Stochastic Adaptive Dynamics of a Simple Market as a Non-Stationary Multi-Armed Bandit Problem

Yann BRAOUEZEC

ESILV
Département d'ingénierie financière
92916 Paris La Défense Cedex
yann.braouezec@devinci.fr

ABSTRACT. *We develop a dynamic monopoly pricing model as a non-stationary multi-armed bandit problem. At each time, the monopolist chooses a price in a finite set and each customer decides stochastically but independently to visit or not his store. Each customer is characterized by two parameters, an ability-to-pay and a probability to visit. Our problem is non-stationary for the monopolist because each customer modifies his probability with experience. We define an ex-ante optimal price for our problem and then look at two different ways of learning this optimal price. In the first part, assuming the monopolist knows everything but the ability-to-pay, we suggest a simple counting rule based on purchase behavior which allows him to obtain enough information to compute the optimal price. In the second part, assuming no particular knowledge, we consider the case in which the monopolist uses an adaptive stochastic algorithm. When learning is easy (difficult), our simulations suggest that the monopolist (does not) choose the optimal price on each sample path.*

KEYWORDS: *Multi-armed bandit problem, adaptive learning, stochastic market dynamics, exploration-exploitation trade-off, non-stationarity*

1. Introduction[1]

Economic theorists have studied markets equilibrium problem as a mathematical problem: existence, unicity and stability of the solution. Stability gives rise to the so-called tatonnement models in which the adaptive dynamics of the price is supposed to reflect the learning process of the walrasian auctioneer. As noted in Kirman et *al.* (2001), this learning process does (unfortunately) not always lead to the equilibrium price, but the main issue is that this auctioneer does not exists in most markets! In reality, firms adjusts their price as a function of their knowledge and their experience (e.g., Lesourne, 1992): the market should thus be modeled as a self-organization process.

In recent years, self-organization (*i.e.*, learning and evolution) processes have been extensively studied in economic theory but mostly in a game-theoretic framework (e.g., Fudenberg-Levine, 1998, Weibull, 1996). On the contrary, in this paper, we shall consider a market model as a non-strategic dynamic decision model embedded in a multi-armed bandit problem.

In economic models of incomplete information related to the bandit problem (e.g. Aghion *et al.* 1991, Arthur, 1993, Easley and Kieffer, 1988, McLennan, 1979, Rothschild, 1974, Schmalensee, 1975), to each decision (*i.e.*, bandit) is associated a *stationary* probability distribution of the payoff but a parameter of this distribution is *unknown* to the decision-maker, *e.g.,* the mean and/or the variance. Two main ways of "solving" this multi-decisions problem have been conducted.

1) The "Bayesian-optimal" learning approach.

2) The "adaptive" learning approach.

In the first approach, the decision-maker forms a prior probability distribution over the unknown parameter, which is revised using Bayes rule, and his objective is to maximize a given criteria, typically an expected discounted sum of profits. Since learning is not *per se* the objective of the decision-maker, it may not be optimal for her to perfectly identify the true value of the unknown parameter. As a consequence, she may choose a sub-optimal decision in the long-run, as has been shown among others by Rothschild (1974). From a behavioral point of view, this Bayesian-optimal learning approach requires a high degree of rationality from the decision-maker since she is not only supposed to revise their beliefs using Bayes rule, but also to solve a very complex dynamic optimization problem (see *e.g.,* Easley and Kieffer, 1988).

In the second approach, much less rationality is expected from the decision-maker since she just follows a simple "rule of thumb", that is, she applies an adaptive algorithm (stochastic or not) to choose her decision. Interestingly enough, this adaptive

[1] I thank the participants of the "probabilités numériques" seminar of the Department of Probability of University of Paris 6 in November 2006, of Mash 2007, and of the seminar series of University of Southern Denmark, december 2008. Special thanks to Patrick Leoni. I also thank two anonymous referee for their comments. The usual disclaimer applies.

learning approach has been both studied in the fifties and sixties by mathematicians, psychologists and economists.

– By (applied) mathematicians, as a way to "solve" a two-armed (stationary) bandit problem. The aim was to design an adaptive algorithm (with finite memory) that may be deterministic, as in Robbins (1952) and Robbins (1956), or stochastic as in Samuels (1968), which maximizes the probability to choose (in the long-run) the best bandit (see Cover and Hellman, 1970 and Narendra and Thathachar, 1989 chapter 3 for an overview).

– By psychologists, as a way to model the observed behavior of human beings such as the "probability matching phenomenon". The experimental protocol was in general a simple two-armed bandit problem (or even a one-armed bandit) in which the subject is repeatedly put into a position where he must give one of two possible responses, and where there are two outcomes, e.g., "success" and "failure". The probability matching phenomenon is the fact that subjects match the underlying probability, *i.e.*, if p is the probability of reward in the bandit 1, then, after some periods, the subject chooses this bandit 1 with a frequency close to the probability[2] p.

– By economists, as a way to model (the so-called) bounded rationality, *i.e.*, when agents do not behave in the way predicted by utility theory (see Simon, 1959, p. 260-261).

The common point in these various literatures on learning (optimal-Bayesian or adaptive) is that the decision-maker has *no impact* on his environment because it is *stationary*. In the two-armed bandit theory of market pricing considered in Rothshlid (1974) or in Schmalensee (1975), the monopolist is learning over time while customers are assumed to behave in a stationary way. It seems nevertheless quite safe to assume that in a typical (retail) market, sellers (try to) learn their demand function in order to maximize their profit while customers (try to) learn the cheapest seller in order to minimize their expenditure. As a consequence, the market evolves in a non-stationary way. In this paper, we provide a *non-stationary multi-armed bandit theory* of monopoly pricing in which both side of the market (monopolist and customers) is learning over time, so that the evolution of the market is a *self-organization* process.

In our model, at each time, the monopolist chooses a price from a finite set and each customer decides stochastically, but independently, to visit or not the store of the monopolist. When a given customer visits the store at some time and when the price at this time is acceptable for her, she buys one unit of the good. We say in this case that she is "rewarded". When she visits the store but when the price is too high for her, she doesn't buy, she is thus "not rewarded". Assuming the well-known (psychological) *law of effect*, which asserts that the event "reward" increases the probability of the re-

[2] See e.g. Vulkan 2000 for an economist's perspective on probability matching. See also Cane (1962) and Restle (1959). From an "economic" point of view, this probability matching behavior is clearly sub-optimal. In 1954, in an interesting paper, Merril Flood suggested that this behavior might reflect the fact that subjects believe that the environment is non-stationary, although it is not. See also a related paper of Herbert Simon, (1956, p. 267-272).

warded response (see e.g., Staddon and Horner 1989), it implies that when a customer visits and buys, she increases her probability to visit. Conversely, when she visits and doesn't buy, she decreases her probability to visit. To use the terminology used in behavioral psychology, the event "visit and no-purchase" is *negatively reinforced* while the event "visit and purchase" is *positively reinforced.* Since the probability to visit the store of each customer is time-varying, the expected profit of the monopolist is also time-varying; he faces thus a *non-stationary* multi-armed bandit problem.

