Nash versus Reinforcement Learning on a Search Market: some Similarities and Differences between Individual and Social learning

Eric Darmon* — Roger Waldeck**

** CREM – CNRS and University of Rennes 1*
7, place Hoche, F-35065 Rennes Cedex
eric.darmon@univ-rennes1.fr

*** ENST-Bretagne, Département LUSSI*
Technopôle Brest Iroise CS 83818, F-29238 Brest Cedex 3
and ICI -EA2652- Université de Bretagne Occidentale
roger.waldeck@enst-bretagne.fr

ABSTRACT. *In this paper, we consider a simple search market extended from H. Varian's (Amer. Econ. Rev., 1980) classical model and ask whether less informed sellers are able to learn such sophisticated price strategies in this framework. We therefore compare the choices made by adaptive sellers (reinforcement learning) to those made by Nash sellers. We confront the results of two types of learning models: individual learning (where sellers only observe their own price performance) and social learning (where sellers observe the pricing experiments of other sellers). In the case of individual learning, we show that although sellers are not able to learn the Nash price distribution, they are able to qualitatively mimic Nash predictions when buyers' search information varies. In the case of social learning, first results suggest that the process is highly path dependent. Again, the choices made by adaptive sellers do not converge to the Nash equilibrium. In addition, some (but not all) qualitative properties are no longer preserved.*

KEYWORDS: *imperfect information, Nash equilibrium, mixed strategies, reinforcement learning, individual and social learning*

1. Introduction

Starting from (Stigler, 1969), an abundant literature has analyzed the consequences of imperfect and costly information on the structure and performance of decentralized markets. Recently, this approach has been applied to the debate on electronic markets (see *e.g.* (Brynjolfsson *et al.*, 2000)) by questioning whether increasing information transparency in markets would necessarily improve market competition. More generally, two types of consumers ("informed" and "uninformed") are usually distinguished according to the magnitude of their search costs. To compute the Nash equilibrium for the market game, one needs to assume common knowledge about the structure of the game and in particular about buyers' search characteristics. From this knowledge, sellers can deduce their optimal pricing strategies. Assuming in addition common knowledge of rationality, they may deduce a Nash equilibrium. However, without the two previous assumptions, one should express worries about how a Nash equilibrium distribution can be computed or learned.[1]

This paper tackles this issue and asks whether adaptive sellers learn to play a mixed Nash equilibrium. In this paper, we deliberately focus on sellers' learning. That is, we consider that buyers' behaviors are ruled by a simple and exogenously given rule which is fixed sample size search (cf. (Varian, 1980))[2]. In this context, the Nash equilibrium in mixed strategies is compared to the outcomes derived from Reinforcement Learning. We consider two types of learning: Individual Learning and Social Learning. Concerning Individual Learning, we briefly recall some results obtained under Individual Learning (Waldeck *et al.*, 2006).[3] This analysis reveals that sellers do not learn to play the Nash equilibrium in a strict sense. Rather, the key point is that when buyers' search characteristics vary, the qualitative predictions of Nash equilibrium are remarkably preserved. This shows that two types of rather extreme rationality models may lead to the same predictions and that the Nash equilibrium concept, which is one of the core concept of economic theory in the last decades, still performs well to depict choices made by less rational sellers.

Given these results, we consider in greater details the case with Social Learning. An extensive theoretical literature has considered the choices made by agents under social learning or social influences. In most of these models with pure informational externalities, agents need to make a choice based on the value of a state of nature. Hence, learning is about an unknown state of an event. Agents get a private signal about this state and try to infer from other agents' choices/actions (public information) the true state of the nature in order to make an optimal choice (cf. *e.g.*

---

[1] One should note that instrumental rationality can also be soften by considering Quantal Response Equilibria.

[2] In a richer framework, buyers' search behavior should be endogenous and limited rationality should also prevail on the buyer side (Kirman *et al.*, 2001). But since the focus of the paper is the comparison between Nash and a learning model in a search market game, we take one of the earliest model treated in the literature (Varian, 1980) as a starting point.

[3] We develop here two specific cases and recall some earlier results in order to compare them to those obtained with social learning.

the seminal paper of (Bikhchandani et *al.,* 1992)). This literature examines under what conditions informational cascades may appear when agents are learning from each other (Gale, 1996). In pure models of informational externalities, agents influence each other when forming their beliefs but the actions taken by some agents do not directly influence the payoff of others. However, in our game theoretical framework, the actions taken by some agents directly affect the payoffs of other agents. In addition if agents can observe the price set by other players and the resulting profits, they could learn something from it. For these reasons, we considered an implementation of social learning closer to (Vriend, 2000). Using a Genetic Algorithm to compare Individual to Social Learning, (Vriend, 2000) shows how observing the strategies and payoffs of other players changes the market outcome in a standard Cournot game. The same issue has also been considered experimentally. In the same type of game (Cournot competition with or without product differentiation, extended to Bertrand competition with product differentiation), (Altavilla *et al.*, forth.) reports how different information conditions in a market game may influence the behaviors of sellers and the final market outcome. The key finding is that providing full information to sellers about the profits and choices made by other sellers leads to a situation closer to the competitive equilibrium.

Using a computational approach, our paper considers this issue on a search market and compares the qualitative and quantitative properties of Social Learning to those of Individual Learning and Nash. It highlights that under Social Learning, Nash qualitative properties are only preserved for some ranges of parameters. For other ranges, especially when the proportion of informed buyers is large, social learning generates path dependency and uniqueness of posted prices (no price dispersion).

Section 2 presents briefly the model. Section 3 sums up the propositions deduced at Nash equilibrium. Section 4 tests whether these hypotheses still hold with adaptive sellers using individual learning. Section 5 tests the same hypotheses in the context of social learning. Section 6 discusses and concludes.

2. The model[4]

The characteristics of the model inherit those of (Varian, 1980); we consider a posted price market on which S sellers and B buyers are operating. Each buyer wishes to buy a unit good at a maximum price of $v > 0$. With no loss of generality, let us assume that $v = 100$. A fraction a $(0 < a < 1)$ of these consumers is informed. According to (Varian, 1980), these consumers visit the whole set of sellers (or equivalently know the lowest-price seller). We consider an extension of

[4] Due to lack of space, it is not possible to reproduce the description of the reinforcement learning process. This description is available at the URL http://e.darmon.free.fr/ffsmarket/. Results are obtained through numerical simulations. The program and the source code are available on request and are also available at the above address

this model, where informed buyers sample randomly k $(1 < k \leq S)$ sellers. The remaining population (uninformed consumers) randomly selects one seller.

Every seller produces at the same cost that will be normalized to 0. The choice of the reservation price and of the production cost has no impact and only alters the magnitude of agents' payoffs. There is no capacity constraint.

One simulation is a succession of T market rounds[5]. Each round can be divided into three steps: 1) sellers set prices simultaneously; 2) buyers visit sellers and transact; 3) sellers compute their profit for the current round and reward the corresponding pricing strategy.

From step 1 onward, prices (p_t) are posted and set according to a reinforcement learning process[6]. The working of this process can be simply expressed: each seller is endowed with a finite set of pricing strategies ranging from $c = 0$ to $v = 100$.[7] Prices are discrete and chosen from the set $\{0,1,2,...,100\}$ and a reward $(F^i_{t,s})$ is assigned to each strategy *i* by seller s $(s = 1,...,S)$ in period t. At each period, sellers test a pricing strategy, and record the payoff generated by that strategy. During the subsequent round, the larger the reward of a pricing strategy, the larger its selection probability. However, this choice always exhibits some randomness in order for sellers to explore the whole set of possible strategies. Otherwise, sellers could be locked-in to some specific pricing strategies, which would artificially inhibit learning. The initialization and selection rules are described respectively by Equations (1) and (2).

$$F^i_{t=0,s} = F_{0,s} = \delta(Bv) \quad (\forall i, \forall s, \delta \in]0,1]) \tag{1}$$

$$prob_{seller\ s}\{\text{select Strategy i}\} = \frac{e^{F^i_{t,s}/\tau}}{\sum_{j=0}^{100} e^{F^j_{t,s}/\tau}} \text{ with } \tau > 0 \tag{2}$$

Equation (1) defines the initial rewards associated with each pricing strategy at the first round. These are initialized in reference to the maximal profit (Bv) that sellers can generate on this market. Parameter δ reflects the initial degree of optimism of sellers. When equal to 1, sellers expect that each rule will generate the maximum payoff. On the contrary, when equal to 0, they expect each rule to yield a zero profit.

---

[5] We are here motivated by the stationary positions of the system and not on the whole price dynamics. Consequently, T has been set so as to ensure that the system has converged and that the observed data (price distribution, average and standard deviation) no longer evolve over a long period of time. These data are computed from the last 100 rounds. If not specified, the default number of rounds is 1000.

[6] Cf. (Sutton, 1991) for a general presentation of these processes and of their properties; Cf. (Kirman *et al.*, 2001) for an application to market dynamics.

[7] Thus sellers know consumers' maximal willingness to pay $v = 100$. Otherwise, since prices above v generate zero profit, those prices would rapidly disappear during the learning process.

Equation (2) depicts how a seller selects one pricing strategy among the set of all available strategies. This stochastic rule respects the general principle of learning: strategies that are endowed with a larger reward have a larger probability of being selected[8]. Those that generated lower profits, will still be regularly evaluated (i.e. played with a small but non zero probability). Parameter τ sets the exploration-exploitation trade-off of a seller: when this parameter decreases, sellers restrict the set of best strategies effectively used and as a corollary explore alternative pricing strategies less frequently.

