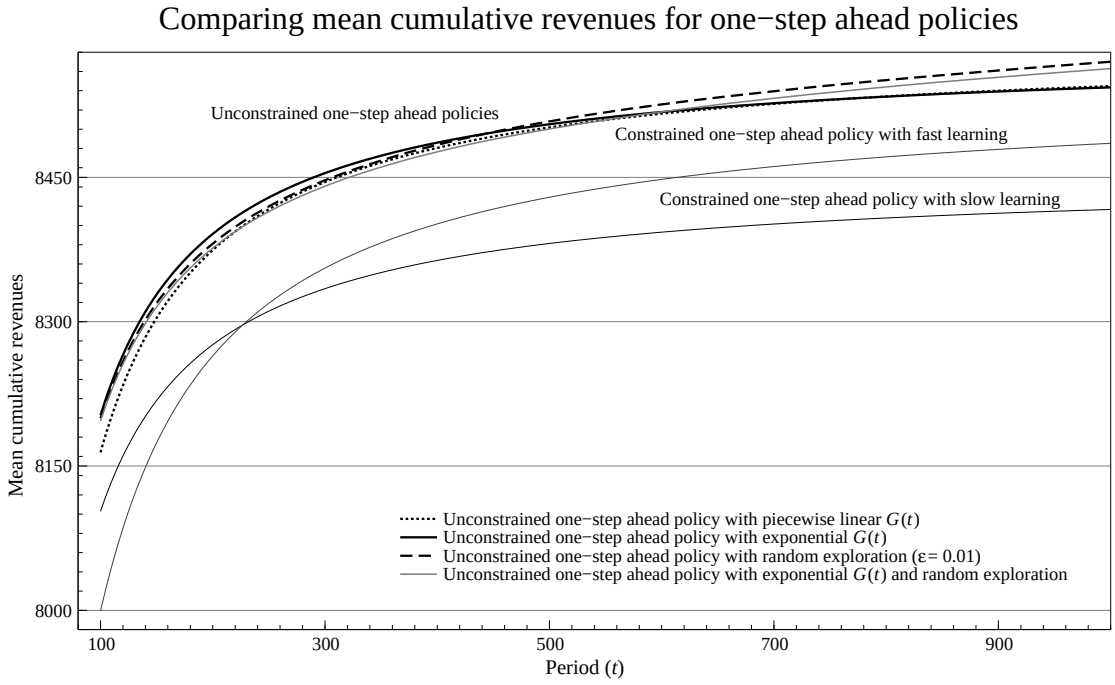

Figure 2: Expected mean cumulative revenues for several one-step ahead policies (optimal expected revenue per period under known parameters values is $8,906$).

In Figure 3, we plot the mean paths of the estimated slope β, for the different strategies. We observe that all strategies produce biased estimates of β for all $t = 1, \ldots, 1000$. This effect is real and is supported by theory. The reason for this will be discussed in Subsection 6.1. However, for the myopic rule with random exploration and the one-step ahead rules with random exploration, the bias tends to go to zero, as t grows, what was already expected based on the discussion about adaptive control with random perturbation presented in Subsection 6.1. An additional strategy, which sets random prices at all periods $t = 1, \ldots, 1000$, was also simulated and produced unbiased estimates for β. However, its revenue perfor-

mance was very poor, since it never uses the produced estimates for optimization purposes. For the one-step ahead policies, the bias is quite significant. However, because of asymmetry in the revenue function, the loss incurred by a negative bias is not as harmful as that incurred by a positive bias. This phenomenon has been previously observed in the inventory literature as for example by Silver and Rahnama (1987).

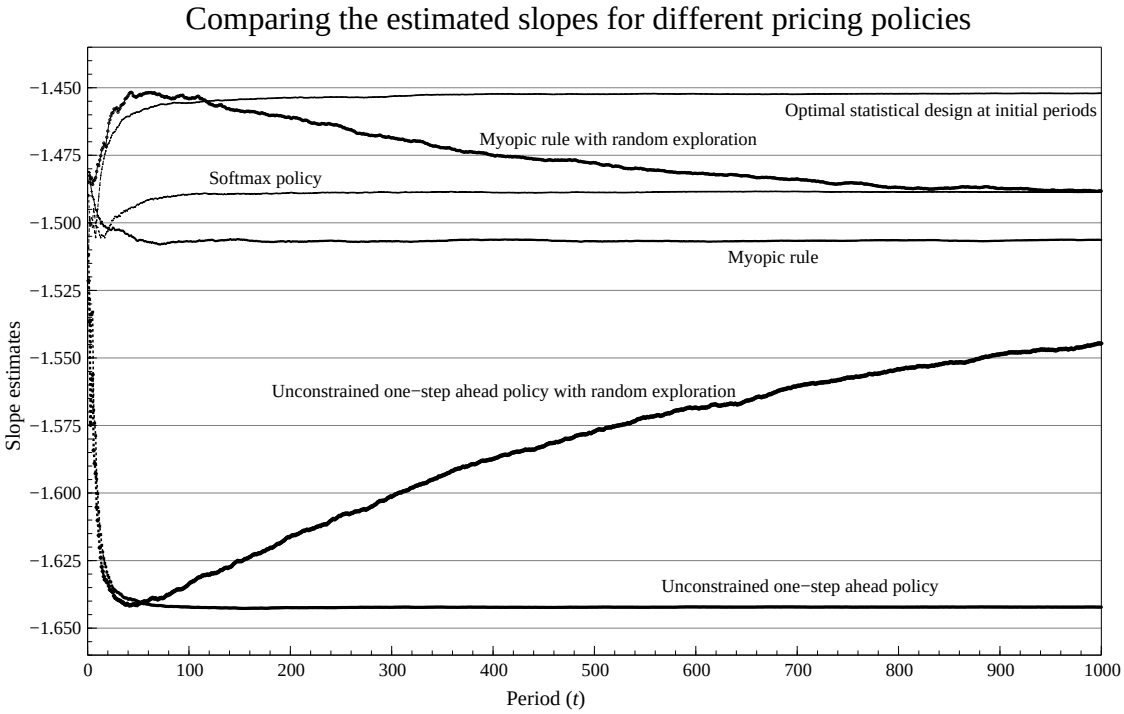

Figure 3: Mean estimates for the slope coefficient β in the log-linear model.

