Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Simultaneously Learning and Optimizing Using Controlled Variance Pricing

Arnoud V. den Boer, Bert Zwart

To cite this article:
Arnoud V. den Boer, Bert Zwart (2014) Simultaneously Learning and Optimizing Using Controlled Variance Pricing. Management Science 60(3):770-783. http://dx.doi.org/10.1287/mnsc.2013.1788

MANAGEMENT SCIENCE
Vol. 60, No. 3, March 2014, pp. 770–783
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1788

Simultaneously Learning and Optimizing Using Controlled Variance Pricing

Arnoud V. den Boer
Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands; and
University of Amsterdam, 1098 XH Amsterdam, The Netherlands, a.v.d.boer@tue.nl

Bert Zwart
Centrum Wiskunde and Informatica, 1098 XG Amsterdam, The Netherlands; and Department of Mathematics,
VU University Amsterdam, 1081 HV Amsterdam, The Netherlands, bert.zwart@cwi.nl

Price experimentation is an important tool for firms to find the optimal selling price of their products. It should be conducted properly, since experimenting with selling prices can be costly. A firm, therefore, needs to find a pricing policy that optimally balances between learning the optimal price and gaining revenue. In this paper, we propose such a pricing policy, called controlled variance pricing (CVP). The key idea of the policy is to enhance the certainty equivalent pricing policy with a taboo interval around the average of previously chosen prices. The width of the taboo interval shrinks at an appropriate rate as the amount of data gathered gets large; this guarantees sufficient price dispersion. For a large class of demand models, we show that this procedure is strongly consistent, which means that eventually the value of the optimal price will be learned, and derive upper bounds on the regret, which is the expected amount of money lost due to not using the optimal price. Numerical tests indicate that CVP performs well on different demand models and time scales.

Keywords: dynamic pricing; sequential decision problems; statistical learning
History: Received March 16, 2010; accepted June 4, 2013, by Assaf Zeevi, stochastic models and simulation.
Published online in *Articles in Advance* December 10, 2013.

1. Introduction and Literature

1.1. Introduction

A firm that sells products or delivers services faces the problem of determining the selling price that generates the highest revenue. This price is generally unknown to the seller, but by experimenting with the selling price, its value might be learned. Price experimentation, however, may be costly, since it means that suboptimal prices are chosen, and thus less revenue is generated. The seller should, therefore, find a balance between price experimentation and revenue maximization. The problem of finding this optimal balance is the subject of this paper.

The demand for the product during a certain time period is modeled as a random variable. This random variable depends on the chosen price and on certain unknown parameters. Common examples include a Normal, Poisson, or Bernoulli distributed demand, with expectation a linear or logit function of the price. To learn the price that generates the highest revenue, the seller needs to obtain good estimates for these unknown parameters. This is done by experimenting with different prices in different selling periods. In this way, the seller accumulates information about the demand function and can then determine the optimal selling price.

At the beginning of each selling period—which may be a day or a week, but also minutes or hours—the seller decides on a price. On one hand, he wishes to set the price close to what is optimal according to his current knowledge of the demand function. On the other hand, the prices should vary, to learn about the relation between demand and price. This collection of price decisions for each selling period is called a pricing policy. The goal of the seller is to find a pricing policy that exhibits sufficient price experimentation to learn the value of the unknown demand parameters, but that also is not too costly.

An intuitively appealing pricing policy is to set the new price at each period equal to the price that would be optimal if the current parameter estimates were correct. Such a policy is usually called passive learning, myopic pricing, or certainty equivalent pricing, for obvious reasons: each decision is made as if the current parameter estimates are equal to the true values. In this paper, we show that this policy, although intuitively appealing, is not suitable: the seller may never learn the value of the optimal price. The reason is that certainty equivalent pricing puts too much emphasis on instant revenue maximization, and not enough emphasis on collecting information.

To solve this issue, we propose a new dynamic pricing policy called controlled variance pricing (CVP).

The idea is to select the certainty equivalent price, unless it lies in a certain, slowly shrinking, taboo interval around the average prices chosen so far. This taboo interval guarantees sufficient price dispersion, and leads to a good balance between experimentation and optimization.

We formulate our pricing policy for a broad class of demand functions, which includes many demand models that are used in practice. Moreover, we do not require that the seller has complete knowledge about the structural form of the demand distribution, but only about the relation between the first two moments and the selling price. This makes the model a little more robust compared with models where a complete demand distribution is assumed. The unknown parameters are estimated by maximum quasi-likelihood estimation (MQLE); this is a generalization of maximum likelihood estimation (MLE) to distributions where only the first two moments are known. For many demand models that are used in practice, MQLE and MLE are similar; this enables us to compare our results with other pricing policies in the literature that use MLE.

We show analytically that CVP will eventually provide the correct value of the unknown parameters, and thus the value of the optimal price. We also provide bounds on the speed of convergence. Furthermore, we obtain an asymptotic upper bound on the regret, which measures the expected amount of money lost due to not using the optimal price. In particular, we show that the regret after T time periods is $O(T^{1/2+\delta})$, where $\delta > 0$ is arbitrarily small. This bound is close to $O(\sqrt{T})$, which in several settings has been shown to be the best achievable asymptotic upper bound of the regret (see, e.g., Besbes and Zeevi 2011, Broder and Rusmevichientong 2012, Kleinberg and Leighton 2003). Apart from this theoretical result, we also numerically compare the performance of CVP with other existing pricing policies from the literature. These numerical experiments suggest that CVP performs well for different demand functions and time scales.

1.2. Literature Review

The problem we are considering is a dynamic pricing problem with unknown demand. There is a wide range of literature on this topic. The existing operations research literature can be categorized into "classical" parametric, Bayesian parametric, and nonparametric approaches. We also mention some relevant economics literature on this subject, and connections to literature on multiarmed bandit problems and stochastic approximation.

1.2.1. Parametric Approaches with Non-Bayesian Estimation Methods. In these studies it is assumed that the unknown demand function is in a parameterized family of distributions; the unknown parameters are estimated by classical statistical methods such as maximum likelihood or least squares estimation. Examples include Lobo and Boyd (2003), Carvalho and Puterman (2005a, b), Bertsimas and Perakis (2006), Besbes and Zeevi (2009), and Broder and Rusmevichientong (2012).

Lobo and Boyd (2003) maximize the expected discounted revenue of a linear demand model over a short time horizon. They formulate the pricing problem as a dynamic program and propose a tractable convex approximation. No analytical results are given on the performance of this pricing policy, but it is numerically compared with two other policies: certainty equivalent pricing and certainty equivalent pricing with added random perturbations. Simulation results on a short time horizon (10 periods) suggest that the convex approximation performs better than the other two policies.

The papers by Carvalho and Puterman (2005a, b) propose one-step-ahead pricing. Using a Taylor expansion of the expected revenue for the next period, the price is chosen that approximately maximizes the sum of the revenues in the next two periods. This is in contrast to certainty equivalent pricing, where only the expected revenue of one period is maximized. In Carvalho and Puterman (2005b), this idea is applied to a binomial demand function with expectation a logit function of the price, whereas in Carvalho and Puterman (2005a), a log-normal demand model is considered. Neither paper provides analytical results on the performance of the policies, but both show numerical simulations suggesting that the one-step-ahead policy performs well.

Bertsimas and Perakis (2006) consider a dynamic pricing problem with unknown, linear demand and finite inventory. They formulate pricing policies based on dynamic programming heuristics, and also apply their techniques to a setting with competition.

A Bernoulli demand distribution is assumed by Broder and Rusmevichientong (2012). They propose a policy called the MLE-cycle, and show that it achieves the asymptotically optimal bound $O(\sqrt{T})$ on the regret. In this pricing policy, the time horizon is divided into learning phases, during which certain a priori determined prices are chosen, and optimization phases, during which the certainty equivalent price is chosen. A policy somewhat similar in structure was studied in Besbes and Zeevi (2009), where performance bounds are derived in a certain asymptotic regime.

1.2.2. Parametric Approaches with Bayesian Learning. In Bayesian approaches to dynamic pricing with unknown demand, the unknown parameters are learned via Bayesian updates of a certain belief distribution. Such an approach is studied by Lin (2006), Araman and Caldentey (2009), Farias and

van Roy (2010), and Harrison et al. (2012). All these papers assume a single unknown parameter and propose several pricing heuristics.

A general framework for stochastic control of linear regression models is investigated by Easley and Kiefer (1988), Kiefer and Nyarko (1989), and Aghion et al. (1991). They study the asymptotic behavior of different Bayesian learning policies, and show that in some settings, the beliefs converge to a limit that in expectation may differ from the correct value.

1.2.3. Nonparametric Approaches. Robust or nonparametric approaches do not assume a known parametric functional form of the demand, and generally investigate how to maximize revenue in worst-case scenarios. Such approaches can be found in Kleinberg and Leighton (2003), Cope (2007), Lim and Shanthikumar (2007), Eren and Maglaras (2010), and Besbes and Zeevi (2009).

1.2.4. Economics Literature. Related research from the economics literature is found in Taylor (1974), Anderson and Taylor (1976), McLennan (1984), Balvers and Cosimano (1990), and Keller and Rady (1999). Taylor (1974) shows consistency of certainty equivalent pricing in a linear demand model with one unknown parameter. Anderson and Taylor (1976) consider a linear model with two unknown parameters; they perform simulations to investigate if a certainty equivalent policy is inconsistent when the objective is to steer the demand to a certain fixed, a priori chosen value. A difference with our setting is that we try to steer the expected demand to a level (namely, the expected demand level that occurs at the optimal price) that is not known a priori: its value is learned over time. Lai and Robbins (1982) prove that in this model, the certainty equivalent policy is not strongly consistent. They also propose a strongly consistent modification of the certainty equivalent policy. McLennan (1984) studies a dynamic pricing problem with unknown demand and assumes that there are only two possible values that the unknown parameters can take. Balvers and Cosimano (1990) and Keller and Rady (1999) study optimal learning in the case where the unknown parameters of a linear demand function are not static, but change over time.

1.2.5. Multiarmed Bandit Problem and Stochastic Approximation. Dynamic pricing with unknown demand has similarities to the classical multiarmed bandit problem: both are sequential decision problems under uncertainty, for which one wants to determine a decision policy that optimally balances exploration and exploitation. Some differences between these problems are that we have an uncountable action space, whereas bandit problems often assume a finite action space, and we assume a known structural relation between reward and action. Lai and Robbins (1985), Gittins (1989), Vermorel and Mohri (2005), Cesa-Bianchi and Lugosi (2006), and Powell (2010) provide an introduction to these types of problems, and discuss several decision policies. Typically, these policies have to balance between exploration and exploitation; a similar trade-off is encountered in this study.

A study from the multiarmed bandit literature that is interesting to mention is Goldenshluger and Zeevi (2009). They study a particular type of a single-armed bandit problem. In one of the policies they propose, called nearly myopic policy, the action taken at each time step is equal to the myopic action minus a factor that converges to zero. This is similar to the taboo interval approach taken in this study (see §3).