At the end of the period, the reward of the pricing strategy currently played is revised. There are two types of learning, either individual or social learning.

2.1. Individual learning

Individual learning (further IL) means that a seller (say *s*) can only learn from his own experience and not from the experiences of other sellers. In that case, he only considers the performance generated by the pricing strategy he has chosen during the round (hence its profit $\pi_{t,s}$) and reinforces the reward attached to that rule, Assuming that *s* played strategy *i* at time *t*, this rewarding mode is described by Equation (3a) :

$$F^i_{t+1,s} = F^i_{t,s} + \alpha\left(\pi^i_{t,s} - F^i_{t,s}\right) \text{ with } \alpha \in \left]0,1\right] \tag{3a}$$

This specification ensures that rewards converge to some weighted average of the profits generated by that strategy. Coefficient α measures the weight an agent assigns to his last experience compared to its previous experiences. As α increases, his past experiences are skipped more rapidly.

2.2. Social learning

Beside his own experience, a seller can also learn from the experiences of other sellers by observing their prices and profits. That is what is meant here by Social Learning (further SL). To keep things simple, we suppose that sellers can perfectly

[8] To implement the selection process, we consider rewards that are normalized on the unit interval: $\tilde{F}^i_t = \frac{F^i_t - F^{Min}_t}{F^{Max}_t - F^{Min}_t}$ where $F^{Max}_{t,s}$ (resp. $F^{Min}_{t,s}$) is the maximum (resp.minimum) fitness observed by the seller s in period *t*. Such a transformation maintains the rank of the payoffs and makes the choice of the exploration-exploitation parameter independent of the absolute magnitude of the reward which facilitates inter-simulation comparisons.

observe the profits of other sellers[9]. In the most general formulation, it is common to suppose that a seller may not necessarily attribute the same credibility to his own experiences as to those of other sellers. For that reason, we will suppose that sellers may voluntarily discount the experiences of other sellers. This effect is captured by Parameter λ. Assuming that a seller s' played strategy j $\left(j \neq i\right)$ at time t, social learning by seller s is depicted by Equation (3b).

$$F_{t+1,s}^{j} = F_{t,s}^{j} + \alpha\lambda\left(\pi_{t,s'}^{j} - F_{t,s}^{j}\right) \text{ with } \lambda \in [0,1] \tag{3b}$$

This process is repeated for all the other sellers s' $\left(s' \neq s\right)$. If a price strategy has been played by two sellers other than s, we consider the average of the profit generated by that rule and update the corresponding pricing rule once, using (3b). Finally, if a price strategy has been played by seller s and by at least one other seller, we suppose that s gives more importance to his own experiences, and therefore skips the experience of other sellers[10]. λ is the discount rate that a seller applies to the experiences of other sellers. If λ=0, the seller completely discards other sellers' experiences (individual learning). On the contrary, if λ=1, the seller rewards other experiences' with the same weight as his own experiences.

3. Nash predictions

Varian (1980) established the Nash distribution of posted prices at mixed equilibrium when informed consumers visit the whole set of sellers. (Waldeck, forth.) extended this model by allowing informed buyers to sample only a fraction k of the set of sellers and by studying the effect of information on prices and price dispersion. These results in (Waldeck, forth.)[11] can be summarized as follows:

Proposition 1. The cumulative distribution of posted prices is equal to $F(p;a,k) = 1 - \left(\frac{(1-a)(v-p)}{kap}\right)^{\frac{1}{k-1}}$ on the support $\left[b(a,k),v\right]$ with $b(a,k) = (1-a)v/((k-1)a+1)$.

[9] A further extension would be to consider the case of noisy or incomplete observation of these profits.

[10] Another less radical approach would have been to suppose that the seller considers an average of the two profits weighted in favour of his profit. However, we suspect that such a refinement would not provide a substantial value, but would add a new parameter.

[11] Proposition 1 is shown by Varian. The effect of a on posted market prices are included in (Stahl, 1989) paper on sequential search and a weak converge theorem is shown when $k = S \to \infty$.

Proposition 2. The average posted price decreases with the fraction of informed consumers (a) and increases with k .

Proposition 3. The mean price paid by informed consumers decreases with the fraction of informed consumers (a) and with k .

Proposition 4. The dispersion of posted prices is a left skewed inverse U-shaped function of the fraction of informed consumers. As the number of firms is higher than 8, the peak of the dispersion is reached for a close to 1 $(a \approx 0.99)$

Keeping a constant, the standard deviation of the NSE is a right skewed inverse U-shaped function of k . Whenever $k < 8$, the variance increases in k . The decrease is usually at a low pace and the variance remains significant for large k [12].

The first proposition establishes the existence of a unique symmetric equilibrium in mixed strategy. This distribution is bimodal, which reflects the presence of two types of consumers. Any time a seller posts a high price (close to the monopoly price), he targets the population of non informed buyers (and captures their whole surplus). Conversely, anytime he sets a low price, his probability of selling to the whole market increases. Such a seller then targets with a large probability the population of informed consumers that actively look for competitive prices. Propositions 2 to 4 show that improving market transparency (by increasing the fraction of informed buyers) has a double effect: as expected, it lowers the average price paid by both informed and non informed buyers. However, one unexpected effect is that price dispersion *increases* [13] (see the proof in (Waldeck, forth.)).

4. Nash versus Adaptive Sellers: Individual Learning

We compare the choice made by adaptive sellers to that made by Nash sellers according to four criteria: *i)* distribution of posted prices (Proposition 1); *ii)* average of posted prices; *iii)* average of prices accepted by informed buyers (Propositions 2 and 3); *iv)* dispersion of posted prices (Proposition 4). To better understand the results, let us first highlight two cases *à la* Varian $(k = S)$. We considered here the cases $a = 0.2$ ("low" fraction of informed buyers) and $a = 0.8$ ("high" fraction of informed buyers). We have chosen these two polar values so as to better identify the effects of the change in a . However, the observations made will be generalized to the whole range of parameters a and k .

---

[12] This result (variations with respect to k) is not proven analytically but only stands as a numerical result.

[13] The assertion that price dispersion increases in a is nevertheless dependent on the number of firms sampled by informed consumers). For example, with only two firms, price dispersion would decrease at a = 70%, for 3 firms at around 80% and for more than 3 firms at levels above 90%.

4.1. Two representative cases

4.1.1. Case $a = 0.2$

Let us assume that $S = 20$ and $B = 1000$. In the first case, the fraction of informed buyers is relatively low $(a = 0.2)$. The density of the stationary distribution (resp. the cumulative distribution) of posted prices is displayed in Figure 1a (resp. 1b):

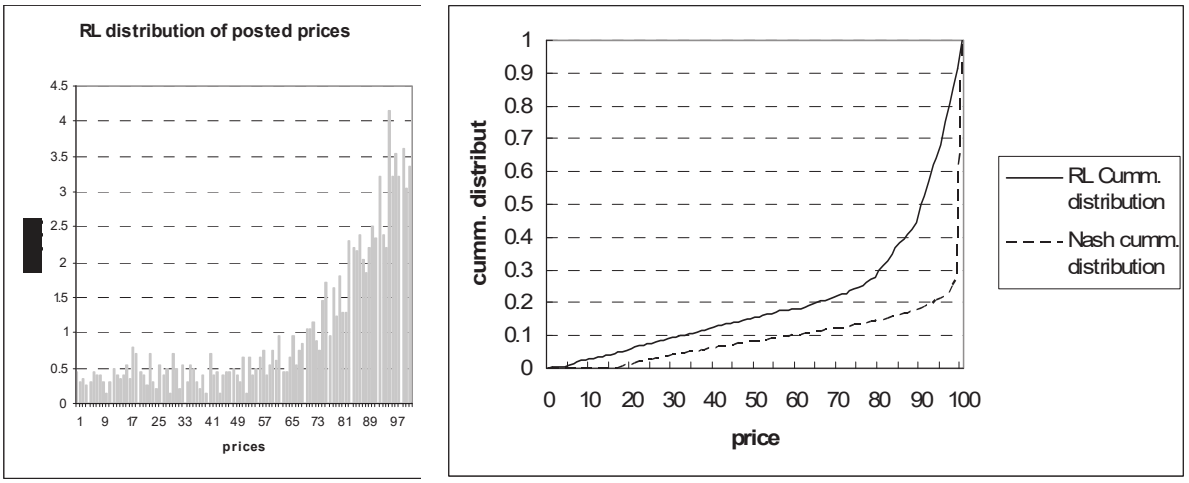

Figure 1a and 1b. *Posted price distribution played by RL agents (left); Nash and RL posted price cumulative distribution (right); The RL distribution has been computed using the last 100 rounds; simulations performed with the following parameter set* $\{a = 0.2; \tau = 0.05; \delta = 0.2; \alpha = 0.8; T = 1000\}$

Figure 1b compares the theoretical Nash distribution to that played by adaptive sellers. It shows that these two distributions do not coincide. Put differently, with $a = 0.2$, choices made by adaptive sellers do not converge to those depicted by the Nash equilibrium in mixed strategies. Kolmogorov-Smirnoff adequacy tests confirm this intuition statistically[14]. In addition, the average and the standard deviation of the distribution differ from the theoretical ones.