Finally, Figure 4 shows the mean paths of selected prices for some of the different strategies. For almost all the policies, the prices get stuck at a fixed level after 80 periods. For the myopic rule with random exploration, although the prices are set initially in a level above the optimal price $p^* = 0.667$, they tend to approach p^* as t grows and there is more exploration about the true slope value. For the one-step ahead rule specifically, the prices tend to go up and down, with the variations around p^* decreasing as more information is added. It illustrates the idea behind the one-step ahead policies: as more information is obtained, the marginal value of extra information decreases and the algorithm values more the maximization of immediate revenues. Note that, although the estimates β_t are biased, as shown in Figure 3, the mean prices in the one-step ahead policies converge to levels very close to the optimal price $p^* = 2/3$. This can be explained by

the nonlinearity in the function $p_t = -1/\beta_{t-1}$, so that $E[p_t] \neq -1/E[\beta_{t-1}]$. For the two constrained one-step ahead policies, note the smoother evolution of the chosen prices, when compared to the price paths for the unconstrained one-step ahead rules. The constrained one-step ahead policy with fast learning presents a higher price variation during the initial periods than the constrained one-step ahead policy with low learning. This fact was already expected, given the higher weight for the learning component (second term in the right-hand-side of equation (10)) in the fast-learning case.

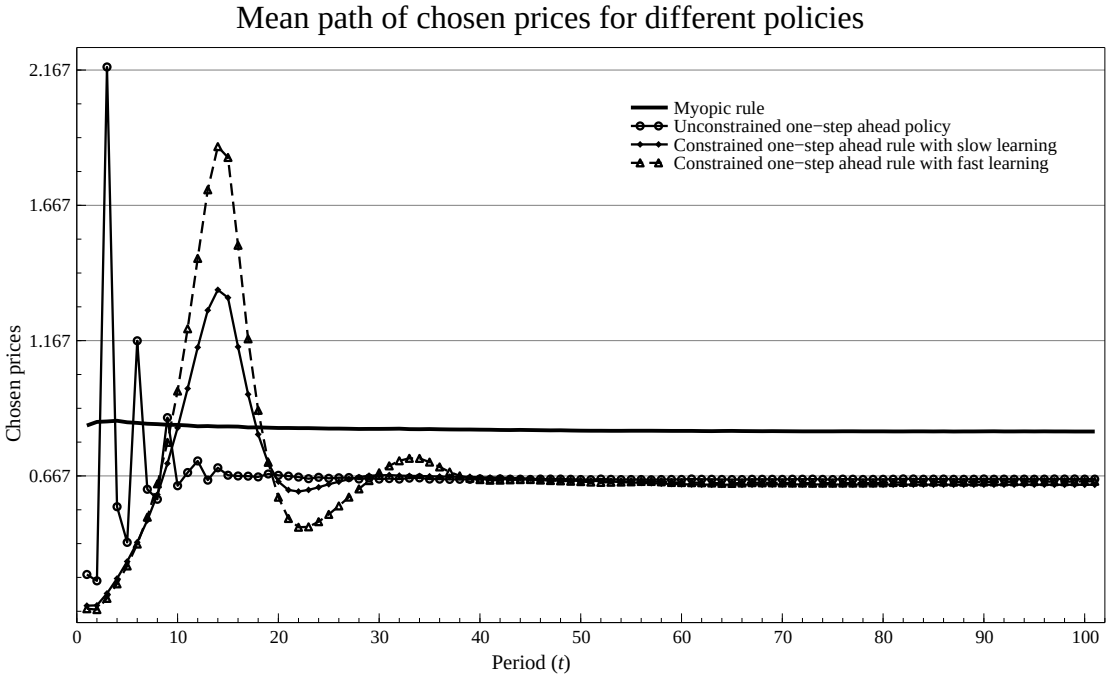

Figure 4: Comparison of chosen prices for several policies (optimal price $p^* = 1/1.5$).

6 Random Prices and Estimation Bias

This section focuses on two important technical issues that underlie the observed bias of the regression parameters in the previous section and the derivation of the Taylor series expansion which is the basis for the one step ahead rules in Sections 2, 3 and 4. Initially, we discuss the source of bias in the sequencial parameter estimates. We then present a discussion for why the one-step-ahead policies are still valid, even in the presence of estimate bias.

6.1 Bias in Parameter Estimates θ

We now discuss why the estimates of the regression parameters may be biased when prices are chosen adaptively. This phenomenon was observed in Figure 3 which showed that estimates of the regression parameter β did not converge to their true value.