There is also a connection with the classical stochastic approximation literature (Robbins and Monro 1951, Kiefer and Wolfowitz 1952), in which one attempts to determine the unknown maximizer of a function. A difference is that in these settings, the performance is measured by the quality of the estimate of the maximizer, and not by the cumulative costs that leads to the estimate.

1.2.6. Consistency of Linear Regression and Maximum Quasi-Likelihood Estimation. Our results on the strong consistency of CVP are based on den Boer and Zwart (2012), who establish sufficient conditions for consistency of MQLE, and provide bounds on mean square convergence rates. Their results are partly based on Lai and Wei (1982), who provide sufficient conditions for strong consistency of recursive least squares estimation in a linear model with adapted design. In particular, writing $\lambda_{\min}(t)$, $\lambda_{\max}(t)$ for the smallest and largest eigenvalues of the design matrix, they showed that $\lambda_{\min}(t) \to \infty$ almost surely (a.s.) and $(\log \lambda_{\max}(t))/\lambda_{\min}(t) \to 0$ a.s. are sufficient conditions for strong consistency of the least squares estimates (under some additional assumptions on the noise terms). This result is the counterpart of Lai et al. (1979), where it was shown that $\lambda_{\min}(t) \to \infty$ a.s. is both necessary and sufficient for strong consistency in the case of a fixed design. Results by Nassiri-Toussi and Ren (1994) indicate that in the adapted case the extra requirement $(\log \lambda_{\max}(t))/\lambda_{\min}(t) \to 0$ a.s. cannot simply be removed.

Our analysis of CVP is valid for a large class of demand models. This is in contrast to Lobo and Boyd (2003), Carvalho and Puterman (2005a, b), and Bertsimas and Perakis (2006), where the analysis is restricted to specific models (e.g., linear) or distributions (e.g., log-normal). In addition, CVP is the first parametric approach to dynamic pricing with unknown demand where only knowledge on the first two moments of the demand distribution is required.

The expected demand can depend on two unknown parameters; this is more natural than a single unknown parameter, which is assumed in the Bayesian approaches to dynamic pricing mentioned above.

Another feature of CVP is that it balances learning and optimization at each decision moment, enabling convergence of the prices to the optimal price. This differs from policies that strictly separate the time horizon in exploration and exploitation phases, as in Broder and Rusmevichientong (2012) or Besbes and Zeevi (2009); here only the average price converges to the optimal price. Moreover, in these latter type of policies, the number of exploration prices that need to be chosen beforehand increases when the number of unknown parameters increases. CVP only requires one variable to choose, independent of the number of unknown parameters (namely, the constant c in, e.g., Proposition 2, for fixed α). This makes the method potentially suitable for extensions to models with multiple products. (Such an extension can be constructed by requiring a lower bound on the smallest eigenvalue of the design matrix P_n, defined in Lemma 1 (see §3.2), instead of on the sample variance of the prices $\mathrm{Var}(p)_t$.) Another difference between CVP and these policies is that CVP uses all available historical data to form parameter estimates, whereas the analysis of the algorithms by Broder and Rusmevichientong (2012) and Besbes and Zeevi (2009) uses only data from the exploration phases.

CVP is intuitively easy to understand by price managers and easy to implement in decision support systems. Numerical experiments suggest that CVP performs well on different time scales and demand models.

The rest of this paper is organized as follows. The demand model and some notation are described in §2, followed by a short discussion on the model assumptions (§2.1) and the method to estimate the unknown parameters (§2.2). In §3, we show that certainty equivalent pricing is not consistent, which motivates the introduction of CVP. We show that under this policy the parameter estimates converge to the true value, and we show that the regret admits the upper bound $O(T^{1/2+\delta})$, where T is the number of time periods and $\delta > 0$ is arbitrarily small. While preparing the final version of this paper, we learned that Keskin and Zeevi (2013) show that CVP is part of a larger family of algorithms that achieve the same regret. In §4, CVP is numerically compared with another pricing policy from the literature on different time scales and different demand functions. Conclusions and directions for future research are provided in §5. The appendix contains the proofs of the propositions in this paper.

2. Model and Notation

We consider a monopolist firm that sells a single product. Time is discretized, and time periods are denoted by $t \in \mathbb{N}$. At the beginning of each time period, the firm determines a selling price $p_t \in [p_l, p_h]$. The prices $0 < p_l < p_h$ are the minimum and maximum prices that are acceptable to the firm. After setting the price, the firm observes a realization d_t of the demand $D_t(p_t)$, which is a random variable, and collects revenue $p_t \cdot d_t$. We assume that the inventory is sufficient to meet all demand; i.e., stockouts do not occur.

The random variable $D_t(p_t)$ denotes the demand in period t against selling price p_t. Given the selling prices, the demand in different time periods is independent, and for each $t \in \mathbb{N}$ and $p_t = p \in [p_l, p_h]$, $D_t(p_t)$ is distributed as $D(p)$, for which we assume the following parametric model:

$$E[D(p)] = h(a_0^{(0)} + a_1^{(0)} p),$$
$$\mathrm{Var}[D(p)] = \sigma^2 v(E[D(p)]). \tag{1}$$

Here, $h\colon \mathbb{R}_+ \to \mathbb{R}_+$ and $v\colon \mathbb{R}_+ \to \mathbb{R}_{++}$ are both twice continuously differentiable known functions, with $\dot{h}(x) := \partial h(x)/\partial x > 0$ for all $x \geq 0$. Furthermore, σ and $a^{(0)} = (a_0^{(0)}, a_1^{(0)})$ are unknown parameters with $\sigma > 0$, $a_0^{(0)} > 0$, $a_1^{(0)} < 0$, and $a_0^{(0)} + a_1^{(0)} p_h \geq 0$.

Write $e_t = D(p_t) - E[D(p_t) \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}]$. We make the technical assumption on the demand that for some $r > 3$,

$$\sup_{t \in \mathbb{N}} E[|e_t|^r \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}] < \infty \quad \text{a.s.} \tag{2}$$

The expected revenue collected in a single time period where price p is used is denoted by $r(p) = p \cdot h(a_0^{(0)} + a_1^{(0)} p)$; to emphasize the dependence on the parameter values, we write $r(p, a_0, a_1) = p \cdot h(a_0 + a_1 p)$ as a function of p and (a_0, a_1).

We assume that there is an open neighborhood $U \subset \mathbb{R}^2$ of $(a_0^{(0)}, a_1^{(0)})$ such that for all $(a_0, a_1) \in U$, $r(p, a_0, a_1)$ has a unique maximizer

$$p(a_0, a_1) = \arg\max_{p_l < p < p_h} \{p \cdot h(a_0 + a_1 p)\},$$

and such that $r''(p(a_0, a_1), a_0, a_1) < 0$. This ensures that the optimal price $p_{\mathrm{opt}} = p(a_0^{(0)}, a_1^{(0)})$ is unique and well defined, and lies strictly between p_l and p_h.

The marginal costs of the sold product equal zero; therefore, maximizing profit is equivalent to maximizing revenue. Note that a situation with positive marginal costs $c > 0$ can easily be captured by replacing p by $p - c$.

A pricing policy ψ is a method that for each t generates a price $p_t \in [p_l, p_h]$, based on the previously chosen prices $p_1, \ldots, p_{t-1}$ and demand realizations $d_1, d_2, \ldots, d_{t-1}$. This p_t may be a random variable.

The performance of a pricing policy is measured in terms of regret, which is the expected revenue loss caused by not using the optimal price p_{opt}. For a pricing policy ψ that generates prices $p_1, p_2, \ldots, p_T$, the regret after T time periods is defined as

$$\text{Regret}(T, \psi) = E\left[\sum_{t=1}^{T} r(p_{\text{opt}}, a^{(0)}) - r(p_t, a^{(0)})\right].$$

The objective of the seller is to find a pricing policy ψ that maximizes the total expected revenue over a finite number of T time periods. This is equivalent to minimizing Regret(T, ψ). Note, however, that the regret cannot directly be used by the seller to find an optimal policy, since its value depends on the unknown parameters $a^{(0)}$.

Notation. With $\log(t)$ we denote the natural logarithm. If $x_1, x_2, \ldots, x_t$ is a sequence, then $\bar{x}_t := (1/t)\sum_{i=1}^{t} x_i$ denotes the sample mean and $\text{Var}(x)_t = (1/t)\sum_{i=1}^{t}(x_i - \bar{x}_t)^2$ the sample variance. For a vector $x \in \mathbb{R}^n$, x^T denotes the transpose and $||x||$ denotes the Euclidean norm of x. For nonrandom sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$, $x_n = O(y_n)$ means that there exists a $K > 0$ such that $|x_n| \le K|y_n|$ for all $n \in \mathbb{N}$.

2.1. Discussion of Model Assumptions

We do not assume complete knowledge about the demand distribution, only about the first two moments. This makes the demand model a little more robust to misspecifications.

In Equation (1) we assume that the variance of demand is a function of the expectation. This holds for many common demand models, like Bernoulli (with $v(x) = x(1-x)$, $\sigma = 1$), Poisson ($v(x) = x$, $\sigma = 1$), and Normal ($v(x) = 1$ and arbitrary $\sigma > 0$) distributions. All these examples also satisfy the moment condition (2).

The assumptions on the functions h and v and parameters σ, $a_0^{(0)}$, and $a_1^{(0)}$ imply that the expected demand is strictly decreasing in the price, and that the variance is strictly positive; i.e., demand is nondeterministic. These are both natural assumptions on the demand distribution.

The assumption on the existence and uniqueness of $\arg\max_{p_l<p<p_h}\{p \cdot h(a_0 + a_1 p)\}$ for all a_0, a_1 in an open neighborhood U of $a^{(0)}$ is satisfied by many functions h that are used in practice to model the relation between price and expected demand; examples are $h(x) = x$, $h(x) = \exp(x)$, and $h(x) = (1 + \exp(-x))^{-1}$. A sufficient condition to satisfy this assumption is that the revenue function $r(p, a^{(0)})$ is strictly concave in p and attains its maximum strictly between p_l and p_h.

2.2. Estimating the Unknown Parameters

The unknown parameters $a^{(0)}$ can be estimated with MQLE. This is a natural extension of MLE to settings where only the first two moments of the distribution are known; see Wedderburn (1974), McCullagh (1983), and Godambe and Heyde (1987), and the books by McCullagh and Nelder (1983), Heyde (1997), and Gill (2001).