The literature of learning in games shows that the NSE is in general unstable under a variety of learning rules including reinforcement or best response learning ((Hopkins *et al.*, 2002), (Benaïm et al., 2005)[15]. In addition, previous works compared Nash and RL- predictions to actual human behaviors drawn from laboratory experiences (cf. (Erev *et al.*, 1998)). In doing so, they showed that RL-driven outcomes do efficiently describe actual human behaviors with sufficient accuracy and that their predictive (ex ante) and descriptive (ex post) powers are better than those of equilibrium predictions. If the same ranking applied to the

[14] Adequacy tests measure the distance between the two distributions and test whether this distance is statistically different from 0.

[15] However, we should point out that these authors consider some kind of positive definite adaptive dynamics which is not the case here.

search market analyzed in that paper, we could infer from our observations that the Nash equilibrium would not reproduce the choices made by actual sellers. However, there are three limiting points before concluding in this way. First, one should note that (Erev *et al.*, 1998) used a different formulation of RL for both the probabilistic selection rule and the fitness updating rule[16]. Second, even bypassing the previous remark, one should reproduce Erev et Roth's experiment in the specific search market considered here, and examine whether Erev et Roth's results can be extended to that setting. Third, an appropriate fit of data does not necessarily mean that subjects actually behave as reinforcement learners but only that everything happens *as if* they do.

4.1.2. Case $a = 0.8$

Let us then consider a second situation where the fraction of informed sellers increases to $a = 0.8$ (Figures 2a. and 2b.). Comparing again the RL- to the Nash-distribution with individual learning, one can make two observations. First, the two distributions are, as previously, not identical (Figure 2b). This is confirmed by adequacy tests. Second, the RL distribution when $a = 0.8$ has a bimodal shape (Figure 2a). A first peak is located at $p = v$, while a second peak is located close to the Bertrand pricing strategy[17]. Such a conclusion did not appear in the previous case $\left(a = 0.2\right)$ but is consistent with the phenomena explaining a lower average price together with higher price dispersion for the RL distribution. In other words, when the fraction of informed buyers increases, extreme prices will become equally attractive: when he posts a low price, a seller may capture the whole market. But, posting a high price insures a high profit margin while selling only to uninformed consumers. With a small market share of informed consumers $\left(a = 0.2\right)$, low prices are not attractive, so that the distribution is more likely to be mono-modal. As a increases, "low price" strategies are more attractive since informed buyers are more numerous. Sellers will learn it and, as in the Nash equilibrium, price dispersion will increase with a.

---

[16] They are two differences between our learning rule and the one used by (Erev *et al.*, 1998). First, the updating rule of the fitness described by equation [2] considers an average of the past fitness with the current profit, whereas profits cumulate to fitness in Erev and Roth's paper. Second, their probabilistic choice rule [3] is a simple proportional rule of current fitness. We choose to keep to the logit model because of its axiomatic foundations coming from the psychological literature (e.g. (De Palma *et al.*, 1989) for a review).

[17] If all buyers are informed, competition by price undercutting will lead prices to $p = 1$ in the case of a discrete price grid. A firm increasing its price from $p = 1$ will make no sale whereas a firm decreasing its price to $p = 0$ will make zero profit. This is Bertrand competitive equilibrium. In the case of a continuous distribution, the symmetric Bertrand equilibrium is $p = 0$.

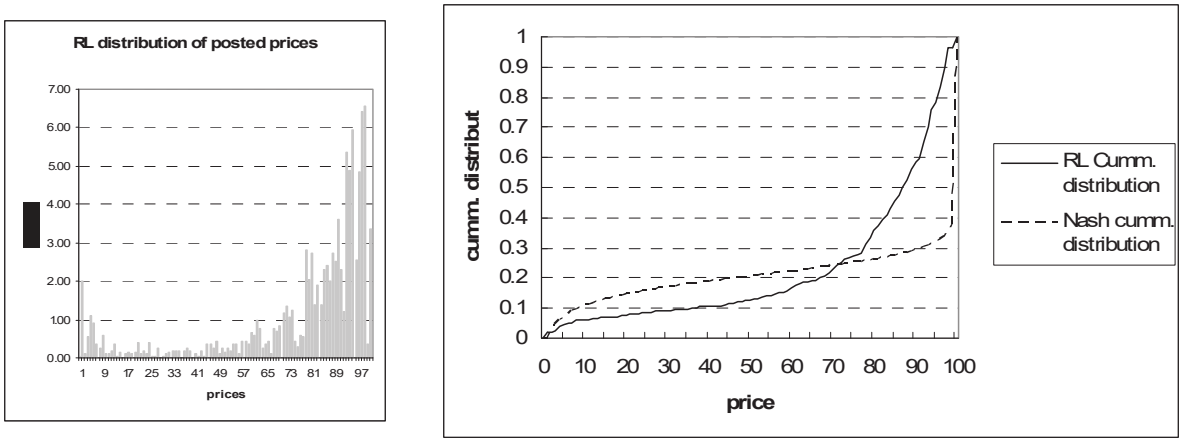

Figure 2a and 2b. *Posted price distribution played by RL agents (left); Nash and RL posted price cumulative distribution (right); the RL distribution has been computed using the last 100 rounds; simulations performed with the following parameter set* $\{a = 0.8; \tau = 0.05; \delta = 0.2; \alpha = 0.8\}$

Table 1 reports the mean and standard deviation statistics in the case of both adaptive (individual learning) and Nash sellers for the two previous cases:

	Mean posted prices		**Mean prices accepted by informed buyers**		**Standard deviation of posted prices**	
	Nash	**RL IL**	Nash	**RL IL**	Nash	**RL IL**
$a = 0.2$	91.3	**78.7**	34.9	**18.0**	20.1	**23.9**
$a = 0.8$	79.7	**77.0**	5.0	**6.6**	34.7	**24.6**

Table 1. *(Individual Learning) Comparison of average posted price and price dispersion for* $a = 0.2$ *and* $a = 0.8$ *(data computed from the last 100 periods of the two previous simulations*[18]*; simulations performed with the following parameter set* $\{\tau = 0.05; \delta = 0.2; \alpha = 0.8, T = 1000\}$ *)*

Comparing the price dataset obtained when $a = 0.2$ to that obtained when $a = 0.8$, reveals that as the proportion of informed buyers increases, *i)* the mean price accepted by uninformed buyers decreases (mean posted price, cf. Proposition 2); *ii)* the mean price accepted by informed buyers also decreases (cf. Proposition 3); and *iii)* price dispersion increases. This last conclusion is in agreement with analytical findings. Indeed, Proposition 4 establishes that price dispersion increases as long as $a < 0.99$. However, Nash variations are more pronounced than RL variations both for average prices and for price dispersion.

From these simple observations, we conclude that for our two examples although sellers characterized by individual learning do not learn to play the Nash equilibrium in a strict sense, their behaviors mimic the characteristics of the Nash outcomes for

[18] We simply considered the two previous simulations here, since a multishot analysis reveals that inter-simulations variations are very small (cf. Figure 3 hereafter).

the first and second order statistics. This conclusion may reinforce the validity of Nash equilibrium for comparative static in economic analysis. In some sense, our result may be restated in the following way: the drivers of prices and price dispersion which is profit maximization will be learned by adaptive sellers although in a imperfect way. Thus for example, the Nash price dispersion depends on the number of visits made by informed buyers (k). An increase in k leads to a linear increase in the market size gained by the lowest priced firm but to an exponential decrease in the probability of being the lowest price. At Nash equilibrium, firms' optimal reaction will lead to a shift in the distribution of posted price with more weight given to the extreme bounds of the support of the price distribution, thereby increasing price dispersion for low k but increasing the average posted price for all $k > 1$. However for large k, low pricing will become concentrated at the marginal cost so that low pricing will be discouraged in favor of pricing near the monopole price. Price dispersion decreases for k large enough. But, low pricing is still a profitable strategy which may be heavily reinforced if none of the other firms adopt it. Since the preceding arguments hold for Nash but also for RL, the shift in the distribution for RL is similar to that for Nash, albeit not in the same proportions.

4.2. Generalization

Previous observations only hold in the case of the two specific simulations presented above. More precisely, they are dependent on the specific values of coefficients a and k and on the specific set of learning parameters (δ, τ, α). To assess the robustness of this result, we need to consider a larger set of parameters a and k and to generalize the previous analysis to a wider spectrum of learning parameters.

To explore the first point, let us consider first Varian's model $(k = S)$ and make a sensitivity analysis on parameter *a* independently. For that, we considered discrete values $a = \{0.1, 0.2, ..., 0.9\}$ inside the range of definitions of parameter a. We randomly chose this parameter and made 400 simulations of the process. First, we implemented an adequacy test for each iteration. Even when considering large acceptance levels, these tests systematically reject the equality assumption between the Nash (described by Proposition 1) and the RL distributions. Second, we analyzed the evolution of the average posted price, of the average price accepted by informed buyers, and of the price dispersion with respect to a.

Figure 3 corroborates graphically the observations made in the previous section. In accordance with Propositions 2 and 3, the two averages (posted prices and prices accepted by informed buyers) decrease as a increases. In addition, in accordance with Proposition 4, price dispersion increases with a.