To understand the reason for the bias of $\hat{\beta}_t$, the estimate of β based on t observations, consider the usual ordinary least squares estimator $\hat{\delta}_t = (Z_t'Z_t)^{-1}Z_t'Y_t$, where Z_t is the $t \times 2$ design matrix $[[1\ p_1]', [1\ p_2]', \ldots, [1\ p_t]']'$ and Y_t is the $t \times 1$ response vector $[\log q_1\ \log q_2\ \ldots\ \log q_t]'$. Although the parameter vector $\delta = [\alpha\ \beta]'$ is recursively estimated with the Kalman filter, given the choice of prior $N(\delta_0, \sigma^2 P_0)$ employed in our simulations, the resulting estimate $\hat{\delta}_t$ is numerically equivalent to the ordinary least squares (OLS) estimator $(Z_t'Z_t)^{-1}Z_t'Y_t$. According to the classical regression analysis theory (see Draper and Smith, 1998) when Z_t is fixed, the expected value of δ is equal to

$$E\{\hat{\delta}_t\} = E\{(Z_t'Z_t)^{-1}Z_t'Y_t\} = (Z_t'Z_t)^{-1}Z_t'E\{Y_t\} = (Z_t'Z_t)^{-1}Z_t'E\{Z_t\delta + \upsilon_t\}, \quad (27)$$

where $\upsilon_t = [\epsilon_1\ \epsilon_2\ \ldots\ \epsilon_t]'$. Because $E\{\upsilon_t\} = 0$, we conclude that $E\{\hat{\delta}_t\} = \delta$, and hence $\hat{\delta}_t$ is unbiased for fixed Z_t.

As we discussed above, the sequence of prices $p_1, \ldots, p_t$ is usually random, and hence the design matrix Z_t is not fixed. Therefore, the classical theory for OLS estimation is not valid in this case specifically, and we cannot ensure the unbiasedness of $\hat{\delta}_t$. Besides, following the same derivation in (27), we have

$$E\{\hat{\delta}_t\} = \delta + E\{(Z_t'Z_t)^{-1}Z_t'\upsilon_t\}, \quad (28)$$

and if Z_t and υ_t were independent, it is easy to show that the second term in (28) would be zero and the OLS estimator would be still unbiased. However, because the price p_k set at period k depends on the estimate $\hat{\delta}_{k-1}$, and the estimate $\hat{\delta}_{k-1}$ depends on the history of disturbances $\epsilon_1, \ldots, \epsilon_{k-1}$, we conclude that the p_k depends on $\epsilon_1, \ldots, \epsilon_{k-1}$. Therefore, the random variables Z_t and υ_t are not independent and we cannot guarantee that $E\{(Z_t'Z_t)^{-1}Z_t'\upsilon_t\} = 0$, so it is expected that $\hat{\theta}_t$ is biased for finite t.

Although $\hat{\delta}_t$ is biased for finite t, one may be interested in the behavior of $\hat{\delta}_t$ as $t \to \infty$. As is well known in the econometrics literature, when Z_t is random, under

some regularity conditions, the estimate $\hat{\delta}_t$ is strongly consistent (i.e., converges almost surely) to the true parameter δ, in the sense that $\hat{\delta}_t \xrightarrow{a.s.} \delta$ as $t \to \infty$, as discussed in White (2001). These conditions, interpreted in the context of the pricing model, are that the random sequence of prices $\{p_t : t = 1, \ldots, \infty\}$ satisfies a strong Law of Large Numbers and that there exists a $\Delta > 0$, such that the sequence of minimum eigenvalues $\lambda_{min,t}$ of $Z_t'Z_t$ satisfies $\lambda_{min,t} > \Delta$ for $t = 1, \ldots, \infty$ with probability 1. However, for the one-step ahead rules the sequences of prices $\{p_t : t = 1, \ldots, \infty\}$ does not satisfy either of these two conditions. In particular, because Z_t contains a column of ones and for each simulation replicate the prices approach a constant value as indicated in Figure 4, the smallest eigenvalue of $Z_t'Z_t$ converges to zero as t goes to infinity. Besides, looking at different sample paths for different simulations (not shown here), the price level to which the sequences of prices converges varies across the simulation replicates. Therefore, the Law of Large Numbers does not apply to the price sequence. We conclude that the usual conditions for consistency of the ordinary least squares estimator are not satisfied, and we cannot guarantee that $\hat{\delta}_t \xrightarrow{a.s.} \delta$ as t goes to infinity.

Some of these issues have been addressed in the adaptive control literature by Campi and Kumar (1998), Chen and Guo (1988), Kumar (1990) and Sternby (1977) who show that the parameter estimates $\hat{\delta}_t \xrightarrow{a.s.} \delta_\infty$, where δ_∞ depends on the random path of the state variable which in this example is the demand sequence $\{q_t\}_{t=1}^{\infty}$. Further $\delta_\infty \neq \delta$. To guarantee the consistency of the estimates $\hat{\delta}_t$ to the true parameter vector δ in adaptive control problems Campi and Kumar (1998), and Chen and Guo (1988) suggest the addition of persistent yet infrequent random perturbations to the control law. These perturbations should be small in magnitude and sufficiently infrequent, so that they do not incur a high extra cost. Specifically for our dynamic pricing problem, the addition of random perturbations can be accomplished by choosing the price p_t according to the objective function (10) with probability $1 - \eta_t$, and drawing p_t from a uniform distribution with probability η_t, with η_t very low. Although the addition of the random experimentation guarantees the consistency of $\hat{\delta}_t$, it does not improve the performance of the one-step ahead rules over short horizons.

The use of biased parameter estimates also has some precedence in the inventory literature as for example in Silver and Rahnama (1987).