Given prices $p_1, \ldots, p_t$ and demand realizations $d_1, \ldots, d_t$, the MQLE of $(a_0^{(0)}, a_1^{(0)})$, denoted by $\hat{a}_t = (\hat{a}_{0t}, \hat{a}_{1t})$, is the solution to the two-dimensional equation

$$l_t(\hat{a}_t) = \sum_{i=1}^{t} \frac{\dot{h}(\hat{a}_{0t} + \hat{a}_{1t}p_i)}{\sigma^2 v(h(\hat{a}_{0t} + \hat{a}_{1t}p_i))}\begin{pmatrix}1\\ p_i\end{pmatrix} \cdot (d_i - h(\hat{a}_{0t} + \hat{a}_{1t}p_i)) = 0. \tag{3}$$

If the probability density function (or probability mass function in the case of discrete demand distribution) of $D(p)$ can be written in the form $\exp(\sigma^{-1}(d\theta - g(\theta)))$, where θ is some function of $h(a_0 + a_1 p)$, then (3) corresponds to the maximum-likelihood equations (Wedderburn 1974). Many demand distributions that are used in practice, such as Poisson, Bernoulli, and Normal distributions, fall into this class (see, e.g., McCullagh and Nelder 1983, Gill 2001). In case of Normal distributed demand with h the identity function, (3) is also equivalent to the Normal equations of ordinary least squares, namely,

$$l_t(\hat{a}_t) = \sum_{i=1}^{t}\begin{pmatrix}1\\ p_i\end{pmatrix}(d_i - \hat{a}_{0t} - \hat{a}_{1t}p_i) = 0. \tag{4}$$

The solution to the quasi-likelihood equations may, in general, not always be unique. A standard way to select the "right" solution is to pick the solution with the lowest mean-square error (see Heyde 1997, §13.3). In our numerical results (see §4), we did not encounter problems with multiple solutions of (3).

3. Controlled Variance Pricing

3.1. Inconsistency of Certainty Equivalent Pricing

An intuitively natural pricing policy is to estimate after each time period the unknown parameters, and to set the next price equal to the price that is optimal with respect to these estimates. More precisely, choose two different initial prices $p_1, p_2 \in [p_l, p_h]$; after $t \ge 2$ time periods, calculate the MQLE estimators $\hat{a}_t$ with (3), and set the next price p_{t+1} equal to

$$p_{t+1} = \arg\max_{p\in[p_l, p_h]} r(p, \hat{a}_{0t}, \hat{a}_{1t}). \tag{5}$$

This pricing policy is known under the name certainty equivalent pricing, myopic pricing, or passive learning.

Under different settings, certainty equivalent policies are known to produce suboptimal outcomes; see, e.g., the simulation results of such policies in Lobo

and Boyd (2003) and Carvalho and Puterman (2005a, b). Anderson and Taylor (1976) studied a linear system $y_t = a_0^{(0)} + a_1^{(0)} x_t + \epsilon_t$ with unknown parameters $a_0^{(0)}$ and $a_1^{(0)}$ and input variables x_t; the objective is to steer y_t to a desired value y_*. The certainty equivalent policy is to set $x_{t+1} = \min\{x_{\max}, \max\{x_{\min}, (y_* - \hat{a}_{0t})/\hat{a}_{1t}\}\}$, where $\hat{a}_{0t}$ and $\hat{a}_{1t}$ are the least squares estimates of $a_0^{(0)}$ and $a_1^{(0)}$, based on observations $(x_1, y_1), \ldots, (x_t, y_t)$, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum admissible values for x_t. Lai and Robbins (1982) showed that there are parameter values such that using this certainty equivalent policy, the controls x_t converge with positive probability to a value different from the optimal control $x = (y_* - a_0^{(0)})/a_1^{(0)}$; this implies that the certainty equivalent policy is not strongly consistent. (Interestingly, in a Bayesian setting the certainty equivalence policy is strongly consistent, as shown by Chen and Hu 1998.) The proof idea of Lai and Robbins (1982) can easily be extended to our case, when h is the identity and v is constant. In that case, the expected demand is a linear function of the price, and the MQLE equations (3) are equivalent to the Normal equations for ordinary linear regression. A difference with Lai and Robbins (1982) is that they posed specific conditions on p_1, p_2, p_l, and p_h; in our result, these assumptions are left out.

Proposition 1. *Suppose that demand is normally distributed with constant variance and expected demand a linear function of the price* (*i.e.*, $h(x) = x$, $v(x) = 1$), *and suppose that certainty equivalent pricing is used. Then, with positive probability,* p_t *does not converge to* p_{opt}.

The details of the proof of this proposition can be found in the appendix. The idea of the proof is to show by induction that with positive probability, $p_t = p_h$ for all $t \geq 3$. Since it was assumed that $p_{\text{opt}} < p_h$, the proposition then follows.

Proposition 1 shows that certainty equivalent pricing is not strongly consistent for a linear demand function with constant variance. Its scope, however, is somewhat limited in the sense that it does only partially describe the asymptotic behavior of the policy. It is proven that with a positive, but possibly very small probability, the prices converge to $p_h \neq p_{\text{opt}}$. If this would happen in practice, the price manager would simply increase p_h. Moreover, simulations suggest that p_t may also converge to a value strictly between p_l and p_h, and that the limit price is with probability one different from p_{opt}.

3.2. Controlled Variance Pricing

An intuition for what goes wrong with the certainty equivalent policy is that the prices p_t converge "too quickly" to a certain value. As a result, not enough new information is obtained to further improve the parameter estimates, and thus they will not converge to the correct values. The key idea is to control the speed at which the prices converge. This is done by constructing a lower bound on the sample variance of the chosen prices. In particular, we require that at each time period t, $\text{Var}(p)_t \geq ct^{\alpha-1}$, for some $c > 0$ and $\alpha \in (0, 1)$.

The pricing policy we propose, CVP, chooses at each time period the certainty equivalent price (5), unless this means that the lower bound on the sample variance of the prices $\text{Var}(p)_{t+1} \geq c(t+1)^{\alpha-1}$ is not satisfied. In that case, the next price should be chosen not too close to the average price chosen so far; in particular, p_{t+1} is then not allowed to lie in the interval

$$\text{TI}(t) = \left(\bar{p}_t - \sqrt{c[(t+1)^\alpha - t^\alpha]\frac{t+1}{t}},\ \bar{p}_t + \sqrt{c[(t+1)^\alpha - t^\alpha]\frac{t+1}{t}}\right), \tag{6}$$

which is referred to as the *taboo interval* (TI) at time t. Choosing p_{t+1} outside the taboo interval creates extra price dispersion by guaranteeing $\text{Var}(p)_{t+1} \geq c \cdot (t+1)^{\alpha-1}$ (see Proposition 2).

Pricing Policy: Controlled Variance Pricing

Initialization. Choose initial prices $p_1, p_2 \in [p_l, p_h]$, $p_1 \neq p_2$. Choose $\alpha \in (0, 1)$ and
$c \in (0, 2^{-\alpha}(p_1 - p_2)^2 \min\{1, (3\alpha)^{-1}\})$.

For all $t \geq 2$:

Step 1: Estimation. Calculate the MQLE estimates $\hat{a}_t$ according to (3).

Step 2: Pricing. If
(a) there is no solution $\hat{a}_t$, or
(b) $\hat{a}_{0t} \leq 0$ or $\hat{a}_{1t} \geq 0$, or
(c) $\hat{a}_{0t} + \hat{a}_{1t} p < 0$ for some $p \in [p_l, p_h]$,
set $p_{t+1} \in \{p_1, p_2\}$ such that
$|p_{t+1} - \bar{p}_t| = \max(|p_1 - \bar{p}_t|, |p_2 - \bar{p}_t|)$.

Now assume $\hat{a}_t$ exists and $\hat{a}_{0t} > 0$, $\hat{a}_{1t} < 0$, $\hat{a}_{0t} + \hat{a}_{1t} p \geq 0$ $(p \in [p_l, p_h])$.

Set

$$p_{t+1} = \arg\max_{p \in [p_l, p_h]} r(p, \hat{a}_t) \tag{7}$$

if this results in $\text{Var}(p)_{t+1} \geq c(t+1)^{\alpha-1}$.

Else, set

$$p_{t+1} = \arg\max_{p \in [p_l, p_h] \setminus \text{TI}(t)} r(p, \hat{a}_t), \tag{8}$$

where $\text{TI}(t)$ is the taboo interval (6) at time t.

In cases (a)–(c), we choose one of the initial prices p_1, p_2, that is most far away from $\bar{p}_t$. This ensures that the bound on the variance $\text{Var}(p)_t \geq ct^{\alpha-1}$ remains valid. The upper bound on the constant c ensures that $[p_l, p_h] \setminus \text{TI}(t)$ is nonempty for $t \geq 2$, and that $\text{Var}(p)_t \geq ct^{\alpha-1}$ is satisfied for $t = 2$.

A basic requirement for any pricing policy is that the price p_t should converge to the optimal price p_{opt}, and thus the sample variance $\text{Var}(p)_t$ should converge

to zero. The speed at which the sample variance goes to zero turns out to be strongly related to the quality of the parameter estimates: in particular, the parameter estimates $\hat{a}_t$ converge quickly to the correct values $a^{(0)}$ if the sample variance $\mathrm{Var}(p)_t$ converges slowly to zero; then the price, however, converges slowly, which may be costly. We here observe a trade-off between exploration (quick convergence of parameter estimates to the correct values) and exploitation (quick convergence of prices to the optimal price). The balance between exploration and exploitation is captured in the parameter α of CVP. The following proposition establishes a relation between α and the sample variance of the prices:

Proposition 2. *With CVP,* $\mathrm{Var}(p)_t \geq ct^{\alpha-1}$ *for all* $t \geq 2$.

This assertion follows directly from the construction of CVP. The details are given in the appendix.

The results from the literature on consistency and convergence rates of parameter estimates that we use are stated in terms of the eigenvalues of the design matrix. The following lemma relates these eigenvalues to the sample variance of the prices. Its proof is straightforward and contained in the appendix.

Lemma 1. *Let* $\lambda_{\max}(t)$ *and* $\lambda_{\min}(t)$ *be the largest and smallest eigenvalues of the design matrix*

$$P_t = \begin{pmatrix} t & \sum_{i=1}^t p_i \\ \sum_{i=1}^t p_i & \sum_{i=1}^t p_i^2 \end{pmatrix}, \quad (t \geq 2),$$

where $p_1, \ldots, p_t \in [p_l, p_h]$ *and* $p_1 \neq p_2$. *Then* $\lambda_{\max}(t) \leq (1+p_h^2)t$ *and* $t\mathrm{Var}(p)_t \leq (1+p_h^2)\lambda_{\min}(t)$.

In the following proposition and theorem, we assume that CVP with $\alpha \in (1/2, 1)$ is used. We show that a solution $\hat{a}_t$ to the estimation equations (3) eventually exists, and the parameter estimates $\hat{a}_t$ converge to the correct value $a^{(0)}$. In addition, we provide an upper bound on the mean square convergence rate in terms of the parameter α.

