Simultaneously Learning and Optimizing using Controlled Variance Pricing

Arnoud den Boer
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands, boer@cwi.nl
Bert Zwart
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands, VU University Amsterdam, Department of Mathematics, De Boelelaan 1081a, 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. The key idea of the policy is to enhance the certainty equivalent pricing policy with a taboo interval around the average 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 Controlled Variance Pricing performs well, on different demand models and time scales.

Key words: dynamic pricing; sequential decision problems; statistical learning

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 sub-optimal 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 are a normal, Poisson, or Bernoulli distributed demand, with expectation a linear or logit function of the price. In order 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 then can 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 the 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, in order 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 in order 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 to 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 to 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 amount of money lost due to not knowing 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 2009b, Broder and Rusmevichientong 2010, Kleinberg and Leighton 2003). Apart from this theoretical result, we also numerically compare the performance of CVP with another existing pricing policies from the literature. These numerical experiments suggest that CVP performs well for different demand functions and time scales.

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 OR 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 multi-armed bandit problems and stochastic approximation.

**Parametric approaches based on classical estimation methods.** In these studies it is assumed that the unknown demand function is in a parametrized family of distributions; the unknown parameters are estimated by classical statistical methods such as maximum likelihood or least squares estimation. Examples are Lobo and Boyd (2003), Carvalho and Puterman (2005b,a), Bertsimas and Perakis (2006), Besbes and Zeevi (2009a) and Broder and Rusmevichientong (2010).

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 to two other policies: certainty equivalent pricing, and certainty equivalent pricing with added random perturbations. Simulation results on a short time horizon (ten periods) suggest that the convex approximation performs better than the other two policies.

The papers by Carvalho and Puterman (2005b,a) 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. Both papers do not provide analytical results on the performance of the policies. It is however numerically compared to several other policies; these simulations suggest 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 (2010). They propose a policy called 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 (2009a), where performance bounds are derived in a certain asymptotic regime.

**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 (2009), Harrison et al. (2010). All these papers assume a single unknown parameter, and propose several pricing heuristics. We mention in particular a policy proposed by Harrison et al. (2010). Similar to the Controlled Variance Pricing policy developed in this paper, they suggest a policy based on the idea of enhancing the certainty equivalent policy with a taboo interval. A difference is that they employ taboo intervals of fixed length, whereas we use intervals that shrink at a certain, specifically chosen speed which optimizes the expected revenue and ensures the convergence of the prices to the optimal price. A fixed-length taboo interval corresponds to choosing \( \alpha = 1 \) in Section 3.2; we show however that an optimal balance of exploration and exploitation is achieved when \( \alpha \) is chosen close to 1/2.

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.

**Nonparametric approaches.** Robust or nonparametric approaches do not assume a known parametric functional form of the demand, and generally investigate how to minimize worst case scenarios. Such approaches can be found in Kleinberg and Leighton (2003), Cope (2006), Lim and Shanthikumar (2007), Eren and Maglaras (2008) and Besbes and Zeevi (2009a).
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) which 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.

Multi-armed bandit problem and stochastic approximation. Dynamic pricing with unknown demand has similarities with the classical multi-armed bandit problem: both are sequential decision problem 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 type 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 multi-armed 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 which converges to zero. This is similar to the taboo interval approach taken in this study (see further Section 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.
Consistency of linear regression and maximum quasi-likelihood estimation

Our results on the strong consistency of Controlled Variance Pricing are based on Chang (1999), who establishes sufficient conditions for consistency of maximum quasi-likelihood estimation. His results are 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 \( \lambda_{\text{min}}(t), \lambda_{\text{max}}(t) \) for the smallest and largest eigenvalue of the design matrix, they showed that \( \lambda_{\text{min}}(t) \to \infty \) a.s. and \( (\log \lambda_{\text{max}}(t))/\lambda_{\text{min}}(t) \to 0 \) 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 \( \lambda_{\text{min}}(t) \to \infty \) 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 \( (\log \lambda_{\text{max}}(t))/\lambda_{\text{min}}(t) \to 0 \) a.s. cannot simply be removed.

Our analysis of Controlled Variance Pricing is valid for a large class of demand models. This is in contrast to Lobo and Boyd (2003), Carvalho and Puterman (2005b,a) 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 (2010) or Besbes and Zeevi (2009a); here only the average price converges to the optimal price. Moreover, in this 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 \( \alpha \)); this makes the method potentially suitable for extensions to models with multiple products. 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 (2010) and Besbes and Zeevi (2009a) only uses 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 is described in Section 2, followed by a short discussion on the model assumptions (Section 2.1) and the method to estimate the unknown parameters (Section 2.2). We show in Section 3 that certainty equivalent pricing is not consistent, which motivates the introduction of Controlled Variance Pricing (CVP). We show that under this policy the parameter estimates converge to the true value, and 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. In Section 4, CVP is numerically compared to another pricing policy from the literature, on different time scales and different demand functions. Conclusions and directions for future research are provided in Section 5. The Appendix, Section 6, contains the proofs of the propositions in the 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 price that are acceptable to the firm. After setting the price, the firm observes a realization $d_t$ of the demand, 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. stock-outs 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:

$$E[D(p)] = h(a_0^{(0)} + a_1^{(0)} p),$$

$$\text{Var}[D(p)] = \sigma^2 v(E[D(p)]).$$

Here $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $v : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ are both twice continuously differentiable known functions, with $\dot{h}(x) := \frac{\partial h(x)}{\partial x} > 0$ for all $x \geq 0$. Furthermore, $\sigma$ 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 $e_t = d_t - E[D(p_t) \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}]$. We make the technical assumption on the demand that for some $r > 2$,

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

(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 $p_{\text{opt}} = p(a_0^{(0)}, a_1^{(0)})$ 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 $\psi$ 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 $p_{\text{opt}}$. For a pricing policy $\psi$ 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 $\psi$ that maximize the total expected revenue over a finite number of $T$ time periods. This is equivalent to minimizing $\text{Regret}(T, \psi)$. Note however that the regret can not 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 $\bar{x}_t := \frac{1}{t} \sum_{i=1}^{t} x_i$ denotes the sample mean and $\text{Var}(x)_t = \frac{1}{t} \sum_{i=1}^{t} (x_i - \bar{x}_t)^2$ 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 non-random sequences $(x_n)_n, (y_n)_n$, $x_n = O(y_n)$ means that there exists a $K > 0$ such that $|x_n| \leq K|y_n|$ for all $n$.

**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 distributions (\( v(x) = 1 \) and arbitrary \( \sigma > 0 \)). All these examples also satisfy the moment condition (2).

The assumptions on the functions \( h, v \) and parameters \( \sigma, a_0^{(0)}, a_1^{(0)} \) imply that the expected demand is strictly decreasing in the price, and that the variance is strictly positive; i.e. is non-deterministic. 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 \( h(x) = x, h(x) = \exp(x), h(x) = (1 + \exp(-x))^{-1} \). A sufficient condition to satisfy this assumption is that the revenue function \( r(p, a^{(0)}) \) is strictly concave in \( p \), and attains it maximum strictly between \( p_l \) and \( p_h \).