6.2 Validity of the one-step-ahead rule

A crucial assumption in the derivations for Theorem 1 is that the sequence of prices $\{p_1, p_2, \ldots, p_{T-2}\}$ is fixed, or, if it is random, the dependence between p_t and p_{t+k} vanishes as $k \to \infty$. However, if we employ the optimization rule recursively by maximizing the objective function $\hat{F}_t(p_t)$ in (10), at each period t, the optimum price p_t will be a function of the estimates α_{t-1}, β_{t-1} and σ^2_{t-1}, which are random variables calculated using the sequence of prices $\{p_1, p_2, \ldots, p_{t-1}\}$. Therefore, the price p_t will also be a random variable and will depend on the sequence $\{p_1, p_2, \ldots, p_{t-1}\}$. If the initial estimate $[\hat{\alpha}_0 \ \hat{\beta}_0 \ \hat{\sigma^2}_0]'$ is random, and we use the objective function in (10) to recursively update the prices, the whole sequence $\{p_t : t = 1, 2, \ldots, T\}$ will be random. The randomness of the price sequence compromises the derivation of Theorem 1. In fact, as discussed in Subsection 6.1 and illustrated in the simulations in Section 3, the bias of $\hat{\beta}_t$, bias $E\{\hat{\beta}_t - \beta\}$, does not converge to zero as t increases. Fortunately, although the assumption of non-randomness of $\{p_1, p_2, \ldots, p_T\}$ does not hold, the Taylor series approximation, on which the one-step ahead policies are based, may still be valid. In this subsection, we give a informal discussion of why the one-step ahead rules work well even though the assumptions on which they are based do not hold.

To understand the problems caused by the inconsistency of $\hat{\theta}_t$, consider the following approximation, based on the Taylor expansion in (31), presented in the proof of Theorem 1 in the Appendix.

$$\begin{aligned} E\{R^*_T(p_T(\beta_{T-1}))\} \doteq R^*_T(p_T(\beta)) + \partial_{\beta_{T-1}} R^*_T(p_T(\beta)) E\{\beta_{T-1} - \beta\} \\ + \frac{1}{2}\partial^2_{\beta_{T-1}} R^*_T(p_T(\beta)) E\{[\beta_{T-1} - \beta]^2\}. \end{aligned} \tag{29}$$

Because of the inconsistency of β_{T-1}, the term $E\{[\beta_{T-1} - \beta]^2\}$ in (29) is not equal to the variance β_{T-1} anymore. In this case, we have $E\{[\beta_{T-1} - \beta]^2\} = \mathrm{MSE}_{\beta_{T-1}} \neq \mathrm{Var}(\beta_{T-1})$, even for large T. Therefore, the approximation in (29) can be rewritten as

$$\begin{aligned} E\{R^*_T(p_T(\beta_{T-1}))\} \doteq R^*_T(p_T(\beta)) + \partial_{\beta_{T-1}} R^*_T(p_T(\beta)) E\{\beta_{T-1} - \beta\} \\ + \frac{1}{2}\partial^2_{\beta_{T-1}} R^*_T(p_T(\beta)) \mathrm{MSE}_{\beta_{T-1}}, \end{aligned}$$

and the objective function to be maximized in the recursive pricing procedure

should be

$$\hat{F}_t(p_t) = p_t e^{\alpha_{t-1}+p_t\beta_{t-1}} M_{t-1} + \frac{G(t)}{2}\frac{M_{t-1}e^{-1}e^{(\alpha_{t-1})}}{\beta_{t-1}^3}\mathrm{MSE}_{\beta_t}(p_t), \tag{30}$$

where α_{t-1}, β_{t-1} and $M_{t-1} = \exp[\hat{\sigma}_{t-1}^2/2]$ are the estimates for α, β and $M = \exp[\sigma^2/2]$, based on the information available at the end of period $t-1$. We wrote $\mathrm{MSE}_{\beta_t} = \mathrm{MSE}_{\beta_t}(p_t)$ to emphasize that the mean square error at the end of period t depends on the price p_t.

To implement the optimization/learning policy based on maximizing $\hat{F}_t(p_t)$ in (30), we need an expression for $\mathrm{MSE}_{\beta_t}(p_t)$, which may be very hard to obtain in explicit form. In Section 5, we implemented the one-step ahead rule in the simulations by maximizing at each period t, $t = 1, \ldots, T$, the objective function in (10), where we replace the unconditional $\mathrm{MSE}_{\beta_t}(p_t)$ by the conditional $\sigma^2_{\beta_t}(p_t)$.

In order to evaluate the approximation of the unconditional mean square error $\mathrm{MSE}_{\beta_t}(p_t)$ by the conditional variance $\sigma^2_{\beta_t}(p_t)$, assuming fixed prices, we can use the generated paths for β_t, $t = 1, \ldots, T$, in the Monte Carlo experiment. Figure 5 presents the comparison between the mean $\overline{\sigma^2}_{\beta_t}(p_t)$ of the estimates for $\sigma^2_{\beta_t}(p_t)$ and the estimate $\widehat{\mathrm{MSE}}_{\beta_t}(p_t)$, for the unconditional mean square error $\mathrm{MSE}_{\beta_t}(p_t)$, obtained from the simulations. The upper graph in Figure 5 shows the evolution of $\widehat{\mathrm{MSE}}_{\beta_t}(p_t)$ and $\overline{\sigma^2}_{\beta_t}(p_t)$ over time. Note that both decay at the same rate, although the $\overline{\sigma^2}_{\beta_t}(p_t)$ is slightly higher than $\widehat{\mathrm{MSE}}_{\beta_t}(p_t)$ for all time periods. The lower graph in Figure 5 show the scatter plot of $\widehat{\mathrm{MSE}}_{\beta_t}(p_t)$ versus $\overline{\sigma^2}_{\beta_t}(p_t)$. According to the graph, there is an approximate linear relationship between these two measures. Besides, the corresponding regression line has slope 1.0274, intercept -0.0527 and $R^2 = 0.9975$. Therefore, the approximation $\mathrm{MSE}_{\beta_t}(p_t) \doteq K^{-1}\sigma^2_{\beta_t}(p_t)$, with K very close to one, is justified empirically. These empirical results suggest that the objective function in (30) can be reasonably approximated by the objective function in (10), used the simulations.