Proposition 3. *Let* $\alpha > 1/2$. *A solution* $\hat{a}_t$ *to* (3) *eventually exists, and* $\hat{a}_t \to a^{(0)}$ *a.s. In addition, if we define*

$$T_\rho = \sup\{t \in \mathbb{N} \mid \textit{there is no solution } \hat{a}_t \textit{ of } (3) \textit{ such that } ||\hat{a}_t - a^{(0)}|| \leq \rho\}, \quad (9)$$

then there exists a $\rho_0 > 0$ *such that* $E[T_{\rho_0}^{1/2}] < \infty$ *and*

$$E[||\hat{a}_t - a^{(0)}||^2 \mathbf{1}_{t>T_{\rho_0}}] = O\left(\frac{\log t}{t^\alpha}\right), \quad (10)$$

where $\mathbf{1}_{t>T_{\rho_0}}$ *denotes the indicator function of the event* $t > T_{\rho_0}$.

This proposition can be derived from den Boer and Zwart (2012), where strong consistency and convergence rates for quasi-likelihood estimates are discussed. Theorem 1 of den Boer and Zwart (2012) implies the assertion $E[T_{\rho_0}^{1/2}] < \infty$. (Note that in their Theorem 1, the required condition $1/2 < r\alpha - 1$ is valid for all $1/2 < \alpha \leq 1$, because of our moment condition (2), with $r > 3$.) Its proof is based on the Leray–Schauder theorem, and on results concerning the moments of last-time random variables of the form $\sup\{n \in \mathbb{N} \mid |S_n| \geq cn^\alpha\}$, where S_n is a martingale, $c > 0$, and $1/2 < \alpha \leq 1$; see Lemma 4 and Proposition 1 of den Boer and Zwart (2012). The convergence rates (10) follow from Theorem 2 and Remark 2 of den Boer and Zwart (2012), together with our Lemma 1. The proof of their Theorem 2 relies on bounds for the solutions of certain quadratic equations (their Lemma 6), and on an extension of the a.s. bounds of Lai and Wei (1982) to bounds in expectation (their Proposition 2).

Quite some machinery is thus needed to prove Proposition 3 in its fullest generality, which is contained in den Boer and Zwart (2012). However, in the relevant special case of normally distributed demand with a linear demand function, the assertions of Proposition 3 can be shown with a considerably easier proof. The reason is that in this case, the MQLE estimates $\hat{a}_t$ are equal to the least-squares linear regression estimates, and strong consistency follows by adapting results from Lai and Wei (1982) to derive

$$E[||\hat{a}_t - a^{(0)}||^2] = O\left(\frac{\log t}{t^\alpha}\right). \quad (11)$$

A proof of (11) is shown in the appendix.

Proposition 3 enables us to calculate the following upper bound on the Regret:

Theorem 1.

$$\mathrm{Regret}(T, \mathrm{CVP}) = O(T^\alpha + T^{1-\alpha}\log T),$$

provided $\alpha > 1/2$.

We use a Taylor-series expansion of the revenue function $r(p)$ to show that $|r(p) - r(p_{\mathrm{opt}})| = O((p - p_{\mathrm{opt}})^2)$. The implicit function theorem is invoked to obtain $|p(a) - p_{\mathrm{opt}}| = O(||a - a^{(0)}||)$. The theorem can then be derived from Proposition 3, and the rate at which the size of the taboo interval converges to zero. The details of the proof are given in the appendix.

3.3. Discussion

3.3.1. Discussion of Bound on Regret. The term $T^{1-\alpha}\log T$ in Theorem 1 comes from bounds (10) on the quality of the parameter estimates, and the term T^α comes from the length of the taboo interval (6). The parameter α captures the trade-off between learning

and optimization. If α is large, then much emphasis is put on learning: the parameters converge quickly to their correct values, but due to the large size of the taboo interval, the prices converge slowly. If α is small, then the emphasis is on optimizing instant revenue: the taboo interval is then very small, thus the next-period price is close to the certainty equivalent optimal price; however, for small α, only a relatively slow convergence of the parameter estimates is guaranteed by Proposition 3. The optimal choice of α in Theorem 1 clearly is 1/2, but since α should be larger than 1/2, we get the following:

Corollary 1.

$$\text{Regret}(T, \text{CVP}) = O(T^{1/2+\delta}),$$

with $\alpha = 1/2 + \delta$, *for arbitrarily small* $\delta > 0$.

This result would be a little more elegant if the term T^{δ}, $\delta > 0$, could be removed. The results of den Boer and Zwart (2012), however, require $\alpha > 1/2$, and it appears that, in general, this requirement cannot easily be removed. In some special cases, e.g., $h(x) = x$, $v(x) = 1$, or $h(x) = (1 + \exp(-x))^{-1}$, $v(x) = x(1-x)$, Remark 5 of den Boer and Zwart (2012) implies that Proposition 3 also holds for $\alpha = 1/2$. In these cases, CVP has $\text{Regret}(T, \text{CVP}) = O(\sqrt{T}\log T)$, which slightly improves Corollary 1.

The bound from Corollary 1 is nevertheless quite close to $O(\sqrt{T})$. In several settings it has been shown that this is the best achievable asymptotic upper bound on the Regret (see, e.g., Besbes and Zeevi 2011, Broder and Rusmevichientong 2012, Kleinberg and Leighton 2003).

3.3.2. Generality of Result. Concerning the generality of the result, we note that Proposition 3 holds for any pricing policy that guarantees $\text{Var}(p)_t \geq ct^{\alpha-1}$. The relation between Regret and $\text{Var}(p)_t$ through the parameter α, as in Theorem 1, depends, however, on the specifics of the used pricing policy.

A lower bound on $\text{Regret}(t)$ in terms of $\text{Var}(p)_t$ can easily be constructed. For example, if the revenue function $r(p)$ is strictly concave in p, then one can show that $\text{Regret}(t, \psi) \geq k\sum_{i=1}^{t}(p_i - p_{\text{opt}})^2$ for some positive constant k and any policy ψ. Since $\arg\min_{p\in\mathcal{P}}\sum_{i=1}^{t}(p_i - p)^2 = \bar{p}_t$, this implies that $t\,\text{Var}(p)_t \leq k^{-1}\text{Regret}(t, \psi)$ a.s. To derive an upper bound on the Regret in terms of the growth rate of $\text{Var}(p)_t$ for arbitrary policies ψ seems much less straightforward. It is an interesting direction for future research to completely characterize the relation between Regret and empirical variance.

Remark 1 (Differences with the Multiperiod Control Problem). For the multiperiod control problem mentioned in §3.1, Lai and Robbins (1982) showed that one can achieve $\text{Regret}(T) = O(\log T)$. This problem is very much akin to the dynamic pricing problem with linear demand function: in the first problem, the optimal control $x(\hat{a}_{0t}, \hat{a}_{1t})$ as a function of the parameter estimates equals $(y^* - \hat{a}_{0t})/\hat{a}_{1t}$, in the latter problem, the optimal price $p(\hat{a}_{0t}, \hat{a}_{1t})$ equals $-\hat{a}_{0t}/(2\hat{a}_{1t})$ (for the moment neglecting bounds on x and p and assuming $\hat{a}_{0t}$ and $\hat{a}_{1t}$ have the correct sign). When $y^* = 0$, these optimal controls only differ by a factor of 2. An intuitive explanation of why we do not achieve $\text{Regret}(T) = O(\log T)$ in the dynamic pricing problem, despite the similarities with the multiperiod control problem, is the presence of what Harrison et al. (2012) call "indeterminate equilibria." An indeterminate equilibrium occurs if there are estimates $(\hat{a}_0, \hat{a}_1)$ such that the average observed output at $x(\hat{a}_0, \hat{a}_1)$ "confirms" the correctness of these estimates, i.e., if $(\hat{a}_0, \hat{a}_1)$ satisfies $a_0^{(0)} + a_1^{(0)}x(\hat{a}_0, \hat{a}_1) = \hat{a}_0 + \hat{a}_1 x(\hat{a}_0, \hat{a}_1)$. It is not difficult to show that there are infinitely many indeterminate equilibria, both in the multiperiod control problem and the dynamic pricing problem. In the multiperiod control problem, each indeterminate equilibrium $(\hat{a}_0, \hat{a}_1) \neq (a_0^{(0)}, a_1^{(0)})$ still gives an optimal control $x(\hat{a}_0, \hat{a}_1) = x(a_0^{(0)}, a_1^{(0)})$, whereas in the dynamic pricing problem, each indeterminate equilibrium $(\hat{a}_0, \hat{a}_1) \neq (a_0^{(0)}, a_1^{(0)})$ yields a suboptimal price $p(\hat{a}_0, \hat{a}_1) \neq p(a_0^{(0)}, a_1^{(0)})$. This means that in the multiperiod control problem, convergence of the parameter estimates to any arbitrary indeterminate equilibrium implies convergence of the controls to the optimal control; whereas in the dynamic pricing problem, *only* convergence of the parameter estimates to the "true" indeterminate equilibrium $a^{(0)}$ implies convergence of the controls to the optimal control. This makes the dynamic pricing problem structurally more complex than the multiperiod control problem.

Remark 2 (Applicability to Other Sequential Decision Problems). In §§3.1 and 3.2, we discuss inconsistency of certainty equivalent pricing and study the performance of CVP in the specific context of dynamic pricing under uncertainty. We believe, however, that the ideas developed in this paper can be applied to many other types of sequential decision problems with uncertainty. We provide a brief sketch of problems for which the controlled variance pricing idea may be a fruitful approach. At each time instance $t \in \mathbb{N}$, the decision maker chooses a control $x_t \in \mathbb{R}^d$, $(d \in \mathbb{N})$, and observes a realization y_t of a random variable $Y(x_t, \theta)$, whose probability distribution depends on x_t and on an unknown parameter θ in a parameter space $\Theta \subset \mathbb{R}^d$; subsequently, a cost $c(x_t, y_t, \theta)$ is encountered. The decision maker estimates θ using historical data $(x_i, y_i)_{i\leq t}$, with an appropriate statistical estimation technique (e.g., MLE). A certainty equivalent control rule then sets the next control to

$x_{t+1} = \arg\min_x E[c(x, Y(x, \hat{\theta}_t), \hat{\theta}_t)]$, where $\hat{\theta}_t$ denotes the estimate of θ at time t. If the quality of the parameter estimates $||\hat{\theta}_t - \theta||$ depends on some measure of dispersion of the controls (e.g., on their sample variance $\mathrm{Var}(x)_t$, as in the dynamic pricing problem, or on $\lambda_{\min}(t)$, as in den Boer and Zwart 2012), then a controlled variance rule sets the next control to $x_{t+1} = \arg\min_x E[c(x, Y(x, \hat{\theta}_t), \hat{\theta}_t)]$ subject to a lower bound on the measure of dispersion. The optimal lower bound depends on the problem characteristics, and captures in some sense the trade-off between estimation and instant optimization, i.e., exploration and exploitation. An extension of the dynamic pricing problem to a setting with multiple products may also be elaborated in this fashion.