### 2.2. Estimating the Unknown Parameters

The unknown parameters \( a^{(0)} \) can be estimated with Maximum Quasi Likelihood estimation. This is a natural extension of Maximum Likelihood Estimation to settings where only the first two moments of the distribution are known; see Wedderburn (1974), McCullagh (1983), 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 maximum quasi-likelihood estimator (MQLE) of \( (a_0^{(0)}, a_1^{(0)}) \), denoted by \( \hat{a}_t = (\hat{a}_0 t, \hat{a}_1 t) \), is the solution to the two-dimensional equation

\[
\ell_t(\hat{a}_t) = \sum_{i=1}^t \frac{\dot{h}(\hat{a}_0 t + \hat{a}_1 t p_i)}{\sigma^2 v(h(\hat{a}_0 t + \hat{a}_1 t p_i))} \left( \frac{1}{p_i} \right) (d_i - h(\hat{a}_0 t + \hat{a}_1 t p_i)) = 0.
\]

(3)

If the probability density function (or probability mass function in case of discrete demand distribution) of \( D(p) \) can be written in the form \( \exp(\sigma^{-1}(d\theta - g(\theta))) \), where \( \theta \) 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 in this class. (see e.g. McCullagh and Nelder 1983, Gill 2001). In case of Normal distributed demand, (3) is also equivalent to the normal equations of ordinary least squares.

### 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 \( t \geq 2 \) time periods, calculate the MQLE estimators \( \hat{a}_t \) with (3), and set the next price \( p_{t+1} \) equal to

\[
p_{t+1} = \arg \max_{p \in [p_l, p_h]} r(p, \hat{a}_0 t, \hat{a}_1 t).
\]

(4)
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 (2005b,a). Anderson and Taylor (1976) studied a linear system $y_t = a_0^{(0)} + a_1^{(0)} x_t + \epsilon_t$ with unknown parameters $a_0^{(0)}, a_1^{(0)}$ and input variables $x_t$; the objective is to steer $y_t$ to a desired value $y^*$. The certainty equivalent policy is to set $x_{t+1} = \min \{ x_{\text{max}}, \max \{ x_{\text{min}}, y^* - \hat{a}_0 t - \hat{a}_1 t \} \}$, where $\hat{a}_0 t, \hat{a}_1 t$ are the least square estimates of $a_0^{(0)}, a_1^{(0)}$, based on observations $(x_1, y_1), \ldots, (x_t, y_t)$, and $x_{\text{min}}, x_{\text{max}}$ 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 $p_1, p_2, p_l, p_h$; 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 $p_{\text{opt}}$.

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 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 $p_h$, but different from $p_{\text{opt}}$, and perhaps even converge to $p_{\text{opt}}$ with probability zero.
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 $t$, $\text{Var}(p)_t \geq c(t+1)^{-1}$, for some $c > 0$ and $\alpha \in (0, 1)$.

The pricing policy we propose, called Controlled Variance Pricing, chooses at each time period the certainty equivalent price (4), unless this means that the lower bound on the sample variance of the prices $\text{Var}(p)_{t+1} \geq c(t+1)^{-1}$ 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

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

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

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

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

Else, set

$$p_{t+1} = \arg \max_{p \in [p_l, p_h]\setminus TI(t)} r(p, \hat{a}_t).$$
where $TI(t)$ is the taboo interval (5) at time $t$.

In cases (a) - (c), we choose one of the initial prices $p_1, p_2$, that is most far away from $p_t$. 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 $[p_1, p_h] \setminus TI(t)$ 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 $p_{\text{opt}}$, 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 $\text{Var}(p)_t$ converges slowly to zero (cf. Proposition 4); 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 $\alpha$ of CVP. The following Proposition establishes a relation between $\alpha$ and the sample variance of the prices:

**Proposition 2.** With CVP, $\text{Var}(p)_t \geq ct^{\alpha-1}$ 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 $\lambda_{\text{max}}(t), \lambda_{\text{min}}(t)$ be the largest and smallest eigenvalue of the design matrix

$$P_t = \left( \frac{t}{\sum_{i=1}^t p_i \sum_{i=1}^t p_i^2} \right), \quad (t \in \mathbb{N})$$

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

In the following Propositions 3, 4 and Theorem 1, 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 $a^{(0)}$. 

PROPOSITION 3. A solution $\hat{a}_t$ to (3) eventually exists, and $\hat{a}_t \to a^{(0)}$ a.s., provided $\alpha > 1/2$.

The assertion follows by application of Theorem 1.1 of Chang (1999).

The speed at which the parameter estimates converge to the true value, is related to the sample variance of the prices $\text{Var}(p)_t$. The relation is characterized in the following Proposition.

PROPOSITION 4. For $\rho > 0$, let

$$T_\rho = \sup\{t \in \mathbb{N} | \text{there is no solution } \hat{a}_t \text{ of (3) such that } ||\hat{a}_t - a^{(0)}|| \leq \rho\}. \quad (8)$$

Then for all sufficiently small $\rho > 0$, $E[T_\rho] < \infty$ and

$$E \left[ ||\hat{a}_t - a^{(0)}||^2 \right] = O \left( \frac{\log t}{t^{\alpha}} \right), \quad (t > T_\rho), \quad (9)$$

provided $\alpha > 1/2$.

Lai and Wei (1982) studied the a.s. convergence rates $||\hat{a}_t - a^{(0)}||^2$ for ordinary linear regression estimates (i.e. with $h(x) = x, v(x) = 1$). The main part of the proof of Proposition 4 is to expand their result to convergence in expectation, and to the more general MQLE setting we consider. The full proof is given in the Appendix.

The convergence rate in Proposition 4 enables us to calculate the following upper bound on the Regret:

THEOREM 1.

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

provided $\alpha > 1/2$.

We use a Taylor series expansion of the revenue function $r(p)$ to show $|r(p) - r(p_{opt})| = O((p - p_{opt})^2)$. The implicit function theorem is invoked to obtain $|p(a) - p_{opt}| = O(||a - a^{(0)}||)$. The theorem can then be derived from Proposition 4 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 (9) on the quality of the parameter estimates, the term $T^\alpha$ comes from the length of the taboo interval (5). The parameter $\alpha$ captures the trade-off between learning and optimization. If $\alpha$ 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 $\alpha$ 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 \( \alpha \), only a relatively slow convergence of the parameter estimates is guaranteed by Proposition 4. The optimal choice of \( \alpha \) in Theorem 1 clearly is \( 1/2 \), but since \( \alpha \) should be larger than \( 1/2 \), we get the following Corollary 1.

\[
\text{Regret}(T, CVP) = O\left(T^{1/2 + \delta}\right),
\]

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

This result would be a little more elegant if the term \( T^\delta \), \( \delta > 0 \), could be removed. The strong consistency result of Chang (1999) however requires \( \alpha > 1/2 \), and this requirement appears crucial to his result. In the case of linear demand with constant variance, i.e. \( h(x) = x, v(x) = 1 \), it follows from Lai and Wei (1982) that Propositions 3 and 4 also hold for \( \alpha = 1/2 \), which implies that then CVP has \( \text{Regret}(T, CVP) = O\left(\sqrt{T \log T}\right) \). This suggest that, at least in some cases, Corollary 1 can slightly be improved.

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 2009b, Broder and Rusmevichientong 2010, Kleinberg and Leighton 2003).