7 Conclusions

In this paper, we have prescribed and analyzed methods for setting prices in the presence of demand function parameter uncertainty focussing especially on short planning horizons. Our contributions are both practical and technical with each

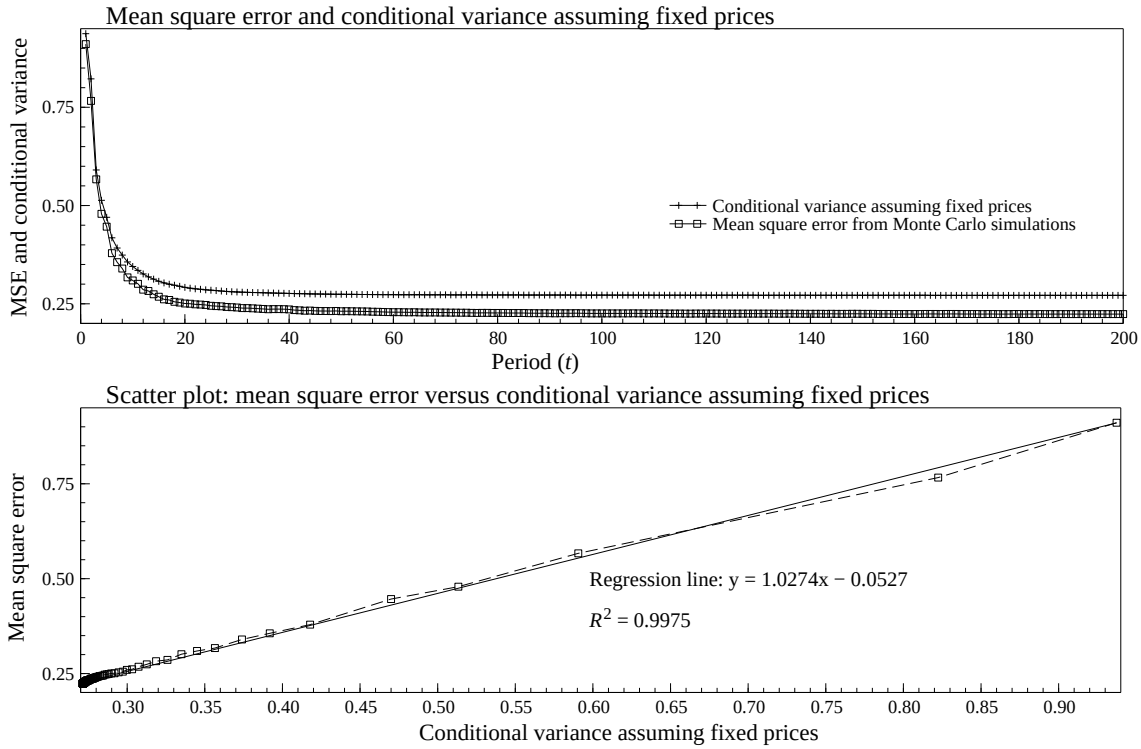

Figure 5: Comparing the mean of the variance estimates based on fixed prices to the true mean square error.

of these aspects being important.

The key practical issue addressed by this paper arises from the fact that the demand function is never known in practice!! Thus the prudent decision maker must maker an explicit tradeoff between variance reduction and revenue optimization. Theorems 1, 2 and 3 in this paper makes that tradeoff rigorous by using statistical asymptotic theory to approximate the MDP value function.

Extensive simulations produce important managerial implications and deep insights into the mathematical foundations of active learning.

The key managerial insights are:

- Myopic policies perform poorly for all planning horizons.
- Myopic policies plus occasional random price changes outperform myopic policy over the long term but not over the short term.
- Active learning (such as the one step ahead rules) are better than all other

approaches over all planning horizons.

- It is possible to constrain price changes each period and still drastically improve over myopic policies but of course, unconstrained policies produce greater revenue.

The bottom line is that managers should use active learning and if its not possible, at least be willing to experiment with some price changes to learn about the demand function. The methods in this paper suggest how to do this experimentation and would have been of use to Intrawest's management when it sought to increase revenue by varying prices.

From a technical perspective we have observed that the adaptive rules lead to biased parameter estimates. Even though the demand function is never estimated accurately, active learning still produces good revenue streams. This also suggests that one should consider biased parameter estimates when combining estimation with optimization. Bias is present in all active learning (adaptive control). This bias is due to randomness in *both* prices and noise. Under repeated simulations, we would still get biased parameter estimates unless prices were fixed and the only source of variability was the random disturbance in the demand function. We have shown why this is the case and also why one step ahead methods still produce excellent results.

The authors are investigating several extensions of this model.

- Empirically testing the methods of this paper in real or simulated markets.
- Including other explanatory variables in the demand function that might be fixed (seasonal dummy variables, time trend, day of the week) or random (competitor prices, market indicators) covariates. In particular, by regarding the constant in (1) as *market size*, we can view a time trend as a changing market size and investigate its implications on price choice throughout the planning horizon.
- For low demand items, Poisson, binomial or other generalized linear models may be more appropriate demand distribution models. A first step in this direction is pursued in Carvalho and Puterman (2005).

- Allowing model parameters to change over time following a state space model or a step-change model.
- Exploring enhanced price setting mechanisms that may yield higher revenues or have reduced biased.
- Allowing for heterogeneity in markets by using mixture models (see, for example, Jacobs, Jordan, Nowlan and Hinton, 1991, Hastie, Tibshirani and Friedman, 2001, and Carvalho and Tanner, 2005), where we increase the number of components as we observe more data. In this case, we expect that the number of components or basis functions J will be an increasing function of the number of available observations t.