4. Numerical Evaluation

We numerically compare the performance of CVP with the policy MLE-cycle, which was introduced by Broder and Rusmevichientong (2012). We test Normal, Poisson, and Bernoulli distributed demand, since these are commonly used demand models in practice. For each distribution, we test two functions h that model the relation between price and expected demand. The function v need not be specified, since it is already determined by the demand distribution: $v(x) = 1$ for Normal demand, $v(x) = x$ for Poisson demand, and $v(x) = x(1-x)$ for Bernoulli demand. All six sets of demand distribution and the function h are listed in Table 1. For each set, we randomly generate 10,000 different instances of parameters a_0 and a_1. For Normal demand we also generate a value for σ; for Poisson and Bernoulli demand, $\sigma = 1$. The parameters are drawn from a uniform distribution. The support of these uniform distributions is chosen such that the optimal price lies between 3 and 8. For Normal demand, an additional requirement is that $h(a_0 + a_1 p_{\mathrm{opt}}) - 3\sigma > 0$ and $\sigma/(h(a_0 + a_1 p_{\mathrm{opt}})) > \frac{1}{20}$. This implies that at the optimal price, the probability that demand is negative is small (less than 0.135%), and the coefficient of variation at the optimal price is not extremely small (at least $\frac{1}{20}$). For Bernoulli demand, an additional requirement is $h(a_0 + a_1 p) \in (0, 1)$ for all $p_l \le p \le p_h$. Table 2 lists summary statistics for the chosen parameter values.

For both policies that we compare, the lowest and highest admissible price are set to $p_l = 1$ and $p_h = 10$, respectively. The policy CVP uses $\alpha = 0.5001$, and initial prices $p_1 = 4$ and $p_2 = 7$. For the constant c in the taboo interval, we try three different values: 1, 3, and 5. The exploration prices of MLE-cycle are set to $p_1 = 4$ and $p_2 = 7$. We vary the number of exploration phases per cycle. In particular, we try one, two, and three consecutive exploration phases (n consecutive exploration phases means that during the $2n$ exploration periods in each cycle, the price alternates between p_1 and p_2).

For each set of instances, we calculate the average relative regret over 10,000 instances. Thus, for each instance, corresponding to a choice of parameter values, we measure the relative regret

$$\frac{\sum_{t=1}^{T} r_{\mathrm{opt}} - r(p_t)}{T r_{\mathrm{opt}}} \times 100\%,$$

and then we average over all instances:

$$\frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \frac{\sum_{t=1}^{T} r_{\mathrm{opt}} - r(p_t)}{T r_{\mathrm{opt}}} \times 100\%.$$

This quantity is measured for $T \in \{10, 50, 100, 500, 1{,}000\}$. The results are listed in Table 3. In this table, the header CVP(c) denotes the policy CVP with constant c and the header MLE-c(n) denotes the policy MLE-cycle with n consecutive exploration phases.

The results from Table 3 suggest that CVP performs comparably to MLE-cycle, or even better. This holds for all tested time scales $T = 10, 50, 100, 500$, and 1,000, and all six sets of problem instances. If we

Table 1 **Problem Sets, with Parameter Range**

Distribution	$h(x)$	a_0	a_1	σ
1. Normal	x	$[0.1, 20]$	$[-a_0/11, -a_0/16]$	$[1/20, 1/3] \cdot (a_0 + a_1 p_{\mathrm{opt}})$
2. Normal	$x^{3/4}$	$[0.1, 20]$	$[-a_0/11, -a_0/14]$	$[1/20, 1/3] \cdot (a_0 + a_1 p_{\mathrm{opt}})^{3/4}$
3. Poisson	$\exp(x)$	$[11/3, 20]$	$[-1/3, -1/8]$	1
4. Poisson	x	$[11/3, 20]$	$[-a_0/11, -a_0/16]$	1
5. Bernoulli	$(1 + \exp(-x))^{-1}$	$[\log(-3a_1 - 1) - 3a_1, \log(-8a_1 - 1) - 8a_1]$	$[-1, -4/9]$	1
6. Bernoulli	$x^{3/4}$	$[0.8, 1.1]$	$[-a_0/11, -a_0/14]$	1

consider the results for $T = 1{,}000$, we see that CVP outperforms MLE-cycle on all problem sets except 1. On a shorter time scale, $T = 100$, this holds for all problem sets. One of the reasons for this difference is that MLE-cycle only uses the data from the exploration phases to form parameter estimates, whereas CVP uses all the available historical data.

Note that for some instances—in particular, problem set 3—the relative regret decreases very fast: CVP(1) has regret below 1% already from $T = 50$. The sets 5 and 6, with Bernoulli distributed demand, show a more slowly decreasing relative regret.

We also see that the optimal value of the constant c in CVP depends on T. For example, in problem set 6, $c = 1$ performs best for $T = 10, 50, 100$, whereas $c = 5$ is the best choice for $T = 500, 1{,}000$.

We wish to emphasize that these results are not meant as an exhaustive comparison between the numerical performance of CVP and MLE-cycle. In that case, we should also have fine-tuned the value of the exploration prices p_1 and p_2. The simulation results, nevertheless, are an indication that CVP may perform well in practical applications.

Table 2 Sample Statistics of Parameters

	a_0	a_1	σ	p_{opt}
		Problem set 1		
Max	19.9973	−0.0066	3.2775	7.9993
Mean	10.0518	−0.7712	0.9652	6.5984
Min	0.1050	−1.8004	0.0042	5.5002
Std	5.7519	0.4517	0.7246	0.7187
		Problem set 2		
Max	19.9989	−0.0075	2.8178	7.9998
Mean	10.0050	−0.8125	0.8181	7.0703
Min	0.1009	−1.8044	0.0037	6.2860
Std	5.7400	0.4704	0.6135	0.4964
		Problem set 3		
Max	19.9995	−0.1250	1.0000	7.9991
Mean	11.8249	−0.2286	1.0000	4.7182
Min	3.6669	−0.3333	1.0000	3.0004
Std	4.7345	0.0600	0	1.3508
		Problem set 4		
Max	19.9983	−0.2353	1.0000	7.9991
Mean	11.8751	−0.9094	1.0000	6.6062
Min	3.6687	−1.8095	1.0000	5.5006
Std	4.7217	0.3762	0	0.7230
		Problem set 5		
Max	9.8596	−0.4445	1.0000	7.9998
Mean	4.8056	−0.7255	1.0000	5.3353
Min	0.3068	−1.0000	1.0000	3.0006
Std	1.9504	0.1606	0	1.4570
		Problem set 6		
Max	1.1000	−0.0574	1.0000	8.0000
Mean	0.9497	−0.0770	1.0000	7.0780
Min	0.8000	−0.0997	1.0000	6.2858
Std	0.0866	0.0088	0	0.4952

Table 3 Average Relative Regret

t	CVP(1) (%)	CVP(3) (%)	CVP(5) (%)	MLE-c(1) (%)	MLE-c(3) (%)	MLE-c(5) (%)
			Problem set 1: Normal demand, $h(x) = x$			
10	5.0	5.0	5.0	7.6	8.9	8.6
50	3.2	3.1	3.2	5.0	6.7	7.2
100	2.9	2.9	2.9	3.9	5.3	6.5
500	2.7	2.7	2.7	2.0	3.0	4.1
1,000	2.7	2.6	2.7	1.5	2.2	3.2
			Problem set 2: Normal demand, $h(x) = x^{3/4}$			
10	6.8	7.2	7.5	9.4	11.0	10.4
50	4.0	3.7	3.8	7.0	8.6	8.9
100	3.2	2.8	2.8	5.9	7.0	8.1
500	1.9	1.4	1.4	3.4	4.0	5.2
1,000	1.4	1.0	1.0	2.6	3.0	4.1
			Problem set 3: Poisson demand, $h(x) = \exp(x)$			
10	2.3	2.7	3.3	5.8	7.7	9.1
50	0.9	1.3	1.9	3.1	6.3	7.3
100	0.6	1.0	1.4	2.3	5.0	6.6
500	0.3	0.4	0.7	1.2	2.9	4.2
1,000	0.2	0.3	0.5	0.8	2.2	3.3
			Problem set 4: Poisson demand, $h(x) = x$			
10	8.1	8.6	9.1	9.4	9.5	8.7
50	5.5	5.5	5.6	8.5	8.1	8.0
100	4.8	4.5	4.3	7.6	6.9	7.3
500	3.4	2.7	2.4	4.9	4.2	4.8
1,000	2.8	2.1	1.9	3.9	3.2	3.8
			Problem set 5: Bernoulli demand, $h(x) = (1 + \exp(-x))^{-1}$			
10	18.4	18.5	18.3	21.0	19.2	20.7
50	9.5	10.0	10.5	15.8	17.1	18.0
100	6.8	7.2	7.6	13.5	14.4	16.5
500	3.6	3.5	3.5	8.6	8.9	11.0
1,000	2.8	2.5	2.5	6.8	6.9	8.7
			Problem set 6: Bernoulli demand, $h(x) = x^{3/4}$			
10	11.3	11.5	11.6	11.4	12.3	10.4
50	9.2	9.8	10.1	11.1	10.8	10.4
100	8.0	8.3	8.4	11.0	10.3	10.0
500	5.8	5.4	5.0	9.9	8.1	7.9
1,000	5.0	4.4	3.9	9.0	6.8	6.5

5. Conclusions and Future Research

It is important for firms to find the optimal selling price for their products. Price experimentation can be a valuable tool to find this revenue-maximizing price; however, it may be costly if it is not executed in a proper way. A firm, therefore, needs a pricing policy that optimally balances obtaining information and gaining revenue. Controlled variance pricing is such a pricing policy, and it is proposed in this paper. The key idea is to enhance certainty equivalent pricing with a slowly shrinking taboo interval around the average price chosen up to that moment. We have shown that the parameter estimates then converge to the true values, and the chosen prices to the optimal price. The Regret after T time periods satisfies the upper bound $O(T^{1/2+\delta})$ for arbitrarily small $\delta > 0$. The intuition behind CVP is easy to understand, the method is easily implemented, and is applicable to a wide range of demand distributions. Numerical

tests indicate that CVP performs well on several time scales and demand models.

From a theoretical perspective, interesting future research is to study whether the term T^δ in the upper bound on the Regret can be removed, and to further study the relation between Regret and $\mathrm{Var}(p)_t$, as discussed in §3.3.2. From an application point of view, it would be useful to develop a heuristic that suggests which parameter c in the taboo interval (6) is approximately optimal.

Acknowledgments

The authors thank Anton Kleywegt for providing useful literature references and for introducing them to consistency problems in certainty equivalent pricing. The authors also thank Kacha Dzhaparidze for several useful discussions. The comments and suggestions of the department editor, associate editor, and anonymous referees have led to an improved version of the paper. The research of Bert Zwart is partly supported by an NWO VIDI grant and an IBM faculty award. Arnoud V. den Boer was with Centrum Wiskunde and Informatica (CWI), Amsterdam, while this paper was written.