There is an important difference between stationary and non-stationary environment. In a stationary environment, the decision-maker can always *separate* the exploration phase from the exploitation one. As a consequence, using a simple stopping rule (optimal or not), he can choose the best bandit in the long-run with a probability arbitrarily close to one. Unfortunately, in a non-stationary environment, as we shall see, not only this separation is not possible but the notion of best bandit is undefined.

In the first part of this paper, we define an *ex-ante optimal decision* for our non-stationary multi-armed bandit problem and we provide a discussion of learning in stationary and non-stationary environments. In the second part, assuming that the monopolist knows everything but the ability-to-pay, we exhibit a "counting learning rule" based on purchase behavior which allows him to obtain enough information to compute the optimal price. In the third part, we relax the assumption that the monopolist knows everything but the ability-to-pay. We thus assign the monopolist a *stochastic reinforcement learning rule* suggested by Arthur (1993) in a multi-armed (stationary) bandit problem (also used by Posch (1997), Hopkins and Posch (2005), Beggs (2005)) among others in a game-theoretic framework) which doesn't require any particular knowledge except environmental feedback. In that context, as in Kirman *et al.* (2001) among others, the randomness of the market reflects the learning process of the various agents. At the beginning, decisions are taken stochastically but as time evolves, agents tend to choose a decision with probability one. Our numerical simulations reveal that in the long-run:

- Each customer visits the store with a probability zero or one.
- The monopolist may choose a sub-optimal price in the long-run.

The market can indeed locks-in to absorbing state which are not only sub-optimal from the monopolist point of view, but also from the customers point of view. Finally, based on numerical simulations result, we exhibit an "easy-learning" environment in which the monopolist choose always the monopoly price in the long-run.

2. The model

The (discrete time) dynamics of our model is the following one. At each time t, the monopolist chooses a price $P_j > 0$, net of production cost, in the finite following set:

$$\mathbf{P} = \{P_1, P_2, ...P_J\} \qquad [1]$$

$$P_j < P_{j+1} \ \forall j = 1, 2...J-1 \qquad [2]$$

and each buyer $i \in I \equiv \{1, 2...N\}$ decides *independently* whether or not to come to visit the store. The buyers who has decided to come will now buy one unit of the good if the price at time t is lower than (or equal) to her willingness to pay and won't buy in the opposite case. Let $n_t \leq N$ be the number of customers who bought one unit of the good at time t. The profit of the monopolist is thus equal to :

$$\Pi_{j,t} = P_j.n_t \qquad [3]$$

Note that since each price $P_j > 0$ is net of production cost, the profit is non-negative. At the end of time t, each buyer adapts her probability to visit as to whether or not she has bought the good.

2.1. *The decision rule of the buyer*

Each buyer $i \in I$ is characterized by the two following parameters:

1) an ability-to-pay $v_i \in \mathbb{R}^+$.

2) an adaptive probability $\theta_{i,t} \in [0, 1]$ to visit the monopolist.

We assume that the willingness-to-pay of each buyer is constant, which is equivalent to say that preferences of buyers stable over time. However, the probability to visit the store of the monopolist evolves over times. In their early book on stochastic models of buying behavior, Massy *et al.* (1970) suggested that the factors affecting purchase probabilities were:

– Experience, *i.e.*, feedback from past purchases.

– Influence of exogenous market forces.

– Population heterogeneity.

In what follows, we assume that a given customer tend to visit the store more frequently when she buys one unit of the good, and less frequently when she doesn't buy. The purchase probability is modified with personal experience only, and there is no interaction effect (*i.e.*, no strategic behavior, no imitation) and/or no collective learning possibility[3].

Let $P_t \in \mathbf{P}$ be the price chosen at time t by the monopolist, and let $z_{i,t} = 1$ ($z_{i,t} = 0$) the case in which customer i visits (does not visit) the store at time t. Formally, we assume that each customer i use the following stochastic reinforcement algorithm to modify her probability to visit $\theta_{i,t}$:

$$\theta_{i,t+1} = \begin{cases} \theta_{i,t}.(1-\beta_i) & \text{if} \quad (z_{i,t}=1, v_i < P_t) \\ \theta_{i,t} + (1-\theta_{i,t})\beta_i & \text{if} \quad (z_{i,t}=1, v_i \geq P_t) \\ \theta_{i,t} & \text{if} \quad (z_{i,t}=0) \end{cases} \qquad \forall i \in I \qquad [4]$$

where $\beta_i \in]0,1]$ is a parameter. This reinforcement algorithm is called a *Linear Reward Penalty* in the learning automaton literature (see e.g. Narendra and Thathachar, 1989) and is quite natural: when a customer buys one unit of the good at time t, she increases her probability to visit and decreases it when she doesn't buy. Given the learning rule [4], it may thus be the case that in the long-run, a customer will visit the store with probability one while another one will eventually stop to visit him because he has experienced too many high prices. Note importantly that:

– $\{0\}$ is an *absorbing state* of the stochastic process but not $\{1\}$.

– the size of the increment is a decreasing function of β_i. Given the relation [4], the increment of the stochastic process $\{\theta_{i,t}\}$ is:

$$\Delta\theta_{i,t+1} = \begin{cases} -\beta_i\theta_{i,t} & \text{if} \quad (z_{i,t}=1, v_i < P_t) \\ (1-\theta_{i,t})\beta_i & \text{if} \quad (z_{i,t}=1, v_i \geq P_t) \\ 0 & \text{if} \quad (z_{i,t}=0) \end{cases} \qquad \forall i \in I \qquad [5]$$

We can thus interpret β_i as a "reactivity parameter". The higher β_i, the more reactive is the customer since the magnitude of the increment is an increasing function of β_i.

2.2. *Monopoly price in pure strategy : the complete information case*

As in most of the literature on adaptive learning, we shall assume in this paper the monopolist doesn't discount the future. As we shall see later, this implies that the "exploration costs" are zero. To derive an optimal price (*i.e.*, the monopoly price) in our dynamic market model, we assume complete information; the monopolist knows the set of customers, the ability to pay v_i, the reinforcement learning rule [4], the learning parameter β_i and (if necessary), the initial condition $\theta_{i,0}$ for each i. Since

[3] One could imagine that one customer who have decided to visit the store a time t reveals the price at another customer who have decided not to visit at time t.

the probability to visit the store of each customer i is not constant over time, as we already said, the monopolist faces a *non-stationary multi-armed bandit problem*, which complicates the normative analysis since there is *a priori* no natural way to define an optimal price. We shall here focus on the optimal price in pure strategy. As in game theory, we call *pure strategy* the decision (*i.e.*, a price $P_j \in \mathbf{P}$) which is chosen by the monopolist with probability 1 at each time $t \in \mathbb{N}$. Let

$$I_j = \{i \in I : v_i \geq P_j\} \qquad [6]$$

be the set of customers such that their ability-to-pay is higher than this price P_j and let

$$I_j^c = \{i \in I : v_i < P_j\} \qquad [7]$$

be its complementary. Finally, let $\text{Card } I_j$ be the cardinal of I_j. Without loss of generality, we assume that $\text{Card } I_j \neq \emptyset$ for all $j = 1, 2...J$.