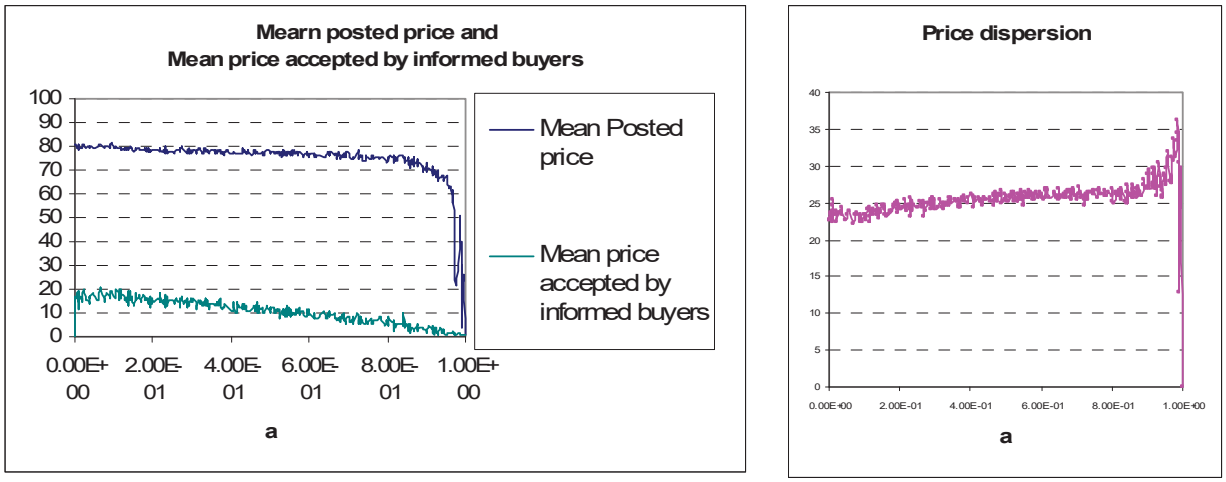

Figure 3. *Average posted price, average price accepted by informed buyers, price dispersion; 400 simulations with random* a *and* $\{\tau = 0.1; \delta = 1; \alpha = 0.8\}$

We checked these propositions statistically[19]. With respect to a and k, we ran 100 simulations for each parameter's configuration (a,k). For each simulation, we computed the aggregate mean and variance of the price distribution over the last 100 periods. We thus come to the following results on averages over the 100 simulations (Waldeck and Darmon, 2006):

Result 1. (Price distribution, Individual Learning) The RL posted price distribution never converges to the Nash posted price distribution.

Result 2. (Average posted price, Individual Learning) The RL mean price is a decreasing function of the proportion of informed consumers in the market.

Result 3. (Standard deviation for RL, Individual Learning) RL Standard deviation conforms to the Nash prediction with respect to a change in *a* except when *k=2*. A one-sided Wilcoxon ranked test (at 5%) shows that 83% of figures reported a variation in agreement with the NSE prediction whereas only 3% showed an opposite variation to NSE.

In (Waldeck and Darmon, 2006), we evaluated the robustness of Results 1 to 3. For that, we needed to consider weaker hypotheses. For instance, we showed that the statistically robust result is that the RL mean price is "non increasing in a" while the Nash prediction is that the mean price is *strictly* increasing in a. One key reason for doing this is that RL-outcomes inherently generate some statistical randomness. This is caused by the persistence of explorative behaviors which introduce some noise into the observations derived from the simulations (unlike Nash predictions).

[19] (Waldeck and Darmon, 2006) details the whole procedure and reports sensitivity tests with respect to learning parameters.

5. Social Learning

Let us now consider Social Learning. For all the experiments reported here, we supposed that $\lambda = 1$. Although we need to consider less extreme values for this parameter in the future, this situation is qualitatively interesting as it measures the effect of social learning when fully introduced. As previously, let us first consider the two previous cases ($a = 0.2$ and $a = 0.8, k = S$) *à la* Varian (1980). This will suggest some general intuitions that will be generalized afterwards by a multi-shot analysis (5.2). We then discuss the results and highlight some factors that could account for the differences observed between individual and social learning (5.3).

5.1. Two single cases

5.1.1. Case $a = 0.2$

Figure 4 reports the distribution of posted prices as $a = 0.2$ when SL is introduced. This has to be compared to Figure 1 that presents the same case under IL.

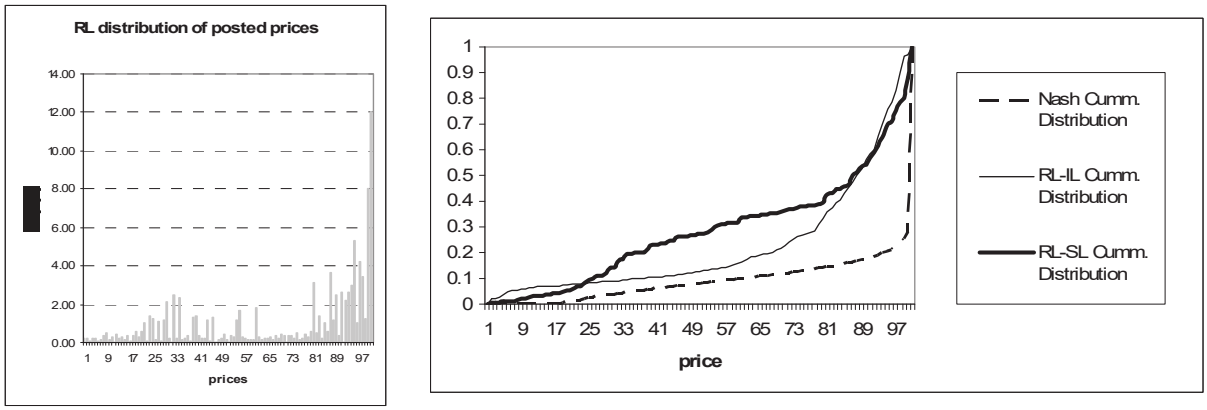

Figure 4a and 4b. *Posted price distribution played by RL agents (left) under SL; Nash and RL posted price cumulative distribution under IL and SL(right); the RL distributions have been computed using the last 100 rounds; simulations performed with the following parameter set* $\{a = 0.8; \tau = 0.05; \delta = 0.2; \alpha = 0.8\}$ *(Il and SL) and with* $\{\lambda = 1\}$ *for (SL)*

From Figure 4a, we observe that the market price distribution is still dispersed. Compared to Figure 2a (IL case), we notice that this distribution has a less regular shape. Using Figure 4b, we can now compare three elements (Nash, Individual Learning and Social Learning). Figure 4b plots the cumulative price distributions in these three cases. Graphically, we can see here that these distributions do not coincide. That is, the choices made by adaptive sellers characterized by Social Learning converge neither to those made by Nash sellers, nor to those made by adaptive sellers characterized by Individual Learning only.

5.1.2. Case $a = 0.8$

We now consider the second case when $a = 0.8$. Figure 5 reports the price distribution of two simulations using the same set of parameters.

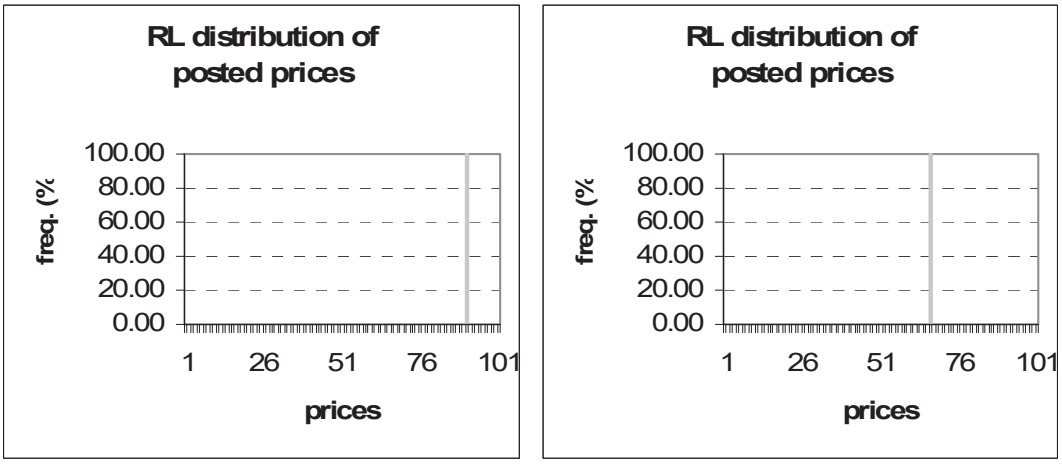

Figure 5a and 5b *Posted price distribution played by RL agents under SL; two different simulations; the RL distributions have been computed using the last 100 rounds; simulations performed with the following parameter set* $\{a = 0.8; \tau = 0.05; \delta = 0.2; \alpha = 0.8; \lambda = 1\}$ *(SL)*

Unlike previous cases, these two figures highlight two points. First, we note that when $a = 0.8$, the process is highly *path dependent*. That is, the same set of parameters (especially with the same exploration-exploitation trade off parameter τ) leads to different price structures. Second, we observe that the market price distribution is no longer dispersed. Instead, the effect of Social Learning in this configuration is to make the prices posted by each seller converge to a unique price. Path dependency and no price dispersion appear as a combination of *i)* Social Learning and *ii)* a particular market structure $(a = 0.8)$.