Appendix

Proof of Proposition 1. The proof is similar to §2 of Lai and Robbins (1982), with the difference that we do not make assumptions on the values of p_1, p_2, p_l, and p_h.

Without loss of generality, assume $p_1 < p_2$, define $a = ((p_h - p_1)^2 + (p_h - p_2)^2)p_h^{-1}$, and recall that $e_t = D(p_t) - E[D(p_t) \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}]$. Write $\sigma^2 = E[e_t^2]$ for all $t \in \mathbb{N}t$. Let $\delta > 0$, and consider the event

$$A = \left\{ \begin{array}{l} (p_2 - 2p_h)e_1 + (2p_h - p_1)e_2 \geq -a_1^{(0)}(p_2 - p_1)2p_h \\ (p_1 - p_h)e_1 + (p_2 - p_h)e_2 \geq (-2p_h a_1^{(0)} - a_0^{(0)} + \delta)a \\ \qquad + (2p_h - p_1 - p_2) \\ \delta|\bar{e}_t| \leq \delta \quad \text{for all } t \geq 3 \end{array} \right\}.$$

We first show that for sufficiently large δ, the event A occurs with strictly positive probability. The first two inequalities of A are satisfied when

$$\begin{pmatrix} p_2 - 2p_h & 2p_h - p_1 \\ p_1 - p_h & p_2 - p_h \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \geq \begin{pmatrix} -a_1^{(0)}(p_2 - p_1)2p_h \\ a(-2p_h a_1^{(0)} - a_0^{(0)} + \delta) + (2p_h - p_1 - p_2)\delta \end{pmatrix}. \quad (12)$$

The determinant of the coefficient matrix equals $(p_2 - 2p_h) \cdot (p_2 - p_h) + (p_1 - p_h)(p_1 - 2p_h)$, which is strictly positive. A solution to this linear system therefore exists, and (12) happens with positive probability. Let $B \subset \mathbb{R}^2$ be a bounded subset of the solutions (e_1, e_2) of (12), such that $P((e_1, e_2) \in B) > 0$. Choose $\delta > \sqrt{8}\sigma + \sup_{(e_1,e_2)\in B} \frac{1}{3}|e_1 + e_2|$. It follows from the Kolmogorov inequality (see, e.g., Chow and Teicher 2003, Theorem 6, p. 133) that for any $\epsilon > \sqrt{8}\sigma$,

$$\begin{aligned} P\left(\sup_{3\leq t}\left|\frac{1}{t-2}\sum_{i=3}^{t} e_i\right| > \epsilon\right) &\leq \sum_{j=1}^{\infty} P\left(\sup_{2^j < t \leq 2^{j+1}}\left|\frac{1}{t-2}\sum_{i=3}^{t} e_i\right| > \epsilon\right) \\ &\leq \sum_{j=1}^{\infty} P\left(\sup_{2^j < t \leq 2^{j+1}}\left|\sum_{i=3}^{t} e_i\right| > (2^j - 1)\epsilon\right) \\ &\leq \sum_{j=1}^{\infty} P\left(\sup_{1 \leq t \leq 2^{j+1}}\left|\sum_{i=3}^{t} e_i\right| > (2^j - 1)\epsilon\right) \\ &\leq \sum_{j=1}^{\infty} \frac{1}{(2^j - 1)^2 \epsilon^2}\sigma^2 2^{j+1} \leq \sum_{j=1}^{\infty} 8\sigma^2\epsilon^{-2}2^{-j} \\ &= 8\sigma^2\epsilon^{-2} < 1, \end{aligned}$$

since $2^{j+1}/(2^j - 1)^2 \leq 8 \cdot 2^{-j}$, $j \geq 1$. This implies

$$\begin{aligned} &P(|\bar{e}_t| \leq \delta \text{ for all } t \geq 3) \\ &= P\left(\sup_{t\geq 3}|\bar{e}_t| \leq \delta\right) \\ &= P\left(\sup_{t\geq 3}\left|\frac{e_1 + e_2}{t} + \frac{1}{t}\sum_{i=3}^{t} e_i\right| \leq \delta\right) \\ &\geq P\left((e_1, e_2) \in B \text{ and } \sup_{t\geq 3}\left|\frac{1}{t}\sum_{i=3}^{t} e_i\right| \leq \delta - \sup_{(e_1,e_2)\in B}\frac{1}{3}|e_1 + e_2|\right) \\ &= P((e_1, e_2) \in B) \cdot P\left(\sup_{t\geq 3}\left|\frac{1}{t}\sum_{i=3}^{t} e_i\right| \leq \left(\delta - \sup_{(e_1,e_2)\in B}\frac{1}{3}|e_1 + e_2|\right)\right) \\ &\geq P((e_1, e_2) \in B) \cdot P\left(\sup_{t\geq 3}\left|\frac{1}{t-2}\sum_{i=3}^{t} e_i\right| \leq \left(\delta - \sup_{(e_1,e_2)\in B}\frac{1}{3}|e_1 + e_2|\right)\right) \\ &> 0. \end{aligned}$$

This proves that for δ sufficiently large, the event A occurs with probability $P(A) > 0$.

If for some t, $\hat{a}_{1t} \geq 0$ or $\hat{a}_{0t} \leq 0$, then clearly the parameter estimates have the wrong sign; it would be foolish for a price manager to use the certainty equivalent price $\arg\max_{p_l \leq p \leq p_h} p \cdot (\hat{a}_{0t} + \hat{a}_{1t}p)$ in that case. We therefore assume that $p_{t+1} = p_h$ whenever $\hat{a}_{1t} \geq 0$. (Alternatively, one might impose some extra conditions in the set A, and still use $p_{t+1} = \arg\max_{p_l \leq p \leq p_h} p \cdot (\hat{a}_{0t} + \hat{a}_{1t}p)$ when the estimates have the wrong sign.)

We show by induction that on the event A, $p_{t+1} = p_h$ for all $t \geq 2$.

Case $t = 2$. The line through the points $(p_i, a_0^{(0)} + a_1^{(0)}p_i + e_i)$, $i = 1, 2$ has slope $\hat{a}_{12} = a_1^{(0)} + (e_2 - e_1)(p_2 - p_1)^{-1}$ and intercept $\hat{a}_{02} = (e_1 p_2 - e_2 p_1)(p_2 - p_1)^{-1}$. If $\hat{a}_{12} \geq 0$, then $p_3 = p_h$. If $\hat{a}_{12} < 0$, then $p_3 = \hat{a}_{02}/(-2\hat{a}_{12}) \geq p_h$ is implied by

$$(e_1 p_2 - e_2 p_1)(p_2 - p_1)^{-1} \geq -2\big(a_1^{(0)} + (e_2 - e_1)(p_2 - p_1)^{-1}\big)p_h,$$

which, by multiplying $(p_2 - p_1)$ and rearranging terms, is equivalent to the condition

$$(p_2 - 2p_h)e_1 + (2p_h - p_1)e_2 \geq -a_1^{(0)}(p_2 - p_1)2p_h.$$

Case $t \geq 3$. Suppose that for all $i = 3, \ldots, t$, $p_i = p_h$. Then $\bar{p}_i = p_h - ((2p_h - p_1 - p_2)/i)$ $(3 \leq i \leq t)$. Defining $C_t = \sum_{i=1}^{t}(p_i - \bar{p}_t)e_i$ and $V_t = \sum_{i=1}^{t}(p_i - \bar{p}_t)^2$, the least squares estimates are

$$\begin{pmatrix} \hat{a}_{0t} \\ \hat{a}_{1t} \end{pmatrix} = \begin{pmatrix} a_0^{(0)} \\ a_1^{(0)} \end{pmatrix} + \begin{pmatrix} \bar{e}_t - \bar{p}_t C_t/V_t \\ C_t/V_t \end{pmatrix}.$$

For all $t \geq 2$, V_t and C_t can be rewritten as

$$V_t = \sum_{i=2}^{t}\frac{i-1}{i}(p_i - \bar{p}_{i-1})^2, \qquad C_t = \sum_{i=2}^{t}\frac{i-1}{i}(p_i - \bar{p}_{i-1})(e_i - \bar{e}_{i-1}),$$

and by some algebra and an induction argument, it follows that

$$V_t = V_2 + (2p_h - p_1 - p_2)^2\left(\tfrac{1}{2} - t^{-1}\right),$$
$$C_t = C_2 + (2p_h - p_1 - p_2)(\bar{e}_t - \bar{e}_2),$$

where $V_2 = \frac{1}{2}(p_2 - p_1)^2$ and $C_2 = \frac{1}{2}(p_2 - p_1)(e_2 - e_1)$.

If $\hat{a}_{1t} \geq 0$, then $p_{t+1} = p_h$.

Now suppose $\hat{a}_{1t} < 0$. Then,

$$\begin{aligned}
p_{t+1} = p_h
&\Leftrightarrow \quad \frac{\hat{a}_{0t}}{-2\hat{a}_{1t}} \geq p_h \\
&\Leftrightarrow \quad \hat{a}_{0t} \geq -2\hat{a}_{1t}p_h \\
&\Leftrightarrow \quad a_0^{(0)} + \bar{e}_t - \bar{p}_t C_t/V_t \geq -2p_h a_1^{(0)} - 2p_h C_t/V_t \\
&\Leftrightarrow \quad \bar{e}_t + (2p_h - \bar{p}_t)C_t/V_t \geq -2p_h a_1^{(0)} - a_0^{(0)} \\
&\Leftrightarrow \quad \left(p_h + \frac{2p_h - p_1 - p_2}{t}\right)C_t/V_t \geq -2p_h a_1^{(0)} - a_0^{(0)} - \bar{e}_t.
\end{aligned}$$

Observe that on the event A, $-\bar{e}_t < \delta$, $p_h + (2p_h - p_1 - p_2)/t \geq p_h$, $V_t \leq V_2 + (2p_h - p_1 - p_2)^2 \cdot \frac{1}{2}$, $C_t = C_2 + (2p_h - p_1 - p_2)\cdot(\bar{e}_t - \bar{e}_2) \geq C_2 + (2p_h - p_1 - p_2)(-\delta - \bar{e}_2)$, and thus it suffices to show

$$\begin{aligned}
&C_2 + (2p_h - p_1 - p_2)(-\delta - \bar{e}_2) \\
&\quad \geq (-2p_h a_1^{(0)} - a_0^{(0)} + \delta)\left(V_2 + (2p_h - p_1 - p_2)^2 \cdot \tfrac{1}{2}\right)p_h^{-1};
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\tfrac{1}{2}(p_2 - p_1)(e_2 - e_1) - (2p_h - p_1 - p_2)\bar{e}_2 \\
&\geq (-2p_h a_1^{(0)} - a_0^{(0)} + \delta)\left(\tfrac{1}{2}(p_2 - p_1)^2 + (2p_h - p_1 - p_2)^2 \cdot \tfrac{1}{2}\right)p_h^{-1} \\
&\quad + (2p_h - p_1 - p_2)\delta.
\end{aligned}$$