Proposition 1 Assume that $\theta_{i,0} > 0$ for all i. If $P_j \text{Card} I_j \neq P_k \text{Card} I_k$ for $j \neq k$, then, there exists a unique (*ex-ante*) optimal price in pure strategy.

Proof : consider the case in which the monopolist chooses the price P_j in pure strategy. Given the reinforcement learning rule [4] and provided that $\theta_{i,0} > 0$ for all $i \in I$, customers who belong to I_j will visit the store in the long-run with probability one, while customers who does not belong to I_j won't visit him in the long-run with probability one. This implies that :

$$\lim_{t \to \infty} \theta_{i,t} = \theta_{i,\infty} \begin{cases} 1 & \forall i \in I_j \\ 0 & \forall i \notin I_j \end{cases} \qquad [8]$$

The long-run expected profit is thus :

$$\begin{aligned} \lim_{t \to \infty} \mathbb{E}(\Pi_t(P_j)) &= \lim_{t \to \infty} \sum_{i \in I_j} P_j \, \theta_{i,t} = P_j \sum_{i \in I_j} \theta_{i,\infty} && [9] \\ &= P_j \text{ Card } I_j := \Pi_\infty(P_j) && [10] \end{aligned}$$

If all the $\Pi_\infty(P_j)$ are different, then, there exists a price $P_m \in \mathbf{P}$, the monopoly price, such that $\Pi_\infty(P_m) > \Pi_\infty(P_j)$ for all $j \neq m$ □

In the complete information case, the monopolist can compute the long-run (deterministic) profit $\Pi_\infty(P_j) = P_j \text{ Card } I_j$ associated to each price P_j since he knows $\text{Card } I_j$. As a consequence, he will charge this price P_m at each time $t \in \mathbb{N}$. Note importantly that what is really needed to find the (uniform) monopoly price is not the knowledge of the v_i, but only the knowledge of $\mathrm{Card}\ I_j$ for all j, that is, how many customers buy at price P_j if they were all visiting the store. Of course, if the monopolist were allowed to price discriminate, *i.e.*, to charge a different price to a different customer, then, perfect discrimination would require the perfect knowledge of the v_i. We can now compute the (usual) Marshallian total surplus in the long-run, $W_\infty(p_m)$, defined as follows:

$$W_\infty(P_m) = \sum_{i \in I_m} (v_i - P_m) + P_m \mathrm{Card}\ I_m \qquad [11]$$

Defining $W_\infty(P_m)$ is important since it defines a normative benchmark for the market welfare.

2.3. *Learning in stationary and non-stationary environment: exploration-exploitation trade-off*

We shall now assume that the monopolist doesn't have a complete information on his environment, in particular, the v_i are unknown but the number of customers and their reinforcement learning rule [4] may also be unknown. When $\theta_{i,t}$ is constant over time for each i, the environment of the monopolist is *stationary*, while when $\theta_{i,t}$ is time-varying, the environment is *non-stationary*.

2.3.1. *Learning in stationary environment*

Consider the stationary case, as for example in Schmalensee (1975) or Arthur (1993). If the monopolist knows that θ_i are constant over time for each i, since future is not discounted, his problem is not difficult because he can separate the *exploration* phase from the *exploitation* one by using the following *stopping rule*.

1) **Exploration**. Choose (consecutively) $n \in \mathbb{N}$ times (with n finite) the price P_j and then estimates the (unknown) expected profit by its empirical mean $\overline{\Pi}_{j,n}$. Repeat for $j = 1, 2..J$.

2) **Exploitation**. For $t \geq nJ$, choose with probability one the price associated with the highest empirical mean.

By the law of large number, he knows that if n is high enough, he will discover the monopoly price with a probability arbitrarily close to one. In a stationary environment, the monopolist can thus behave like a classical (or even Bayesian) statistician. In particular, there is no need to use some stochastic adaptive algorithm. It is important to realize that n can be arbitrarily large because exploration cost is zero since the future is not discounted by assumption. When the future is discounted, given a criteria to optimize, choosing n "high" may not be optimal. Rothschild (1974) considers explicitly the discounted case in a stationary two-armed bandit problem from a Bayesian learning point of view in which the objective of the decision-maker is to maximize the expected discounted sum of profits. He shows that there exists an optimal stopping rule which best solves the exploration-exploitation conflict, and that this optimal stopping rule may lead the decision-maker to choose infinitely often a suboptimal price. This result[4] has been called *incomplete learning in the long run* in the economic literature.

In a stationary environment, when the future is not discounted, it doesn't matter whether the decision-maker starts the exploration phase with the price P_1 or P_J provided that all prices are tried. When this is done, with probability arbitrarily close to one, the monopolist will choose (forever) the monopoly price after the exploration phase.

Schmalensee (1975) and more recently Arthur (1993) have considered a stationary multi-armed bandit problem in which the monopolist uses some stochastic reinforcement algorithm. They show that in general, the monopolist doesn't discover the monopoly price in the long-run: with positive probability, he will eventually choose infinitely often a sub-optimal price[5]. In a *stationary* framework, this kind of stochastic reinforcement algorithm is not very convincing because we can easily find the optimal price using the above simple statistical decision rule. As we shall see now, when the environment is *non-stationary*, such a stopping rule is not implementable, which means that stochastic reinforcement learning may find its roots in a non-stationary environment.

2.3.2. *Non-stationarity: "exogenous" versus "endogenous"*

If stationarity environment is clear, non-stationary environments are not. As we shall see, non-stationary environments may sometimes be indeed stationary! In a multi-armed bandit problem, the environment is stationary for the decision-maker if the probability distribution of a given bandit is time-independent. In a simple two-armed bandit problem, in which each bandit i is a Bernoulli distribution of parameter θ_i for $i = 1, 2$ (*i.e.*, the probability of "success"), stationarity implies that θ_i is time-independent. On the contrary, the environment is non-stationary if θ_i is time-dependent, *i.e.*, if θ_i may vary over time. But θ_i may vary over time in many different ways.

The "exogenous" form of non-stationarity is the case in which the parameter evolves from say time t to time $t + 1$ according to an "exogenous specific rule". In their book on "Learning Automata", Narendra and Thathachar[6] (1989) distinguish two forms of (exogenous) non-stationary environments:

[4] Easley and Kieffer (1988) considers the general mathematical version of this stochastic optimal control problem with Bayesian adaptation of beliefs, and show that, in a stationary multi-armed bandit problem, complete or incomplete learning depends on the discount factor. When the discount factor is very close to one, they show that complete learning is optimal.

[5] To be precise, Arthur (1993) gives a version of his algorithm in which he discovers the optimal price with probability one.

[6] See chapter 7 "Non stationary environments", p. 230.

1) Markovian switching environment.

2) State dependent non-stationary environments.

A Markovian switching environment may be the one in which $\theta_{i,t}$ follows a simple Markov chain, e.g., $\theta_{i,t} \in \{0.2, 0.6\}$ and $\mathbb{P}(\theta_{i,t+1} = 0.2|\theta_{i,t} = 0.2) = p_i$ and $\mathbb{P}(\theta_{i,t+1} = 0.6|\theta_{i,t} = 0.6) = q_i$. A state dependent non-stationary environment may the one in which $\theta_{i,t}$ evolves say as a deterministic function of time. For example, $\theta_{i,t+1} = f(\theta_{i,t})$ for some (fixed) function f such that $f(\theta_{i,t}) \in]0,1[$ for all t.