Appendix

Proof of Theorem 1. By using a Taylor's series expansion for $R_T^*(p_T(\cdot))$ around the true parameter β, we have

$$\begin{aligned} R_T^*(p_T(\beta_{T-1})) &= R_T^*(p_T(\beta)) + \partial_{\beta_{T-1}} R_T^*(p_T(\beta))[\beta_{T-1} - \beta] \\ &+ \frac{1}{2}\partial^2_{\beta_{T-1}} R_T^*(p_T(\beta))[\beta_{T-1} - \beta]^2 + \frac{1}{6}\partial^3_{\beta_{T-1}} R_T^*(p_T(\beta))[\beta_{T-1} - \beta]^3 \\ &+ \frac{1}{24}\partial^4_{\beta_{T-1}} R_T^*(p_T(\bar{\beta}))[\beta_{T-1} - \beta]^4, \end{aligned}$$

where $\bar{\beta}$ is located between β_{T-1} and β, and $\partial^r_{\beta_{T-1}}$ denotes the r-th derivative with respect to β_{T-1}. Taking expectations with respect to the random variable β_{T-1}, we obtain

$$\begin{aligned} E\{R_T^*(p_T(\beta_{T-1}))\} &= R_T^*(p_T(\beta)) + \partial_{\beta_{T-1}} R_T^*(p_T(\beta)) E\{\beta_{T-1} - \beta\} \\ &+ \frac{1}{2}\partial^2_{\beta_{T-1}} R_T^*(p_T(\beta)) E\{[\beta_{T-1} - \beta]^2\} + \frac{1}{6}\partial^3_{\beta_{T-1}} R_T^*(p_T(\beta)) E\{[\beta_{T-1} - \beta]^3\} \\ &+ \frac{1}{24} E\{\partial^4_{\beta_{T-1}} R_T^*(p_T(\bar{\beta}))[\beta_{T-1} - \beta]^4\}. \end{aligned} \quad (31)$$

The second term in the right-hand-side of (31) is equal to zero, provided that β_{T-1} is unbiased for β. The third term is equal to

$$\frac{1}{2}\partial^2_{\beta_{T-1}} R_T^*(p_T(\beta)) E\{[\beta_{T-1} - \beta]^2\} = \frac{1}{2}\partial^2_{\beta_{T-1}} R_T^*(p_T(\beta)) \mathrm{Var}[\beta_{T-1}].$$

The fourth term in (31) is equal to zero because β_{T-1} is normally distributed, so the third central moment is zero. Finally, for the fifth term, employing Jensen's and Cauchy-Schwarz inequalities

$$\begin{aligned}|E\{\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))[\beta_{T-1}-\beta]^4\}| &\leq E\{|\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))|[\beta_{T-1}-\beta]^4\} \\ &\leq E\{|\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))|^2\}^{1/2} E\{[\beta_{T-1}-\beta]^8\}^{1/2}.\end{aligned}$$

We can show that $E\{|\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))|^2\} = O(1)$, so it does not diverge as the sample size $(T-1)$, used in the estimation of β_{T-1}, goes to infinity. Besides, $\mathrm{Var}[\beta_{T-1}]^{-1/2}[\beta_{T-1}-\beta] \sim \mathrm{N}(0,1)$, so that $E\{\mathrm{Var}[\beta_{T-1}]^{-4}[\beta_{T-1}-\beta]^8\} = \mu_8$, where μ_8 is the 8-th central moment of a standard normal random variable. We then have $E\{[\beta_{T-1}-\beta]^8\} = \mathrm{Var}[\beta_{T-1}]^4\mu_8$, and therefore

$$|E\{\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))[\beta_{T-1}-\beta]^4\}| \leq E\{|\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))|^2\}^{1/2}\mathrm{Var}[\beta_{T-1}]^2\mu_8^{1/2}.$$

We know that, $\mathrm{Var}[\beta_{T-1}] = \sigma^2 P_{T-1,2,2}$, with P_{T-1} has the form $(Z'Z)^{-1}$, where Z is the corresponding design matrix for the regression model in (1). If the magnitude of the rows in the design matrix Z do not change as $(T-1)$ goes to infinity, we have $\mathrm{Var}[\beta_{T-1}] = O((T-2)^{-1})$, in the sense that it goes to zero at order $(T-2)^{-1}$ when the sample size n goes to infinity. Hence,

$$|E\{\partial^4_{\beta_{T-1}} R^*_T(p_T(\bar{\beta}))[\beta_{T-1}-\beta]^4\}| = O((T-2)^{-2}),$$

and

$$\begin{aligned}&E\{R^*_T(p_T(\beta_{T-1}))\} = R^*_T(p_T(\beta)) \\ &+ \frac{1}{2}\partial^2_{\beta_{T-1}} R^*_T(p_T(\beta))E\{[\beta_{T-1}-\beta]^2\} + O((T-2)^{-2}).\end{aligned}$$

By differentiating (6) twice with respect to β_{T-1}, we have

$$\begin{aligned}\partial^2_{\beta_{T-1}} R^*_T(p_T(\beta_{T-1})) &= -2\frac{M}{\beta^3_{T-1}}\exp(\alpha-\beta/\beta_{T-1}) \\ &+ 4\frac{M\beta}{\beta^4_{T-1}}\exp(\alpha-\beta/\beta_{T-1}) - \frac{M\beta^2}{\beta^5_{T-1}}\exp(\alpha-\beta/\beta_{T-1}),\end{aligned}$$

and

$$\partial^2_{\beta_{T-1}} R^*_T(p_T(\beta_{T-1}))\Big|_{\beta_{T-1}=\beta} = \frac{Me^{(\alpha-1)}}{\beta^3}\sigma^2_{\beta_{T-1}}.$$