3.3.2. Generality of Result. Concerning the generality of the result, we note that Propositions 3 and 4 hold 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 \( \alpha \), 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 \( \text{Regret}(t, \psi) \geq k \sum_{i=1}^{t} (p_i - p_{opt})^2 \), for some positive constant \( k \) and any policy \( \psi \). Since \( \arg \min_{p \in P} \sum_{i=1}^{t} (p_i - p)^2 = \bar{p}_t \), this implies \( t \text{Var}(p)_t \leq k^{-1} \text{Regret}(t, \psi) \) a.s. To derive an upper bound on the Regret in terms of the growth rate of \( \text{Var}(p)_t \), for arbitrary policies \( \psi \), 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 Section 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 function of the
parameter estimates equals \( (y^* - \hat{a}_0) / \hat{a}_1 \), in the latter problem the optimal price \( p(\hat{a}_0, \hat{a}_1) \) equals 
\(-\hat{a}_0 / (2\hat{a}_1)\) (for the moment neglecting bounds on \( x, p \) and assuming \( \hat{a}_0, \hat{a}_1 \) have the correct sign). When \( y^* = 0 \), these optimal controls only differ by a factor 2. An intuitive explanation why \( \text{Regret}(T) = O(\log T) \) is not possible in the dynamic pricing problem, despite the similarities with the multiperiod control problem, is the presence of what Harrison et al. (2010) call 'uninformative equilibria'. An uninformative equilibrium occurs if there are estimates \( (\hat{a}_0, \hat{a}_1) \) such that the average output \( y \) at \( x(\hat{a}_0, \hat{a}_1) \) 'confirms' the correctness of these estimates, but where \( x(\hat{a}_0, \hat{a}_1) \) is not equal to the optimal control; more precisely, if \( (\hat{a}_0, \hat{a}_1) \) satisfies \( x(\hat{a}_0, \hat{a}_1) \neq x(\hat{a}_0(0), \hat{a}_1(0)) \) and \( \hat{a}_0(0) + \hat{a}_1(0) x(\hat{a}_0, \hat{a}_1) = \hat{a}_0 + \hat{a}_1 x(\hat{a}_0, \hat{a}_1) \). It is not difficult to show that there are no uninformative equilibria in the multiperiod control problem, and infinitely many in the dynamic pricing problem. The dynamic pricing problem is therefore structurally more complex than the multiperiod control problem. This offers an intuitive explanation why \( \text{Regret}(T) = O(\log T) \) can be achieved in the latter problem, whereas the optimal \( \text{Regret} \) in the first problem grows with rate at least \( \sqrt{T} \).

**Remark 2** *(Applicability to other Sequential Decision Problems).* In Sections 3.1 and 3.2 we discuss inconsistency of certainty equivalent pricing, and study the performance of Controlled Variance pricing, 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, for example inventory control, or admission control problems in queueing networks. 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 \( x_t \in \mathbb{R} \), 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 \( \theta \) in a parameter space \( \Theta \subset \mathbb{R}^d \); subsequently a cost \( c(x_t, y_t, \theta) \) is encountered. The decision maker estimates \( \theta \) using historical data \( (x_i, y_i)_{i \leq t} \), with an appropriate statistical estimation technique (e.g. Maximum Likelihood estimation). 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 \( \theta \) 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. their sample variance \( \text{Var}(x)_t \), as in the dynamic pricing problem), then a controlled variance rule sets the next control to \( x_{t+1} = \arg\min_x E[c(x, Y(x, \bar{\theta}_t), \bar{\theta}_t)] \) 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, e.g. exploration and exploitation.

**4. Numerical Evaluation**
We numerically compare the performance of Controlled Variance Pricing to the policy MLE-cycle, which was introduced by Broder and Rusmevichientong (2010). 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 models 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 combinations of demand distribution and the function $h$ are listed in Table 1. For each set, we randomly generate 10,000 different instances of parameters $a_0, a_1$. For Normal demand we also generate a value for $\sigma$; 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 $h(a_0 + a_1 p_{\text{opt}}) - 3\sigma > 0$ and $\frac{\sigma}{h(a_0 + a_1 p_{\text{opt}})} > \frac{1}{20}$. 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 $p_l \leq p \leq p_h$. Table 2 list summary statistics for the chosen parameter values.

For both policies that we compare, the lowest and highest admissible price are set to $p_l = 1$, $p_h = 10$. The policy CVP uses $\alpha = 0.5001$, and initial prices $p_1 = 4, p_2 = 7$. 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 $p_1 = 4, p_2 = 7$. Similar to the experiments in Broder and Rusmevichientong (2010), we vary the number of exploration phases per cycle. In particular we try 1, 2, and 3 consecutive exploration phases. ($n$ consecutive exploration phases means that during the $2n$ exploration periods in each cycle, the price alternates between $p_1$ and $p_2$).

<table>
<thead>
<tr>
<th>Distr.</th>
<th>$h(x)$</th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Normal</td>
<td>$x$</td>
<td>[0.1, 20]</td>
<td>$[-a_0, -a_0]$</td>
<td>$[\frac{1}{20}, \frac{1}{3}]$</td>
</tr>
<tr>
<td>2. Normal</td>
<td>$x^{3/4}$</td>
<td>[0.1, 20]</td>
<td>$[-a_0, -a_0]$</td>
<td>$[\frac{1}{20}, \frac{1}{3}]$</td>
</tr>
<tr>
<td>3. Poisson</td>
<td>$\exp(x)$</td>
<td>$[\frac{1}{2}, 20]$</td>
<td>$[-a_0, -a_0]$</td>
<td>$[\frac{1}{2}, \frac{1}{3}]$</td>
</tr>
<tr>
<td>4. Poisson</td>
<td>$x$</td>
<td>$[\frac{1}{2}, 20]$</td>
<td>$[-a_0, -a_0]$</td>
<td>$[\frac{1}{2}, \frac{1}{3}]$</td>
</tr>
<tr>
<td>5. Bernoulli</td>
<td>$(1 + \exp(-x))^{-1}$</td>
<td>$[\log(-3a_1 - 1) - 3a_1, \log(-8a_1 - 1) - 8a_1]$</td>
<td>$[-a_0, -a_0]$</td>
<td>$[-1, \frac{-a_1}{\pi}]$</td>
</tr>
<tr>
<td>6. Bernoulli</td>
<td>$x^{3/4}$</td>
<td>[0.8, 1.1]</td>
<td>$[-a_0, -a_0]$</td>
<td>$[\frac{-a_0}{11}, \frac{-a_0}{14}]$</td>
</tr>
</tbody>
</table>
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

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

and then we average over all instances:

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

This quantity is measured for $T \in \{10, 50, 100, 500, 1000\}$. The results are listed in Table 3. In these tables, the header CVP($c$) denotes the policy CVP with constant $c$, the header MLE-c($n$) denotes the policy MLE-cycle with $n$ consecutive exploration phases.