In these two forms of environments, non-stationarity is exogenous since the decision-maker does not have any impact on his environment. In fact, the environment is labeled non-stationary because the parameter $\theta_{i,t}$ varies over time, but the law of evolution of the parameter is stationary, *i.e.*, the transition matrix of the Markov chain and the function f are time-independent and indeed *exogenous*. One could even consider the transition matrix and the function f to be time-dependent as long as their law of evolution are stationary.

The “endogenous” form of non-stationarity is much more difficult and concerns the evolution of a “system” (*i.e.*, a game, a market, a social network...) which is the result of many *interdependent* decisions processes. Let us give two well-known examples.

– The El-Farol bar problem[7], in which say every sunday N (e.g., N =100) persons go a bar (El-Farol). Each person decides independently to visit or not the bar. The problem is that the bar is very small and each person does not enjoy to be there when it is crowed. If, say more than 70% of the population go to the bar, it is crowded, and each person is worse than if she stayed at home. On the contrary, if less than 70% of the population go to the bar, each person enjoys to be there, *i.e.*, no regret to come. Given a learning rule for each person to decide whether or not to come at week $n + 1$ based on the known[8] frequency of the population that went to the bar at week n, the evolution of this system is (endogenously) non-stationary because the decision process of each person depends on the percentage of people in the bar, which itself depends on the decision process of each person.

– The repeated “average game[9]”, in which N persons have to choose at each time a number between 0 and 100 and the winner is the one who chooses the closest number to the mean of all chosen numbers multiplied by a parameter p lower (or equal) to one[10]. At the end of each game, the winner number is revealed. Given a learning

[7] See e.g., Arthur (1994, p. 406-411)

[8] One may assume that if a person don’t come to the bar at week n, she learns say on monday whether the bar was crowded or not.

[9] See e.g., Nagel (1995, p. 313-1326).

[10] When $p < 1$, if we assume that both the rule of the game and the fact that all the players reason like the modeler are common knowledge, then, 0 is the unique Nash equilibrium in pure strategy that can be “eductively” (i.e., by a pure mental process) obtained by iterated dominance. Note that learning over time could also be a way to “converge” toward the Nash equilibrium.

rule for each player to choose some number for the game $n + 1$ based on the result of the game n, the evolution of this system is (endogenously) non-stationary because the decision process of each person depends on the winner number, which itself depends on the decision process of each person.

Assuming a fixed (adaptive) learning process for each agent, the modeler (who may know the various learning processes) still does not know the law that governs the evolution of the system. This is why we talk about endogenous non-stationarity. In general, for such a system, the evolution is (strongly) path-dependent and there are many (admissible) "stationary state". Such a system is generally called a *self-organization process*.

2.3.3. *Learning in non-stationary environment*

Consider now the case in which $\{\theta_{i,t}\}$ evolves over time as described in equation (4) and assume that the monopolist starts the exploration phase by charging n times (consecutively) the highest price P_J. If n is high enough, it may be the case that before the end of the exploration phase, customers who do not belong to I_J won't visit the store anymore. Since $\{0\}$ is an *absorbing state* of the stochastic process $\{\theta_{i,t}\}$, customers who belong to I_J will be the only remaining customers after the exploration phase. In that case, P_J becomes the "ex-post" optimal price but is not in general equal to the ex-ante optimal price. A better idea could be to start the exploration phase with the lowest price P_1, and then P_2 and so on, but this will obviously lead him to the same problem if all the price are tried.

There is thus a fundamental difference between the stationary and the non-stationary environment, namely the existence of an *irreversibity effect*[11], (or a *lock-in effect*) in the non-stationary environment. This irreversibility effect implies that it is not possible to separate the exploration phase from the exploitation one because the gathering information process affects the object of learning, or, in different terms, learning affects what is to be learned. This problem is very well known in game theory (see e.g. Fudenberg-Levine (1998)) but not, as far as we know, in a multi-armed bandit problem (but see Banks *et al.*, 1997). Interestingly enough, we discovered recently that a number of psychologists, neuroscientists and biologists investigate the exploration-exploration trade-off both in stationary and non-stationary environments.

In Daw *et al.* (2006) (see also Cohen *et al.*, 2007 or Groß *et al.*, 2008), they consider a non-stationary multi-armed bandit problem in which the mean payoffs change randomly and independently from trial to trial. They consider the following three learning processes.

– The ϵ-greedy rule, in which the decision-maker chooses with a probability $(1-\epsilon)$ the bandit (believed) to be the best and with a probability ϵ, another bandit.

[11] Note that this irreversibility effect depends of the decision rule of the customers, in which $\{0\}$ is an absorbing state but not $\{1\}$.

– The soft-max rule, in which one bandit is chosen stochastically on the basis of its relative performance or its estimated expected value.

– a "modified" soft-max rule in which a bandit that have not been chosen receive an "awarding bonus" that increases its probability of being chosen.

Daw *et al.*, (2006) found empirical (strong) evidence that subjects (in their experiment) use the soft-max rule.

3. Learning by counting: only the ability-to-pay of the customers are unknown

We shall assume in this section that only the ability to pay v_i are unknown. In particular, the reinforcement algorithm given by Equation [4] together with the parameter β_i, the initial condition $\theta_{i,0}$ for all i and the number of customers are assumed to be known by the monopolist. In this incomplete information setting, given a set of prices $\mathbf{P}$, how can the decision-maker find the monopoly price P_m if *only* the v_i are *unknown*?

A natural way to find the monopoly price is to *infer* (*i.e.*, learn) $\mathrm{Card}\ I_j$ from the purchase behavior of the customers. Note that this is in some sense equivalent to the (old) revealed preference problem in which one try to infer (unobservable) preferences from purchase behavior. For example, if a given customer i buys at a given price P_j but doesn't buy at the price P_{j+1}, the monopolist can infer that $i \in I_j \cap I_{j+1}^c$. Given the set of prices $\mathbf{P}$, knowing that $i \in I_j \cap I_{j+1}^c$ is equivalent to know v_i. As we said earlier, what is really needed here is not the perfect knowledge of the v_i but only $\mathrm{Card}\ I_j$.

Suppose the monopolist charges a given price P_j. If *all* the customers visit the store, he just have to count the number of customers who buy at that price to obtain $\mathrm{Card}\ I_j$. If he can do that for all $j = 1, 2...J$, he will be able to compute the associated profit $P_j\ \mathrm{Card}\ I_j$ for each j. The problem is however not so simple because the monopolist faces a non-stationary problem: if he charges a given price P_j consecutively too many times, it may be the case that customers who belong to I_j^c will disappear with probability one. We shall now propose a "counting rule" which allows the monopolist to infer $\mathrm{Card}\ I_j$ for all j when the learning parameter β_i is small enough for all i.

Counting rule

1) **Process for increasing loyalty**: charge consecutively n times (with n finite) the lowest price P_1.