Rewriting the left-hand side, we get the condition

$$\begin{aligned}
&e_1(p_1 - p_h) + e_2(p_2 - p_h) \\
&\geq (-2p_h a_1^{(0)} - a_0^{(0)} + \delta)\left(\tfrac{1}{2}(p_2 - p_1)^2 + (2p_h - p_1 - p_2)^2 \cdot \tfrac{1}{2}\right)p_h^{-1} \\
&\quad + (2p_h - p_1 - p_2)\delta \\
&= (-2p_h a_1^{(0)} - a_0^{(0)} + \delta)a + (2p_h - p_1 - p_2)\delta. \quad \square
\end{aligned}$$

Proof of Proposition 2. We prove the assertion by induction. For $t = 2$, observe that the upper bound $c \leq 2^{-\alpha}(p_1 - p_2)^2$ on the constant c implies $\mathrm{Var}(p)_2 = (p_1 - p_2)^2/2 \geq c2^{\alpha-1}$. Now let $t \geq 2$ and suppose that $\mathrm{Var}(p)_t \geq ct^{\alpha-1}$. If (3) has no solution, $\hat{a}_{0t} \leq 0$, $\hat{a}_{1t} \geq 0$, or $\hat{a}_{0t} + \hat{a}_{1t}p < 0$ for some $p \in [p_l, p_h]$, then $|p_{t+1} - \bar{p}_t| = \max(|p_1 - \bar{p}_t|, |p_2 - \bar{p}_t|) \geq |p_1 - p_2|/2$. Observe that for all $t \geq 2$ and $\alpha \in (0,1)$, $(t+1)^\alpha - t^\alpha \leq \alpha t^{\alpha-1}$. Together with the bound $c \leq 2^{-\alpha}(3\alpha)^{-1}(p_1 - p_2)^2$, this implies

$$\begin{aligned}
(t+1)\,\mathrm{Var}(p)_{t+1} &= t\,\mathrm{Var}(p)_t + \frac{t}{t+1}(p_{t+1} - \bar{p}_t)^2 \\
&\geq ct^\alpha + c[(t+1)^\alpha - t^\alpha] = c(t+1)^\alpha.
\end{aligned}$$

If (7) is chosen, then automatically $\mathrm{Var}(p)_{t+1} \geq c(t+1)^{\alpha-1}$.

If (8) is chosen, then by construction of the taboo interval (6), we have

$$\begin{aligned}
(t+1)\,\mathrm{Var}(p)_{t+1} &= t\,\mathrm{Var}(p)_t + \frac{t}{t+1}(p_{t+1} - \bar{p}_t)^2 \\
&\geq ct^\alpha + c[(t+1)^\alpha - t^\alpha] = c(t+1)^\alpha. \quad \square
\end{aligned}$$

Proof of Lemma 1. From $\lambda_{\max}(t) + \lambda_{\min}(t) = \mathrm{trace}(P_t) = t(1 + \bar{p}_t^2) > 0$ and $\lambda_{\max}(t)\lambda_{\min}(t) = \det(P_t) = t^2\,\mathrm{Var}(p)_t > 0$, it follows that $\lambda_{\min}(t) > 0$. Together with $\bar{p}_t^2 \leq p_h$, we thus have $\lambda_{\max}(t) \leq t(1 + \bar{p}_t^2) \leq t(1 + p_h^2)$. Furthermore, $\lambda_{\min}(t) = \lambda_{\max}(t)^{-1}\det(P_t) = \lambda_{\max}(t)^{-1}t^2\,\mathrm{Var}(p)_t \geq (1 + p_h^2)^{-1}t \cdot \mathrm{Var}(p)_t$. $\square$

Proof of (11), in Case of Normally Distributed Linear Demand. We assume $D(p) \sim N(a_0^{(0)} + a_1^{(0)}p, \sigma^2)$ for some $\sigma > 0$, and show

$$E[||\hat{a}_t - a^{(0)}||^2] = O\left(\frac{\log t}{t^\alpha}\right).$$

Write $q_t = \sum_{i=1}^t e_i\binom{1}{p_i}$ and $Q_t = q_t P_t^{-1} q_t$. By (4) we have $\hat{a}_t = a^{(0)} + P_t^{-1}q_t$ for $t \geq 2$, and thus

$$\begin{aligned}
E[||\hat{a}_t - a^{(0)}||^2] = E[||P_t^{-1}q_t||^2] &\leq E[||P_t^{-1/2}||^2||P_t^{-1/2}q_t||^2] \\
&\leq E[Q_t](1 + p_h)^2 c^{-1}t^{-\alpha},
\end{aligned}$$

where we used $||P_t^{-1/2}||^2 = \lambda_{\max}(P_t^{-1/2})^2 = \lambda_{\min}(P_t^{-1}) \leq (1 + p_h)^2/(t\,\mathrm{Var}(p)_t) \leq (1 + p_h)^2 c^{-1}t^{-\alpha}$ a.s., by Lemma 1 and Proposition 2. It thus suffices to show $E[Q_t] = O(\log t)$. We have

$$\begin{aligned}
Q_t &= (q_{t-1}^T + (1, p_t)e_t)P_t^{-1}\left(q_{t-1} + \binom{1}{p_t}e_t\right) \\
&= Q_{t-1} + q_{t-1}^T(P_t^{-1} - P_{t-1}^{-1})q_{t-1} + (1, p_t)P_t^{-1}\binom{1}{p_t}e_t^2 \\
&\quad + 2(1, p_t)P_t^{-1}q_{t-1}e_t. \qquad (13)
\end{aligned}$$

By the Sherman–Morrison formula (Bartlett 1951),

$$P_t^{-1} = P_{t-1}^{-1} - \left(P_{t-1}^{-1}\binom{1}{p_t}(1, p_t)P_{t-1}^{-1}\right)\Big/\left(1 + (1, p_t)P_{t-1}^{-1}\binom{1}{p_t}\right),$$

which implies

$$\begin{aligned}
&q_{t-1}^T(P_t^{-1} - P_{t-1}^{-1})q_{t-1} \\
&= -\left(q_{t-1}^T P_{t-1}^{-1}\binom{1}{p_t}(1, p_t)P_{t-1}^{-1}q_{t-1}\right)\Big/\left(1 + (1, p_t)P_{t-1}^{-1}\binom{1}{p_t}\right) \\
&= -((1, p_t)P_{t-1}^{-1}q_{t-1})^2\Big/\left(1 + (1, p_t)P_{t-1}^{-1}\binom{1}{p_t}\right) \leq 0 \text{ a.s.} \qquad (14)
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&E\left[(1, p_t)P_t^{-1}\binom{1}{p_t}e_t^2\right] \\
&= E\left[(1, p_t)P_t^{-1}\binom{1}{p_t}E[e_t^2 \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}]\right] \\
&\leq \sigma^2 E\left[(1, p_t)P_t^{-1}\binom{1}{p_t}\right], \qquad (15)
\end{aligned}$$

and

$$E[2(1,p_t)P_t^{-1}q_{t-1}e_t]$$
$$= E[2(1,p_t)P_t^{-1}q_{t-1}E[e_t \mid p_1,\ldots,p_{t-1},d_1,\ldots,d_{t-1}]] = 0. \quad (16)$$

Combining Equations (13)–(16), we obtain for all $t \geq 2$,

$$E[Q_t] = E[Q_2] + E\left[\sum_{j=3}^{t} Q_j - Q_{j-1}\right]$$
$$\leq E[Q_2] + \sigma^2 \sum_{j=3}^{t} E\left[(1,p_j)P_j^{-1}\begin{pmatrix}1\\ p_j\end{pmatrix}\right].$$

Sylvester's determinant theorem, $\det(I+AB) = \det(I+BA)$, for matrices A, B, implies

$$\det(P_{t-1}) = \det(P_t)\det\left(I - P_t^{-1}\begin{pmatrix}1\\ p_t\end{pmatrix}(1,p_t)\right)$$
$$= \det(P_t)\left(1-(1,p_t)P_t^{-1}\begin{pmatrix}1\\ p_t\end{pmatrix}\right),$$

and thus

$$E[Q_t] \leq E[Q_2] + \sigma^2 \sum_{j=3}^{t} E\left[\frac{\det(P_j)-\det(P_{j-1})}{\det(P_j)}\right].$$

Write $y_j = \det(P_j)/\det(P_2)$, $j \geq 2$, and $R_t = (\log y_t)^{-1} \cdot \sum_{j=3}^{t}(y_j - y_{j-1})/y_j$, $t \geq 3$. We show by induction that $R_t \leq C$ for all $t \geq 3$, where $C = 1 + 1/\log y_3$. For $t = 3$, this follows immediately. Now let $t > 3$ and assume $R_{t-1} \leq C$. Define $g(y) = (1 - y_{t-1}/y + R_{t-1}\log y_{t-1})/\log y$, and observe $R_t = g(y_t)$. If $R_{t-1} \leq C - 1/\log y_t$, then $R_t = g(y_t) \leq 1/\log y_t + R_{t-1} \leq C$. If $R_{t-1} > C - 1/\log y_t$, then $R_{t-1} > 1 + 1/\log y_3 - 1/\log y_t \geq 1$, and

$$\frac{\partial g(y)}{\partial y} = \frac{1}{y(\log y)^2}\left[-1 + \frac{y_{t-1}}{y}(1+\log y) - \log(y_{t-1})R_{t-1}\right] < 0.$$

The term between square brackets is strictly smaller than zero, which follows from $R_{t-1} > 1$, $y_{t-1} \geq 1$, and the fact that $(1+\log z)/z$ is decreasing in z, for $z \geq 1$. It follows that $R_t = g(y_t) \leq \max_{y \geq y_{t-1}} g(y) = g(y_{t-1}) = R_{t-1} \leq C$. This completes the proof of the fact $R_t \leq C$ for all $t \geq 3$. As a result,

$$E[Q_t] \leq E[Q_2] + \sigma^2 C \log(\det(P_t)/\det(P_2))$$
$$\leq E[Q_2] + \sigma^2 C \log((1+p_h^2)^2 t^2/\det(P_2)) = O(\log t),$$

where we used Lemma 1 and $\det(P_t) \leq \lambda_{\max}(t)^2$. □

Proof of Theorem 1. Since $r(p,a)$ is twice continuously differentiable in p, it follows from a Taylor-series expansion that, given a, for all $p \in [p_l, p_h]$, there is a $\tilde{p} \in [p_l, p_h]$ on the line segment between p and p_{opt}, such that

$$r(p,a) = r(p_{\text{opt}},a) + r'(p_{\text{opt}},a)(p-p_{\text{opt}}) + \tfrac{1}{2}r''(\tilde{p},a)(p-p_{\text{opt}})^2,$$

where $r'(p,a)$ and $r''(p,a)$ denote the first and second derivatives of r with respect to p. The assumption $p_l < p_{\text{opt}} < p_h$ implies $r'(p_{\text{opt}}) = 0$, and with $K := \sup_{p\in[p_l,p_h]}|r''(p)| < \infty$, it follows that

$$|r(p) - r(p_{\text{opt}})| \leq \frac{K}{2}(p-p_{\text{opt}})^2, \quad (p \in \mathcal{P}). \quad (17)$$