### Table 2 Sample statistics of parameters

<table>
<thead>
<tr>
<th>Problem set 1</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>19.9973</td>
<td>-0.0066</td>
<td>3.2775</td>
<td>7.9993</td>
</tr>
<tr>
<td>mean</td>
<td>10.0518</td>
<td>-0.7712</td>
<td>0.9652</td>
<td>6.5984</td>
</tr>
<tr>
<td>min</td>
<td>0.1050</td>
<td>-1.8004</td>
<td>0.0042</td>
<td>5.5002</td>
</tr>
<tr>
<td>std</td>
<td>5.7519</td>
<td>0.4517</td>
<td>0.7246</td>
<td>0.7187</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem set 2</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>19.9989</td>
<td>-0.0075</td>
<td>2.8178</td>
<td>7.9998</td>
</tr>
<tr>
<td>mean</td>
<td>10.0050</td>
<td>-0.8125</td>
<td>0.8181</td>
<td>7.0703</td>
</tr>
<tr>
<td>min</td>
<td>0.1009</td>
<td>-1.8044</td>
<td>0.0037</td>
<td>6.2860</td>
</tr>
<tr>
<td>std</td>
<td>5.7400</td>
<td>0.4704</td>
<td>0.6135</td>
<td>0.4964</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem set 3</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>19.9995</td>
<td>-0.1250</td>
<td>1.0000</td>
<td>7.9991</td>
</tr>
<tr>
<td>mean</td>
<td>11.8249</td>
<td>-0.2286</td>
<td>1.0000</td>
<td>4.7182</td>
</tr>
<tr>
<td>min</td>
<td>3.6669</td>
<td>-0.3333</td>
<td>1.0000</td>
<td>3.0004</td>
</tr>
<tr>
<td>std</td>
<td>4.7345</td>
<td>0.0600</td>
<td>1.3508</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem set 4</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>19.9983</td>
<td>-0.2353</td>
<td>1.0000</td>
<td>7.9991</td>
</tr>
<tr>
<td>mean</td>
<td>11.8751</td>
<td>-0.9094</td>
<td>1.0000</td>
<td>6.6062</td>
</tr>
<tr>
<td>min</td>
<td>3.6687</td>
<td>-1.8095</td>
<td>1.0000</td>
<td>5.5006</td>
</tr>
<tr>
<td>std</td>
<td>4.7217</td>
<td>0.3762</td>
<td>0.7230</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem set 5</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>9.8596</td>
<td>-0.4445</td>
<td>1.0000</td>
<td>7.9998</td>
</tr>
<tr>
<td>mean</td>
<td>4.8056</td>
<td>-0.7255</td>
<td>1.0000</td>
<td>5.3353</td>
</tr>
<tr>
<td>min</td>
<td>0.3068</td>
<td>-1.0000</td>
<td>1.0000</td>
<td>3.0006</td>
</tr>
<tr>
<td>std</td>
<td>1.9504</td>
<td>0.1606</td>
<td>1.4570</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem set 6</th>
<th>a0</th>
<th>a1</th>
<th>σ</th>
<th>popt</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>1.1000</td>
<td>-0.0574</td>
<td>1.0000</td>
<td>8.0000</td>
</tr>
<tr>
<td>mean</td>
<td>0.9497</td>
<td>-0.0770</td>
<td>1.0000</td>
<td>7.0780</td>
</tr>
<tr>
<td>min</td>
<td>0.8000</td>
<td>-0.0997</td>
<td>1.0000</td>
<td>6.2858</td>
</tr>
<tr>
<td>std</td>
<td>0.0866</td>
<td>0.0088</td>
<td>0.0000</td>
<td>0.4952</td>
</tr>
</tbody>
</table>
### Table 3  Average relative regret

**Problem set 1: Normal demand, \( h(x) = x \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5.0 %</td>
<td>5.0 %</td>
<td>5.0 %</td>
<td>7.6 %</td>
<td>8.9 %</td>
<td>8.6 %</td>
</tr>
<tr>
<td>50</td>
<td>3.2 %</td>
<td>3.1 %</td>
<td>3.2 %</td>
<td>5.0 %</td>
<td>6.7 %</td>
<td>7.2 %</td>
</tr>
<tr>
<td>100</td>
<td>2.9 %</td>
<td>2.9 %</td>
<td>2.9 %</td>
<td>3.9 %</td>
<td>5.3 %</td>
<td>6.5 %</td>
</tr>
<tr>
<td>500</td>
<td>2.7 %</td>
<td>2.7 %</td>
<td>2.7 %</td>
<td>2.0 %</td>
<td>3.0 %</td>
<td>4.1 %</td>
</tr>
<tr>
<td>1000</td>
<td>2.7 %</td>
<td>2.6 %</td>
<td>2.7 %</td>
<td>1.5 %</td>
<td>2.2 %</td>
<td>3.2 %</td>
</tr>
</tbody>
</table>

**Problem set 2: Normal demand, \( h(x) = x^{3/4} \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6.8 %</td>
<td>7.2 %</td>
<td>7.5 %</td>
<td>9.4 %</td>
<td>11.0 %</td>
<td>10.4 %</td>
</tr>
<tr>
<td>50</td>
<td>4.0 %</td>
<td>3.7 %</td>
<td>3.8 %</td>
<td>7.0 %</td>
<td>8.6 %</td>
<td>8.9 %</td>
</tr>
<tr>
<td>100</td>
<td>3.2 %</td>
<td>2.8 %</td>
<td>2.8 %</td>
<td>5.9 %</td>
<td>7.0 %</td>
<td>8.1 %</td>
</tr>
<tr>
<td>500</td>
<td>1.9 %</td>
<td>1.4 %</td>
<td>1.4 %</td>
<td>3.4 %</td>
<td>4.0 %</td>
<td>5.2 %</td>
</tr>
<tr>
<td>1000</td>
<td>1.4 %</td>
<td>1.0 %</td>
<td>1.0 %</td>
<td>2.6 %</td>
<td>3.0 %</td>
<td>4.1 %</td>
</tr>
</tbody>
</table>

**Problem set 3: Poisson demand, \( h(x) = \exp(x) \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.3 %</td>
<td>2.7 %</td>
<td>3.3 %</td>
<td>5.8 %</td>
<td>7.7 %</td>
<td>9.1 %</td>
</tr>
<tr>
<td>50</td>
<td>0.9 %</td>
<td>1.3 %</td>
<td>1.9 %</td>
<td>3.1 %</td>
<td>6.3 %</td>
<td>7.3 %</td>
</tr>
<tr>
<td>100</td>
<td>0.6 %</td>
<td>1.0 %</td>
<td>1.4 %</td>
<td>2.3 %</td>
<td>5.0 %</td>
<td>6.6 %</td>
</tr>
<tr>
<td>500</td>
<td>0.3 %</td>
<td>0.4 %</td>
<td>0.7 %</td>
<td>1.2 %</td>
<td>2.9 %</td>
<td>4.2 %</td>
</tr>
<tr>
<td>1000</td>
<td>0.2 %</td>
<td>0.3 %</td>
<td>0.5 %</td>
<td>0.8 %</td>
<td>2.2 %</td>
<td>3.3 %</td>
</tr>
</tbody>
</table>

**Problem set 4: Poisson demand, \( h(x) = x \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.1 %</td>
<td>8.6 %</td>
<td>9.1 %</td>
<td>9.4 %</td>
<td>9.5 %</td>
<td>8.7 %</td>
</tr>
<tr>
<td>50</td>
<td>5.5 %</td>
<td>5.5 %</td>
<td>5.6 %</td>
<td>8.5 %</td>
<td>8.1 %</td>
<td>8.0 %</td>
</tr>
<tr>
<td>100</td>
<td>4.8 %</td>
<td>4.5 %</td>
<td>4.3 %</td>
<td>7.6 %</td>
<td>6.9 %</td>
<td>7.3 %</td>
</tr>
<tr>
<td>500</td>
<td>3.4 %</td>
<td>2.7 %</td>
<td>2.4 %</td>
<td>4.9 %</td>
<td>4.2 %</td>
<td>4.8 %</td>
</tr>
<tr>
<td>1000</td>
<td>2.8 %</td>
<td>2.1 %</td>
<td>1.9 %</td>
<td>3.9 %</td>
<td>3.2 %</td>
<td>3.8 %</td>
</tr>
</tbody>
</table>