2) **Counting process**: at time $n + j$, $j = 1, 2...J - 1$, charge the price P_{j+1} and count the number of customers who buy at that price.

When n is high enough, at the end of the process for increasing loyalty, all customers who belong to I_1 visit the store with a probability arbitrarily close to one. Now, during the counting process, (at least) all the customers of $I_1 \cap I_2^c$ don't buy the good when they visit the store. So, from time $t = n + 1$ to time $T = n + J - 1$, $\theta_{i,t}$ is a decreasing function of t for all $i \in I_1 \cap I_2^c$. However, in the particular case in which the β_i are small enough, customers who belong to $I_1 \cap I_2^c$ will still visit the store with a probability arbitrarily close to one at time $T = n + J - 1$. This is the essential idea of the proposition 2.

Proposition 2. Assume the monopolist uses the above counting rule. For simplicity, we assume that $\theta_{i,0} = \theta_0$ and $\beta_i = \beta$ for all $i \in I$. Given a set of prices $\mathbf{P}$, $\forall \epsilon > 0$ (with $\epsilon < 2(1 - \theta_0)$), if $\beta \leq 1 - \left[2\left(\frac{1-\epsilon}{2-\epsilon}\right)\right]^{\frac{1}{J-1}}$ and $n \geq \dfrac{\ln(\frac{\epsilon}{2(1-\theta_0)})}{\ln(1-\beta)}$, at time $T = n + J - 1$, then $\theta_{i,t} \geq 1 - \epsilon$, $\forall i \in I_1$, $\forall t \in \{n+1, ...n+J-1\}$; the counting process can be done exactly with a probability arbitrarily close to one.

Proof. See the appendix.

A different way to obtain the same result is to consider the case in which the parameter $\beta_{i,t}$ is a decreasing function of time t for all i. If we assume as in Lamberton-Pagès-Tarrès (2004) that:

$$\lim_{t \to \infty} \beta_{i,t} \to 0 \quad \text{with} \quad \sum_{t \geq 0} \beta_{i,t} \to \infty$$

we can also prove that there exists a finite n such that the counting process can be done exactly with a probability arbitrarily close to one.

We've seen previously that the size of the increment of the stochastic process $(\theta_{i,t})_{t \in \mathbb{N}}$ (see Equations [5]) is an increasing function of β_i. Assuming that β_i is very small is thus equivalent to say that the learning process is very slow. When the process for increasing loyalty is finished, all the probabilities to visit the store are arbitrarily close to one. Since all the β_i are very small, the monopolist can try all the price as if the various probabilities to visit were constant, this is why the monopolist can infer $\text{Card}\, I_j$ from purchase behavior.

When the various $\beta_i \in]0, 1[$, *i.e.*, not necessarily small, assuming $\theta_{i,0} > 0$ for all i, he may still use the following counting rule. Let an arbitrarily small $\epsilon > 0$.

1) Charge consecutively n times the lowest price P_1 such that $\theta_{i,n} > 1 - \epsilon$ for all i (note that n is a function of $\theta_{i,0}$ and β_i for all i and ϵ)

2) At time $n + 1$, charge the price P_J and count the number of customers who buy at that price, *i.e.*, obtain $\text{Card}\, I_J$

3) Charge now consecutively $n_J \geq 1$ times the lowest price P_1 such that the probability $\theta_{i,n+n_J+1} > 1 - \epsilon$ for all i

4) At time $n + n_J + 2$, charge the price P_{J-1} and count the number of customers who buy at that price, *i.e.*, obtain $\text{Card}\, I_{J-1}$

5) Charge now $n_{J-1} \geq 1$ times the lowest price P_1 such that the probability $\theta_{i,n+n_J+2+n_{J-1}} > 1 - \epsilon$ for all i

6) Repeat until $\mathrm{Card}\ I_j$ for all j.

Of course, the possibility to learn $\mathrm{Card}\ I_j$ from purchase behavior crucially depends on the perfect knowledge on the ability-to-pay v_i *and* the learning rule of the customers. In practice, it is not clear why, for example, the ability-to-pay should be known. When the monopolist knows nothing about the customers' characteristics, since he observes the profit associated the price he charges, he can still use a version of a soft-max rule to (try to) learn the monopoly price. This is what we shall now investigate.

4. Learning by reinforcement: the monopolist may be "totally ignorant"

4.1. *The reinforcement rule*

In a stationary multi-armed bandit problem, we saw that the decision-maker can first estimate the unknown parameters (typically the mean) and then choose the price with the highest estimated mean. Using this stopping rule will allow him to discover the optimal price with a probability arbitrarily close to one. However, in the non-stationary case, such a stopping rule does not work, as we have seen previously, because the decision-maker cannot *separate* the exploration phase from the exploitation one. In such an environment, one way to "solve" the problem may be achieved using a soft-max rule, *i.e.*, a stochastic decision rule. Although there are many candidates, we have chosen the stochastic adaptive algorithm suggested by Arthur (1993).

Let $S_{j,t-1}$ be the *strength* assigned to the price P_j at time $t-1$ and suppose P_j is the chosen price at time t and that n_t customers bought one unit of the good at time t. Let $\Pi_{j,t} = P_j.n_t$ be the profit of the monopolist at time t. The (simple) adaptive rule is the following one :

$$S_{j,t} = S_{j,t-1} + \Pi_{j,t} \qquad [12]$$

$$S_{k,t} = S_{k,t-1} \qquad \forall k \neq j \qquad [13]$$

The strength of each price reflects its past performance, *i.e.*, its cumulated profit. Since our monopolist can't separate the exploration phase from the exploitation one, it is quite natural to assume that he uses an *adaptive randomized rule* which allows him to both "explores and exploits". Let $\eta_{j,t}$ to be the *normalized* strength at each time t:

$$\eta_{j,t} = \frac{S_{j,t}}{\sum_k S_{k,t}} \qquad \forall j = 1, 2...J \qquad [14]$$

Since $\eta_{j,t}$ is a normalized weight, we now assume that the monopolist draws randomly a given price P_j at time t with a probability equal to $\eta_{j,t}$. Since $\eta_{j,t}$ evolves over time according to the performance of price P_j, this stochastic reinforcement algorithm may be a way to "solve" the exploitation-exploration trade-off.

In our model, $\Pi_{j,t} \geq 0$ implies that $\eta_{j,t} \geq 0$. If $\Pi_{j,t}$ could be negative, $\eta_{j,t}$ could also be negative, and our reinforcement algorithm wouldn't work. However, when $\Pi_{j,t}$ can be negative, one can still use, as Kirman *et al.*, (2001) among others, the following "logit" version of our stochastic reinforcement rule:

$$\eta_{j,t} = \frac{e^{\lambda S_{j,t}}}{\sum_{k=1}^{N} e^{\lambda S_{k,t}}} \qquad \forall j = 1, 2...J \tag{15}$$

where λ is a parameter. Of course, this reinforcement algorithm depends on the parameter λ while the one we use is parameter-free.