The following table reports two different cases for the same parameter configuration $\{\tau = 0.05; \delta = 1; \alpha = 0.8, T = 1000\}$:

	Mean Posted Prices		**Mean prices accepted by informed buyers**		**Standard deviation of posted prices**	
	Nash	**RL SL**	Nash	**RL SL**	Nash	**RL SL**
$a = 0.2$	91.3	**72.0**	34.9	**20.1**	20.1	**20.8**
$a = 0.8$ (simulation #1)	79.7	**90.0**	5.0	**90.0**	34.7	**0**
$a = 0.8$ (simulation #2)	79.7	**63.3**	5.0	**63.3**	34.7	**0**

Table 2. *Comparison of average posted price and of price dispersion for* $a = 0.2$ *and* $a = 0.8$ *(data computed from the last 100 periods of the two previous*

simulations; simulations performed with the following parameter set $\{\tau = 0.05; \delta = 1; \alpha = 0.8, T = 1000\}$ *); Social Learning*

From this table, we can already see that variations of average posted prices with respect to a may have an ambiguous impact. When comparing the average price reached when $a = 0.2$ to that reached when $a = 0.8$ (simulation #1), we obtain two opposite variations: using simulation #1, we may conclude that the average price increases (72.0<90.0), while using simulation #2, we may conclude inversely (72.0>63.3). Concerning price dispersion, this indeterminacy associated with this path dependency vanishes since the posted price distribution reduces to a single price when $a = 0.8$, whatever the simulation considered. These three examples show that the qualitative predictions of Nash with respect to a variation of a *i)* do not always hold with Social Learning when considering the average posted price and *ii)* are violated when dealing with price dispersion. These are striking differences from the IL case. As previously, we need to confirm these preliminary observations by a multi-shot analysis.

5.2. Generalization

To assess the robustness of the above observations, we built a dataset that contains 25 repeated simulations for each (a,k) combination[20]. Using this dataset, we checked if the results obtained under Individual Learning extend to Social Learning. Consequently, we tested first the convergence of RL price distributions under SL to the Nash distribution. Then, we evaluated if the qualitative predictions of Nash (with respect to a variation in the market structure i.e. a and k) still hold under SL.

To test the convergence to Nash, we performed a Kolmogorov-Smirnov test (with a 95% confidence level) on each individual simulation. This test rejected the null hypothesis (equality between the Nash and the Social Learning distributions) for each of the 25 simulations of each configuration (a,k). Similar to Result 1 obtained under IL, we deduce Result 4.

Result 4 (Price distribution, Social Learning). The price distribution under Social Learning does not converge to the Nash distribution whatever the (a,k) parameters.

Turning to the properties of first-order statistics, the first observation is that the average posted price is a decreasing function of a only for values of a lower than 0.5, as evidenced by Figure 6. Even when averaging over 25 simulations, we no

20 All other parameters $\{\alpha = 0.8; \tau = 0.05; \delta = 0.2; \lambda = 1; T = 1000\}$ are constant. The tests and charts presented in this section are all deduced from this dataset.

longer observe a decrease in average price when the proportion of informed consumers increases.

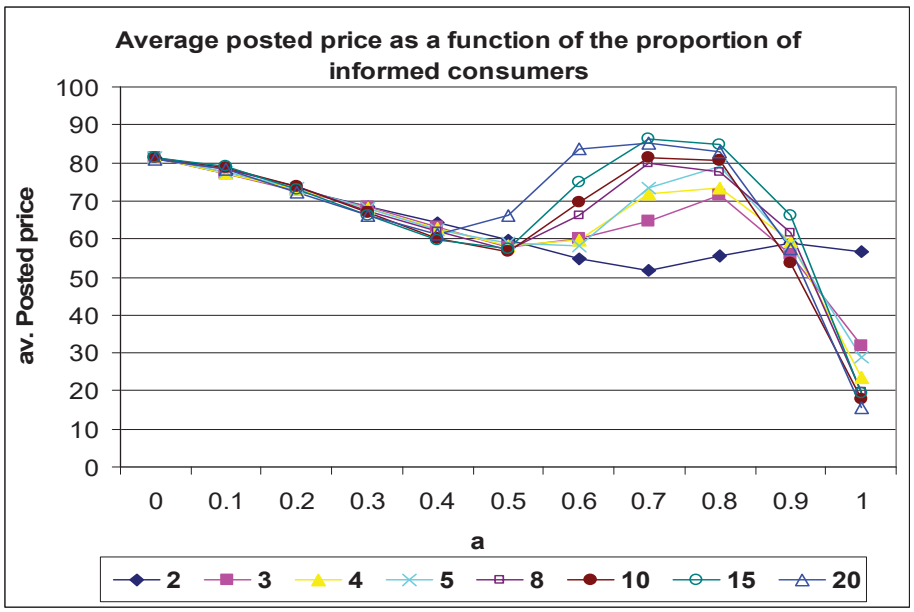

Figure 6. *Average posted price for different (a;k) configurations*

NOTE. – Each point represents an average of the 25 simulations for each (a,k) configuration.

In addition, Figure 7 below plots the average posted price for $k=10$ as a function of a (25 simulations for each value of parameter a) for both individual and social learning. The IL curve has a regular decreasing shape. Moreover, the figures are quite close together. This shows that each of the 25 simulations converged to approximately the same average price. The SL curve shows that there is a large dispersion of average prices across two simulations, as a is larger than 0.5. Such inter-simulation heterogeneity confirms that, when SL is considered, the process is path dependent for some values of parameter a. For example, average prices may converge to a large corridor of values ranging from 40 to 100 for $a=0.8$ depending on this path. For $a=1$, all consumers are informed. The Nash prediction is a Bertrand equilibrium (equal to 1 in the case of $a=1$).Yet, average posted prices with SL are different from those observed at Nash equilibrium. Again, such difference shows the emergence of apparently collusive behaviors caused by price imitations. The imitation of highly successful strategies such as capturing a large share of informed consumers will lead to a unique price. However, which price is chosen becomes a matter of chance.

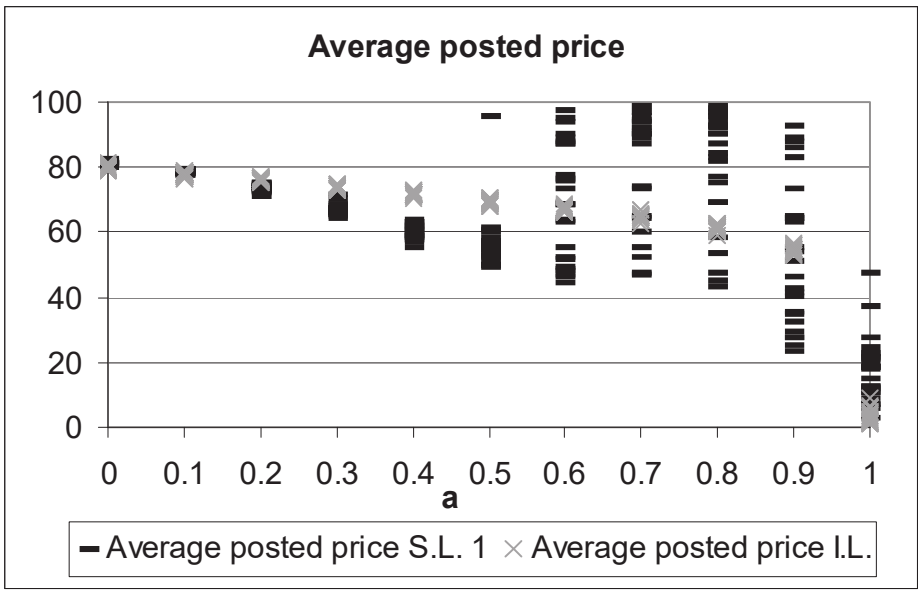

Figure 7. *Average posted price; comparison between IL and SL for different values of a, k=10. (25 simulations for each parameter a and learning type)*

In addition, by performing a one-sided Wilcoxon ranked test (with a precision level set to 5%), 98 % of the values reported a variation in line with Nash under IL[21] whereas this score was only 64% under SL. Moreover, none of the figures reported a significant increase under IL whereas 13.5% did so under SL.

Concerning price dispersion, Figure 8 compares IL to SL. Whereas IL data were in line with Nash predictions for standard deviation, this is no longer the case for SL, where price dispersion is now decreasing. Moreover, low dispersion is present for values of a larger than 0.6. This indicates that prices become are more concentrated around the average posted price for larger values of a and that in some cases, we may observe a convergence to a unique price. Moreover, for $a = 1$, price dispersion is equal to zero in 22 simulations over 25 with SL[22]. This means that with SL, firms learn to fix a unique price but *above* the Bertrand equilibrium.

[21] This means that 98% of the values exhibited a significant decrease in average price when a increases from a to $a+0.1$ (for $a \neq 1$).

[22] The other 3 simulations show almost no dispersion of prices.