**Problem set 5: Bernoulli demand, \( h(x) = (1 + \exp(-x))^{-1} \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>18.4 %</td>
<td>18.5 %</td>
<td>18.3 %</td>
<td>21.0 %</td>
<td>19.2 %</td>
<td>20.7 %</td>
</tr>
<tr>
<td>50</td>
<td>9.5 %</td>
<td>10.0 %</td>
<td>10.5 %</td>
<td>15.8 %</td>
<td>17.1 %</td>
<td>18.0 %</td>
</tr>
<tr>
<td>100</td>
<td>6.8 %</td>
<td>7.2 %</td>
<td>7.6 %</td>
<td>13.5 %</td>
<td>14.4 %</td>
<td>16.5 %</td>
</tr>
<tr>
<td>500</td>
<td>3.6 %</td>
<td>3.5 %</td>
<td>3.5 %</td>
<td>8.6 %</td>
<td>8.9 %</td>
<td>11.0 %</td>
</tr>
<tr>
<td>1000</td>
<td>2.8 %</td>
<td>2.5 %</td>
<td>2.5 %</td>
<td>6.8 %</td>
<td>6.9 %</td>
<td>8.7 %</td>
</tr>
</tbody>
</table>

**Problem set 6: Bernoulli demand, \( h(x) = x^{3/4} \)**

<table>
<thead>
<tr>
<th>t</th>
<th>CVP(1)</th>
<th>CVP(3)</th>
<th>CVP(5)</th>
<th>MLE-c (1)</th>
<th>MLE-c (3)</th>
<th>MLE-c (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11.3 %</td>
<td>11.5 %</td>
<td>11.6 %</td>
<td>11.4 %</td>
<td>12.3 %</td>
<td>10.4 %</td>
</tr>
<tr>
<td>50</td>
<td>9.2 %</td>
<td>9.8 %</td>
<td>10.1 %</td>
<td>11.1 %</td>
<td>10.8 %</td>
<td>10.4 %</td>
</tr>
<tr>
<td>100</td>
<td>8.0 %</td>
<td>8.3 %</td>
<td>8.4 %</td>
<td>11.0 %</td>
<td>10.3 %</td>
<td>10.0 %</td>
</tr>
<tr>
<td>500</td>
<td>5.8 %</td>
<td>5.4 %</td>
<td>5.0 %</td>
<td>9.9 %</td>
<td>8.1 %</td>
<td>7.9 %</td>
</tr>
<tr>
<td>1000</td>
<td>5.0 %</td>
<td>4.4 %</td>
<td>3.9 %</td>
<td>9.0 %</td>
<td>6.8 %</td>
<td>6.5 %</td>
</tr>
</tbody>
</table>
The results from Table 3 suggest that CVP performs comparable to MLE-cycle, or even better. This hold for all tested time scales $T = 10, 50, 100, 500$ and 1000, and all six sets of problem instances. If we consider the results for $T = 1000$, we see that CVP outperforms MLE-cycle 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 $T = 500, 1000$.

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 $p_1, p_2$. The simulation results nevertheless are an indication that CVP may perform well in practical applications.

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 (CVP) is such a pricing policy, and 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 if the term $T^\delta$ in the upper bound on the Regret can be removed, and to further study the relation between Regret and $\text{Var}(p)_t$, as discussed in Section 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 (5) is approximately optimal.
6. Appendix

6.1. Proof of Proposition 1

The proof is similar to Section 2 of Lai and Robbins (1982), with the difference that we do not make assumptions on the values of \( p_1, p_2, p_t, p_h \).

W.l.o.g. assume \( p_1 < p_2 \), define \( a = (p_h - p_1)^2 + (p_h - p_2)^2 \), and recall that \( e_t = d_t - E[D(p_t) \mid p_1, \ldots, p_{t-1}, d_1, \ldots, d_{t-1}] \). Write \( \sigma^2 = E[e_t^2] \), for all \( t \in \mathbb{N} \). Let \( \delta > 0 \), and consider the event

\[
A = \left\{ \begin{array}{l}
(p_2 - 2p_h)e_1 + (2p_h - p_1)e_2 \geq -a_1^{(0)}(p_2 - p_1)2p_h \\
(p_1 - p_h)e_1 + (p_2 - p_h)e_2 \geq (2p_ha_1^{(0)} - a_0^{(0)} + \delta)a + (2p_h - p_1 - p_2)\delta
\end{array} \right. \}
\]

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

\[
\left( \begin{array}{c}
(p_2 - 2p_h, 2p_h - p_1) \\
(p_1 - p_h, p_2 - p_h)
\end{array} \right) \left( \begin{array}{c}
(e_1) \\
(e_2)
\end{array} \right) \geq \left( \begin{array}{c}
-a_1^{(0)}(p_2 - p_1)2p_h \\
(a(2p_ha_1^{(0)} - a_0^{(0)} + \delta) + (2p_h - p_1 - p_2)\delta)
\end{array} \right).
\]

The determinant of the coefficient matrix equals \((p_2 - 2p_h)(p_2 - p_h) + (p_1 - p_h)(p_1 - 2p_h)\), which is strictly positive. A solution to this linear system therefore exists, and (10) happens with positive probability. Let \( B \subset \mathbb{R}^2 \) be a bounded subset of the solutions \((e_1, e_2)\) of (10), s.t. \( P((e_1, e_2) \in B) > 0 \). Choose \( \delta > \sqrt{8}\sigma + \sup_{(e_1, e_2) \in B} \frac{1}{4} |e_1 + e_2| \). It follows from the Kolmogorov inequality (see e.g. Chow and Teicher 2003, Theorem 6, page 133) that for any \( \epsilon > \sqrt{8}\sigma \),

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

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

\[
P(|e_t| \leq \delta \text{ for all } t \geq 3) = P(\sup_{t \geq 3} |e_t| \leq \delta) = P\left( \sup_{t \geq 3} \left| \frac{e_1 + e_2}{t} + \frac{1}{t} \sum_{i=3}^{t} e_i \right| \leq \delta \right) \geq P\left( (e_1, e_2) \in B \text{ and } \sup_{t \geq 3} \left| \frac{1}{t} \sum_{i=3}^{t} e_i \right| \leq \delta - \sup_{(e_1, e_2) \in B} \frac{1}{3} |e_1 + e_2| \right) = P((e_1, e_2) \in B) \cdot P\left( \sup_{t \geq 3} \left| \frac{1}{t} \sum_{i=3}^{t} e_i \right| \leq \left( \delta - \sup_{(e_1, e_2) \in B} \frac{1}{3} |e_1 + e_2| \right) \right) \geq P((e_1, e_2) \in B) \cdot P\left( \sup_{t \geq 3} \left| \frac{1}{t - 2} \sum_{i=3}^{t} e_i \right| \leq \left( \delta - \sup_{(e_1, e_2) \in B} \frac{1}{3} |e_1 + e_2| \right) \right) > 0.
\]
This proves that for $\delta$ 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 $\arg\max_{p_t \leq p \leq p_h} p \cdot (\hat{a}_{0t} + \hat{a}_{1t} p)$ 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 $p_{t+1} = \arg\max_{p_t \leq p \leq p_h} p \cdot (\hat{a}_{0t} + \hat{a}_{1t} p)$ when the estimates have the wrong sign).