On an open neighborhood of p_{opt} in $\mathcal{P}$, $p_{\text{opt}} = p(a^{(0)})$ is the unique solution to $r'(p,a^{(0)}) = 0$. Since by assumption $r''(p,a_0,a_1)$ exists and is nonzero at the point $(p(a^{(0)}),a^{(0)})$, it follows from the implicit function theorem (see, e.g., Duistermaat and Kolk 2004) that there are open neighborhoods U of $p(a^{(0)})$ in $\mathbb{R}$ and V of $a^{(0)}$ in $\mathbb{R}^2$, such that for each $a \in V$ there is a unique $p \in U$ with $r'(p,a) = 0$. Moreover, the mapping $V \to U$, $a \mapsto p(a)$ is continuously differentiable in V. Consequently, for all $a \in V$, there is a $\tilde{a} \in V$ on the line segment between a and $a^{(0)}$, such that

$$p(a) = p(a^{(0)}) + \left.\frac{\partial p(a)}{\partial a^T}\right|_{\tilde{a}}(a-a^{(0)}),$$

which implies that we can choose V such that for all $a \in V$,

$$|p(a) - p(a^{(0)})| = O(||a-a^{(0)}||). \quad (18)$$

Let $\rho > 0$ such that $\{a\colon ||a-a^{(0)}|| \leq \rho\} \subset V$ and $||a-a^{(0)}|| \leq \rho \Rightarrow a_0 > 0$, $a_1 < 0$, and let T_ρ be as in (9). This implies that for all $t \in \mathbb{N}$,

$$E[|p_t - p_{\text{opt}}|^2] = E[|p_t-p_{\text{opt}}|^2 \cdot \mathbf{1}_{t>T_\rho}] + E[|p_t-p_{\text{opt}}|^2 \cdot \mathbf{1}_{t\leq T_\rho}]$$
$$\leq E[|p_t-p_{\text{opt}}|^2 \cdot \mathbf{1}_{t>T_\rho}] + (p_h-p_l)^2 P(t \leq T_\rho)$$
$$\leq E[|p_t-p_{\text{opt}}|^2 \cdot \mathbf{1}_{t>T_\rho}] + (p_h-p_l)^2 \frac{E[T_\rho^{1/2}]}{t^{1/2}}, \quad (19)$$

where $\mathbf{1}_A$ denotes the indicator function of the event A.

Since $r'(p(a^{(0)}),a^{(0)}) = 0$ and $r''(p(a^{(0)}),a^{(0)}) < 0$, it follows from continuity arguments that $r'(p(a),a) = 0$ and $r''(p(a),a) < 0$ for all a in an open neighborhood of $a^{(0)}$. This implies that if $||\hat{a}_t - a^{(0)}||$ is sufficiently small and t sufficiently large, $\arg\max_{p\in[p_l,p_h]\setminus \mathrm{TI}(t)} r(p,\hat{a}_t)$ lies on the boundary of the taboo interval $\mathrm{TI}(t)$. It follows that there is a $\rho > 0$ such that for all $t > T_\rho$,

$$|p_t - p(\hat{a}_t)| \leq |\mathrm{TI}(t)| \quad (20)$$

where $|\mathrm{TI}(t)|$ denotes the length of the taboo interval $\mathrm{TI}(t)$. Combining (18), (20), and (19),

$$E[|p_t-p_{\text{opt}}|^2]$$
$$\leq E[|p_t-p_{\text{opt}}|^2 \cdot \mathbf{1}_{t>T_\rho}] + (p_h-p_l)^2\frac{E[T_\rho^{1/2}]}{t^{1/2}},$$
$$\leq 2E[|p_t-p(\hat{a}_t)|^2 \cdot \mathbf{1}_{t>T_\rho}] + 2E[|p(\hat{a}_t)-p(a^{(0)})|^2 \cdot \mathbf{1}_{t>T_\rho}]$$
$$+ (p_h-p_l)^2\frac{E[T_\rho^{1/2}]}{t^{1/2}},$$
$$= O\left(E[|\mathrm{TI}(t)|^2] + E[||\hat{a}_t-a^{(0)}||^2 \cdot \mathbf{1}_{t>T_\rho}] + (p_h-p_l)^2\frac{E[T_\rho^{1/2}]}{t^{1/2}}\right),$$
$$= O\left(t^{\alpha-1} + \frac{\log t}{t^\alpha}\right),$$

by Proposition 3, from which follows

$$\text{Regret}(T) = \sum_{t=1}^{T} E[r(p_{\text{opt}}) - r(p_t)] = O\left(\sum_{t=1}^{T} E[(p_t-p_{\text{opt}})^2]\right)$$
$$= O\left(\sum_{t=1}^{T} t^{\alpha-1} + t^{-\alpha}\log t\right)$$
$$= O(T^\alpha + T^{1-\alpha}\log T). \quad \square$$

References

Aghion P, Bolton P, Harris C, Jullien B (1991) Optimal learning by experimentation. *Rev. Econom. Stud.* 58(4):621–654.

Anderson TW, Taylor JB (1976) Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica* 44(6):1289–1302.

Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5): 1169–1188.

Balvers RJ, Cosimano TF (1990) Actively learning about demand and the dynamics of price adjustment. *Econom. J.* 100(402): 882–898.

Bartlett MS (1951) An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Statist.* 22(1):107–111.

Bertsimas D, Perakis G (2006) Dynamic pricing: A learning approach. Hearn D, Lawphongpanich S, eds. *Mathematical and Computational Models for Congestion Charging* (Springer, New York), 45–79.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.

Besbes O, Zeevi A (2011) On the minimax complexity of pricing in a changing environment. *Oper. Res.* 59(1):66–79.

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.

Carvalho AX, Puterman ML (2005a) Dynamic optimization and learning: How should a manager set prices when the demand function is unknown? IPEA Discussion Paper 1117. Instituto de Pesquisa Economica Aplicada, Brasilia.

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

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, New York).

Chen K, Hu I (1998) On consistency of Bayes estimates in a certainty equivalence adaptive system. *IEEE Trans. Automatic Control* 43(7):943–947.

Chow YS, Teicher H (2003) *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed. (Springer Verlag, New York).

Cope E (2007) Bayesian strategies for dynamic pricing in e-commerce. *Naval Res. Logist.* 54(3):265–281.

den Boer AV, Zwart B (2012) Mean square convergence rates for maximum quasi-likelihood estimators. Working paper, University of Technology, Eindhoven, The Netherlands. https://www.researchgate.net/publication/257985753_Mean_square_convergence_rates_for_maximum_quasi-likelihood_estimators.

Duistermaat JJ, Kolk JAC (2004) *Multidimensional Real Analysis: Differentiation*, Cambridge Studies in Advanced Mathematics, Vol. 86 (Cambridge University Press, Cambridge, UK).

Easley D, Kiefer NM (1988) Controlling a stochastic process with unknown parameters. *Econometrica* 56(5):1045–1064.

Eren SS, Maglaras C (2010) Monopoly pricing with limited demand information. *J. Revenue Pricing Management* 9:23–48.

Farias VF, van Roy B (2010) Dynamic pricing with a prior on market response. *Oper. Res.* 58(1):16–29.

Gill J (2001) *Generalized Linear Models: A Unified Approach* (Sage Publications, Thousand Oaks, CA).

Gittins JC (1989) *Multi-Armed Bandit Allocation Indices*, Wiley Interscience Series in Systems and Optimization (John Wiley & Sons, New York).

Godambe VP, Heyde CC (1987) Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* 55(3):231–244.

Goldenshluger A, Zeevi A (2009) Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* 19(4):1603–1633.

Harrison JM, Keskin NB, Zeevi A (2012) Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Sci.* 58(3):570–586.

Heyde CC (1997) *Quasi-Likelihood and Its Application*. Springer Series in Statistics (Springer Verlag, New York).

Keller G, Rady S (1999) Optimal experimentation in a changing environment. *Rev. Econom. Stud.* 66(3):475–507.

Keskin NB, Zeevi A (2013) Dynamic pricing with an unknown linear demand model: Asymptotically optimal semi-myopic policies. Working paper, University of Chicago Booth School of Business, Chicago. http://faculty.chicagobooth.edu/bora.keskin/pdfs/DynamicPricingUnknownDemandModel.pdf.

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23(3):462–466.

Kiefer NM, Nyarko Y (1989) Optimal control of an unknown linear process with learning. *Internat. Econom. Rev.* 30(3):571–586.

Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *Proc. 44th IEEE Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 594–605.

Lai TL, Robbins H (1982) Iterated least squares in multiperiod control. *Adv. Appl. Math.* 3(1):50–73.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Lai TL, Wei CZ (1982) Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* 10(1):154–166.

Lai TL, Robbins H, Wei CZ (1979) Strong consistency of least squares estimates in multiple regression II. *J. Multivariate Anal.* 9(3):343–361.

Lim AEB, Shanthikumar JG (2007) Relative entropy, exponential utility, and robust dynamic pricing. *Oper. Res.* 55(2):198–214.

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

Lobo MS, Boyd S (2003) Pricing and learning with uncertain demand. Working paper, Stanford University, Stanford, CA. http://www.stanford.edu/~boyd/papers/pdf/pric_learn_unc_dem.pdf.

McCullagh P (1983) Quasi-likelihood functions. *Ann. Statist.* 11(1):59–67.

McCullagh P, Nelder JA (1983) *Generalized Linear Models* (Chapman & Hall, London).

McLennan A (1984) Price dispersion and incomplete learning in the long run. *J. Econom. Dynam. Control* 7(3):331–347.

Nassiri-Toussi K, Ren W (1994) On the convergence of least squares estimates in white noise. *IEEE Trans. Automatic Control* 39(2):364–368.

Powell WB (2010) The knowledge gradient for optimal learning. Cochran JJ, Cox LA Jr, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, New York).

Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

Taylor JB (1974) Asymptotic properties of multiperiod control rules in the linear regression model. *Internat. Econom. Rev.* 15(2):472–484.

Vermorel J, Mohri M (2005) Multi-armed bandit algorithms and empirical evaluation. Gama J, Camacho R, Brazdil P, Jorge A, Torgo L, eds. *Proceedings of the 16th European Conference on Machine Learning*, Lecture Notes in Computer Science, Vol. 3720 (Springer-Verlag, Berlin), 437–448.

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61(3):439–447.