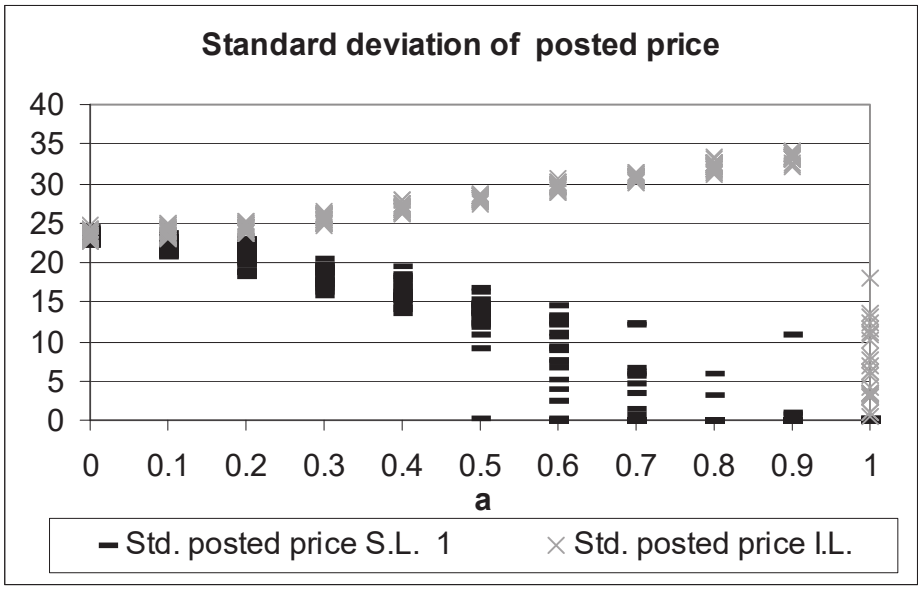

Figure 8. *Posted Price dispersion; comparison between IL and SL for different values of a, k=10.(25 simulations for each parameter a and learning type)*

The above qualitative conclusions have been statistically generalized to other values of k, hence Result 5.

Result 5. (Variations of mean posted price and price dispersion with respect to a, Social Learning) Social learning is different from Nash and Individual Learning with respect to a variation in a. This is true both for the average posted price and price dispersion. Price dispersion reduces with larger values of a and prices converge to a unique price in some cases.

Let us now examine if the qualitative predictions of Nash equilibrium still stand when considering variations in parameter k. Table 3 shows the average posted price for Social Learning as compared to Nash.

a \ k	2		3		4		5		8		10		15	
0	100	81.2	100	81.3	100	81.3	100	81.2	100	81.1	100	81.3	100	81.3
0.1	90	77.1	91	77.1	91	77.2	91	77.6	92	78.3	93	78.8	94	79.0
0.2	81	73.1	82	72.8	84	73.0	85	73.3	87	73.9	88	73.7	90	73.0
0.3	72	68.3	75	68.6	77	68.2	79	68.0	82	67.3	84	67.0	87	66.3
0.4	64	64.3	68	63.3	71	62.9	73	62.4	79	62.1	81	60.3	85	59.8
0.5	55	**59.8**	60	57.8	65	58.1	68	58.9	75	57.0	78	56.8	83	57.5
0.6	46	**54.6**	53	**60.0**	59	**59.6**	63	58.0	71	66.3	75	69.6	80	75.0
0.7	37	**51.7**	46	**64.5**	52	**71.9**	57	**73.4**	67	**80.0**	71	**81.3**	78	**86.3**
0.8	27	**55.4**	37	**71.3**	45	**73.5**	51	**79.0**	63	**77.7**	67	**80.4**	75	**84.7**
0.9	16	**59.0**	27	**56.0**	35	**59.2**	42	**58.7**	56	**61.5**	62	**53.6**	71	**66.1**
1.0	1	**56.6**	1	**32.1**	1	**23.7**	1	**28.7**	1	**19.9**	1	**17.9**	1	**19.2**

Table 3. *Average posted price under Nash (shaded cells) and RL-Social Learning for different values of a and k*

NOTE. – The dark cells are the theoretical prediction of Nash. The light cells are for social learning. Each number represents an average of the 25 simulations simulation for a specific (a,k) configuration. Bold numbers indicate that the mean price of social learning is larger than the Nash mean price. One should recall that this statistics (average of posted prices) exhibits important inter-simulations differences (path dependency), for high values of the coefficient a.

We also plotted the average posted price and the price dispersion for different values of parameter k (cf. Figure 9).

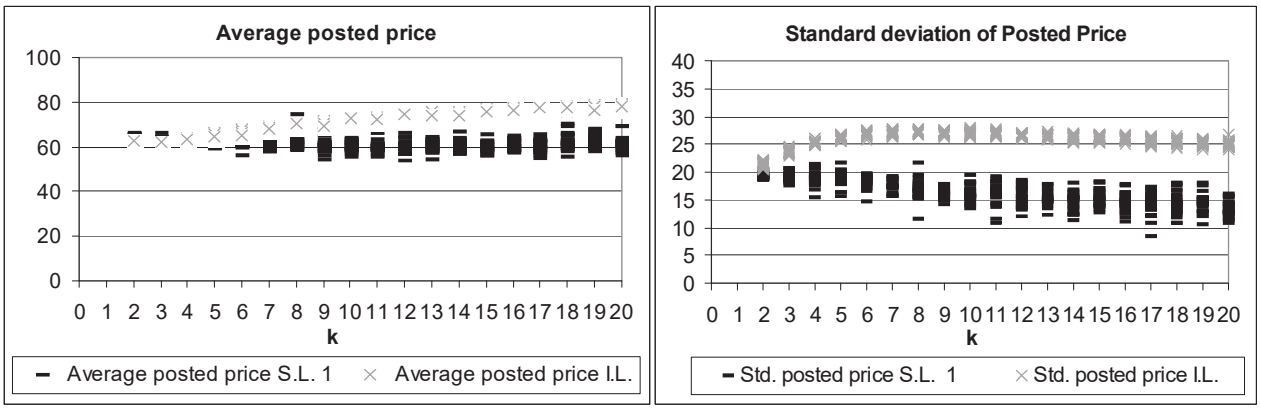

Figure 9. *Average price and price dispersion; comparison between IL and SL for different values of k, a=0.4. (25 simulations for each parameter a and learning type)*

For SL, average prices seem relatively stable with respect to variations of k. This is confirmed by performing a Wilcoxon ranked test (at 5%). This test evidences that only 6.2 % of the simulations are significantly increasing with k (vs 81% for IL) and 5% are significantly decreasing with k (vs 5.6% for IL). With SL, price dispersion is decreasing with k. Again, Nash predictions are clearly rejected in this case. As evidenced by Figure 10, considering higher values of a $(a = 0.9)$ does not change this result much.

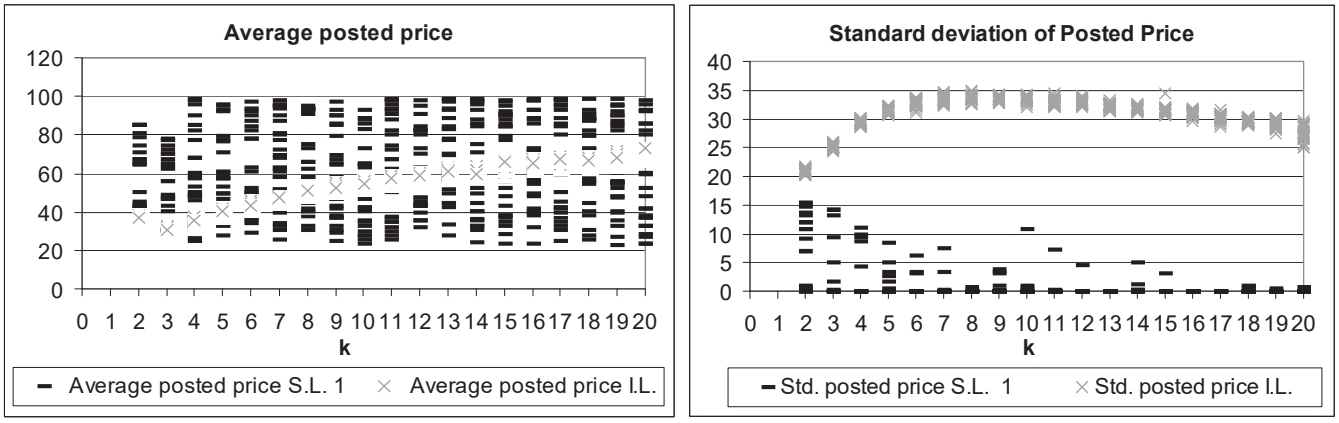

Figure 10. *Average price and price dispersion; comparison between IL and SL for different values of k, a=0.9. (25 simulations for each parameter k and learning type)*

Again, this figure reflects the path dependency of the process for high values of a. This path dependency is a striking difference compared with IL for which any price dependency was totally absent, and a given (a,k) configuration leads approximately to the same outcome across simulations. Result 6 groups these different observations.

Result 6. (Variations of the mean posted price and of price dispersion with respect to *k*, Social Learning). Social learning is different from Nash and individual learning with respect to a variation in k both for average posted price and price dispersion. Price dispersion reduces with larger values of a converging to a unique price for large values of k. Moreover, the resulting average prices are path dependent in this range of values of (a,k).

5.3. Discussion

In this section, we examine some factors on which the previous lock-in effect could be sensitive. More precisely, we need to explain why we observe path dependency for large values of a and no such path dependency for smaller values. First, we shall highlight the effects of the type of competition faced by sellers in these two cases. Second, we shall analyze the effects of the learning parameters on the result.

5.3.1. Effects of the type of competition

First of all, let us note that for low a (less than 0.5), some diversity in the pricing rules played by sellers remained. This can be explained by the type of competition faced by sellers. When most consumers are uninformed, the best-price seller gets both his share of uninformed consumers [23] and share of informed consumers. On the contrary, a high price seller gets his fraction of uninformed sellers only. However, the competitive advantage of posting the lowest price is not very large in this case. Indeed, the gain of low pricing obtained on informed consumers may not be offset by the additional gain of high pricing when targeting uninformed consumers only. Hence, some diversity in the price posted by sellers remains.