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

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

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

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

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

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

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

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

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

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

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

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

$$p_{t+1} = p_h \Leftrightarrow \frac{\hat{a}_{0t}}{-2\hat{a}_{1t}} \geq p_h$$
\[
\iffalse a_{0l} \geq -2 \hat{a}_{1l} p_h 
\iffalse a_0^{(0)} + \hat{c} - \bar{p}_t C_t / V_t \geq -2 p_h a_1^{(0)} - 2 p_h C_t / V_t 
\iffalse \hat{c}_t + (2 p_h - \bar{p}_t) C_t / V_t \geq -2 p_h a_1^{(0)} - a_0^{(0)} 
\iffalse (p_h + \frac{2 p_h - p_1 - p_2}{t}) C_t / V_t \geq -2 p_h a_1^{(0)} - a_0^{(0)} - \hat{c}_t.
\]

Observe that on the event \( A \), \(- \hat{c}_t < \delta, p_h + \frac{2 p_h - p_1 - p_2}{t} \geq p_h, V_t \leq V_2 + (2 p_h - p_1 - p_2)^2 \cdot \frac{1}{2}, C_t = C_2 + (2 p_h - p_1 - p_2)(\hat{c}_t - \hat{c}_2) \geq C_2 + (2 p_h - p_1 - p_2)(-\delta - \hat{c}_2) \) and thus it suffices to show
\[
C_2 + (2 p_h - p_1 - p_2)(-\delta - \hat{c}_2) \geq (-2 p_h a_1^{(0)} - a_0^{(0)} + \delta)(V_2 + (2 p_h - p_1 - p_2)^2 \cdot \frac{1}{2}) p_h^{-1},
\]
i.e.
\[
\frac{1}{2} (p_2 - p_1)(e_2 - e_1) - (2 p_h - p_1 - p_2) \hat{c}_2 \\
\geq (-2 p_h a_1^{(0)} - a_0^{(0)} + \delta)(\frac{1}{2} (p_2 - p_1)^2 + (2 p_h - p_1 - p_2)^2 \cdot \frac{1}{2} p_h^{-1} + (2 p_h - p_1 - p_2) \delta.
\]

Rewriting the lefthandside we get the condition
\[
e_1(p_1 - p_h) + e_2(p_2 - p_h) \\
\geq (-2 p_h a_1^{(0)} - a_0^{(0)} + \delta)(\frac{1}{2} (p_2 - p_1)^2 + (2 p_h - p_1 - p_2)^2 \cdot \frac{1}{2} p_h^{-1} + (2 p_h - p_1 - p_2) \delta \\
= (-2 p_h a_1^{(0)} - a_0^{(0)} + \delta) \alpha + (2 p_h - p_1 - p_2) \delta.
\]

### 6.2. Proof of Proposition 2

We proof 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 \( \text{Var}(p)_2 = \frac{(p_1 - p_2)^2}{2} \geq c 2^{\alpha - 1} \).

Now let \( t \geq 2 \) and suppose that \( \text{Var}(p)_t \geq c 2^{\alpha - 1} \).

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 \( p \in [p_1, p_h] \), then \( |p_{t+1} - \bar{p}_t| = \max(|p_1 - \bar{p}_t|, |p_2 - \bar{p}_t|) \geq \frac{|p_1 - p_2|}{2} \).

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
\[
(t + 1) \text{Var}(p)_{t+1} = t \text{Var}(p)_t + \frac{t}{t + 1} (p_{t+1} - \bar{p}_t)^2 \geq ct^\alpha + c [(t + 1)^\alpha - t^\alpha] = c(t + 1)^\alpha.
\]

If (6) is chosen, then automatically \( \text{Var}(p)_{t+1} \geq c(t + 1)^{\alpha - 1} \).

If (7) is chosen, then by construction of the taboo interval (5) we have
\[
(t + 1) \text{Var}(p)_{t+1} = t \text{Var}(p)_t + \frac{t}{t + 1} (p_{t+1} - \bar{p}_t)^2 \geq ct^\alpha + c [(t + 1)^\alpha - t^\alpha] = c(t + 1)^\alpha.
\]
6.3. Proof of Lemma 1
From \( \lambda_{\text{max}}(t) + \lambda_{\text{min}}(t) = \text{trace}(P_t) = t(1 + \beta^2) > 0 \) and \( \lambda_{\text{max}}(t) \lambda_{\text{min}}(t) = \det(P_t) = t^2 \text{Var}(p) > 0 \), it follows that \( \lambda_{\text{min}}(t) > 0 \). Together with \( \beta^2 \leq p_h \) we thus have \( \lambda_{\text{max}}(t) \leq t(1 + \beta^2) \leq t(1 + p^2_h) \). Furthermore, \( \lambda_{\text{min}}(t) = \lambda_{\text{max}}(t)^{-1} \det(P_t) = \lambda_{\text{max}}(t)^{-1} t^2 \text{Var}(p) \geq (1 + p^2_h)^{-1} t \text{Var}(p). \)

6.4. Proof of Proposition 4
Let \( \rho > 0 \). The assertion \( E[T_\rho] < \infty \) was derived by Chang (1999), we will show (9). For \( t > T_\rho \), let \( \hat{a}_t \) be a solution to the MQLE equations (3). A Taylor expansion of \( l_t(\hat{a}_t) \) around \( a^{(0)} \) yields

\[
0 = \frac{1}{t} l_t(\hat{a}_t) = \frac{1}{t} l_t(a^{(0)}) + \left( \frac{1}{t} \nabla l_t(a^{(0)}) + R_t(\hat{a}_t) \right) (\hat{a}_t - a^{(0)}) \quad \text{a.s.,}
\]

where \( R_t(\hat{a}_t) \) is a \( 2 \times 2 \) random matrix, and, writing \( g(r) = \frac{\hat{h}(r)}{\text{as} \hat{h}(r)} \),

\[
\nabla l_t(a^{(0)}) = \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) (1, p_i) \epsilon_i - \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) (1, p_i) \hat{h}(a^{(0)}_i + a^{(0)}_i p_i).
\]

The first term in (12) is a matrix of which each entry is a martingale. By the strong law of large numbers (see e.g. Csörgő 1968) and the fact that \( \inf_{p \in \mathcal{P}} g(a^{(0)}_i + a^{(0)}_i p) > 0 \), we have \( \nabla l_t(a^{(0)}) \sim - \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) (1, p_i) \hat{h}(a^{(0)}_i + a^{(0)}_i p_i) \) a.s. Moreover, from Proposition 3 and the definition of the CVP policy it follows that

\[
\frac{1}{t} \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) (1, p_i) \hat{h}(a^{(0)}_i + a^{(0)}_i p_i) \quad (13)
\]

converges a.s., and \( ||R_t(\hat{a}_t)|| \to 0 \) a.s., and thus

\[
E[||\hat{a}_t - a^{(0)}||] = O \left( E[||\nabla l_t(a^{(0)})^{-1} l_t(a^{(0)})||] \right) =
O \left( E \left[ \left\| \left( \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) (1, p_i) \hat{h}(a^{(0)}_i + a^{(0)}_i p_i) \right)^{-1} \sum_{i=1}^{t} g(a^{(0)}_i + a^{(0)}_i p_i) \left( \frac{1}{p_i} \right) \epsilon_i \right| \right] \right).
\]