Remark[12]. The reinforcement learning rule of the customers (given by Equation [4]) and the reinforcement learning rule of monopolist (given by Equation [12] to [14]) are different. The learning rule of the monopolist is proportional to the performance of a decision, *i.e.*, the *profit*, while the learning rule of the customers is not proportional to the performance of the decision to visit, *i.e.*, to the *surplus*. Let $S_{i,1,t}$ be the strength associated to the decision " visit" of customer i ($S_{i,0,t}$ is the strength associated to the decision "no visit"). Assume now that for all i:

$$S_{i,1,t} = S_{i,1,,t-1} + (v_i - P_t) \tag{16}$$

$$S_{i,0,t} = S_{i,0,t-1} \tag{17}$$

where $P_t \in \mathbf{P}$ and let

$$\theta_{i,t} = \frac{e^{\beta_i S_{i,1,t}}}{e^{\beta_i S_{i,1,t}} + e^{\beta_i S_{i,0,t}}} \qquad \forall i = 1, 2...N \tag{18}$$

be the probability to visit the store at time t of the customer i, where β_i is a parameter. When $(v_i - P_t) > 0$, *i.e.*, when the customer i visits and buys, her probability to visit increases while when $(v_i - P_t) < 0$, its probability to visit decreases. In such a model, the decision rule of the customers and of the monopolist are both proportional to the "performance".

[12] This remark has been suggested by an anonymous referee.

4.2. ***Related technical literature***

Kirman *et al.* (2001) offer a market model in which all agents are learning over time. They study the asymptotic behavior of their stochastic process using a mean field approximation method. Although this is not strictly speaking a probabilistic proof, this gives an interesting simple computational method to study the long-run behavior of the stochastic process. Related to the bandit problem, Lamberton *et al.*, (2004) provide rigorous probabilistic proofs of the asymptotic behavior of a 2-armed bandit problem when the decision-maker uses a linear-reward-inaction stochastic algorithm but in a stationary environment. Arthur (1993) study a stationary multi-armed bandit problem using the linear-reward-inaction algorithm (14) in which the strengths are renormalized using the renormalization sequence $C_t = C\ t^v$ (see also Laslier-Topol-Walliser 2001). Arthur (1993) shows that when $0 < v < 1$, the stochastic process converges toward to a non-optimal vertex of the simplex with positive probability while when $v = 1$, it converges to the optimal vertex with probability one. In the proofs, Arthur (1993) relies on a very general theorem from Pemantle (1990). Hopkins and Posch (2005) study the algorithm used by Arthur (1993) but in a game theoretic framework and provide a discussion on the applicability of the Pemantle (1990) theorem (See also Beggs (2005) and Pementle (2007)). They suggest that Arthur (1993) theorems may be correct although the proofs are not since he relies on Pemantle (1990) theorem while the conditions are not met. We won't discuss further these technicalities since we provide only numerical results in this section.

4.3. ***Numerical results***

4.3.1. *Overview*

We shall study the market behavior from a numerical simulation point of view. Let $\eta_t = (\eta_{1,t}, ...\eta_{J,t}) \in \Delta(\mathbb{R}^J)$, where $\Delta(\mathbb{R}^J)$ is the unit-simplex of $\mathbb{R}^J$ and $\theta_t = (\theta_{1,t}, ...\theta_{N,t}) \in [0,1]^N$. Let

$$(\eta_t, \theta_t) \in \Delta(\mathbb{R}^J) \times [0,1]^N \qquad [19]$$

where (η_t, θ_t) is the state vector of our stochastic process and $\Delta(\mathbb{R}^J) \times [0,1]^N$ is the state space. On each sample path, our numerical simulations reveal that the *long-run behavior* of the stochastic process have the two following properties:

– all the customers either come with probability zero or one;

– the monopolist chooses a price with probability one.

This means that in the long-run, the evolution of the market becomes deterministic: the monopolist chooses a price with probability one, and all the customers come either with probability one or zero. To understand the intuition behind this numerical observation, note that the *magnitude of the increment*

– $\Delta\theta_{i,t}$ is stationary.

– $\Delta\eta_t$ is not stationary. The reason is simple: given the adaptive rule (12), since $\sum_j S_{j,t}$ is an increasing function of time t, the higher is $\sum_j S_{j,t}$ for a given profit Π_j, the lower are the increments $\Delta\eta_t$. As a consequence, the increments of the process η_t are a decreasing function of time t.

From a behavioral point of view, this means that the "reactivity" of the customers is constant over time while the "reactivity" of the monopolist is a decreasing function of time t. While the monopolist still chooses each price with positive probability in the intermediate run, customers tend to separate in two groups:

– those who visit the store with probability one and buy the good[13];

– those who disappeared.

Recalling that $\{0\}$ is an absorbing state of the stochastic process $(\theta_{i,t})_{t\in\mathbb{N}}$ but not $\{1\}$, it may be the case, as we said, that a customer comes (and buys) with a probability equal to one in the intermediate run, but eventually disappears in the long-run. This feature comes from the *path-dependency property* of the market.

We ran the numerical simulations with the following idea in mind. When the profit associated to the monopoly price is *not* sufficiently higher than all the others, it may be difficult for the decision-maker to discover the price P_m in the long-run on each sample path. We thus study two cases, one in which learning is "difficult", and one in which learning is "easy". We have chosen in both cases the following parameters :

– $J = N = 5$

– $P_i = v_i,$ with $v_i \in \mathbb{R}^+$ but is finite for all *i*

– $\beta_i = 0.2$

– $\theta_{i,0} = 0.5$ for all *i*

– $S_{j,0} = 1 \Longleftrightarrow \eta_{j,0} = 0.2$ for all j.

4.3.2. *Case in which learning is "difficult"*

As we said previously, learning is "difficult" when the profits associated to each price are very close to each others. A natural candidate for difficult learning is the case in which the various prices are between 0 and 1, that is:

$$P_i \in [0,1] \quad \text{for all } i = 1...5 \qquad [20]$$

Consider the following set of prices.

$$\mathbf{P} = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4 = 0.53; P_5 = 0.75\} \qquad [21]$$

[13] Note that it may be the case that in the intermediate run, a customer visits the store with probability one but disappear in the long-run

It is thus easy to check that the monopoly price is :

$$P_m = P_2 = 0.35 \qquad [22]$$

Depending on the history, we observe that the monopolist chooses in the long-run a price in $\mathbf{P}$ except $P_1 = 0.12$. For a non-negligible fraction of sample paths, the monopolist chooses the price P_5 in the long-run and as a consequence, customer 5 is the only remaining one. The reason lies in the *self-reinforcing* property of the market dynamics: the higher is the choice probability of the price P_5, *ceteris paribus*, the lower will be the probability to visit of customers 1 to 4. But the lower is the probability to visit of customers 1 to 4, the higher will be the choice probability of the price P_5 since P_5 becomes the optimal price. Note that when the monopolist chooses this price P_5 in the long-run, this stationary state is highly sub-optimal from the customers point of view, but also from the monopolist point of view. The Marshallian total surplus associated to the monopoly price is $W(P_m) = 0.75 + 1.4 = 2.15$ while it is $W(P_5) = 0.85$ in the stationary state. The loss of welfare is thus considerable!