For a larger proportion of informed consumers, the gap between the performance of a high pricing strategy (thus getting only a smaller mass of uninformed consumers) and a low pricing strategy (thus getting additionally an increased mass of informed consumers) increases. Since sellers observe the payoff of other sellers with SL, they will instantaneously observe which price leads to the highest profit. Depending on what this price is at a given moment of time, the process gets rapidly

[23] Uninformed consumers sample randomly one seller at each period. However, since we consider a large number of buyers, on average, each seller gets a mass of $B(1-a)/S$ of uninformed consumers.

locked into this price. Yet, since sellers never completely stop exploring alternative strategies $(\tau = 0.05)$, some jumps to a new price may occur at different moments in time. Now, a unique seller exploring by chance a new price will be imitated only if the change is to a lower price. A new lock-in will appear at this new price. However if price margins become too small, eventually a jump to a higher price may appear much in the spirit of Edgeworth cycles. The existence of these jumps is corroborated by looking at price dynamics. Figure 11 reports the evolution of the average posted price over 5000 market periods.

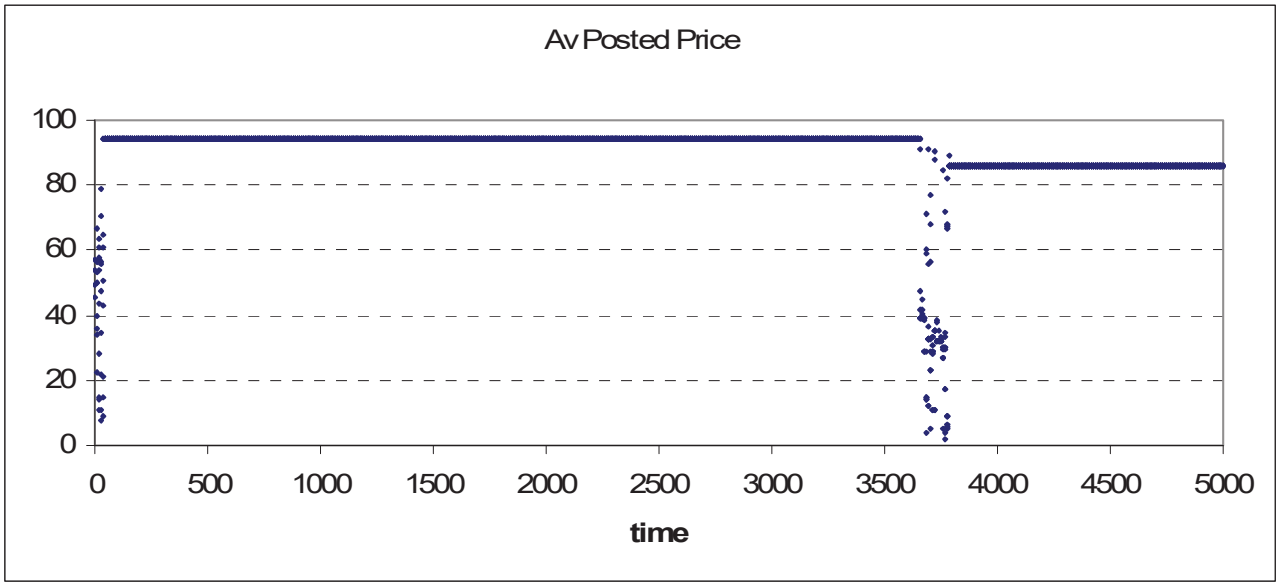

Figure 11. *Evolution of the average price, T= 5000, (a,k)=(0.9;19)*

As one can see, firms experience many price strategies during the first rounds (up to period 100). As an effect of learning, these experiments stop and then all firms are fixed at the same price. However, after 3700 periods, new experiments take place, and firms' posted prices stabilize to a new level.

5.3.2. Effects of the learning parameters

Since our main objective was to test whether the Nash equilibrium was robust to adaptive learning, we need now to check that the lock-in observed with SL and high values of Parameter a was not an artifact of a particular parameter set ruling the behavior of the reinforcement learning process. This process is characterized by three parameters: δ (initial fitness), τ (exploration-exploitation parameter), and α (reward updating parameter). We analyze the effect of these parameters for the case $(a,k) = (0.9,19)$ for which lock-in was manifest.

To evaluate the effect of the magnitude of initial fitness (ruled by parameter δ), we report 100 simulations form randomly drawn initial fitness. We observe the same kind of path dependency and convergence to a unique price (Figure 12).

Hence, the magnitude of initial fitness does not affect the general result of convergence to a unique price. [24]

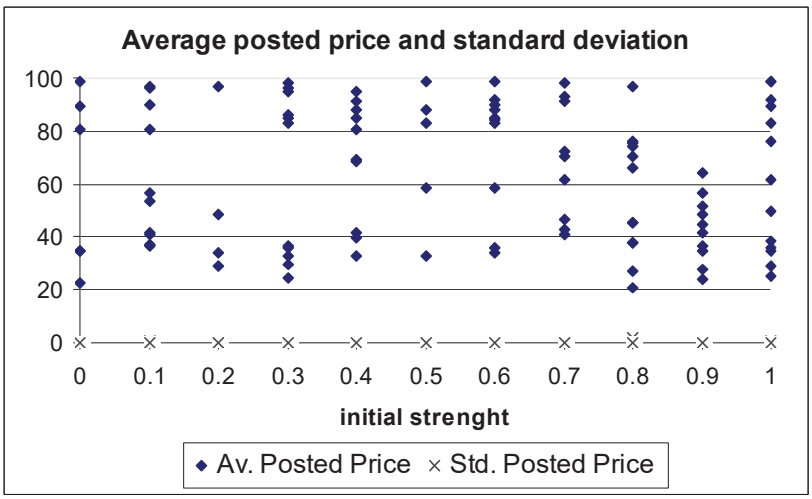

Figure 12. *Average posted price and standard deviation as a function of initial strength; 100 runs simulations for (a,k)= (0.9;19) with random draw of* δ .

Increasing the exploration-exploitation parameter (τ , called 'temperature' in the RL literature) will lead to more numerous experiments and to more frequent switches. As expected, Figure 13 shows that the lock-in effect is dependent on this parameter. Increasing the temperature coefficient decreases the path dependency and generates higher levels of price dispersion. That is, with a higher temperature, firms learn to set different prices more frequently[25]. In general, the bifurcation point is around $\tau = 0.07$.

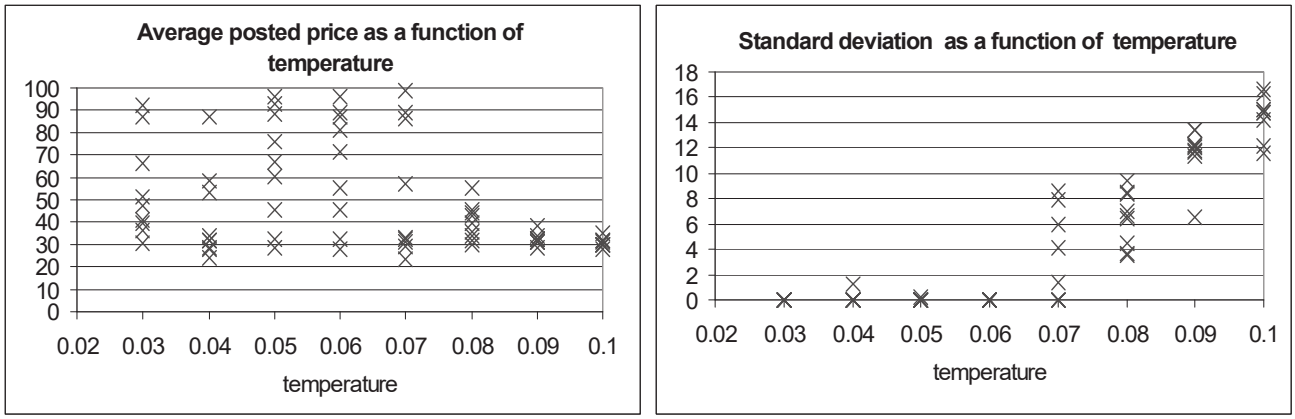

Figure 13. *Average price and price dispersion, for* $\alpha = 0.8$ *, 9 simulations for each parameter configuration (temperature* τ *, reward updating parameter* α *)*

[24] However, by increasing the initial expectation on profits, we should expect a higher price than for less optimistic expectations, on average computed over several simulations. We have not yet tested whether this expectation was confirmed.

[25] This qualitative observation is true for different values of α (reward updating coefficient).

Although firms learn to set different prices when firms experience more strategies, one has to note that these prices generate an identical performance level (i.e. profit) which indeed is consistent with SL[26].