Let \( w_i := g(a^{(0)}_i + a^{(0)}_i p_i) \hat{h}(a^{(0)}_i + a^{(0)}_i p_i) \), and \( \epsilon_i = \epsilon_i \hat{h}(a^{(0)}_i + a^{(0)}_i p_i)^{-1} \). Notice that \( 0 < \inf_{i \in \mathbb{N}} w_i \leq \sup_{i \in \mathbb{N}} w_i < \infty \), and \( (\epsilon_i)_{i \in \mathbb{N}} \) is a martingale difference sequence w.r.t. its natural filtration. We show that for \( t \geq T_\rho \),

\[
E \left[ \left\| \left( \sum_{i=1}^{t} w_i \left( \frac{1}{p_i} \right) (1, p_i) \right)^{-1} \sum_{i=1}^{t} w_i \left( \frac{1}{p_i} \right) \epsilon_i \right| \right] \leq \frac{K \log t}{t^\alpha},
\]

for some \( K > 0 \). Note that the left hand side can be written as \( E \left[ \left\| (X_t^T W_t X_t)^{-1} X_t^T W_t E_t \right\| \right] \), where \( X_t \) is a \( t \times 2 \) matrix whose \( i \)-th row consists of \( (1, p_i) \), \( W_t \) is a \( t \times t \) matrix with \( w_1, \ldots, w_t \) on the diagonal and all other entries zero, and \( E_t \) is the vector \( (\epsilon_1, \ldots, \epsilon_t)^T \).
Let \( v_{\min}(X_t^T W_t X_t) \) and \( v_{\min}(X_t^T X_t) \) be normalized eigenvectors corresponding to \( \lambda_{\min}(X_t^T W_t X_t) \) and \( \lambda_{\min}(X_t^T X_t) \). Since for all \( z \in \mathbb{R}^2 \),

\[
Z^T (X_t^T W_t X_t) Z = \sum_{i=1}^{t} w_i (x_i^T z_i)^2 x_i^T z_i \in \left[ \inf_{i \in \mathbb{N}} w_i, \sup_{i \in \mathbb{N}} w_i \right] \sum_{i=1}^{t} (x_i^T z_i)^T x_i^T z_i = \left[ \inf_{i \in \mathbb{N}} w_i, \sup_{i \in \mathbb{N}} w_i \right] Z^T (X_t^T X_t) Z,
\]

it follows that

\[
\lambda_{\min}(X_t^T X_t) = \min_{\|v\|=1} v^T (X_t^T X_t) v \leq v_{\min}(X_t^T W_t X_t) v_{\min}(X_t^T X_t) \leq v_{\min}(X_t^T W_t X_t) v_{\min}(X_t^T X_t) \cdot \frac{1}{\inf_{i \in \mathbb{N}} w_i} = \lambda_{\min}(X_t^T W_t X_t) \cdot \frac{1}{\inf_{i \in \mathbb{N}} w_i}
\]

and

\[
\lambda_{\max}(X_t^T W_t X_t) \leq \text{trace}(X_t^T W_t X_t) \leq \text{trace}(X_t^T X_t) \cdot \sup_{i \in \mathbb{N}} w_i \leq \lambda_{\max}(X_t^T X_t) \cdot 2 \cdot \sup_{i \in \mathbb{N}} w_i.
\]

Then

\[
\|\hat{a}_t - a(0)\|^2 = \|X_t^T W_t X_t)^{-1} X_t^T W_t E_t\|^2 \leq \|X_t^T W_t X_t)^{-1/2}\|^2 \cdot \|X_t^T W_t X_t)^{-1/2} X_t^T W_t E_t\|^2 \leq \lambda_{\min}(X_t^T W_t X_t)^{-1} E_t^T W_t^T X_t (X_t^T W_t X_t)^{-1} X_t^T W_t E_t \leq \frac{1}{\inf_{i \in \mathbb{N}} w_i} \lambda_{\min}(X_t^T X_t)^{-1} Q_t^w \ a.s.,
\]

where \( Q_t^w := E_t^T W_t^T X_t (X_t^T W_t X_t)^{-1} X_t^T W_t E_t \). We show \( E[Q_t^w] = O(\log \lambda_{\max}(X_t^T W_t X_t)) \). Since by Proposition 2 and Lemma 1, \( \lambda_{\min}(X_t^T W_t X_t)^{-1} \leq \lambda_{\min}(X_t^T X_t)^{-1} \cdot \frac{1}{\inf_{i \in \mathbb{N}} w_i} \leq \frac{1+p^2}{\inf_{i \in \mathbb{N}} w_i} \frac{1}{\epsilon^{1-\epsilon}} \ a.s., \) using (14) we then obtain \( E[\|\hat{a}_t - a(0)\|^2] = O(\frac{\log 1}{\epsilon^{1-\epsilon}}) \).

Write \( V_t = (X_t^T W_t X_t)^{-1} \) and \( x_t = (1, p_t) \). Since \( X_t^T W_t X_t = X_{t-1}^T W_{t-1} X_{t-1} + (w_t^{1/2} x_t)^T (w_t^{1/2} x_t) \), it follows from the Sherman-Morrison formula (Bartlett 1951) that \( V_t = V_{t-1} - w_{t-1} x_t^T V_{t-1} x_t \), and thus

\[
x_t^T V_t = x_t^T V_{t-1} - \frac{(w_{t-1} x_t^T V_{t-1} x_t)}{1 + w_{t-1} x_t^T V_{t-1} x_t} = x_t^T V_{t-1} / (1 + w_{t-1} x_t^T V_{t-1} x_t).
\]

Then \( Q_t^w \) equals

\[
Q_t^w = \sum_{i=1}^{t} x_i^T w_i \epsilon_i V_t (\sum_{i=1}^{t} x_i w_i \epsilon_i) = \sum_{i=1}^{t} x_i^T w_i \epsilon_i V_t (\sum_{i=1}^{t} x_i w_i \epsilon_i) + x_t V_t x_t \epsilon_t^2 + 2 x_t V_t (\sum_{i=1}^{t-1} x_i w_i \epsilon_i) w_t \epsilon_t = Q_{t-1}^w + (\sum_{i=1}^{t-1} x_i^T w_i \epsilon_i) (\frac{-w_{t-1} x_t^T V_{t-1}}{1 + w_{t-1} x_t^T V_{t-1} x_t}) (\sum_{i=1}^{t-1} x_i w_i \epsilon_i) \]

\[ + x_t^T V_t x_t w_t^2 \epsilon_t^2 + 2 \frac{x_t^T V_{t-1}}{1 + w_t x_t^T V_{t-1} x_t} \left( \sum_{i=1}^{t-1} x_i w_i \epsilon_i \right) w_t \]
\[ = Q_{t-1}^w - \frac{w_t (x_t^T V_{t-1} \sum_{i=1}^{t-1} x_i w_i \epsilon_i)^2}{1 + w_t x_t^T V_{t-1} x_t} + x_t^T V_t x_t w_t^2 \epsilon_t^2 + 2 \frac{x_t^T V_{t-1}}{1 + w_t x_t^T V_{t-1} x_t} \left( \sum_{i=1}^{t-1} x_i w_i \epsilon_i \right) w_t \epsilon_t. \]