This negative result lies in the fact that learning is "difficult" because the profit associated to each price are very close to each others. As a consequence, the adaptive algorithm used by the monopolist is unable to discriminate the optimal price from the others[14]. If the profit associated to the monopoly price was much higher than the others, the adaptive algorithm would (perhaps) discover it in the long run. Let us now consider this case.

4.3.3. *Case in which learning is "easy"*

Consider the following set of prices.

$$\mathbf{P}' = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4 = 0.53; P_5 = 10\} \qquad [23]$$

It is easy to check that P_5 is the monopoly price. In such a case, as soon as P_5 is chosen, the dynamics becomes *self-reinforcing* and the monopolist chooses quite quickly P_5 with a probability close to one. To see this more clearly, assume that P_5 is chosen at date $t = 0$ and that customer 5 comes to visit the store at this time[15]. Since only customer 5 purchases the good, the profit of the monopolist is equal to $P_5 = 10$. Using the adaptive rule (12) and given the initial condition $S_{j,0} = 1$ for all j, the monopolist will choose the price P_5 at time $t = 1$ with a probability equal to $(11/15) \approx 0.73$: the dynamics becomes then *self-reinforcing*. As a consequence, on each sample path, we observe that the monopolist chooses the (optimal) price $P_5 = 10$ in the long-run. Of course, this is the easy case since the profit realized with this price is much higher than all the other one.

[14] This observation remains valid whatever the parameter β_i is.

[15] Note that $\theta_{5,t}$ is a non-decreasing function of time t.

Consider now the set of prices **P'** but in which $P_5 \in]0.53; 10[$. Clearly, when P_5 decreases from 10 to 0.53, learning becomes more and more difficult since, *ceteris paribus*, the profit associated to this highest price P_5 becomes lower. When $P_5 = 0.75$, as we have seen previously, the monopolist is not always able to discriminate the optimal price $P_5 = 0.75$ from the others. On the contrary, when $P_5 = 10$, the monopolist is always able to discriminate the optimal price from the others. It thus natural to ask if there exists a "critical" price P^* such that:

– if $P_5 \geq P^*$, the monopoly price is always chosen in the long-run.

– if $P_5 < P^*$, the monopoly price is not always chosen in the long-run.

We have tested numerically many prices and we find that this critical price is (around) $P^* = 1.9$, which means that when we consider the set of price $\mathbf{P}' = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4 = 0.53; P_5\}$, as long as $P_5 > 1.9$, the monopolist always discover the monopoly price in the long-run.

It is now natural to ask whether or not this critical price $P^* = 1.9$ is robust to a perturbation of the environment. To this end, many perturbations can be done. For example, on can change say the price P_4, or the number of customers, or the number of prices, say, by adding another price P'. As we shall see, this critical price is in general not robust to such a perturbation. Note importantly that in all the following cases, P_5 is the optimal price.

1) Consider now the set of prices

$$\mathbf{P} = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4 = 0.53; P_5 = 2\} \quad [24]$$

in which, except for P_4, only one customer is such that $v_i = P_i$. For the price P_4, there are now 3 customers such that their ability-to-pay is 0.53. Numerical simulations reveal that the monopolist is not able to discover the monopoly price in the long-run: P_4 or P_5 may be chosen in the long-run. As a consequence, 1.9 is not anymore the critical price.

2) Let $P_i = v_i$ for $i = 1...5$ as in the previous case. However, we now the following set of prices

$$\mathbf{P}' = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4 = 0.53; P' = 1.8; P_5 = 2;\} \quad [25]$$

In this example, prices P' and P_5 provide (high) profit compare to the other prices when customer 5 visits the store. Numerical simulations reveal that both $P' = 1.8$ and $P_5 = 2$ may be chosen by the monopolist in the long-run. As a consequence, 1.9 is not anymore the critical price.

3) Consider the following set of prices with $P_i = v_i$ for all i.

$$\mathbf{P}'' = \{P_1 = 0.12; P_2 = 0.35; P_3 = 0.42; P_4; P_5 = 2\} \quad [26]$$

When $P_4 = 0.53$, we are back to the previous environment and we know that the monopolist will discover the optimal price. However, when $P_4 = 0.9$, numerical simulations reveal that the monopolist may eventually chooses the price $P_4 = 0.9$ in the long-run. As a consequence, 1.9 is not anymore the critical price.

Fundamentally, this exercise of robustness reveals that the stochastic algorithm used by the monopolist is efficient only when the "performance" associated to the optimal price is much more higher than all the others. When this is not the case, sub-optimal price(s) may be chosen in the long-run. In a stationary setting, one way to obtain the convergence toward the best price, as in Arthur (1993), may be to decrease the speed of the learning process via a renormalization sequence. Unfortunately, in a non-stationary setting, this does not work because the learning process affects, in a path dependent way, what is to be learned. As Cohen *et al.,* (2007) state in their paper:

> "To date, no universally optimal algorithm has been described that prescribes how to trade-off between exploration and exploitation in non-stationary environments, and it is not clear that doing so is possible." Cohen *et al.,* (2007) p 935.

We ran many simulations of this market model with different set of prices and the "qualitative" behavior of our market invariably remains the same: in the long-run, each customer comes either with a probability 0 or 1 and the monopolist chooses a price with probability one.

5. Conclusion

We believe that our "simple" framework is interesting since it highlights the use of adaptive stochastic algorithm in a non-stationary environment, *i.e.*, in which it is not possible to separate the exploitation phase from the exploration one. When everything but the ability-to-pay is known by the monopolist, we've shown that the monopolist can infer from purchase behavior, by using a counting rule, all the information needed to deduce the optimal price. When he is "totally ignorant" about the customers' characteristics, a natural way to learn the optimal price may be achieved by using an adaptive stochastic algorithm in which the probability to choose a given price is proportional to its (past) performance. When the performance of the different prices are "close" enough, the monopolist may eventually choose a sub-optimal price in the long-run since the algorithm is not able to discriminate the optimal price from the others. In the particular case in which the performance of the optimal price is much higher than all the others, then, the monopolist will choose this price in the long-run. Our results are indeed numerical and not analytical. It thus remains to prove the "convergence" of our market stochastic process but also to design an adaptive stochastic algorithm which would allow the monopolist to choose the optimal price in the long-run with probability one. As far as we know, we don't know any general theorems that could be applied to prove the "convergence" of our market model since models involving adaptive stochastic algorithms are mathematically difficult (see e.g., Pemantle, 2007) and critically depend on the stochastic algorithm which is considered (see e.g., Lamberton *et al.,* 2004).

6. References

Aghion, P. Bolton, P. Harris, C. and Jullien B., “Optimal Learning by Experimentation”, *Review of Economic Studies*, 621-654, 1991.

Arthur B., On designing Economic Agents that Behave like Human Agents, *Journal of Evolutionary Economics*, vol 3, pp1-23, 1993.

Arthur B, “Inductive Reasoning and Bounded Rationality”, *American Economic Review*, Papers and Proceedings), 84, pp. 406-411, 1994.

Banks, J, Olson M, Porter, D., "An experimental Analysis of the Bandit Problem ", *Economic Theory*, Vol. 10 Issue 1, pp. 55-77, 1997.