The last learning parameter is the coefficient used to update the fitness of the rule used (α). The lower this parameter, the longer the memory of the sellers about the performance of past prices i.e. the slower a seller 'records' new profits into his fitness. We should expect that the magnitude of this parameter does not affect the general result (path dependency, convergence to a unique price) mainly because *i)* this parameter affects both individual learning (Eq. 3a) and Social Learning (Eq. 3b) in the same way and *ii)* all sellers are endowed with the same coefficient ('synchronous' learning). This is effectively the case as shown by Fig. 14:

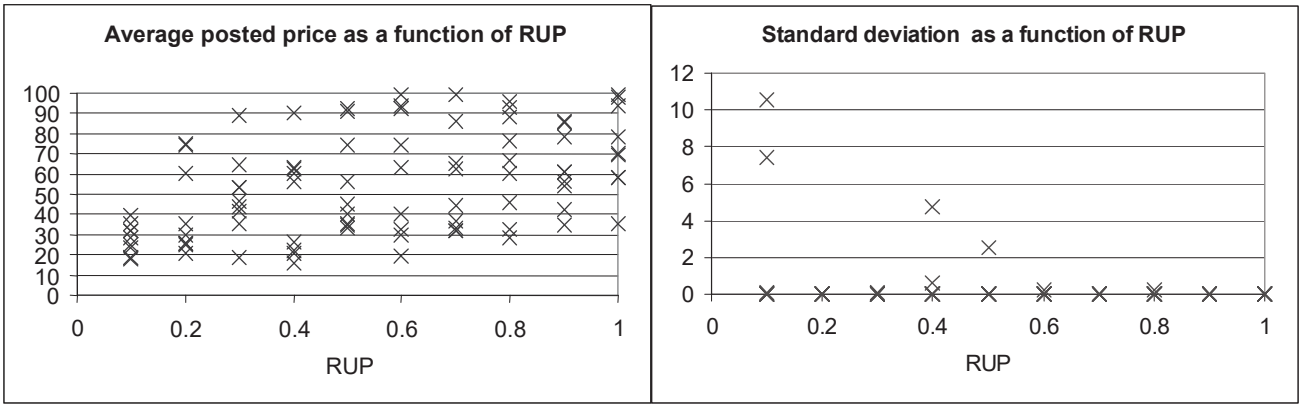

Fig. 14. *Average price and standard deviation for* τ =0.05 *and* α *ranging from 0 to 1. (9 simulations for each parameter configuration (temperature* τ *, reward updating parameter* α *))*

However in some cases, price dispersion remains (for $(\alpha = 0.1)$ 3 simulations out of 9 and for $(\alpha = 0.2)$ two out of 9). One explanation is the following: exploration may appear at a randomly chosen period of time. Since statistics are computed over 100 periods of time, it may be that by chance dispersion appeared. From these observations, we can infer that the only determining learning parameter is the exploration-exploitation parameter τ. This is confirmed by Figure 15. There is a bifurcation point at $\tau = 0.07$ where the uniqueness of posted prices happens in less than 28% of the cases. As expected, such path dependency is weakened when sellers choose to experiment new strategies more often. For lower values of $\tau \leq 0.05$ with a majority of cases with a unique price, the reward updating coefficient seem not to impact on the probability of convergence to a unique price.

[26] We performed a Fisher-Snedecor test of mean equality on our dataset. Tests report that for a low temperature coefficient, this convergence is achieved by fixing a unique price whereas for a higher temperature, price dispersion may still persist.

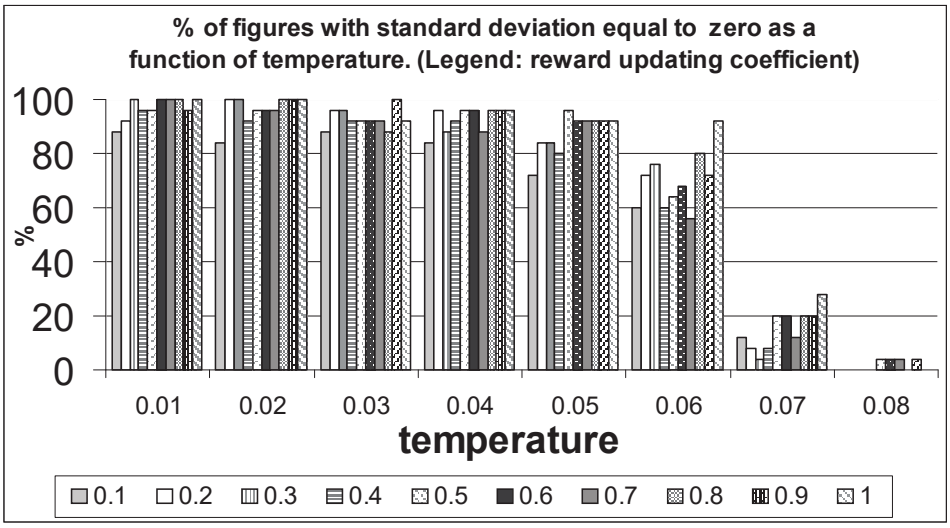

Fig. 15. *Proportion of cases exhibiting a convergence to a unique price as a function of the temperature and the reward updating parameter in the case (a,k)= (0.9;19) (25 simulations for each parameter configuration (temperature τ , reward updating parameter α))*

6. Conclusion

The aim of this paper was to compare Nash to two extreme forms of adaptive (reinforcement) learning: individual learning and social learning. We used an extension of the simple stylized market of Varian (1980). In that context, Individual Learning and Nash lead to the same *qualitative* conclusions with respect to a change in the consumers' information technology (proportion of informed consumers, number of visits made by informed consumers), despite the fact that the two distributions do not converge in a quantitative sense. On the contrary, for identical learning parameters, social learning leads to more contrasted results. Again, the price distributions do not coincide. Further, some of the qualitative predictions of the Nash equilibrium no longer hold, especially when the fraction of informed sellers is large. In that case, Social Learning generates both path dependency and uniqueness of posted prices. We provide some explanations to account for these differences. Some may be due to the changing nature of the competition process between adaptive sellers. Others are due to the lack of enough experimentations in sellers' behaviors. Increasing sellers' explorative behavior may then contribute to decreasing the lock-in describing sellers' price selection strategy. However, everything being equal i.e. with the same parameter set (exploration-exploitation parameter in particular), such a kind of path dependency did not appear with Individual Learning only.

Although the Nash equilibrium concept is highly demanding (concerning its requirements on rationality and information), its predictive capacity is well confirmed as far as *Individual Learning* is concerned. This may not be the case when considering Social Learning since conclusions are more dependent on some parameters values. These relate either to buyers' behaviors (that shape the type of competition on the market, and the sellers' rewards structure) or to sellers' behaviors

(exploration-exploitation trade-off in particular).Yet in another context (search market game, price competition), we here retrieve some results of (Vriend, 2000) that pointed out some differences between individual and social learning in a Cournot-game. While we did not attribute the explanation for the difference between the two types of learning to some "spite effect", our paper reinforces Vriend's by illustrating how two forms of learning models may lead to different market outcomes. There are several possible extensions to this work. An immediate one is to analyze more precisely the effect of varying "degrees" of social learning (i.e. various λ parameters) and further to consider heterogeneous coefficients among sellers, making some sellers more "receptive" to social influences than others. Another direction is to consider a less perfect type of social learning. Despite the fact that prices played by other sellers can always be perfectly observed, profits cannot. Hence, one could consider noisy observations of profits levels instead of perfect observations. Finally, as already pointed in the literature, the choice between Individual and Social Learning in this kind of game remains necessarily ad hoc. In that sense, we still lack some experimental evidence about how players decide in a market game. Hence, the appropriate learning assumption is still an open debate which of course may be highly dependent on the context or structure of the market game.

This work was launched during the ELICCIR project (CNRS support "Systèmes Complexes en SHS"). We would like to thank Olivier Bruno, and two anonymous referees for their detailed and careful comments on a previous version of this article. The usual caveat applies.

References

Altavilla C., Luini L. and Sbriglia P., forth. Social Learning in Market Games. *To appear in the Journal of Economic Behavior and Organization*,forth.

Benaïm M., Hofbauer J. and Hopkins E., 2005. Learning in Games with Unstable Equilibria. *Unpublished manuscript*,2005.

Bikhchandani S., Hirshleifer D. and Welch I., 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *The Journal of Political Economy*, 100(5),1992, 992-1026.

Brynjolfsson E. and Smith M., 2000. Frictionless Commerce? A Comparison of Internet and Conventional Retailers. *Management Science*, 46(4),2000, 563-585.

De Palma A. and Thisse J.-F., 1989. Les modèles de choix discrets. (14),1989, 151-191.

Erev I. and Roth A. E., 1998. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88(4),1998, 848-881.

Gale D., 1996. What have we learned from social learning? *European Economic Review*, 40(3-5),1996, 617-628.

Hopkins E. and Seymour R., 2002. The stability of price dispersion under seller and consumer learning. *International Economic Review*, 19(1),2002, 46-76.

Kirman A. P. and Vriend N. J., 2001. Evolving Market Structure: An ACE Model of Price Dispersion and Loyalty. *Journal of Economic Dynamics and Control*, 25(3-4),2001, 459-502.

Stahl D. O., 1989. Oligopolistic Pricing with Sequential Consumer Search. *American Economic Review*, 79,1989, 700-712.

Stigler G. J., 1969. The economics of Information. *Journal of Political Economy*, 69,1969.

Sutton R. G., 1991, Reinforcement Learning. Cambridge University Press.

Varian H., 1980. A Model of Sales. *American Economic Review*, 70,1980, 651-659.

Vriend N. J., 2000. An Illustration of the Essential Difference between Individual and Social Learning, and Its Consequences for Computational Analyses. *Journal of Economic Dynamics and Control*, 24(1),2000, 1-19.

Waldeck R., forth. Search and Price Competition. *Journal of Economic Behavior and Organization*,forth.

Waldeck R. and Darmon E., 2006. Can Boundedly Rational Sellers Learn To Play Nash? *Forthcoming in the Journal of Economic Interaction and Organisation*,2006.

Waldeck R. and Darmon E., forth. Can Boundedly Rational Sellers Learn To Play Nash? *Journal of Economic Interaction and Organisation*(forthcoming),forth.