By telescoping the sum we obtain
\[ Q_T^w = Q_{t_0}^w - \sum_{t=t_0}^{T} \frac{w_t (x_t^T V_{t-1} \sum_{i=1}^{t-1} x_i w_i \epsilon_i)^2}{1 + w_t x_t^T V_{t-1} x_t} + \sum_{t=t_0}^{T} x_t^T V_t x_t w_t^2 \epsilon_t^2 + 2 \sum_{t=t_0}^{T} \frac{x_t^T V_{t-1}}{1 + w_t x_t^T V_{t-1} x_t} \left( \sum_{i=1}^{t-1} x_i w_i \epsilon_i \right) w_t \epsilon_t, \quad (15) \]

where \( u_t := \frac{x_t^T V_{t-1}}{1 + w_t x_t^T V_{t-1} x_t} \left( \sum_{i=1}^{t-1} x_i w_i \epsilon_i \right) w_t \). Using \( E[u_t \epsilon_t] = E[u_t E[\epsilon_t \mid \epsilon_1, \ldots, \epsilon_{t-1}]] = 0 \) and \( E[x_t^T V_t x_t w_t^2] = E[x_t^T V_t x_t E[\epsilon_t^2 \mid \epsilon_1, \ldots, \epsilon_{t-1}]] \leq E[x_t^T V_t x_t w_t^2] \sup_{t \in N} E[\epsilon_t^2] \), we can derive from (15) that \( E[Q_T^w] = O \left( E \left[ \sum_{t=t_0}^{T} x_t^T V_t x_t w_t^2 \right] \right) \). By Sylvester’s determinant theorem \( \det(I + AB) = \det(I + BA) \) we have
\[ \det(X_{t-1}^T W_{t-1} X_{t-1}) = \det(X_t^T W_t X_t - (w_t^{1/2} x_t)^T (w_t^{1/2} x_t)) \]
\[ = \det(X_t^T W_t X_t) \det(I - (X_t^T W_t X_t)^{-1} (w_t^{1/2} x_t)^T (w_t^{1/2} x_t)) \]
\[ = \det(X_t^T W_t X_t) (1 - w_t x_t^T V_t x_t), \quad (16) \]

and thus \( w_t x_t^T V_t x_t = \frac{\det(X_t^T W_t X_t)}{\det(X_{t-1}^T W_{t-1} X_{t-1})} \). From the Claim, formulated and proven below, applied to the sequence \( (\det(X_t^T W_t X_t))_{t \geq t_0} \), we obtain
\[ \sum_{t=t_0}^{T} w_t x_t^T V_t x_t = \sum_{t=t_0}^{T} \frac{\det(X_t^T W_t X_t) - \det(X_{t-1}^T W_{t-1} X_{t-1})}{\det(X_{t-1}^T W_{t-1} X_{t-1})} \leq S \log \det(X_T^T W_T X_T) \]
for some \( S > 0 \) with \( E[S] < \infty \). This implies \( E[Q_T^w] = O(\log \lambda_{\max}(X_T^T W_T X_T)) \).

**Claim.** Let \( (x_n)_{n \in N} \) be a monotone increasing unbounded positive sequence and \( S_n = \frac{1}{\log x_n} \sum_{k=1}^{n} \frac{x_k - x_{k-1}}{x_k} \), where \( x_0 := 0 \). Let \( N = \inf \{ n \in N \mid S_n > 1, x_n > e \} \). If \( N = \infty \) then \( \sup_{n \geq N} S_n \leq 1 \); if \( N < \infty \) then \( \sup_{n \geq N} S_n \leq S_N \).

**Proof of claim.** The case \( N = \infty \) is trivial. Suppose \( N < \infty \), and assume by induction that \( S_{n-1} \leq S_N \) for some \( n - 1 \geq N \). Define \( g(x) = \frac{1}{\log x} \frac{x - x_{n-1}}{x} + \frac{\log x_{n-1}}{\log x} S_{n-1} \). Then \( 1 \leq S_{n-1} \) and \( x_{n-1} > e \) imply
\[ \frac{\partial g(x)}{\partial x} = \frac{1}{x(\log x)^2} \left[ -1 + \frac{x_{n-1}}{x}(1 + \log x) - \log(x_{n-1})S_{n-1} \right] < 0, \quad (x \geq x_{n-1}), \]
and thus \( S_n = g(x_n) \leq \max_{x \geq x_{n-1}} g(x) = g(x_{n-1}) = S_{n-1} \leq S_N \). This proves the claim.
6.5. 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

\[
r(p, a) = r(p_{opt}, a) + r'(p_{opt}, a)(p - p_{opt}) + \frac{1}{2} r''(\tilde{p}, a)(p - p_{opt})^2,
\]

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

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

(17)

On an open neighborhood of \( p_{opt} \) in \( \mathcal{P} \), \( p_{opt} = p(a^{(0)}) \) 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. Duijsetmaat and Kolk 2004) that there are open neighborhoods \( U \) of \( p_{opt} \) 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 \( V \rightarrow U, a \mapsto p(a) \), 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

\[
p(a) = p(a^{(0)}) + \frac{\partial p(a)}{\partial a^T} \bigg|_{\tilde{a}} (a - a^{(0)}),
\]

which implies that for all \( a \in V \)

\[
|p(a) - p(a^{(0)})| = O(||a - a^{(0)}||).
\]

(18)

Let \( \rho > 0 \) such that \( \{a : ||a - a^{(0)}|| \leq \rho\} \subseteq V \) and \( ||a - a^{(0)}|| \leq \rho \implies a_0 > 0, a_1 < 0 \), and let \( T_\rho \) be as in (8). Chang (1999) showed that \( ET_\rho < \infty \). This implies that for all \( t \in \mathbb{N} \),

\[
E[|p_t - p_{opt}|] = E[|p_t - p_{opt}| \cdot 1_{t > T_\rho}] + E[|p_t - p_{opt}| \cdot 1_{t \leq T_\rho}]
\]

\[
\leq E[|p_t - p_{opt}| \cdot 1_{t > T_\rho}] + (p_h - p_l)P(t \leq T_\rho)
\]

\[
\leq E[|p_t - p_{opt}| \cdot 1_{t > T_\rho}] + (p_h - p_l) \frac{ET_\rho}{t},
\]

(19)

where \( 1_A \) denotes the indicator function.

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 \( ||\tilde{a}_t - a^{(0)}|| \) is sufficiently small and \( t \) sufficiently large, \( \arg\max_{p \in [p_l, p_h] \cap TI(t)} r(p, \tilde{a}_t) \) lies on the boundary of the taboo interval \( TI(t) \). It follows that there is a \( \rho > 0 \) such that for all \( t > T_\rho \),

\[
|p_t - p(\tilde{a}_t)| \leq |TI(t)|
\]

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

$$E[|p_t - p_{opt}|] \leq E[|p_t - p_{opt}| \cdot 1_{t>T_\rho}] + (p_h - p_l) \frac{ET_\rho}{t},$$

$$\leq E[|p_t - p(\hat{a}_t)| \cdot 1_{t>T_\rho}] + E[|p(\hat{a}_t) - p(a^{(0)})| \cdot 1_{t>T_\rho}] + (p_h - p_l) \frac{ET_\rho}{t},$$

$$= O\left(E[|TI(t)|] + E[|\hat{a}_t - a^{(0)}| \cdot 1_{t>T_\rho}] + (p_h - p_l) \frac{ET_\rho}{t}\right),$$

$$= O\left(\sqrt{t^{\alpha-1} + \frac{\log t}{t^\alpha}}\right),$$

by Proposition 4, from which follows

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

$$= O\left(\sum_{t=1}^{T} t^{\alpha-1} + t^{-\alpha} \log t\right)$$

$$= O\left(T^\alpha + T^{1-\alpha} \log T\right).$$

**Acknowledgments**

We thank Anton Kleywegt for providing useful literature references, and for introducing us to consistency problems in certainty equivalent pricing. We furthermore thank Kacha Dzhaparidze for several useful discussions. The comments and suggestions of the area 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.

**References**