Beggs A “On the convergence of reinforcement learning”, *Journal of Economic Theory*, vol. 122(1), 1-36, May 2005.

Cane V , “Learning and Inference”, *Journal of the Royal Statistical Society*, pp. 183-200, 1962.

Cohen J, McClure S, Yu A, “Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration”, *Philosophical Transactions of the Royal Society B*, 362, pp. 933-942, 2007).

Cover T, Hellman, “The Two-Armed Bandit Problem with Time Invariant Finite Memory”, *IEEE Transactions on Information Theory*, pp 185-195, 1970.

Easley D, Kieffer N, “Controlling a Stochastic Process with Unknown Parameters”, *Econometrica*, pp 1045-1064, 1988.

Daw N, O’Doherty J, Dayan P, Seymour B, Dolan R, “Cortical substrates for exploratory decisions in humans”, *Nature*, Vol 44, pp. 876-879, 2006.

Flood M, "Environmental non-stationarity in a sequential decision-making experiment", in *Decisions processes*, by C. H. Coombs, R. L. Davis, Robert McDowell Thrall, Wiley, (1954).

Fudenberg D., Levine D., *The Theory of Learning in Games*, MIT Press, 1998.

Hopkins E, “Adaptive Learning Models of Consumer Behavior”, working paper, 2006.

Hopkins E, Posch M, “Attainability of Boundary Points under Reinforcement Learning” *Games and Economic Behavior*, 53, 110-125, 2005.

Groß R, Houston A, Collins E, McNamara J, Dechaume-Moncharmont F.X, Francks N, “Simple learning rules to cope with changing environments”, *Journal of the Royal Society Interface*, Oct 6;5(27), pp. 1193-1202, 2008.

Kirman A., Weisbuch G., Market Organization and Trading Relationships, *Economic Journal*, 411-436, 2001.

Lamberton D., Pagès G., Tarrès P., When can the two-armed bandit algorithm be trusted ?, *Annals of Applied Probability*, 1424-1454, 2004.

Laslier, J. F., Topol, R., Walliser, B., “A Behavioral Learning Process in Games”, *Games and Economic Behavior*, 37, 340-366, 2001.

Lesourne J., “*The Economics of Order and Disorder*”. Oxford University Press, 1992.

Massy W., Montgomery D., Morisson D., *Stochastic Models of Buying Behavior*, MIT Press, 1970.

McLennan, A., "Price dispersion and incomplete learning in the long-run", *Journal of Economic Dynamics and Control*, 7, 331-347, 1984.

Nagel R.M, “Unraveling in guessing games : an experimental study”, *American Economic Review*, pp. 1313-1326, 1995.

Narendra K., Thathachar M, *Learning Automaton*, Prentice-Hall, 1989.

Pemantle R., “Non convergence to unstable points in urn models and stochastic approximations”, *Annals of Probability*, 698-712, 1990.

Pemantle R., “A survey of random processes with reinforcement”, *Probability Survey*, Vol 4, pp 1-79, 2007.

Posch, M., “Cycling in a Stochastic Learning Algorithm for Normal Form Games”, *Journal of Evolutionary Economics*, pp. 193-207, 1997.

Restle F, “A Survey and classification of learning models”, in *Studies in mathematical learning theory*, Robert R. Bush and William K. Estes, Editors. Stanford, California: Stanford University Press, 1959.

Robbins H., “Some Aspect of the Sequential Design of Experiments”, *Bulletin of the American Mathematical Society*, 527-535, 1952.

Robbins H., “A Sequential Decision Problem with Finite Memory”, *Proceeding of the National Academy of Science*, pp 920-923,1956.

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

Samuels B (1968), “Randomized Rules for the Two-Armed Bandit with Finite Memory”, *Annals of Mathematical Statistics*, pp 2103-2107.

Schmalensee R., Alternative Models of Bandit Selection ", *Journal of Economic Theory*, 333-342, 1975.

Herbert Simon, “A comparison of game theory and learning theory”, *Psychometrica*, pp. 267-272, 1956.

Simon H., “ Theories of decision-making in economics and behavioral sciences”, *American Economic Review*, Vol 49, pp 253-283, 1959.

Staddon J, Horner J.M.,” Stochastic choice models: a comparison between Bush-Mosteller and a source-independent reward-following model”, *Journal of the Experimental Analysis of Behavior*, vol 52, pp. 57-64, 1989.

Vulkan N, “An Economist’s perspective on probability matching”, *Journal of Economic Survey*, Vol 14, pp. 101-118, 2000.

Weibull J., *Evolutionary Game Theory*, MIT Press, 1995.

7. Appendix

Proof of proposition 2 : assume the monopolist charges the lowest price P_1 during n times consecutively. Fix $\epsilon > 0$ and recall that $\theta_{i,0} = \theta_0$ and $\beta_i = \beta$ for all $i \in I$. Clearly, customers $i \in I_1$ buy when they visit the store. As a consequence, by Equation (4), $\forall i \in I_1$, $\theta_{i,t+1} = \theta_t + (1-\theta_t)\beta = \theta_t(1-\beta) + \beta$ for $t \leq n$. Since this is a first order linear difference equation, this implies that at time $t = n$, $\theta_{i,n} \equiv \theta_n = 1 - (1-\beta)^n \, (1-\theta_0) \; \forall i \in I_1$. We now look for $n \in \mathbb{N}$ such that $\theta_n = 1 - (1-\beta)^n \, (1-\theta_0) \geq 1 - \frac{\epsilon}{2}$, which gives $n \geq \dfrac{\ln(\frac{\epsilon}{2(1-\theta_0)})}{\ln(1-\beta)}$.

During time $t = n+1, n+2 ... n+J-1$, the price $P_2, P_3, ... P_J$ are charged. Clearly, some customers will come, won't buy and will thus decrease the probability to visit. Consider customers who belong to $I_1 \cap I_2^c$. During time $n+1$ until $n+J-1$, their probability to visit will monotonically decrease. At time $n+J-1$, we thus have that $\theta_{i,n+J-1} = (1-\beta)^{J-1}\theta_{i,n}$ where $\theta_{i,n} \geq 1 - \frac{\epsilon}{2}$. We now look on β such that $\displaystyle\sum_{t=n+1}^{n+J-1} |\Delta\theta_{i,t}| \leq \frac{\epsilon}{2}$, which is equivalent to $(1-\beta)^{J-1}(1-\frac{\epsilon}{2}) \geq 1-\epsilon$. The condition on β is thus $\beta \leq 1 - \left[2\left(\frac{1-\epsilon}{2-\epsilon}\right)\right]^{\frac{1}{J-1}}$. Let $T = n+J-1$. Since $\theta_{i,T} < \theta_{k,T} \; \forall \, i \in I_1 \cap I_2^c$, and $\forall \, k \in I_j \; j \geq 2$, we thus have proved that $\theta_{i,T} \geq 1-\epsilon$ for all $i \in I_1$. Since ϵ is arbitrarily small, all customers visit the store at each date $t = n+1, n+2 ... n+J-1$ with probability arbitrarily close to one. The monopolist can thus infer $\mathrm{Card}\, I_j$ for all j from purchase behavior □