Therefore,

$$E[R_T^*(p_T(\beta_{T-1}))] = R_T^*(p_T(\beta)) + \frac{1}{2}\frac{Me^{(\alpha-1)}}{\beta^3}\sigma^2_{\beta_{T-1}} + O((T-2)^{-2}),$$

as we wanted to show. □

Proof of Theorem 3. Using a Taylor expansion around the point $\theta_t = \theta$, we can write the following approximation for the expected value of $V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))$[3]

$$\begin{aligned} E_{\theta_t}\left\{V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\right\} =& V_{t+1}(h_{t+1}(\theta, \Sigma_t)) \\ &+ \left[\partial_{\theta_t'} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta}\right] E\{\theta_t - \theta\} \\ &+ \frac{1}{2}E\Big\{(\theta_t - \theta)'\mathbf{A}_{t+1}(\theta_t - \theta)\Big\} + O(t^{-2}). \end{aligned} \tag{32}$$

with,

$$\mathbf{A}_{t+1} = \left[\partial_{\theta_t}\partial_{\theta_t'} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta}\right],$$

where the expectation $E_{\theta_t}[\cdot]$ is calculated with respect to the random variable θ_t. Because θ_t is not necessarily unbiased, the second term in the right-hand side of (32) is not necessarily zero. However, when $\Sigma_t = \mathbf{0}$, from the unbiasedness of the policy $h_{t+1}(\theta_t, \Sigma_t)$, according to Definition 1, and because $V_{t+1}(h_{t+1}(\cdot,\cdot))$ is twice continuously differentiable at $\theta_t = \theta$ and $\Sigma_t = \mathbf{0}$, we conclude that

$$\partial_{\theta_t'} V_{t+1}(h_{t+1}(\theta_t, \mathbf{0}))\Big|_{\theta_t=\theta} = \mathbf{0}. \tag{33}$$

On the other hand, we know that $\Sigma_t = O(t^{-1})$. Therefore, given the continuity of the second derivative of $V_{t+1}(h_{t+1}(\cdot,\cdot))$ at $\Sigma_t = \mathbf{0}$, we have

$$\partial_{\theta_t'} V_{t+1}(h_{t+1}(\theta_t, \Sigma_t))\Big|_{\theta_t=\theta} = O(t^{-1}), \tag{34}$$

and, because the bias $E\{\theta_t - \theta\} = O(t^{-1})$, we conclude that the second term in the right-hand side of (32) is $O(t^{-2})$.

For the first term in (32), a first order Taylor expansion implies

$$\begin{aligned} V_{t+1}(h_{t+1}(\theta, \Sigma_t)) = & V_{t+1}(h_{t+1}(\theta, \mathbf{0})) \\ &+ \left[\partial_{\mathrm{vec}(\Sigma_t)} V_{t+1}(h_{t+1}(\theta, \Sigma_t))\Big|_{\Sigma_t=\Sigma_t^*}\right]\mathrm{vec}(\Sigma_t), \end{aligned} \tag{35}$$

[3]For more details on approximation of moments, see for example Lehmann (1999).

where $\mathrm{vec}(\Sigma_t)$ is a vector obtained by stacking the columns of Σ_t and $\mathrm{vec}(\Sigma_t^*)$ is a convex combination of $\mathbf{0}$ and $\mathrm{vec}(\Sigma_t)$. From the continuity of $\partial_{\mathrm{vec}(\Sigma_t)} V_{t+1}(h_{t+1}(\theta, \Sigma_t))$ at $\Sigma_t = \mathbf{0}$ and from the fact that $\Sigma_t = O(t^{-1})$, we conclude that the second term in the right-hand side of (35) is also $O(t^{-2})$. Therefore, combining (32) and (35), we can write the approximation

$$\begin{aligned} E_{\theta_t}\left\{V_{t+1}(h_{t+1}(\theta_t,\Sigma_t))\right\} &= V_{t+1}(h_{t+1}(\theta, \mathbf{0})) \\ &\quad + \frac{1}{2} E\Big\{(\theta_t - \theta)' \mathbf{A}_{t+1} (\theta_t - \theta)\Big\} + O(t^{-2}). \end{aligned} \tag{36}$$

Finally, note that

$$E\Big\{(\theta_t - \theta)' \mathbf{A}_{t+1} (\theta_t - \theta)\Big\} = \mathrm{trace}\big[E\big\{(\theta_t - \theta)(\theta_t - \theta)'\big\}\mathbf{A}_{t+1}\big] = \mathrm{trace}\big[\Sigma_t \mathbf{A}_{t+1}\big],$$

concluding the proof. □

REFERENCES

C. Anderson and Z. Hong. Reinforcement Learning with Modular Neural Networks for Control. *Proceedings of NNACIP'94*, the IEEE International Workshop on Neural Networks Applied to Control and Image Processing, 1994.

Y. Aviv and A. Pazgal. Pricing of Short Life-Cycle Products through Active Learning, *Technical Report*, Olin School of Business, Washington University, 2002.

Y. Aviv and A. Pazgal. A Partially Observed Markov Decision Process for Dynamic Pricing, *Technical Report*, Olin School of Business, Washington University, 2002.

K. Azoury. Bayes Solution to Dynamic Inventory Models under Unknown Demand Distributions, *Management Science* 31, 1150-1160, 1985.

R. Balvers and T. Cosimano. Actively Learning about Demand and the Dynamics of Price Adjustment, *The Economic Journal* 100, 882-898, 1990.

M. Campi and P. Kumar. Adaptive Linear Quadratic Gaussian Control: The Cost-Biased Approach Revisited, *University of Illinois at Urbana-Champaign Technical Report, http://black.csl.uiuc.edu/prkumar*, 1998.

A. Carvalho and M. Puterman. Learning and Pricing in an Internet Environment with Binomial Demands, *Journal of Revenue and Pricing Management* 3, 320--336, 2005.

A. Carvalho and M. Tanner. Modeling nonlinear time series with local mixtures of generalized linear models. *The Canadian Journal of Statistics* 18, 97-114, 2005.

H. Chen and L. Guo. A Robust Stochastic Adaptive Controller, *IEEE Transactions on Automatic Control* 33, 1988.

E. Cope. Non-parametric Strategies for Dynamic Pricing in e-Commerce, *Technical Report*, Sauder School of Business, University of British Columbia, 2004.

T. Dietterich and X. Wang. Batch Value Function Approximation via Support Vectors, *Dietterich, T. G. Becker, S., Ghahramani, Z. (Eds.) Advances in Neural Information Processing Systems 14*, Cambridge, MA: MIT Press, 2003.

X. Ding, M. Puterman and A. Bisi. The Censored Newsvendor and the Optimal Acquisition of Information, *Operations Res*earch 50, 517-527, 2002.

N. Draper and H. Smith. *Applied Regression Analysis*, Wiley Series in Probability and Statistics, 1998.

D. Easley and N. Kiefer. Controlling a Stochastic Process with Unknown Parameters, *Econometrica* 56, 5, 1045-1069, 1988.

D. Easley and N. Kiefer. Optimal Learning with Endogenous Data, *International Economics Review* 30, 4, 963-978, 1989.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modeling Based on Generalized Linear Models (Springer Series in Statistics)*, Springer-Verlag, 1994.

J. Forbes and D. Andre. Real-Time Reinforcement Learning in Continuous Domain, *AAAI Spring Symposium on Real-Time Autonomous Systems*, 2000.

G. Gallego and G. van Ryzin. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons, *Management Science* 40, 8, 999-1020, 1994.

A. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1994.

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer, 2001.

C. Hu, W. Lovejoy and S. Shafer. Comparison of Some Suboptimal Control Policies in Medical Drug Therapy, *Operations Resesearch* 44, 696-709, 1996.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton. Adaptive mixtures of local experts, *Neural Computation* 3, 79-87, 1991.

K. Kalyanam. Pricing Decisions Under Demand Uncertainty: A Bayesian Mixture Model Approach, *Marketing Science*, 1996.

N. Kiefer and Y. Nyarko. Optimal Control of an Unknown Linear Process with Learning, *International Economics Review* 30, 3, 571- 586, 1989.

P. Kumar. Convergence of Adaptive Control Schemes Using Least-Squares Parameter Estimates, *IEEE Transactions on Automatic Control*, 1990.

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

E. Lehmann. *Elements of Large-Sample Theory*, Springer, 1999.

M. Lobo and S. Boyd. Pricing and Learning with uncertain demand, *Working Paper*, 2003.

W. Lovejoy. Myopic Policies for Some Invneotory Models with Uncertain Demand Distributions, *Management Science* 36, 724-738, 1990.

R. Martinez. Pricing in a Congestible Service Industry with a Focus on the Ski Industry, *Unpublished MSc Thesis*, Sauder School of Business, University of British Columbia, 2003.

N. Petruzzi and M. Dada. Dynamic Pricing and Inventory Control with Learning, *Naval Research Logistics* 49, 304-325, 2002.

C. Raju, Y. Narahari and K. Kumar. Learning dynamic prices in multi-seller electronic retail markets with price sensitive customers, stochastic demands, and inventory replenishments, *Indian Institute of Science Working Paper*, 2004.

M. Rothschild. A Two-Armed Bandit Theory of Market Pricing, *Journal of Economic Theory* 9, 185-202, 1974.

H. Scarf. Some Remarks on Baye's Solution to Inventory Problem, *Naval Research and Logistics* 7, 591-596, 1960.

E. Silver and M. Rahnama. Biased Selection of the Inventory Reorder Point when Demand Parameters are Statistically Estimated. *Engr. Cost and Prod. Econ.* 12, 283-292,1987.

J. Treharne and C. Sox. Adaptive Inventory Control for Non-stationary Demand and Partial Information, *Management Science* 48, 607-624, 2002.

R. Sutton and G. Barto. *Reinforcement Learning*. MIT Press, 2nd edition, 1998.

J. Sternby. On Consistency for the method of least squares using martingale theory, *IEEE Transactions on Automatic Control*, 1977.

J. Tsitsiklis. An Analysis of Temporal-Difference Learning with Function Approximation, *IEEE Transactions on Automatic Control*, 1997.

H. White. *Asymptotic Theory for Econometricians*. Academic Press, 2001.

Ipea – Institute for Applied Economic Research

PUBLISHING DEPARTMENT

Coordination
Cláudio Passos de Oliveira

Supervision
Everson da Silva Moura
Reginaldo da Silva Domingos

Typesetting
Bernar José Vieira
Cristiano Ferreira de Araújo
Daniella Silva Nogueira
Danilo Leite de Macedo Tavares
Diego André Souza Santos
Jeovah Herculano Szervinsk Junior
Leonardo Hideki Higa

Cover design
Luís Cláudio Cardoso da Silva

Graphic design
Renato Rodrigues Buenos

The manuscripts in languages other than Portuguese published herein have not been proofread.

Ipea Bookstore

SBS – Quadra 1 – Bloco J – Ed. BNDES, Térreo
70076-900 – Brasília – DF
Brazil
Tel.: + 55 (61) 3315 5336
E-mail: livraria@ipea.gov.br

Composed in Adobe Garamond 11/13.2 (text)
Frutiger 47 (headings, graphs and tables)
Brasília – DF – Brazil

Ipea's mission
Enhance public policies that are essential to Brazilian development by producing and disseminating knowledge and by advising the state in its strategic decisions.

ISSN 1415-4765

ipea Institute for Applied Economic Research

Secretariat of Strategic Affairs

BRAZILIAN GOVERNMENT BRASIL